**Joint estimation of surface gloss and 3D shape**

**Gizem Kucukoglu** Department of Psychology, New York University

**Wendy J. Adams** Department of Psychology, University of Southampton

**Michael S. Landy** Department of Psychology, New York University

Center of Neural Science, New York University

(Order of authors has not been decided)

Title as of now: Joint estimation of surface gloss and 3D shape

**Abstract**

**Introduction**

**Methods**

**Stimuli**

We generated a physical and a rendered set of stimuli for both gloss and depth. The details of how each set of stimuli is generated are explained in detail in the following sections.

**Physical Depth Stimuli**

The physical stimuli for depth were 3-D printed ellipsoid arrays. Each ellipsoid array had one of eleven possible depth levels. We denote the physical values of depth as *di* (*i* = 1, 2, …, 11). We defined the stimulus using a Cartesian coordinate system where x- and y- axes lay fronto-parallel to the observer. The z-axis was parallel to the viewing direction. Twenty five spheres with x and y radii of 1cm were centered on a 5×5cm square grid. The spheres were then jittered randomly in the range of ±0.2cm in both x and y directions. To set the depth value of the ellipsoid array, the z radii of the spheres were scaled randomly in the range [0, *di*], where *di* = (*i* + 1)2/10 cm*.* This results in depth values in the range of [0.4, 4.4] cm. To 3-D print these shapes on a flat surface, we sliced the 3-D model in half through the x-y plane, creating a half-ellipsoid array. This does not affect the depth values of the ellipsoids as we scaled the z-radii in both top and bottom half of the ellipsoids.

The 3-D models of the objects were prepared using Blender 2.72b software and then imported into Print Wizard software provided by the Stratasys Mojo printer. The stimuli were printed using a Stratasys Mojo Desktop 3D Printer at NYU LaGuardia Studio. The 3-D printed stimuli were not hollow and were printed using the ABS*plus*model material in ivory color.

**Physical Gloss Stimuli**

The physical gloss stimuli were spray painted ping-pong balls. Each painted ball had one of eleven possible gloss levels. First, we spray painted each ping-pong ball with Rust-Oleum® Specialty Camouflage Spray Paint in Khaki color. This paint produces an ultra-flat finish.

To create a range of gloss levels we mixed matte and glossy varnishes in varying ratios and applied the desired varnish mixture on the painted ping-pong balls using spray bottles . We used Liquitex High Gloss Varnish and Liquitex Matte Varnish to prepare the varnish mixtures. For a given gloss level, the matte/glossy varnish mixture always added up to 10ml. For 10% gloss coating, we mixed 1ml glossy varnish with 9ml of matte varnish. Similarly, for 20% gloss, we mixed 2ml glossy varnish with 8ml of matte varnish. All the measurements were made using a 0.1ml sensitive syringe. The gloss levels had a range of [0, 100] % where physical gloss level had 10% increments. 0% gloss was the ping-pong ball with camouflage paint only and 100% gloss corresponded to the ping-pong ball with camouflage paint covered with glossy varnish only (10ml gloss varnish + 0ml matte varnish).

**Matching Physical and Rendered Gloss**

We came up with a novel method to quantify the physical gloss levels we created. We took photos of the physical gloss stimuli under a known, real-world illumination and using a fitting algorithm we fitted three parameters of Ward bi-directional distribution function (BRDF). We then used these parameters to set the gloss levels of computer-rendered stimuli (see next section).

A BRDF defines how light is reflected off of a surface. Ward BRDF (Ward, 1992) has four parameters where ρd determines diffuse reflectance, ρs determines specular reflectance and αx and αy determine the spread (roughness) of the highlights. When objects have isotropic reflections, αx = αy. In our fitting procedure we used Ward BRDF because it is close to the physical truth and it has simple parameters which are easy to manipulate.

*Determining the illumination field:* Using a SpheroCam HDR camera with 16mm fish-eye lens we captured a 360° panoramic, high dynamic range (HDR) image of a room. The aperture was set to f/5.6, shutter speed was 1/8 second and ISO speed was 200. The resolution of the image was 5396×2698. The room was well-lit using various sources of illumination like desk and floor lamps located in different places across the room. We chose an indoor location as we needed the light in the captured image to be constant over time. In an outdoor environment due to conditions like sun's motion or weather, the illumination might change over time. As we needed to photograph multiple objects under the exact same illumination, using an indoor location avoided this problem.

*Photographing physical gloss stimuli:* After taking the photo of the room we placed each physical gloss stimulus (painted Ping-Pong ball) in the same room and photographed them one by one. By doing so the room acted as an illumination field for the painted balls. Each painted Ping-Pong ball was stuck on a tripod and the tripod was placed in the same exact location where the SpheroCam stood when it took the photo of the room. This was done because we measured the illumination of the room from the viewpoint of the SpheroCam and to replicate this illumination we placed each Ping-Pong ball in the same location. SpheroCam was placed 27.5 cm away from the tripod with the stimulus. This distance ensured that the stimulus was in focus. The stimulus on the tripod was 114.5cm from the floor which matched the distance from floor to the center of the SpheroCam lens. We photographed every stimulus using the same aperture and exposure settings as used in photographing the room. Instead of taking a full 360° photo, we only took a photo of the 30° slice which had the stimulus in it. All photos were high dynamic range.

*Generating reference photos (I don’t like this subtitle):* Before the fitting procedure we created a sample rendered scene and matched the photos of each stimulus to this sample render. This was done to avoid the noise that small position shifts across the photos might cause. We rendered an initial scene where the photo of the room was the illumination field and there was a rendered sphere with size and position as similar to the photos as possible. All the renderings were done using RenderToolbox3 (cite) for MATLAB using Mitsuba renderer. The reflectance of the sample rendered sphere was chosen manually to look as close to the photo as possible. This was done to provide us visual guidance when matching the size and location of the sphere in the render and the photo. We did not need the gloss level to be correctly matched at this point.

We then ran an fminsearch with x-scale, y-scale, x-position and y-position were the four parameters to be fitted. In each iteration the algorithm created an affine transformation matrix with a set of four parameters, applied this matrix to the photo and calculated a pixel-wise sum of squared error between the sample render and the photo of a given gloss stimulus. The search converged when the four parameters that give the minimum error was found. For each set of best fitting parameters we checked whether the locations of the highlights on the render were centered on the locations of the highlights in the photos. As a result, each gloss stimulus photo was registered to the same reference scene and the position of the sphere in each photo was now the same.

*Gloss fitting procedure:* The reference photos generated in the previous step were in turn used to find the Ward BRDF parameters that match the gloss level of a given photo. First, we converted all of our reference renders to grayscale and applied a circular mask to each render to discard the pixel information coming from the background. For a given gloss level, we ran a 2-D grid search where the two parameters of the grid were ρs for specularity and α for highlight spread. The diffuse parameter ρd was always equal to 1-ρs. (This was done to ensure that the renderer does not clip the ρs + ρd to be equal to 1). As our objects had isotropic reflections αx and αy were equal to α at all times. In every iteration of the grid search the algorithm rendered a high resolution panoramic image with a rendered sphere located in the illumination field. The rendered sphere’s gloss level was determined by the parameters picked from the search grid. The image was then converted to grayscale and a circular mask was applied to discard the pixels outside of the sphere. Before calculating error we normalized both the reference photo and the rendered image by the mean luminance of each image respectively. In each iteration error was calculated as the pixel-wise sum of squared error between the reference photo and the rendered image of the sphere. The algorithm worked such that the function is evaluated at each grid point. Then grid is centered on the previous best value, and the range of each parameter is shrunk by 35%. The search repeats until the best fitting value changes by less than 0.001 for at least five iterations in a row.

We ran the gloss fitting for each gloss level separately. As a result, for each gloss level, we had three parameters (ρs, ρd = 1- ρs and α) that matched the physical gloss level of the painted Ping-Pong balls to the Ward BRDF parameters of a rendered sphere with the same illumination. However, these parameters were calculated on the normalized photo-render pairs so we first undid the normalization step. Each best fit parameter was multiplied by the ratio of mean luminance of photo and mean luminance of render to undo the normalization. To make the albedo of the surfaces the same, we averaged across all the diffuse parameter fits and used the resulting value as the diffuse component across all gloss levels. Specular component values were best fit by a linear line checked by leave-one-out cross validation, which ensured that the specular component was changing linearly as a function of gloss value. We adjusted the specular components so that they fall on this best fitting linear line.

*Should I show a plot of the fitted parameters?*

*Make Figure to show residuals*

**Computer-Rendered Stimuli**

The computer-rendered stimuli were ellipsoid arrays with gloss and depth levels that matched the gloss and depth levels of physical stimuli (Fig. #). Physical depth and rendered depth were matched by making the computer-rendered stimuli same size as the physical stimuli. The rendered stimuli were twenty five spheres arranged in a 5×5cm square grid. A given ellipsoid had x and y radii of 1cm with a ±0.2cm random jitter in both x and y directions. The z-radii of a given sphere randomly scaled in the range [0, *di*] cm where *i* = [2,10] excluding the lowest and the highest depth values in the physical range resulting in nine depth levels. This was done to provide a larger range for the observers during the task to allow for over- or underestimation of depth level. The computer-rendered stimuli lay fronto-parallel to the observer.

For each depth level, nine gloss level stimuli were rendered. For a given gloss level we used the three fitted Ward BRDF parameters obtained using the fitting method explained in the previous section. We excluded lowest (0%) and highest (100%) gloss levels from the rendered stimuli set to allow for over- or underestimation of gloss level during the task.

The images were rendered under two illumination conditions taken from the SYNS Dataset (citation? or link). For each illumination condition we had 81 stimuli (9 depth x 9 gloss levels) which resulted in a total of 162 computer-rendered stimuli. Rendering was done using RenderToolbox and Mitsuba.



**Figure #:** Two examples of computer-rendered stimuli. Left: lower depth and gloss levels. Right: Higher depth and gloss levels. Computer rendered stimuli were 5x5 cm random array of ellipsoids varying in depth, presented in two different light ‑fields. The depth levels of the computer-rendered stimuli matched that of the physical stimuli: it ranged from 0.8 cm to 4.0cm in increments of 0.4cm. The glossiness of the computer-rendered stimuli was also matched to the physical gloss levels we generated. We first determined the surface reflectance parameters of the Ward BRDF model that minimizes the difference between photographs of the painted spheres and rendered spheres. We then used these fitted Ward BRDF parameters to generate our computer-rendered stimuli.

We tone-mapped and inverse-gamma corrected each rendered stimulus using a piecewise function. The bottom 67% of the luminance range was mapped to range of [0,171] linearly and the top 33% of the luminance range was mapped to the range of [172,255] with an asymptotic logistic function. Should I add equations? As the luminance range was very different between the two illumination conditions we used two tone-mapping functions that satisfy luminance mapping above. Same function was applied to all stimuli under the same illumination condition. Instead of linearizing our monitors we inverse gamma corrected each stimuli using the gamma measured from the monitor calibration.

**Software and Apparatus**

The experimental software was written in MATLAB using Psychophysics ToolboxVersion 3 {Kleiner:2007ui}. The experiment ran on an iMac with 3.1 GHz Intel Core i5 equipped with AMD Radeon HD 6970M graphics card. The display used for the experiment was a Dell P780 CRT monitor with a resolution 1024×768. Prior to the experiment the display was calibrated for any non-uniformity in luminance. Observers were seated 55cm away from the display with their head resting on a chin-rest.

The physical gloss and depth stimuli were attached on black cardboards and were placed to the left of the observer 55cm away from the observer. All physical stimuli were labeled with numbers [1,11] where 1 corresponded to 0% gloss and 0.4cm depth and 11 corresponded to 100% gloss and 4.4cm depth. When the observers looked to their left, the depth stimuli lay fronto-parallel to the observer. Both sets of physical stimuli were lit by a 24 Watts Satco 40” Accent Light attached on the ceiling of the display case. The light fixture span the full length of the display case, illuminating each stimulus from the top.

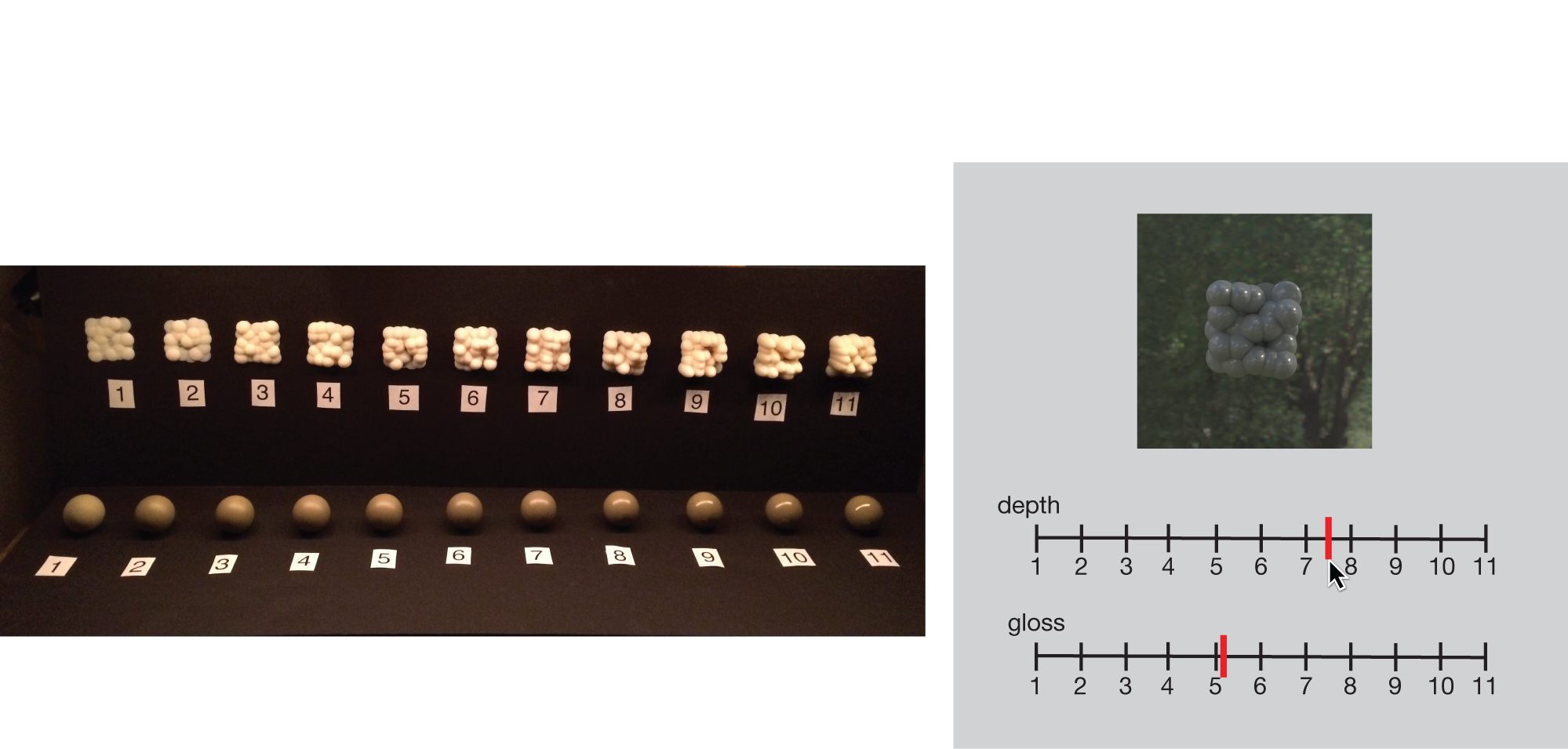
The monitor was encased in a black tunnel to block the light from physical stimulus reaching the monitor. The black tunnel was built from thick black foam boards and it extended to both sides of the chin-rest. While viewing the monitor from the chin-rest, the tunnel blocked right and left side of the observer’s view.

Observers viewed the computer-rendered stimulus on the monitor monocularly and physical stimuli with both eyes. Instead of an eye-patch, we built a circular occluder in front of the chin-rest for monocular viewing. This made it easier for the observer to look back and forth between the monitor and the physical stimuli without having to constantly remove and put on the eye patch.

Add photo of apparatus?

**Procedure**

We used a matching task in which a computer-rendered stimulus was presented on the screen at a given trial with two slider bars underneath. One slider bar was for matching gloss level and the other was for depth level. The observer's task was to pick one gloss level and one depth level from the physical stimuli sets which match the gloss and depth level of the stimulus on the screen and indicate the matched levels on the slider bars (Figure #a).



**Figure #:** (a) Physical stimuli set: The physical depth and gloss stimuli were placed to the left of the observer and labeled with numbers 1-11 in increasing depth/gloss. Observers viewed the physical stimuli with both eyes. (b) Task: In a given trial observers monocularly viewed a computer-rendered stimulus and chose a matching depth level and a gloss level from the physical stimuli set. They indicated the matching depth and gloss levels on the corresponding slider bar. A red marker indicated the observer’s choice which they can drag using a mouse.

The computer-rendered stimulus was always presented fronto-parallel to the observer and it was always viewed monocularly through a circular occluder. The physical stimuli sets for gloss and depth were viewed with both eyes and observers were allowed to move their head to look at them from different angles. The observer could look back and forth between the rendered and physical stimuli as many times as they needed until they made their decision. The numbers on the sliders matched that of the labels on the physical stimuli and ranged from 1 to 11 continuously. Observers used the mouse to indicate their responses on the slider bars. They were also allowed to pick values in between two gloss or two depth levels. Once the observer was happy with their decision they hit space bar on the keyboard to continue to the next trial. Each observer completed 162 trials in a two, 45 minute sessions. The order of computer-rendered stimuli were randomized across all trials.

**Observers**

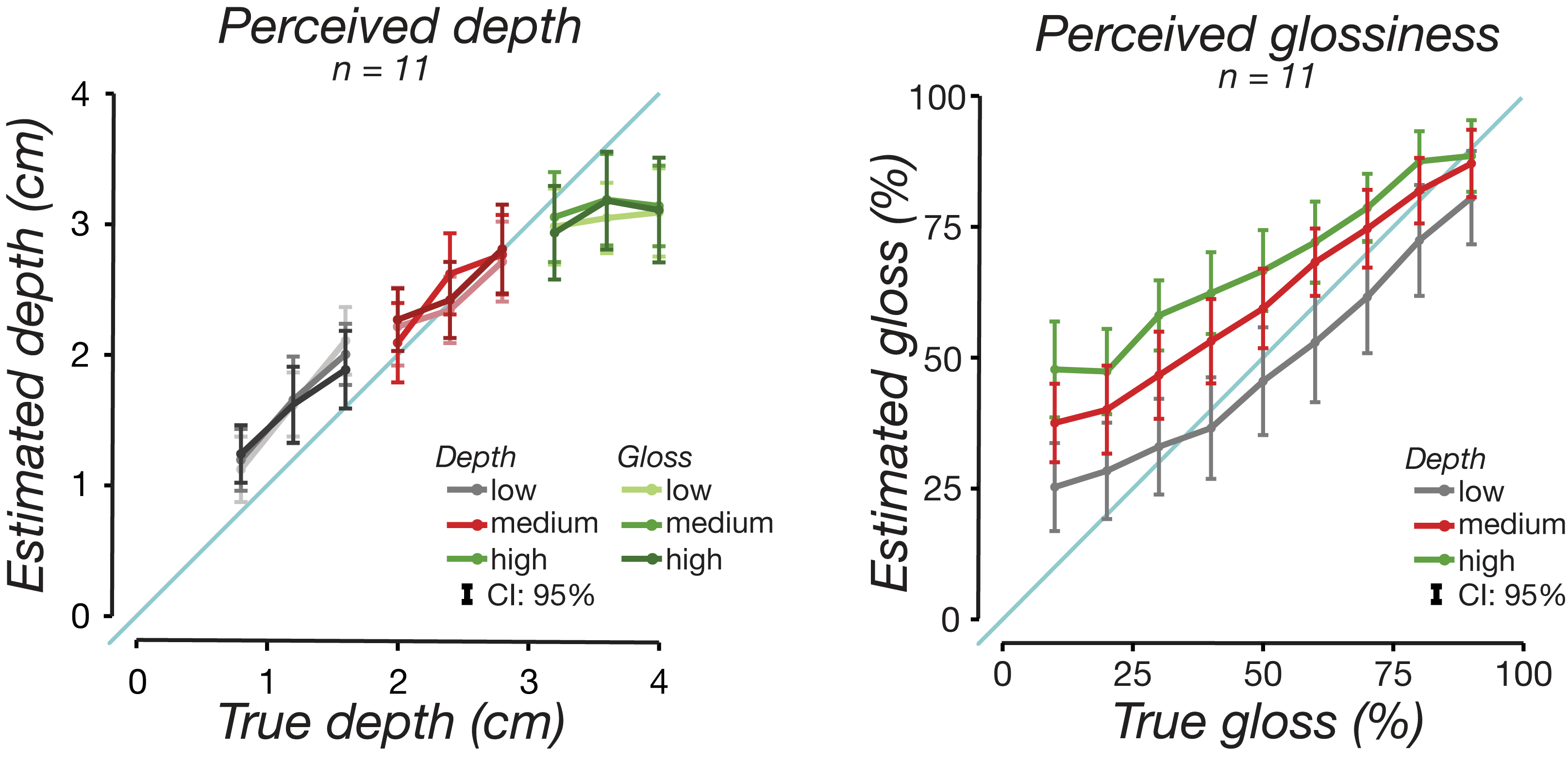
Eleven observers participated in this study. All observers, except for GK (one of the authors), were unaware of the purpose of the experiment. All observers had normal or corrected-to-normal vision. Say something about IRB here.

**Results**

Perceived depth and glossiness averaged across all observers are shown in Figure #a and Figure #b respectively. The results are binned into three gloss and three depth levels, where low depth (gray) corresponds to depth values [0.8, 1.2, 1.6] cm, medium depth (red) is [2.0, 2.4, 2.8] cm and high depth (green) is [3.2, 3.6, 4.0] cm. Similarly for gloss, low gloss (light shades) corresponds to gloss values of [10, 20, 30] %, medium gloss (medium shades) is [40, 50, 60] % and high gloss (dark shades) is [70, 80, 90] %.

Figure #a shows observers' depth estimates as a function of true depth where colors represent the three different depth levels and the saturation of the colors represent different gloss levels. If the observer perfectly matched perceived depth to true depth, the data would fall on the blue identity line. As can be seen from this figure, perceived depth is overestimated for low depth levels (gray) and underestimated for high depth levels (green). There is no difference in estimated depth across different gloss levels.

Figure #b shows observers' gloss estimates as a function of true glossiness where colors show different depth levels. Estimated glossiness changes depending on the depth level, where stimuli with high depth (green) are judged to be glossier and stimuli with low depth (gray) are underestimated in glossiness.

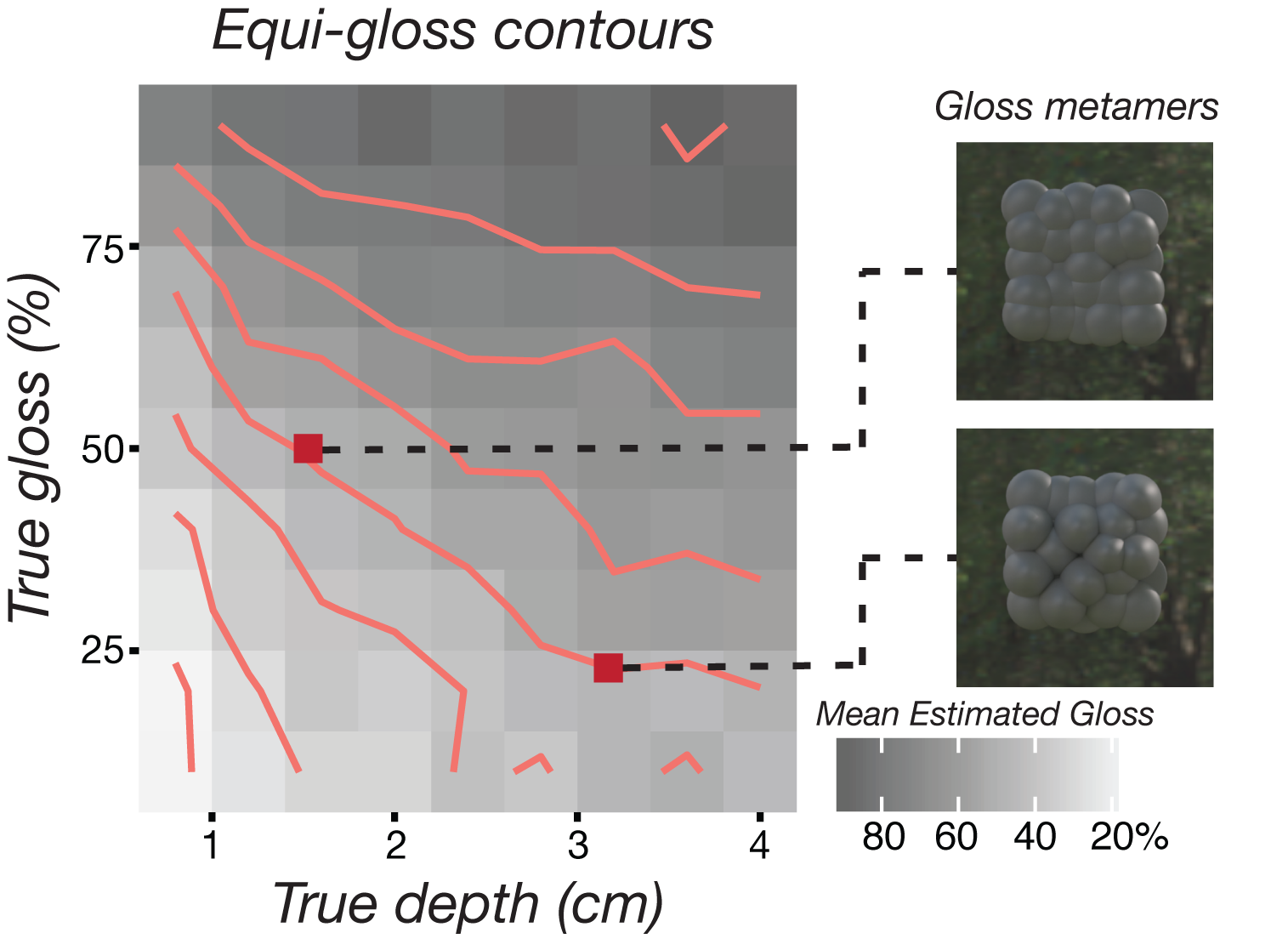


**Figure #:** Results for the matching task averaged across eleven observers. We converted observers' numerical matches back to centimeters for depth and percentages for gloss. Gloss and depth levels were binned into three groups: low, medium, high. Low gloss: [10, 20, 30] %, medium gloss: [40, 50, 60] %, high gloss: [70, 80, 90] %. Low depth: [0.8, 1.2, 1.6] cm, medium depth: [2.0, 2.4, 2.8] cm, high depth: [3.2, 3.6, 4.0] cm. (a) Perceived depth as a function of true depth averaged across all participants. Low depth (gray) values are overestimated in depth and high depth (green) values are underestimated in depth. There is no difference in depth estimates across different gloss levels. (b) Perceived gloss as a function of true gloss averaged across all participants. Objects with high depth (green) are perceived to be glossier than their true gloss and objects with low depth (gray) are perceived to be less glossy than they are. All error bars are 95% confidence intervals.

We used linear mixed modeling (LMM) to linearly fit estimated gloss and estimated depth while accounting for repeated measurements per participant (Bolker et al., 2009).  Random intercepts and slopes per participant were estimated for each main effect to reduce the influence of between-subjects variance on fixed effects estimation (Barr, 2013). For both gloss estimate and depth estimate the factors were gloss and depth levels. The factors were included in the model as summation contrast coded predictors, in other words, fixed effects. Type II Wald chi-square tests were used to test for significant deviation of the estimated parameters from zero, and significant interactions were followed up by ... To compensate for deviations from the normal distribution, estimated depth curves (Fig. #a) were Box-Cox transformed with lambda = 0.83.

We used lme4 package which estimated model parameters with LaPlace approximation (Bates, Mächler, Bolker, & Walker, 2015). (For each model, we conducted 10,000 parametric bootstrap runs to approximate reliable 95% confidence intervals as well as p-values to assess whether parameter estimates significantly differed from zero. - I didn’t do this but might need to.)

The results show a significant main effect of depth on estimated depth (𝛸2(1, N=11) = 3716.04, p<0.001) but no significant main effect of gloss or gloss-depth interaction. For estimated gloss we found a significant main effect of gloss (𝛸2(1, N=11) = 2775.75, p<0.001), depth (𝛸2(1, N=11) =719.25, p<0.001) and gloss-depth interaction (𝛸2(1, N=11) = 27.54, p<0.001).



**Figure #:** Equi-gloss contours for each true gloss-depth pair averaged across all observers (N=11). The red lines represent the equi-gloss contour lines. Any two gloss-depth pair that falls on a given line are perceived to be equally glossy. The perceived gloss level is indicated by the gray shading in the background where darker shades mean higher gloss level. The two red squares show a pair of gloss metamers: The stimulus that has 20% gloss and 3.2 cm depth (bottom) is perceived to be as glossy as 50% gloss and 1.6 cm depth (top).

As we collected a gloss estimate value for each gloss-depth pair, we can create a contour plot from this data that shows the equi-gloss contours averaged across all observers. Figure # shows this contour plot for true gloss as a function of true depth where the red lines are equi-gloss contours and the gray shading show the average gloss estimate for a given gloss-depth pair. Any two gloss values that fall on the same equi-gloss contour are, on average, perceived to be equally glossy.

**Image Statistics**

*Image set:* We generated a large set of our ellipsoid array stimulus to calculate various image statistics. We used 11 indoor and 11 outdoor light fields from the SYNS Dataset. For each light field we had 11 depth levels and for each depth level we had 11 gloss levels. We had 22 ellipsoid arrays for each depth-gloss-illumination triplet. This resulted in 2×11×11×11×22 = 58,564 images. The stimuli were rendered using Render Toolbox with Mitsuba. Each image was rendered hyperspectrally and then converted to grayscale.

*Statistics:* For each rendered image we calculated various image statistics. Contrast was calculated using root mean square (RMS) contrast, skew was calculated …., kurtosis was calculated by … In addition to these we calculated the Eero’s code stuff goes here and cite.