This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Repurposing the Image Generative Potential: Exploiting GANs to Grade Diabetic Retinopathy

Isabella Poles[1], Eleonora D'Arnese[1], Luca G. Cellamare[1], Marco D. Santambrogio[1], Darvin Yi[2]
[1] Politecnico di Milano, Italy [2] University of Illinois at Chicago, USA
{isabella.poles, eleonora.darnese, marco.santambrogio}@polimi.it
lucagiuseppe.cellamare@mail.polimi.it dyi9@uic.edu

Abstract

Diabetic Retinopathy (DR) is a common cause of irreversible vision loss in the working-age population. Automatic DR grading allows ophthalmologists to provide timely treatment to numerous patients. However, developing a robust grading model needs large, balanced, and annotated data, which poses challenges in the collection. Moreover, data augmentation often fails to generate diverse data, necessitating alternative approaches such as Generative Adversarial Networks (GANs). However, GANs often operate with low-resolution images as a result of their costly training process. Therefore, we present a novel method that repurposes the discriminator of an unconditional Progressive GAN, leveraging the generative knowledge gained for DR grading. Furthermore, a new Log-Likelihood Inception Distance (LLID) metric estimates the similarity between one synthesized and a set of real images, thereby capturing human judgment more effectively. Our method is validated through extensive experiments on three public datasets, outperforming the baseline classifiers' performance by 12.5% and 14.33% average accuracy on small data regimes and when combined with state-of-the-art methods on large datasets, respectively. Additionally, LLID reproduces the comprehension ability of most of our Visual Turing Test participants, enabling differentiation between a synthesized image and a set of reference images with 82.88% accuracy. This confirms the quality of generated images and the metric consistency with human decision-making mechanisms.

1. Introduction

Diabetic Retinopathy (DR) is a severe complication of diabetes mellitus and a leading cause of blindness in working-age adults worldwide [38]. The international protocol categorizes DR progression into normal, mild, moderate, severe Non-Proliferative DR (NPDR), and PDR levels [33]. Fundus images captured by non-mydriatic colorful retinal cameras are commonly used to provide clinicians essential characteristics for grading DR [13]. However, early DR symptoms and subtle differences between consecutive stages pose challenges for accurate diagnosis [20]. Moreover, manual characterization of multiple patients is inefficient and prone to fatigue and misdiagnosis in the long run.

Recently, a significant focus has been posed on developing automatic grading models for DR [14]. Deep Learning (DL) has been a promising alternative to traditional machine learning approaches and handcrafted feature extraction [29]. However, training Convolutional Neural Networks (CNNs) and transformers often requires a large and diverse dataset. So, transfer learning has emerged as an effective solution to this challenge by leveraging knowledge distilled from other datasets [15]. Unfortunately, it typically assumes labeled and highly balanced source retinal data, which may not always be the case [8]. Nevertheless, common data augmentation techniques limit improving model generalization due to the augmented dataset's similarity to the original one [2].

In this context, generative DL models, such as Generative Adversarial Networks (GANs) [12], offer a solution to the scarcity of fundus retina data. Starting from a latent vector z, GANs can generate synthetic data $G(z)$ following the probability distributions of real samples ($G(z) \sim P_{act}(x)$). Image generation is achieved by optimizing in a *min-max* game a generative G and a discriminative D network that differentiates between real $D(x)$ and synthesized $D(G(z))$ data. GANs have successfully generated diverse retinal fundus images, incorporating vessel tree semantic information to control image realism and DR severity [5, 39]. *Coyner et al.* addressed the Vanilla GANs limitations in handling high-resolution images with a Progressive GAN (ProGAN) and in evaluating generated image quality with the Euclidean distance [7]. While *Noguchi et al.* achieved higher generated retinal fundus images Fréchet Inception Distance (FID) by leveraging a pre-trained generator [24]. However, the knowledge repurposing of the individual GAN architecture components to address new tasks has not been deeply investigated [31, 37]. To exemplify, the D model can learn and preserve real-world image representations while G is synthesizing new ones. However, exploiting it as an image classification network to take advantage of the knowledge it has gained has never been analyzed. Moreover, the quality evaluation of generated images poses significant challenges. While FID has been widely adopted due to its sensitivity to small distributional differences, it requires comparing two large image sets, which can be inconsistent with human inspection mechanisms that assess image realism by comparing each image with a set of already-seen images [3].

To address these issues, we propose a methodology to leverage the knowledge of the discriminator of a ProGAN [18] trained to generate high-resolution fundus images, fine-tuning it to grade DR in small-data regimes (Figure 1). Additionally, we introduce a novel metric to estimate the similarity between each GAN-synthesized image and a set of real images, providing a more accurate correlation between human visual perception and metrics for evaluating the quality of GAN-generated samples, as well as more efficient image quality evaluation in clinical practice. Furthermore, we allow accurate DR grading even in high-data regimes, ensembling our strategy with selected state-of-the-art models by an unweighted pooling function.

The main contributions of this paper are:

- An efficient strategy to repurpose a pre-trained unconditional GAN fundus image discriminator (Section 3.1) and enable the generalizability of its architecture for image classification tasks in low data regimes (Section 3.3).
- A novel quantitative metric, called Log-Likelihood Inception Distance (LLID), for evaluating GAN performance that captures the knowledge of human judgment by comparing the similarity between each synthesized image and a set of real images (Section 3.2).
- An ensemble learning technique for DR grading that combines the proposed repurposing strategy with CNN models to grade DR in different data regimes (Section 3.3).

2. Related Work

This study encloses two key research areas: DL for DR grading and fundus image generation. We provide an in-depth overview of relevant research in each domain.

Deep Learning for DR grading. In recent years, DL has emerged as a promising solution to address the limited efficiency and generalization posed by manual feature extraction from fundus images to grade DR [25]. *Gayathri et al.* used CNNs to automate feature extraction while machine learning classifiers to grade DR [10]. Image enhancement and classification have been combined in a three-branch neural network to perform grading reliably, also with low-quality images [16]. Furthermore, attention modules have been employed to focus on essential parts of the retinal image to improve the CNN-based detection performance [1]. Given the time-consuming and economically expensive retinal image labeling, transfer learning has emerged to relax the DL need for high-dimensional annotated datasets. In this context, *Li et al.* employed a fine-tuned CNN to extract features from retinal fundus images while using support vector machines for grading a small DR dataset [22]. Similarly, *Zhang et al.* combined five pre-trained CNNs for feature vector representation of DR images but trained an ensemble classifier on top of these CNNs to grade DR [35]. Finally, *Tymchenko et al.* adopted transfer learning in a three-headed ensemble CNN, achieving superior results by different model ensembles and trimmed mean predictions from five-ary DR fundus image classification [30]. Although these advancements enhance DR grading accuracy and efficiency, reducing huge dataset needs, working on a diverse data regime remains a firm requirement.

Deep Learning for Fundus Image Generation. GAN and diffusion models can tackle the scarcity of large and diverse annotated fundus image datasets. Early two-step approaches generated vessel masks followed by GAN fundus image one [6, 36]. Subsequently, DR-GAN incorporated multi-scale spatial and channel attention modules to improve the DR-related lesions FID in generated images [39], while *Odena et al.* extended the GAN framework for simultaneous generation and classification [26]. More recently, diffusion models have gained attention due to their comparable, if not better, generation and training stability than GAN. *Sojung et al.* used diffusion models for fundus image generation with artery/vein masks for vessel segmentation and classification [11], while DiffMIC used a novel conditional guidance and regularizations for image denoising and classification [34]. However, they are less effective than GANs in generating semantically meaningful latent representations, thus sometimes requiring conditioning during the generation. In addition, they provide less privacy and maintain GANs faster at predicting images since they must perform multiple inference steps to generate only one sample. Given the computational burden of training these models from scratch, transfer learning has primarily focused on fine-tuning pre-trained GANs on new datasets [24, 32]. Differently, *Salimans et al.* proposed a semi-supervised classifier by incorporating GAN-generated samples during training. However, comprehensive exploration of GAN pre-training remains limited, especially for fundus image discriminative purposes [27]. To exemplify, although *Vineeta et al.* developed a classifier for age-related macular degeneration using adversarial training as a regularization term, they did not investigate its use as a pre-training step nor for DR grading applications [9].

We have identified three key issues requiring attention. *Firstly*, DR grading relies on supervised learning, requiring many labeled samples not always readily available for fun-

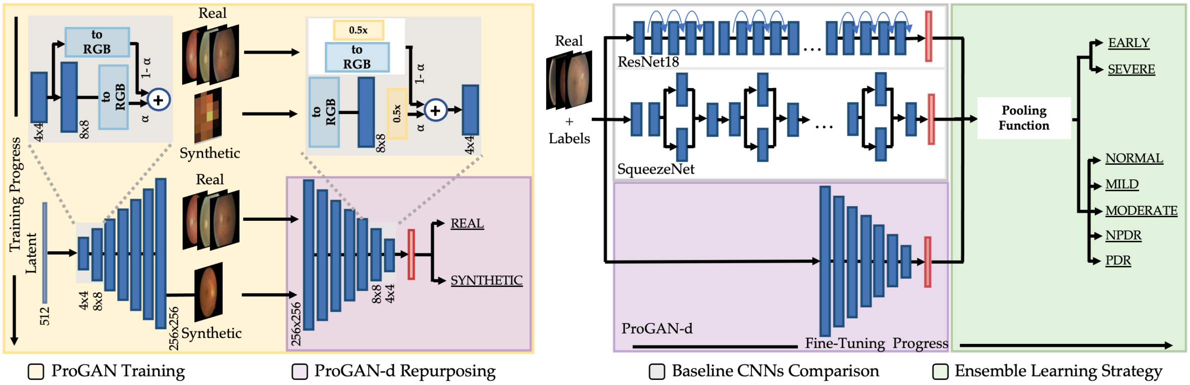

Figure 1. High-level description of the proposed methodology from the ProGAN training, the ProGAN discriminator repurposing (ProGAN-d), the ResNet18 and SqueezeNet baselines comparison, to the ensemble binary and five-ary classification.

dus images. *Secondly*, previous research has overlooked the potential of using a pre-trained GAN discriminator to discern image features. *Finally*, while FID is widely adopted in GANs evaluation, concerns exist in its consistency with human inspection. To the best of our knowledge, our work is the first and only attempt to propose a **DR grading method** that **leverages the classification knowledge acquired by the discriminator of a pre-trained GAN during unconditional image generation** and introduces a **novel inception-based distance metric** to estimate the similarity between each synthesized image and a set of real images.

3. Method

This Section describes the DL architecture and the loss function for image generation. It details the novel metric to evaluate image-generated quality. Finally, it describes the proposed repurposing strategy for DR grading.

3.1. Fundus Image Generation

This work proposes a more lightweight version of the original ProGAN, exploiting the unconditional synthesizing and learning ability to alleviate DL models' annotated datasets' constraints and distinguish large structures with fine details.

DL model architecture. The ProGAN generator model uses a 512-element latent vector of Gaussian random noise as a starting point. Subsequently, blocks of 3×3 convolutional layers, pixel normalization, LeakyReLU activation function (slope of 0.2), and 1×1 convolution mapping the RGB image are progressively added until the target image dimension of 256×256 is reached. During the progressive resolution increase, the output of each new layer is combined with the output of the previous one, upsampled by nearest neighbor interpolation to the current higher resolution. Furthermore, when the resolution transition happens, the new layers are smoothly faded to prevent the previous ones from a sudden transition, as shown in Fig.1, where α is the fading control parameter linearly interpolated over multiple training iterations. Subsequently, the discriminator model reverse engineers how the generator network was built. It starts with an RGB image that passes through convolutional layers using average pooling as downsampling. The current downsampled image and the previous are weighted, starting with a full weighting for the downsampled raw input and linearly transitioning to a full weighting for the interpreted output of the new input layer block. Downsampling convolutional layers are added until a single output is reached, where a sigmoid function determines whether the generated fundus image resembles a fake or real one. Finally, the original ProGAN is modified to obtain its lower-capacity ProGAN version: the convolution layer feature maps are halved at the 16×16 resolution and divided by 4 in subsequent ones. In contrast, the target resolution is set to 256×256 considering computational constraints.

Loss function. We trained our ProGAN by minimizing a generator and a discriminator loss. The adversarial loss defines the generator loss as:

$$L_{adv} = \mathbb{E}_{z \sim P(z)} \left[logD \left(G \left(z \right) \right) \right] \,, \tag{1}$$

where $D(G(z))$ is the discriminator's evaluation for the noise z from the generator, and $P(z)$ is the noise distribution. The discriminator loss is defined as the sum of an adversarial loss, a gradient penalty loss, and a drift loss:

$$L_D = L_{adv} + L_{GP} + L_{drift} \,. \tag{2}$$

In particular, L_{adv} identifies the adversarial loss consisting of the Wasserstein loss that measures the distance between the distributions of the real and synthesized images:

$$\begin{aligned} L_{adv} = {} & \mathbb{E}_{x \sim P_{act}(x)} \left[log \left(D \left(x \right) \right) \right] \\ & + \mathbb{E}_{z \sim P(z)} \left[log \left(1 - D \left(G \left(z \right) \right) \right) \right] \,, \end{aligned} \tag{3}$$

where $\mathbb{E}_{x \sim P_{act}(x)}$, and $\mathbb{E}_{z \sim P(z)}$ are the expected values over real data instances and random inputs to the generator, respectively. The L_{GP} loss of Equation (2) allows for more stable training enforcing gradients of the discriminator output with respect to the inputs to have unitary norm:

$$L_{GP} = \lambda \cdot \mathbb{E}_{\tilde{x}\sim P(\tilde{x})}\left[\left(\left\|\nabla_{\tilde{x}} D\left(\tilde{x}\right)\right\|_2 - 1\right)^2\right] , \tag{4}$$

where $\tilde{x}$ is a random interpolation of the real x and a generated $G(z)$ images, while λ is set to 1. Finally, the L_{drift} drift loss term of Equation (2) can be expressed as follows:

$$L_{drift} = \mathbb{E}_{x\sim P_{act}(x)}\left[D\left(x\right)^2\right] + \mathbb{E}_{z\sim P(z)}\left[D\left(G\left(z\right)\right)^2\right] , \tag{5}$$

moves the discriminator output far from zero.

3.2. Evaluating GAN Generated Fundus Image Quality

To meet the requirements of novelty and clinical plausibility of the GAN-generated samples, a Visual Turing Test (VTT) was designed, and the FID metric was extended for better alignment with human decision-making mechanisms.

Visual Turing Test. We conducted a VTT to explore visually coherent synthetic images, which are difficult for experts to classify as synthetic without prior knowledge. Participants, including five expert ophthalmologists and 15 different ophthalmology researchers with a minimum of five and two years of clinical experience, respectively, and 32 laypeople, participated using a web-based quiz platform. The platform presented images with a 50% chance of being real or synthetic, and participants were classified as “Real” or “Synthetic”. The dataset comprised images generated by our ProGAN unconditionally trained on images representing all DR degrees. However, during the VTT, the participants were not required to grade DR since it would have introduced too much complexity.

To assess participant performance, we computed correct and wrong classification percentages for images categorized as real and synthetic, respectively. In particular, we analyzed the ability of each of the participants to correctly discriminate between “Real” by the True Negative (TN) and “Synthetic” by the True Positive (TP) rates. This analysis provides insights into the perceptual realism of synthetic images.

LLID: A Novel GAN Quality Evaluation Metric. The FID is a widely used quality metric for evaluating the fidelity of GAN-generated images compared to real images. It quantitatively measures the similarity between the two distinct distributions of the inception embeddings of the real $\mathbf{r}$ and generated $\mathbf{g}$ image sets, obtained as activations from the penultimate layer of an Inception-V3 network. The two image set distributions are mathematically represented as multi-dimensional Gaussians characterized by mean $\mathbf{m}$ and covariance $\mathbf{C}$ parameters. The FID is computed as the distance between these two Gaussian distributions:

$$FID = \left\|\mathbf{m_r} - \mathbf{m_g}\right\|^2 + Tr\left(\mathbf{C_r} + \mathbf{C_g} - 2\sqrt{\mathbf{C_r}\mathbf{C_g}}\right) . \tag{6}$$

To enhance the capacity of generative models to replicate human decision-making processes that consist of comparing a new single image with the prior knowledge of a set of known samples, we propose the Log-Likelihood Inception Distance (LLID) metric. Indeed, unlike FID, which replicates an operator struggling to discern differences between two sets of images, LLID allows for comparing an individual image with a reference image collection. This aligns more closely with real-world scenarios where an operator is more easily tasked when distinguishing between an actual or synthesized image presented during a VTT with the prior knowledge of a set of images than between a set of actual or synthesized images and the set of images constituting their prior knowledge. The LLID metric quantifies the logarithm of the likelihood of a given sample under the distribution of inception-extracted features within the reference image collection:

$$LLID = -log\left(\mathcal{N}\left(\mathbf{m_r}, \mathbf{C_r}\right) \cdot f\left(G\left(z\right)\right)\right) , \tag{7}$$

where $\mathcal{N}\left(\mathbf{m_r}, \mathbf{C_r}\right)$ is the multivariate normal distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$, while $f\left(G\left(z\right)\right)$ is the feature representation of the generated sample obtained by passing it through the Inception-V3 network. The LLID scores are computed for each sample from the VTT using a reference real distribution estimated from $10k$ fundus images used to train the proposed ProGAN. After normalizing the LLID scores between 0 and 1, real samples are expected to have higher likelihoods under the real distribution than fake samples. However, it should be noted that a low LLID score under the real distribution does not necessarily indicate a bad or unrealistic image. For this reason, the LLID scores were used to create an *Inception-based classifier* using a threshold of 0.5. Samples with scores above the threshold were classified as real, while those below the threshold were classified as fake. Finally, FID represents a distance metric, while LLID is a likelihood one. Therefore, lower FID but higher LLID scores indicate better image quality and closer resemblance to real images regarding visual quality, diversity, and realism.

3.3. Discriminator Repurposing for DR Grading

Our premise is based on the observation that GANs can generate meaningful intermediate images by interpolating different classes, suggesting that the learned image features exist on a manifold where new classes can also reside [12]. Therefore, if the discriminator learns these features while classifying real and synthesized images, it can transfer this knowledge for grading real DR images. To leverage this, we repurpose the pre-trained ProGAN discriminator (ProGAN-d) for binary and five-ary DR fundus image classification in small-data regimes. Moreover, we employ an ensemble learning strategy to enhance model performance.

DL model architecture. The DR classifier employs the same network architecture as ProGAN-d (Section 3.1). In the case of binary classification tasks and image generation, no further modifications are required for the last classifier layer. However, during five-ary DR classification, the final ProGAN-d sigmoid layer is replaced with a softmax layer capable of handling multi-class classification tasks. During the fine-tuning step, the weights of the ProGAN-d are frozen, and only the final linear layer classifier is trained using the pre-trained backbone parameters.

Loss function. Unlike the original ProGAN training, we let our ProGAN-d learn by minimizing the Cross-Entropy (CE) loss function due to its better suitability for learning a pure classification task instead of an adversarial one. In particular, we define it as $CE(P^*|P) = \sum_{i=1}^{n} P^*(i) \cdot log\, P(i)$, where P^* and P are the predicted and true class distribution, while i refers to the image class, and $n = 2$ or $n = 5$ in the case of binary or five-ary image classification.

Model Learning Strategy. An ensemble learning strategy combining predictions from multiple models was used to allow the end-user to take advantage of the low-capacity ProGAN-d even in high-data regimes, whether required. Firstly, the ProGAN discriminator was repurposed, and the pre-trained ResNet18 and SqueezeNet models, trained on ImageNet, were fine-tuned. Secondly, the output probabilities predicted by each model were combined using the $\bar{y} = \frac{1}{3} \sum_{i=0}^{2} y_i$ unweighted pooling function, where i represents the ProGAN-d, SqueezeNet, and ResNet18 models ($i \in [0, 2]$), y_i denote their respective output probabilities. At the same time, equal weight was given to each prediction. The selection of SqueezeNet and ResNet18 was based on their comparable and lower number of learnable parameters to the ProGAN-d one and baselines.

4. Experiments

This Section presents the datasets employed and the results from the generation and classification experiments. The experiments employ PyTorch (1.12.1). The training was run on a 48 GB RAM PNY NVIDIA Quadro RTX 6000 GPU, with Adam and a learning rate of $3 \cdot 10^{-3}$. A decay rate from 0 to 0.99 for averaging gradients leads the learning rate to $1 \cdot 10^{-8}$ as an asymptotic value. The ProGAN was trained with mini-batch discrimination, equalized learning rate, random image cropping, and horizontal flip as data augmentation.

4.1. Datasets Description and Pre-processing

Our method utilizes three public fundus image datasets: EyePACS [8], APTOS 2019 (APTOS19) [19], and DDR [21]. All dataset images have been graded by a clinician on a 0-4 scale based on the Early Treatment Diabetic Retinopathy Study scale [28]. However, they have diverse sizes and suffer from class imbalance issues. In particular, EyePACS, consisting of 88,702 RGB fundus images, shows that the first less severe DR degrees account for the higher number of images being the 73.48%, 6.96%, and 15.07% of the dataset, while the most severe remaining classes are the less frequent being the 2.48% and 2.01% of the total dataset. On the other hand, APTOS19 and DDR, accounting for 3,662 and 12,522 RGB fundus images, show as the most populated DR severity classes the first and the third, followed by the mild and PDR, respectively, with the two remaining classes covering 13.3% and 6.91% of the two datasets, respectively. We leverage EyePACS to train the generation and classification models, while the APTOS19 and DDR prove our strategy's generalizability during image classification. We utilize 35,126 EyePACS samples for GAN training. 70/20/10 images were picked randomly from all the datasets for training/validation/testing during ProGAN and DR grading tasks, maintaining the original label imbalance during training while balancing the test sets. Furthermore, since the images come from various sources, resulting in variations in lighting conditions and resolutions, we normalize each channel by subtracting the 0.5 mean and dividing by the 0.5 standard deviation and resize them into 256×256 given the 48GB GPU memory, accelerating the model convergence.

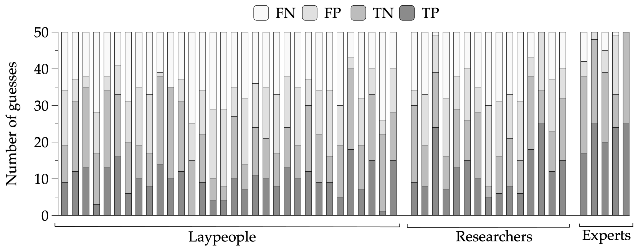

Figure 2. False Negative (FN), False Positive (FP), True Negative (TN), and True Positive (TP) for each of the laypeople, researcher, and expert ophthalmologist VTT participants.

4.2. GAN Fundus Image Generation Results

We evaluated our ProGAN fundus image generation on EyePACS using the VTT in Section 3.2. We then compared the results with five other generative methods using the FID.

Visual Turing Test Evaluation. The results in Figure 2 show the participants' accuracy in correctly discerning real from synthesized images using 50 samples during VTT. Our real and ProGAN-generated images achieved a recognition rate of 51.1%, confirming their remarkable fidelity, considering the theoretical random guess of $\sim 50\%$. However, the classification performance varied among real and synthetic

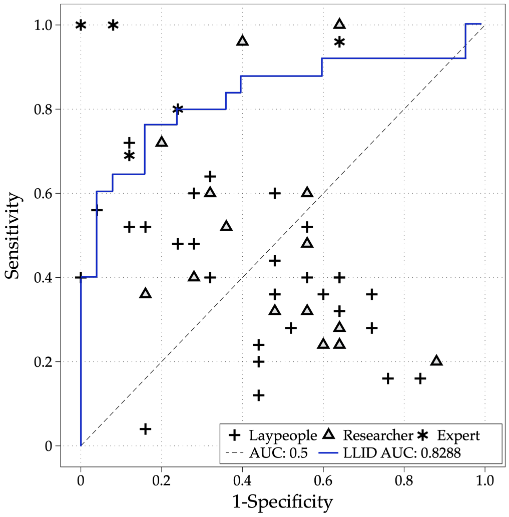

Figure 3. Comparison between ROC curve and VTT results.

images (Figure 3). The average sensitivity and specificity were 45.77% and 57.38%, respectively. In this context, it is visible that expert and researchers ophthalmologists had significantly higher correct response rates than laypeople, indicating better sensitivity and specificity.

Table 1. FID and LLID comparing our method and baselines. ‘†’ and ‘‡’ respectively denote the [4], [39] results.

The best performances are underlined.

	Model	FID	LLID
baseline	CGAN [23]	$19.45^{\dagger}$	0.72
baseline	Pix2Pix [17]	$15.24^{\dagger}$	0.83
baseline	Tub-sGAN [36]	$9.67^{\dagger}$	-
baseline	RF-GAN2 [4]	$7.03^{\dagger}$	-
baseline	DR-GAN[39]	$4.53^{\ddagger}$	-
ours	**ProGAN**	**4.06**	**0.92**

Baselines Performance Comparison. Table 1 presents a comparative performance analysis of our proposed ProGAN model against CGAN [23], Pix2Pix [17], Tub-sGAN [36], RF-GAN2 [4], and DR-GAN [39]. The FID values respectively denote the [4], [39] results since Tub-sGAN, RF-GAN2, and DR-GAN come close-source, while the LLID metric has been computed training CGAN, Pix2Pix, and our ProGAN. Specifically, our ProGAN demonstrates superior performance, achieving a FID score of 4.06, 5× lower than the FID score of 19.45 obtained by the CGAN model. The CGAN's lower performance can be attributed to the imbalance in conditioned information, which results in the generator prioritizing dominant conditions while neglecting minority ones. In contrast, our ProGAN exhibits a marginal improvement of only 0.47 FID compared to DR-GAN. This slight enhancement can be attributed to the improved control and preservation of input information offered by DR-GAN reconstruction loss with respect to the CGAN. These findings indicate that the high-resolution image regime in which our ProGAN approach lives allows for generating highly accurate retinal fundus images with FID scores comparable to state-of-the-art methods. Although the FID measurement provides valuable insights into the visual quality of the generated images, it is questionable how it aligns with human judgment since it compares two sets of images rather than one with respect to a set. For this reason, we decorated the FID results of our ProGAN with the LLID ones first to demonstrate the LLID ability to measure image quality. More in detail, it is visible how the resulting LLID trend aligns with the FID one since both Pix2Pix and our ProGAN show higher LLID performances than the CGAN, while our model confirms the most accurate in the generation process with 0.92 as LLID. Subsequently, to evaluate the LLID capability of resembling human judgment, we implemented and exploited an *Inception-based classifier*, as described in Section 3.2. Figure 3 shows this classification method's Receiver Operator Characteristic (ROC) curve compared to the performance of the VTT participants. The Area Under the Curve (AUC) of the LLID classifier equals 82.88%, which is significantly higher than random guessing. Also, it is clear how our LLID classifier outperformed most of the laypeople participants and the ophthalmology researchers. Conversely, all the expert ophthalmologists matched or outperformed the LLID classifier. This significant result demonstrates how the novel-implemented quality distance metric can reproduce the comprehension ability of most participants to distinguish a synthesized image from a set of real images. Therefore, it could help physicians or researchers to gain more precise insights into the generation quality of individual images rather than being constrained to the quality of two sets of images. Finally, the *Inception-based classifier* further confirms how our ProGAN can generate accurate images since they are enough to trick a person with some experience with retinal fundus imaging and be considered a resource for training ophthalmologists in the fundus image analysis practice.

4.3. DR Grading Results

We evaluate whether using the proposed LLID metric to assess the ProGAN in subsequent classification tasks yields better classification results. Then, we evaluate the efficacy of the repurposing strategy by computing the accuracy of our ProGAN-d to binary and five-ary classify the different DR degrees and by comparing its performances with four shallow and deep state-of-the-art networks, namely SqueezeNet, ResNet18, InceptionV3, and VGG11. In particular, we combined the lowest (0, 1) and the severer (2, 3, 4) DR degrees for the binary task, while for the multiclass case, we consider the five classes singularly. Finally, we show the ensemble learning strategy results de-

Table 2. Ablation Study: LLID values at different training epochs and relative best binary and five-ary classification results. Best results underlined.

#Epochs	LLID	Best Binary Accuracy	Best Five-ary Accuracy
$100k$	0.74	0.56	0.23
$250k$	0.83	0.58	0.21
$400k$	0.87	0.58	0.41
$550k$	0.89	0.59	0.41
$700k$	$\underline{0.92}$	$\underline{0.61}$	$\underline{0.46}$

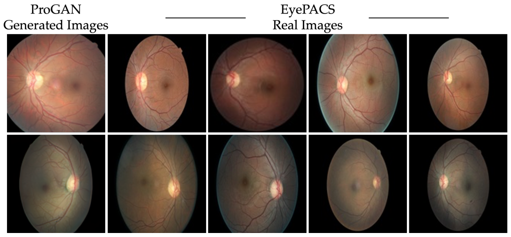

Figure 4. Examples of $700k^{th}$ epoch generated and real images.

scribed in Section 3.3, referring to $\mathcal{E}_i$ where $i \in [1, 4]$ when weighting together the output probabilities of the ProGAN-d and ResNet ($i = 1$), ProGAN-d and SqueezeNet ($i = 2$), ResNet and SqueezeNet ($i = 3$), and all the three models ($i = 4$). We discarded the VGG11 and the InceptionV3 models from the ensembling due to the higher number of training parameters than our ProGAN-d, which could have led them to hide our ProGAN-d model contribution. All classification experiments have been performed retraining all the models on small-data (100, 300 images) and high-data ($1k$, $3k$ and $10k$ images) regimes considering the random and ImageNet weights initialization for the baselines, the random and EyePACS weights initialization for the ProGAN-d, while just the ImageNet and EyePACS weights initialization for the ensembles models.

LLID-based Ablation Study. Table 2 demonstrates how the proposed LLID correlates with the repurposed classifier. In particular, the accuracy and LLID results reported show how letting our ProGAN generate progressively real-quality images as the training progresses from the $100k^{th}$ to the last $700k^{th}$ epoch (Figure 4) brings our ProGAN-d models to more accurately classify real fundus images. Indeed, it is visible how the LLID of the generated images increases from 0.74 to 0.92 as the ProGAN training progresses from $100k$ to $700k$ epochs. The image quality improvement is reflected in the DR classification performances, which show 5% and 23% accuracy improvements, respectively, considering the best binary and five-ary classification results achieved through all the training image set sizes. Given that the model with higher LLID is also the one achieving the best classification performance when repurposed, our study supports the possibility of employing the novel LLID not only as a metric for generation goodness but also as a metric to identify the best model for repurposing.

Binary Fundus Image Classification. Table 3 (a) presents a comparative analysis of the performance obtained in binary fundus image classification while progressively increasing the size of the fine-tuning dataset. Upon initial observation, it shows that utilizing an ImageNet and EyePACS pre-trained network positively impacts all models. It should be noted that the ImageNet weights were obtained by training the state-of-the-art models in a supervised manner, whereas the EyePACS weights were obtained through unsupervised training of the ProGAN network, eliminating the need for labeled data during training. Another noteworthy trend emerges when the training dataset size is increased from 100 to $10k$ images. In particular, our ProGAN-d model is the reference model between depth comparable state-of-the-art models when the number of available training images is between 100 and 300, given the highest 0.61 when fine-tuning our model on the EyePACS datasets. This result is confirmed when generalizing our model training to the DDR and APTOS19 datasets, which brings the highest accuracy of 0.77 and 0.67, respectively. While the training size increases, our ProGAN-d shows the slowest improvement rate when both pre-trained and fine-tuned on EyePACS. Indeed, when fine-tuning our model on the APTOS19 and DDR datasets, our EyePACS pre-trained weights find novel and different patterns of DR features to learn from rather than being stacked with already-embedded EyePACS DR global features. Furthermore, compared to state-of-the-art models, the smaller and domain-different EyePACS dataset than the ImageNet leads the EyePACS pre-trained weights to struggle while updating if the fine-tuning datasets show the same distribution of features and their size increase. Differently, the ImageNet pre-trained weights perform better when fine-tuned on a very large and domain-different dataset, given their better ability to generalize. This aspect is further highlighted when considering more complex baselines such as ResNet18 and VGG11. However, if higher performance would be needed while exploiting our ProGAN-d on high-data regimes, this issue can be mitigated by combining the prediction probabilities in the $\mathcal{E}_4$ ensemble model. As a result, up to 4.67%, 4%, and 2.33% accuracy improvements average between the three individual models are achieved respectively with $1k$, $3k$, and $10k$ fine-tuning images. Table 4 further confirms the robustness of the proposed method. The ProGAN-d model acts as a good contributor to the baseline ensemble, and each ensemble method consistently outperforms its ablated versions and original components across all training dataset sizes.

Five-ary Fundus Image Classification. The classification performance of the five degrees of DR is presented in

Table 3. Fundus image classification results for our methods and baselines on the pre-training and training datasets. R = Random, E = EyePACS, I = ImageNet, A = APTOS19, D = DDR, and "-"are due to the A and D low dimension. The best performances are underlined.

				(a) Binary DR Grading Accuracy					(b) Five-ary DR Grading Accuracy				
Model	Pre-Tr	Tr		100	300	1k	3k	10k	100	300	1k	3k	10k
ResNet18	R	E	baseline	0.49	0.52	0.56	0.57	0.58	0.28	0.31	0.35	0.36	0.38
SqueezeNet	R	E	baseline	0.51	0.54	0.58	0.60	0.61	0.30	0.40	0.42	0.44	0.45
InceptionV3	R	E	baseline	0.48	0.50	0.50	0.51	0.53	0.31	0.33	0.34	0.38	0.41
VGG11	R	E	baseline	0.56	0.58	0.59	0.61	0.69	0.34	0.32	0.37	0.43	0.45
ProGAN-d	R	E	ours	0.44	0.54	0.54	0.58	0.68	0.28	0.33	0.35	0.37	0.41
ResNet18	I	E	baseline	0.56	0.59	0.61	0.69	0.72	0.37	0.40	0.43	0.45	0.50
SqueezeNet	I	E	baseline	0.60	0.61	0.65	0.70	0.70	0.39	0.41	0.40	0.48	0.53
InceptionV3	I	E	baseline	0.50	0.51	0.53	0.56	0.59	0.31	0.32	0.36	0.40	0.44
VGG11	I	E	baseline	0.58	0.63	0.68	0.70	0.74	0.36	0.40	0.44	0.48	0.55
ProGAN-d	E	E	**ours**	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	**0.46**	**0.46**	**0.46**	**0.46**	**0.46**
Other Datasets													
ProGAN-d	E	A	**ours**	0.73	0.77	0.73	0.74	-	0.49	0.53	0.51	0.52	-
ProGAN-d	E	D	**ours**	0.60	0.67	0.72	0.74	0.80	0.48	0.49	0.50	0.45	0.44

Table 4. Ensemble results, where $\mathcal{E}_4$ = ProGAN-d + ResNet18 + SqueezeNet, $\mathcal{E}_3$ = ResNet18 + SqueezeNet, $\mathcal{E}_2$ = ProGAN-d + SqueezeNet, and $\mathcal{E}_1$ = ProGAN-d + ResNet18. The best performances are underlined.

	(a) Binary DR Grading Accuracy					(b) Five-ary DR Grading Accuracy				
Model	100	300	1k	3k	10k	100	300	1k	3k	10k
$\mathcal{E}_4$	0.62	0.68	0.72	0.74	0.75	0.40	0.55	0.45	0.49	0.52
$\mathcal{E}_3$	0.61	0.64	0.68	0.70	0.72	0.37	0.41	0.43	0.46	0.50
$\mathcal{E}_2$	0.61	0.65	0.68	0.71	0.74	0.46	0.46	0.48	0.50	0.55
$\mathcal{E}_1$	0.63	0.66	0.66	0.69	0.72	0.41	0.43	0.44	0.46	0.48

Table 3 (b). The results show similar trends to those of binary classification. Specifically, pre-trained weights improve classification accuracies compared to random weight initialization. Notably, our EyePACS fine-tuned ProGAN-d model consistently achieves the highest classification performance of 0.46 among all models when trained with dataset sizes ranging from 100 to $1k$ images. Furthermore, it demonstrates its compatibility even with the APTOS19 and DDR datasets since up to 0.53 and 0.50 average accuracy have been obtained, respectively. However, analyzing models' performances when trained with larger training datasets, our ProGAN-d model shows a constant or decreasing performance, unlike the baseline models, whose accuracy improves and plateaus when the dataset size reaches $10k$ training images. Similar to the observations made in binary DR grading, the constant behavior through all the dataset sizes is attributable to the challenging generalization ability during the fine-tuning of EyePACS pre-trained models, while the decreasing one is due to the progressively lower number and higher variance of the per-class images in the five-ary classification scenario. To allow a possible end-user to benefit from our model even in higher-data regimes, we combine the prediction probabilities of ResNet and SqueezeNet with those of ProGAN-d. This strategy's effectiveness is demonstrated by the highest accuracy improvements of 2%, 4%, 9% across all the baselines when fine-tuning $\mathcal{E}_2$ with $1k$, $3k$, and $10k$ images, respectively. Furthermore, Table 4 shows that ProGAN-d consistently improves all model predictions, as indicated by the superior performance of $\mathcal{E}_4$, $\mathcal{E}_2$, and $\mathcal{E}_1$ compared to $\mathcal{E}_3$, where only ResNet and SqueezeNet are used. Lastly, when comparing the performances of binary and five-ary DR grading, combining the DR degrees into two classes yields higher accuracy than individually classifying the five DR degrees. This outcome can be attributed to the higher complexity in distinguishing multiple degrees with potentially fine-grained differences compared to two classes with higher inter-class variance.

5. Discussion and Conclusion

This paper shows how repurposing a ProGAN discriminator allows for superiority over conventional DR grading methods, with 12.5% accuracy improvement averaged between binary and five-ary results on smaller datasets. Furthermore, it benefits the end-user with large datasets through ensembling, with 14.33% accuracy improvement averaged between binary and five-ary results. Finally, introducing a novel LLID metric correlates better with human visual quality perception than conventional metrics and classifiers.

Although such encouraging results and the potentiality of repurposing GANs, some considerations should be made. Indeed, during GAN training, the discriminator sees both true and generated data, leading to possible unrealistic features learning and bias introduction. However, we generated images nicely mimicking real data, suggesting they do not negatively impact the discriminator. Moreover, during repurposing, the discriminator only learns the pathology signs from real data, limiting the impact of biases or less realistic features. These results are not definitive, yet they offer valuable insights for future work seeking to exploit the ability of generative models in challenging classification tasks while scaling it to different anatomical districts.

References

[1] Chandranath Adak, Tejas Karkera, Soumi Chattopadhyay, and Muhammad Saqib. Detecting severity of diabetic retinopathy from fundus images using ensembled transformers. *arXiv preprint arXiv:2301.00973*, 2023. 2

[2] Teresa Araújo, Guilherme Aresta, Luís Mendonça, Susana Penas, Carolina Maia, Angela Carneiro, Ana Maria Mendonca, and Aurelio Campilho. Data augmentation for improving proliferative diabetic retinopathy detection in eye fundus images. *IEEE Access*, 8:182462–182474, 2020. 1

[3] Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022. 2

[4] Yu Chen, Jun Long, and Jifeng Guo. Rf-gans: a method to synthesize retinal fundus images based on generative adversarial network. *Computational intelligence and neuroscience*, 2021:1–17, 2021. 6

[5] Pedro Costa, Adrian Galdran, Maria Inês Meyer, Michael David Abramoff, Meindert Niemeijer, Ana Maria Mendonça, and Aurélio Campilho. Towards adversarial retinal image synthesis. *arXiv preprint arXiv:1701.08974*, 2017. 1

[6] Pedro Costa, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abràmoff, Ana Maria Mendonça, and Aurélio Campilho. End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791, 2017. 2

[7] Aaron S Coyner, Jimmy S Chen, Ken Chang, Praveer Singh, Susan Ostmo, RV Paul Chan, Michael F Chiang, Jayashree Kalpathy-Cramer, J Peter Campbell, et al. Synthetic medical images for robust, privacy-preserving training of artificial intelligence: Application to retinopathy of prematurity diagnosis. *Ophthalmology Science*, 2(2), 2022. 1

[8] Jorge Cuadros and George Bresnick. Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of diabetes science and technology*, 3(3):509–516, 2009. 1, 5

[9] Vineeta Das, Samarendra Dandapat, and Prabin Kumar Bora. A data-efficient approach for automated classification of oct images using generative adversarial network. *IEEE Sensors Letters*, 4(1):1–4, 2020. 2

[10] S Gayathri, Varun P Gopi, and Ponnusamy Palanisamy. A lightweight cnn for diabetic retinopathy classification from fundus images. *Biomedical Signal Processing and Control*, 62:102115, 2020. 2

[11] Sojung Go, Younghoon Ji, Sang Jun Park, and Soochahn Lee. Generation of structurally realistic retinal fundus images with diffusion models. *arXiv preprint arXiv:2305.06813*, 2023. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1, 4

[13] Andrzej Grzybowski, Piotr Brona, Gilbert Lim, Paisan Ruamviboonsuk, Gavin SW Tan, Michael Abramoff, and Daniel SW Ting. Artificial intelligence for diabetic retinopathy screening: a review. *Eye*, 34(3):451–460, 2020. 1

[14] Varun Gulshan, Renu P Rajan, Kasumi Widner, Derek Wu, Peter Wubbels, Tyler Rhodes, Kira Whitehouse, Marc Coram, Greg Corrado, Kim Ramasamy, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in india. *JAMA ophthalmology*, 137(9): 987–993, 2019. 1

[15] Misgina Tsighe Hagos and Shri Kant. Transfer learning based detection of diabetic retinopathy from small dataset. *arXiv preprint arXiv:1905.07203*, 2019. 1

[16] Qingshan Hou, Peng Cao, Liyu Jia, Leqi Chen, Jinzhu Yang, and Osmar R Zaiane. Image quality assessment guided collaborative learning of image enhancement and classification for diabetic retinopathy grading. *IEEE Journal of Biomedical and Health Informatics*, 2022. 2

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[19] M Karthick and D Sohier. Aptos 2019 blindness detection. *Kaggle https://kaggle. com/competitions/aptos2019-blindness-detection Go to reference in chapter*, 2019. 5

[20] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S Corrado, Lily Peng, and Dale R Webster. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125(8):1264–1272, 2018. 1

[21] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. 2019. 5

[22] Xiaogang Li, Tiantian Pang, Biao Xiong, Weixiang Liu, Ping Liang, and Tianfu Wang. Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification. In *2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)*, pages 1–11. IEEE, 2017. 2

[23] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 6

[24] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2750–2758, 2019. 1, 2

[25] Mads Fonager Nørgaard and Jakob Grauslund. Automated screening for diabetic retinopathy–a systematic review. *Ophthalmic research*, 60(1):9–17, 2018. 2

[26] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016. 2

[27] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 2

[28] Sharon D Solomon and Morton F Goldberg. Etdrs grading of diabetic retinopathy: still the gold standard? *Ophthalmic research*, 62(4), 2019. 5

[29] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021. 1

[30] Borys Tymchenko, Philip Marchenko, and Dmitry Spodarets. Deep learning approach to diabetic retinopathy detection. *arXiv preprint arXiv:2003.02261*, 2020. 2

[31] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. 2

[32] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: Generating images from limited data. In *European Conference on Computer Vision*, pages 220–236, 2018. 2

[33] Charles P Wilkinson, Frederick L Ferris III, Ronald E Klein, Paul P Lee, Carl David Agardh, Matthew Davis, Diana Dills, Anselm Kampik, R Pararajasegaram, Juan T Verdaguer, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–1682, 2003. 1

[34] Yijun Yang, Huazhu Fu, Angelica I Aviles-Rivero, Carola-Bibiane Schönlieb, and Lei Zhu. Diffmic: Dual-guidance diffusion network for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 95–105. Springer, 2023. 2

[35] Wei Zhang, Jie Zhong, Shijun Yang, Zhentao Gao, Junjie Hu, Yuanyuan Chen, and Zhang Yi. Automated identification and grading system of diabetic retinopathy using deep neural networks. *Knowledge-Based Systems*, 175:12–25, 2019. 2

[36] He Zhao, Huiqi Li, Sebastian Maurer-Stroh, and Li Cheng. Synthesizing retinal and neuronal images with generative adversarial nets. *Medical image analysis*, 49:14–26, 2018. 2, 6

[37] Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained gans for generation with limited data. In *International Conference on Machine Learning*, pages 11340–11351. PMLR, 2020. 2

[38] Yingfeng Zheng, Mingguang He, and Nathan Congdon. The worldwide epidemic of diabetic retinopathy. *Indian journal of ophthalmology*, 60(5):428, 2012. 1

[39] Yi Zhou, Boyang Wang, Xiaodong He, Shanshan Cui, and Ling Shao. Dr-gan: conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images. *IEEE Journal of Biomedical and Health Informatics*, 26(1):56–66, 2020. 1, 2, 6