A Geometric and Photometric Exploration of GAN and Diffusion Synthesized Faces

Matyáš Boháček
Gymnasium of Johannes Kepler
Czech Republic
matyas.bohacek@matsworld.io

Hany Farid
University of California, Berkeley
Berkeley, CA USA
hfarid@berkeley.edu

Abstract

Classic computer-generated imagery is produced by modeling 3D scene geometry, the surrounding illumination, and a virtual camera. As a result, rendered images accurately capture the geometry and physics of natural scenes. In contrast, AI-generated imagery is produced by learning the statistical distribution of natural scenes from a large set of real images. Without an explicit 3D model of the world, we wondered how accurately synthesized content captures the 3D geometric and photometric properties of natural scenes. From a diverse set of real, GAN- and diffusion-synthesized faces, we estimate a 3D geometric model of the face, from which we estimate the surrounding 3D photometric environment. We also analyze 2D facial features – eyes and mouth – that have been traditionally difficult to accurately render. Using these models, we provide a quantitative analysis of the 3D and 2D realism of synthesized faces.

1. Introduction

Following in the footsteps of computer-graphics giant Ivan Sutherland, Martin Newell in 1975 created the now inescapable 3D Utah Teapot that for decades served as a benchmark in computer graphics [4]. Since this time, computer-generated imagery (CGI) has perfected the process of building 3D geometric models, texture-mapping these models, illuminating them with complex lighting, and rendering them to yield highly photo-realistic imagery.

The past few years have seen a radical revolution in photo-realistic rendering. AI-synthesized content has jettisoned the explicit construction of 3D models, lighting, and virtual cameras, and has instead leveraged massive 2D image datasets to effectively learn the statistical distribution of natural images. This type of neural-based rendering is producing stunningly realistic images.

The first AI-generated art was created in 1975 when Harold Cohen created the rule-based AARON program for generating abstract paintings, often compared to the drip paintings of Jackson Pollock [9]. In the intervening decades, AI-generated art has moved away from the classic rule-based approach towards a machine-learning, data-driven approach. This latest revolution in AI-generation was spurred by the development of generative adversarial networks (GANs) [18]. These neural-based computations consist of two main components: a generator and a discriminator. Tasked with, for example, synthesizing an image of a person, the generator starts by laying down a random array of pixels. If the discriminator can distinguish this proffer from a large database of real faces, it provides feedback to the generator for a second round. This process repeats until the discriminator is unable to distinguish the generator's synthesized face from a real face.

Versions 1, 2, and 3 of StyleGAN [22–24] and its precursor ProGAN [21] are some of the most successful techniques for synthesizing realistic faces. Previous work (e.g., [17]) has found that these types of synthesized images contain subtle spectral patterns not found in real photographs (e.g., [41]). While these properties can be forensically exploited (albeit with some limits [12]), they don't impact the visual plausibility of the synthesized faces. And, in fact, a recent set of perceptual studies [28] found that StyleGAN2 faces are nearly indistinguishable from real faces (and even slightly more trustworthy). By comparison, as late as 2016 – with decades longer to perfect photo-realistic rendering – classic CGI-rendered faces were still somewhat distinguishable (albeit not perfectly) from photographic faces [19].

Although highly realistic, StyleGAN does not afford much control over the appearance or surroundings of the synthesized face. By comparison, more recent diffusion-based synthesis affords more rendering control [3, 30, 33]. Trained on hundreds of millions of images (and accompanying text descriptions), each image is progressively corrupted until only visual noise remains. The model then learns to denoise each image by reversing this corruption. This diffusion model can then conditioned to generate an image that

Figure 1. Representative examples of (a) real; (b) GAN-generated; and (c) diffusion-generated faces.

is semantically consistent with a specified category ("cat," "dog," "landscape," etc.), or a more detailed description ("a cat and dog riding a rainbow-colored unicorn on a rolling green landscape").

For example, Google's Imagen[1] is a 4.6-billion parameter text-to-image diffusion model, OpenAI's DALL-E[2] is a 3.5-billion parameter model, and Stability AI's open-source Stable Diffusion[3] with just under 1 billion parameters has made text-to-image synthesis readily accessible.

Given the evolution in image rendering from detailed modeling of 3D geometry and physics to 2D data-driven approaches, we wondered if AI-synthesized images exhibit the same veridicality as CGI images. We describe a series of 3D and 2D analyses of GAN- and diffusion-generated faces in an initial exploration of the 3D geometric, 3D photometric (lighting), and 2D facial-feature – eyes (oculometric) and mouth (oralmetric) – realism of AI-synthesized faces. This study has implications for those on the synthesis side (CGI and AI) and on the forensic side of this revolution in image rendering.

Throughout our analysis we intentionally utilize unadorned computational machinery to analyze the 2D and 3D structure of real and synthesized faces. The rationale for this is that our focus is not to build robust forensic classifiers, but rather to explore the underlying realism of AI-synthesized content. By using basic – mostly linear – techniques, we can discover underlying consistencies and inconsistencies while avoiding latching onto more minor features that may be useful for forensic classification, but are not the focus of our analysis.

[1] https://imagen.research.google
[2] https://openai.com/blog/dall-e
[3] https://stability.ai/blog/stable-diffusion-public-release

2. Related Work

We next review the most comparable works to ours as described in a series of studies examining the semantic and geometric plausibility of DALL-E 2 images.

The first study examined the ability of DALL-E 2 to capture basic relations between simple objects and agents [10]. The physical relations included terms like *in, on, under, near*; the agentic relations included terms like *pushing, pulling, touching, hitting*; and the objects and agents included terms like *box, bowl, teacup* and *child, robot, iguana*. By mixing and matching these terms, sample DALL-E 2 text prompts were created like "a child touching a bowl" and "a spoon in a teacup." Based on perceptual judgements of synthesized images, the authors conclude that the models do not have a complete understanding of basic relations between simple objects and agents.

A second study examined the ability of DALL-E 2 to capture basic grammatical phenomena in human language including *coordination, comparative, negation* [26]. Sample prompts include phrases like "The man is drinking water and the woman is drinking orange juice," "The bowl has more cucumbers than strawberries," "A tall woman without a handbag." The authors conclude that the models do not have a complete understanding of these basic grammatical structures that are generally well understood by young children. A third study found similar gaps in semantic understanding in DALL-E 2 synthesized images [32].

Moving from the semantic to the geometric, a pair of papers examining the consistency of perspective constructs (vanishing points and cast shadows) [15] and lighting [14] found that DALL-E 2 has some – albeit imperfect – understanding of basic geometric and photometric constructs. In particular, it was found that the perspective geometry of vanishing points on planar surfaces (a tiled kitchen floor and counter) and cast shadows (from cubes on a sidewalk) are

locally consistent, but globally (across the entire scene) inconsistent. It was also found that 3D lighting environments on rendered spheres in outdoor settings are again locally consistent, but globally inconsistent.

Our analysis is related to these geometric analyses, but here we focus exclusively on images of faces, arguably the type of content that has raised most concern for potential misuse [5, 6].

3. Datasets

Our dataset consists of 1200 facial images: 400 real; 400 GAN-generated; and 400 diffusion-generated. Shown in Figure 1 are representative examples of these real and synthesized faces.

The real and GAN-based (StyleGAN2 [24]) images are taken from the perceptual study of [28]. The GAN-based images are equally distributed (50 per category) across two apparent genders (women and men) and four apparent races (African American or Black, East Asian, and South Asian, White). A latent representation (VGG [29]) was extracted from each GAN-generated face, from which each synthetically-generated face was matched to a real face (from the StyleGAN2 training dataset) with the closest latent representation (in the $\ell 2$-norm sense). These synthesized and real color images are of size 400×400 pixels.

Adding to this dataset, we used Stable Diffusion (v1.4) [34] to synthesize 50 faces for each of eight demographics with the prompts "a profile photo of a middle-aged {Black, East-Asian, South-Asian, White} {Woman, Man} with a solid background." The images were synthesized at a resolution of 512×512 pixels. We manually replaced any obvious synthesis failures in which, for example, the face was not visible or the face had obvious and significant rendering artifacts.

4. Geometric

We employ the neural-based Detailed Expression Capture and Animation (DECA) model [16] to extract a 3D geometric model from a single RGB image (we also considered the more recent Metric Face (MICA) [43], but found it did not impact our analyses). These models capture individual facial structural differences and expressions. The 3D model is parameterized as a standard mesh with $n = 5023$ vertices and $m = 9976$ faces connecting triples of vertices. Shown in Figure 2 are representative examples of estimated models from three images described in the previous section.

In order to explore possible 3D structural differences between real and synthesized faces, each 3D model is first aligned to a single reference model. This is done by first translating the model to the origin and isotropically scaling it to fit within a unit sphere. The model is then rigidly aligned (rotated and translated) to a specified reference

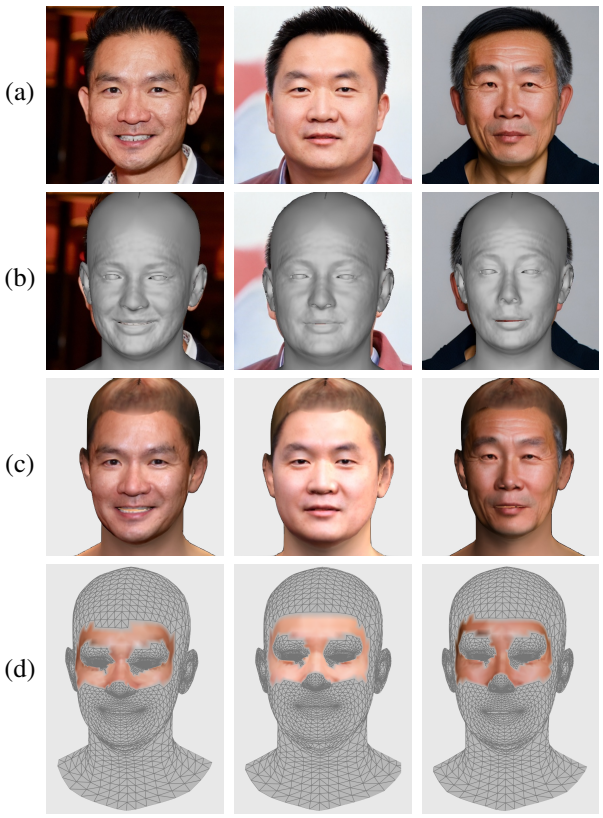

Figure 2. Representative examples of: (a) RGB input image; (b) DECA-generated 3D model superimposed atop the input image; (c) fully the texture-mapped model; and (d) partially texture-mapped model with the portion of the facial texture used by the photometric analysis.

model. Specifically, the 3D rotation matrix R and 3D translation vector $\vec{t}$ that align the 3D vertices ($\vec{q}$) of a model to the corresponding reference vertices ($\vec{p}$) are determined by minimizing the following quadratic error:

$$E(R, \vec{t}) \;=\; \sum_{i=1}^{n} \left(\vec{p}_i - (R\vec{q}_i + \vec{t})\right)^2 . \qquad (1)$$

Note that because the underlying DECA model is derived from a fixed 3D model with n vertices, the i^{th} vertex of the reference model ($\vec{p}_i$) corresponds to the i^{th} vertex of the model to be aligned ($\vec{q}_i$). By using quaternions to represent the transformation (R, $\vec{t}$), a standard least-squares estimation can be used to estimate the optimal rigid alignment [20].

Once aligned, the set of 400 real and 400 GAN-generated 3D models are subjected to a principal component analysis (PCA) [1], where each model is represented as a $3n \times 1$ vector corresponding to the n 3D coordinates of the model's vertices. We find that the top 15 principal components (PCs) capture 99% of the variance of these 800

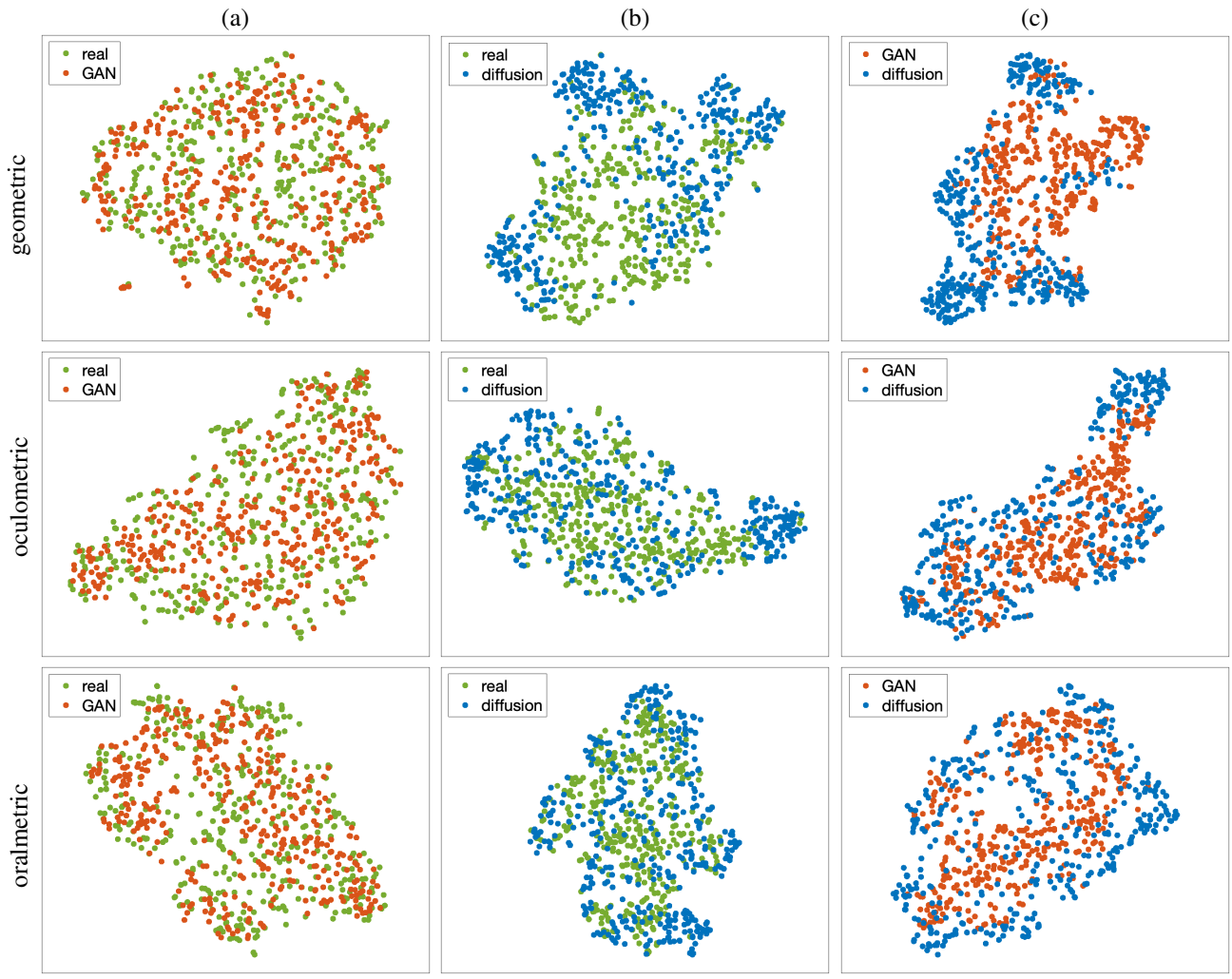

Figure 3. A t-SNE representation of the 15-D PCA parameterization of the facial geometry (top row), the 23-D PCA parameterization of the oculometric shape (middle), and the 22-D PCA parameterization of the oralmetric shape (bottom) comparing (a) real and GAN-generated faces, (b) real and diffusion-generated faces, and (c) GAN- and diffusion-generated faces.

models. Each model is projected onto these top 15 PCs to yield a 15-D geometric representation of each face.

Shown in the top row of Figure 3(a) is a 2D t-SNE [38] projection of this PC representation, from which we see no discernible grouping of the real and GAN faces.

To further explore the potential differences between the 3D geometry of real and GAN faces, we next trained a logistic regression (LR) on the 15-D PC representations. Given our relatively small dataset, here – and throughout – we report on the training accuracy; given the use of a simple LR classifier, the chance of over-fitting is less significant.

Averaged over 100 repetitions, the average classification accuracy for the real and GAN faces is 65.1% and 66.3% with a standard deviation of 1.1% and 1.0% (Table 1, row 1), where chance classification in this balanced dataset is 50%. With accuracy only slightly better than chance, we see that the 3D geometry of GAN-generated faces is generally consistent with real faces.

This entire process was repeated for the set of 400 real and 400 diffusion-generated 3D models. Shown in the top row of Figure 3(b) is the t-SNE projection of this PC representation, from which we see a bit more clustering of the real and diffusion faces as compared to the GAN faces in panel (a). This is confirmed by the average LR classification accuracy increasing to 77.9% and 83.4% for the real and synthetic faces with a standard deviation of 1.2% and 0.7% (Table 1 row 1).

We also compared the GAN- to the diffusion-synthesized images. Shown in the top row of Figure 3(c) is the t-SNE projection of this PC representation, revealing partial grouping similar to the real and diffusion-generated faces in panel (b). The average LR classification confirms this

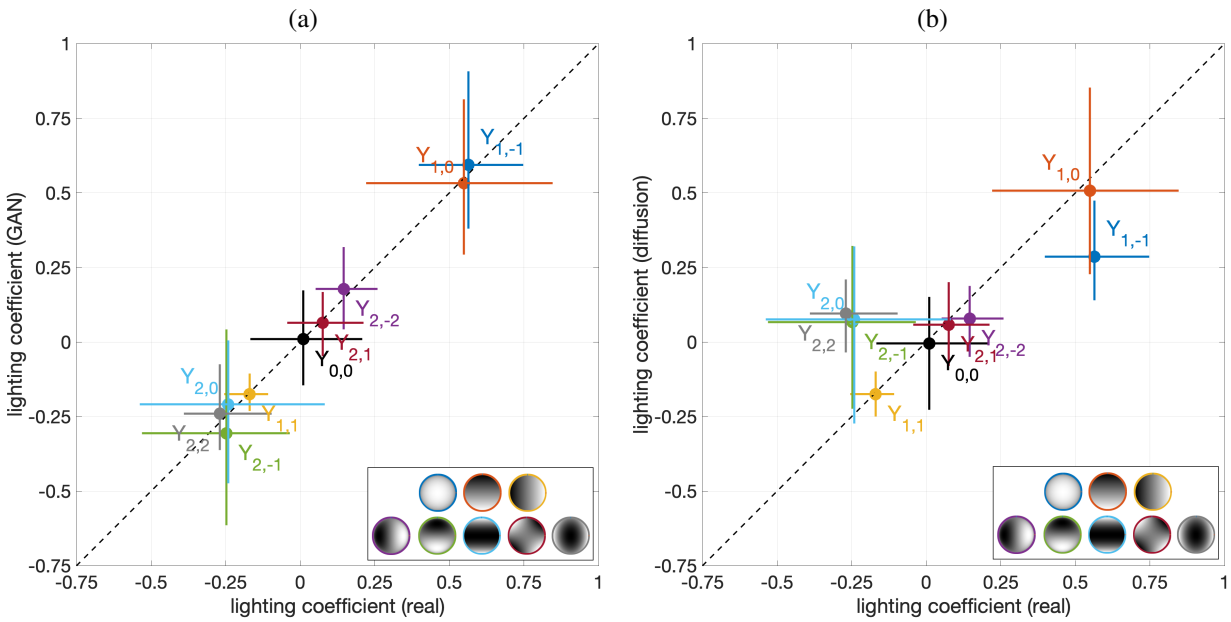

Figure 4. Median, and 35% and 65% quantiles, of the nine lighting coefficients for the (a) GAN- and (b) diffusion-generated faces, as compared to the real faces. $Y_{n,*}$ corresponds to the n^{th}-order spherical harmonics: shown in the legend are $Y_{1,*}$ (top) and $Y_{2,*}$ (bottom).

grouping with an average accuracy of 81.9% and 85.3% for the GAN and diffusion faces with a standard deviation of 0.5% and 0.6% (Table 1 row 1).

As compared to diffusion-generated faces, GAN-generated faces more closely mimic the 3D geometry of real faces, and there appears to be a geometric difference between GAN and diffusion faces.

5. Photometric

We next explore possible 3D environmental lighting differences between real and synthesized faces. These photometric properties are quantified using a spherical-harmonic representation [2, 31]. This method assumes a convex Lambertian surface of constant reflectance illuminated by a distant light source(s), and a known 3D geometry. Shown in Figure 2(d) are examples of texture-mapped 3D models (from Section 4) filtered to a small region around the eyes and nose that can reasonably be assumed to satisfy these assumptions (i.e., primarily uniform skin texture with no hair, facial hair, eyes, or teeth). The 2D textures are automatically filtered using MediaPipe's [27] facial keypoint estimation to isolate the region of interest.

From the facial appearance (texture) and underlying 3D geometry, we estimate the 9-D environmental lighting coefficients as described in [25].

Shown in Figure 4 are the median, and 35% and 65% quantiles, of the nine lighting coefficients for the GAN- and diffusion-generated faces as compared to the real faces. For the basis spherical harmonics depicted in the figure legend, the in-plane horizontal and vertical axis corresponds to the x-axis and z-axis (the camera optical axis) and the positive y-axis is facing into the page.

With a correlation of $r^2 = 0.99$, the photometric properties of GAN-generated faces are highly similar to real faces. With a correlation of $r^2 = 0.58$, diffusion-generated faces are less similar to real faces. Notable deviations are found in the first-order term ($Y_{1,-1}$) corresponding to the lighting from above or below the photographer, and three of the second-order lighting terms ($Y_{2,-1}$, $Y_{2,0}$, and $Y_{2,2}$) corresponding to differences in illumination in front of and behind the camera. These differences relate to what is perhaps the most common illumination patterns in natural photos: the dominant light source is usually from above, and photographers typically position themselves between their subject and the dominant light source to avoid the glare that results from back lighting. These patterns don't seem to be fully respected in diffusion-synthesized faces.

As in the previous section, we train a logistic regression on the 9-D lighting coefficients. Averaged over 100 repetitions, classification accuracy for the real and GAN faces is 58.4% and 59.4% with a standard deviation of 1.0% and 1.2%. Classification accuracy for the real and diffusion faces is significantly higher at 78.2% and 80.1% with a standard deviation of 0.7% and 0.9%. Lastly, classification between GAN and diffusion faces has a similar accuracy of 75.2% and 81.5% with a standard deviation of 0.6% for both classes (Table 1, row 2).

As compared to the 3D geometric properties described in the previous section, GAN-generated faces are even more consistent with real faces. On the other hand, the geometric and photometric properties of diffusion-generated faces are similarly distinct from real faces.

6. Oculometric

We next explore potential differences in the 2D eye shape in real and synthesized faces. MediaPipe's keypoint estimator [27] is used to identify the left and right eye. A bounding box is then specified encompassing the eye and eyelid. Each eye is then scaled to a fixed resolution of 20×32 pixels and converted from RGB to grayscale, Figure 5.

The left and right eyes, packed into a single image, for 400 real and 400 GAN faces are subjected to a PCA. We find that the top 23 principal components capture 99% of the variance. Each pair of eyes is projected onto these top 23 PCs to yield a 23-D representation of the ocular shape.

Shown in the middle row of Figure 3(a) is a 2D t-SNE projection of this PC representation. Although there is no apparent grouping, a logistic regression reveals some differences. Averaged over 100 LR repetitions, the accuracy for the real and GAN eyes is 71.6% and 76.6% with a standard deviation of 2.3% and 0.8% (Table 1, row 3). These shape differences are greater than the 3D geometric and 3D photometric differences.

This process is repeated for the set of 400 real and 400 diffusion eyes. Shown in the middle row of Figure 3(b), is the t-SNE representation, revealing more grouping. The average LR classification (again averaged over 100 repetitions) accuracy is 70.6% and 67.9% with a standard deviation of 0.7% and 0.9% (Table 1 row 3).

Lastly, we compare GAN- to diffusion-generated eyes. Shown in the middle row of Figure 3(c) is the t-SNE projection of this PC representation, revealing some grouping. The average LR classification confirms this grouping with an average accuracy of 83.8% and 75.7% for the GAN and diffusion eyes with a standard deviation of 1.4% and 2.3% (Table 1 row 3).

Overall, the 2D ocular features of GAN faces are less consistent with real images than the 3D geometric and photometric properties. On the other hand, ocular features of diffusion faces are more consistent with real images than the 3D properties.

7. Oralmetric

We next explore possible 2D mouth shape differences between real and synthesized faces. MediaPipe's keypoint estimator [27] is used to identify the mouth. A rectangular bounding box is specified horizontally by the corners of the mouth and vertically by MediaPipe's philturm and chin crease keypoints. Each bounding box is then scaled to 140×70 pixels, and converted from RGB to grayscale, Figure 5.

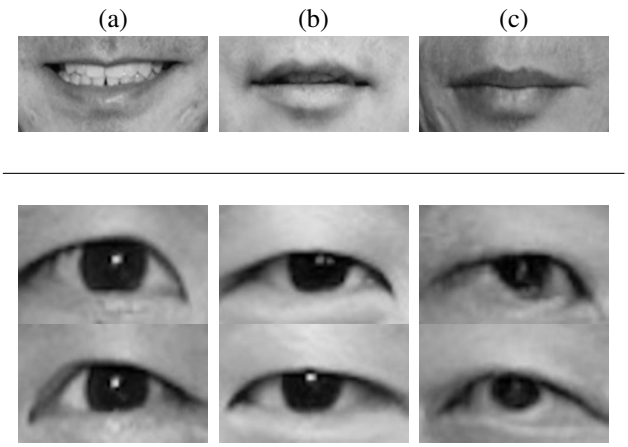

Figure 5. Representative examples of cropped (a) real, (b) GAN-generated, and (c) diffusion-generated mouths and eyes used in the oralmetric and oculometric analyses.

Subjecting the 400 real and 400 GAN-generated mouths to a PCA yields a basis of size 22 capturing 99% of the variance. Each mouth is projected onto these top 22 PCs, yielding a 22-D representation of the oral shape. Shown in the bottom row of Figure 3(a) is a 2D t-SNE projection of this PC representation. Although there is no apparent grouping, a logistic regression reveals some differences. Averaged over 100 LR repetitions, the accuracy for the real and GAN mouths is 65.8% and 68.0% with a standard deviation of 0.9% and 1.2% (Table 1, row 4).

This process was repeated for the set of 400 real and 400 diffusion faces. Shown in the bottom row of Figure 3(b), is the t-SNE representation, revealing more grouping. This is supported by an increase in the average LR classification of 72.0% and 71.5% with a standard deviation of 1.0% and 0.8% (Table 1 row 4).

Finally, we compared GAN- to diffusion-generated mouth. Shown in the bottom row of Figure 3(c) is the t-SNE projection of this PC representation and the average LR classification is 86.1% and 81.5% with a standard deviation of 0.6% for both classes (Table 1 row 4).

Overall, the trend from the previous sections continue: as compared to diffusion faces, GAN faces are more consistent with real faces. For GAN faces, the 3D geometric and photometric properties are more consistent with real faces than the 2D ocular and oral features; for diffusion faces, this is reversed with the 2D facial features more consistent than the 3D properties.

feature	real	GAN	real	diff	GAN	diff
geometric	65.1	66.3	77.9	83.4	81.9	85.3
photometric	58.4	59.4	78.2	80.1	75.2	81.5
oculometric	71.6	76.6	70.6	67.9	83.8	75.7
oralmetric	65.8	68.0	72.0	71.5	86.1	81.5
combined	75.0	78.1	74.6	74.5	84.8	82.0

Table 1. Two-way classification accuracy (%) to distinguish between real and GAN, real and diffusion, and GAN and diffusion faces. Accuracy is averaged over 100 LR repetitions trained on different facial features (rows 1-4) and a all features (row 5).

8. All together now

Across 3D geometric and photometric properties and 2D facial features, both GAN- and diffusion-synthesized faces share strong commonalities with real faces, with GAN faces exhibiting more realism than diffusion faces. At the same time, we see more significant differences between GAN and diffusion faces. In this final analysis, the 15 geometric, 9 lighting, 22 ocular, and 22 oral features are combined into a single LR classifier. As shown in the last row of Table 1, this combined classifier affords only a slight improvement in distinguishing real from synthesized. In particular, the combined LR accuracy for real versus GAN is only slightly better than the oculometric accuracy, and the combined LR accuracy for real versus diffusion is slightly worse than the 3D geometric and 3D photometric accuracy.

These results imply that the differences across the various 3D and 2D properties are not independent.

9. Discussion

We find that GAN-synthesized faces are more consistent with real faces than diffusion-synthesized faces with respect to 3D geometric and photometric properties and 2D oral facial features, while the 2D ocular features of diffusion faces are slightly more realistic. Somewhat counter to our findings, the authors in [11] find that diffusion models can achieve superior image quality to GANs. While we focus on 3D and 2D facial features, these authors evaluate a broad category of images using an inception score image-quality metric [35] which measures the diversity and distinctiveness of synthesized images. This different metric and categorical focus likely explains our different conclusions.

Our focus has been an exploration of the photo-realism of AI-synthesized faces in which we intentionally employ mostly linear techniques so as to focus on basic facial properties. We have little doubt that more sophisticated classifiers may be able to utilize these basic findings to yield a high-performing forensic classifier.

We have only considered two of the most popular synthesis techniques. It remains to be seen if related techniques like 3D-aware GANs [8] or GAN-based text-to-image [36] will produce similarly photo-realistic images.

Even in these relatively early days of synthetic media, AI-generated faces are highly realistic and have arguably surpassed the photo-realism of classic computer-generated imagery (CGI). Having jettisoned the need for highly detailed 3D models and computationally-intensive rendering, AI-generated content is also significantly less labor intensive (once the system has been trained). AI-generated content does require significantly more data. However, with massive datasets freely available (e.g., https://laion.ai), access to data is no longer a rate-limiting step.

On the other hand, synthetically-generated content can produce bizarre and implausible imagery. But, because synthesis is so effortless and fast, brute-force synthesis will eventually generate a desired and highly photo-realistic image. It seems likely therefore that AI-synthesized content will eventually surpass CGI in terms of usability and photo realism. It remains to be seen if a hybrid rendering approach can take advantage of CGI's fine control/physical models, and AI's ease/flexibility.

Regardless of the current state of synthesized content, we contend that if the trends continue, AI synthesis will eventually generate content that passes through the uncanny valley, yielding images (and eventually audio and video) that are perceptually indistinguishable from reality. This will no doubt be considered a major success for the machine learning and computer vision communities, but will also raise complex privacy, legal, and ethical questions and concerns.

On the privacy and legal fronts, recent investigations of diffusion-based models have revealed that they are capable of producing identical or nearly identical images found in the model's training set [7,37]. This apparent memorization has privacy and legal implications. On the privacy front, if – as has previously been reported [13] – sensitive images find their way into a model's training set, the synthesis process can leak this type of sensitive data. On the legal front, if a model's training set contains copyrighted images obtained without the appropriate permission [40], the regeneration of a copyrighted image could be considered a violation of intellectual property law.

On the ethical front, although OpenAI placed reasonable safeguards on DALL-E's ability to generate abusive, harmful, or NSFW content, Stability AI initially placed no such restrictions on their Stable Diffusion. As a result, almost immediately after its release, their synthesis engine was used to create all forms of NSFW imagery including those involving children. In response, the company's founder, Emad Mostaque, said "Ultimately, it's peoples' responsibility as to whether they are ethical, moral and legal in how they operate this technology" [39].

We disagree with this seeming avoidance of responsibility for how one's technology is being weaponized. There

are reasonable and practical measures that can be put in place that allow for continued technological advances while placing safeguards to mitigate predictable harms. With respect to downstream detection of synthetic media, invisible and robust watermarks can be embedded into synthesized content. These watermarks can be baked into the synthesis engines by watermarking all of the images in the training dataset, after which the synthesis engine will generate content that contains the same watermark(s) [42]. While certainly not a perfect solution, the task of mitigating harms from synthetic media should not be left to only the forensics community, but should also begin to be addressed on the synthesis side.

Acknowledgment

This work was supported in part by an unrestricted gift from Meta.

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. 3

[2] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003. 5

[3] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. arXiv:2103.10951, 2021. 1

[4] James F Blinn and Martin E Newell. Texture and reflection in computer generated images. *Communications of the ACM*, 19(10):542–547, 1976. 1

[5] Shannon Bond. AI-generated fake faces have become a hallmark of online influence operations. https://www.npr.org/2022/12/15/1143114122/ai-generated-fake-faces-have-become-a-hallmark-of-online-influence-operations, 2022. 3

[6] Shannon Bond. That smiling LinkedIn profile face might be a computer-generated fake. https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles, 2022. 3

[7] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. arXiv:2301.13188, 2023. 7

[8] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3D generative adversarial networks. In *International Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 7

[9] Paul Cohen. Harold Cohen and AARON. *AI Magazine*, 37(4):63–66, 2016. 1

[10] Colin Conwell and Tomer Ullman. Testing relational understanding in text-guided image generation. arXiv:2208.00005, 2022. 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 7

[12] Chengdong Dong, Ajay Kumar, and Eryun Liu. Think twice before detecting GAN-generated fake images from their spectral domain imprints. In *International Conference on Computer Vision and Pattern Recognition*, pages 7865–7874, 2022. 1

[13] Benj Edwards. Artist finds private medical record photos in popular ai training data set. Ars Technica (https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/), 2023. 7

[14] Hany Farid. Lighting (in)consistency of paint by text. arXiv:2207.13744, 2022. 2

[15] Hany Farid. Perspective (in)consistency of paint by text. arXiv:2206.14617, 2022. 2

[16] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH*, 40(4):88:1–88:13, Aug. 2021. 3

[17] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pages 3247–3258, 2020. 1

[18] 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

[19] Olivia Holmes, Martin S Banks, and Hany Farid. Assessing and improving the identification of computer-generated portraits. *ACM Transactions on Applied Perception*, 13(2):1–12, 2016. 1

[20] Berthold KP Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of Optical Society of America A*, 4(4):629–642, 1987. 3

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. arXiv:1710.10196, 2017. 1

[22] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Neural Information Processing Systems*, 2021. 1

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *International Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *International Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1, 3

[25] Eric Kee and Hany Farid. Exposing digital forgeries from 3-D lighting environments. In *IEEE International Workshop on Information Forensics and Security*, pages 1–6. Institute of Electrical and Electronics Engineers, 2010. 5

[26] Evelina Leivada, Elliot Murphy, and Gary Marcus. DALL-E 2 fails to reliably capture common syntactic processes. arXiv:2210.12889, 2022. 2

[27] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for building perception pipelines. arXiv:1906.08172, 2019. 5, 6

[28] Sophie J Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022. 1, 3

[29] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Association*, 2015. 3

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[31] Ravi Ramamoorthi and Pat Hanrahan. On the relationship between radiance and irradiance: Determining the illumination from images of a convex Lambertian object. *Journal of the Optical Society of America*, 18(10):2448–2459, 2001. 5

[32] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models. arXiv:2210.10606, 2022. 2

[33] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *International Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, June 2022. 3

[35] 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. 7

[36] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. arXiv:2301.09515, 2023. 7

[37] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? Investigating data replication in diffusion models. arXiv:2212.03860, 2022. 7

[38] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 4

[39] James Vincent. Anyone can use this AI art generator – that's the risk. The Verge (https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data), 2022. 7

[40] James Vincent. Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content. The Verge (https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit), 2023. 7

[41] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot... for now. In *International Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020. 1

[42] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *International Conference on Computer Vision and Pattern Recognition*, pages 14448–14457, 2021. 8

[43] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*. Springer International Publishing, October 2022. 3