Eurographics Workshop on Visual Computing for Biology and Medicine (2024)
L. Garrison and D. Jönsson (Editors)

Short Paper

Exploring Drusen Type and Appearance using Interpretable GANs

C. Muth[1], O. Morelle[2,3], R. G. Raidou[1], M. W. M. Wintergerst[3], R. P. Finger[3,4], and T. Schultz[2,5]

[1]TU Wien, Austria; [2]B-IT and Department of Computer Science, University of Bonn, Bonn, Germany; [3]Department of Ophthalmology, University Hospital Bonn, Bonn, Germany; [4]Department of Ophthalmology, University Medical Center Mannheim, Heidelberg University, Mannheim, Germany; [5]Lamarr Institute for Machine Learning and Artificial Intelligence, Bonn, Germany

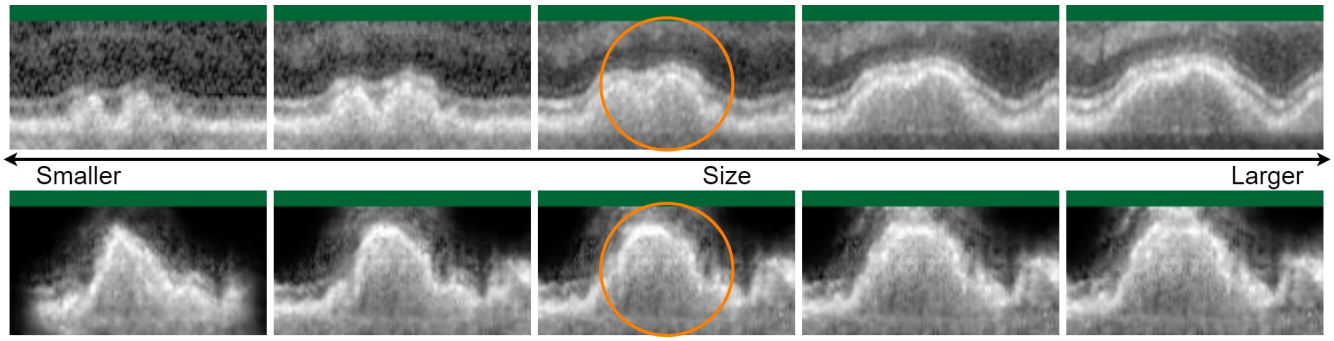


Figure 1: *Each row represents a path depicting the shrinkage (to the left) and growth (to the right) of drusen, w.r.t. a reference state in the middle (in the orange circle). The first row shows a path of multiple drusen, taken from the StyleGAN that was trained for 2000 epochs without weighted sampling, and the path model is Warp [TTP21] in $\mathcal{Z}$ space. The second row shows a path of a single druse, taken from the GAN that was trained for 100 epochs with weighted sampling, and the path model is PDE [SKSW23] in $\mathcal{Z}$ space. The green bar on top of the images indicates that the respective latent codes fall in the support of the GAN's Gaussian prior ($p > 0.05$), and are therefore expected to align with the training data.*

Abstract

We propose an algorithmic pipeline that uses interpretable Generative Adversarial Networks (GANs) to visualize the variability of the visual appearance of drusen in Optical Coherence Tomography (OCT). Drusen are accumulations of extracellular debris between Bruch's membrane and the retinal pigment epithelium of the eye. They are a hallmark of age-related macular degeneration (AMD)—the most common cause of vision loss in the elderly. Imaging the morphology of drusen with OCT reveals different subtypes, which might have different relevance for disease severity and the risk of progression. We compare two GAN architectures and three recently proposed methods for the unsupervised discovery of interpretable paths in their latent space with respect to their ability to visualize natural variations in drusen appearance. We also introduce a color code that indicates generated images that extrapolate beyond the training data and should, therefore, be interpreted with caution. Our results suggest that, even when trained on cross-sectional data, GANs can recover smooth and anatomically plausible variations of drusen that are in agreement with changes over time that are known from longitudinal observations.

CCS Concepts

• ***Computing methodologies*** → ***Computer vision; Neural networks;*** • ***Applied computing*** → ***Health informatics;***

1. Introduction

Age-related macular degeneration (AMD) is a medical condition that leads to a gradual worsening of vision in the center of the visual field with age. It can severely affect the ability to perform daily life activities and worsen a person's quality of life. Important diagnostic markers of AMD are *drusen*, which are deposits or clusters of extracellular material between Bruch's membrane and the retinal pigment epithelium. Imaging drusen cross-sections with spectral domain optical coherence tomography (SD-OCT) has provided clinicians with a classification of these deposits into different sub-types. Yet, a more detailed investigation of the drusen ultrastructure might improve our ability to understand and track the disease, or even to predict its progression [KKIT08].

Recent research on interpretable generative adversarial networks (GANs) has demonstrated their ability to discover relevant modes of variation in natural images. For instance, these variations include changes in hair length and color, gender, head pose, or skin tone in

DOI: 10.2312/vcbm.20241187

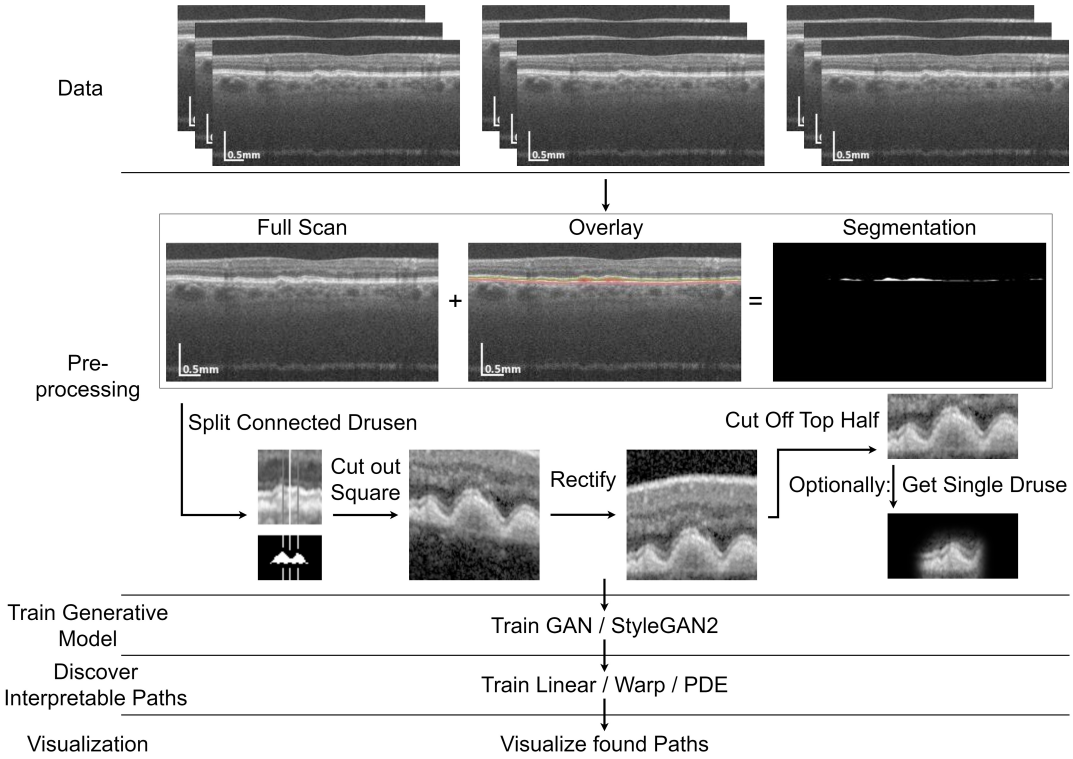


Figure 2: *Schematic overview of our proposed pipeline.*

images of human faces [VB20, TTP21, SKSW23]. In this work, we explore the potential of such methods to smoothly and interpretably visualize the variability in drusen appearance, as shown in Figure 1.

In this work, we introduce an *algorithmic pipeline* that involves the generation of suitable image patches based on drusen instance segmentation. Within this scope, we also compare alternative approaches to training interpretable GANs in an unsupervised manner, including paths that are linear or nonlinear in latent space. We conclude that GANs are able to synthesize OCT images of drusen with realistic variations that resemble those observed in longitudinal studies, despite the fact that our training set was cross-sectional.

2. Related Work

Previous work has used GANs on OCT images for a variety of applications, including anomaly detection, domain adaptation, super-resolution and de-noising, and data augmentation for classification or segmentation tasks [KACRC22]. GANs have also been used to produce synthetic fundus images related to AMD [BJP*19]. To our knowledge, we are the first to use GANs to generate synthetic images of drusen in OCT and to identify interpretable modes of variation in their appearance.

Several efforts have been made to discover and manipulate latent spaces in GANs that control and interpret variations in generated images—very few of which focus on finding interpretable features on medical images [FBK*20, SSP22]. Within other general domains of application, Voynov and Babenko [VB20] investigate linear paths, Tzelepis et al. [TTP21] search for non-linear paths based on learned warping functions, and Song et al. [SKSW23] use potential flows for location dependent paths. In their approach called AdvStyle, Yang et al. [YCW*21] search in the style space of StyleGANs. Additionally to binary attributes, such as gender and hair length they also find non-binary attributes, such as anime style. Another way to find interpretable directions in GANs is described by Härkönen et al. [HHLP20]. They use Principal Component Analysis (PCA) on a specific layer to find directions changing this layer. As the features are found layer-wise, it is possible to distinguish between high and low-level features. Interestingly, this technique even finds biases in trained GANs. For example, adding makeup changes images of women a lot, but only a little for men.

3. Methodology

Our proposed pipeline is illustrated in Figure 2. It creates image patches around the individual drusen in an internal, not publicly available OCT dataset, trains a GAN architecture on those patches, and discovers interpretable paths in their latent space. Our implementation (with a detailed explanation of all steps) is available under https://github.com/ChristianMuth5/IntDrusen.

Drusen are computed based on a state-of-the-art automatic segmentation of Bruch's membrane (BM) and the retinal pigment epithelium (RPE) [MWFS23]. The BM can intersect the patch at various angles, which is however not helpful to distinguish between drusen subtypes. Therefore, we remove this undesired variability by rectification, moving the individual columns of the image patch down to align the BM with the lower boundary. In particular, each A-scan (image column) is shifted

such that a Gaussian-smoothed ($\sigma = 5$) version of the BM becomes a horizontal line. Drusen in close proximity are often merged in the segmentation mask as shown in Figure 2. To isolate drusen instances, we look for peaks in the RPE height with a minimum distance of 20 pixels and a minimum height of 3 pixels. Drusen are separated if the valley between two neighboring peaks is at least 9 pixels below both peaks.

Subsequently, we cut out an image patch centered on each druse. Since rectification moves the drusen to the bottom half of the initially square patch, we can remove the top half. The resulting patches might still contain neighbors of the drusen on which they have been centered. We create two variants of the training dataset: In the first, we keep those patches as they are, which allows us to track interactions between neighboring drusen. In the second, our goal is to focus on the morphology of individual drusen. We, therefore, use the segmentation mask of the central druse and smoothly blend out the surroundings.

On these two datasets, we trained two types of generative models, DCGAN [RMC15] and StyleGAN2 [KLA*20]. These were selected as being the most common in the chosen papers for finding interpretable paths. Both follow the GAN approach described by Goodfellow et al. [GPAM*14]. Since StyleGANs require square images, we pad our rectangular ones with black bars on the top and bottom. To increase the variance in our dataset, we augment it by randomly flipping it across the vertical axis. However, to make sure the GAN captures the natural variability in the data, rather than our augmentations, we refrain from using augmentations that change the pixel values or that might introduce features that are not present naturally, such as elastic deformations.

Drusen sizes are not well distributed in our dataset, with over 90% belonging to the smallest of 10 classes. To investigate the impact of the data distribution, we train both without and with weighted sampling on epoch level. For the weighted sampling case, we split our data into 5 classes, corresponding to drusen volume, which is the number of pixels in the segmentation mask, respectively $[0,100),[100,300),[300,600),[600,1500),[1500,\infty)$. The DCGAN has been trained for 100 epochs with binary cross entropy loss and a latent vector size of 20 as the dataset is not very complex. The StyleGAN has been trained for 1000 epochs with and without weighted sampling, and 2000 epochs without weighted sampling. Other hyperparameters have been selected as `gamma=0.1024`, `cbase=16384`, `mapdepth=2`, `glr=0.0025`, and `dlr=0.0025`. All hyperparameters used in this study were selected based on the recommendations from the respective papers, where the employed models have been first presented. The number of epochs was kept low to accommodate the testing of various combinations.

On each of those generative models, we train three methods for discovering interpretable paths in latent space: the *GANLatentSpace* [VB20], *WarpedGANSpace* [TTP21], and *PDETraversal* [SKSW23]. We hereby refer to them as *Linear*, *Warp*, and *PDE*, respectively. We apply them to the latent space $\mathcal{Z}$ of both generative models and the style space $\mathcal{W}$ of the StyleGAN. The three selected approaches differ in how paths are calculated. Linear searches for linear paths given by linear directions in space. Warp discovers non-linear paths that follow the gradients of Radial Basis Function (RBF)-based warping functions of the space. PDE views the multi-dimensional input space as a dynamic potential landscape, in which the paths can be understood as gradient flows. This allows the paths to flow differently depending on the initial location of the input vector. The parameters were chosen following the suggestions from the authors of the corresponding papers where the employed models have been described. For choosing the number of epochs used in each approach, we have them run for a similar time. This results in 50,000, 20,000, and 100,000 for GANLatentSpace, WarpedGANSpace, and PDETraversal when trained on the DCGAN, and respectively 20,000, 20,000, and 50,000 when trained on the StyleGAN2.

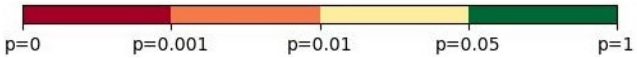


Figure 3: *Proposed mapping of p-values to color. Green indicates that the latent code is likely well-represented by the training data, while red indicates less reliable alignment with the training data, so that the resulting images should be interpreted with caution.*

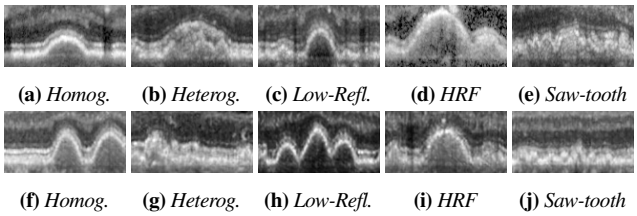
Finally, the resulting paths are visualized as animations or by placing keyframes side by side. A path can be seen as a function $f(\mathbf{x},d)$, with original latent space position $\mathbf{x}$ and a signed distance d along the path. When showing frames side by side, we show 5 images in each row: $f(\mathbf{x},-2\epsilon)$, $f(\mathbf{x},-\epsilon)$, $f(\mathbf{x},0)$, $f(\mathbf{x},\epsilon)$, and $f(\mathbf{x},2\epsilon)$, where ϵ is a step size parameter. This is done for all paths we show. Animations show images within the same range, but sample d more densely.

Since the latent space of the generative models is sampled during training with a multivariate Gaussian prior, latent codes with low probability under that Gaussian are no longer reliably aligned with the distribution of the training data, and resulting images should therefore be interpreted with caution. We detect those cases by computing p-values of a Chi-square test that accounts for the Gaussian prior. For better visualization we propose a colored bar on top of images depicting the p-values, following the color bar in Figure 3, indicating images from green to red depending on the level of data support. These color bars are missing for paths in style space $\mathcal{W}$, where the training distribution cannot be specified in closed form, so there is no simple way to calculate p.

4. Results

We present results from training our system on a set of 112 OCT scans of patients with drusen, taken from a dataset that was previously used for drusen segmentation [MWFS23]. Each scan consists of around 100 B-scans (2D slice images), with a resolution of 496×512 pixels. The raw data consists of multiple equally spaced B-scans per patient that stem from SD-OCT. To verify that our method is capable of generating a diverse set of drusen, we used an established classification [KKIT08] to select drusen that were present in the training data and manually searched for respectively generated drusen. A comparison of drusen found in the training and generated data is shown in Figure 4.

We trained 10 different generative models. Eight were combinations of GAN/StyleGAN, with/without weighted sampling, on images of multiple/single drusen. Additionally, 2 StyleGANs were trained for 2,000, instead of 1,000 epochs, without weighted sampling, on images of multiple/single drusen. Weighted sampling increases the number of large drusen but also leads to some images being repeated. As diverse drusen types are present without weighted sampling, we conclude that it is not worth training with weighted sampling. We found no obvious qualitative difference between training StyleGANs for 1,000 or 2,000 epochs.



(a) *Homog.* **(b)** *Heterog.* **(c)** *Low-Refl.* **(d)** *HRF* **(e)** *Saw-tooth*

(f) *Homog.* **(g)** *Heterog.* **(h)** *Low-Refl.* **(i)** *HRF* **(j)** *Saw-tooth*

Figure 4: *Drusen examples from different categories for our training data (top row) and respective images generated with a StyleGAN2 without weighted sampling (bottom row).*

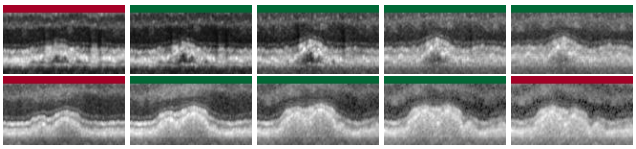


Figure 5: *A path showing the growth of multiple drusen, taken from a StyleGAN trained for 2000 epochs without weighted sampling. The path model is Linear in $\mathcal{Z}$ space. For interpretation of the color bars, please see Figure 3.*

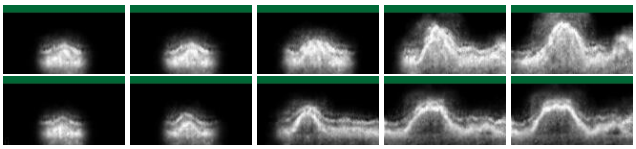


Figure 6: *A path showing the growing path for a single drusen, taken from the GAN trained for 100 epochs with weighted sampling. The path model is PDE in $\mathcal{Z}$ space. For interpretation of the color bars, please see Figure 3.*

We visually examined paths in the latent spaces of the 4 GANs and in the latent and style spaces of the 6 StyleGANs. The resulting 16 spaces were searched with all three methods (Linear, Warp, and PDE), resulting in 48 path models. Drusen size was a feature that we observed in many paths. Examples can be seen for multiple drusen in Figure 5, and for single drusen in Figure 6. Figure 1 shows two paths where, in addition to the middle drusen, also the right drusen change in size. These smooth variations are anatomically plausible, since an increase of drusen area and volume, as well as a confluence of drusen over time, have been observed in a longitudinal study [YWR*11].

In the multiple drusen setting, shifts in sizes are another type of feature we observed. Their corresponding paths can be described as different drusen in an image, where one is growing and the other(s) is (are) shrinking. Figure 7 shows a path where the left drusen is growing, and others are shrinking. In this path at the beginning, all drusen seem to grow. Shifts in sizes also occur mixed with other features. An example of another feature is given in Figure 8: a change in overall image brightness. This has been observed rarely compared to size, with less than one path per path model showing this feature.

When working with StyleGANs, we obtained more convincing results when working in latent space $\mathcal{Z}$ than in style space $\mathcal{W}$. In particular, when applying the PDE and warp based methods in $\mathcal{W}$, it

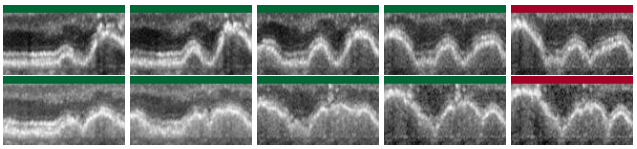


Figure 7: *A path showing the shift in sizes for multiple drusen, taken from the StyleGAN trained for 2000 epochs without weighted sampling. The path model is Linear in $\mathcal{Z}$ space. For interpretation of the color bars, please see Figure 3.*

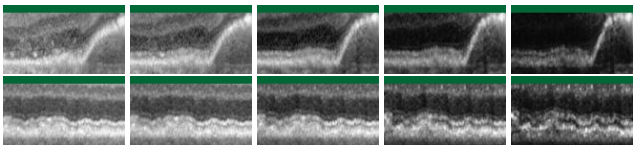


Figure 8: *A path showing the lighting feature for multiple drusen, taken from the StyleGAN that was trained for 2000 epochs without weighted sampling. The path model is Warp in $\mathcal{Z}$ space. For interpretation of the color bars, please see Figure 3.*

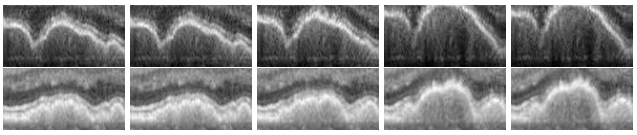


Figure 9: *A path showing abrupt change feature for multiple drusen, taken from the StyleGAN trained for 1000 epochs with weighted sampling. The path model is PDE in $\mathcal{W}$ space.*

was sometimes difficult to see any changes along paths, or paths would show sudden fast changes, as indicated in Figure 9. Here, images stay almost the same for some time, then suddenly change, e.g., showing slight growth, and after that do not change anymore.

The StyleGANs sometimes (about 1% of cases) show artifacts such as bright spots, repeating patterns, or having drusen drawn on the black bar at the top. Similarly, when searching StyleGANs in $\mathcal{Z}$ space with the PDE-based approach, we occasionally encountered blurry images that no longer depict retinal layers or drusen, as in Figure 10. This also happened for three out of four GANs when searching with Linear. We also observed other artifacts, such as missing parts of the image. Examples can be seen in Figure 10 in the first two images of the second row. We believe that they are a result of the well-known difficulties of training GANs, which occasionally lead to inaccurate or unrealistic images. Fortunately, these cases are rare and easily ignored by experts when interpreting our visualizations.

According to the p-values of the Chi-square test, the paths in Figures 5 and 7 lead into parts of latent space that are no longer well-supported by the training data. Both were created by the Linear method. The same observation held true for all paths of all generative models that were found with the Linear method, so we conclude that the PDE and Warp methods are better at staying in the part of the space the generative model was trained on. In our analysis of paths with implausible images such as Figure 10, about half stay in space, while the other half leads slightly out of it, with $p \approx 0.01$.

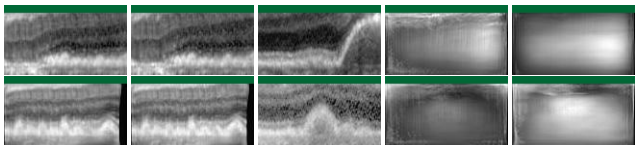


Figure 10: *Illustrations of the rare remaining failure cases of the StyleGAN (2000 epochs without weighted sampling). The path model is PDE in $\mathcal{Z}$ space. In both rows, the last two images no longer show plausible retinal structures. In the second row, the first two images are missing information on the right image boundary. For interpretation of the color bars, please see Figure 3.*

Based on our experimental comparison, we conclude that paths generated by the PDE and Warp methods yield good results, showing diverse drusen subtypes that change smoothly and contain consistent, human-interpretable features. This was also true when searching StyleGANs with the Linear method in $\mathcal{Z}$ space, while searching StyleGANs with Linear in $\mathcal{W}$ space did not work well. Overall, pathfinding was successful in disentangling different modes of variation, even though some paths remained showing a combination of multiple human-interpretable features.

5. Conclusion

Our work is the first to use GANs to explore the generation of drusen images and for the data-driven discovery of human-interpretable features that define their appearance. We successfully generated synthetic images of diverse types of drusen and discovered smooth and realistic modes of variation in them. We tested GAN and StyleGAN models with varying training parameters and found that both types of models are useful for this task.

While standard GANs can be trained faster, StyleGANs have been widely investigated in the context of GAN inversion [XZY*22], which would allow us to identify points in the GAN's latent space that amount to an approximation of a given real drusen image. Even though our current dataset is cross-sectional, i.e., it does not track the same patients over a longer period of time, it does contain a large number of drusen at diverse stages of the disease. We believe that this allowed our method to discover paths that correspond to previously described changes of drusen over time, in particular, changes in size and confluence of smaller drusen into larger ones.

Our main interest is to construct latent spaces that can serve as a basis for modeling the temporal evolution of drusen, and to ultimately aid the prediction of overall disease progression. Discussing our current results with our clinical co-authors indicates that, even when trained on cross-sectional data, GANs are able to learn latent representations of drusen in which continuous paths reflect plausible changes over time. A more comprehensive and quantitative evaluation will require longitudinal data, which was not available for the current study.

Acknowledgments

Partly funded by the Federal Ministry of Education and Research within the project "BNTrAinee" (funding code 16DHBK1022).

References

[BJP*19] Burlina P. M., Joshi N., Pacheco K. D., Liu T. A., Bressler N. M.: Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA ophthalmology 137*, 3 (2019), 258–264. 2

[FBK*20] Fetty L., Bylund M., Kuess P., Heilemann G., Nyholm T., Georg D., Löfstedt T.: Latent space manipulation for high-resolution medical image synthesis via the stylegan. *Zeitschrift für Medizinische Physik 30*, 4 (2020), 305–314. 2

[GPAM*14] Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y.: Generative adversarial networks, 2014. 3

[HHLP20] Härkönen E., Hertzmann A., Lehtinen J., Paris S.: Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems 33* (2020), 9841–9850. 2

[KACRC22] Kugelman J., Alonso-Caneiro D., Read S. A., Collins M. J.: A review of generative adversarial network applications in optical coherence tomography image analysis. *Journal of Optometry 15* (2022), S1–S11. 2

[KKIT08] Khanifar A. A., Koreishi A. F., Izatt J. A., Toth C. A.: Drusen ultrastructure imaging with spectral domain optical coherence tomography in age-related macular degeneration. *Ophthalmology 115*, 11 (2008), 1883–1890. 1, 3

[KLA*20] Karras T., Laine S., Aittala M., Hellsten J., Lehtinen J., Aila T.: Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 8110–8119. 3

[MWFS23] Morelle O., Wintergerst M. W., Finger R. P., Schultz T.: Accurate drusen segmentation in optical coherence tomography via order-constrained regression of retinal layer heights. *Scientific Reports 13*, 1 (2023), 8162. 2, 3

[RMC15] Radford A., Metz L., Chintala S.: Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015). 3

[SKSW23] Song Y., Keller A., Sebe N., Welling M.: Latent traversals in generative models as potential flows. *arXiv preprint arXiv:2304.12944* (2023). 1, 2, 3

[SSP22] Schön J., Selvan R., Petersen J.: Interpreting latent spaces of generative models for medical images using unsupervised methods. In *MICCAI Workshop on Deep Generative Models* (2022), Springer, pp. 24–33. 2

[TTP21] Tzelepis C., Tzimiropoulos G., Patras I.: Warpedganspace: Finding non-linear rbf paths in gan latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 6393–6402. 1, 2, 3

[VB20] Voynov A., Babenko A.: Unsupervised discovery of interpretable directions in the GAN latent space. In *Proceedings of the 37th International Conference on Machine Learning* (13–18 Jul 2020), vol. 119 of *Proc. of Machine Learning Research*, pp. 9786–9796. 2, 3

[XZY*22] Xia W., Zhang Y., Yang Y., Xue J.-H., Zhou B., Yang M.-H.: Gan inversion: A survey. *IEEE transactions on pattern analysis and machine intelligence 45*, 3 (2022), 3121–3138. 5

[YCW*21] Yang H., Chai L., Wen Q., Zhao S., Sun Z., He S.: Discovering interpretable latent space directions of gans beyond binary attributes. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 12177–12185. 2

[YWR*11] Yehoshua Z., Wang F., Rosenfeld P., Penha F., Feuer W., Gregori G.: Natural history of drusen morphology in age-related macular degeneration using spectral domain optical coherence tomography. *Ophthalmology 118*, 12 (2011), 2434–2441. 4