Saliency3D: A 3D Saliency Dataset Collected on Screen

Yao Wang[*,†]
University of Stuttgart
Stuttgart, Germany
yao.wang@vis.uni-stuttgart.de

Qi Dai[*,‡]
Schaeffler
Shanghai, China
daqure@outlook.com

Mihai Bâce[‡]
KU Leuven
Leuven, Belgium
mihai.bace@kuleuven.be

Karsten Klein
University of Konstanz
Konstanz, Germany
karsten.klein@uni-konstanz.de

Andreas Bulling
University of Stuttgart
Stuttgart, Germany
andreas.bulling@vis.uni-stuttgart.de

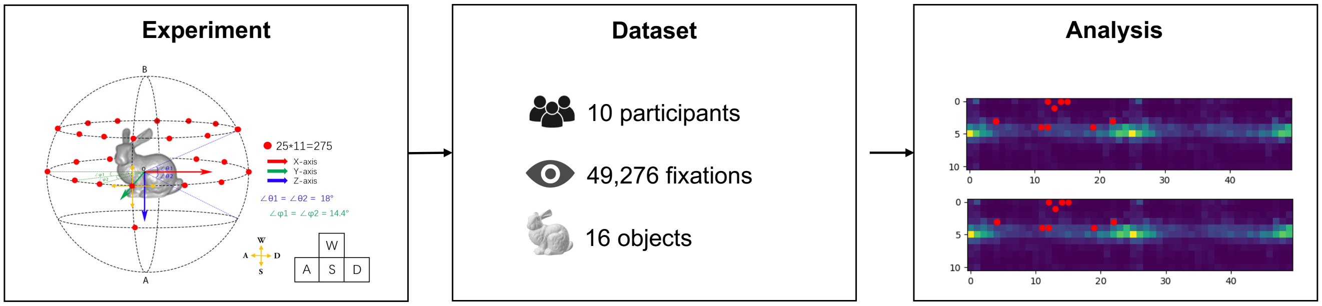

Figure 1: We propose a novel interactive experimental design to collect 3D saliency data using an eye tracker deployed on a screen. Red dots represent observation points restricted to a spherical surface. Using this method, we collected a 3D saliency dataset comprising 10 participants looking at sixteen 3D objects and analysed human gaze behaviour on our dataset.

ABSTRACT

While visual saliency has recently been studied in 3D, the experimental setup for collecting 3D saliency data can be expensive and cumbersome. To address this challenge, we propose a novel experimental design that utilises an eye tracker on a screen to collect 3D saliency data, which could reduce the cost and complexity of data collection. We first collected gaze data on a computer screen and then mapped the 2D points to 3D saliency data through perspective transformation. Using this method, we propose Saliency3D, a 3D saliency dataset (49,276 fixations) comprising 10 participants looking at sixteen objects. We examined the viewing preferences for objects and our results indicate potential preferred viewing directions and a correlation between salient features and the variation in viewing directions.

CCS CONCEPTS

• **Human-centered computing** → *Laboratory experiments*; **HCI theory, concepts and models**.

KEYWORDS

3D saliency, eye-tracking study, gaze behavior

ACM Reference Format:
Yao Wang, Qi Dai, Mihai Bâce, Karsten Klein, and Andreas Bulling. 2024. Saliency3D: A 3D Saliency Dataset Collected on Screen. In *2024 Symposium on Eye Tracking Research and Applications (ETRA '24), June 4–7, 2024, Glasgow, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3649902.3653350

1 INTRODUCTION

Visual saliency describes how certain features in a visual stimulus stand out and capture human attention [Wang et al. 2016]. The recent focus of saliency research is on 2D images [Cornia et al. 2016; Fosco et al. 2020; Liu and Han 2018]. In recent years, there has been an increasing interest in studying saliency in 3D environments, such as how people explore VR [Hu 2020; Sitzmann et al. 2018], and stereoscopic scenes [Fang et al. 2014; Wang et al. 2013]. Ramenahalli and Niebur [2013] explored the method of computing 3D saliency from 2D images. Additionally, depth information played an important role in the identification of visually salient regions in images [Desingh et al. 2013]. Wang et al. [2018] designed an experiment to observe 3D printed stimuli from multiple views. However, they limited the view within 90° due to their experiment setting – printed 3D objects are cumbersome and inflexible.

To overcome these limitations, we present a novel interactive design to collect 3D saliency data using an eye tracker deployed on a standard 2D computer screen. To simulate the manipulation of

[*] Both authors contributed equally to this research
[†] Corresponding author
[‡] A significant part of this work was conducted while at the University of Stuttgart

ETRA '24, June 4–7, 2024, Glasgow, United Kingdom

a real, 3D object, participants could freely switch views by pressing keys on a keyboard, while gaze data were collected using a screen-based eye tracker and then mapped to 3D saliency. We validated our data by studying two hypotheses informed by prior research: 1) humans show clear preference towards certain perspectives [Blanz et al. 1999] and 2) the existence of a bias towards facial features [Bindemann et al. 2005]. Our results confirmed that participants did have a consistent preferred viewing perspective and a strong face bias when viewing 3D objects.

The contributions of this paper are twofold: (1) We propose a novel experimental design to collect 3D saliency data with a screen-based eye tracker. (2) We collect *Saliency3D*, a 3D saliency dataset comprising 10 participants looking at sixteen 3D objects. Our dataset and code are publicly available at https://doi.org/10.18419/darus-4101.

2 RELATED WORK

The concept of visual saliency has been studied extensively over the past few decades. Numerous experiments have been conducted to investigate the various salient features.

Human viewing behaviour. The oculomotor system defines human viewing behavior with three major systems: the fixation-saccade system, the vestibule ocular system (VOR), and the smooth pursuit system [Wang et al. 2018]. In the fixation-saccade system, the eyes can maintain stability while humans fixate gaze [Martinez-Conde et al. 2004]. Moreover, the human eyes exhibit eye movements called saccades, which make quick and ballistic movements between two fixations in a very short time [Majaranta and Bulling 2014].

Salient feature. Previous studies have demonstrated that saliency strongly correlates with low-level and high-level features, influencing human visual attention [Cong et al. 2018; Kummerer et al. 2017; Xu et al. 2014; Zhang et al. 2008]. Reinagel and Zador [1999] showed that image regions with higher spatial contrast often attract more attention from humans. Furthermore, Baddeley and Tatler demonstrated that high-frequency edges dominate predicting fixation positions [Baddeley and Tatler 2006]. Studies conducted by Engmann et al. [2009]; Itti et al. [1998]; Jost et al. [2005] revealed that color is a significant factor that influences human visual attention. In addition, high-level salient features, such as semantic information, abstract concepts, and task requirements, also play an important role in capturing human attention. For example, faces strongly attract human attention [Strohm et al. 2023], even if the faces are unrelated to the goal of the experimental task [Bindemann et al. 2005; Langton et al. 2008].

3D saliency. Compared to 2D saliency, 3D saliency brings more factors that affect human viewing behaviour [Lavoué et al. 2018; Wang et al. 2013]. Three-dimensional shapes and lighting significantly influence attention [Lavoué et al. 2018], while human visual attention is influenced by depth information [Desingh et al. 2013; Lang et al. 2012]. Recently, Bruckert et al. [2023] used the BubbleView [Kim et al. 2017] metaphor to crowdsource visual attention data on 3D graphical content. Hu et al. [2021, 2020, 2019] analyzed and predicted fixations in virtual reality environments. While 3D saliency has been studied extensively, there is currently no easy-to-use method to collect 3D saliency data using commonly available hardware. We propose an experimental design that requires a 2D computer monitor and a remote eye tracker.

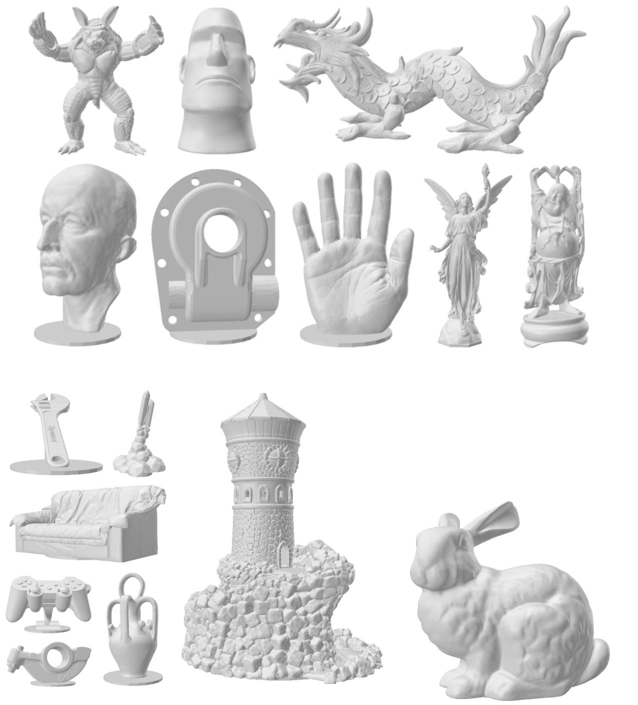

Figure 2: Sixteen selected stimuli [CzernO 2021; Turk et al. 2003; Wang et al. 2018]

3 SALIENCY3D DATASET

3.1 Data collection

Stimuli. The selected stimuli should be rich in both high-level and low-level features [Henderson and Hollingworth 1999; Itti et al. 1998]. Sixteen 3D objects are selected as stimuli (see Figure 2). Eleven objects (*Dragon, Hand, Planck, Sofa, Space_shuttle, Spanner, Vase, Watchtower, Casting, Game_controller, Rockarm*) are from Wang et al. [2018], and four objects (*Bunny, Happy_buddha, Armadillo, Lucy* are from The Stanford Models [Turk et al. 2003]. One of the stimuli *face* was selected from website [CzernO 2021].

Participants. We recruited 10 participants (6 male, 4 female) from the local university[1]. All participants reported normal or corrected-to-normal vision. Participants were asked to use a mounted chin rest to minimise the influence of head movements on gaze data quality. They were compensated for their participation and could stop anytime without adverse consequences. All personal information was fully pseudonymised.

Apparatus. We used a desktop computer and a 24.5-inch monitor with a 1920 × 1080 pixel resolution to display the stimuli. Participants' gaze data was collected using the Eyelink-1000 Plus Desktop

[1] The university ethics committee approved our study prior to data collection.

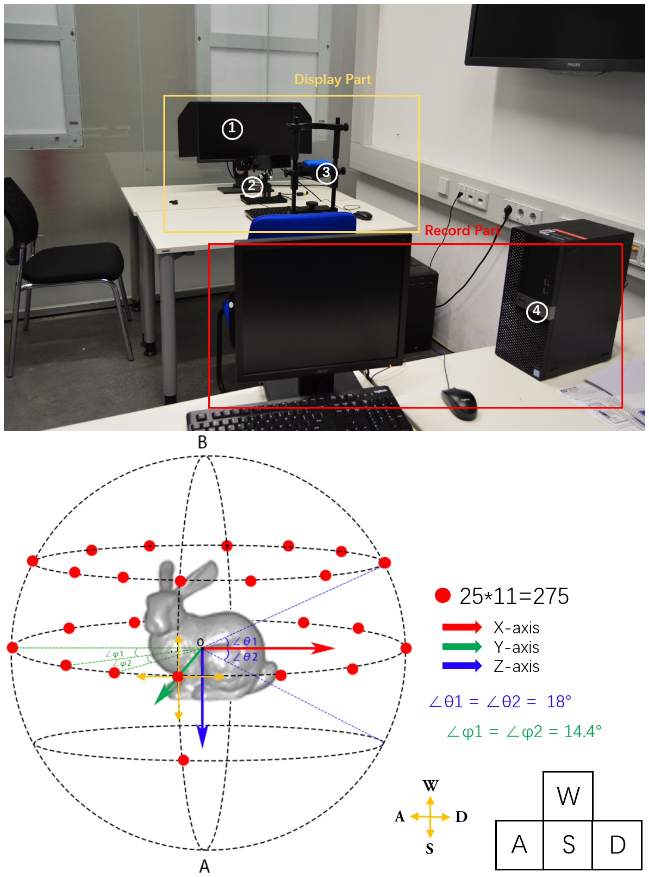

Figure 3: Top: Experiment setup. Numbers represent ① a monitor, ② eye tracker, ③ chin rest, and ④ host computer. Bottom: Interactive experiment design. The red dots represent some observation perspectives. The observation perspectives are restricted to a spherical surface with a radius of 300 mm.

Mount Eye Tracker running at 2,000 Hz and providing an accuracy of 0.5° after proper calibration. The distance from the display to the eye is set to 85 cm. Figure 3 depicts the display setup, eye tracker, chin rest, and host computer. We developed a web-based interactive experimental platform to observe the multi-view stimuli. Each participant observed twelve 3D objects, each for two minutes. The order of presentation for the objects was randomised. We ran a web browser in full-screen mode with Three.js[2] to render the objects, which was embedded in the WebLink recording software provided by the manufacturer[3].

Procedure. The participants used the keyboard to change the viewing direction to observe the 3D object from different directions. To maintain the depth information of the stimuli surface and the saliency distribution, the relative distance between the camera and the stimuli must remain fixed, regardless of the observation point's changes [Lang et al. 2012]. Additionally, the angle intervals of the observation points distributed around the Z-axis should be consistent [Wang et al. 2018]. To achieve this, we distribute the viewpoints on the surface of a sphere, with the centre of the sphere being the centre of the 3D object and the radius being 300 mm. As shown in Figure 3 (right), the sphere is divided into 11 slices along the X-axis, and the angle between every two slices is 18° along the Z-axis ($\angle\theta 1$ and $\angle\theta 2$). Furthermore, each slice has 25 evenly distributed observation points, with a difference of 14.4° between every two adjacent points ($\angle\varphi 1$ and $\angle\varphi 2$). The spherical coordinate system (r, θ, φ) is used to switch the camera's viewing direction, where φ refers to the angle change along the X-axis, and θ denotes the angle change along the Z-axis. The coordinate position of the camera can be expressed as:

$$(x, y, z) = 300\,mm \cdot (\cos\varphi \sin\theta, \sin\varphi \sin\theta, \cos\theta) \qquad (1)$$

The four keys 'W', 'S', 'A', and 'D' were used to switch the viewing direction, where 'W' was used to decrease θ, 'S' to increase θ, 'A' to decrease φ, and 'D' to increase φ. To avoid the possible bias caused by participants being inclined to look at the centre of the screen, the initial viewing direction was randomly set for each round. Each 3D object was observed for 2 minutes. The participants were asked to use the keyboard to switch the viewing direction to observe the region of interest in the 3D stimulus. After observation, participants were asked to choose five preferred viewing directions and prioritise them using the numeric keys 'NUM1' – 'NUM5' on the keypad. Participants can rest and move freely after observing each stimulus for up to 1 minute. Prior to presenting each new object to the participant, the eye tracker was properly calibrated.

3.2 Data Processing

Mapping Gaze from 2D to 3D. The raw gaze data contains coordinates, start and end timestamps, and duration. Records of key presses are loaded to align the view directions and gaze data. After each key press, the viewing direction remains the same until the next key press. With the initial position known, each key press corresponds to a unique viewing direction, i.e., each 'A' or 'D' key press corresponds to $\pm 14.4^\circ \angle\varphi$, and each 'W' or 'S' key press corresponds to $\pm 18^\circ \angle\theta$.

The viewing duration is calculated as the interval between successive key presses. We employ the ray-casting method to determine the 3D gaze mapped on the model's surface [Roth 1982]. The intersection point of the ray with the 3D object in the spatial coordinate system represents the actual three-dimensional position observed by the observer. Then, we compute the inverse model-view-projection operation based on the spatial location of the viewpoint to map the gaze points from the screen to the spatial coordinate system:

$$v_{model} = (P \cdot V \cdot M)^{-1} \cdot v_{clip}, \qquad (2)$$

where M, V, and P represent the object, view, and projection matrices, v_{clip} represents the gaze points in the screen coordinate system, and v_{model} represents the gaze points in the spatial coordinate system. See the Appendix for a detailed deduction of screen-spatial coordinate system transformation. Due to the lack of depth information, the transformed result is a ray originating from the current viewpoint. Therefore, we calculate the coordinates of the

[2] https://threejs.org/
[3] https://www.sr-research.com/weblink/

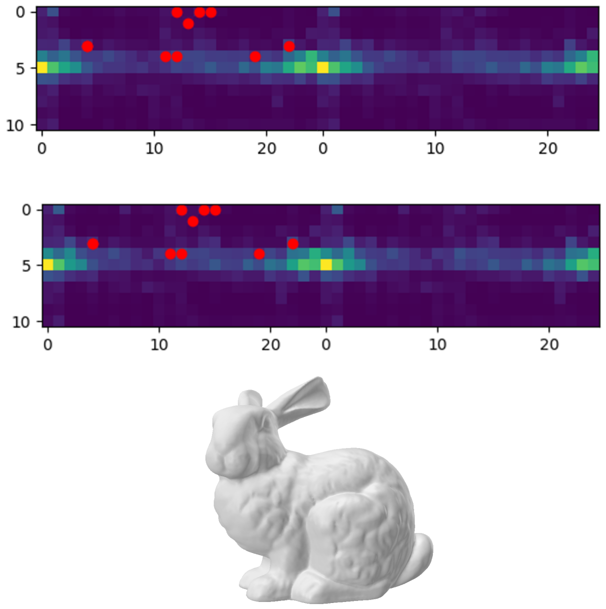

Figure 4: Number of fixations (top) and fixation duration (middle) in different viewing perspectives for *Bunny*, and the object under the most viewed perspective (bottom). Red dots depict the initial perspective of each participant.

intersection of this ray and the model surface, representing the 3D gaze data corresponding to the 2D gaze data under the given viewing perspective.

4 RESULTS AND DISCUSSION

4.1 Hypothesis

H1: It was previously observed that different viewing directions are not equally effective at revealing shapes, and a clear preference for certain views is expressed [Blanz et al. 1999]. To examine the consistency of human viewing behavior on the same stimulus, we formulated two hypotheses:

- H1.1: Most people preferred a viewing perspective when observing a 3D stimulus.
- H1.2: The initial viewing perspective does not affect the most preferred viewing pespective.

H2: Humans are more interested in facial features when viewing images [Bindemann et al. 2005]. We assume the existence of a face bias on 3D objects.

4.2 Findings

Viewing Preference. Figure 4 depicts the gaze distribution of *Bunny* from the aspect of number of fixations (top) and fixation duration (middle). The x-axis ranges between $[0, 25]$, representing $\angle\varphi$ in $[0, 360]$, and the y-axis ranges between $[0, 10]$, representing $\angle\theta$ in $[0, 180]$. Many fixations are distributed at $[0, 5]$ and other viewing directions near this position. While the observer may switch to one viewing direction several times, the fixation duration in this direction is stable. Thus, the number of fixations cannot be regarded as the sole criterion to indicate the preferred viewing perspective. Moreover, the fixation duration is concentrated under the viewing perspective $[0, 5]$. It strongly indicates the existence of the most preferred viewing perspective of the *Bunny* (see Figure 4), which appears in other 3D objects as well (see Appendix). We hereby confirm that the preferred viewing perspective of 3D objects among people exists (H1.1). Meanwhile, the experimental result shows that this observational preference did not change with the first impression. In the experiment, the initial viewing directions of the 10 observers were randomly generated. As shown in figure 4, all initial positions are marked by red dots. This means that preference is not affected by the first impression but rather a general observational behavior (H1.2). Thus, the preferred viewing direction is not influenced by the initial perception of the object but rather by a fundamental observational behavior.

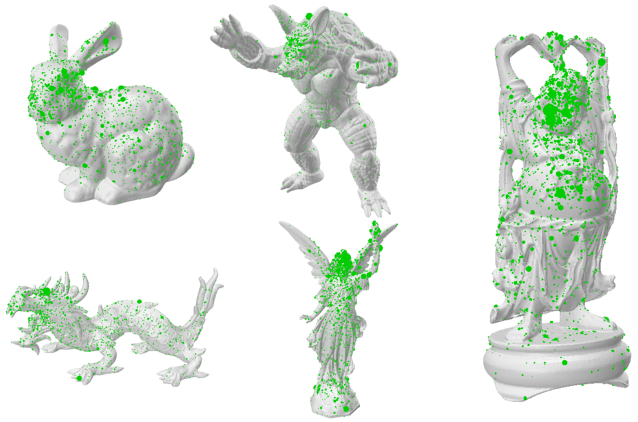

Figure 5: Five stimuli with facial features. Participants have a strong face bias on 3D objects.

Face Bias. We use fixation density [Wang et al. 2023] to study the face bias on 3D objects. The fixation density (FD) is calculated as the accumulated number of gaze fixations divided by the covering area of fixation targets. Five objects with facial features are selected (*Bunny, Happy_buddha, Lucy, Armadillo, Dragon*). As shown in table 1, r_{gaze} denotes the ratio of the number of gazes in the face region to the number of all gazes. r_{area} denotes the ratio of the face region's surface area to the object's total surface area. The fixation density on the face region can be represented by:

$$(Gaze_{total} \times r_{gaze})/(Area_{total} \times r_{area}), \tag{3}$$

where $Gaze_{total}/Area_{total}$ represents the average density of fixations on the object. $FD = r_{gaze}/r_{area}$ quantifies the fixation density on the face region for the whole object. *Lucy* and *Happy_buddha* have the highest *FD*, 7.05 and 5.33 respectively, while the *dragon* has the lowest fixation density of 1.63. A *FD* larger than 1 suggests the existence of face bias, while a higher *FD* indicates a stronger face bias. The *FD* of human faces (*Lucy* and *Happy_buddha*) is higher than any other objects, which confirms the hypothesis H2.

Table 1: *Lucy* and *Happy_buddha* have the highest mean fixation density, indicating a strong face bias during observation. The object with the highest fixation density of each participant is shown in bold. FD: Fixation Density.

name	Armadillo		Happy_buddha		Lucy		Bunny		Dragon	
r_{area}	9.48%		4.26%		3.07%		24.5%		31.4%	
	r_{gaze}	FD	r_{gaze}	FD	r_{gaze}	FD	r_{gaze}	FD	r_{gaze}	FD
P1	38.4%	4.05	41.3%	9.70	42.6%	**13.88**	50.6%	2.07	79.7%	2.54
P2	44.1%	4.65	27.2%	**6.38**	15.6%	5.07	45.2%	1.85	57.0%	1.82
P3	22.4%	2.36	19.8%	**4.65**	6.7%	2.18	31.5%	1.28	40.8%	1.30
P4	28.1%	2.97	9.6%	2.26	11.6%	**3.77**	50.2%	2.05	46.7%	1.49
P5	41.8%	4.41	28.4%	6.66	43.0%	**14.00**	50.0%	2.04	52.0%	1.65
P6	44.1%	**4.65**	17.0%	4.00	14.1%	4.60	39.0%	1.59	50.0%	1.59
P7	28.3%	2.98	20.7%	4.87	24.4%	**7.93**	33.7%	1.38	42.5%	1.35
P8	41.9%	4.42	24.8%	**5.82**	6.0%	1.94	36.9%	1.51	43.8%	1.40
P9	29.2%	3.08	14.2%	3.33	16.6%	**5.41**	42.1%	1.72	50.4%	1.60
P10	39.4%	4.15	23.9%	5.62	35.9%	**11.68**	46.4%	1.90	48.1%	1.53
mean FD	3.77 (0.7)		5.33 (4.25)		**7.05 (21.16)**		1.74 (0.08)		1.63 (0.13)	
mean Ranking	2.5 (0.5)		1.9 (0.54)		**1.6 (0.71)**		4.2 (0.18)		4.7 (0.23)	

4.3 Limitation

In the study of Wang et al. [2018], the models observed by the participants were 3D-printed models. Conversely, the study in our work utilizes computer-generated 3D models for participant observation. Further research is required to prove whether the rendered 3D objects on the screen can fully substitute 3D-printed objects for 3D saliency studies.

5 CONCLUSION

This paper proposes a novel design for collecting gaze data for viewing 3D objects from screens. Using this method, we collected Saliency3D, a 3D saliency dataset comprising 10 participants looking at 16 3D objects. Furthermore, we investigated the viewing preferences for 3D objects. For most models, the differences in the significant characteristics under different observation angles have a strong linear correlation with the observation angle difference. Moreover, we found that participants have a strong face bias on 3D objects. Our work shows the potential to collect 3D saliency datasets more cheaply and efficiently.

ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161. M. Bâce was funded by the Swiss National Science Foundation (SNSF) through a Postdoc.Mobility Fellowship (grant number 214434) while at the University of Stuttgart. A. Bulling was funded by the European Research Council (ERC) under grant agreement 801708.

The authors thank Yue Jiang for paper editing support and Luis A. Leiva for helpful comments on this paper.

REFERENCES

Roland J Baddeley and Benjamin W Tatler. 2006. High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision research* 46, 18 (2006), 2824–2833.

Markus Bindemann, A Mike Burton, Ignace TC Hooge, Rob Jenkins, and Edward HF De Haan. 2005. Faces retain attention. *Psychonomic bulletin & review* 12, 6 (2005), 1048–1053.

Volker Blanz, Michael J Tarr, and Heinrich H Bülthoff. 1999. What object attributes determine canonical views? *Perception* 28, 5 (1999), 575–599.

Alexandre Bruckert, Mona Abid, Matthieu Perreira Da Silva, and Patrick Le Callet. 2023. Could the bubbleview metaphor be used to infer visual attention on 3D graphical content?. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. 2018. Review of visual saliency detection with comprehensive information. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 10 (2018), 2941–2959.

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2016. A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 3488–3493.

CzernO. 2021. Easter Island dude. https://sketchfab.com/3d-models/easter-island-dude-156387085d1843b2bd9427de7de178d9.

Karthik Desingh, K Madhava Krishna, Deepu Rajan, and CV Jawahar. 2013. Depth really Matters: Improving Visual Salient Region Detection with Depth. In *BMVC*. 1–11.

Sonja Engmann, M Bernard, Thomas Sieren, Selim Onat, Peter König, and Wolfgang Einhäuser. 2009. Saliency on a natural scene background: Effects of color and luminance contrast add linearly. *Attention, Perception, & Psychophysics* 71, 6 (2009), 1337–1352.

Yuming Fang, Junle Wang, Manish Narwaria, Patrick Le Callet, and Weisi Lin. 2014. Saliency detection for stereoscopic images. *IEEE Transactions on Image Processing* 23, 6 (2014), 2625–2636.

Camilo Luciano Fosco, Anelise Newman, Patr Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, and Zoya Bylinskii. 2020. How Much Time Do You Have? Modeling Multi-Duration Saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 4472–4481.

John M Henderson and Andrew Hollingworth. 1999. High-level scene perception. *Annual review of psychology* 50, 1 (1999), 243–271.

Zhiming Hu. 2020. Gaze analysis and prediction in virtual reality. In *Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VR)*. 543–544.

Zhiming Hu, Andreas Bulling, Sheng Li, and Guoping Wang. 2021. FixationNet: Forecasting Eye Fixations in Task-Oriented Virtual Environments. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 27, 5 (2021), 2681–2690.

Zhiming Hu, Sheng Li, Congyi Zhang, Kangrui Yi, Guoping Wang, and Dinesh Manocha. 2020. DGaze: Cnn-based gaze prediction in dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 26, 5 (2020), 1902–1911.

Zhiming Hu, Congyi Zhang, Sheng Li, Guoping Wang, and Dinesh Manocha. 2019. SGaze: a data-driven eye-head coordination model for realtime gaze prediction. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 25, 5 (2019), 2002–2010.

Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1254–1259.

Timothée Jost, Nabil Ouerhani, Roman Von Wartburg, René Müri, and Heinz Hügli. 2005. Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding* 100, 1-2 (2005), 107–123.

Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. 2017. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 5 (2017), 1–40.

Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. 2017. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 4789–4798.

Congyan Lang, Tam V Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan. 2012. Depth matters: Influence of depth cues on visual saliency. In *European Conference on Computer Vision*. Springer, 101–115.

Stephen RH Langton, Anna S Law, A Mike Burton, and Stefan R Schweinberger. 2008. Attention capture by faces. *Cognition* 107, 1 (2008), 330–342.

Guillaume Lavoué, Frédéric Cordier, Hyewon Seo, and Mohamed-Chaker Larabi. 2018. Visual attention for rendered 3D shapes. *Computer Graphics Forum* 37, 2 (2018), 191–203.

Nian Liu and Junwei Han. 2018. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *IEEE Transactions on Image Processing* 27, 7 (2018), 3264–3274.

Päivi Majaranta and Andreas Bulling. 2014. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*. Springer, 39–65.

Susana Martinez-Conde, Stephen L Macknik, and David H Hubel. 2004. The role of fixational eye movements in visual perception. *Nature reviews neuroscience* 5, 3 (2004), 229–240.

Sudarshan Ramenahalli and Ernst Niebur. 2013. Computing 3D saliency from a 2D image. In *2013 47th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 1–5.

Pamela Reinagel and Anthony M Zador. 1999. Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems* 10, 4 (1999), 341.

Scott D Roth. 1982. Ray casting for modeling solids. *Computer graphics and image processing* 18, 2 (1982), 109–144.

Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2018. Saliency in VR: How Do People Explore Virtual Environments? *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 1633–1642.

Florian Strohm, Mihai Bâce, and Andreas Bulling. 2023. Usable and Fast Interactive Mental Face Reconstruction. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 1–15.

Greg Turk, Mark Levoy, Brian Curless, and Andrew Gardner. 2003. the Stanford Model. http://graphics.stanford.edu/data/3Dscanrep/.

Junle Wang, Matthieu Perreira Da Silva, Patrick Le Callet, and Vincent Ricordel. 2013. Computational model of stereoscopic 3D visual saliency. *IEEE Transactions on Image Processing* 22, 6 (2013), 2151–2165.

Xi Wang, Sebastian Koch, Kenneth Holmqvist, and Marc Alexa. 2018. Tracking the gaze on objects in 3D: How do people really look at the bunny? *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–18.

Xi Wang, David Lindlbauer, Christian Lessig, Marianne Maertens, and Marc Alexa. 2016. Measuring the visual salience of 3d printed objects. *IEEE Computer Graphics and Applications* 36, 4 (2016), 46–55.

Yao Wang, Mihai Bâce, and Andreas Bulling. 2023. Scanpath Prediction on Information Visualisations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2023), 1–15.

Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. 2014. Predicting human gaze beyond pixels. *Journal of vision* 14, 1 (2014), 28–28.

Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. 2008. SUN: A Bayesian framework for saliency using natural statistics. *Journal of vision* 8, 7 (2008), 32–32.

A APPENDIX

A.1 Screen-Spatial Coordinate Transformation

The Model-View-Projection (MVP) Matrix transforms 3D coordinates into 2D coordinates on the screen. The Model matrix transforms the 3D coordinate point from the local coordinate system to the world coordinate system:

$$M = T \cdot R \cdot S \tag{4}$$

$$v_{\text{world}} = M \cdot v_{\text{model}} \tag{5}$$

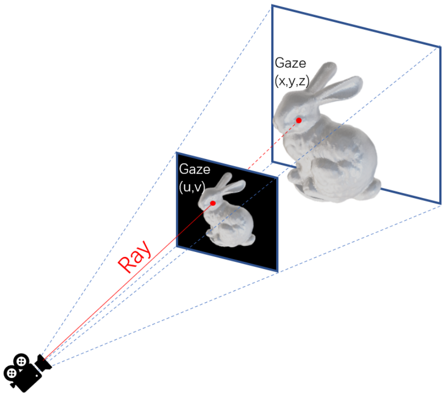

Figure 6: Ray casting. Starting from the camera as the origin point, a ray passes through a 2D coordinate on the screen. The location at which this ray intersects the 3D stimulus determines the spatial coordinate.

where T,R, and S represent the translation, rotation, and scaling matrices. Due to the use of perspective projection instead of orthogonal projection, the spatial coordinates projected on the screen need projection transformation. Figure 6 depicts the visible range of the camera as the frustum, and its volume can be determined by the field of view, the near plane, and the far plane. The projection matrix realises the transformation between 3D coordinates and 2D coordinates under the corresponding viewing direction:

$$P = \begin{bmatrix} \frac{2n}{r-l} & 0 & \frac{r+l}{r-l} & 0 \\ 0 & \frac{2n}{t-b} & \frac{t+b}{t-b} & 0 \\ 0 & 0 & -\frac{f+n}{f-n} & -\frac{2fn}{f-n} \\ 0 & 0 & -1 & 0 \end{bmatrix}. \tag{6}$$

v_{clip} represents the 2D gaze coordinates corresponding to the screen resolution:

$$v_{clip} = P \cdot V \cdot M \cdot v_{\text{model}} \, . \tag{7}$$

where M, V, and P represent the object, view, and projection matrices, and v_{model} represents the 3D coordinates in the space coordinate system. Moreover, x,y in v_{clip} are corresponded to the resolution of the screen, and z is normalised as 1.

We use the inverse transformation of perspective projection to transform 2D gaze into 3D gaze coordinates:

$$v_{model} = (P \cdot V \cdot M)^{-1} \cdot v_{clip}. \tag{8}$$

A.2 Figures

This appendix contains the number of fixations and fixation duration in different viewing directions for the object *Armadillo* (Figure 7), *Casting* (Figure 8), *Dragon* (Figure 9), *Face* (Figure 10), *Game_controller* (Figure 11), *Hand* (Figure 12), *Happy_budda* (Figure 13), *Lucy* (Figure 14), *Planck* (Figure 15), *Rockarm* (Figure 16), *Sofa* (Figure 17), *Space_shuttle* (Figure 18), *Spanner* (Figure 19), *Vase* (Figure 20), and *Watchtower* (Figure 21).

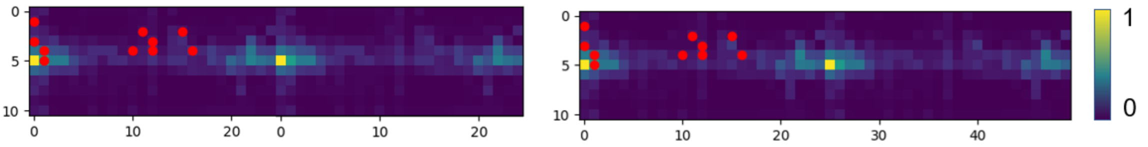

Figure 7: [Armadillo] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing perspective of each participant.

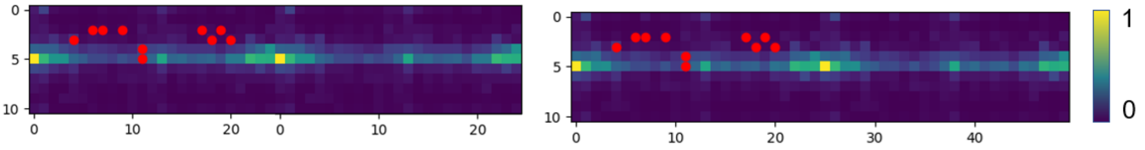

Figure 8: [casting] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing direction of each participant.

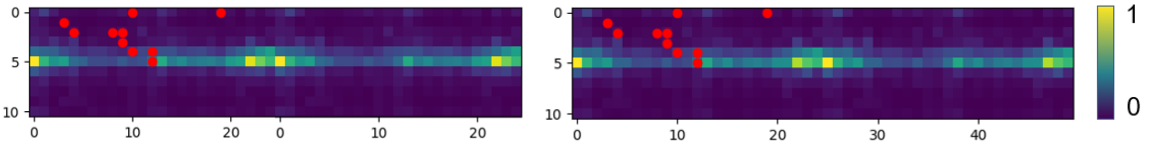

Figure 9: [dragon] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing direction for each participant.

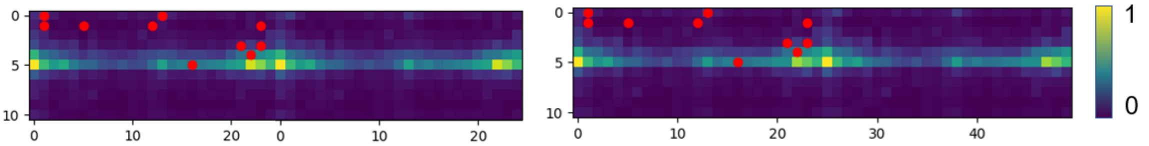

Figure 10: [Face] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing direction of each participant.

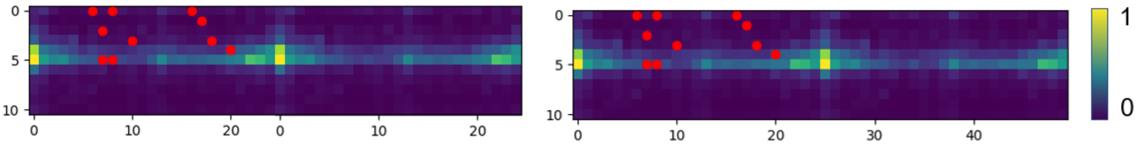

Figure 11: [Game_controller] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing direction of each participant.

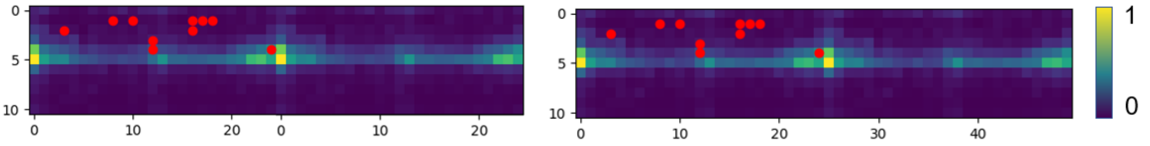

Figure 12: [Hand] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing direction of each participant.

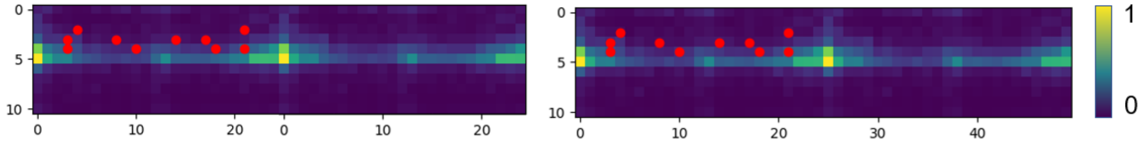

Figure 13: [Happy_budda] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing direction of each participant.

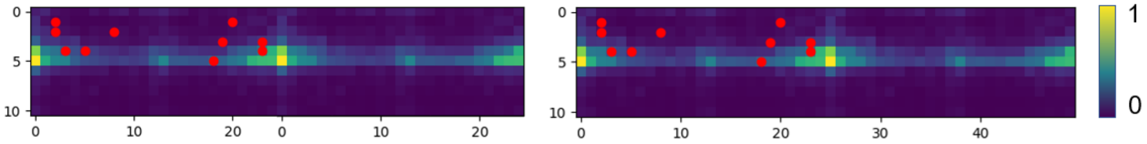

Figure 14: [Lucy] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing direction of each participant.

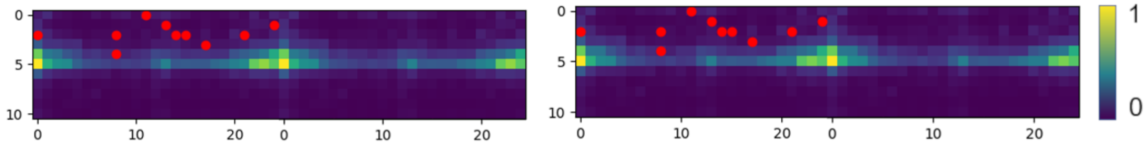

Figure 15: [Planck] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing direction of each participant.

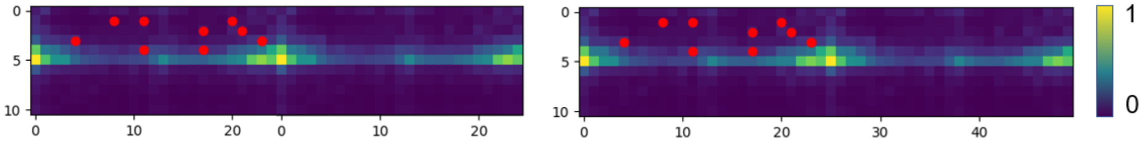

Figure 16: [Rockarm] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing direction of each participant.

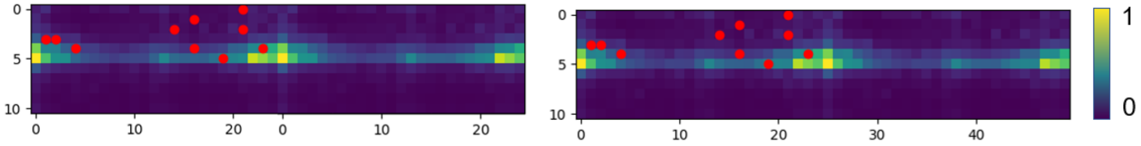

Figure 17: [Sofa] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing direction of each participant.

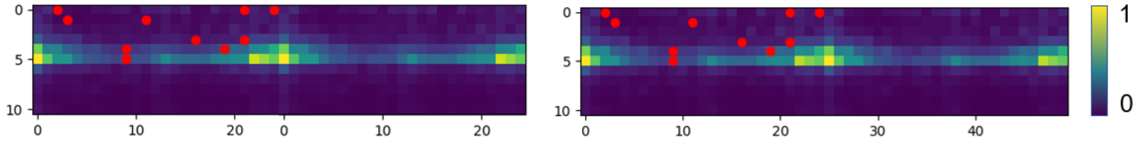

Figure 18: [Space_shuttle] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing perspective of each participant.

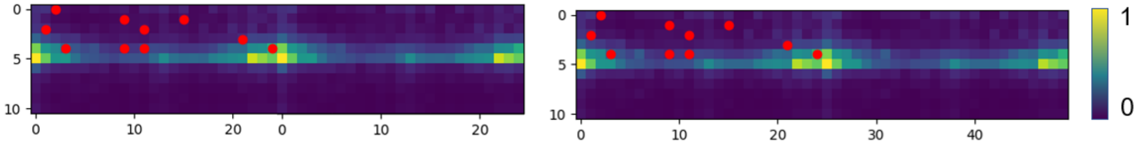

Figure 19: [Spanner] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing perspective of each participant.

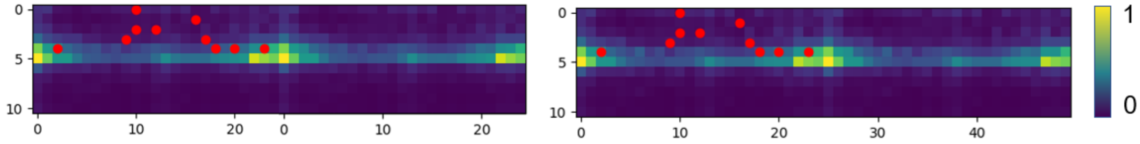

Figure 20: [Vase] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing perspective of each participant.

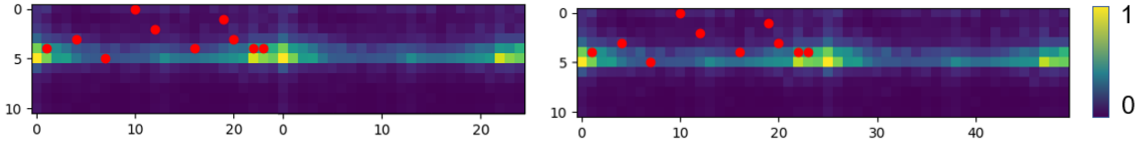

Figure 21: [Watchtower] Left: number of fixations across viewing perspectives, Right: fixation duration across viewing perspectives. The red points depict the initial viewing perspective of each participant.