# Measuring the Visual Salience of 3D Printed Objects

**Xi Wang, David Lindlbauer, Christian Lessig, Marianne Maertens, and Marc Alexa** ■ *Technical University of Berlin*

**In an effort to validate the assumption that the saliency found in flat stimuli can be related to a 3D scene, the authors set up an experiment that examines if visually salient features exist for genuine 3D stimuli. They then validate computational models for the visual saliency of 3D objects.**

Visual saliency describes the idea that certain features in a visual stimulus stand out more than others and are more likely to attract an observer's attention. For flat stimuli such as 2D images, numerous experiments have shown that human observers are more likely to shift their gaze to such visually salient features.[1] A common assumption in the literature is that the saliency found in flat stimuli can be related to the underlying 3D scene. Although this assumption might seem intuitive, it has not been validated experimentally.

To evaluate this assumption, we set up an experiment that examines if visually salient features exist for genuine 3D stimuli. In this article, we describe that experiment and its analysis. Specifically, we asked whether different human observers consistently fixate on similar points on the surface of a given physical object under constant surface reflectance and fixed illumination. We used 3D printed objects as stimuli and tracked the observers' gazes while they were inspecting the objects' surfaces.

From the pupil position data, we then extracted the fixations during the first few seconds of the visual inspection. Because we know the object's geometry, we can relate the observers' fixations to gaze positions on the objects. Next, we analyzed the fixation data to address two questions. First, we tested whether human observers show consistency in their fixation patterns for the same object. Such consistency would be expected if visual salient features exist for physical objects and if these features guide fixation behavior. Second, we tested whether an algorithmic model of visual saliency, known as *mesh saliency,* can accurately predict human fixations.

To test the consistency between different human observers, we propose an analysis in which the observed fixation patterns on an object are tested against sequences of fixations that are unrelated to the geometry, yet are psychophysically plausible. These test sequences were generated from fixation sequences recorded for the same subject but using a different object. The resulting fixation sequences are realistic because they share the underlying oculomotor characteristics, but they are unrelated to the geometry of the object for which they serve as test sequences. To compare the generated and real fixation sequences, we quantified their similarity by computing the difference in eye-ray space.

Our results show a higher amount of agreement for fixations on the same object between observers compared with generated test sequences. This indicates the existence of visually salient features on 3D objects. The result also suggests that gaze directions systematically and meaningfully vary with the external stimulus, which is a necessary precondition for further analyses of human viewing behavior for 3D stimuli.

Because the data appears to be meaningful, we used it to investigate the predictive power of mesh saliency,[2–4] which is the only model with a psychophysical experiment supporting it. Mesh saliency extracts a measure of visual saliency from the local geometry of an object's mesh representation,

 Published by the IEEE Computer Society 

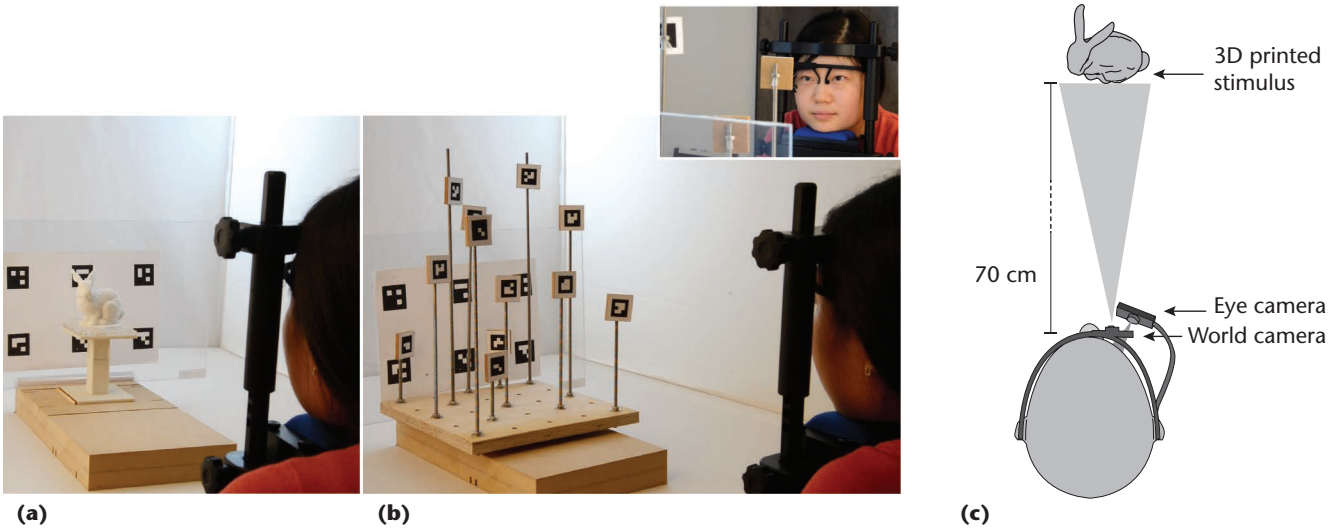


**Figure 1. Experimental setup. (a) Participant viewing one test object. (b) The calibration rig used to establish a mapping between the gaze position and object space. (c) Top view of the experimental setup, with the observer on the bottom and the object on the top.**

not taking surface scattering and view-dependent phenomena into account. The method has been validated against human fixations on flat stimuli of synthetically generated images of objects[5] and against points manually selected by human subjects.[6] As pointed out in earlier work,[7] the latter definition of feature points should not be mistaken for visually salient points in the sense of features that would trigger low-level human vision.

Lastly, to validate mesh saliency using fixation data for genuinely 3D stimuli, we compared the algorithmic predictions against permutations of the values across the mesh's vertices. If mesh saliency was indeed a good predictor of the fixation positions, then it would perform better than the permutations of itself. Our results show that this is not the case.

We believe that our experiment is an important first test of the assumption that theoretical concepts of human perception derived from experiments with 2D images also hold for the perception of 3D objects.

## Experimental Method

For our experiment, we recruited 30 unpaid participants (eight female and 22 male), all of whom were students. Their ages ranged from 19 to 31 years (median = 25), and all had normal or corrected-to-normal vision (based on self-reports). Four participants had previously participated in experiments involving eye tracking.

### *Stimuli and Apparatus*

The experiment was conducted in a quiet room. The stimuli were presented inside a 1 m × 1 m × 1 m box that was placed on a table and covered in white fabric (see Figure 1). The box was illuminated evenly, and illumination was kept constant across all objects and participants.

The participants were seated in front of the box, and their heads were placed on a chin rest 70 cm away from the stimuli. To measure gaze position, participants were equipped with the monocular Pupil eye tracker.[8] The eye tracker and software were calibrated to establish the subject-dependent correspondence between the raw gaze data and 3D object space using a custom procedure.

The 3D objects used in the experiments were 3D printed to ensure we had highly accurate realizations of their geometries (approximately 0.1 mm) and the input mesh used for printing was readily available digitally. We used a Contex DESIGNmate Cx SLS 3D printer, which produces homogeneous and highly diffused surfaces. Our experiment included 15 different objects to cover a range of different stimuli, from abstract to humanoid. Each object was presented in one of three predefined viewing directions. Orientations were randomly distributed among participants and the different objects and were ensured through a custom-built base with a dent for each orientation.

The online Web extra provides a more detailed video of the experimental setup (see www.cg.tu-berlin.de/research/projects/visual-saliency-on-3d-objects/).

### *Null Hypothesis and Design*

The goal of our experiment was to study human observers' fixation patterns while they were visually exploring 3D objects. In particular, we designed the experiment to test the following hypothesis, stated as null hypothesis:

Fixations on a 3D object are distributed randomly across the object's surface.

Experimentally, the hypothesis claims that observers produce their own idiosyncratic fixation patterns. A similar null hypothesis for visual saliency[9,10] has been repeatedly refuted when images were used as stimuli.[1] Not being able to refute this hypothesis for objects would be worrisome and would seriously call into question the validity of images as experimental stimuli.[11]

### Task

The participants were instructed to look at and visually explore each object. The objects were grouped into blocks consisting of the successive presentation of three different objects. After each block, participants were asked one question about one of the objects. An example mock question is "Was the dragon's mouth open?" We introduced this task to reduce the variability in the viewing patterns between observers, who might otherwise invent their own tasks.[12]

On average, the participants answered 79 percent of the questions correctly, with a standard deviation (SD) of 18 percent. Participants responded with "I am not sure" in nine out of 150 questions. We did not observe any systematic inability to answer questions for individual participants.

### Procedure

At the beginning of the experiment, each observer completed a demographic questionnaire. Thereafter, participants were introduced to the experimental setup and equipped with the eye tracker. The experimenter explained the calibration routine and the sequence of events during the experiment, which was as follows.

The objects were presented one at a time for a duration of approximately five seconds in blocks of three objects, with short breaks between subsequent blocks. Each participant completed five blocks, resulting in the presentation of 15 objects per participant.

Prior to each block, we calibrated the experimental setup. The calibration procedure consisted of two subsequent sequences of looking at the calibration targets (see Figure 1b) with the first one being the actual calibration and the second one a verification. If the error exceeded a threshold of 1.5 degrees, the calibration was deemed too inaccurate and repeated. To prevent participants from seeing the objects prior to data collection, we blocked their view with a moveable screen after the calibration and in between the presentations.

The objects were presented in random order. Each observer was presented with each object only once to avoid potentially confounding effects of habituation or boredom for repeated presentations of the same object. We familiarized participants with the procedure by having them complete a brief training session with one sample object (a teapot) prior to the first block.

Gaze position was recorded for the first few seconds after the onset of the stimulus. The exact time is a parameter in the data-processing pipeline.

### Data Collection and Representation

We collected both video data from the eye tracker's scene camera (30 fps) and gaze data (120 fps). The raw gaze data is in the form of estimated pupil center positions in the pupil camera's coordinate system at discrete time values.

In this article, we refer to data according to the following scheme. Superscripts starting with a lowercase i indicate the object, and subscripts starting with n indicate the subject. This conforms to a top view of the experimental setup, with the observer being on the bottom of the image and the object on the top (see Figure 1c). Hence, pupil data for object i collected from subject n at time sample $t_k$ is $\mathbf{g}_n^i(t_k): \mathbb{N} \mapsto \mathbb{R}^2$.

## From Pupil Positions to 3D Gaze Locations

Our procedure that relates measured pupil positions to 3D gaze locations on an object relies on an eye tracker with a world and eye camera, as already described, and a fiducial marker in a fixed relative position to the object (see Figure 1). We analyzed measured pupil positions in the eye camera image and extracted fixations as positions that remain within a fixed radius $\rho$ for a sufficiently long period of time $\tau$ (the minimum fixation duration). The fiducial marker's image in the eye tracker's world camera is used to determine the mapping from the object space to the world camera coordinate system. The mapping parameters from the world to the eye camera coordinates are determined in the calibration phase. Together with the known geometry, this setup lets us determine highly accurate gaze positions on the object.

Here, we provide an overview of the mapping from measured pupil positions to gaze locations on an object. For a detailed discussion, see earlier work.[13]

### From Pupil Positions to Fixations

To determine a fixation position $\mathbf{f}_n^i \in \mathbb{R}^2$ for object $i$ and subject $n$ in the eye camera image, we consider a sequence of measured pupil positions $\mathbf{g}_n^i(t_k): \mathbb{N} \mapsto \mathbb{R}^2$ at consecutive sampling times. If

$\mathbf{g}_n^i(t_k)$ remain in a small region of radius $\rho$ for a duration of at least $\tau$ milliseconds, then these are considered fixations $\mathbf{f}_n^i$ at the mean location of the $\mathbf{g}_n^i(t_k)$. We consider only the first $t$ seconds after stimuli onset for fixations because we are interested in spontaneous visual reactions. Together, the parameters $\rho$, $\tau$, and $t$ control the mapping from measured pupil positions to fixations $\mathbf{f}_n^i$.

For our analysis, we chose $\rho = 0.5$ degrees, which gave us a good balance between fixation length and data stability (illustrated in Figure 2). Using this parameter setting resulted in an average fixation duration of 315 ms (SD = 41 ms) for $\tau = 100$ ms and 346 ms (SD = 41 ms) for $\tau = 150$ ms. Additionally, we chose $t = 1.5$ seconds. On average, these parameter settings result in 3.75 fixations (SD = 0.31) per object and participant.

### From Fixations to 3D Gaze Positions

We employ the fiducial marker's image in the world camera and its known relative position to the object to estimate the position $\mathbf{t}_n^i$ and orientation $\mathbf{R}_n^i$ of object $i$ relative to subject $n$. This provides a mapping from points $\mathbf{x}_i \in \mathbb{R}^3$ in object space to points $\mathbf{w}_n^i \in \mathbb{R}^3$ in the world camera coordinate system of subject $n$:

$$\mathbf{w}_n^i = \mathbf{R}_n^i \mathbf{x}_i + \mathbf{t}_n^i . \tag{1}$$

To relate a fixation $\mathbf{f}_n^i$ to the corresponding gaze location on object $i$, we use the inverse of the mapping from 3D points $\mathbf{w}_n \in \mathbb{R}^3$ in the world camera coordinates to 2D pupil positions $\mathbf{p}_n$ in the eye camera image. This mapping is independent of the object because the world camera coordinate system serves as a reference, with the object dependence being described by Equation 1. The mapping is projective and in homogeneous coordinates and hence is given by

$$s\begin{pmatrix}\mathbf{p}_n\\1\end{pmatrix} = \mathbf{Q}_n\begin{pmatrix}\mathbf{w}_n\\1\end{pmatrix}, \mathbf{Q}_n \in \mathbb{R}^{3\times 4}, \tag{2}$$

where $s$ is a scaling factor. We determine $\mathbf{Q}_n$ from a set of correspondences $\{\mathbf{p}_i, \mathbf{w}_i\}$ obtained during calibration, where subjects are asked to fixate on fiducial markers on a custom build 3D rig (see Figure 1). Because $\mathbf{Q}_n$ is a projective transformation, it can be factored into an intrinsic camera matrix $\mathbf{A}_n^Q$ and a rigid transformation $\mathbf{T}_n^Q = \left(\mathbf{R}_n^Q, \mathbf{t}_n^Q\right) \in \mathbb{R}^{3\times 4}$ consisting of a rotation $\mathbf{R}_n^Q$ and a translation $\mathbf{t}_n^Q$. This allows us to associate with each fixation $\mathbf{f}_n^i$ a ray $\mathbf{r}_n^i$ in the world camera space whose direction is defined by

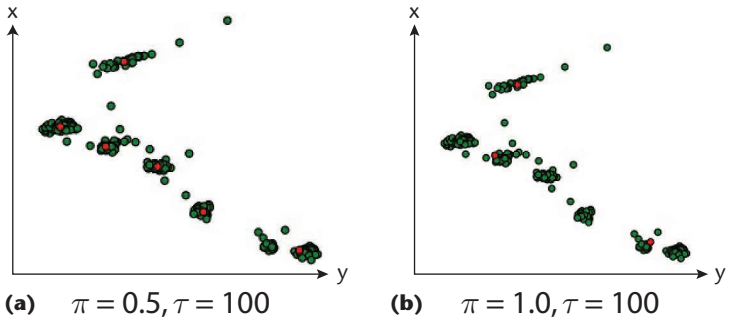


**Figure 2. Fixation classification for two values of $\rho$. Red dots indicate classified fixations, and green dots show raw data points. (a) A good clustering of fixations is obtained for $\rho = 0.5$ degrees, (b) whereas $\rho = 1.0$ degrees results in merged groups that visually do not belong together. The data here is a subset of fixations on one object from one participant.**

$$\mathbf{f}_n^i = \mathbf{A}_n^Q \mathbf{R}_n^Q \mathbf{r}_n^i . \tag{3}$$

The ray origin $\mathbf{o}_n$ is $\mathbf{t}_n^Q$, and the depth along it is indeterminate because $\mathbf{A}_n^Q$ is a projection. Pupil positions have limited accuracy for determining viewing direction[12] and eye tracking introduces additional measurement errors, so each fixation is in fact associated with a cone of possible gaze positions. The opening angle can be determined experimentally,[13] and we denote it by $c$.

To find a subject's gaze location $\mathbf{v}_n^i$ on object $i$ for a fixation $\mathbf{f}_n^i$, we determine the vertex $\mathbf{v}_a$ of the object's mesh representation $M$ whose mapping

$$\hat{\mathbf{p}}_a = \mathbf{Q}_n\left(\mathbf{R}_n^i \mathbf{v}_a + \mathbf{t}_n^i\right) \tag{4}$$

to pupil coordinates is closest to the fixation $\mathbf{f}_n^i$. For this, we first consider the set

$$\Gamma_c\left(\mathbf{f}_n^i\right) = \left\{ \mathbf{v}_a \in M \,\middle|\, \frac{\left(\hat{\mathbf{f}}_n^i\right)^T \mathbf{M}_n^Q \hat{\mathbf{p}}_a}{\left(\left(\hat{\mathbf{f}}_n^i\right)^T \mathbf{M}_n^Q \hat{\mathbf{f}}_n^i\right)^{1/2} \left(\hat{\mathbf{p}}_a^T \mathbf{M}_n^Q \hat{\mathbf{p}}_a\right)^{1/2}} > \cos c \right\}$$

of all vertices $\mathbf{v}_a$ in the cone with the opening angle $c$, where $\hat{\mathbf{f}}_n^i = \left(\mathbf{f}_n^i, 1\right)$ is the fixation in homogeneous coordinates. The sought-after gaze location $\mathbf{v}_n^i$ corresponding to fixation $\mathbf{f}_n^i$ is then the vertex in $\Gamma_c\left(\mathbf{f}_n^i\right)$ closest to the eye, which is given by $\mathbf{M}_n^Q = (\mathbf{A}_n^Q(\mathbf{A}_n^Q)^T)^{-1}$. (See related work[13] for a derivation.)

## Analysis

We tested the processed data against the null hypothesis (fixations are randomly distributed) in three steps: First, we defined a measure of agreement between two fixation sequences on the same object. Second, we generated test fixation sequences that are unrelated to a particular object in the experiment. And third, we tested whether

there is more agreement between real fixations than between real and test fixations.

### Comparing Fixation Sequences on the Same Object

Our analysis is based on fixation sequences. A fixation sequence $v_n^i$ is given as a set of vertex indices of the mesh representing object $i$. With $v_n^i$, $v_m^i$ being two such sets, we want to measure the amount of agreement between the two sequences. We denote this similarity as $s\left(v_n^i, v_m^i\right)$.

It is generally difficult to measure the distance between geometric sets. We therefore accept that the measure is asymmetric—that is, in general $s\left(v_n^i, v_m^i\right) \neq s\left(v_m^i, v_n^i\right)$. Now, recall that the vertices that are computed from measured data are known to be inaccurate. This means that the computed vertices are unlikely to be exactly identical for identical fixations, and small distances between two fixations in either image space or on the object's surface are unlikely to be meaningful. Consequently, it is most meaningful to consider two fixations identical when they are closer than some threshold.

This threshold is well defined in terms of the angular deviation between eye rays, which we have experimentally determined to be 0.8 degrees on average. Therefore, we will relate two vertices to the angle between the rays from the eye to each of the vertices.

The rays from eye to vertex depend on the eye's position relative to the object coordinate system, and different subjects have different eye positions. Fortunately, given the geometry of our setup, a slight change in eye position has only a negligible effect on the angle between two eye rays. This is because the distance between the eye and object is much larger than the distance between the two vertices on the object (at least for vertices potentially considered as the same fixation). We exploit this observation and measure the angle between eye rays from one of the two eye positions connected to the two data sets. With $\mathbf{o}_n$ as the center of rays for subject $n$, we can generate the eye ray for any point $\mathbf{x}$ in world coordinates based on the eye center as

$$\mathbf{r}_n(\mathbf{x}) = \frac{\mathbf{o}_n - \mathbf{x}}{\left\|\mathbf{o}_n - \mathbf{x}\right\|} \quad . \tag{5}$$

To compare two vertices in the sequences $v_n^i$, $v_m^i$, we consider the rays from the eye of subject $n$ for a vertex $\mathbf{v} \in v_n^i$ in the fixation sequence for object $i$ as well as a vertex $\mathbf{w} \in v_m^i$ in the fixation sequence of subject $m$, whose eye center is different.

The cosine of the angle between the view rays is

$$\mathbf{r}_j(\mathbf{v})^{\mathrm{T}}\, \mathbf{r}_j(\mathbf{w}) . \tag{6}$$

For vertices with associated eye rays that differ by an angle smaller than the defined threshold $\delta$, we consider the corresponding fixations identical.

Because we used a free viewing paradigm, we do not assume that fixations occur in an orderly sequence. We therefore compare each fixation in sequence $v_n^i$ to all fixations in sequence $v_m^i$ within the same time window $\tau$ regardless of their actual temporal position in their own sequence.

Our similarity measure is therefore

$$s\left(v_n^i, v_m^i\right) = \sum \left| \left\{ \mathbf{v} \in v_n^i, \mathbf{w} \in v_m^i \,\middle|\, \mathbf{r}_j(\mathbf{v})^{\mathrm{T}}\, \mathbf{r}_i(\mathbf{w}) > \cos(\delta) \right\} \right| . \tag{7}$$

As with other parameters, we varied $\delta$ to verify the stability of our result with respect to the particular choice made. Unless otherwise mentioned, we used $\delta$ = 1 degree. We generally suggest choosing a $\delta$ that is slightly larger than the measured angular deviation ($\delta$ = 0.8 degrees in our case). The number of fixations might vary with changes in $\delta$, but for the similarity measure, the test and real sequences must have the same number of fixations. Each sequence pair is compared twice in both directions due to the measurement asymmetry.

### Generation of Test Sequences

To analyze viewing behavior, we compare the fixation sequence $\mathbf{f}_n^i$ from subject $n$ for object $i$ against a mock fixation sequence. The mock fixation sequences are generated by projecting the observed sequence $\mathbf{f}_n^i$ onto another object $j$. We denote $v_n^{i \mapsto j}$ the test sequence for object $j$ generated by intersecting the eye rays computed from the fixation sequences $f_n^j$ with object $j$.

When projecting onto the true object, we ignore fixations that have no intersection with the object. This provides us with a set of fixated vertices that is unrelated to the visual stimulus created by $j$ (assuming that objects $i$ and $j$ are sufficiently different from each other). The mock sequences exhibit all properties of the fixation sequences that arise from normal oculomotor functioning, such as characteristic dwell times and velocities, and hence are more realistic than randomly generated sequences. In particular, because the physiological processes underlying the generation of eye movements are not well understood, it would be difficult to generate fair random sequences.

### Agreement of Fixations for the Same Visual Stimulus

To test whether different human observers tend to generate similar fixations for the same visual

stimulus, we compare the fixation sequences of two observers $n$ and $m$ for a fixed object $i$. Specifically, we ask whether the real sequence $v_n^i$ is more similar to the real sequence $v_m^i$ or to a test sequence $v_m^{j\to i}$ generated for observer $m$ from an object $j \neq i$. This is illustrated in Figure 3, where the real sequence of subject $n$ is compared with the real and test sequences of subject $m$.

The comparison was insensitive to the number of fixations because we used only sequences $v_m^{j\to i}$ that had the same number of fixations as the real sequence $v_m^i$. Summarizing, a single trial evaluates

$$\operatorname{sgn}\left(s\left(v_n^i, v_m^i\right) - s\left(v_n^i, v_m^{j\to i}\right)\right), \text{ s.t.} \left|v_m^i\right| = \left|v_m^{j\to i}\right|. \quad (8)$$

Note that this expression may evaluate to zero because ties are possible. According to the null hypothesis, there should be no difference between the test and real data, implying that there is no preference for the sign value in the trial.

In the experiment, each participant viewed each of the 15 objects only once. This results in 90 = $\mathbf{P}_2^{10}$ trials per object (because the comparisons are asymmetric). For each trial, we randomly selected a test sequence that satisfied the condition that the number of fixations is equal. Discounting the ties, we compared all remaining trial results to a cumulated binomial distribution to estimate the chance probability of that result.

We deliberately designed our similarity measure so that repeated fixations would result in a higher score. One reason for this is that repeated fixations of the same position might indicate a higher visual saliency of that object part. It is also possible that, because of our accuracy limits, two close features might not be properly resolved. In other words, spatially close features can result in the same measurement. We therefore repeated the experiment for all objects varying the parameters $\rho$, $\tau$, $t$, and $\delta$. In general, the results are stable with respect to these parameters. Figure 4 illustrates the resulting $p$ values.

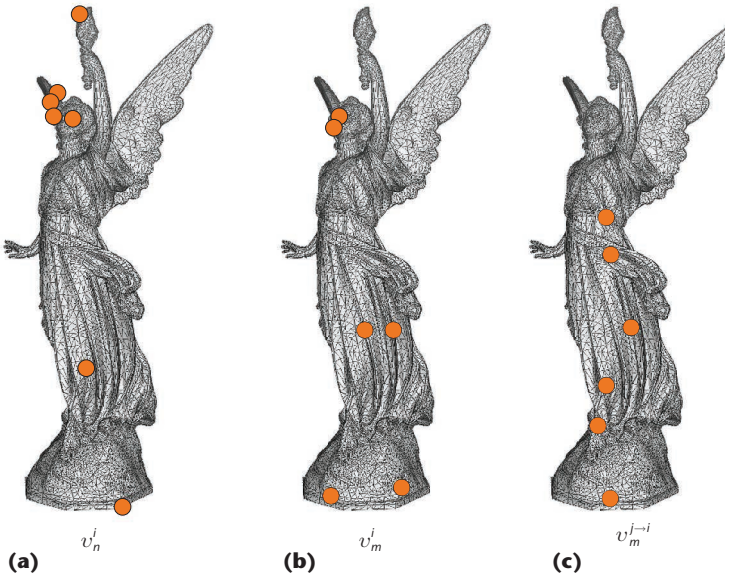


**Figure 3. Sample fixations for participants $n$ = 8 and $m$ = 13. The right image shows a sequence generated by taking a fixation sequence gathered while a different object $j$ was presented to subject $m$ and projecting this data on the current object $i$.**

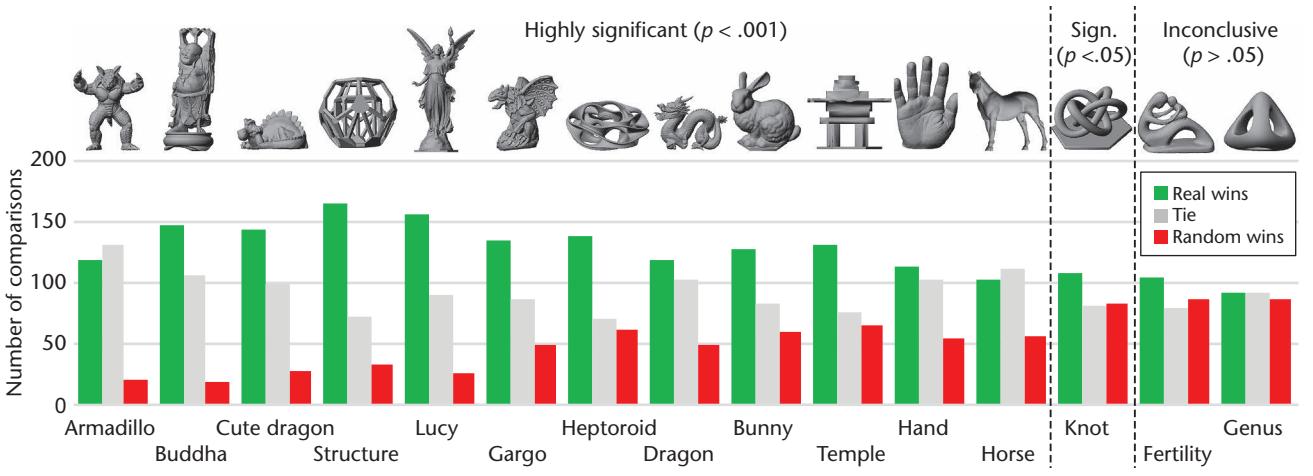


**Figure 4. Results of the trials between the measured data from one subject and either measured or generated data for subject $m$ (over all participants.) The green bars represent the trials in which measured data agreed more with measured data, the red bars show higher agreement between measured data and test data, and gray represents ties. For all but three objects, we find significant agreement across subjects for nearly identical stimuli. For nearly featureless objects, the results are less significant or inconclusive.**

For most of the objects, fixation patterns agree between observers. This is particularly the case for the more complex shapes with distinctive features. Only a few cases were less conclusive. We suspect that smoother or simpler shapes have fewer features that stand out so that fixations might be attracted by the occluding contours of the objects, and thus the agreement between observers drops to chance levels.

## Validating Computational Saliency Models

Our results indicate that the fixations we gathered in our experiment agree across subjects for similar stimuli. This means that the data is suitable for testing the validity of computational models of saliency.

The main tool we suggest for such an analysis is permutation of the values generated by the computational model. If a computational model has predictive power for fixations on the object, it will provide significantly higher saliency values for our data than its own permutations. Permutation tests are generally considered strong tests for significance,[14] and they directly yield $p$ values as the normalized rank of the original data relative to the permutations.

### *Unbiased Permutations of Saliency Values*

Computational saliency models provide, in one way or another, a scalar value for each point on an object's surface that describes the point's visual saliency. Our goal is to provide another scalar function over the surface such that the expected value for a sample drawn under the same conditions, as in our experiment, is unchanged. There seem to be (at least) two possible assumptions on how mesh saliency values are sampled:

- Samples are drawn uniformly from the surface. This assumption reflects that saliency models are object-based, and permutations could be interpreted as alternative view-independent, global, object-based saliency models (such as mesh saliency[2]).
- Samples are drawn uniformly from the visible part of the surface. This assumption reflects that fixations are based on a single view (that is, the probability of fixating on vertices on the back-side of the object is zero).

We can derive unbiased permutations for these two assumptions.

We assume here that the object is represented by a triangle mesh $\mathcal{M} = (\mathbf{V}, \mathcal{T}) \subset \mathbb{R}^3$. Quantities defined on a mesh are commonly given for each vertex or for each triangle. Although often not explicitly specified, these values can be extended to every point on the surface using basis functions $b_l(\mathbf{z})$, $\mathbf{z} \in \mathcal{M}$, where $l$ is the index of the mesh element—that is, a vertex or a triangle. To make this concrete, consider values given in vertices, which are linearly interpolated over triangles (more complex basis functions are possible and can be treated similarly).

The basis functions, together with the saliency values $s_l$ per vertex $l$, define the saliency over the piecewise linear surface as

$$s(\mathbf{z}) = \sum_l s_l b_l(\mathbf{z}),\ \mathbf{z} \in \mathcal{M}\,. \tag{9}$$

The expected value for a sample drawn uniformly from the surface $\mathcal{M}$ is then

$$E_{\mathcal{M}} = \int_{\mathcal{M}} \sum_l s_l b_l(\mathbf{z})\, d\mathbf{z} = \sum_l s_l \int_{\mathcal{M}} b_l(\mathbf{z})\, d\mathbf{z}\,. \tag{10}$$

To compute the expectation in the projection, we need the surface normals $\mathbf{n}(\mathbf{z})$, the eye ray $\mathbf{r}(\mathbf{z})$ (see Equation 5), and the binary information $v(\mathbf{z}) \in \{0, 1\}$ on whether the surface point $\mathbf{z}$ is visible. Then we can adjust Equation 10 for projection and visibility and get

$$E_P(\mathbf{z}) = \sum_l s_l \int_{\mathcal{M}} \mathbf{n}(\mathbf{z})^{\mathrm{T}}\, \mathbf{r}(\mathbf{z})\, v(\mathbf{z})\, b_l(\mathbf{z})\, d\mathbf{z}\,. \tag{11}$$

To make this concrete for linear interpolation, consider the area vector to vertex $l$:

$$a_l = \frac{1}{2} \sum_{(l,l',l'') \in \mathcal{T}} (\mathbf{v}_l - \mathbf{v}_{l'}) \times (\mathbf{v}_l - \mathbf{v}_{l''}) \tag{12}$$

encoding both vertex normal and associated area.[15] The area vector lets us succinctly write the integrals in these expectations for the case of piecewise linear basis functions. We have

$$B_l = \int_{\mathcal{M}} b_l(\mathbf{z})\, d\mathbf{z} = \frac{1}{3} \|a_l\|\,, \tag{13}$$

and we approximate the case for projections under the assumption that directional variation of the ray from the origin to the surface is small for a single triangle (and considering only visible vertices) as

$$B_l = \int_{\mathcal{M}} \mathbf{n}(\mathbf{z})^{\mathrm{T}}\, \mathbf{r}(\mathbf{z})\, b_l(\mathbf{z})\, d\mathbf{z} = \frac{1}{3} a_l^{\mathrm{T}} \mathbf{r}(\mathbf{v}_l)\,. \tag{14}$$

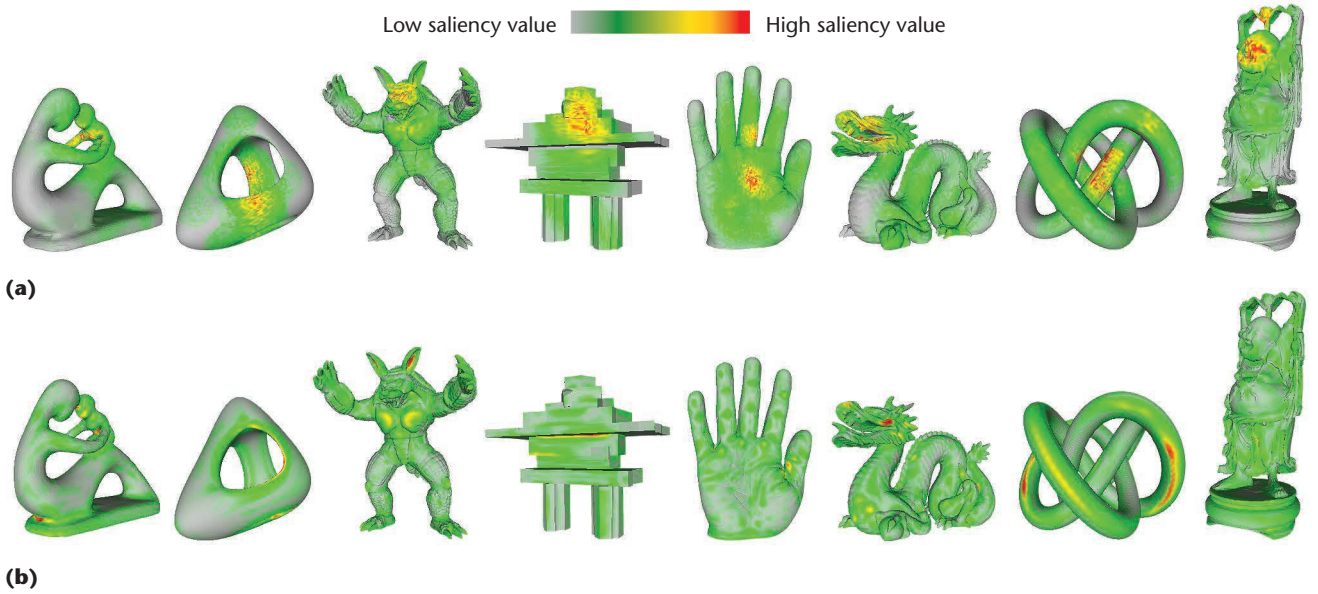


**Figure 5. Mesh saliency test. We compared the heat maps generated from (a) the collected fixation data to (b) the computed mesh saliency values on a representative set of objects used in the experiment.**

Note how Equations 13 and 14 differ only in the projection of the area vector.

We can easily keep the expected saliency value constant by taking permutations of the expected values $\{s_l B_l\}$ for the individual elements. Let $\{\pi[l]\}$ be a permutation of mesh elements. Then we assign permuted saliency values as

$$s'_l = s_{\pi[l]} \frac{B_{\pi[l]}}{B_l}. \quad (15)$$

### *Testing Mesh Saliency*

We compute mesh saliency values following the original approach of Chang Ha Lee and his colleagues[2] and obtain saliency values over the entire mesh using piecewise linear basis functions. Given a set of fixations $v_{ij}$, we need to compute a saliency score. Although it is clear that we want to sum over the contributions $\mathbf{v} \in v_n^i$ of individual fixations, there is no single correct way for considering each fixation.[5]

We opt for simply taking the saliency values $s(\mathbf{z})$ and summing them up over the set of fixations $v_n^i$. The reason is that the permutations we compute should yield the same expected value for point-wise samples, yet not necessarily for other means of collecting values. In particular, integrating over a small area on the surface (such as within a cone related to the measurement error) has different characteristics for mesh saliency compared with its permutations: mesh saliency provides smoothly varying values (see Figure 5b), so all saliency values are rather similar, while the permutations generate high-frequency noise. Area integrals for the permutations tend to be the same everywhere, whereas area integrals for the true mesh saliency depend on the fixation. This would introduce bias.

Based on summing up the saliency values of the closest vertices for all fixation sequences on an object, we calculate the score for mesh saliency values and 100,000 of its permutations. The rank of the mesh saliency score among all of its permutations yields the $p$ value. We perform this experiment for permutations adjusted to the object's surface as well as its projections.

As a sanity check, we created heat maps from the fixation data (see Figure 5a) and verified that they perform significantly better than their permutations. The results of this test for mesh saliency show that it generally fails to outperform its own permutations in a statistically significant way. As an example, Figure 6 shows the result for our preferred value for dispersion $\rho$ = 0.5 degrees and permutations adjusted for area. We considered objects in different viewing directions individually. Permutations adjusted for the visible projected area lead to comparable results, but they contained higher variations because of the large influence of the small projected area of fixations close to occluding contours. Results from using permutations of mesh saliency on the whole object yielded similar results.

The process of gathering fixations on 3D objects is more complex than for flat stimuli. We believe that the (expected) result of agreement between subjects for the fixations for most objects indicates that our experimental setup is meaningful and avoids excessive noise.

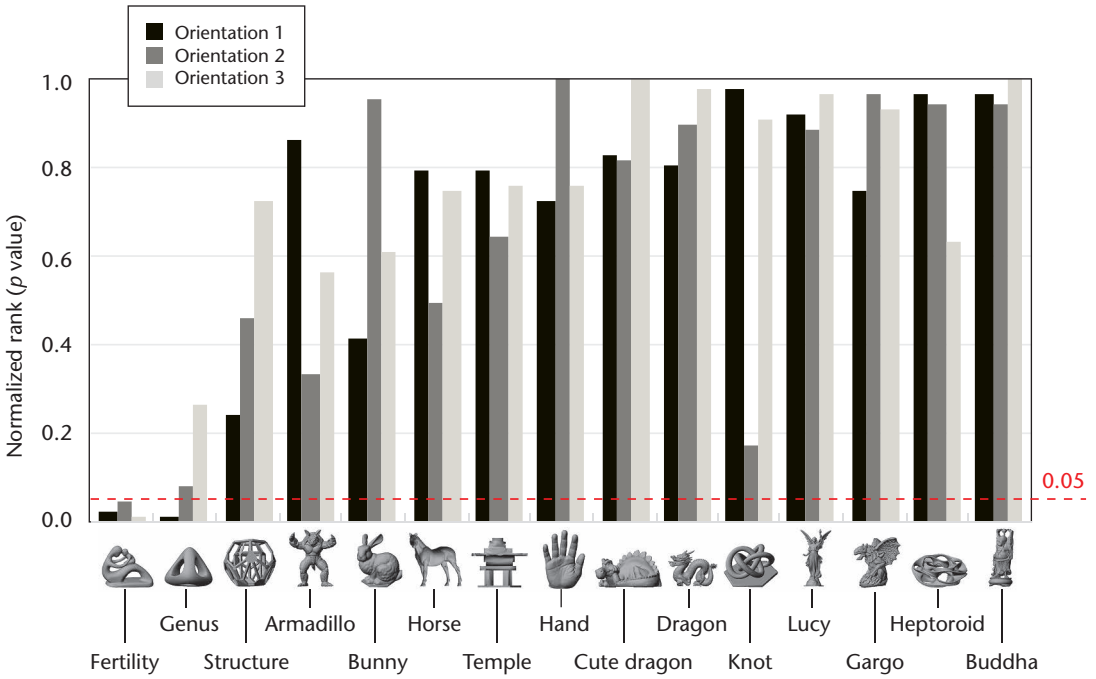


**Figure 6. Mesh saliency matching our recorded gaze data. This figure shows data for a dispersion value of 0.5 degrees. Permutations were computed with respect to the surface area.**

We have made the data we collected available on our website (www.cg.tu-berlin.de/research/projects/visual-saliency-on-3d-objects) so that other researchers may check its validity and use it for their visual saliency models for 3D objects.

The data is already useful in its current form, but we believe that varying experimental conditions further is important to learning about the invariance properties of visual saliency. In particular, our future work will look into varying the illumination conditions and the material properties of the stimuli. Whether fixations would still agree across such changes is still is an open question.

From this perspective, one might expect that similar parameters are varied in the generation of flat stimuli based on rendering. Compared with 3D printing or setting up lights physically, it appears more convenient to change material properties or lights in rendering. Yet, to our knowledge, this is not usually done. We suspect that such parameters might also affect the results of on-screen eye-tracking experiments. Such differences in the choice of the experimental setup might explain why our results differ from those of previous research with respect to the predictive power of mesh saliency.[5]

We would be curious to see how other computational models of visual saliency perform on our test data. Quite generally, we believe that predicting visual saliency independent of viewing conditions will be ill-posed; we probably need at least the orientation of the object toward the observer. These considerations are the main reason why we resisted the temptation to simply fit a saliency model to our data.

## Acknowledgments

*We thank Kenneth Holmqvist for valuable advice on design decisions and potential problems in eye-tracking experiments. Furthermore, we thank Felix Haase for help with performing the experiment. This work has been supported by the ERC through grant ERC-2010-StG 259550 (XSHAPE) and by the German Research foundation through grant DFG MA5127/1-1.*

## References

1. A. Borji et al., "Analysis of Scores, Datasets, and Models in Visual Saliency Prediction," *Proc. 2013 IEEE Int'l Conf. Computer Vision* (ICCV), 2013, pp. 921–928.
2. C.H. Lee, A. Varshney, and D.W. Jacobs, "Mesh Saliency," *ACM Trans. Graphics*, vol. 24, no. 3, 2005, pp. 659–666.
3. J. Wu et al., "Mesh Saliency with Global Rarity," *Graphical Models*, vol. 75, no. 5, 2013, pp. 255–264.
4. R. Song et al., "Mesh Saliency via Spectral Processing," *ACM Trans. Graphics*, vol. 33, no. 1, 2014, article no. 6.
5. Y. Kim et al., "Mesh Saliency and Human Eye Fixations," *ACM Trans. Applied Perception*, vol. 7, no. 2, 2010, article no. 12.
6. H. Dutagaci, C.P. Cheung, and A. Godil, "Evaluation

of 3D Interest Point Detection Techniques," *Proc. 4th Eurographics Conf. 3D Object Retrieval* (3DOR), 2011, pp. 57–64.
7. X. Chen et al., "Schelling Points on 3D Surface Meshes," *ACM Trans. Graphics*, vol. 31, no. 4, 2012, article no. 29.
8. M. Kassner, W. Patera, and A. Bulling, "Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-Based Interaction," *Proc. 2014 ACM Int'l Joint Conf. Pervasive and Ubiquitous Computing Adjunct Publication* (UbiComp), 2014, pp. 1151–1160.
9. C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, no. 4, 1985, pp. 219–227.
10. L. Itti and C. Koch, "Computational Modelling of Visual Attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, 2001, pp. 194–203.
11. J.J. Koenderink, "Virtual Psychophysics," *Perception*, vol. 28, 1999, pp. 669–674.
12. K. Holmqvist et al., *Eye Tracking: A Comprehensive Guide to Methods and Measures*, Oxford Univ. Press, 2011.
13. X. Wang et al., "Accuracy of Monocular Gaze Tracking on 3D Geometry," *Proc. Workshop on Eye Tracking and Visualization* (ETVIS), 2015; www.vis.uni-stuttgart.de/etvis/etvis2015/papers/etvis15_wang.pdf.
14. E.L. Lehmann and J.P. Romano, *Testing Statistical Hypotheses*, 3rd ed., Springer, 2005.
15. M. Alexa and M. Wardetzky, "Discrete Laplacians on General Polygonal Meshes," *ACM Trans. Graphics*, vol. 30, no. 4, 2011, article no. 102.

***Xi Wang*** *is a PhD candidate in the Computer Graphics Group, Faculty of Electrical Engineering and Computer Science, at the Technical University of Berlin. Her research focuses on combining the advantages of digital manufacturing and computer graphics for investigating human perception. Wang has an MS in computer science and engineering from Shanghai Jiao Tong University and an MS in computer science from the Technical University of Berlin. Contact her at xi.wang@tu-berlin.de.*

***David Lindlbauer*** *is a PhD candidate in the Computer Graphics Group, Faculty of Electrical Engineering and Computer Science, at the Technical University of Berlin. His research interests include dynamically altering optical properties of interactive devices. Lindlbauer has an MS in interactive media from the University of Applied Sciences Upper Austria. Contact him at david.lindlbauer@tu-berlin.de.*

***Christian Lessig*** *is a postdoctoral fellow in the Computer Graphics Group at the Technical University of Berlin. His research interests include efficient and reliable numerical techniques for image synthesis and computational tools for science and engineering. Lessig has a PhD in computer graphics from the University of Toronto. Contact him at christian.lessig@tu-berlin.de.*

***Marianne Maertens*** *is the head of an Emmy-Noether junior research group in the Faculty of Electrical Engineering and Computer Science at the Technical University of Berlin. Her research interests include the perception of material properties and the lightness of surfaces. Maertens has a PhD in psychology from the Max Planck Institute of Cognitive Neuroscience. Contact her at marianne.maertens@tu-berlin.de.*

***Marc Alexa*** *is a professor in the Faculty of Electrical Engineering and Computer Science and heads the Computer Graphics Group at the Technical University of Berlin. His research interests include creating, processing, and manufacturing shapes as well as developing intuitive interfaces for these tasks. Alexa has a PhD in computer science from the Darmstadt University of Technology. Contact him at marc.alexa@tu-berlin.de.*