Supplementary material: Adversarial robustness in discontinuous spaces via alternating sampling & descent

Rahul Venkatesh
Stanford University
Stanford, CA, USA
rmvenkat@stanford.edu

Eric Wong*
University of Pennsylvania
Philadelphia, PA, USA
exwong@cis.upenn.edu

Zico Kolter*
Carnegie Mellon University and
Bosch Center for AI
Pittsburgh, PA, USA
zkolter@cs.cmu.edu

In this document, we present the following supporting information for the main paper.

- Visualization of different image variations generated by linearly sampling 3D scene parameters using our dataset creation pipeline described in Sec. 6.1 of the main paper (Figs. 2, 6, 7, 8, 9, 10), and the different semantic classes present in the dataset (Fig. 1).
- Range of values of different 3D scene parameters used in dataset creation and adversarial attack generation (Table. 1a). For more details about precisely how these parameters are used in rendering system architecture we refer the reader to the pytorch3D documentation [4] and the Phong lighting model [3] for additional technical details.
- Comparison with alternative black box attack methods on the 3D scene parameter setting – we report *RTA* of both SPSA and Boundary Attacks (BA) on the 3D scene parameter setting in Table 1b below. We find that these are not competitive attacks on this setting.
- A visualization of adversarial images generated by (1) altering 3D scene parameters (Fig. 3a and Fig. 3b) and (2) inpainting an adversarial patch (Fig. 4)

References

[1] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017. 2

[2] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pages 1802–1811, 2019. 2

[3] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975. 1

[4] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 1

[5] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, pages 5025–5034. PMLR, 2018. 2

* = Equal advising
Uesato et.al., ICML 2018
Brendel et.al., ICLR 2018

No.	p_i	min_{p_i}	max_{p_i}	unit
1	Camera Azimuth	-60	60	*degrees*
2	Camera Elevation	-30	30	*degrees*
3	Camera Tilt	-40	40	*degrees*
4	Camera Distance	0.2	0.3	*meters*
5	Light Azimuth	-25	25	*degrees*
6	Light Elevation	-25	25	*degrees*
7	Light Distance	0.5	0.8	*degrees*
8	Light Ambient Hue	60	180	*color*
9	Light Specular Hue	60	180	*color*
10	Light Diffuse Hue	60	180	*color*
11	Light Ambient Saturation	0	1	*ratio*
12	Light Specular Saturation	0	1	*ratio*
13	Light Diffuse Saturation	0	1	*ratio*
14	Light Ambient Value	0.5	1	*ratio*
15	Light Specular Value	0	1	*ratio*
16	Light Diffuse Value	1	1	*ratio*
17	Material Shininess	0.2	1	*ratio*

(a) Parameter ranges

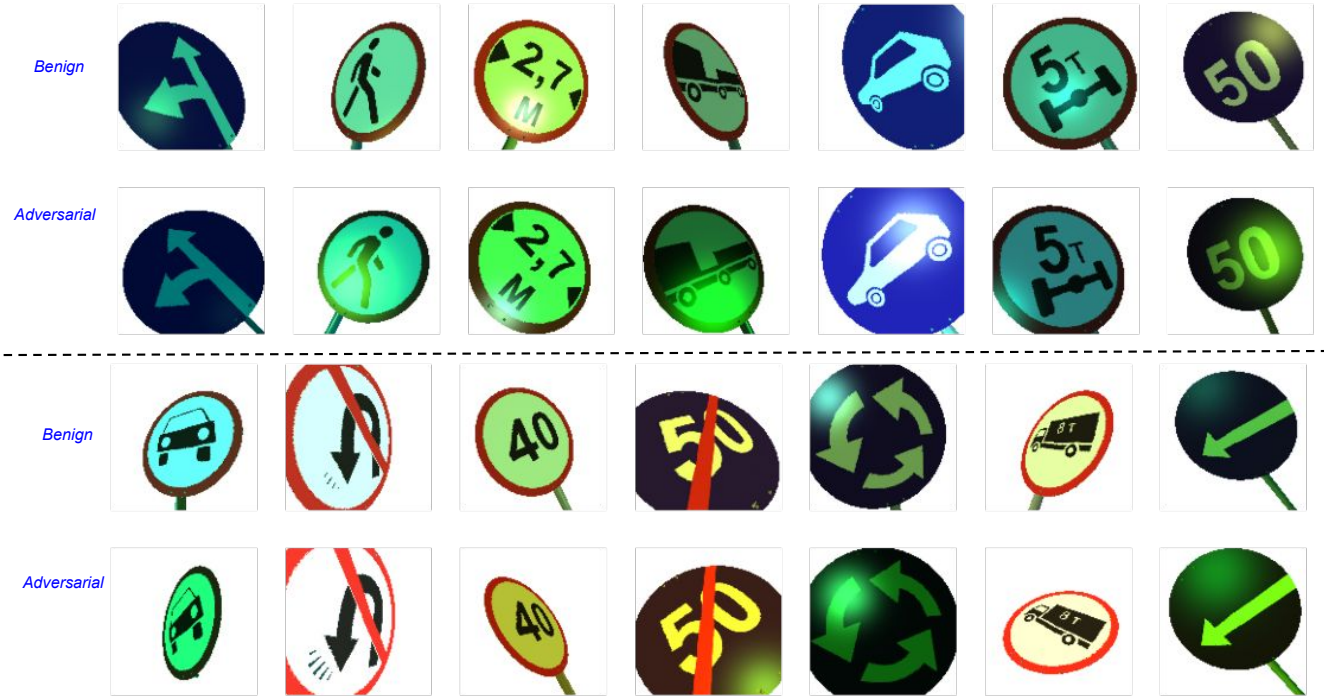
Attack	*RTA* ↓
RS [2]	0.60
BA [1]	0.59
RS+PGD	0.51
SPSA [5]	0.56
CMA	0.53
ASD	**0.46**

(b) *RTA* on an undefended model

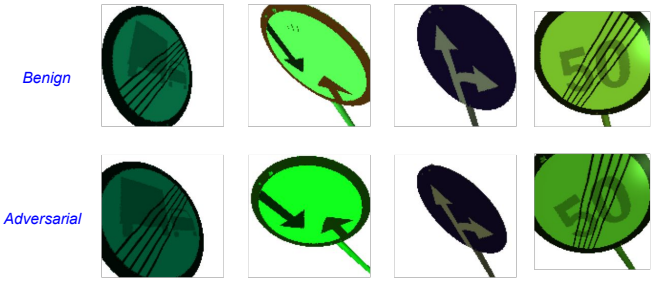
Table 1. ***Left:*** Ranges of different parameters used for generating the dataset. Note that for $|p| = 7$, the light distance signifies the intensity of the light source (directional light in this case.). However, for $|p| = 17$, we use a constant value (0.5) for this parameter since the light intensity parameters i.e. color values (brightness) in HSV space already account for variations in intensity. ***Right:*** Testing various attacks on an undefended model (Table. 1 in paper) – BA (Boundary attack) and SPSA underperform.

Figure 1. Different traffic sign classes in the dataset.

Figure 2. Varying Hue (color gamut) and brightness of a directional light source.



(a) Adversarially attacking an undefended model (using *ASD*).



(b) Adversarially attacking a defended model (using *ASD*).

Figure 4. Adversarial patch images generated by attacking an undefended CIFAR-10 model using *ASD*.

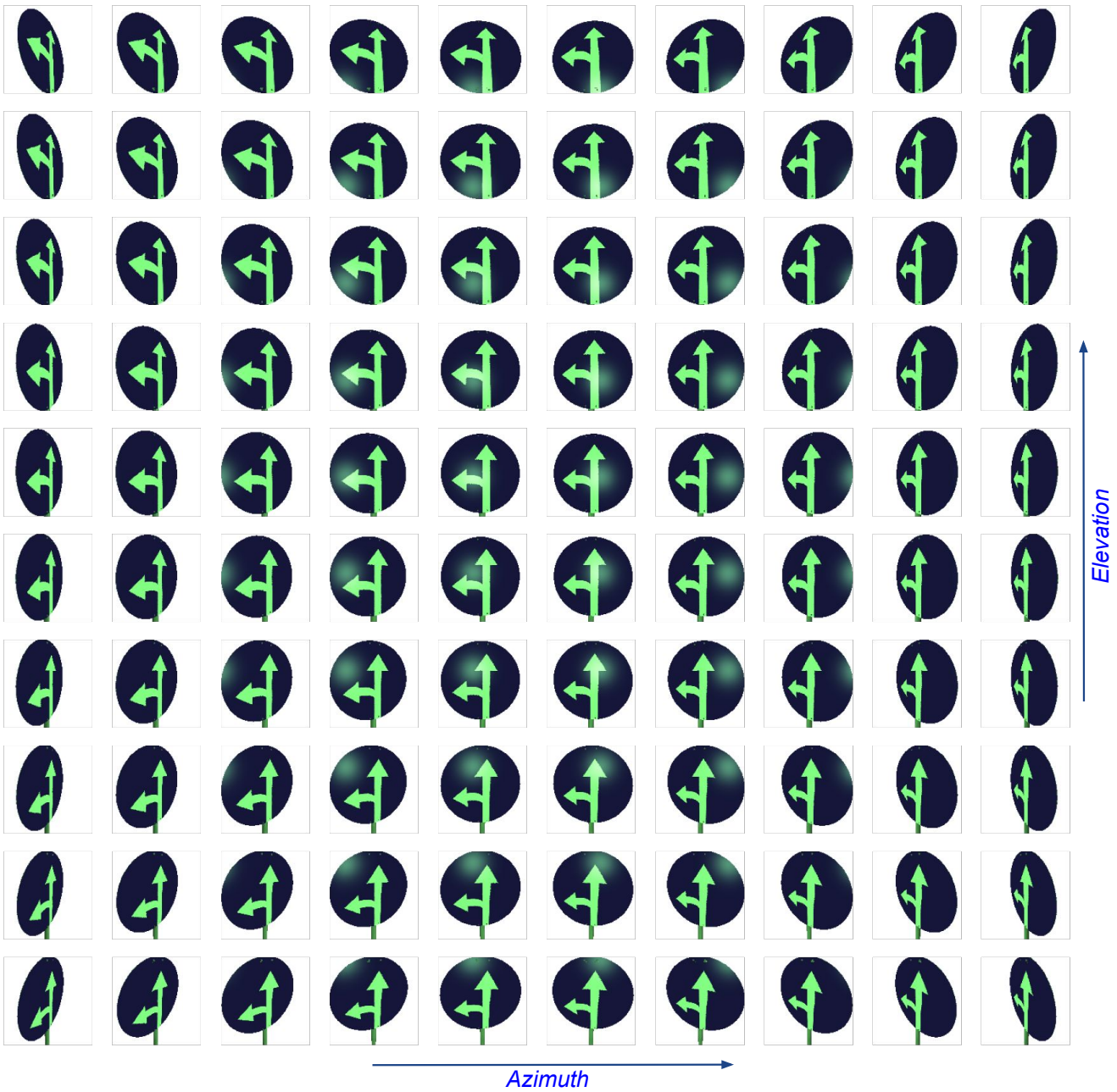


Figure 5. *Visualizing physical effects of 3D scene parameter perturbations:* **camera azimuth vs elevation**.

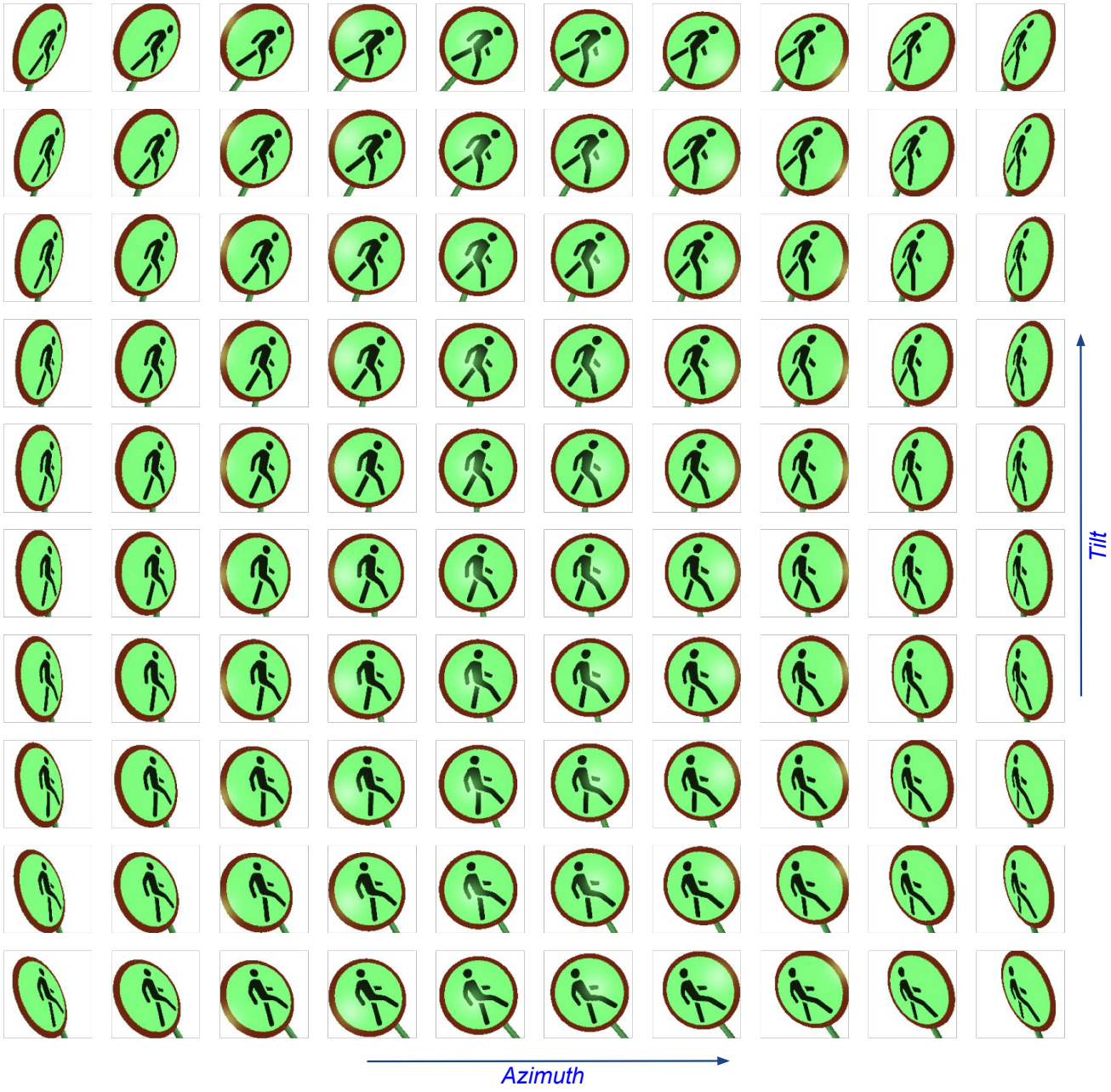


Figure 6. *Visualizing physical effects of 3D scene parameter perturbations:* **camera azimuth vs tilt**.

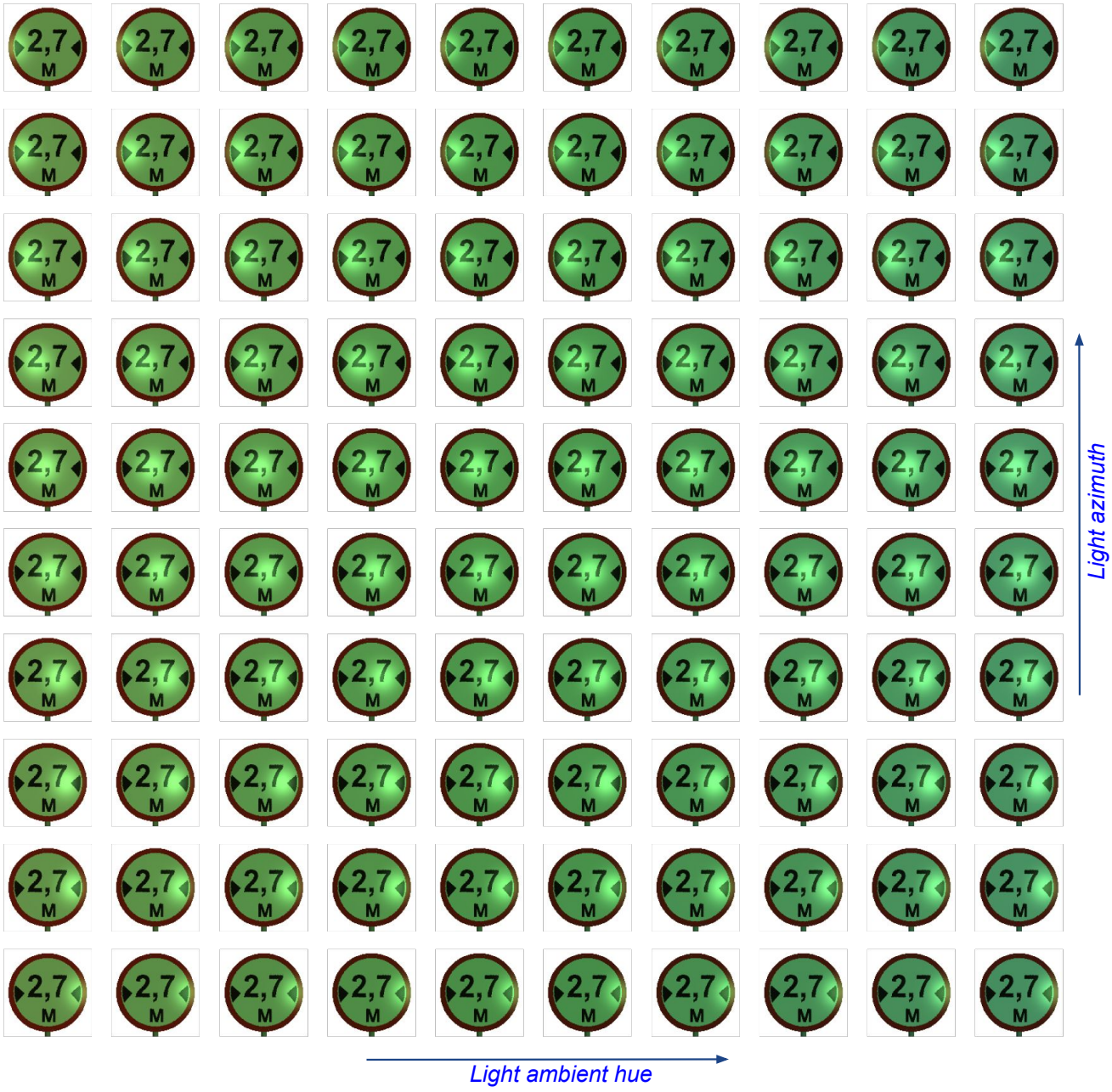


Figure 7. *Visualizing physical effects of 3D scene parameter perturbations:* **light ambient hue (color) vs azimuth**.

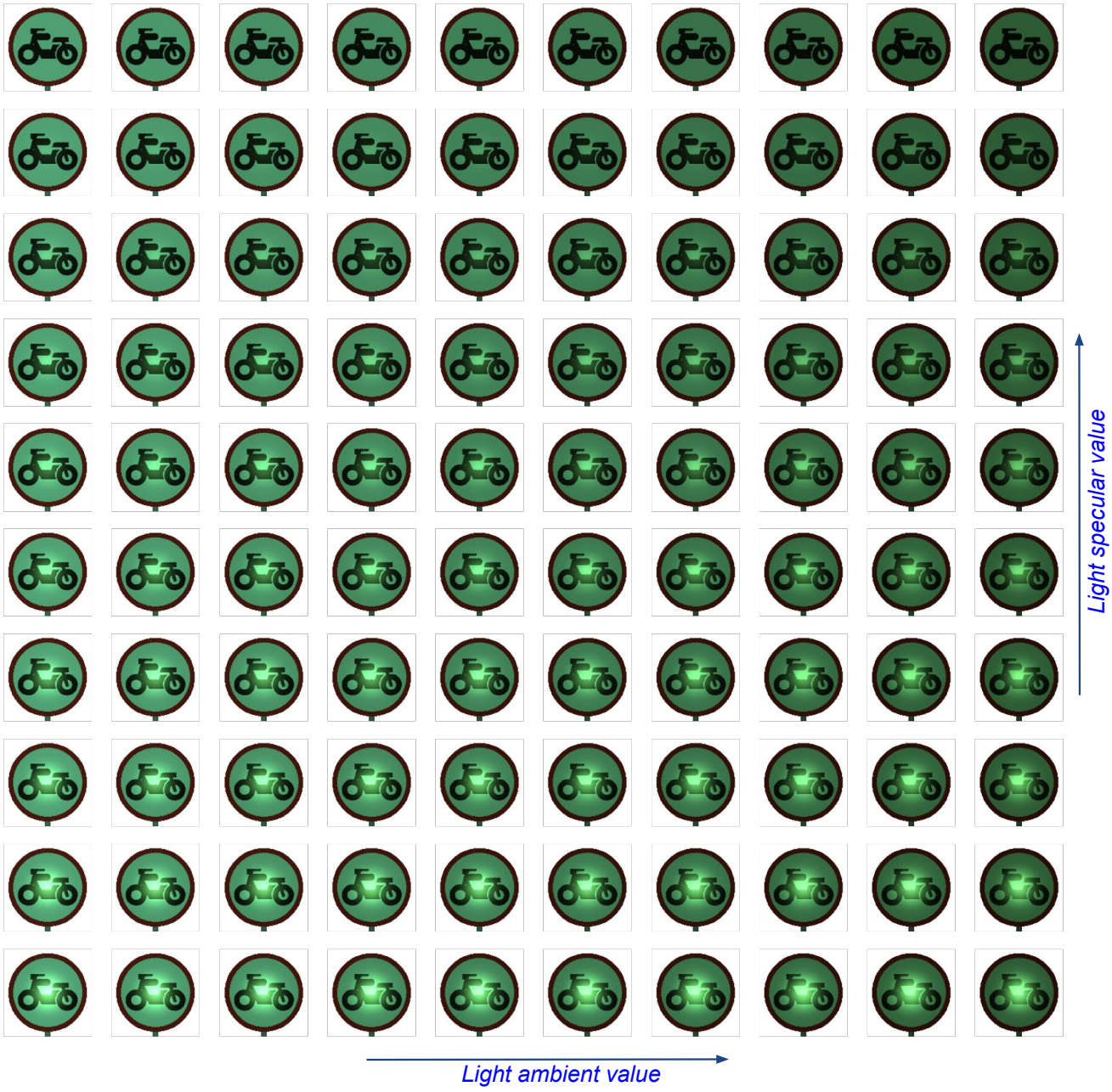


Figure 8. *Visualizing physical effects of 3D scene parameter perturbations:* **light ambient value (brightness) vs light specular value**.

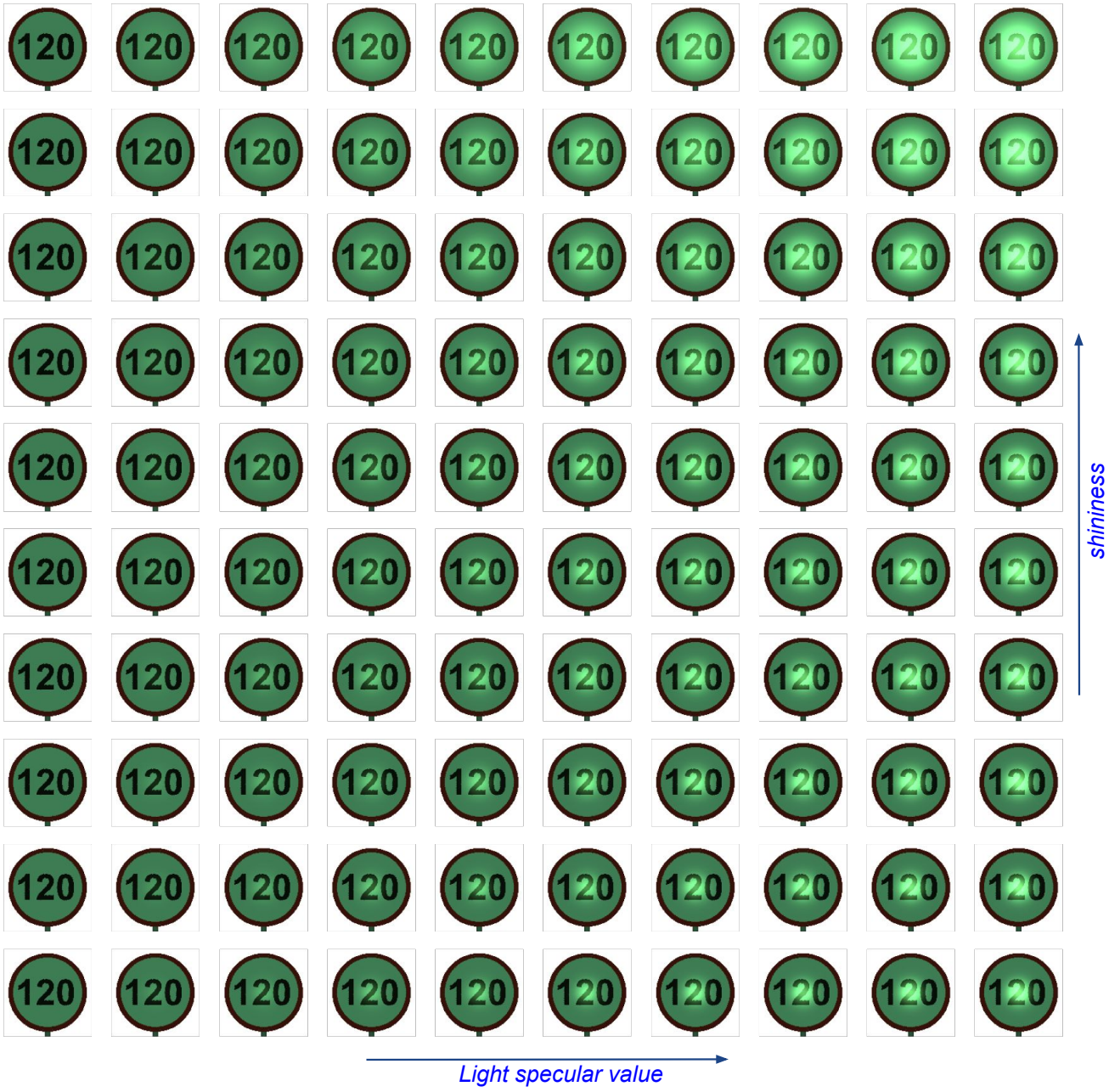


Figure 9. *Visualizing physical effects of 3D scene parameter perturbations:* **light specular value (brightness) vs material shininess**.

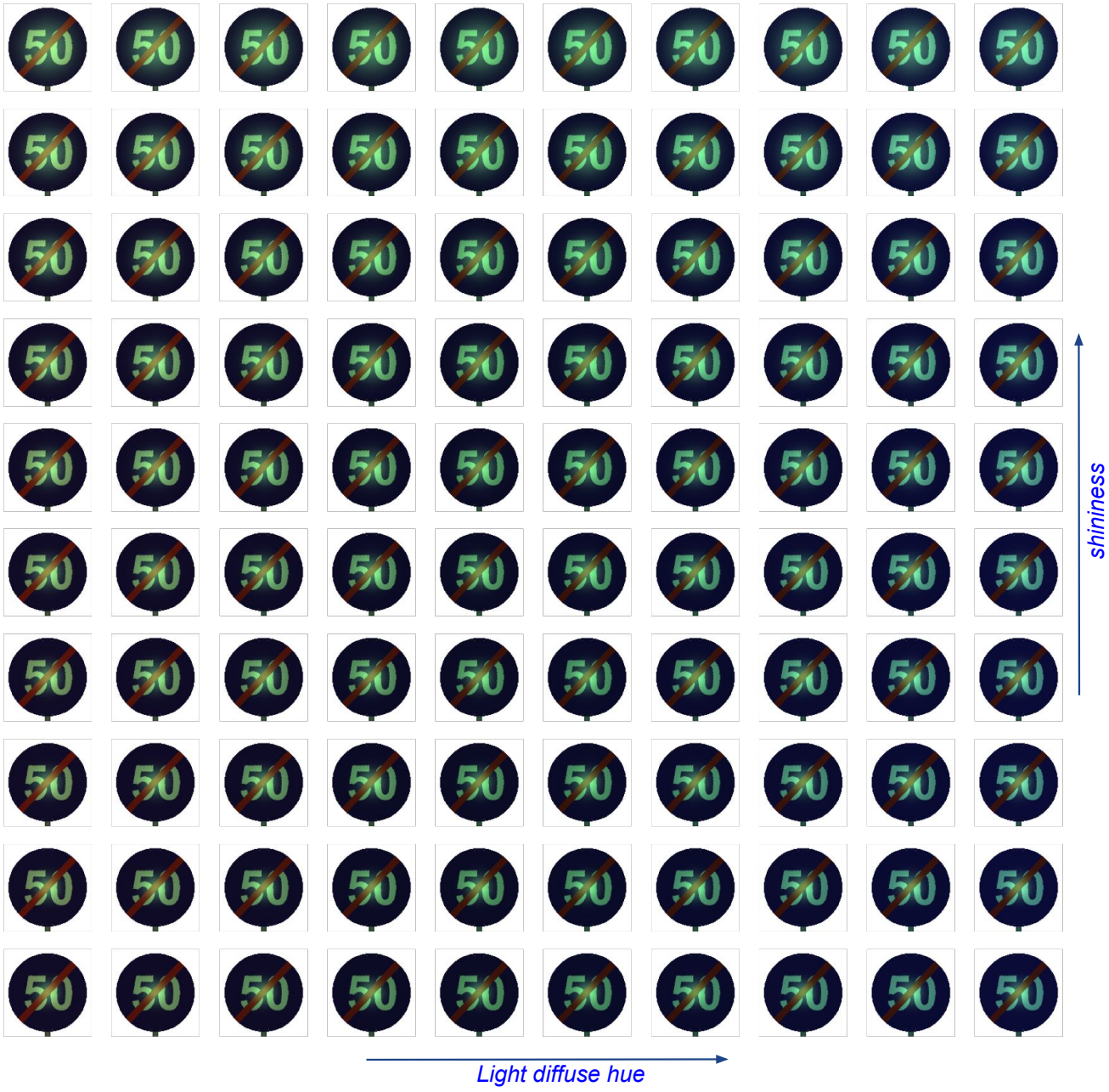


Figure 10. *Visualizing physical effects of 3D scene parameter perturbations:* **light diffuse hue (brightness) vs material shininess**.