LIGHTING (IN)CONSISTENCY OF PAINT BY TEXT

Hany Farid
Department of Electrical Engineering and Computer Sciences
School of Information
University of California, Berkeley
hfarid@berkeley.edu

ABSTRACT

Whereas generative adversarial networks are capable of synthesizing highly realistic images of faces, cats, landscapes, or almost any other single category, paint-by-text synthesis engines can – from a single text prompt – synthesize realistic images of seemingly endless categories with arbitrary configurations and combinations. This powerful technology poses new challenges to the photo-forensic community. Motivated by the fact that paint by text is not based on explicit geometric or physical models, and the human visual system's general insensitivity to lighting inconsistencies, we provide an initial exploration of the lighting consistency of DALL·E-2 synthesized images to determine if physics-based forensic analyses will prove fruitful in detecting this new breed of synthetic media.

Keywords Photo Forensics · DALL·E-2 · Text-to-Image

1 Introduction

As OpenAI moves to expand access to their impressive DALL·E-2 paint-by-text synthesis engine[1], concerns have been raised as to how this new technology might be misused [1] (see also Google's Imagen[2] and Parti[3]). Paint by text synthesizes high-resolution images from a single text prompt [2, 3, 4, 5], affording control of semantic content seemingly limited only by our imagination. This latest introduction into the synthetic media arena creates new challenges for the photo-forensic community.

Paint by text, and other learning-based synthesis engines, are not based on explicit modeling of a 3-D scene, lighting, or camera. It is perhaps not surprising, therefore, that some aspects of scene geometry, including perspective geometry, are not always preserved in these synthesized images [6].

Although these synthesized images appear highly natural, the human visual system is generally insensitive to lighting inconsistencies [7] and – as with perspective geometry – may not necessarily notice implausible, unnatural, or inconsistent lighting. We provide an initial exploration of the

---

[1] https://openai.com/dall-e-2
[2] https://imagen.research.google
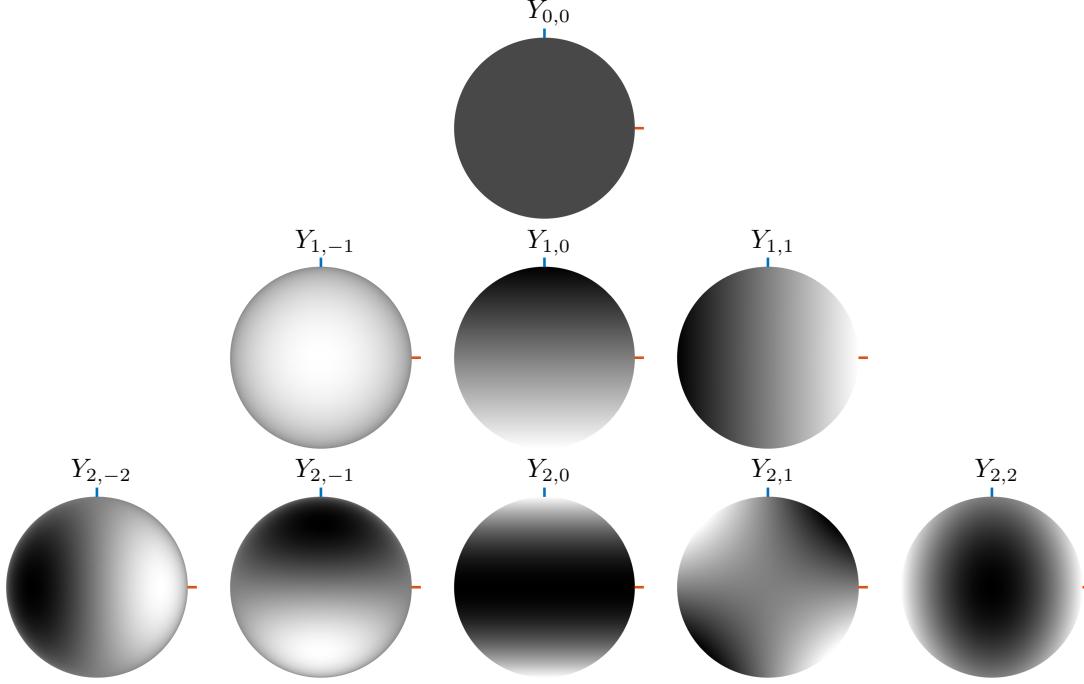[3] https://parti.research.google

Figure 1: The first three orders of spherical harmonics as functions on the sphere. Shown from top to bottom are the zeroth-order spherical harmonic, $Y_{0,0}(\cdot)$; the three first-order spherical harmonics, $Y_{1,*}(\cdot)$; and the five second-order spherical harmonics, $Y_{2,*}(\cdot)$. The red and blue axis corresponds to the positive x- and z-axis (corresponding to the camera optical axis), and the positive y-axis is facing into the page.

lighting consistency of DALL·E-2 synthesized images to determine if lighting (modeled using spherical-harmonic based lighting environments [8, 9]) are globally consistent with natural images, and if lighting is locally consistent within a synthesized image.

This analysis reveals that while global illumination is – with a few exceptions – relatively consistent with natural images, lighting within an image can be highly variable. These observations may prove forensically useful in distinguishing synthesized from photographed images.

2 Photo Forensics: Lighting Analysis

Under the assumption of a single distant light source (e.g., the sun), the appearance I_k of a convex Lambertian surface of constant reflection can be modeled as $I_k = \vec{L}^T \cdot \vec{N}_k + A$, where $\vec{L}$ is the 3-D vector denoting the orientation to the light source, $\vec{N}_k$ is the 3-D surface normal corresponding to pixel k, and A is a scalar term embodying the ambient light[4]. Despite it simplicity, this lighting model has been used to forensically analyze images for lighting consistency [10, 11].

This lighting model can be extended to accurately capture more complex lighting environments, with a number of arbitrarily placed light sources. As originally described in [8, 9], and forensically employed in [12, 13, 14, 15], under the assumption of distant lighting, the appearance of a convex Lambertian surface of constant reflectance can be modelled with a spherical Fourier basis. Specifically, a point on a surface with 3-D surface normal $\vec{N}_k$, is imaged to pixel location k with pixel

---

[4]This lighting model assumes the angle between the light $\vec{L}$ and surface normal $\vec{N}_k$ is between $0°$ and $90°$.

intensity given by:

$$
\begin{aligned}
I_k \;=\; & l_{0,0}\pi Y_{0,0}(\vec{N}_k) + \\
& l_{1,-1}\tfrac{2\pi}{3}Y_{1,-1}(\vec{N}_k) + l_{1,0}\tfrac{2\pi}{3}Y_{1,0}(\vec{N}_k) + l_{1,1}\tfrac{2\pi}{3}Y_{1,1}(\vec{N}_k) + \\
& l_{2,-2}\tfrac{\pi}{4}Y_{2,-2}(\vec{N}_k) + l_{2,-1}\tfrac{\pi}{4}Y_{2,-1}(\vec{N}_k) + l_{2,0}\tfrac{\pi}{4}Y_{2,0}(\vec{N}_k) + l_{2,1}\tfrac{\pi}{4}Y_{2,1}(\vec{N}_k) + l_{2,2}\tfrac{\pi}{4}Y_{2,2}(\vec{N}_k), (1)
\end{aligned}
$$

where, the spherical harmonics $Y_{*,*}(\cdot)$ – embodying the lighting environment – are parameterized as a function of the 3-D surface normal $\vec{N}_k = \begin{pmatrix} x_k & y_k & z_k \end{pmatrix}$:

$$
\begin{aligned}
& Y_{0,0}(\vec{N}_k) = \tfrac{1}{\sqrt{4\pi}} && Y_{1,-1}(\vec{N}_k) = \sqrt{\tfrac{3}{4\pi}}y_k && Y_{1,0}(\vec{N}_k) = \sqrt{\tfrac{3}{4\pi}}z_k \\
& Y_{1,1}(\vec{N}_k) = \sqrt{\tfrac{3}{4\pi}}x_k && Y_{2,-2}(\vec{N}_k) = 3\sqrt{\tfrac{5}{12\pi}}x_k y_k && Y_{2,-1}(\vec{N}_k) = 3\sqrt{\tfrac{5}{12\pi}}y_k z_k && (2) \\
& Y_{2,0}(\vec{N}_k) = \tfrac{1}{2}\sqrt{\tfrac{5}{4\pi}}(3z_k^2 - 1) && Y_{2,1}(\vec{N}_k) = 3\sqrt{\tfrac{5}{12\pi}}x_k z_k && Y_{2,2}(\vec{N}_k) = \tfrac{3}{2}\sqrt{\tfrac{5}{12\pi}}(x_k^2 - y_k^2).
\end{aligned}
$$

As shown in Figure 1, the zeroth-order spherical harmonic, $Y_{0,0}(\cdot)$, embodies the ambient lighting, the first-order spherical harmonics, $Y_{1,*}(\cdot)$, embody directional lighting (top/down, front/back, and left/right), and the second order harmonics, $Y_{2,*}(\cdot)$, embody higher-order, directional lighting effects.

Note that the expression of pixel intensity, Equation (1), is linear in the nine lighting environment coefficients, $l_{0,0}$ to $l_{2,2}$. Given 3-D surface normals at $p \geq 9$ points, the lighting environment coefficients can be estimated as the least-squares solution to the following system of linear equations:

$$
\begin{pmatrix}
\pi Y_{0,0}(\vec{N}_1) & \tfrac{2\pi}{3}Y_{1,-1}(\vec{N}_1) & \cdots & \tfrac{\pi}{4}Y_{2,2}(\vec{N}_1) \\
\pi Y_{0,0}(\vec{N}_2) & \tfrac{2\pi}{3}Y_{1,-1}(\vec{N}_2) & \cdots & \tfrac{\pi}{4}Y_{2,2}(\vec{N}_2) \\
\vdots & \vdots & \ddots & \vdots \\
\pi Y_{0,0}(\vec{N}_p) & \tfrac{2\pi}{3}Y_{1,-1}(\vec{N}_p) & \cdots & \tfrac{\pi}{4}Y_{2,2}(\vec{N}_p)
\end{pmatrix}
\begin{pmatrix}
l_{0,0} \\ l_{1,-1} \\ \vdots \\ l_{2,2}
\end{pmatrix}
=
\begin{pmatrix}
I_1 \\ I_2 \\ \vdots \\ I_p
\end{pmatrix}
$$

$$
A\vec{l} \;=\; \vec{b}, \tag{3}
$$

where A is the $p \times 9$ matrix containing the sampled spherical harmonics, $\vec{l}$ is the 9×1 vector of unknown lighting environment coefficients, and $\vec{b}$ is the $p \times 1$ vector of intensities at p pixel locations. The least-squares solution to this system is $\vec{l} = \left(A^T A\right)^{-1} A^T \vec{b}$. This estimation can be performed for each of three color channels in an RGB image.

While the accurate estimation of 3-D surface normals may not always be possible from a single image, basic geometric shapes like a sphere afford a simple way to extract 3-D surface normals. A sphere photographed from any orientation will be imaged as an ellipse; in images recorded with a narrow field of view, these ellipses will be well approximated by a circle, particularly near the center of the image. We will assume that the image of a sphere is circular, and therefore the 3-D surface normal at any point on the sphere can be determined directly from the pixel coordinate (x_k, y_k) as:

$$
\vec{N}_k \;=\; \begin{pmatrix} x_k - c_x & y_k - c_y & \sqrt{r^2 - (x_k - c_x)^2 + (y_k - c_y)^2} \end{pmatrix}, \tag{4}
$$

where r and (c_x, c_y) are the radius and center of the image-based circular projection of a sphere, and $\vec{N}_k$ is scaled to unit length: $\vec{N}_k / \|\vec{N}_k\|$.

Shown in Figure 2(a) is an image of a concrete garden sphere overlaid with a fitted circle (see below) and a subset of the estimated 3-D surface normals. Shown in panel (b) of this figure is a sphere rendered with the estimated lighting environment, where we can see that the overall color and positioning of the lighting has been captured.
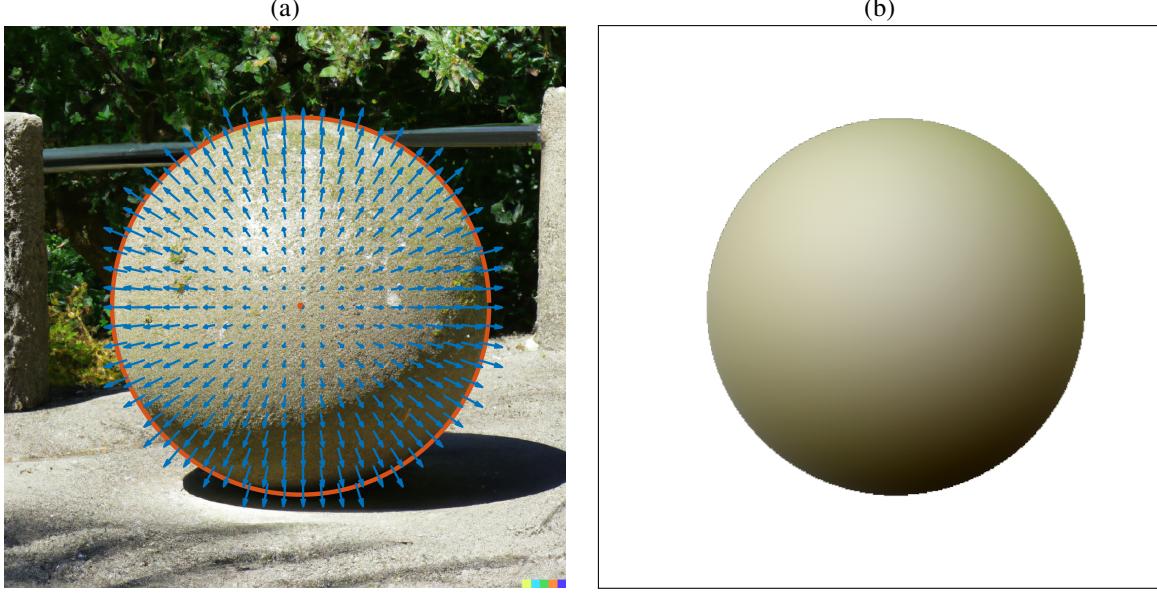
(a)   (b)

Figure 2: Estimating 3-D lighting environment from a sphere: (a) a DALL·E-2 synthesized image of a concrete garden sphere overlaid with a fitted circle (red) and estimated 3-D surface normals (blue); and (b) a rendered sphere illuminated with the estimated lighting environment.

An EM-based approach is used to fit a circle (center and radius) to the image of a sphere, Figure 2(a). An RGB image is first converted to grayscale and histogram equalized (to boost edge contrast) followed by a standard gradient-based edge detection and an intensity threshold to yield a binary-valued image. The resulting binary image highlights the salient edges, including – but not exclusively – the circle's boundary. An expectation/maximization (EM) algorithm [16] is employed to iteratively segment the edge pixels into those belonging to the circle's boundary, and to estimate the circle center and radius.

In the E-step, the residual error between each edge pixel k is computed as the shortest distance between the pixel at location (x_k, y_k) and the current estimate of the circle with center (c_x, c_y) and radius (r):

$$\delta_k \;\; = \;\; |(x_k - c_x)^2 + (y_k - c_y)^2 - r^2|. \tag{5}$$

The probability that any pixel k, with residual error δ_k, is associated with the circle's boundary is computed as:

$$w_k \;\; = \;\; \frac{e^{-\delta_k^2/2\sigma^2}}{e^{-\delta_k^2/2\sigma^2} + \epsilon}, \tag{6}$$

where σ is the variance of the underlying normal distribution, and ϵ is the uniform probability that pixel k is not associated with the circle's boundary (i.e., pixel k belongs to an outlier model).

In the M-step, a weighted total least-squares estimation [17] is used to re-estimate the circle's center (c_x, c_y) and radius (r) as the minimal eigenvalue-eigenvector $(\vec{v})$ of $M^T W^T W M$, where:

$$M = \begin{pmatrix} x_1^2 + y_1^2 & x_1 & y_1 & 1 \\ x_2^2 + y_2^2 & x_2 & y_2 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_n^2 + y_n^2 & x_n & y_n & 1 \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n, \end{pmatrix} \tag{7}$$

Figure 3: Representative examples of garden spheres used to evaluate the consistency of global lighting environments between synthesized images (top row) and photographed images (bottom row).

where n is the total number of identified edge pixels, and w_k is the probability that pixel k is associated with the circle's boundary, as computed in the E-step, Equation (6). The circle's center and radius are extracted from the solution $\vec{v}$ as follows:

$$c_x \;=\; -\,\frac{v_2}{2v_1} \qquad c_y \;=\; -\,\frac{v_3}{2v_1} \qquad r \;=\; \sqrt{\frac{v_2^2 + v_3^2}{4v_1^2} - \frac{v_4}{v_1}}, \tag{8}$$

where v_i denotes the ith component of the 4-D eigenvector $\vec{v}$.

The E- and M-steps are iteratively evaluated until subsequent estimates of the circle center and radius are each less than a threshold of 0.5 pixels. On each iteration, the variance σ in Equation (6) is updated to $\sum_{k=1}^{n} w_k r_k^2 / \sum_{k=1}^{n} w_k$. The EM estimation is bootstrapped by manually annotating the imaged sphere's approximate center and radius.

3   Paint by Text: Lighting Analysis

A total of 50 images were synthesized using DALL·E-2 with the text prompt "a photo of a concrete sphere in a garden." In most cases, the resulting image was semantically consistent with the prompt. The synthesized images were manually curated to remove any images that did not show the sphere in its entirety (due to clipping or occlusion) or was not relatively uniform in appearance. A second set of 50 images were downloaded following a web-based image search with the same text prompt "a concrete sphere in a garden." The same curation was applied to the surfaced images, along with a resolution constraint of at least 1024×1024 pixels – the same resolution as the synthesized images. The context of these images – primarily garden-supply stores – made it highly likely the images were photographic and not computer-generated or synthesized.

All 100 images were manually cropped loosely around the sphere and uniformly resized to 600×600 pixels. Shown in Figure 3 are four representative examples of these synthesized (top row) and photographed (bottom row) images.
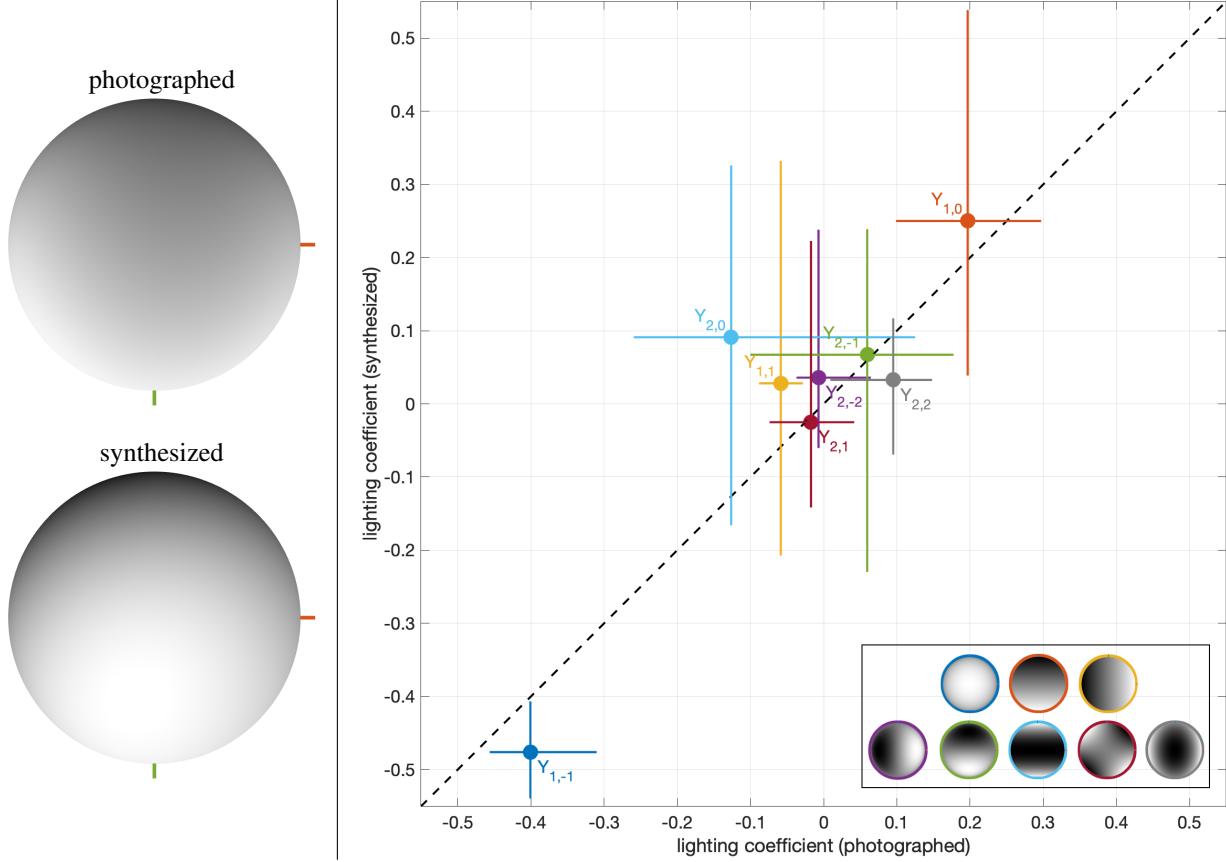
Figure 4: Shown on the right are the environmental lighting coefficients for photographed (horizontal axis) and synthesized (vertical axis) images, with a correlation of $R^2 = 0.80$. Each data point corresponds to the median lighting coefficient (50% quantile) and the horizontal and vertical bars correspond to the 35% and 65% quantile. The first-order, $Y_{1,*}$, and second-order, $Y_{2,*}$, coefficients are normalized by the zeroth-order coefficient $Y_{0,0}$. Shown on the left is a sphere rendered with the median photographic (top) and synthesized (bottom) lighting environment (these spheres are displayed on a shared intensity range and the red and green axes correspond to the x- and y-axis and the z-axis (camera optical axis) is facing into the page – note these spheres are rendered with a different viewpoint than the legend to highlight the differences between them).

After fitting a circle to the image of the sphere, the 3-D surface normals are estimated, Figure 2; given the relatively high image resolution, the normals are sampled at every other pixel. The image is then spatially filtered with a 9-tap median filter to minimize the impact of non-uniformities in the sphere's reflectance. The lighting environment, Equation (3), is then estimated for each image, yielding a triple of 9-D lighting environment coefficients, one per color channel. Because there was no substantive difference in the lighting environments across color channels, here we will report the lighting environments from a grayscale version of the input images.

Shown in Figure 4 is a scatter plot of median (50% quantile) lighting environment estimates from the 50 photographed (horizontal axis) and 50 synthesized (vertical axis) images. To remove the impact of overall scene brightness, the lighting coefficients are divided by the magnitude of the zeroth-order term $Y_{0,0}(\cdot)$, leaving three first-order coefficients $Y_{1,*}(\cdot)$ and five second-order coefficients $Y_{2,*}(\cdot)$. The horizontal and vertical bars correspond to the 35% and 65% quantile. For the most part, the synthesized and photographed lighting environments are well correlated, with a $R^2 = 0.80$ and with most of the coefficient medians lying near the dashed, zero-intercept, unit-slope line.

†

Figure 5: Representative examples of synthesized images used to evaluate the consistency of lighting within a scene.

Note that the coefficients on the $Y_{1,-1}$ spherical-harmonic term, corresponding to top/down illumination, is generally negative, corresponding to the typical natural illumination from above. The coefficients on the $Y_{1,0}$ term, corresponding to front/back illumination, is generally positive, corresponding to illumination behind the camera (an illumination in front of the camera places the subject in deep shadow or silhouette, and so photographers typically orient their subject so that they are illuminated with a light source behind the camera).

The most notable deviation between synthesized and photographed lighting is the second-order term $Y_{2,0}(\cdot)$ which is significantly larger for synthesized images (0.09) than for photographed images (-0.13). As shown in Figure 1 (center, bottom row), this second-order spherical harmonic corresponds to illumination from in front of and behind the camera. This somewhat unnatural illumination pattern is prevalent in synthesized, but not in photographed images (the negative coefficient on this spherical harmonic corresponds to an intensity-inverted version of this pattern). Shown in Figure 4 is a sphere rendered with the median photographic (top-left panel) and synthesized (bottom-left panel) lighting environment, revealing this front/back illumination difference (note these spheres are rendered with a different viewpoint than those in Figure 1 to highlight the differences between them).

For all lighting coefficients, there is significantly more variation in the environmental lighting for the synthesized images than photographed images, as seen by the difference in the vertical (simulated) and horizontal (photographed) bars. This larger variation motivates the next analysis in which we examine if the lighting within an image is consistent.

A total of 10 images were synthesized using DALL·E-2 with the text prompt "a photo of a garden with gray matte spheres scattered on the ground." Shown in Figure 5 are four representative examples of these synthesized images. From each image, the lighting environment was estimated from five unoccluded spheres. By comparing the lighting coefficients of all pairs of spheres, each image contributes 10 lighting comparisons ($_5C_2 = (5 \times 4)/2 = 10$). Shown in Figure 6 are three scatter plots corresponding to the correlation between the zeroth-, first-, and second-order spherical harmonics for these pair-wise comparisons within an image. Unlike the previous analysis, these lighting coefficients are not normalized because the scale of the coefficients should be the same within an image (assuming, per our model, distant light sources). With an R^2 of 0.54, 0.69, and 0.65 for the zeroth-, first- and second-order harmonics, the within-image lighting consistency is weaker than the global across-image – synthesized to photographed – consistency ($R^2 = 0.80$), Figure 4.

The lower correlation for the zeroth-order coefficients may be due to differences in the underlying reflectance of the synthesized spheres. Although the first-order terms $Y_{1,-1}$ and $Y_{1,1}$ are fairly well
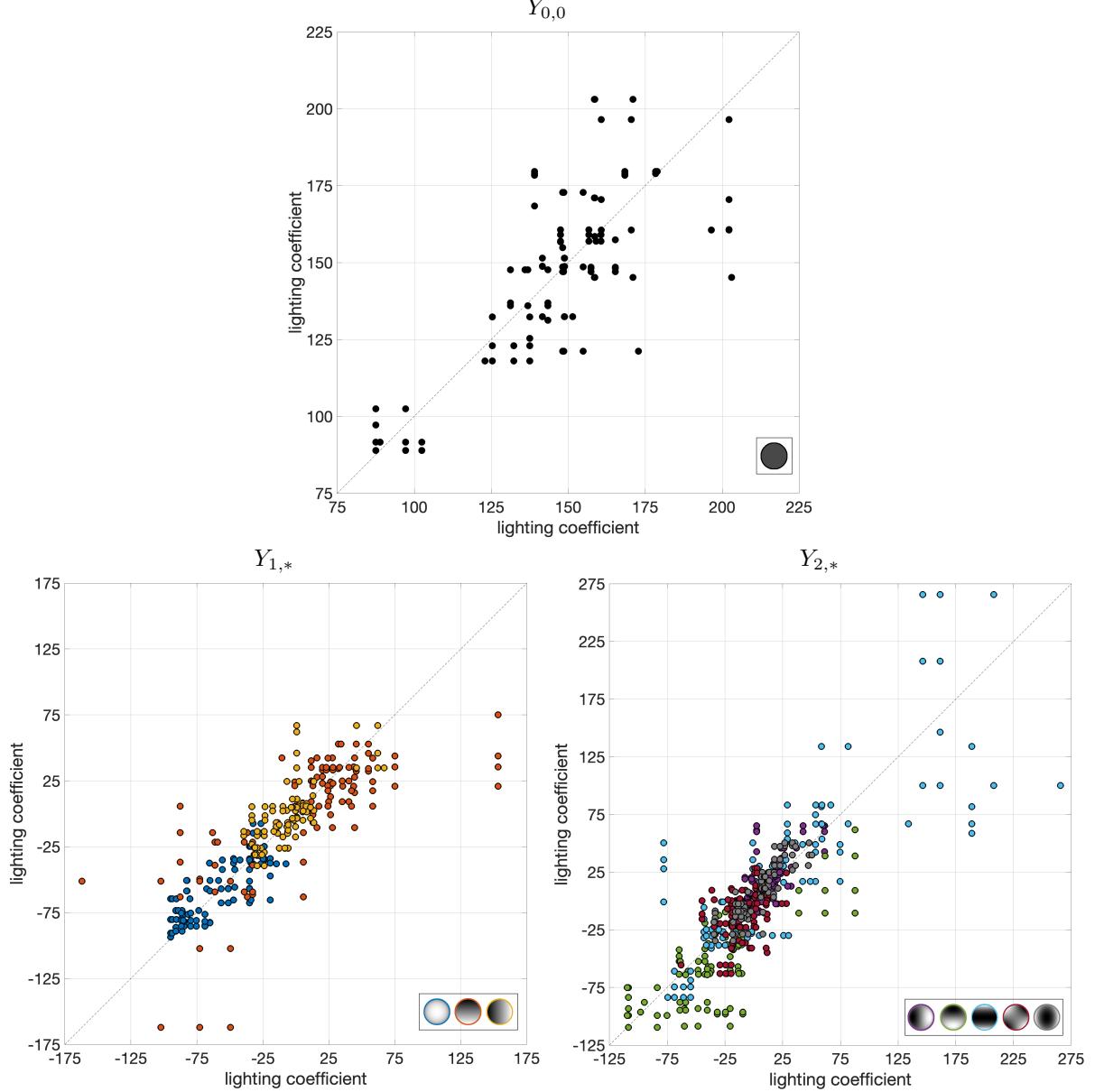
Figure 6: The environmental lighting coefficients for all pairs of spheres in a single synthesized image. The correlation (R^2) of the zeroth- (top), first- (bottom left), and second-order (bottom right) spherical-harmonic coefficients is 0.54, 0.69, and 0.65.

correlated, the $Y_{1,0}$ term, corresponding to lighting from front/back, is significantly less correlated than the other first-order terms. Similarly, the second-order term $Y_{2,0}$ is significantly less correlated than the other second-order terms. These deviations are consistent with the previous analysis that found larger deviations in the lighting terms corresponding to these front/back lighting terms.

4 Discussion

Even in the absence of explicit 3-D modeling of a camera, scene geometry, and lighting – as found in traditional CGI-rendering – paint-by-text synthesis is capable of creating remarkably realistic

images. We previously observed that specific perspective geometry is not always respected in these synthesized images [6]. Here we observe that globally, lighting is relatively consistent with natural photographed images, with a slight front/back (relative to the camera) lighting bias in synthesized images. We also observe that lighting within an image is somewhat consistent, but with some significant deviations.

Because the images used in this study are fairly generic, there is no reason to believe these observations are specific to the relatively small number of analyzed images. It remains to be seen, however, if these observations will generalize to arbitrary indoor and outdoor scenes. This will require pairing photographed images with synthesized images from a broad range of different settings containing objects of known 3-D geometry.

Estimating 3-D lighting environments requires, of course, estimation of 3-D surface normals. While this is trivial for simple geometric shapes, it is not always easy or even possible to accurately estimate 3-D geometry from arbitrary objects in a single image. Three-dimensional modeling of faces [18] and bodies [19], however, has become increasingly more reliable and accurate, allowing for the recovery of 3-D lighting environments from people in images. When 3-D models cannot be built, a 2-D lighting environment [12] (or direction [20, 10]) can be estimated by observing that five of the nine spherical harmonics – $Y_{0,0}(\cdot)$, $Y_{1,-1}(\cdot)$, $Y_{1,1}(\cdot)$, $Y_{2,-2}(\cdot)$, and $Y_{2,2}(\cdot)$, Equation (2) – do not depend on the z-component of the required 3-D surface normal. As such, a 2-D, reduced dimensional, lighting environment can be estimated from only 2-D surface normals easily extracted from the boundary of arbitrary objects (where the z-component of the surface normal is 0).

The trend in paint by text has been that increasing model size yields increasingly more realistic images. Given this trend, it remains to be seen if larger models will more accurately capture the naturalness and consistency in lighting environments. Until that time, however, physics-based forensic analyses should prove useful in analyzing this new breed of synthetic media.

Acknowledgement

References

[1] Bobby Allyn. Surreal or too real? Breathtaking AI tool DALL-E takes its images to a bigger stage. https://www.npr.org/2022/07/20/1112331013/dall-e-ai-art-beta-test, 2022.

[2] 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.

[3] 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, 2021.

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

[5] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han,

Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. arXiv:2206.10789, 2022.

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

[7] Yuri Ostrovsky, Patrick Cavanagh, and Pawan Sinha. Perceiving illumination inconsistencies in scenes. *Perception*, 34:1301–1314, 2005.

[8] 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 A*, 18:2448–2559, 2001.

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

[10] Micah K. Johnson and Hany Farid. Exposing digital forgeries by detecting inconsistencies in lighting. In *7th Workshop on Multimedia and Security*, pages 1–10, 2005.

[11] Tiago Carvalho, Hany Farid, and Eric Kee. Exposing photo manipulation from user-guided 3-D lighting analysis. In *SPIE Symposium on Electronic Imaging*, 2015.

[12] Micah K. Johnson and Hany Farid. Exposing digital forgeries in complex lighting environments. *IEEE Transactions on Information Forensics and Security*, 2(3):450–461, 2007.

[13] Christian Riess and Elli Angelopoulou. Scene illumination as an indicator of image manipulation. In *International Workshop on Information Hiding*, pages 66–80, 2010.

[14] 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, 2010.

[15] Tiago José De Carvalho, Christian Riess, Elli Angelopoulou, Helio Pedrini, and Anderson de Rezende Rocha. Exposing digital image forgeries by illumination color classification. *IEEE Transactions on Information Forensics and Security*, 8(7):1182–1194, 2013.

[16] Arthur P. Dempster, Nan M.. Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[17] Walter Gander, Gene H. Golub, and Rolf Strebel. Least-squares fitting of circles and ellipses. *BIT Numerical Mathematics*, 34(4):558–578, 1994.

[18] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. In *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, volume 40, 2021.

[19] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578, 2016.

[20] Peter Nillius and Jan-Olof Eklundh. Automatic estimation of the projected light source direction. In *Computer Vision and Pattern Recognition*, volume 1, 2001.