Visual Attention Predictability

Andrey Arguedas Espinoza
Computer Science Engineering School
Costa Rica Institute of Technology
San Jose, Costa Rica
and12rx12@gmail.com

***Abstract*—In this paper, we do an extensive research on the topic of visual attention for the Virtual Reality world. Our main objective is to understand how we can understand and model the user's visual attention behaviors, how can we store it to generate data and how we can analyze it and interact with this data for a better understanding of the human behaviour in the Virtual Reality environments. We do a summary of the state of the art on projects that have implemented computational techniques such as saliency maps, eye tracking and regression models used on VR projects to understand and process visual attention in order to improve VR experiences.**

***Index Terms*—Virtual Reality, Gaze, Gaze Prediction, Eye Tracking, Visual Attention, Regression Model, Saliency Maps, FOVE.**

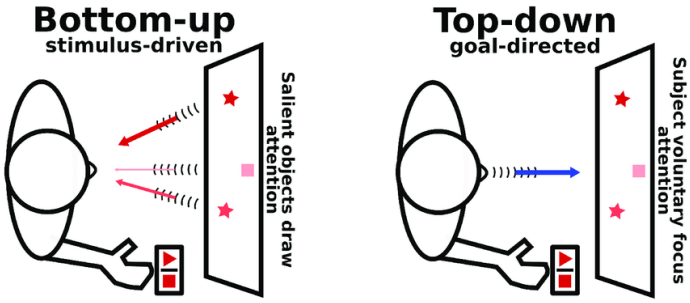

Fig. 1. Top Down Attention vs Bottom Up Attention. Taken from [2]

I. Introduction

In the Virtual Reality (VR) world understanding the user's visual attention has become more important with the past of time. Different projects have shown that with the data retrieved from the user's experience multiple things can be done in order to improve, update or analyze the VR experiences.

In this paper we do an extensive research on the visual attention topic for the VR world, the need to understand it correctly, how it is model computationally and how we can transform it into data that we can use to predict it's future development. Also we do a review on recent projects that have implemented visual attention or gaze prediction to create unique solutions on VR environments.

II. Research

According to [1] a rich stream of visual data (between 10^8-10^9 bits) enters our eyes every second. Processing this data in real time is a difficult task, without the help of clever mechanisms to reduce the amount of erroneous visual data [2]. Current projects that needs to work with visual attention such as object recognition, scene interpretation or visual predictability rely on visual data that has been transformed in such a way as to be tractable [2].

For Virtual Reality environments we can define the term **attention** as a general concept covering all factors that influence visual selection mechanisms, whether they are scene-driven bottom-up (BU) or expectation-driven top-down (TD) [2]. Visual attention we can define it as a mechanism to shape what we see, and allow for concurrent selection of relevant information and inhibition of other information [3].

Meanwhile, **gaze** is defined as a coordinated motion of the eyes and head, and it has often been used as a proxy for attention. For instance, an user has to interact with surrounding objects and control the gaze to perform a task while moving in the environment. In this sense, gaze control engages vision, action, and attention simultaneously to perform a sensorimotor coordination necessary for the required task. [2]

Saliency we can define it as a method to intuitively characterize some parts of a scene which could be objects or regions that appear to an observer to stand out to their relative neighbor parts [2], for example in a 3D VR scene maybe an user can drive it's attention to an interactable object in a wall and ignore the rest of the wall.

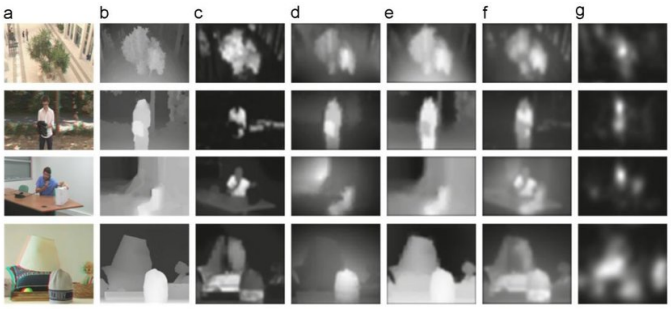

Fig. 2. Example of the state of the art on 3D images processing to create saliency maps compatible with 3D assets: (a) input 3D image (b) disparity map (c) 2D image saliency map (d) depth saliency map (e) visual comfort based saliency map (f) 3D saliency map (g) human eye fixation density map. Taken from [4]

The research of the visual attention is very extensive and it goes to different areas of science as neuroscience and human vision, on computer science, in 2013 researchers did an state of the art publication on numerous computational models devised for predicting the allocation of attention based on the distribu-

tion of visual features in a scene [2]. They stated that on visual attention the main concern is how to model it computationally, specially how, when, and why we select behaviorally relevant image regions, to address this problem several definitions and computational perspectives are available. Usually in 3D and VR environments we want to understand visual attention as the eye tracking data that attracted to the most informative scene regions, or those regions that maximize reward regarding a task.

Visual attention models fall into two main categories bottom-up which are mainly based on characteristics on stimulus-driven of visual scene [5], whereas top-down are determined by cognitive phenomena like knowledge, expectations, reward, and current goals. **Bottom-Up** attention is fast, involuntary, and most likely feed-forward. A prototypical example of bottom-up attention is looking at a scene with only one horizontal bar among several vertical bars where attention is immediately drawn to the horizontal bar [6]. Many models fall into this category, but they can only explain a small fraction of eye movements since the majority of fixations are driven by task [7]. **Top-Down** attention is slow, task driven and voluntary [8], eye movements depend on the current task, for example the experiment which participants were asked to watch the same scene (a room with a family and an unexpected visitor entering the room) under different questions such as "estimate the material circumstances of the family..." "what are the ages of the people? [9].

On computational history visual attention has been modeled with different methods such as:

- **Cognitive Models:** Normally they used feature channels color, intensity, and orientation. An image is sub sampled into a Gaussian pyramid and each pyramid level is decomposed into channels for Red (R), Green (G), Blue (B), Yellow (Y), Intensity (I), and local orientations (O). After this the channels are normalized and scale to create a saliency map.
- **Bayesian Models:** They are used usually when we need to predict something on an image or scene, combines sensory evidence with prior constraints. In these models, prior knowledge (scene context) and sensory information (target features) are probabilistic combined according to Bayes' rule (Detect an object of interest).
- **Decision Theoretic Models:** They pretend to imitate biological behavior of visual attention, theory states that perceptual systems evolve to produce decisions about the states of the surrounding environment that are optimal in a decision theoretic sense. The end goal is that visual attention should be driven by optimality with respect to the end task .
- **Information Theoretic Models:** They deal with selecting the most informative parts of a scene and discarding the rest. Usually they are modeled using data analysis techniques.
- **Graphical Models:** These models use graphs to model the structure of visual attention, and treat eye movements as a time series. Their output is a saliency map that also takes in account temporary constraints.
- **Pattern Classification Models:** These are base on machine learning techniques, with data recorded from eye fixations ore previous salient maps they create learning models Usually attention works as the "stimuli-saliency" function to select, reweight, and integrate the input visual stimuli. These models may not be purely bottom-up since they use features that guide top-down attention (faces, text, regions).

Thanks to devices like the **FOVE** VR headset is easier to obtain this data in 3D VR environments [10] thanks to the in-built eye tracker. It is important to understand that the allocation of visual attention for Head Mounted Devices (HMD) is measured in terms of gaze duration on a given scene element (interval of viewing an element without shifting one's gaze) [11].

Fig. 3. FOVE Head Mounted Display with eye-tracking capabilities

On 2018 a really important work from [12] was made, this work was focused on understanding how people explore immersive virtual environments, they stated that previous work was focus on modeling saliency in 2D viewing conditions and VR is very different from these conditions because viewing behavior is governed by stereoscopic vision and by the complex interaction of head orientation, gaze, and other kinematic constraints. Therefore, they captured and analyzed gaze and head orientation data of 169 users exploring stereoscopic, static omni-directional panoramas, for a total of 1980 head and gaze trajectories for three different viewing conditions, with all data obtained they found several important insights, such as the existence of a particular fixation bias, which nowadays is used to adapt existing saliency predictors to immersive VR conditions. To identify fixations, they transformed the normalized gaze tracker coordinates to latitude and longitude in the 360 degree panorama. This was necessary to detect users fixating on panorama features while turning their head. They used thresholding based on dispersion and duration (150 milliseconds) of the fixations. Their dataset was used by many scientists during the next years to work on predicatbility models for gaze prediction on VR projects.

On 2018, David et al [13] was able to compare four methods that can be used to generate 360 degrees saliency maps from eye tracking data obtained by a VR device. The studies of

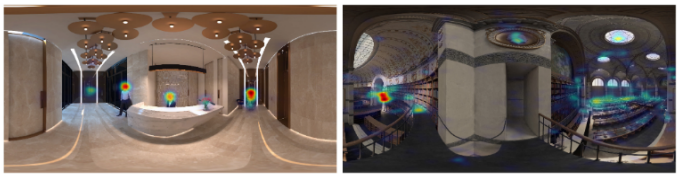

Fig. 4. Saliency Map in a VR scene. Taken from [12]

saliency maps is well established for the 2D world, where it is used to predict gaze [14], image or video compression [15] and others. Meanwhile in the VR world is different, the user is surrounded by a photorealistic virtual scene, due to the spherical nature of 360 degrees content, saliency map generation methods are different. Saliency maps aggregate data from multiple observers into a representative map of human attention. Typically, saliency maps are generated by summing fixations from each observer into a discrete fixation map. This fixation map is convolved with a 2D Gaussian kernel to produce a continuous map highlighting salient regions instead of individual pixels. We will implement their approach of a subsampled version of the modified Gaussian that proved to be sufficient for generating accurate 360 degrees saliency maps fairly quickly, even at image resolutions exceeding 8K [13].

Authors from [16] created a new visual attention user dataset for omnidirectional video (ODV) by investigating behavior of viewers when consuming ODV content. They found that the limitations of the ODV technologies are strongly related to the massive volume of video data that needs to be stored and rendered compared to traditional video. Since HMDs use only a fraction of an ODV at a time, namely viewport, ODV can be optimized by predicting where the viewers' visual attention is concentrated at a given point in time. For this, they used saliency maps, which predict viewer's eye fixations for given content. To understand the salient regions of ODV viewed in HMDs, saliency maps can be estimated either by collecting eye fixations during subjective tests or by using visual attention models. In their research they created a new dataset which include viewport trajectories (VT) and visual attention maps from 17 participants while watching uncompressed ODV, this can be used to obtain visual attention maps without the need for eye tracking devices. For modeling visual attention they were only interested in the user's fixations. They consider a valid fixation, if the Vieport Center Trajectory (VCT) remains almost stable in a certain location for at least 200 milliseconds. This requires clustering the VCT in order to remove influence from minor irrelevant movements and to reduce sensitivity to noise.

In VR, the temporal characteristics of visual attention have also been studied, Sitzmann et al. revealed that observers in virtual reality behave in two different modes: "attention" and "reorientation" [12]. Attention mode refers to the condition when observers focus their attention on some regions, while re-orientation mode is the status when observers shift their attention. Taking this in consideration the "stablished" period of time for VR projects was defined to be 200 ms and it has been used as a guide to track "valid" visual attention across VR projects.

Now that we know that is possible to track visual attention via different methods, therefore we want to know if other studies have use this type of data to predict future eye tracking or optimize different aspects of a 3D scene thanks to the obtained data. In this research path we have found the following contributions.

Gaze prediction is a hot topic since 2018 on the VR world, most of the previous existing gaze prediction methods were based on Bottom-Up models [17] [18], which focus on low-level image features like intensity, color, and orientation, or Top-Down models [19] [20], which take high-level features such as specific tasks and context into consideration. In addition, with recent advances in deep learning, many deep learning-based gaze prediction methods have also been proposed [21]. In the area of virtual reality until 2018, there was little work on gaze prediction.

On 2018 Sitszman et al [12] were able to adapt existing saliency prediction models for 2D images to VR using insights from their data analysis, such as the equator bias they found, and also did an adaptation to analyze head movement in order to model gaze on a HMD without eye-tracking capabilities. With their model they were able to implement a solution in multiple VR problems such as automatic cuts for VR videos, paranorama thumbnails, video for VR marketing and saliency aware image compression for VR. Similarly in 2019 Hu et al [22] were able to predict gaze in real time using an eye-head coordination model with HMD devices that did not have eye-tracking capabilities. Their model (SGaze) is computed by generating a large dataset that corresponds to different users navigating in virtual worlds with different lighting conditions and is a great solution to use when using HMD devices without eye-tracking functionalities.

With the incorporation of eye-tracking HMD works regarding visual attention started to focus on prediction of gaze in the VR world, Hu et al [23] presented the concept of temporal continuity (consistency of users' on-screen gaze position sequences) in visual attention, how to evaluate it and analyze it in free viewing versus task oriented conditions. With the data obtained on their experiment they found that temporal continuity can be applied to future gaze prediction and if it is good, users' current gaze positions can be directly utilized to predict their gaze positions in the future. The finds on their work is crucial in current projects that implements gaze-contingent rendering, advertisement placement, and content-based recommendation in VR environments.

On 2019 the project from Alghofaili et al [24] used an eye-tracking VR headset to use visual attention data to arrange artwork in a virtual museum, posting banners for virtual events or placing advertisements by analyzing visual attention patterns of the users. Their propose consist in a data-driven optimization approach for automatically analyzing visual attention and placing visual elements in 3D virtual environments. With the collected eye-tracking data they train

a regression model for predicting gaze duration, then use the predicted gaze duration output from the regressors to optimize the placement of visual elements with respect to certain visual attention.

On 2020 Hu et al, [23] presented great results with a gaze prediction model for VR under free viewing conditions, using the dataset previously generated on [22] which contains over 4,000,000 gaze positions, with this they created function to evaluate temporal continuity (TC), if TC is good, users' current gaze positions can be directly utilized to predict their gaze positions in the future.They used the angular distance as the evaluation metric between the ground truth and the predicted gaze position, for example, the angle between the user's ground truth line of sight and the predicted line of sight. The smaller the angular distance, the smaller the prediction error and the better the performance.

Fig. 5. Results of gaze prediction from Hu et al [23]. The green cross refers to users'current gaze positions; blue indicates users'gaze positions in future 100ms; yellow denotes gaze positions in future 400ms; red represents gaze positions in future 700ms.

Two years later, an improvement to the state of the art on gaze prediction was achieved by [25], they proposed a neural network and a Long Short Term Memory based model utilizing past HMD pose and gaze data to predict future gaze locations using heavy saliency computation. Their solution considers data from the exhaustive OpenNEEDs dataset which contains 6 Degrees of Freedom (6DoF) data captured in VR experiences with subjects given the freedom to explore the VR scene and/or to engage in tasks. Their solution outperformed Hu's [23] work by predicting gaze in real-time for sub 150ms VR use-cases. Their model is really complex but incredible strong, it is able to predict every frame, that means in VR context is executed 90-120 times per second.

On 2022, Takahashi [26] were able to accommodate 2D items dynamically for more proactive visual exploration based on ongoing search context by analyzing the distribution of eye gaze through an eye-tracking device, in order to infer how the most attractive items lead to the finally wanted ones. To guide dynamic item arrangement, they employed an eye-tracker that takes spatiotemporal eye gaze distribution as input. Their solution was based in compute the spatiotemporal distribution of visual attention by convolving each gaze point in the sequence with a Gaussian weighting kernel. For a visualization technique they used the spatial distribution of visual attention as a heatmap, in which the color changes from blue to green to red as the degree of attention increases. One important aspect of their work four our project is that they increases the priority values of invisible items so that they could replace visible ones that are not of interest with this invisible ones.

On 2023 Liang et al [27], published a paper where they show how to optimally place virtual products in a VR store, their approach consists in take a virtual store and a list of virtual products and then run an optimization algorithm to find the best place for each of the products based on the scene constraints and producing a rational layout. The optimizer consists of a total cost function and an optimizing algorithm, the total cost function is devised to evaluate each generated product placement and consists of an exposure and spatial cost. The product exposure term is based on a Random Forest Regressor trained to predict shoppers' gaze duration on products, and guides the optimizer to improve product exposure in the virtual store. The spatial term encodes the design priors for arranging products reasonably to improve shopping experiences by avoiding crowdedness, ensuring even distribution and keeping visual balance. For each product the algorithm will return a set of coordinates **(x,y,z)** to represent the product's location and a θ value representing the angle orientation of the product. This project is very interesting since they used visual attention theory and saliency maps to fully improve a VR experience.

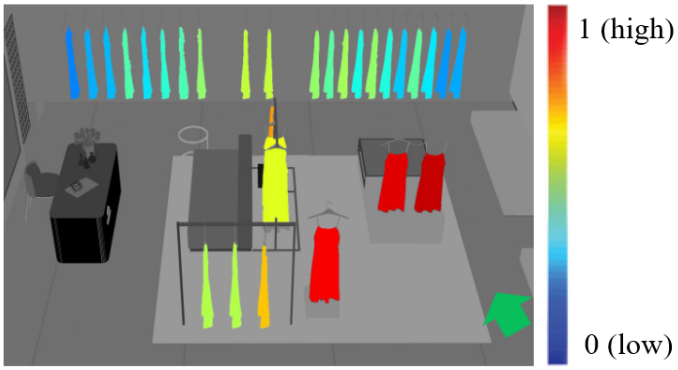

Fig. 6. Heatmap obtained after processing a saliency map on project from Liang et al [27].

Fig. 7. Final result after predicting optimal placement for virtual products on a VR store, from Liang et al [27].

III. Conclusion

We have learned how important is to understand Visual Attention and how humans interact visually with the environments, we can also see how a concept that has been studied for years across multiple disciplines can be constantly updated and used for different top projects in technology. Virtual Reality is a topic that is constantly changing and their is space to improve and create solutions for it, classic concepts from computer science can be adjusted and rethink for VR environments.

We found that there are two main types of Visual Attention and this changes based on the context of the environment of the tasks, this topic also has been really important in computer vision and there are different ways to model it such as Cognitive Models, Bayesian Models, Decision Theoretic Models, Information Theoretic Models, Graphical Models and Pattern Classification Models. Unfortunately their was not much history on how to track visual attention on VR environments, thankfully now we have special HMDs to retrieve real time eye-tracking. It was also important to see how different projects started to model the eye-tracking without this special devices by using regression models and understanding the user's movement. For more recent implementations we have found that with this data we can create innovative solutions and improve the VR environments and experiences for all users.

It was a great experience to learn about an specific topic and research about how has been implemented in recent VR projects, we also think that there is space to keep researching how to understand Visual Attention using techniques as context aware visual recognition systems or procedural scenes generation for VR to track how visual attention behaves on random and dynamically changing environments, we think that generating a dataset of gaze data on this type of environments could lead us to obtain better prediction models.

References

[1] K. Koch, J. McLean, R. Segev, M. A. Freed, M. J. Berry, V. Balasubramanian, and P. Sterling, "How Much the Eye Tells the Brain," *Current biology : CB*, vol. 16, pp. 1428–1434, July 2006.

[2] A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 185–207, Jan. 2013.

[3] K. K. Evans, T. S. Horowitz, P. Howe, R. Pedersini, E. Reijnen, Y. Pinto, Y. Kuzmova, and J. M. Wolfe, "Visual attention," *WIREs Cognitive Science*, vol. 2, no. 5, pp. 503–514, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.127.

[4] A. Rogalska and P. Napieralski, "A problem of assessing visual comfort in virtual reality applications," pp. 43–53, Dec. 2016.

[5] H.-C. Nothdurft, "Salience from feature contrast: additivity across dimensions," *Vision Research*, vol. 40, pp. 1183–1201, June 2000.

[6] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, Jan. 1980.

[7] J. Henderson and A. Hollingworth, "High-Level Scene Perception," *Annual review of psychology*, vol. 50, pp. 243–71, Feb. 1999.

[8] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews. Neuroscience*, vol. 2, pp. 194–203, Mar. 2001.

[9] B. Tatler, N. Wade, H. Kwan, J. Findlay, and B. Velichkovsky, "Yarbus, Eye Movements, and Vision," *i-Perception*, vol. 1, pp. 7–27, July 2010.

[10] A. Duchowski, *Eye Tracking Methodology: Theory and Practice*. Jan. 2007. Journal Abbreviation: Eye Tracking Methodology: Theory and Practice Publication Title: Eye Tracking Methodology: Theory and Practice.

[11] S. P. Liversedge, K. B. Paterson, and M. J. Pickering, "Chapter 3 - Eye Movements and Measures of Reading Time," in *Eye Guidance in Reading and Scene Perception* (G. Underwood, ed.), pp. 55–75, Amsterdam: Elsevier Science Ltd, Jan. 1998.

[12] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in VR: How Do People Explore Virtual Environments?," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, pp. 1633–1642, Apr. 2018.

[13] B. David-John, P. Raiturkar, O. Le Meur, and E. Jain, "A Benchmark of Four Methods for Generating 360° Saliency Maps from Eye Tracking Data," pp. 136–139, Dec. 2018.

[14] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," vol. 20, Nov. 2007.

[15] C. Guo and L. Zhang, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression," *Image Processing, IEEE Transactions on*, vol. 19, pp. 185–198, Feb. 2010.

[16] C. Ozcinar and A. Smolic, "Visual Attention in Omnidirectional Video for Virtual Reality Applications," May 2018.

[17] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *CVPR 2011*, pp. 409–416, June 2011. ISSN: 1063-6919.

[18] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254–1259, Nov. 1998. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[19] A. Borji, D. N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 470–477, June 2012. ISSN: 1063-6919.

[20] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," vol. 19, pp. 545–552, Jan. 2006.

[21] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting Human Eye Fixations via an LSTM-Based Saliency Attentive Model," *IEEE Transactions on Image Processing*, vol. 27, pp. 5142–5154, Oct. 2018. Conference Name: IEEE Transactions on Image Processing.

[22] Z. Hu, C. Zhang, S. Li, G. Wang, and D. Manocha, "SGaze: A Data-Driven Eye-Head Coordination Model for Realtime Gaze Prediction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, pp. 2002–2010, May 2019.

[23] Z. Hu, S. Li, and M. Gai, "Temporal continuity of visual attention for future gaze prediction in immersive virtual reality," *Virtual Reality & Intelligent Hardware*, vol. 2, pp. 142–152, Apr. 2020.

[24] R. Alghofaili, M. S. Solah, H. Huang, Y. Sawahata, M. Pomplun, and L.-F. Yu, "Optimizing Visual Element Placement via Visual Attention Analysis," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, (Osaka, Japan), pp. 464–473, IEEE, Mar. 2019.

[25] G. K. Illahi, M. Siekkinen, T. Kämäräinen, and A. Ylä-Jääski, "Real-time gaze prediction in virtual reality," in *Proceedings of the 14th International Workshop on Immersive Mixed and Virtual Environment Systems*, (Athlone Ireland), pp. 12–18, ACM, June 2022.

[26] S. Takahashi, A. Uchita, K. Watanabe, and M. Arikawa, "Gaze-driven placement of items for proactive visual exploration," *Journal of Visualization*, vol. 25, no. 3, pp. 613–633, 2022.

[27] W. Liang, L. Wang, X. Yu, C. Li, R. Alghofaili, Y. Lang, and L.-F. Yu, "Optimizing Product Placement for Virtual Stores," in *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, (Shanghai, China), pp. 336–346, IEEE, Mar. 2023.