AMERICAN ACADEMY OF OPHTHALMOLOGY®

Deepfakes in Ophthalmology

Applications and Realism of Synthetic Retinal Images from Generative Adversarial Networks

Jimmy S. Chen, MD,[1,] Aaron S. Coyner, PhD,[1,*] R.V. Paul Chan, MD,[2] M. Elizabeth Hartnett, MD,[3] Darius M. Moshfeghi, MD,[4] Leah A. Owen, MD, PhD,[3] Jayashree Kalpathy-Cramer, PhD,[5,6] Michael F. Chiang, MD, MA,[7] J. Peter Campbell, MD, MPH[1]*

Purpose: Generative adversarial networks (GANs) are deep learning (DL) models that can create and modify realistic-appearing synthetic images, or deepfakes, from real images. The purpose of our study was to evaluate the ability of experts to discern synthesized retinal fundus images from real fundus images and to review the current uses and limitations of GANs in ophthalmology.

Design: Development and expert evaluation of a GAN and an informal review of the literature.

Participants: A total of 4282 image pairs of fundus images and retinal vessel maps acquired from a multicenter ROP screening program.

Methods: Pix2Pix HD, a high-resolution GAN, was first trained and validated on fundus and vessel map image pairs and subsequently used to generate 880 images from a held-out test set. Fifty synthetic images from this test set and 50 different real images were presented to 4 expert ROP ophthalmologists using a custom online system for evaluation of whether the images were real or synthetic. Literature was reviewed on PubMed and Google Scholars using combinations of the terms *ophthalmology*, *GANs, generative adversarial networks, ophthalmology, images, deepfakes,* and *synthetic*. Ancestor search was performed to broaden results.

Main Outcome Measures: Expert ability to discern real versus synthetic images was evaluated using percent accuracy. Statistical significance was evaluated using a Fisher exact test, with P values ≤ 0.05 thresholded for significance.

Results: The expert majority correctly identified 59% of images as being real or synthetic ($P = 0.1$). Experts 1 to 4 correctly identified 54%, 58%, 49%, and 61% of images ($P = 0.505$, 0.158, 1.000, and 0.043, respectively). These results suggest that the majority of experts could not discern between real and synthetic images. Additionally, we identified 20 implementations of GANs in the ophthalmology literature, with applications in a variety of imaging modalities and ophthalmic diseases.

Conclusions: Generative adversarial networks can create synthetic fundus images that are indiscernible from real fundus images by expert ROP ophthalmologists. Synthetic images may improve dataset augmentation for DL, may be used in trainee education, and may have implications for patient privacy. *Ophthalmology Science 2021;1:100079* © *2021 Published by Elsevier Inc. on behalf of the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Image-based deep learning (DL) systems developed for ophthalmic diseases[1–7] have achieved diagnostic performance comparable to that of ophthalmologists, but require large amounts of training data. Moreover, it is essential to train on diverse datasets with heterogeneous features present in clinical populations to avoid biased performance in practice.[8,9] Development of these datasets typically requires sharing data across institutions, which can be limited by time, cost, legislation,[10] and privacy regulations.[11] Data- and model-sharing methods including federated[12,13] and distributed[14,15] learning have shown potential in facilitating DL algorithm training without inter-institutional data sharing. However, even if these approaches work as well as developing a multi-institutional dataset, they too may be time-consuming and costly to set up, and still may not provide adequate dataset size and heterogeneity, especially for rare diseases.

Generative adversarial networks (GANs) are DL-based models that can generate realistic-looking fake images, so-called deepfakes.[16] Deepfakes have garnered notoriety in the media for their nefarious applications,[17,18] but recently have been explored in multiple medical domains.[9,19–26] Since ophthalmology has been at the forefront of the DL revolution, there are numerous potential applications of synthetic images, starting with fundus[9,19,20] and OCT.[27–29] Synthetic images can be modified to adjust image features such as pigmentation,[9] image quality,[30] and even disease severity.[31] One of many potential applications is as an

https://doi.org/10.1016/j.xops.2021.100079
ISSN 2666-9145/21

alternative solution to increase the size and diversity of training datasets for DL algorithms.[32,33] However, the potential uses of GANs in ophthalmology remain underexplored, including the utility of creating fully synthetic image datasets, their applications for DL development and medical education, and implications for privacy laws and data sharing. The purpose of our study was 2-fold: (1) to evaluate whether clinicians could discern synthetic fundus images generated by a GAN from real fundus images acquired from a retinopathy of prematurity (ROP) screening program and (2) to review current uses and limitations of GANs in ophthalmology.

Methods

Dataset

This study was approved by the Institutional Review Board at the coordinating center (Oregon Health & Science University) and at each of 7 study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Children's Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, Weill Cornell Medical Center) comprising the Imaging and Informatics in ROP (i-ROP) consortium. This study was conducted in accordance with the Declaration of Helsinki. Written, informed consent was obtained from parents of all enrolled infants.

As part of the i-ROP cohort study conducted from January 2012 to July 2020, 970 subjects with birth weight $<$ 1501 g or gestational age $<$ 31 weeks underwent ROP screening over the course of their infancy. During each screening exam, posterior pole-centered retinal fundus images were acquired using RetCam cameras (Natus). A reference standard diagnosis was applied to each eye exam using previously published methods[34] by 3 to 8 independent ROP experts for zone, stage, and presence of pre-plus or plus disease. A subset of fundus images was selected; images were excluded if they were not centered on the posterior pole or exhibited stage 4 or 5 ROP (partial or total retinal detachment). This dataset was randomly split, retaining the natural distribution of plus disease into training (70%), validation (15%), and testing (15%) data subsets by subject identification number to ensure subjects were unique to the respective datasets.

Image Preparation

For each image, black-and-white retinal vessel maps were generated using a U-Net previously trained on a subset of 200 images from 154 subjects in the i-ROP database;[1] these subjects were not included in any datasets for this study. Low-level pixel information was removed from retinal vessel maps by converting all pixel values below a 10% intensity threshold (i.e., pixel value $<$ 26) to 0 to remove information about choroidal blood vessel patterns and pigmentation, both of which are not easily visible to the naked eye on vessel maps. Finally, a black, circular mask was applied to all retinal fundus images and corresponding retinal vessel maps to standardize the field of view. This same mask was applied to generated retinal fundus images.

GAN Training

Models were built and trained in Python[35] using PyTorch[36] on an Nvidia V100 GPU (Nvidia). We tuned pix2pixHD, a GAN trained to generate large, high-resolution synthetic images from segmented images,[37] using default settings (Fig 1). All fundus images and corresponding vessel maps were loaded, pairwise, into the model during training at a pixel size of 640×480×3. We chose Pix2Pix because we wanted to train a GAN to focus specifically on learning the vascular pattern in a given fundus image. This paired-to-paired image transition can be taken one step further to potentially alter vascular severity to generate novel synthetic images of different severity (i.e., turning a normal image to a plus image).[31] The model was trained for 200 epochs using the Adam optimizer with a β value of 0.5. The learning rate was constant at 3×10^{-4} during the first 100 epochs and then linearly decayed to 0 over the remaining 100 epochs. Discriminator and generator loss functions in the training set were monitored to ensure learning was occurring at an equal rate between objective functions and that overfitting did not occur. After training was completed, retinal fundus images were generated from retinal vessel maps in the validation dataset and were manually reviewed by a non-expert (A.S.C.) for veracity.

Synthetic Image Evaluation

Synthetic fundus images were generated from vessel maps in the test dataset. Of the 880 real retinal fundus images in the test dataset, 50 images were randomly selected for evaluation. Likewise, 50 synthetic retinal fundus images were also selected. This subset of images, 50 real and 50 synthetic, was used for evaluation by practicing ROP ophthalmologists familiar with Retcam images (L.O., M.E.H., D.M., and R.V.P.C.). Using a custom online system,[38] the ophthalmologists reported whether they believed each image was real or synthetic. All images were presented at a resolution of 640×480×3. Expert majority predictions for all images were also calculated; ties between experts were recorded as "synthetic" because this designation represented significant uncertainty around whether an image was perceived as fake or real. Individual experts' predictions and the expert majority predictions were compared with the ground truth.

Data Analysis

All analyses were performed in R (R Foundation).[39] Accuracies of individual experts, as well as the expert majority, were assessed. A Fisher exact test for count data was used to determine whether experts were statistically able to identify synthetic images from real images. Significance was determined at P values $\leq$ 0.05.

Informal Review of GANs in Ophthalmology

An informal review of published GANs was performed to evaluate current uses of synthetic images in ophthalmology. PubMed and Google Scholar were iteratively reviewed for any type of GAN (i.e., conditional, cycle GANs) using a combination of the following terms: *GANs, generative adversarial networks, ophthalmology, images, deepfakes, synthetic.* We additionally performed an ancestor search on included articles to broaden our search.

Results

Image Generation

Overall, 6058 images from 970 subjects were included in this dataset and split into training, validation, and test sets with a roughly equal distribution of plus disease and stage across the sets. The distribution of stages across each set was approximately 45% no stage, 15% stage 1, 15% stage 2, and 5% stage 3. The distribution of plus disease across each set was approximately 80% normal, 15% pre-plus, and 5% plus disease. All real retinal fundus images were

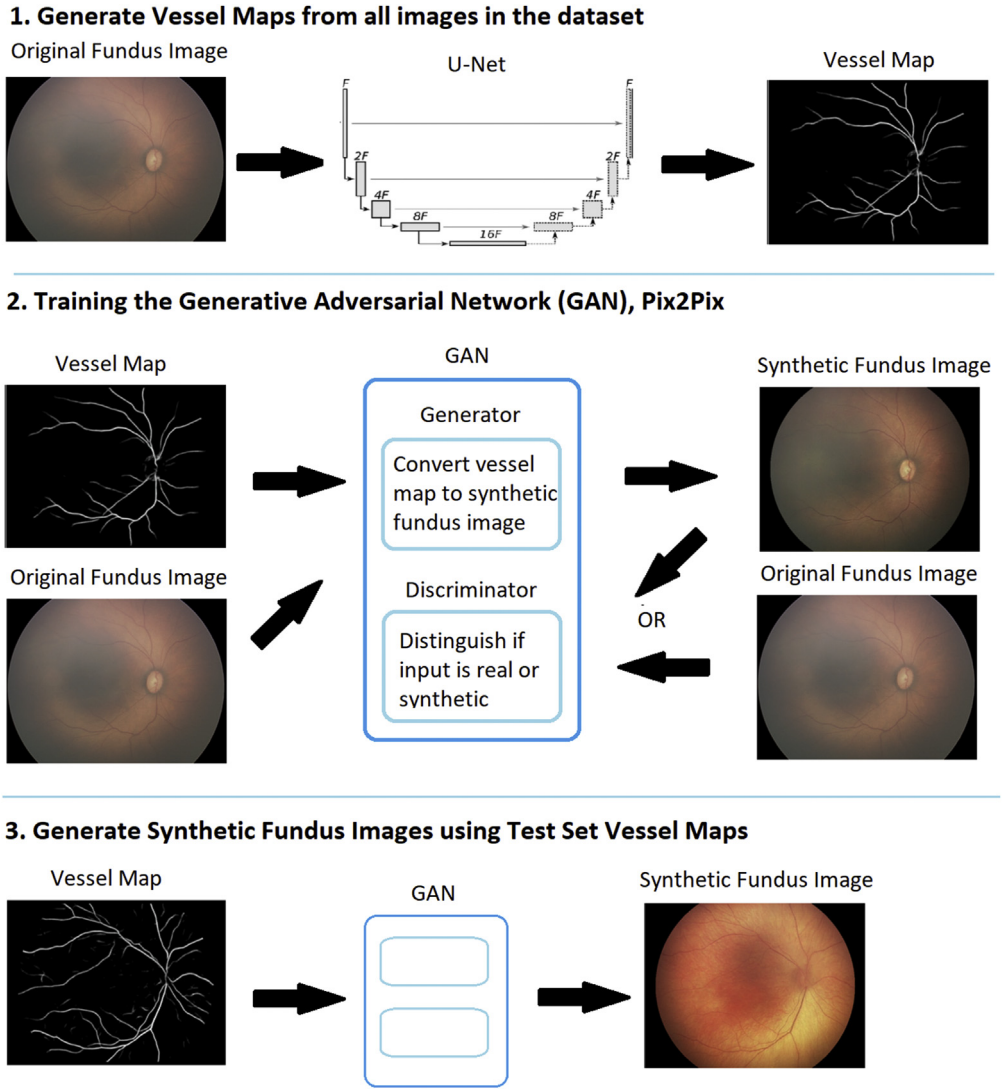

Figure 1. Generative adversarial network (GAN) pipeline for generating synthetic fundus images. First, a U-Net, a convolutional neural network architecture designed to segment image features such as vessels, was used to generate vessel maps from all fundus images in the dataset. Next, paired fundus images and their corresponding vessel maps from the test set were fed as inputs into Pix2Pix, a conditional GAN. This GAN consists of 2 neural networks: (1) a generator that was trained to generate synthetic fundus images from vessel maps and (2) a discriminator that was trained to discriminate between real and synthetic fundus images. After training was completed, vessel maps from the test set were inputted into the GAN and a synthetic fundus image was generated.

successfully segmented into grayscale vessel maps using a U-Net (Fig 2). After training for 200 epochs on 4282 image pairs, the GAN was evaluated for veracity via manual review of images generated from retinal vessel maps in the validation dataset; synthetic retinal fundus images were then generated from all vessel maps in the test dataset (Fig 2). Although most images appeared realistic to a layperson, 5 of the 880 images (0.57%) in the test dataset were obviously unrealistic (Fig 3). This observation seemed to occur only in areas of lower-quality images where retinal vessel information was lacking.

Image Evaluation

Fifty real and 50 synthetic images of similar stage and plus disease distribution as the original dataset were uploaded to a custom online evaluation platform,[38] and 4 ROP experts determined whether the images were real or synthetic. The expert majority correctly identified 59% of images as being real or synthetic; experts 1 to 4 correctly identified 54%, 58%, 49%, and 61% of images, respectively (Table 1). Fisher exact test *P* values for the expert majority and experts 1 to 4 were 0.100, 0.505, 0.158, 1.000, and 0.043, respectively. These results suggest only expert 4 could significantly discern between real and synthetic images, and that, in general, the majority of experts could not.

GANs in Ophthalmology

We found 20 published implementations of GANs specific to ophthalmology. Of these, 11 articles synthesized fundus

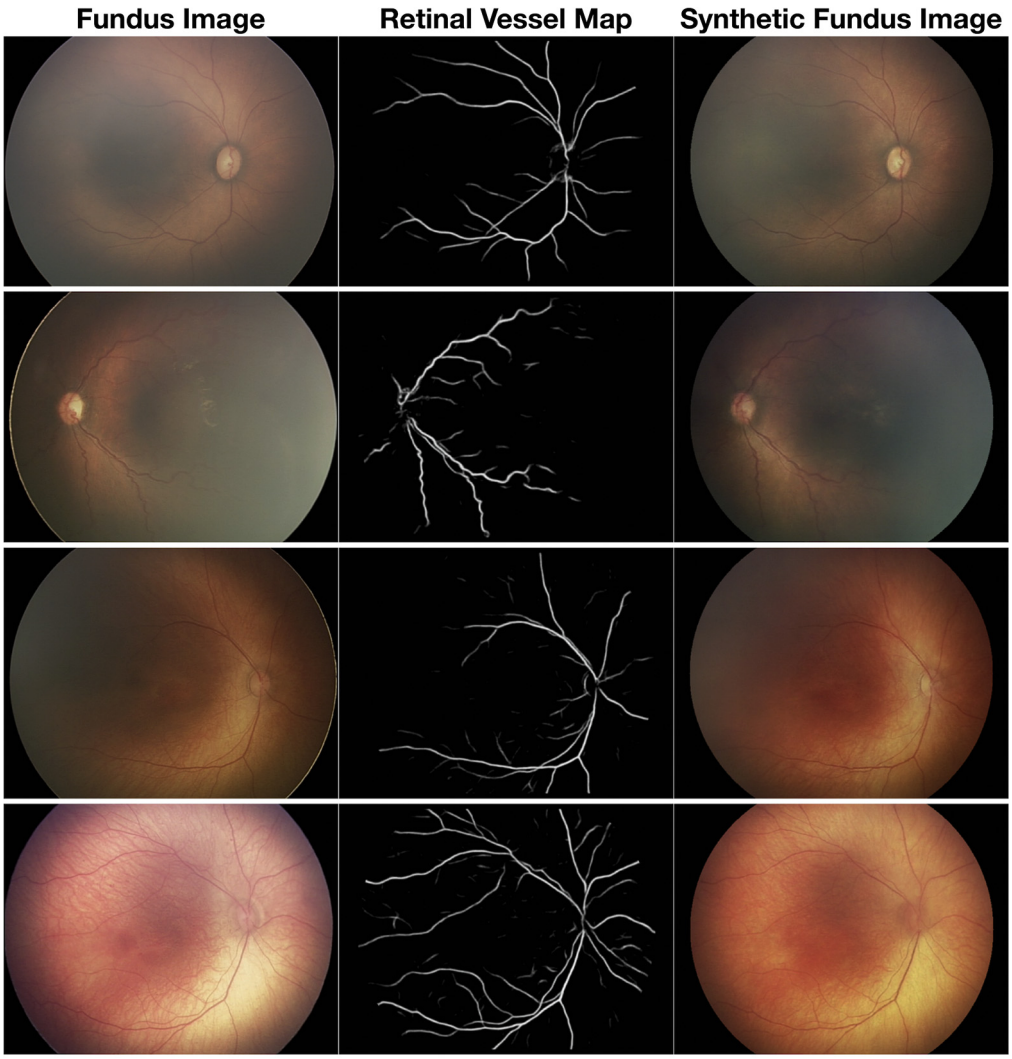

Figure 2. Synthetic retinal images generated from retinal vessel maps. Real retinal fundus images (left) are first segmented into retinal vessel maps (center) using a previously trained U-Net. By using pix2pixHD, a custom implementation of a generative adversarial network (GAN), the retinal vessel maps are then used to generate synthetic retinal fundus images (right).

images,[9,19,20,23,32,37,40–44] 6 articles synthesized OCT images,[27–29,45–47] 2 articles synthesized fluorescein angiography images,[48,49] and 1 article synthesized infrared images[21] (Table 2). The majority of GANs were proof-of-concept studies demonstrating feasibility of generating realistic-appearing synthetic images. Specific implementations of GANs were published in 9 articles for diagnosis of ophthalmic diseases, including diabetic retinopathy (DR),[9,20,32,40] glaucoma,[28,45] age-related macular degeneration,[19,46] and meibomian gland dysfunction.[21]

Discussion

In this study, we demonstrated (1) that a U-Net and GAN pipeline can generate realistic-appearing synthetic fundus images from vessel maps of real fundus images acquired from ROP screening and (2) that the majority of experts are unable to discern between real and synthetic fundus images. We identified multiple examples of GANs, applied a number of ophthalmic imaging modalities and diseases, and review the potential utility of GANs for dataset augmentation to improve the robustness of algorithms, contribute to medical education, and reduce privacy concerns resulting from the sharing and use of patients' images. We additionally discuss the limitations of GANs in clinical use and offer future directions for research.

Dataset Augmentation and Generation

A fundamental requirement of training DL algorithms for clinical deployment in a heterogeneous population is a large, diverse dataset, which may be challenging to acquire from a single institution. However, multi-institutional datasets are also difficult to acquire because of patient privacy regulations and the practicality of storing these data. Although augmentation methods such as image flips and rotations are routinely used to increase the size of training datasets in DL,

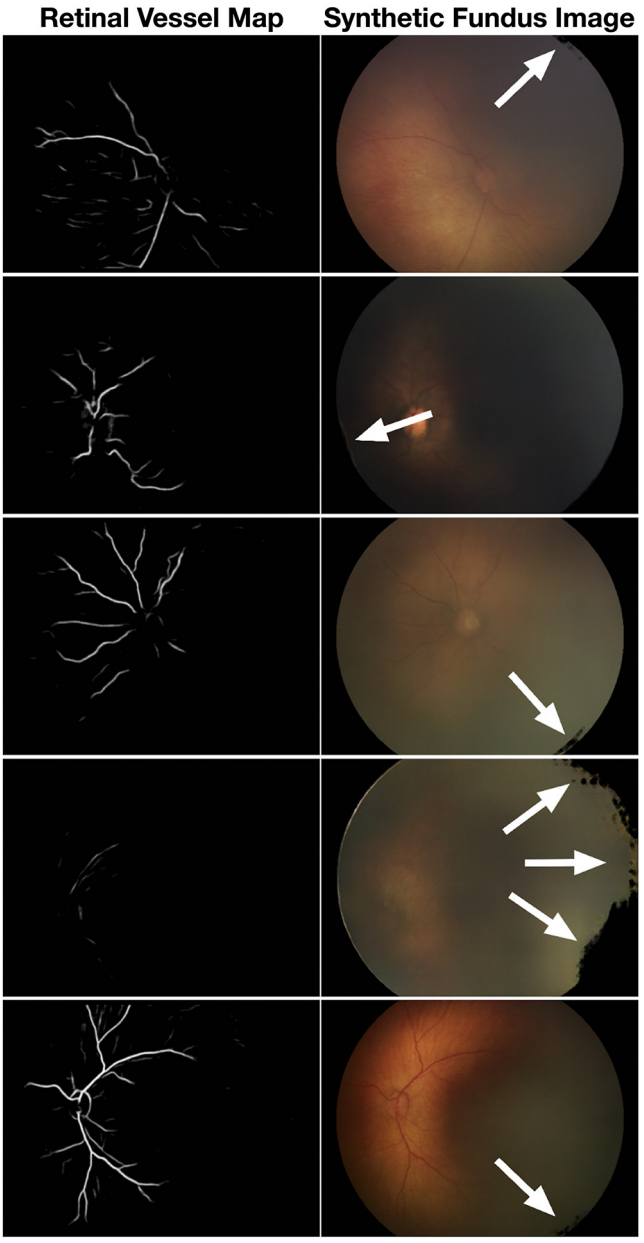

Figure 3. Obvious cases where the generative adversarial network (GAN) did not produce realistic results. A small proportion of test dataset images (0.57%) had clear and obvious markings that indicated they were synthetic images (white arrows).

they do not increase the feature diversity of the data, which in turn affects an algorithm's generalizability. Similar to previous work exploring synthesis of fundus images,[9,19,20,23,32,37,41,43] our GAN was trained on a smaller sample of images and subsequently used to generate 880 synthetic images. These results highlight the potential for GANs to augment the size of the dataset through the combination of real and synthetic data.

Another strategy to increase both the size and diversity of the dataset is the creation of multi-institutional datasets. Although this strategy may marginally increase dataset heterogeneity and size, it is challenging to create datasets that contain examples of every combination of image quality, demographic and ethnic diversity, and so forth that an algorithm might encounter in the clinical population.[4] To address this, synthetic images can be modified during the GAN's image generation process to specifically address biases by augmenting synthetically "unseen" populations to real training data. Potential biases that can be addressed by GANs include underrepresented demographic groups, image acquisition from different devices, and class imbalances, such as limited images available for rare diseases. For example, Burlina et al[9] demonstrated that augmenting synthetic images of darkly pigmented retina images to a predominantly lightly pigmented dataset could decrease DL performance bias toward lightly pigmented retinas for DR. These principles of modifying synthetic images to increase dataset diversity potentially hold true for other population characteristics beyond pigmentation and other image characteristics. Generative adversarial networks can also modify existing vessel maps from "normal" fundus images to demonstrate various degrees of vascular severity for ROP and other diseases that present along a spectrum of severity. Future work is needed to assess the use of these synthetic images modified along a spectrum of disease in DL.

Beyond dataset augmentation, GANs can create completely novel datasets. As part of the proof-of-concept GAN studies by Burlina et al[19] and Zheng et al,[27] both studies demonstrated comparable performance of DL algorithms trained on exclusively synthetic versus exclusively real fundus and OCT images, respectively, although both maintained similar disease distribution across their training sets. Similar to models trained on clinically acquired data, models trained exclusively on synthetic data will need to be validated on data from clinical settings. However, because these models were trained on data distributions present in the original datasets, there may be biased performance against certain demographic groups, which could be addressed by using synthetic images to balance dataset augmentation as described.[9] Future studies should assess whether training GANs on more synthetically balanced datasets in terms of demographics and disease prevalence results in improved testing performance on data representative of the general population.

Medical Education

The rise in importance of big data, artificial intelligence, electronic health records, and tele-health and tele-education in the setting of the Coronavirus Disease 2019 pandemic have all led to calls for changes in the way we educate trainees in ophthalmology.[50–53] Tele-education platforms for ophthalmic imaging have been well documented in the literature.[50,54,55] However, the effectiveness of tele-education platforms to transfer knowledge about disease phenotypes depends on adequate numbers of representative images across the entire disease spectrum. In diseases with a low prevalence of severe cases (i.e., ROP), there may be a dearth of high-quality training images from certain disease phenotypes, cameras, ethnic

Table 1. Confusion Matrix of Expert Determinations of Real versus Synthetic Images

True		Expert Majority		Expert 1		Expert 2		Expert 3		Expert 4	
		Real	*Synthetic*	*Real*	*Synthetic*	*Real*	*Synthetic*	*Real*	*Synthetic*	*Real*	*Synthetic*
Image	Real	35	15	38	12	32	18	43	7	34	16
Type	Synthetic	26	24	34	16	24	26	44	6	23	27

Experts were generally unable to discern between real and synthetic images (accuracy = 54%, 58%, 49%, and 61%, respectively).

subgroups, and patients who have consented to use their images for educational purposes. Synthesizing cases to augment and customize trainee-specific educational experiences may result in improved recognition of more severe cases without having to prospectively identify patients who develop severe disease. Because GANs can also modify an image along a disease spectrum represented in a dataset, synthetic images of the hypothetically "same" patient across various levels of disease severity may improve trainee recognition of disease progression.[31] These images may be used to train trainees/clinicians to stage disease and progression longitudinally. Future work assessing the utility of synthetic images in ophthalmic disease education is warranted.

Table 2. Informal Review of Current Applications of Generative Adversarial Networks in Ophthalmology

Authors, Year	Image Modality	GAN Architecture	Summary of GAN Use Case
Andreini et al, 2018[32]	Fundus	Pix2Pix HD	Synthesis of high-resolution fundus photos using vessel segmentations of publicly available DR image sets.
Wang et al, 2018[37]	Fundus	Conditional GAN	Synthesis of high-resolution fundus photos.
Zhao et al, 2018[23]	Fundus	Custom GAN (Tub-GAN)	Synthesis of fundus photos using 10-20 images.
Burlina et al, 2019[19]	Fundus	ProGAN	Synthesis of fundus images for wet vs. dry AMD. Evaluation of expert ability to discern synthetic vs. real. Trained CNN to identify AMD using datasets of exclusively synthetic or real images.
Niu et al, 2019[40]	Fundus	Custom GAN	Synthesis of lesions specific to diabetic retinopathy.
Odaibo et al, 2019[29]	OCT	Unspecified GAN	Synthesis of retinal OCT images.
Yu et al, 2019[41]	Fundus	Custom GAN, Pix2Pix	Synthesis of high-resolution optic disc photos using a multiple-channel and landmark strategy.
Ha et al, 2020[44]	Fundus	Super-Resolution GAN	Synthesis of high-resolution optic disc photos from low-resolution photos.
Hassan et al, 2020[45]	OCT	Conditional GAN	Predict progression of glaucoma using macular OCT images.
Li et al, 2020[47]	Fluorescein Angiography	Conditional GAN	Synthesis of fluorescein angiography photos from fundus photos.
Liu et al, 2020[46]	OCT	Pix2Pix HD	Synthesis of retinal OCT photos. Evaluation of image quality. Evaluate use of synthetic images to predict treatment response for AMD.
Tavakkoli et al, 2020[49]	Fluorescein Angiography	Conditional GAN	Synthesis of fluorescein angiography photos from fundus photos. Evaluate expert ability to discern synthetic vs. real.
Zheng et al, 2020[27]	OCT	Progressively Grown GAN	Synthesis of retinal OCT images. Evaluation of image quality between real vs. synthetic images. Training a CNN on diagnosis of referral warranting findings using exclusively synthetic or real images.
Zhou et al, 2020[20]	Fundus	GAN	Synthesis of fundus photos that show modification of lesions representative of DR.
Burlina et al, 2021[9]	Fundus	StyleGAN	Synthesis of fundus images of diverse pigmentation for augmentation to a DL algorithm for DR synthesis.
Cheong et al, 2021[47]	OCT	Custom GAN	Synthesis of retinal OCT images with blood vessel shadows removed.
Coyner et al, 2021[42]	Fundus	Pix2Pix HD	Synthesis of high-resolution fundus photos from an ROP screening program.
Khan et al, 2021[21]	Infrared Images	Conditional GAN	Synthesis and processing of infrared images for quantification of irregularities of the meibomian gland.
Wang et al, 2021[43]	Fundus	Custom GAN	Synthesis of diabetic retinopathy image and diagnosis using a multi-channel strategy.
Zheng et al, 2021[28]	OCT	Progressively Grown GAN	Synthesis of anterior-segment OCT images. Evaluation of image quality between real vs. synthetic images. Training a CNN on diagnosis of glaucoma using synthetic vs. real images.

AMD = age-related macular degeneration; CNN = convolutional neural network; DR = diabetic retinopathy; GAN = generative adversarial network; ROP = retinopathy of prematurity.
Overall, 20 published implementations of GANs were found in ophthalmology. These GANs were used to synthesize fundus, OCT, fluorescein angiography, and infrared images. The majority of these GANs were proof-of-concept studies demonstrating feasibility of creating realistic synthetic images.

Privacy

Data privacy laws enacted by the European Union[56–59] to protect patient privacy have other important implications in regulating dataset sharing, which in turn restrict the ability to train more generalizable DL algorithms. In ophthalmology, these challenges are further compounded because the retina and its vasculature are considered protected health information.[60–63] In our study, the synthetic images appeared similar to the original, even though the choroidal vascular patterns ostensibly were fully synthetic. However, GANs trained on paired image-to-image transition may be used to alter the severity of vessel maps (i.e., from normal to pre-plus to plus or vice versa) to generate completely new segmentations and fundus images that are potentially biometrically distinct from the patient's native vasculature. In practice, using retinal vasculature as identifiable data may be problematic to implement, because the retinal appearance can change over time, perhaps more so than other biometric data such as fingerprints. For example, in DR, the purpose of screening using retinal photographs is to detect change in retinopathy status, that is, a change in the way an image looks over time. The degree to which a retinal image can be used to identify a person, especially when that retina looks different over time with age,[64] the presence of disease,[65] and with different cameras, is unclear.

Similar to other DL algorithms, GANs have also been shown to be vulnerable to malicious privacy breaches such as membership attacks, which are adversarial attacks designed to identify which images or patients were used in model training.[66–73] These attacks essentially operate on the premise that DL algorithms perform better on images that they were trained on[74] and depend on whether the attacker has access to the code underlying the model (white-box) or not (black-box).[75] While defense against these attacks remains an active area of research,[71,74] they are costly,[74] and some defense approaches that require re-training the model may even decrease the performance of the original DL algorithm.[75]

Limitations of GANs

Important inherent limitations of GANs exist that require further study before clinical implementation of these algorithms. First, GANs can only synthesize images representing disease phenotypes and imaging features within the training data's distribution.[16] Therefore, the phenotypic spectrum of synthetic images may not represent the full phenotypic variability seen in clinical practice, which is crucial for rare diseases. Second, GANs are often used to improve signal quality or fill in missing information in an image; however, the resulting "improved" image might obscure the presence of real pathology that would have been visible without the artifact or on a better-quality scan/image.[47] Additionally, they can produce so-called image hallucinations, that is, the addition of image features not actually present, which may or may not be useful.[76–79]

Study Limitations

Our study has additional limitations. First, our GAN was trained on RetCam images from North American infants screened for ROP. Future work is needed to evaluate the generalizability of our GAN in other populations and devices. Second, our GAN generated a few images that were clearly unrealistic (Fig 3). These erroneous images were few and were easily identified from our generated dataset, but we speculate that training on larger datasets with varying image quality likely improved the overall quality of synthetic images. Third, we did not ask experts to review images more than once, and therefore did not evaluate the reproducibility of expert evaluation. Although the majority of experts were statistically unable to discern between synthetic and real images, it may be interesting to evaluate whether experts can learn to recognize synthetic versus real images over time. Finally, our GAN was only trained on images from stages 1 to 3 ROP because of the sparsity of images with stages 4 and 5 in our dataset; more prospective data collection is needed to train GANs that can generate realistic-appearing images across the full spectrum of stage, zone, and plus disease.

In conclusion, generative adversarial networks can generate synthetic fundus images that are indiscernible from real fundus images by expert ROP ophthalmologists. Although these synthetic images have many potential applications in DL data augmentation and education, issues surrounding privacy and hallucinations must be further studied before clinical implementation.

Footnotes and Disclosures

Originally received: June 9, 2021.
Final revision: October 1, 2021.
Accepted: October 29, 2021.
Available online: November 16, 2021. Manuscript no. D-21-00097.

[1] Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, Portland, Oregon.

[2] Department of Ophthalmology and Visual Sciences, University of Illinois at Chicago, Chicago, Illinois.

[3] Department of Ophthalmology, John A. Moran Eye Center, University of Utah, Salt Lake City, Utah.

[4] Byers Eye Institute, Horngren Family Vitreoretinal Center, Department of Ophthalmology, Stanford University School of Medicine, Palo Alto, California.

[5] Department of Radiology, Massachusetts General Hospital/Harvard Medical School, Charlestown, Massachusetts.

[6] Massachusetts General Hospital & Brigham and Women's Hospital Center for Clinical Data Science, Boston, Massachusetts.

[7] National Eye Institute, National Institutes of Health, Bethesda, Maryland.

*Drs. Chen and Coyner contributed to this work equally.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s): R.V.P.C.: Scientific Advisory Board − Phoenix Technology Group; Consultant −Novartis, Alcon.

M.F.C.: Consultant − Novartis; Equity owner − Inteleretina.

J.P.C.: Consultant − Boston AI Labs.

D.M.M.: Consultant − Alexion, Congruence Medical Solutions, M3 Global Solutions; Equity owner − Pykus, Grand Legend Technology, Versl, Visunex, Promisight, dSentz.

J.P.C., D.M.M., R.V.P.C.: Research support − Genentech.

R.V.P.C.: Research support − Regeneron.

This work was supported by grant nos. R01 EY19474, R01 EY031331, R21 EY031883, P30 EY10572, P30 EY02687, R01 EY015130, and R01 EY17011 from the National Institutes of Health, and by unrestricted departmental funding and a Career Development Award (J.P.C.) from Research to Prevent Blindness.

HUMAN SUBJECTS: Human subjects were included in this study. This study was approved by the Institutional Review Board at the coordinating center (Oregon Health & Science University) and at each of seven study centers (Columbia University, University of Illinois at Chicago, William Beaumont Hospital, Children's Hospital Los Angeles, Cedars-Sinai Medical Center, University of Miami, Weill Cornell Medical Center) comprising the Imaging and Informatics in ROP (i-ROP) consortium. This study was conducted in accordance with the Declaration of Helsinki. Written, informed consent was obtained from parents of all enrolled infants.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Chen, Coyner, Chan, Hartnett, Moshfeghi, Owen, Kalpathy-Cramer, Chiang, Campbell

Data collection: Chen, Coyner, Chan, Kalpathy-Cramer, Chiang, Campbell

Analysis and interpretation: Chen, Coyner, Chan, Kalpathy-Cramer, Chiang, Campbell

Obtained funding: N/A; Study was performed as part of the authors' regular employment duties. No additional funding was provided.

Overall responsibility: Chen, Coyner, Chan, Hartnett, Moshfeghi, Owen, Kalpathy-Cramer, Chiang, Campbell

Abbreviations and Acronyms:
DL = deep learning; **DR** = diabetic retinopathy; **GAN** = generative adversarial network; **i-ROP** = Informatics in ROP; **ROP** = retinopathy of prematurity.

Keywords:
Deep learning, Generative adversarial networks, Ophthalmology, Synthetic images.

Correspondence:
J. Peter Campbell, MD, MPH, Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, 515 SW Campus Drive, Portland, OR 97239. E-mail: campbelp@ohsu.edu.

References

1. Brown JM, Campbell JP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol*. 2018;136:803−810.
2. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402−2410.
3. Coyner AS, Swan R, Campbell JP, et al. Automated fundus image quality assessment in retinopathy of prematurity using deep convolutional neural networks. *Ophthalmol Retina*. 2019;3:444−450.
4. Chen JS, Coyner AS, Ostmo S, et al. Deep learning for the diagnosis of stage in retinopathy of prematurity: accuracy and generalizability across populations and cameras. *Ophthalmol Retina*. 2021;5:1027−1035.
5. Christopher M, Bowd C, Proudfoot JA, et al. Deep learning estimation of 10-2 and 24-2 visual field metrics based on thickness maps from macula optical coherence tomography. *Ophthalmology*. 2021;128:1534−1548.
6. Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318:2211−2223.
7. Burlina PM, Joshi N, Pacheco KD, et al. Use of deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration. *JAMA Ophthalmol*. 2018;136:1359−1366.
8. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health*. 2019;9, 010318−010318.
9. Burlina P, Joshi N, Paul W, et al. Addressing artificial intelligence bias in retinal diagnostics. *Transl Vis Sci Technol*. 2021;10, 13−13.
10. Chassang G. The impact of the EU general data protection regulation on scientific research. *Ecancermedicalscience*. 2017;11, 709−709.
11. McCallister E, Grance T, Scarfone KA. *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*. Gaithersburg, MD: National Institute of Standards and Technology; 2010. NIST SP 800−122.
12. Brisimi TS, Chen R, Mela T, et al. Federated learning of predictive models from federated electronic health records. *Int J Med Inf*. 2018;112:59−67.
13. Xu J, Glicksberg BS, Su C, et al. Federated learning for healthcare informatics. *J Healthc Inform Res*. 2020:1−19. Online ahead of print.
14. Chang K, Balachandar N, Lam C, et al. Distributed deep learning networks among institutions for medical imaging. *J Am Med Inform Assoc*. 2018;25:945−954.
15. Mehta N, Lee CS, Mendonça LSM, et al. Model-to-data approach for deep learning in optical coherence tomography intraretinal fluid segmentation. *JAMA Ophthalmol*. 2020;138:1017−1024.
16. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. NIPS'14*. Cambridge, MA: MIT Press; 2014:2672−2680.
17. Crystal DT, Cuccolo NG, Ibrahim AMS, et al. Photographic and video deepfakes have arrived: how machine learning may influence plastic surgery. *Plast Reconstr Surg*. 2020;145:1079−1086.
18. Fallis D. The epistemic threat of deepfakes. *Philos Technol*. 2020:1−21 [Online ahead of print].
19. Burlina PM, Joshi N, Pacheco KD, et al. Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA Ophthalmol*. 2019;137:258−264.

20. Zhou Y, Wang B, He X, et al. DR-GAN: conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images. *IEEE J Biomed Health Inform*. 2020, 1−1 [Online ahead of print].
21. Khan ZK, Umar AI, Shirazi SH, et al. Image based analysis of meibomian gland dysfunction using conditional generative adversarial neural network. *BMJ Open Ophthalmol*. 2021;6. e000436.
22. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. *Med Image Anal*. 2019;58:101552.
23. Zhao H, Li H, Maurer-Stroh S, Cheng L. Synthesizing retinal and neuronal images with generative adversarial nets. *Med Image Anal*. 2018;49:14−26.
24. Park JE, Eun D, Kim HS, et al. Generative adversarial network for glioblastoma ensures morphologic variations and improves diagnostic model for isocitrate dehydrogenase mutant type. *Sci Rep*. 2021;11:9912.
25. Kazuhiro K, Werner RA, Toriumi F, et al. Generative adversarial networks for the creation of realistic artificial brain magnetic resonance images. *Tomogr Ann Arbor Mich*. 2018;4: 159−163.
26. Sandfort V, Yan K, Pickhardt PJ, Summers RM. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Sci Rep*. 2019;9, 16884−16884.
27. Zheng C, Xie X, Zhou K, et al. Assessment of generative adversarial networks model for synthetic optical coherence tomography images of retinal disorders. *Transl Vis Sci Technol*. 2020;9, 29−29.
28. Zheng C, Bian F, Li L, et al. Assessment of generative adversarial networks for synthetic anterior segment optical coherence tomography images in closed-angle detection. *Transl Vis Sci Technol*. 2021;10, 34−34.
29. Odaibo SG. Generative adversarial networks synthesize realistic OCT images of the retina. *ArXiv190206676 Cs*. http://arxiv.org/abs/1902.06676. Accessed May 10, 2021.
30. Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of StyleGAN. *ArXiv191204958 Cs Eess Stat*. http://arxiv.org/abs/1912.04958. Accessed June 5, 2021.
31. Coyner AS, Campbell JP, Kalpathy-Cramer J, et al. Retinal fundus image generation in retinopathy of prematurity using autoregressive generative models. *Invest Ophthalmol Vis Sci*. 2020;61(7):2166.
32. Andreini P, Bonechi S, Bianchini M, et al. A two stage GAN for high resolution retinal image generation and segmentation. ArXiv190712296 Cs Eess. http://arxiv.org/abs/1907.12296. Accessed April 21, 2021.
33. Costa P, Galdran A, Meyer MI, et al. End-to-end adversarial retinal image synthesis. *IEEE Trans Med Imaging*. 2018;37: 781−791.
34. Ryan MC, Ostmo S, Jonas K, et al. Development and evaluation of reference standards for image-based telemedicine diagnosis and clinical research studies in ophthalmology. *AMIA Annu Symp Proc AMIA Symp*. 2014;2014:1902−1910.
35. Van Rossum G, Drake Jr FL. *Python Reference Manual*. Amsterdam: Centrum voor Wiskunde en Informatica Amsterdam; 1995.
36. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, et al., eds. *Advances in Neural Information Processing Systems 32*. Red Hook, NY: Curran Associates, Inc.; 2019:8024−8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf. Accessed May 22, 2021.
37. Wang T-C, Liu M-Y, Zhu J-Y, et al. High-resolution image synthesis and semantic manipulation with conditional GANs. *ArXiv171111585 Cs*. http://arxiv.org/abs/1711.11585. Accessed April 28, 2021.
38. Kalpathy-Cramer J, Campbell JP, Erdogmus D, et al. Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis. *Ophthalmology*. 2016;123:2345−2351.
39. R Core Team. *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing*; 2020. https://www.R-project.org/. Accessed March 10, 2021.
40. Niu Y, Gu L, Lu F, et al. Pathological evidence exploration in deep retinal image diagnosis. *Proc AAAI Conf Artif Intell*. 2019;33:1093−1101.
41. Yu Z, Xiang Q, Meng J, et al. Retinal image synthesis from multiple-landmarks input with generative adversarial networks. *Biomed Eng OnLine*. 2019;18:62.
42. Coyner AS, Chen J, Campbell JP, et al. Diagnosability of synthetic retinal fundus images for plus disease detection in retinopathy of prematurity. *AMIA Annu Symp Proc AMIA Symp*. 2021;2020:329−337.
43. Wang S, Wang X, Hu Y, et al. Diabetic retinopathy diagnosis using multichannel generative adversarial network with semi-supervision. *IEEE Trans Autom Sci Eng*. 2021;18:574−585.
44. Ha A, Sun S, Kim YK, et al. Deep-learning-based enhanced optic-disc photography. *PloS One*. 2020;15. e0239913−e0239913.
45. Hassan ON, Sahin S, Mohammadzadeh V, et al. Conditional GAN for prediction of glaucoma progression with macular optical coherence tomography. In: Bebis G, Yin Z, Kim E, et al., eds. *Advances in Visual Computing*. New York: Springer International Publishing; 2020:761−772.
46. Liu Y, Yang J, Zhou Y, et al. Prediction of OCT images of short-term response to anti-VEGF treatment for neovascular age-related macular degeneration using generative adversarial network. *Br J Ophthalmol*. 2020;104:1735.
47. Cheong H, Devalla SK, Pham TH, et al. DeshadowGAN: a deep learning approach to remove shadows from optical coherence tomography images. *Transl Vis Sci Technol*. 2020;9, 23−23.
48. Li W, Kong W, Chen Y, et al. Generating fundus fluorescence angiography images from structure fundus images using generative adversarial networks. https://openreview.net/forum?id=qhZM390B4. Accessed May 10, 2021.
49. Tavakkoli A, Kamran SA, Hossain KF, Zuckerbrod SL. A novel deep learning conditional generative adversarial network for producing angiography images from retinal fundus photographs. *Sci Rep*. 2020;10:21580.
50. Campbell JP, Swan R, Jonas K, et al. Implementation and evaluation of a tele-education system for the diagnosis of ophthalmic disease by international trainees. *AMIA Annu Symp Proc AMIA Symp*. 2015;2015:366−375.
51. Cole E, Valikodath NG, Maa A, et al. Bringing ophthalmic graduate medical education into the 2020s with information technology. *Ophthalmology*. 2021;128:349−353.
52. Valikodath NG, Al-Khaled T, Cole E, et al. Evaluation of pediatric ophthalmologists' perspectives of artificial intelligence in ophthalmology. *J AAPOS*. 2021;25, 164.e1-164.e5.
53. Sharma D, Bhaskar S. Addressing the Covid-19 burden on medical education and training: the role of telemedicine and tele-education during and beyond the pandemic. *Front Public Health*. 2020;8:838.
54. Patel SN, Martinez-Castellanos MA, Berrones-Medina D, et al. Assessment of a tele-education system to enhance

retinopathy of prematurity training by international ophthalmologists-in-training in Mexico. *Ophthalmology*. 2017;124:953–961.
55. Caffery LJ, Taylor M, Gole G, Smith AC. Models of care in tele-ophthalmology: a scoping review. *J Telemed Telecare*. 2017;25:106–122.
56. Policy and investment recommendations for trustworthy Artificial Intelligence | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence. Accessed April 29, 2021.
57. Phillips M. International data-sharing norms: from the OECD to the General Data Protection Regulation (GDPR). *Hum Genet*. 2018;137:575–582.
58. Peloquin D, DiMaio M, Bierer B, Barnes M. Disruptive and avoidable: GDPR challenges to secondary research uses of data. *Eur J Hum Genet*. 2020;28:697–705.
59. Molnár-Gábor F, Korbel JO. Genomic data sharing in Europe is stumbling—Could a code of conduct prevent its fall? *EMBO Mol Med*. 2020;12. e11421.
60. Waheed Z, Usman Akram M, Waheed A, et al. Person identification using vascular and non-vascular retinal features. *Comput Electr Eng*. 2016;53:359–371.
61. Farzin H, Abrishami-Moghaddam H, Moin M-S. A novel retinal identification system. *EURASIP J Adv Signal Process*. 2008;2008:280635.
62. Bolle R, Pankanti S, Jain AK. Biometrics. In: Jain AK, Bolle R, Pankanti S, eds. *Personal Identification in Networked Society*. New York: Kluwer Academic Publishers; 1996:123–141.
63. Bellemo V, Burlina P, Yong L, et al. Generative Adversarial Networks (GANs) for retinal fundus image synthesis. In: Carneiro G, You S, eds. *Computer Vision – ACCV 2018 Workshops*. NY: Springer International Publishing; 2019:289–302.
64. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2:158–164.
65. Korot E, Pontikos N, Liu X, et al. Predicting sex from retinal fundus photographs using automated deep learning. *Sci Rep*. 2021;11:10286.
66. Chen D, Yu N, Zhang Y, Fritz M. GAN-Leaks: a taxonomy of membership inference attacks against generative models. In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. CCS '20. NY: Association for Computing Machinery; 2020:343–362.
67. Zhang J, Zhang J, Chen J, Yu S. GAN enhanced membership inference: a passive local attack in federated learning. In: *ICC 2020-2020 IEEE International Conference on Communications (ICC)*; 2020:1–6. https://ieeexplore.ieee.org/document/9148790. Accessed May 10, 2021.
68. Liu KS, Xiao C, Li B, Gao J. Performing co-membership attacks against deep generative models. In: *2019 IEEE International Conference on Data Mining (ICDM)*. 2019:459–467.
69. Hilprecht B, Härterich M, Bernau D. Monte Carlo and reconstruction membership inference attacks against generative models. *Proc Priv Enhancing Technol*. 2019;2019:232–249.
70. Pyrgelis A, Troncoso C, De Cristofaro E. Knock knock, who's there? Membership inference on aggregate location data. *ArXiv170806145 Cs*. http://arxiv.org/abs/1708.06145. Accessed May 21, 2021.
71. Jia J, Salem A, Backes M, et al. MemGuard: Defending against black-box membership inference attacks via adversarial examples. In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. CCS '19*. NY: Association for Computing Machinery; 2019:259–274.
72. Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. *ArXiv161005820 Cs Stat*. http://arxiv.org/abs/1610.05820. Accessed May 21, 2021.
73. Salem A, Zhang Y, Humbert M, et al. ML-Leaks: model and data independent membership inference attacks and defenses on machine learning models. *ArXiv180601246 Cs*. http://arxiv.org/abs/1806.01246. Accessed May 21, 2021.
74. Huang H, Luo W, Zeng G, et al. DAMIA: Leveraging domain adaptation as a defense against membership inference attacks. *ArXiv200508016 Cs*. http://arxiv.org/abs/2005.08016. Accessed May 14, 2021.
75. Finlayson SG, Bowers JD, Ito J, et al. Adversarial attacks on medical machine learning. *Science*. 2019;363:1287.
76. Cohen JP, Luck M, Honari S. Distribution matching losses can hallucinate features in medical image translation. *ArXiv180508841 Cs*. http://arxiv.org/abs/1805.08841. Accessed May 5, 2021.
77. Fulgeri F, Fabbri M, Alletto S, et al. Can adversarial networks hallucinate occluded people with a plausible aspect? *Comput Vis Image Underst*. 2019;182:71–80.
78. Zhang Y, Tsang I, Luo Y, et al. Copy and paste GAN: face hallucination from shaded thumbnails. ArXiv200210650 Cs. http://arxiv.org/abs/2002.10650. Accessed May 21, 2021.
79. Rajput SS, Arya KV, Singh V, Bohat VK. Face hallucination techniques: a survey. In: *2018 Conference on Information and Communication Technology (CICT)*. Jabalpur, India; 2018:1–6.