Using Deep Generative Models for Glossy Appearance Synthesis and Exploration

Abhinav Reddy Nimma
Colourlab, Department of Computer Science
Norwegian University of Science and Technology
Gjøvik, Norway
abhinavn@stud.ntnu.no

Davit Gigilashvili
Colourlab, Department of Computer Science
Norwegian University of Science and Technology
Gjøvik, Norway
davit.gigilashvili@ntnu.no

***Abstract*—Generating images with realistic material appearance using a physically-based renderer demands significant time and human labor. The images are used in psychophysical experiments to study human perception of material appearance attributes, such as glossiness. Recently, deep learning-based image synthesis models have emerged as a promising approach for generating realistic images with less human supervision. Deep Generative Models are deep learning-based models that learn to generate unique and novel images based on a given training data distribution. Using them for image synthesis is fast and manually less tiresome. An additional benefit these Deep Generative Models offer is latent space encodings that may help to better understand the feature space of gloss and its perception. In this study, we propose to explore the possibility of using Deep Generative Models for realistic image synthesis, focusing on gloss appearance and evaluating the efficiency of such gloss generation process using psychophysical experiments. Additionally, we build tools to extract the latent space of generative models to use them as a feature space representation of gloss appearance and perception. Finally, we analyse the trends and patterns in the learnt feature space to aid gloss appearance modelling.**

***Index Terms*—Gloss Perception, Image Synthesis, Material Appearance Modelling, Learning a Feature Space Representation**

I. Introduction

Perception of material appearance and its properties is fundamental to humans for interacting with the environment. The human visual system (HVS) has complex and sophisticated mechanisms for appearance perception that are a product of millions of years of evolution and remain poorly understood [1], [2]. Gloss – together with color, texture, and translucency – is one of the fundamental attributes of how objects and materials look [3]. Although gloss is primarily understood as a surface reflectance property, the link between instrumentally measured and human perceived gloss is complex and non-monotonic [4], [5]. Multiple handcrafted features have been proposed to predict gloss appearance from image statistics [6]–[8], but handcrafted features are rarely robust enough to account for complex influences from shape, illumination, and observation geometry [9]–[11].

Perceptual studies often involve computer graphics to generate the experimental stimuli. The process of rendering images with glossy surfaces involves understanding the complex interactions between all the intrinsic (optical properties) and extrinsic (environmental) factors. Most images generated using physically-based renderers are labelled using the physical parameter values. This does not help us to understand how the human visual system deciphers gloss appearances and how each factor influences gloss perception in humans. We need a better representation for navigating the gloss appearance space. It is not easy to handcraft features for human gloss perception as it is not fully understood how the human visual system deciphers gloss appearance into individual factors [2], and more efficient feature space is needed. Apart from that, using a physically-based renderer (such as Mitsuba [12]) is both very time-consuming and human labor-intensive. It would be desirable to develop a way to render or generate images with a realistic gloss appearance that requires minimal supervision.

Deep Generative Models have shown promising results in generating realistic images. Image synthesis in deep learning refers to generating images using neural networks. Deep Generative Models are based on deep learning. They learn to generate novel images based on a training data distribution. They first learn to model the distribution in the images in the training data and then use the learnt patterns to generate novel images that are not part of the training dataset. Deep Generative Models are considered unsupervised as they neither need manual supervision during training nor annotations for the data they are being trained on. The learning process is data-driven, i.e., the models learn to form the given data without needing any target labels for the given data. They have demonstrated capabilities in generating realistic novel images that are not part of the training data. If we can generate realistic material appearance using Deep Generative Models, it would save us significant amounts of time and labor. Deep Generative Models try to develop an understanding of the statistical structure in the data distributions. In developing this understanding, Deep Generative Models develop a latent space representation for the data distribution. Thus, apart from aiding in generating images, they also help us encode images into a new latent space. The latent space of these models can be used as a representational space for material appearance attributes.

The models encode the input image into its internal latent space and then decode the latent vector from its internal latent space into output images. During training, the model optimises this encoding and decoding process and learns to model the

979-8-3503-4218-5/23/$31.00

statistical structure in the data distribution of the input images in its internal latent space. This way, in an unsupervised manner, we end up with a new feature space representation of the images in the training dataset. We can use this new feature space to better understand the dataset. It is believed that the HVS exploits statistical structure and regularities in the environment to derive information about our surroundings and develop perception and awareness of the world [13]. The development of latent space in Deep Generative Models is similar, and it is hypothesized that such feature space can eventually be used to model the perception of the HVS.

In this work, we trained a Deep Generative Model with low number of physically-based renderings of glossy objects and synthesized novel images with this model to check whether it can produce realistic images. We report the results of a psychophysical experiment that we conducted to assess the convincingness of the synthesized images. Afterward, we explore the latent space to understand the feature space of gloss and navigate through it in a meaningful manner.

II. RELATED WORK

Several attempts have been made in developing a feature representation for material appearance for surface gloss [7], [14], surface roughness [15], [16], transparency [17], [18], and translucency [19]–[21]. The studies use an analytical approach to find diagnostic image features for material perception. There is a significant challenge in this approach, since the features may not be stable across a broad range of intrinsic and extrinsic factors [1], [19]. An alternative approach in the diagnosis of features for material appearance is a data-driven one [22], [23]. These approaches attempt to extract features of material appearance by modeling the statistical distribution of material appearance across image samples. This approach has demonstrated great potential in modeling human perception [24]. Especially with the rapid progress of deep neural networks to learn patterns from enormous and diverse datasets, data-driven approaches show a significant potential in perception modeling [25]–[27]. Convolutional neural networks can be used to extract features from the images.

For long, deep learning-based techniques were used to analyse images for content objects etc. Recently, with the advancements in deep learning-based techniques, neural networks can generate images from random noise [28], seed [29], or text inputs [30], with remarkable realism. These networks can learn an image generation procedure from the training dataset's images. During training, they model the statistical structure in the distribution of images in the training set and construct an internal latent space representation for all the images in the training dataset. With models that generate accurate, realistic images, the internal latent space can be extracted and used as an efficient and compact feature representation of the distribution of images in the training dataset.

Generative Adversarial Networks (GANs) [31] is a breakthrough architecture on which most of the state-of-the-art Deep Generative Models are based. GANs consist of two deep neural networks: a discriminator and a generator. The task of the generator is to generate images from random input vectors, similar to the training data distribution. Discriminator judges whether the image presented is from the training data distribution or the generator generates it. This way, the generator is forced to get better at synthetic image generation.

StyleGANs can generate various styles at high-resolution [32] and also be able to control the styles in the generated images. For instance, Celeb-A dataset is a collection of high-resolution images of the faces of celebrities. StyleGAN was trained on this dataset. One can fine-tune the faces generated by the model as one wishes. Using the learned inputs to the network, one could control the face's sharpness, the eyebrows' width, and the hair's color. This way, StyleGANs were able to perform high-resolution image synthesis. However, StyleGANs still suffered from multiple issues, like water droplet artefacts and shift-invariance. Blob artefacts have been found in images generated by StyleGANs. StyleGAN2 [29] and StyleGAN2-ADA [33] propose some improvements to tackle these issues. Although StyleGAN2 has solved the issue of high-resolution image synthesis, the problem of requiring enormous-sized datasets to train GANs persists. StyleGAN2-ADA solves the issue of having large datasets and provides a way to train deep generative models on little data [33]. ADA stands for Adaptive Discriminator Augmentation. StyleGAN2 makes use of Adaptive Discriminator Augmentation instead of Stochastic Discriminator Augmentation. This way, StyleGAN2-ADA provides a way to train image synthesis models with limited data.

Some attempts have been made to construct a feature space for material appearance based on deep learning-based models' internal latent space embedding. Storrs *et al.* [24] used Variational Autoencoder (VAE) to model the distribution in images with gloss and matte surfaces . The study has shown that the image features from the internal latent space encoding of trained VAE models correlate well with human gloss perception and even mimic the mistakes that humans make in gloss judgments.

Generative Adversarial Networks (GANs) show improvements over VAEs in realistic image synthesis. Liao *et al.* [34] have generated realistic images of translucent objects with GANs and noticed that structured perceptual attributes emerge in the model's representation. They suggest that Deep Generative Models can discover an efficient and compact feature representation space for material appearance and can be potentially used to mimic the perception model of the HVS.

III. METHODOLOGY

Building upon the literature, we propose to train StyleGAN2-ADA [33] on physically-based renderings of glossy objects. We then evaluate the realism of images generated by the trained model, build tools to encode images into the latent space of the trained model and vice versa, build tools to traverse and analyse the feature space representation to check for gloss appearance attributes and analyse the usability of such feature space in aiding understanding of gloss appearance.

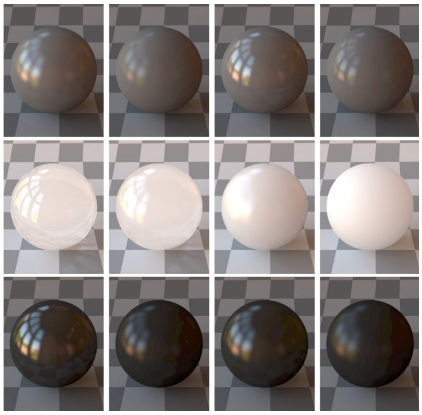

Fig. 1. Some samples from the training dataset.

A. StyleGAN2-ADA

StyleGAN2-ADA [33] is a Generative Adversarial Network designed by researchers at NVIDIA. The implementation provided by NVIDIA in the official GitHub repository is used for all the experiments (https://github.com/NVlabs/stylegan2-ada-pytorch). No specific changes have been made to the network architecture and training procedures. StyleGANs do not use the latent space directly. They first map these latent vectors into an extended latent space before generating an image. In the latent space Z, z is a 512 feature vector. Seed is the number used to generate this 512 feature vector. Then this latent vector is mapped into the extended latent space *W*. A vector w ($w \in W$) is of dimensions $1 \times 14 \times 512$. StyleGAN2-ADA applies data augmentation after the input component for both the generator and the discriminator. StyleGAN2-ADA solves the issue of collecting images to create large-scale datasets. It involves flipping the images, rotating them by a small angle, and zooming in on the image, among others.

B. Dataset

We used 132 physically-based renderings of glossy spherical objects rendered with Mitsuba [12] (can be accessed at https://github.com/davitgigilashvili/GANs4GlossEUVIP). The objects vary in surface roughness, lightness, and translucency – covering a broad range of gloss appearances. To increase the size of the dataset, we performed the augmentations by rotating the image by 90, 180 and 270 degrees, thus quadrupling the size of the dataset to 528 images. The examples of the images that were used for training are shown in Fig. 1.

C. Training

We use model weights from the pre-trained model on the (Flickr-Faces-HQ) FFHQ dataset [29] and transfer learning to train StyleGAN2-ADA to generate images with a realistic gloss appearance. We train the model for 5000kimg (i.e. how many images are evaluated; 528 $\times$ number of epochs). Training such an advanced GAN like StyleGAN2-ADA requires much computational power. We have used two NVIDIA TITAN RTX GPUs to run all our experiments. We train the model to generate images with a resolution of 256×256 pixels. The batch size used for training the model is 32, parallelised over two GPUs. A learning rate of 0.0025 is used for the transfer learning process. It took one day, 17 hours and 42 minutes to train the StyleGAN2-ADA model for 5000 kimg.

D. Image Synthesis

In StyleGAN-based architectures, a mapping network is used to map vectors from latent space Z to extended latent space W. These latent vectors w are directly plugged into the various layers of the network, thus giving us direct control to alter the styles in the images being generated. Since we do not have any understanding of the latent space of the model, to explore this latent space, we need to sample the feature space randomly. To do this, we randomly generate latent vectors from the space. Most random number generators are built on algorithms that start with a base value as an input known as a seed. For the same seed, we always get the same output random value. This helps us to lock random vectors across the experiments. We use seed values from 0 to 2000 and generate corresponding images using the trained StyleGAN2-ADA network. The first step in generating images from the seed involves generating latent vector z from the seed. Later, the latent vector z (1×512 feature space) is mapped into the extended latent space W. The resulting vector w ($w \in W$) is fed to the generator of StyleGAN2-ADA to generate images.

E. Evaluation

We evaluate the images using two methods. The first one involves using an image quality metric called Frechet Inception Distance (FID), which is a popular method to compare real and synthetic images [35]. We calculate FID after every 400 epochs, 50k images are generated from randomly sampling the latent space. FID is calculated on these 50k images by comparing them to the images in the training set.

The second method to evaluate performance was psychophysical experiment, which was hosted at the online QuickEval [36] platform. 19 observers participated in the experiment – mostly researchers and graduate students with substantial knowledge of graphics and appearance. In total, the observers were shown 60 images, 30 real images and 30 synthetic images. The real images were selected from the training set. Some of the synthetic images were those that were trying to mimic the respective real ones, while others corresponded to the random vectors from the latent space. The observers were asked to judge whether the image was real or synthetic. We explained to them that *Real* means that the images were generated using physically-based rendering with human supervision, while *Synthetic* ones were produced by GANs without human supervision. They were instructed to judge the realism of the images solely based on the realism of the gloss on the surface of the sphere.

F. Latent Space Exploration

We use the algorithm discussed above to generate W space latent vectors for all the images in the training dataset. The latent vector z ($z \in Z$) is of size 1×512, and the extended

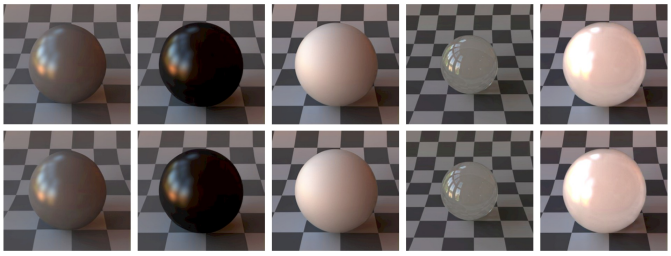

Fig. 2. The first objective is the synthesis of the realistic images. The original images are shown in the top row. They are projected into the extended latent space W. Synthetic images generated from the corresponding w latent vectors are shown in the bottom row that look highly similar to those in the top row.

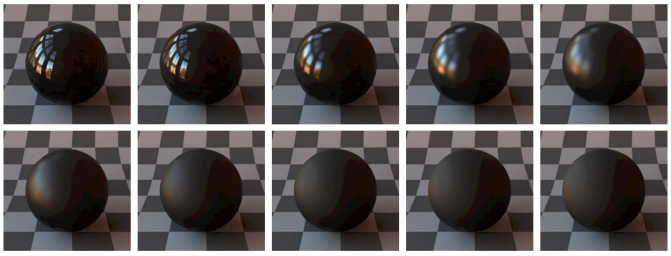

Fig. 3. Interpolations of images (performed in the latent W space) generated between the two target images shown on the left and right of each row.

latent space vectors w are of the size $1 \times 14 \times 512$. We have generated the corresponding latent vector w for every input image in the training dataset. We then use this latent vector w to generate the image. This generated image is referred to as a fake image. The original image is referred to as a real image. The examples are illustrated in Fig. 2.

We perform linear interpolation between the latent space encodings in the W space. To generate interpolations between *Image A* and *Image B*, we first find the latent space encodings of these two images in GAN's latent space. We then perform linear interpolation between the two corresponding latent codes generating a set of new latent codes. We then generate images from these interpolated latent codes. In other words, we can morph between two target images to generate interpolations between these two images. Fig. 3 demonstrates that the interpolations in the latent space W look perceptually meaningful, which indicates that the space is well-developed.

We also explored the directions in the latent space. Exploring directions in the latent space means moving along a specific dimension of the feature space and seeing how it affects the resulting images generated. In this experiment, we limit the directions to primary dimensions in the data, i.e. if the latent space has 512 dimensions, we explore along these 512 directions only. This is a simple algorithm developed from scratch by us to traverse through the latent space of the models. However, there is a significant limitation here. We are only exploring the directions along the primary dimensions. What about the direction with a slope of 45 degrees with the two primary directions? The possible directions are infinite in the data. This can be addressed in future works.

Shen *et al.* [37] propose closed form factorisation, a simple and efficient way to explore latent semantics in GANs to identify interpretable dimensions in the latent space of GANs and to extract the underlying patterns. The algorithm identifies semantically meaningful directions in the latent space by decomposition on the model weights. The output of closed-form factorisation is eigenvectors corresponding to the largest eigenvalues that maximise the objective function. The objective function is to find the directions in the latent space of GANs that reveal explanatory factors. Once we have extracted the interpretable directions in the latent space, the next step is to traverse through these directions to check how each direction impacts the style of the generated images.

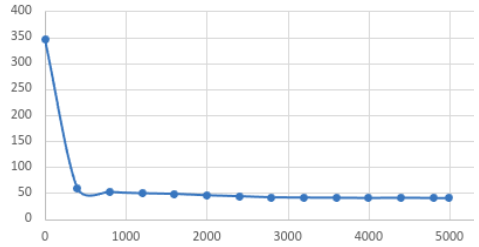

Fig. 4. FID score of images generated (vertical) vs epochs trained (horizontal).

IV. RESULTS

A. Evaluation

Fig. 4 shows how the FID score changes across epochs. As mentioned earlier, a smaller FID score implies that the images generated are closer to the images used for training and thus more realistic. This is a decent score, considering that it is evaluated on 50,000 images randomly sampled from the latent space. By increasing the number of images used for training, we can lower the FID score and thus improve the realism in the images generated. The results of the psychophysical experiments are shown in Table I. 69.02 % of the times observers judged real images as real and 30.98 % of the times observers judged real images as synthetic. When it came to synthetic images, 53.53 % of the times observers judged synthetic images as synthetic and 46.48 % of the times observers judged synthetic images as real. This implies that it was difficult for observers to assess if the images shown were real or synthetic and shows the potential of our models to generate realistic images that can trick humans.

B. Interpretable Directions

We have extracted 512 directions from the latent space and traverse through them. In total, for images generated from seeds 0 to 2000, we have generated the images by moving 5, 10, -5, -10 steps in each of the 512 directions exploited from the latent space. It is not manually possible to analyse all the images extracted, neither fits it within the scope of this paper. Hence, we show some of the significant directions extracted from closed form factorisation. From Fig. 5, we can see that by moving in the direction of the first interpretable direction, we can control the surface roughness and hence, glossiness on the sphere. This way by extracting interpretable

TABLE I
THE RESULTS OF THE PSYCHOPHYSICAL EXPERIMENT. OBSERVERS FOUND IT CHALLENGING TO DISTINGUISH REAL AND SYNTHETIC IMAGES.

	Judged Correctly	Judged Incorrectly
Real	69.02%	30.98%
Synthetic	53.52%	46.48%

directions, we can control the styles in images generated by our StyleGAN2-ADA model. We can see that, the surface roughness changes, making the spheres appear less glossy and more translucent. As the surface becomes smoother, we see that the spheres appear more glossy and less translucent. This is an interesting interaction between translucency and glossiness that automatically appears in the latent space of the model without any human supervision. From Fig. 6 we can see that when moving in the direction of the second extracted direction, we alter the style of translucency in the resulting images. The level of glossiness is more or less constant, but the level of translucency changes. This is very interesting, cause moving in the first direction altered both gloss and translucency in an inversely proportional relation, but moving in the second direction only alters translucency without altering gloss. From Fig. 7 we can see that when moving in the third interpretable direction, we alter the size of the sphere in resulting images. Specular highlights also change slightly, but the change in size is more apparent. Thus, using the extracted directions, we can alter the desired styles like glossiness, translucency or size of the sphere in resulting images. Analysing more directions would give us more control over the appearance attributes and style in synthesized images.

This is a baseline study to demonstrate that the approach can produce realistic images with very limited training set and to make first steps toward explainability. The work has limitations that will be addressed in future works. While fine tuning works for many cases, future work can explore potential changes in the architecture as well as training from scratch on a more specific dataset. Currently we have 512 dimensions that are perceptually non-uniform and exhibit cross-contaminations among perceptual attributes (e.g. size and gloss can change in the same dimension). Dimensionality reduction techniques, such as PCA, can be used to reduce dimensionality of the space from 512 to more manageable and perceptually meaningful dimensions, and psychophysical experiments will be needed to scale each dimension. Besides, we can use differentiable rendering to map the latent space back to the optical properties [38]. In addition to FID, future works can use perceptual loss-based methods for evaluating the results. Finally, although the approach is generalizable, the generated images are limited by the training dataset that the model was exposed to (e.g. single shape and environment map). Future works should include more diverse training datasets with more shapes, materials, and lighting conditions.

V. CONCLUSION

In this study, we have explored two things: 1) the potential of Deep Generative Models for generating images with realistic glossy surfaces from a limited training dataset; and 2) the usability of internal latent space of Deep Generative Models as a compact feature representation space for gloss appearance and perception. We trained StyleGAN2-ADA model to generate images of spheres with realistic glossy surfaces. We built the tools to generate the images from seeds, from z and

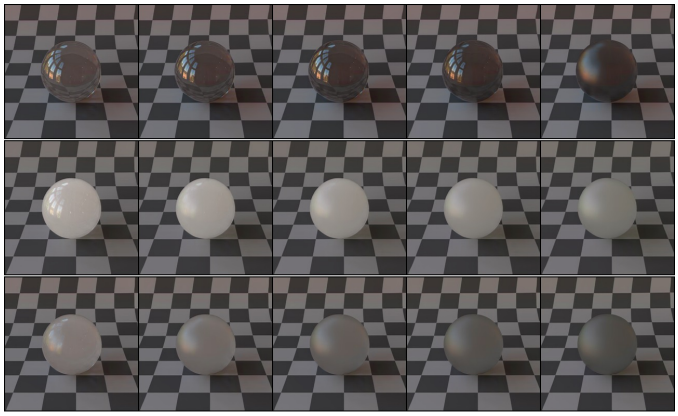

Fig. 5. Seed 6, 7, and 10 (from top to down, respectively). Moving in the direction of first interpretable direction (the direction with largest eigen value). From left to right, 10 steps in positive direction, 5 steps in positive direction, image from seed, 5 steps in negative direction, 10 steps in negative direction.

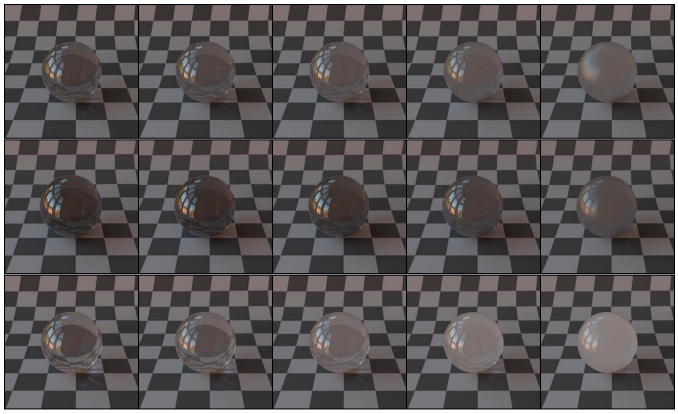

Fig. 6. Seed 1, 6, 13. Moving in the direction of second interpretable direction.

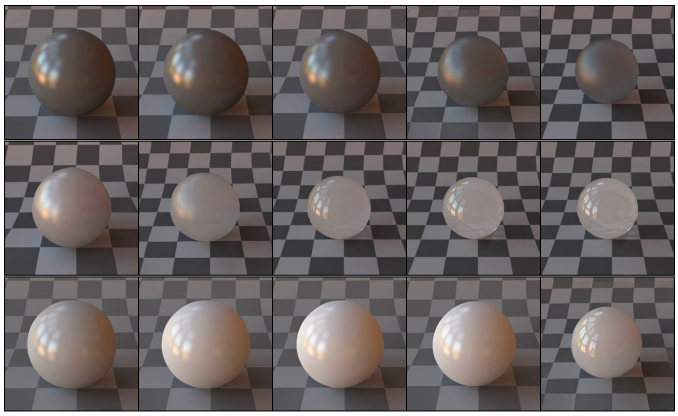

Fig. 7. Seed 1, 15, 16. Moving in the direction of third interpretable direction.

w latent vectors. We have also built the tools to map images to and from the internal latent space of StyleGAN2-ADA. We then analysed usability of this latent space as a feature space for gloss appearance and perception by extracted interpretable directions from the latent space and moving in these directions. It can be seen from our experiments and results that the images generated by StyleGAN2-ADA trick human observers into thinking that these were actually generated by human supervision in a physically based renderer. The results also show that interesting interactions between gloss and translucency emerge in the latent space of the trained model. This space can be used to find relevant features for visual perception of gloss. From linear interpolations between images, we can also see that the latent space is quite well developed. However, there are some limitations – some visual artifacts emerge due to a small dataset size. This implies that the latent space of the model contains some information gaps. Nevertheless, this shows the potential of using Deep Generative Models to generate images with realistic glossy surfaces even with a limited training set and also the potential of latent space of these models to be used as an efficient feature space for gloss appearance. It is known that in neural networks, the initial layers of the model are responsible for constructing low level features, and the final layers of the model are responsible for constructing higher level features. As a future work, the feature space can be further studied to understand which layers of the model influence what parameters of gloss in the synthesized images. Also, psychophyscial experiments need to be conducted to study how human perception correlates with the trends and patterns emerged in the latent space. Overall, using Deep Generative Models for realistic glossy image synthesis shows promising results and certainly merits future research.

REFERENCES

[1] R. W. Fleming, "Visual perception of materials and their properties," *Vision Research*, vol. 94, pp. 62–75, 2014.

[2] L. Sharan, R. Rosenholtz, and E. Adelson, "Material perception: What can you see in a brief glance?" *J. Vis.*, vol. 9, pp. 784–784, 2010.

[3] CIE, *CIE 175:2006 A framework for the measurement of visual appearance*. International Commission on Illumination., 2006.

[4] A. C. Chadwick and R. Kentridge, "The perception of gloss: A review," *Vision research*, vol. 109, pp. 221–235, 2015.

[5] F. B. Leloup, G. Obein, M. R. Pointer, and P. Hanselaer, "Toward the soft metrology of surface gloss: A review," *Color Research & Application*, vol. 39, no. 6, pp. 559–570, 2014.

[6] I. Motoyoshi, S. Nishida, L. Sharan, and E. H. Adelson, "Image statistics and the perception of surface qualities," *Nature*, vol. 447, no. 7141, pp. 206–209, 2007.

[7] F. Pellacini, J. A. Ferwerda, and D. P. Greenberg, "Toward a psychophysically-based light reflection model for image synthesis," in *Proceedings of the ACM SIGGRAPH 2000*, 2000, pp. 55–64.

[8] J.-B. Thomas, J. Y. Hardeberg, and G. Simone, "Image contrast measure as a gloss material descriptor," in *Computational Color Imaging: 6th International Workshop*. Springer, 2017, pp. 233–245.

[9] M. Lagunas, A. Serrano, D. Gutierrez, and B. Masia, "The joint role of geometry and illumination on material recognition," *Journal of Vision*, vol. 21, no. 2:, pp. 1–18, feb 2021.

[10] M. Olkkonen and D. Brainard, "Joint effects of illumination geometry and object shape in the perception of surface reflectance," *i-Perception*, vol. 2, pp. 1014–34, 12 2011.

[11] D. Gigilashvili and A. J. Islam, "The role of shape in modeling gloss," *30th Color and Imaging Conference (CIC30)*, pp. 271–276, 2022.

[12] W. Jakob, "Mitsuba Renderer," 2010, http://www.mitsuba-renderer.org.

[13] K. Storrs and R. Fleming, "Learning about the world by learning about images," *Curr. Dir. Psychol. Sci.*, vol. 30, pp. 120–128, 2021.

[14] P. J. Marlow, J. Kim, and B. L. Anderson, "The perception and misperception of specular surface reflectance," *Current Biology*, vol. 22, no. 20, pp. 1909–1913, 2012.

[15] S. C. Pont and J. J. Koenderink, "Shape, surface roughness and human perception," in *Handbook of Texture Analysis*. World Scientific, 2008, pp. 197–222.

[16] Y.-X. Ho, M. S. Landy, and L. T. Maloney, "How direction of illumination affects visually perceived surface roughness," *Journal of Vision*, vol. 6, no. 5, p. 634–648, 2006.

[17] R. Fleming, F. Jäkel, and L. Maloney, "Visual perception of thick transparent materials," *Psychol. Sci.*, vol. 22, pp. 812–20, 06 2011.

[18] T. Kawabe, K. Maruya, and S. Nishida, "Perceptual transparency from image deformation," *Proc. Natl. Acad. Sci. USA*, vol. 112, 08 2015.

[19] D. Gigilashvili, J.-B. Thomas, J. Y. Hardeberg, and M. Pedersen, "Translucency perception: A review," *Journal of Vision*, vol. 21, no. 8:4, pp. 1–41, 2021.

[20] I. Motoyoshi, "Highlight-shading relationship as a cue for the perception of translucent and transparent materials," *Journal of Vision*, vol. 10:6, pp. 1–11, 07 2010.

[21] B. Xiao, S. Zhao, I. Gkioulekas, W. Bi, and K. Bala, "Effect of geometric sharpness on translucent material perception," *Journal of Vision*, vol. 20:10, pp. 1–17, 07 2020.

[22] H. Tamura, K. E. Prokott, and R. W. Fleming, "Distinguishing mirror from glass: A "big data" approach to material perception," *Journal of Vision*, vol. 22, no. 4:4, pp. 1–22, 2022.

[23] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[24] K. Storrs, B. Anderson, and R. Fleming, "Unsupervised learning predicts human perception and misperception of gloss," *Nature Human Behaviour*, vol. 5, pp. 1–16, 10 2021.

[25] K. Prokott, H. Tamura, and R. Fleming, "Gloss perception: Searching for a deep neural network that behaves like humans," *Journal of Vision*, vol. 21:14, pp. 1–20, 11 2021.

[26] A. O'Toole and C. Castillo, "Face recognition by humans and machines: Three fundamental advances from deep learning," *Annual Review of Vision Science*, vol. 7, 08 2021.

[27] N. Kriegeskorte, "Deep neural networks: A new framework for modeling biological vision and brain information processing." *Annual Review of Vision Science*, vol. 1, pp. 417–446, 2015.

[28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE CVPR*, 2022, pp. 10 684–10 695.

[29] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE CVPR*, 2020, pp. 8110–8119.

[30] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 36 479–36 494, 2022.

[31] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[32] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[33] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," 2020. [Online]. Available: https://arxiv.org/abs/2006.06676

[34] C. Liao, M. Sawayama, and B. Xiao, "Translucency perception emerges in deep generative representations for natural image synthesis," 08 2022.

[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," 2017. [Online]. Available: https://arxiv.org/abs/1706.08500

[36] K. Van Ngo, J. J. Storvik, C. A. Dokkeberg, I. Farup, and M. Pedersen, "Quickeval: a web application for psychometric scaling experiments," in *IQSP XII*, vol. 9396. SCIA, 2015, pp. 1–13.

[37] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in GANs," 2020. [Online]. Available: https://arxiv.org/abs/2007.06600

[38] W. Chen, J. Litalien, J. Gao, Z. Wang *et al.*, "DIB-R++: learning to predict lighting and material with a hybrid differentiable renderer," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 22 834–22 848, 2021.