An intrinsic image network evaluated as a model of human lightness perception

Richard F. Murray [1]***, David H. Brainard*** [2]***, Alban Flachot*** [1]***, and Jaykishan Y. Patel*** [1]
[1] ***York University, Toronto, Canada;*** [2] ***University of Pennsylvania, Philadelphia, U.S.A.***

Abstract

Lightness perception is a long-standing topic in research on human vision, but very few image-computable models of lightness have been formulated. Recent work in computer vision has used artifical neural networks and deep learning to estimate surface reflectance and other intrinsic image properties. Here we investigate whether such networks are useful as models of human lightness perception. We train a standard deep learning architecture on a novel image set that consists of simple geometric objects with a few different surface reflectance patterns. We find that the model performs well on this image set, generalizes well across small variations, and outperforms three other computational models. The network has partial lightness constancy, much like human observers, in that illumination changes have a systematic but moderate effect on its reflectance estimates. However, the network generalizes poorly beyond the type of images in its training set: it fails on a lightness matching task with unfamiliar stimuli, and does not account for several lightness illusions experienced by human observers.

Introduction

Human visual perception begins with the formation of images on the retina, and from this stimulus the visual system constructs useful representations of scene and object properties, including distance, shape, colour, and material. Making these inferences is often a difficult computational problem, and how the human visual system carries out this task is poorly understood. Lightness perception is one such problem that has been studied extensively. *Reflectance* is the proportion light that is reflected by a surface in the visible wavelength range (a purely physical property), and *lightness* is perceived reflectance (a perceptual property). Reflectance estimation is difficult because retinal stimuli are deeply ambiguous, at least locally: the light intensity at a given location on the retina could have been generated by a wide range of combinations of reflectance and illumination, so from pointwise luminance alone it is not possible to make accurate estimates of reflectance.

Decades of work on lightness perception have led to a good qualitative and sometimes quantitative understanding of some aspects of how human vision infers reflectance, including the role of shadow boundaries, occlusion edges, and natural scene statistics. (For reviews, see [1, 2, 3, 4, 5, 6].) However, this work has resulted in few image-computable models that predict human percepts over a wide range of scenes. Here we investigate whether recent progress using deep learning on the closely related problem of intrinsic image estimation in computer vision [7, 8, 9] may provide a new approach to modelling human lightness perception.

Previous work

There have been several attempts recently to use neural networks to estimate reflectance and other scene properties [7, 8, 9]. For example, Li et al. [9] used photorealistic renderings of complex scenes, and trained a network to map colour input images to chromatic reflectance. This network also learned to estimate local 3D shape and lighting conditions. The network was successful enough to support applications like object insertion, where the user adds new objects into a photograph, and the inserted objects are shaded in a way that is consistent with lighting in the rest of the scene.

These impressive results raise interesting questions with regard to modelling human vision. Do these networks really learn general strategies for lightness, or do they learn heuristics that are specific to the training images and do not support robust, general-purpose lightness perception? Are they promising starting points for models of human vision? Here we approach these questions by examining an intrinsic image network using the same kind of lightness perception experiments often used with human observers, and testing whether the network behaves like human observers in important ways. To anticipate, we find that the network is successful and human-like in some ways, but breaks down dramatically in other ways, particularly when generalizing to new kinds of scenes. An earlier version of this work was reported in [10].

Network and training

Network architecture

We used PyTorch to implement a subset of InverseRenderNet [8]. The resulting network was a 30-layer convolutional neural network (CNN) in an hourglass architecture with skip connections. In this architecture, the input image is progressively downsampled to a narrow bottleneck layer with many feature maps, and then progressively expanded back up to its original size. The network mapped an $n \times n$ achromatic luminance input image to an equally sized output image that gave pixelwise estimates of achromatic reflectance. To indicate that the network is derived from but also different from InverseRenderNet, we call it IRNet.

Training images

We used Blender [11], an open-source rendering package, to render 100,000 training images, 5,000 validation images, and 5,000 test images of random geometric objects with achromatic Lambertian surfaces (Figure 1, top row). We used images of simple geometric objects as a first step in exploring the hypothesis that many properties of human lightness perception arise from generic features of 3D scenes, such as cast shadows and occlusion, rather than from more subtle properties of genuine natural

scenes. Each scene contained 20 randomly positioned geometric objects (spheres, cubes, cylinders, icosahedra, cones, and tori). Each object had a one-third probability of being (1) coloured solid grey, with reflectance uniformly drawn from the interval [0.1, 0.9], (2) having a greyscale Voronoi texture, or (3) having a greyscale low-pass noise texture. The background consisted of three planes, intersecting at randomly chosen angles between 80° and 100°, and each independently assigned a randomly chosen reflectance from [0.1, 0.9]. Lighting consisted of an ambient source with fixed intensity, and a directional (i.e., infinitely distant) source whose intensity varied from trial to trial. The direction of the directional source was randomized across scenes, with azimuth randomly choosen between 30° and 60°, and elevation randomly chosen between 50° and 80°. The virtual camera was located at a position with an azimuth randomly chosen between 30° and 60°, and an elevation between 10° and 40°. The camera was directed at a randomly chosen point within 0.7 distance units of the origin (which was approximately where the three background planes intersected). All surfaces were Lambertian, and rendering did not include interreflections. We rendered a 256 × 256 luminance image of each scene, and an equally sized image of the reflectance at each pixel.

Training

We used supervised learning to train the network to infer reflectance images from luminance images. We used the Adam optimizer [12] with a mean-squared error criterion, and a batch size of five images. Batches were randomly sampled without replacement from the 100,000 training images. Training continued until error on the 5,000 validation images had asymptoted, which typically occurred within one or two epochs. We found that training was fast and reliable.

Network evaluation

Reflectance estimates

Figure 1 shows a typical luminance image (row 1, left panel) and corresponding reflectance image (row 1, right panel), as well as the trained network's output (row 2, left panel). The output was a reasonably good estimate of the reflectance image, and in particular there was little intrusion of cast shadows and shading (luminance variation as a function of surface orientation relative to the light source) on the reflectance estimate, with just some residual mottling visible. Figure 1 also shows a scatterplot of the network's estimated vs. true reflectance at 1,000 pixels randomly selected from 100 test images (row 2, right panel), and indicates that the reflectance estimates were strongly correlated with true reflectance. The network's median reflectance error across pixels was 0.03, in reflectance units that range from zero to one.

Figure 1 also shows the outputs of several other algorithms for comparison. The third row shows the results of a null model, whose reflectance estimates are simply an affine transform of the luminance input image. We chose the affine transform parameters by regressing true reflectance against luminance for individual images. This results in a generous estimate of model performance, as the model output is optimized for individual images, relying on knowledge of the true reflectance map for each image. The median pixelwise reflectance error of the null model was 0.11.

The remaining rows of Figure 1 show results for three additional computational models. For each of these models, there

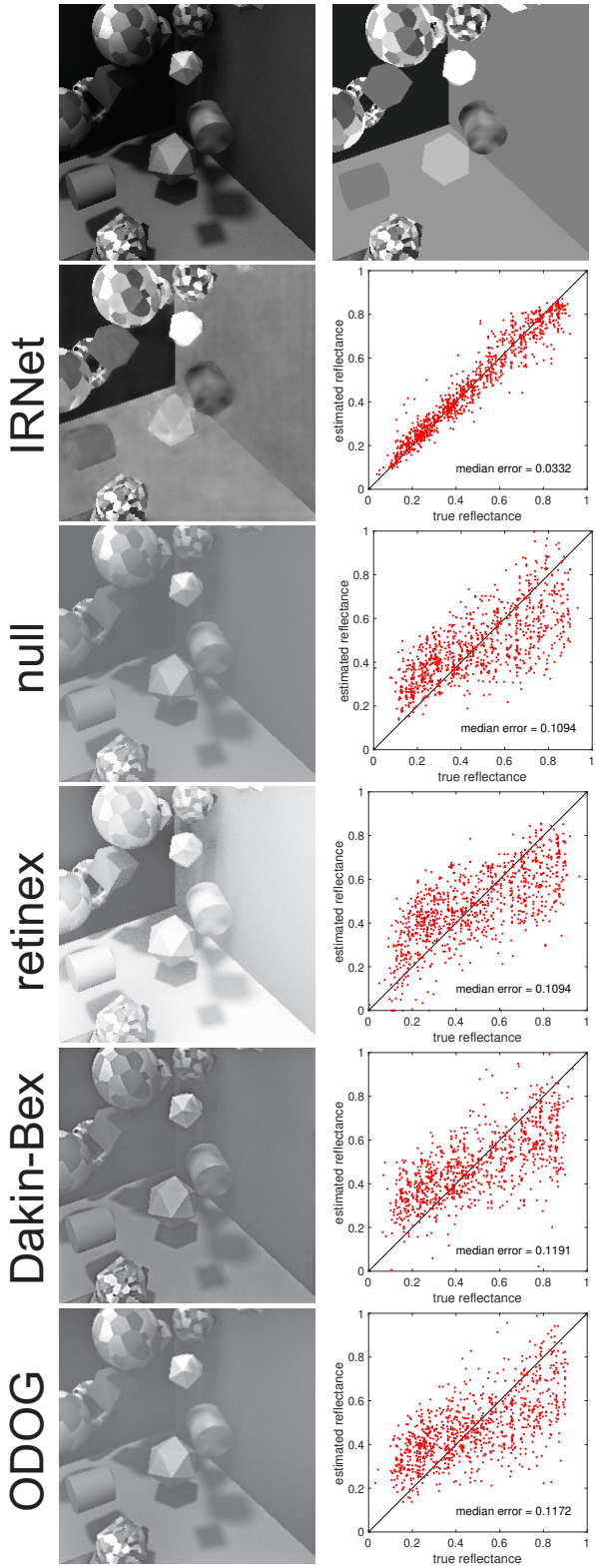

Figure 1. *Rendered stimuli and model outputs. The top row shows a typical luminance image (left panel) and its corresponding reflectance image (right panel). Each remaining row shows the output of a model in response to the luminance image (left panel) and a scatterplot of model output versus true reflectance at 1,000 pixels randomly selected from 100 test images.*

are reasons why they are perhaps not directly comparable to IRNet, but given the dearth of image-computable models of lightness, they provide informative if imperfect points of comparison. McCann [13, 14] presents his variant of retinex as a model of the 'sensation' of lightness; he distinguishes this from perceived reflectance, which he regards as a cognitively inferred property. (For McCann's model, we used the recommended parameter value nIterations=4.) Dakin and Bex [15] describe their bandpass normalization model sometimes as an account of lightness, and sometimes as an account of brightness (i.e., perceived luminance). Blakeslee and McCourt's [16] ODOG model is a model of brightness, but it has been evaluated as a model of lightness as well [17]. The bottom three rows of Figure 1 show the outputs of these models to the luminance image in the top row. In all three cases, cast shadows and shading are clearly visible in the model outputs (left panels), so the models did not fully discount lighting for these stimuli. These models give outputs in arbitrary units, not estimates of absolute reflectance, so the scatterplots in Figure 1 (right panels) show model outputs after an affine transform that gives them the best sum-of-squares prediction of true reflectance for individual images. This affine transform was fitted for each model via linear regression of true reflectance versus model output for individual images, as was done for the null model. Even after this optimal affine transform, all three model outputs were only loosely correlated with true reflectance. The models had about the same median reflectance error as the null model that is simply an affine transform of luminance (0.11), and much higher errors than IRNet (0.03).

Generalization

As a test of IRNet's ability to generalize even minimally beyond the type of images used to train it, we trained and evaluated the network on several variations of the geometric-object image set described above. We rendered images containing (1) only spheres, (2) only cubes, or (3) all six object types in the original image set. We also rendered images where objects had (1) uniform reflectance, with each object assigned a random reflectance from [0.1, 0.9], or (2) all three reflectance patterns in the original image set. Crossing these two factors gave $3 \times 2 = 6$ image sets. For each image set we rendered 100,000 training images, 5,000 validation images, and 5,000 test images. We trained an instance of IRNet on each of the six training sets, and evaluated each trained network on each of the six test sets, for a total of 36 evaluations.

Figure 2 shows the median reflectance error of each trained network on each test set. The network generalized well: the median error did not vary dramatically between estimates for the image type a network was trained on, and for a novel image type. In all cases, the median reflectance error was much less than errors found above for the null model and for previous computational models (around 0.11). We had expected that the variety of surface patterns and 3D shapes in the original image set would provide important information during training, but in fact the network was somewhat robust against mismatch between training and test sets, at least for these fairly small differences.

Thouless ratios

To evaluate the network's degree of lightness constancy, we ran human observers, the network, and the three other computational models discussed above in a lightness matching experiment. Figure 3(a) shows a typical stimulus, rendered in Blender. The reference cube (on the right) was in shadow, and was illuminated with the same ambient lighting intensity on all trials. The reference cube was given a reflectance of 0.1, 0.2, or 0.4 on different trials. The match cube (on the left) was not in shadow, and the intensity of its illumination from a distant point light source varied across seven values from trial to trial.

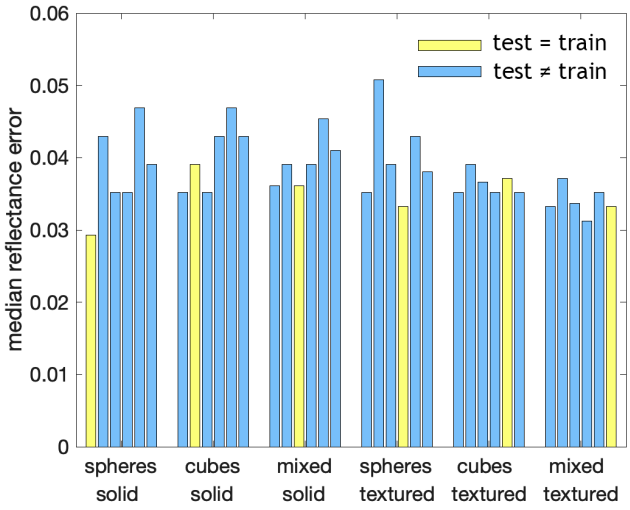

Figure 2. *Generalization error for IRNet. Each cluster of bars shows the median reflectance error for an instance of IRNet trained on a single image set (named in the cluster's label on the x-axis) and tested on six image sets. The top word in the x-axis labels indicates the type of geometric object in the image set (spheres, cubes, or mixed shapes), and the bottom word indicates the surface pattern (solid or mixed textures). The order of test sets within each cluster is the same as the order of training sets indicated on the x-axis, e.g., within each cluster, the leftmost bar shows test error on the "spheres solid" image set, and the rightmost bar shows test error on "mixed textured". Yellow bars show errors where the network was tested on the same type of image it was trained on, and blue bars show errors where the network was tested on a novel type of image.*

In the experiment with human observers, six observers were asked to adjust the reflectance of the match cube so that it appeared to be the same as the reflectance of the reference cube. The experiment consisted of three blocks, each block including all three reflectance conditions of the reference cube and all seven lighting conditions of the match cube. We also included a condition where the stimuli were flipped left-to-right, for a total of $3 \times 7 \times 2 = 42$ trials per block. Stimuli were displayed on a gamma-corrected monitor so that the physical stimulus luminance at each pixel was proportional to rendered image luminance.

In the simulated experiment with IRNet (trained on the original Blender image set, i.e., six geometric shapes and three surface texture patterns) and the other models, we found the models' output on the top face of the match cube for several match cube reflectances, and used interpolation to find the match cube reflectance for which the model gave the same output on the top face of the match cube and the reference cube.

Figure 3 shows results for a typical human observer as well as model observers, plotting match reflectance as a function of the illuminance on the top face of the match cube. Each panel

shows results for the three reference reflectances (0.1 in red, 0.2 in green, 0.4 in blue). For the human observer, match settings decreased moderately as illuminance increased (Figure 3(b)). This is a classic finding, where higher illuminance makes reflectances appear slightly lighter at the match location, and so observers make a lower match setting. We can quantify the degree of lightness constancy with the Thouless ratio, which is an index where 1.0 represents perfect constancy (a horizontal line in a plot like Figure 3), and 0.0 represents simple luminance matching (a line with slope -1) [2, 6, 10]. Thouless ratios for the human data in Figure 3(b) ranged from 0.34 to 0.44 (mean 0.40), and the median Thouless ratio over all six observers and three reference reflectances was 0.43. This is somewhat lower than values typically found in rich, complex scenes [18], but it is a reasonable value for the simple scene used here that provided relatively few lighting cues[1].

Results for IRNet were qualitatively similar to those from human observers, in that match reflectance declined linearly as a function of illuminance (Figure 3(c)). The network's Thouless ratios were in fact higher than those from human observers, ranging from 0.79 to 0.87, indicating better lightness constancy than acheived by humans. We believe that the qualitative similarity is most interesting: the network has partial lightness constancy, and when it fails, it does not fail dramatically or randomly, but has a linear falloff from perfect reflectance matching, just as human observers do.

Results from the other three models were very different (Figure 3(d)-(f)). Match reflectance declined sharply as illuminance increased, meaning that the models largely confounded illuminance and reflectance in these stimuli. Thouless ratios were around zero, so lightness constancy was poor, and the models were close to matching luminance instead of reflectance.

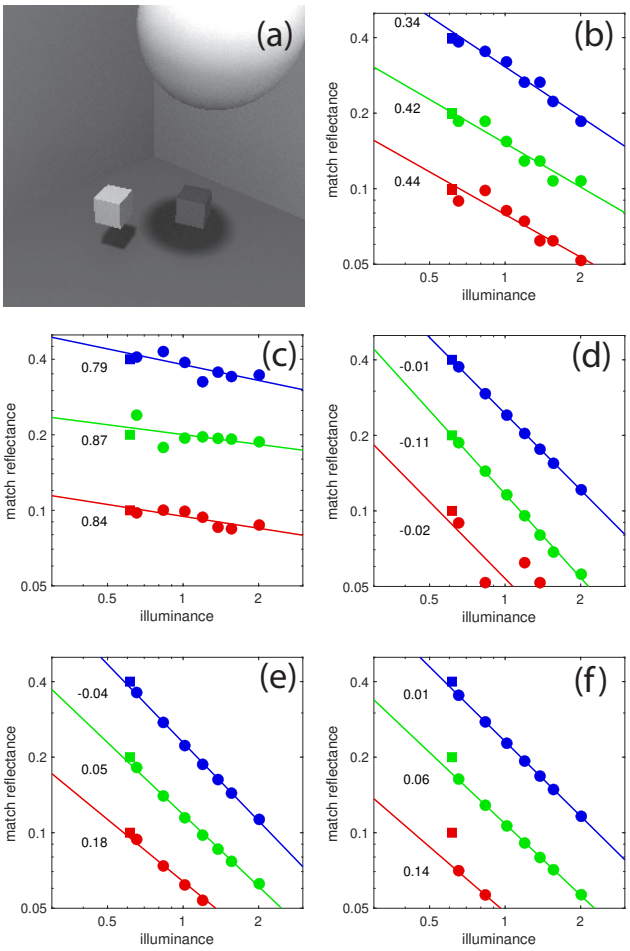

Figure 3. *A lightness matching task. (a) A typical stimulus. The right cube (in shadow) is the reference cube, and the left cube is the match cube. The remaining panels show match reflectance as a function of match illuminance for (b) a typical human observer, (c) IRNet, (d) retinex, (e) Dakin-Bex, and (f) ODOG. Each panel shows results for three reference reflectances: 0.1 in red, 0.2 in green, and 0.4 in blue. The numerical values reported next to each fitted line are Thouless ratios.*

Further generalization

To further test IRNet's ability to generalize beyond the type of image it was trained on, we ran the network in another lightness matching experiment. Here the stimuli were taken from a previous experiment that was run with human observers in virtual reality (Figure 4(a)) [19]. Stimuli were rendered in Unity [20]. We showed a reference patch under fixed lighting intensity, and a match patch under a lighting intensity that varied from trial to trial. Observers adjusted the reflectance of the match patch so that it appeared to be have the same achromatic surface colour as the reference patch. Human observers had good lightness constancy in this task, with Thouless ratios around 0.75 (results not shown).

As in the lightness matching experiment reported above, we measured the IRNet's output at the match patch for several reflectances, and interpolated the match reflectance at which the outputs were the same at the match and reference patches. The network failed dramatically in this matching task. With the lowest reference reflectance, the network's matches were a highly nonlinear function of illuminance (Figure 4(b)). With the other two reference reflectances, matches were approximately linear functions of illuminance, but decreased rapidly as illuminance increased. Thouless ratios were around zero, meaning that the network largely matched luminance, not reflectance. Although the stimuli in this experiment do not seem especially demanding, these results show that they are too far outside the network's training domain for good generalization, and they result in poor lightness constancy.

Lightness illusions

We also evaluated IRNet on several lightness illusions. Murray [6] created 16 $\times$ 16 grid versions of well-known lightness illusions in order to test a Markov random field model of perceived reflectance and illumination. Each grid image had two isoluminant test locations where human observers perceived one test location to be lighter than the other (except in control stimuli). Here we rendered scenes in Blender with a subset of these grid illusion images attached to to one side of a cube (Figure 5, left column). An instance of IRNet trained on the original Blender image set had mixed results in accounting for these illusions (Figure 5, right column). The model's output was consistent with a weak effect in the argyle illusion (row 1), but predicted a stronger illusion in the argyle control stimulus (row 2), where human observers see a reduced illusion. The model predicted a weak simultaneous contrast effect (row 3). The model predicted an effect in the wrong direction for the Koffka-Adelson illusion (row 4), but did account for the snake illusion (row 5) and White's illusion (row 6).

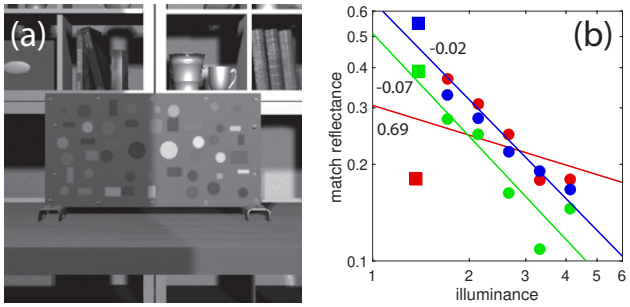

Figure 4. *Another lightness matching task. (a) A typical stimulus from Patel et al.'s [18] lightness matching experiment. The two large circles at the centre of the panel are the reference patch and the adjustable match patch. The luminance edge down the middle of the panel is a shadow boundary. (b) IRNet's match reflectance as a function of match illuminance, for three reference reflectances: 0.18 in red, 0.39 in green, and 0.55 in blue. The numerical values reported next to each fitted line are Thouless ratios.*

Discussion

There are few image-computable models of lightness perception, so recent neural network models are a significant and interesting development. The network we tested here has some strong similarities to human vision, but breaks down when generalizing too far beyond the training set. Overall, these results give some support to a normative view of lightness perception, in which human-like behaviours emerge naturally from a data-driven attempt to estimate surface reflectance from ambiguous images [21]. We did not train the network on human behavioural data, so insofar as it emulates human lightness perception, this is presumably because it was trained to solve a limited version of the same inference problem faced by the human visual system.

One promising direction for future work is to develop architectures that are informed by an understanding of the problem of inferring and representing surface properties such as lightness, instead of using off-the-shelf networks. For example, in typical deep learning approaches to colour constancy, estimates of lighting conditions do not guide estimates of surface colour; rather, lighting conditions are either estimated independently, or are calculated from estimates of surface shape and colour [7, 8, 9]. An interesting alternative would be for a preliminary network to estimate lighting conditions, and for these estimates to guide estimates of surface properties, as is often the case in models of human colour constancy [3]. Photorealistic datasets will also be useful for exploring lightness and colour constancy in more realistic scenes [22], and evaluating the hypothesis suggested earlier that important properties of human lightness perception result from broad, generic properties of natural scenes, such as occlusion and shadow boundaries. These new computational tools provide a valuable approach to leveraging the substantial knowledge of lightness perception that has emerged from many years of research, to construct image-computable models that make testable predictions for complex, naturalistic scenes.

Endnote

1. In this lightness matching experiment, observers could potentially match any of the three visible faces of the cubes. When running the computational models, we specified that the top face should be matched. With human observers, we had no control

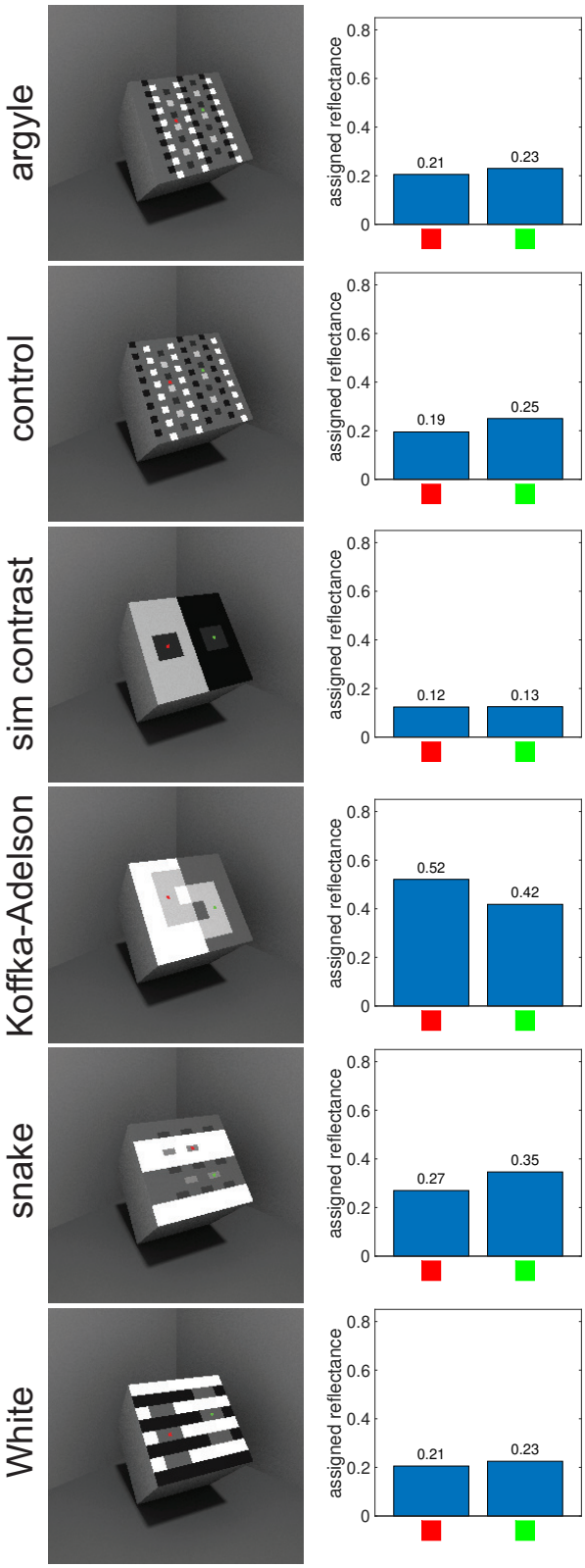

Figure 5. *Left column: lightness illusions attached to one side of a cube. The red and green patches (which were not included in the input to IRNet) indicate isoluminant test locations where the area surrounding the green patch appears lighter to human observers than the area surrounding the red patch. See Murray [6] for larger images of these illusions. Right column: IRNet's response at the test locations indicated by red and green markers. IRNet's responses were often not consistent with illusions seen by human observers.*

over which face or combination of faces were matched. This compliated the analysis of human data. The three faces of the match cube received the same amount of light from ambient illumination, but different amounts of light from the point light source that we varied to manipulate illuminance. As a result, depending on which face of the match cube we used to quantify the match illuminance on a given trial, we would arrive at a different Thouless ratio. We resolved this problem by using the illuminance at the top face of the match cube as the independent variable. This side of the cube faced the point light source most directly, so it had the greatest variation in illuminance from trial to trial, and using this illuminance as the independent variable gave an upper bound on human Thouless ratios. Our main conclusion from the lightness matching experiment was that IRNet had qualitatively similar data to human observers, and achieved better-than-human lightness constancy. These conclusions are supported by comparing the network's performance to an upper bound on human Thouless ratios.

References

[1] E H Adelson. Lightness perception and lightness illusions. In M Gazzaniga, editor, *The new cognitive neurosciences*, pages 339–351. The MIT Press, 2000.

[2] A L Gilchrist. *Seeing black and white*. Oxford University Press, 2006.

[3] D H Brainard and L T Maloney. Surface color perception and equivalent illumination models. *Journal of Vision*, 11(5):1, 2011.

[4] F A A Kingdom. Lightness, brightness and transparency: a quarter century of new ideas, captivating demonstrations and unrelenting controversy. *Vision Research*, 51(13):652–673, 2011.

[5] M D Fairchild. *Color appearance models*. John Wiley & Sons, Ltd., 2013.

[6] R F Murray. A model of lightness perception guided by probabilistic assumptions about lighting and reflectance. *Journal of Vision*, 20(7):28:1–22, 2020.

[7] M Janner, J Wu, T D Kulkarni, I Yildirim, and J Tenenbaum. Self-supervised intrinsic image decomposition. In *Advances in Neural Information Processing Systems 30*, pages 5396–5946, 2017.

[8] Y Yu and W A P Smith. InverseRenderNet: Learning single image inverse rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3163, 2019.

[9] Z Li, M Shafiei, R Ramamoorthi, K Sunkavalli, and M Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020.

[10] R F Murray, D H Brainard, J Y Patel, E Weiss, and K Y Patel. An intrinsic image network with properties of human lightness perception. In *Color and Imaging Conference 30*, pages 225–230, 2022.

[11] Blender Online Community. Blender, version 2.92.0, 2021.

[12] D P Kingma and J L Ba. Adam: a method for stochastic optimization. In *Proceedings of the Third International Conference on Learning Representations*, 2015.

[13] J J McCann. Lessons learned from Mondrians applied to real images and color gamuts. In *Proceedings of the IS&T/SID Seventh Color Imaging Conference*, pages 1–8. Scottsdale, AZ, November 1999, 1999.

[14] B Funt, F Ciurea, and J J McCann. Retinex in MATLAB. In *Proceedings of the IS&T/SID Eighth Color Imaging Conference*, pages 112–121. Scottsdale, AZ, November 2000, 2000.

[15] S C Dakin and P J Bex. Natural image statistics mediate brightness 'filling in'. *Proceedings of the Royal Society B*, 270:2341–2348, 2003.

[16] B Blakeslee and M E McCourt. A multiscale spatial filtering account of the White effect, simultaneous brightness contrast and grating induction. *Vision Research*, 39:4361–4377, 1999.

[17] T Betz, R Shapley, F A Wichmann, and M Maertens. Noise masking of White's illusion exposes the weakness of current spatial filtering models of lightness perception. *Journal of Vision*, 15(14):1:1–17, 2015.

[18] K Y Patel, A P Munasinghe, and R F Murray. Lightness matching and perceptual similarity. *Journal of Vision*, 18(5):1:1–13, 2018.

[19] K Y Patel, L M Wilcox, L T Maloney, K A Ehinger, J Y Patel, E Wiedenmann, and R F Murray. Lightness constancy in reality, in virtual reality, and on flat-panel displays. *Vision Sciences Society Annual Meeting*, 2022.

[20] Unity Technologies. Unity, version 2020.1.5f1, 2020.

[21] D Corney and R B Lotto. What are lightness illusions and why do we see them? *PLOS Computational Biology*, 3(9):e180, 2007.

[22] M Roberts, J Ramapuram, A Ranjan, A Kumar, M A Bautista, N Paczan, and R Webb. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10912–10922, 2021.