The Effects of Visual Cues on Classification Perception in Immersive Scatterplots

Fumeng Yang, James Tompkin, Lane Harrison, and David H. Laidlaw

Abstract—Immersive visualization in virtual reality (VR) allows us to exploit visual cues for perception in 3D space, yet few existing studies have measured the effects of visual cues. Across a desktop monitor and a head-mounted display (HMD), we assessed scatterplot designs which vary their use of visual cues—motion, shading, perspective (graphical projection), and dimensionality—on visualizations of two different sets of data. We conducted a user study with a summary task in which 32 participants estimated the classification accuracy of an artificial neural network from the scatterplots. We find that no variable alone explains all the variance in estimation error. Visual motion cues generally reduce participants' estimation error; besides this motion, using other cues may increase participants' estimation error. The results also indicate the complexity of combining different visual cues and applying them to different datum. Using an HMD generally results in slightly larger estimation errors than using a desktop monitor; however, using particular combinations of visual cues, a VR HMD performs as well as a desktop monitor. Using an HMD, adding visual motion cues, providing a third dimension, or showing a more complicated dataset leads to longer response times. In summary, by studying participants as they interpret the output from a complicated machine learning model, we advance our understanding of the effects of visual cues for immersive analytics.

Index Terms—virtual reality, cluster perception, information visualization, immersive analytics, dimension reduction, classification

1 Introduction

Technologies such as virtual and augmented reality (VR/AR) allow immersive approaches to data visualization and decision-making [1]. While suitable for displaying inherently spatial 3D data (e.g., digital elevation models or isosurfaces), more abstract data raise questions about designing, presenting, and interacting with information visualization in 3D space.

One key challenge here is assessing the effects of *visual cues*, which are global or local properties that help people perceptually prioritize objects and regions [2], [3]. There are two fundamental classes of visual cues: *primary cues* providing physiological percepts (e.g., stereopsis and accommodation), and *pictorial cues* first observed by graphic artists and used to depict 3D depth in 2D pictures (e.g., occlusion, perspective, texture, and shading) [2]. Visual cues are crucial to human perception and cognition affecting 2D visual comparison [4], [5], depth perception [6], [7], [8], 3D length [9], spatial relationships [2], [10], [11], spatial judgments [10], [12], and shape understanding [13], [14], [15].

Prior work on information visualization in VR/AR has not focused on the effects of visual cues [16], [17], [18]; for instance, how pictorial cues affect depth and spatial perception. Previous studies have presented stimuli that differ markedly between what is shown on a desktop monitor and in VR/AR [18], [19], [20]. Some studies used different colors and shading across the desktop monitor and VR for dimension-reduction results [19], [20], and other studies presented 2D scatterplot matrices and 3D scatterplots within the same experiment [18].

- *Fumeng Yang, James Tompkin, and David H. Laidlaw are with the Department of Computer Science, Brown University, Providence, RI. E-mails: {fy, james_tompkin, david_laidlaw}@brown.edu.*
- *Lane Harrison is with the Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA. Email: lharrison@wpi.edu.*

To further complicate matters, prior studies of visual cues from other domains may not directly apply to information visualization. Tasks used in these studies include navigating a simulated 3D world or comparing a handful of virtual objects [21], which are categorically different from visualization tasks performed on thousands of visual elements representing different data points [10]. Low-level perceptual tasks (e.g., 3D distance judgment) also differ from high-level tasks based on *ensemble coding* of visualizations (e.g., cluster perception) [22]. As such, understanding the effects of visual cues in immersive visualization via a robust analytical framework is necessary to facilitate a broader and more appropriate usage of VR/AR techniques.

In this paper, we explicitly study the effects of visual cues on people's task performance in immersive visualization. Specifically, we examined four visual cues: motion, perspective (graphic projection), shading, and dimensionality. We chose these cues for two reasons: they added minimal visual clutter, and they showed stronger effects on people's perception and cognition than many other cues reported in prior studies [1], [7], [12]. We compared these cues across an immersive VR environment using a head-mounted display (HMD) and a traditional, non-VR environment using a desktop monitor. We also used two sets of data to examine the effects of different data properties on participants' task performance [23].

We used scatterplots as the central visualization. Scatterplots are a common visualization supporting both low-level object-centric tasks [23] and high-level visual aggregation [22], [23]. Scatterplots are used widely to show different data features such as correlation [24], anomalies [25], [26], clusters [27], [28], [29], and dimension-reduction results [30], [31], [32], [33]. They can be viewed on a desktop monitor [34], [35], [36], [37], in VR [38], [39], [40], [41], [42], [43], and more recently in AR [16], [44]. Scatterplots show the spatial

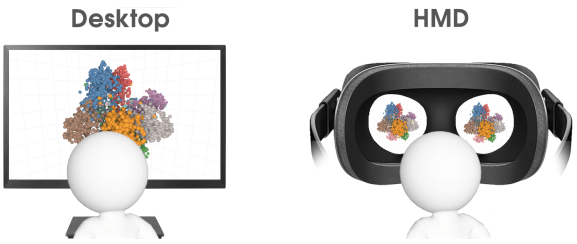

Figure 1. **Main task**. In each trial, participants estimated classification performance of the neural network from a scatterplot, which was then removed and replaced with an interface for inputting a number. Participants finished a series of trials where the scatterplots were rendered using different combinations of visual cues.

distribution of a set of data points [23] and are suitable for studying visual cues based on spatial coordinates [45], [46], [47]. Previous studies compared scatterplots between a desktop monitor and in VR/AR and examined interaction techniques [16], [43], [48] and different encoding channels [49] in VR; however, these studies have not focused on quantifying the impacts of visual cues, particularly pictorial cues applying to all data points, on a task of making sense of complex data at an abstract, conceptual level [50].

We aimed to move beyond low-level and abstract tasks and be more reflective of real-world analytical tasks. We drew inspiration from recent successes that used scatterplots to analyze the outputs and activations of a neural network [51] to identify misclassified instances, training effects, and hidden structures [52], [53]. Tangling these activities, we designed a task where participants assessed a neural network's classification performance from a scatterplot showing the last hidden layer's the outputs (see Fig. 3), and they responsed by providing their estimation of the classification accuracy. This task requires participants to interpret a visualization of class groupings, compare it to their mental model of a particular accuracy level (e.g., using low-level perceptual features), and collapse all their observations and understanding to into one number as their response. We surmise that this task is more representative of a real-world analytical task than other low-level perception tasks (e.g., reading a value).

We present three main research contributions:

- Measurement of the quantitative effects of four commonly-used visual cues across a desktop monitor and in a VR HMD on two sets of data. These cues generally had small effects on estimation error; device and visual motion had large effects on response time.
- Measurement of the quantitative interaction effects between cues, supporting *cue-integration* theory. That is, people combine multiple cues to improve their estimate of a property [54]. Visual cues interacted with each other in complicated ways, especially for estimation error.
- A performance-ranking list for all the tested scatterplots from different combinations of cues and data.

This work contributes to the emerging field of immersive visualization, connecting the huge gamut of choices in visual-cue design to their possible impact on both low- and high-level facets of visualization interpretation. To support these contributions, our experimental materials and system, the neural networks used, datasets, participants' data, and analysis scripts are posted at https://doi.org/10.17605/OSF.IO/PKUVZ.

2 Task Design

The task of assessing neural network classification performance was realized by participants estimating classification accuracy from a scatterplot of the last hidden layer's output (Fig. 1). The goal of this study is to investigate the role of visual cues in visualization task across modalities.

2.1 Task Justification

To facilitate our goals, we considered four types of validity [55], [56]—*internal validity*, *construct validity*, *external validity*, and *ecological validity*. We describe how the task supports each type of validity below.

Internal Validity How well the evidence supports the cause-and-effect relationship between cause and effect is known as internal validity. This task appropriately reflects how well people can visually inspect classification results, and therefore helps establish such a relationship between visual cues and classification perception. Specifically, a visualization of the last hidden layer's output of a neural network represents the model's classification performance (Fig. 3). To measure how well participants perceive classification performance, we sought a quantitative metric. Available metrics include accuracy, precision, and loss [57]. We used accuracy for simplicity; it is the percentage of correctly classified instances. As such, we asked participants to estimate classification accuracy. This design allows us to control a variety of visual properties by generating multiple datasets along the same computational pipeline (see Section 3.3). Other approaches, such as using a number of different datasets may introduce too much variance for a controlled experiment.

Construct Validity To establish that our experiment measured what it purports to be measuring—a property known as construct validity—we carefully considered how participants would interpret the task. In addition to the experimental instructions, we first conducted a pilot study, where we confirmed that participants were able to understand the task and make responses near the true accuracy given scatterplots of the last hidden layers' outputs (Section 4.6). We also required participants to have at least passing knowledge in machine learning and visualization (Section 4.6). Last, we included both training and practice for participants to calibrate their understanding of classification accuracy with scatterplot appearance (Section 4.3).

External Validity The generalizability to different scenarios is often known as external validity, and this task supports external validity because it is a kind of cluster perception in multi-class scatterplots. Assessing classification performance using a visual representation could be considered a function of cluster detection with uncertainty, in which factors such as data properties (e.g., outliers and cluster distance), cognitive (e.g., expertise), and primarily perceptual components (e.g., visual cues) could shift participants' estimates. The task may also generalize to other non-supervised processes and further applies to accessing performance of other mathematical, statistical, or computational models.

Ecological Validity Generalizability to real-life settings is often known as ecological validity and is a specific type of external validity. This task support ecological validity because it is an example taken from a set of similar tasks

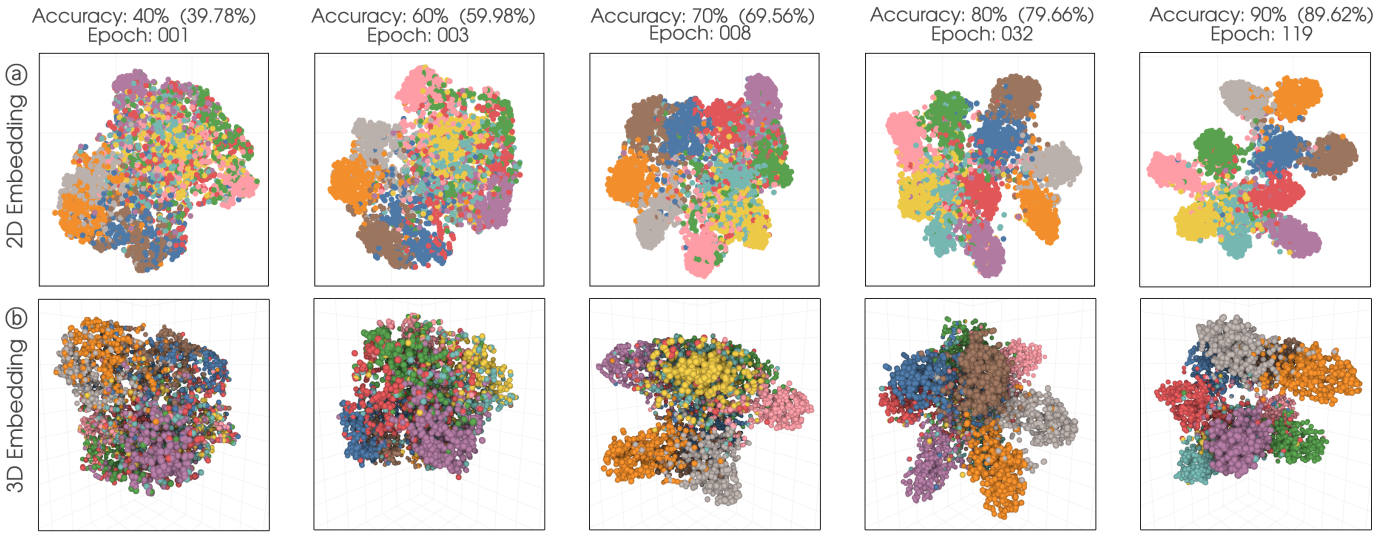

Figure 2. **Examples of last hidden layer outputs and classification accuracy**. As we trained the neural network with more epochs, the visual properties of the last hidden layer outputs change systematically with the classification accuracy. ⓐ The top row shows examples of Orthographic ∘ Flat Shading ∘ Image Data ∘ 2D Embedding. ⓑ The bottom row shows examples of Perspective ∘ Ambient Occlusion ∘ Image Data ∘ 3D Embedding. Each column shows 2D or 3D embedding (dimension-reduction) results for the same high-dimensional dataset, which is a last hidden layer output.

that users plausibly undertake in real-world visual analytics. While estimating classification accuracy is somewhat artificial, seeing last hidden layer outputs facilitates the understanding of the neural network's performance and assists in model diagnostics [51], [52], [53]. Previous studies have demonstrated that the same visualizations helped identify the misclassified instances and understanded relationship between different classes (clusters) [51], [52].

2.2 Task Limitations

While we intended to imitate a realistic task combining multiple facets of perceiving clusters, our attempts may have introduced indirect effects on our experimental framework. These inherent indirections may span from data generation to participants' interpretation; and approximation errors in each computational step may propagate along the rendering pipeline. The last hidden layer output is an approximation of the model's internal representation. And the original high-dimensional last hidden layer output was further approximated via a dimension-reduction procedure, and participants could have been partially estimating the performance of the dimension-reduction technique. They may have different interpretations of the same stimuli, which may or may not correspond to the same accuracy number in their responses. Furthermore, although our student participants have at least passing knowledge in both machine learning and visualization, they are still considered non-expert users and may have to rely more on visual properties of the scatterplots to estimate accuracy. More experienced users would display different behaviors and may rely more on their prior knowledge. While it is difficult, if not impossible, to completely separate and gauge all the indirect effects, the task still measures participants' perception of classification performance under uncertainty.

2.3 Task Construction

To generate the scatterplots, we first trained a neural network on the training set and saved the intermediate neural network after each training epoch (Fig. 3). For each intermediate neural network, we used the test set as the input and calculated its last hidden layer output. We then used this output as the input for dimension reduction and further rendered a scatterplot. We showed the scatterplot to participants as the stimulus and asked them to estimate the classification accuracy of the neural network on the test set. They input a percentage as their answer, and we compared this answer to the ground truth accuracy as a measure of the effectiveness of the scatterplot.

3 Visual Cues

3.1 Background and Cue Selection

We considered testing across different devices and selected four visual cues based on the literature, where the four visual cues showed strong effects on people's performance in perceptual and cognitive tasks. We also included different sets of data to understand the generalizability of the results. In total, we manifested six variables within our experiment: (1) *device*, (2) *visual motion*, (3) *graphical projection*, (4) *shading*, (5) *dimensionality*, and (6) *data model*. This section describes our motivation in light of the literature. In addition, We

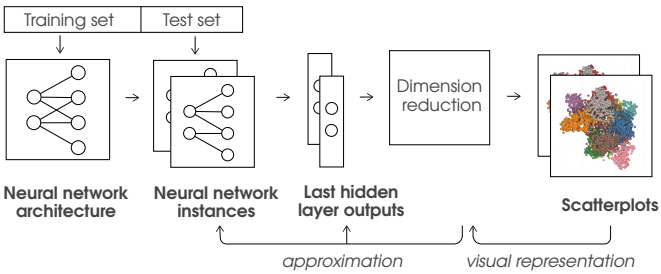

Figure 3. **Task construction**. We collected a set of intermediate neural networks of different classification performance from the training process; we calculated their last hidden layers' outputs, performed dimension-reduction, and created corresponding scatterplots as the stimuli. Colors represent ground truth classes.

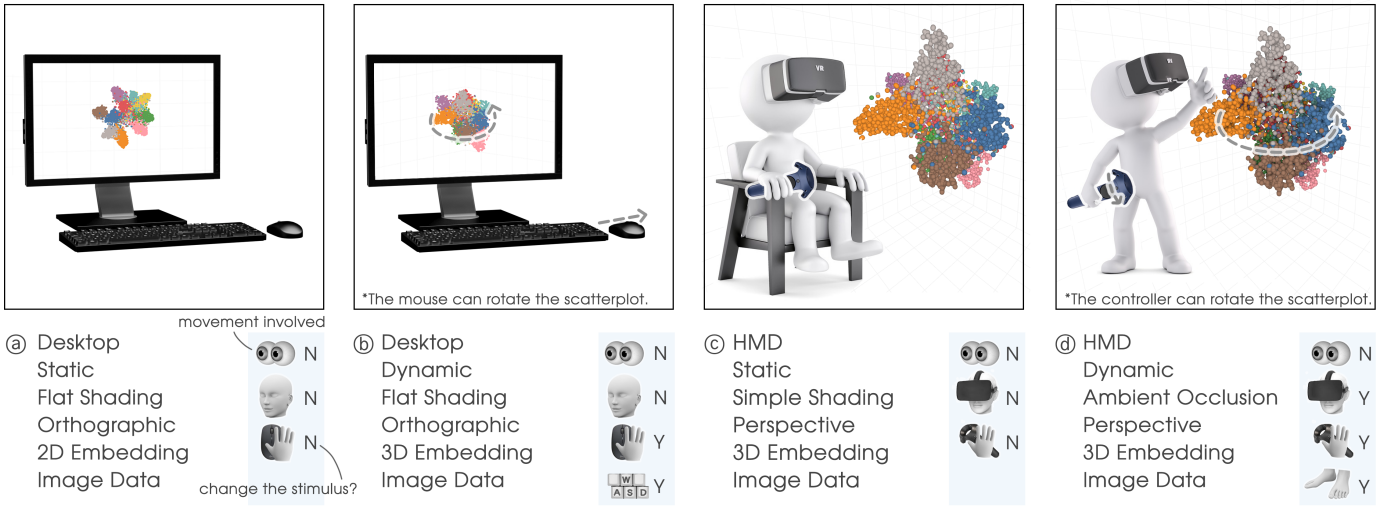

Figure 4. ***Device* , *motion*, and movements involved**. We had both Desktop (ⓐⓑ) and HMD (ⓒⓓ) conditions. Participants saw static scatterplots in Static conditions (ⓐⓒ); and in Dynamic, they interacted and rotated the scatterplots via a mouse and a keyboard (Desktop, ⓑ) or by walking and using the VR controller (HMD, ⓓ). Participants could move their pupils, head, hands, feet, and they could also move virtually using a keyboard or a mouse; between Static and Dynamic, these movements may or may not trigger a change of the stimuli on the display.

describe all cues collected from the literature as a table in Appendix C, and recommend readers to refer to that table.

Device Several surveys on immersive visualization, stereoscopic displays, and human performance [58], [59], [60], [61] agreed that a stereoscopic display may improve participants' performance over a desktop (e.g., [62], [63], [64]), especially for more difficult tasks [65]. Others reported mixed [49], [66], [67], [68], [69] or negative results [70], and the effects were subject to individual differences [71], [72], [73] and tasks [58], [59], [60]. In particular, the studies on graph visualization showed that a VR environment may have positive [74], [75] or neutral effects [17], [76], [77] on task performance and completion time compared to a desktop monitor. Other studies compared scatterplots [19], [20], [78] across a desktop monitor and an HMD, reporting mixed results.

We varied *device* and included a desktop monitor to establish a set of baseline performance for comparison (see Fig. 4). Certain visual cues only present in an immersive environment (e.g., immersiveness and presence [79], [80]); and therefore we collectively considered them as contextual cues for the immersive environment.

Visual Motion Prior studies found that motion alone shows stronger effects than stereoscopy alone [81], [82], while others found that motion is less effective than stereoscopy with head-tracking [63], [83]; more studies reported that motion and stereoscopy may have the same effectiveness [84]. In addition to improving depth discrimination, motion allows people to change the camera position (their vantage point) or their perspective to gain more information about the visualization and the underlying data; motion may also help people make sense of dimension reduction results that orient differently in the low-dimensional space. Motion creates dynamic, interactive visualizations, which are less familiar than static, fixed 2D visualizations to participants. These different facets motivated our comparison of different motion levels on different devices. While motion as a sensory cue could be detected beyond human vision system, like the proprioceptive and vestibular systems, for our goal of facilitating visual analytics, we primarily considered those motion cues related to vision that could cause a change in the visualization on the display.

Shading As a pictorial cue, shading affects people's perception of shape, depth, and spatial relationships [6] in a 3D space. Therefore, improving shading has became a general theme in the fields of particle visualization (e.g., [45], [46], [47], [85]) and computer graphics for the purposes of showing more details and enhancing user experience [86]. A previous study found that illumination improved completion time for tasks like graph path tracing [75]. Other studies reported that visual realism might hurt the performance in an immersive environment because the realistic details may be distracting [87] or cause anxiety [88]. Information visualization often shows abstract data [89] and commonly uses solid, plain colors. However, if data is in 3D space, varying shading may affect depth perception [90], enhance presence, and improve task performance. As such, we compared different levels of shading.

Graphical Projection Previous studies also show the impact of different graphical projection methods [12], [91]. Orthographic projection preserves size information in depth, using lines orthogonal to the projection plane. It is widely used in computer-aided design (CAD), 3D modeling software (e.g., Autodesk 3ds Max), and desktop-based visualizations of 3D data [28] for more precise presentation. Perspective projection alters an object's depicted size with its distance from the center of the projection, often used in VR to generate stereoscopy and immersiveness to improve spatial judgment [2] and distance perception [12], [91]. Orthographic projection is less common for VR, and its effects on immersive analytics is currently unknown. We therefore varied graphical projection methods.

Dimensionality Reducing a high-dimensional dataset to a 3D space is necessary for our purpose of exploring immersive visualization, and a data projection to 3D is often more precise than a data projection to 2D (e.g., the Kullback-Leibler divergence [92] is smaller). However, visualization in a 3D space often causes perception discrepancies (e.g., [51], [52], [53]). To better understand the effects of visual cues

and establish a fair baseline, we used the same dimension reduction procedure to generate both 2D and 3D data projections and rendered scatterplots for them (see Section 3.3 "Dimension Reduction").

Data Model To understand the generalizability of the results, we considered the effects of different data properties. The literature reports that the number of data points affects estimating numerosity [93], [94]; the complexity of a graph visualization (e.g., the number of nodes) affects path tracing performance [74], [95]; the complexity of a scene affects visual search in an immersive virtual environment [88]; the shape of clusters affects cluster perception [27]. Here data model refers to a training set and a neural network as a pair to generate a set of hidden layer outputs. One training set along with a specific neural network architecture generated a series of trained neural networks of different classification accuracies (Fig. 3). The term data model encompasses the differences in data distribution, complexity, cluster shape, and other properties (e.g., Figs. 2 and 7).

3.2 Cue Manipulation

Device We used two devices: a desktop with a monitor, and the same desktop with an HMD (HTC Vive Pro), denoted as Desktop and HMD, respectively.

Visual Motion Our experiment had two levels of motion related to vision: the first level always used static stimuli with a fixed view position, denoted by Static Stimuli (or Static); and the second level allowed participants changing their perspective to update the stimuli via to use movement and rotation, denoted by Dynamic Stimuli (or Dynamic). While motion as a depth cue can be generated from *motion parallax*, *motion perspective*, or *kinetic effect* [96], both a desktop monitor and an HMD can easily provide all three via interaction techniques. We collectively called them *visual motion* cues, and the difference between Static and Dynamic conditions is if these *visual motion* cues could cause a change in the stimulus on the display.

In Static conditions, participants saw static images rendered from the same fixed camera position in both Desktop and HMD conditions. Participants sat in a chair when using the HMD (Fig. 4c); this was to imply "not moving" and avoid further issues if motion sickness or nausea occurs. In all Dynamic conditions, participants were able to change their camera position using a keyboard and a mouse (Desktop, Fig. 4b) or by head tracking and walking in the room (HMD, Fig. 4d). Participants were able to rotate a scatterplot around its center by dragging a mouse (Desktop, Fig. 4b) or using a VR controller (HMD, Fig. 4d). Each of these movements could change the stimulus shown on the display.

Allowing motion in a plausible setup for both devices could activate other human sensory systems beyond vision; for example, walking and moving one's head may activate the proprioceptive and vestibular systems. In our experiment, these proprioceptive and vestibular information did not change the stimuli; therefore they did not directly affect how participants *saw* the stimuli. In other words, they are *non-visual motion cues* which are not directly related to the task centered around the visualization, but they might be utilized by participants to help judge distance or depth. It might be impossible to eliminate these intrinsic properties of an HMD nor approximating them precisely on a desktop; however, we could still quantify the plausible effects of these properties if we measure each combination of *device* and *visual motion*, and these differences are captured in the interaction effects between the two variables.

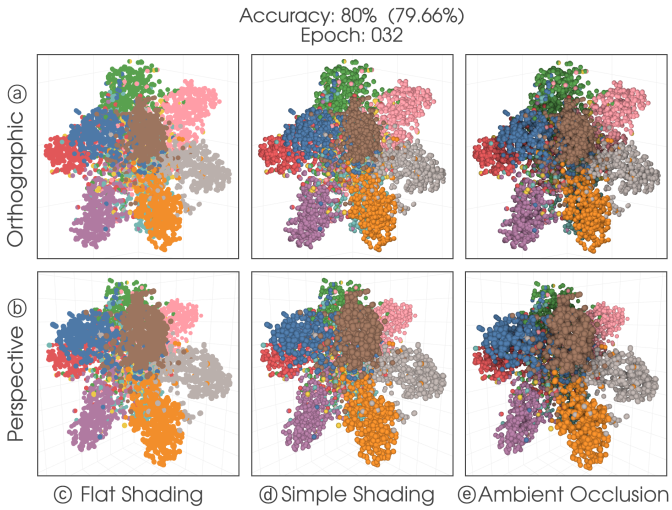

Figure 5. ***Shading* and *graphic projection***. ⓐ The top row shows examples of Orthographic projection for three different shading levels: Flat Shading, Simple Shading, and Ambient Occlusion. ⓑ The bottom row shows examples of Perspective projection. All six scatterplots here show the same underlying dataset from the set of Image Data, generated from the same last hidden layer output and dimension reduction procedure.

Shading Our experiment had three levels of *shading*: the first level used solid colors, denoted as Flat Shading (Fig. 5c); the second level utilized the commonly-used Phong lighting model, denoted as Simple Shading (Fig. 5d); and the third level used the Phong lighting model with ambient occlusion, a more advanced rendering technique [21], denoted as Ambient Occlusion or A.O. (Fig. 5e).

All Flat Shading conditions used only solid, plain colors. All Simple Shading conditions used ambient and diffuse terms in the Phong shading model. All Ambient Occlusion conditions used *screen-space ambient occlusion* (SSAO) [47] and normal reconstruction from the depth buffer. Many techniques have been developed to compute ambient occlusion, especially for scatterplots (or particle visualization) [47], but screen-space ambient occlusion is known for its efficiency and acceptable results, making it suitable for a VR HMD [97] requiring 90 frames per second for both eyes [98].

Graphic Projection Our experiment explored both orthographic and perspective projections, denoted as Orthographic and Perspective, respectively. The design and implementation are illustrated in Fig. 6. The implementation of perspective projection was straightforward. The implementation of orthographic was followed up by perspective projections to preserve binocular disparity in VR. As a result, participants saw a Static or Dynamic plane showing orthographic projection, and there was no stereoscopy on this plane.

Dimensionality We tested both 2D and 3D datasets. Because we used t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimension reduction, we denoted them as 2D Embedding and 3D Embedding, respectively. The 2D Embedding results are semantically related to the 3D Embedding results (Figs. 2a-b, 7a-b), sharing similar visual

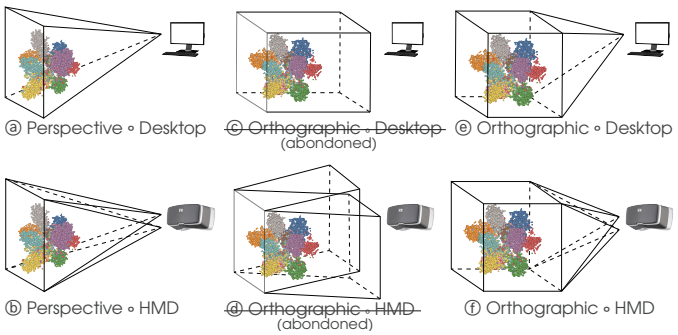

Figure 6. **Perspective and orthographic projections**. Perspective projection is straightforward in Desktop (ⓐ) and HMD (ⓑ) conditions. However, trivially replacing perspective with orthographic would cause incorrect perception in HMD conditions (ⓓ). Therefore, we implemented a single orthographic projection followed by the corresponding perspective projection(s) for each device (ⓔⓕ).

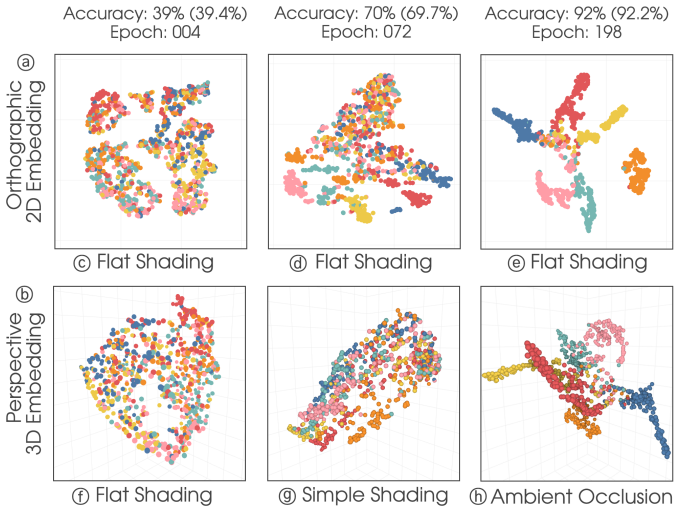

Figure 7. **Examples of Text Data**. ⓐ The top row shows examples of 2D Embedding across different accuracy levels. ⓑ The bottom row shows examples of 3D Embedding of the same hidden layer outputs across different accuracy and shading levels.

properties such as cluster shape and local structure. In our experiment, 2D Embedding was only used with Static ◦ Orthographic ◦ Flat Shading, which looked very similar to commonly seen information visualizations using simple and flat colors (Figs. 2a and 7a). This comparison across 2D and 3D embedding was similar to the study by Sedlmair et al. [28], and it improved compatibility between the two devices, compared to the approach of using scatterplot matrices as a baseline by Kraus et al. [18]. A 2D dataset does not have the third dimension for computing depth. Although it was not impossible to artificially add shading to a 2D dataset (e.g., setting z to 0), this caused serious z-buffer fighting or strongly hinted the drawing order as a third, non-existent dimension, which made it difficult to distinguish 2D Embedding from 3D Embedding. Thus, we ruled out these combinations of visual cues and dimensionality. Similarly, we also matched the dimensionality of the reference frames.

Data Model We selected two data models to capture the variability in data and visual properties of the resulting scatterplots: (1) Text Data, generated from the process of training the text dataset bAbI [99] on a memory neural network (MemNN) [100] and (2) Image Data, generated from training the image dataset CIFAR-10 [101] on a residual neural network [102]. These two data models vary in the number of data points, the number of classes, the shape of resulting clusters, and the neural network architecture. Furthermore, both models yielded a wide range of classification accuracies that are suitable for experimental purposes. Finally, the two sets were not used widely in education, such that the potential participants had not seen them before the experiment. Other commonly-seen data models like MNIST [103] or Fashion-MNIST [104] trained on a convolutional neural network (e.g., CNN) were used extensively in education, and the resulting range of accuracy was very small (Fig. 8a).

For Text Data, we used the training set with 10,000 samples (called "single supporting fact") and 200 epochs; we used the test set with 1,000 samples (6 classes) as the input to the trained neural networks. For Image Data, we used the training set with 50,000 images and 150 epochs. We randomly split the test set with 10,000 images into two folds with 5,000 images (10 classes) for each as the input of the trained neural networks. We present the statistics for collected classification accuracies in Figs. 8b and c, respectively.

3.3 Control and Extraneous Variables

Other differences between a desktop monitor and an immersive environment may also affect task performance. We considered and controlled a set of influential factors; they were *control variables*. A couple of variables introduced by hardware differences were impossible to eliminate; these we refer to as *extraneous variables*. Here we primarily discuss three of the control variables: *dimension reduction*, *visual content*, and *interactions*. We believe these three variables were most important for our experiment and would be more interesting to readers. As remarked in Section 3.1, we summarize all manipulated, control, and extraneous variables in Appendix C.

Dimension Reduction We used t-SNE to reduce the dimensionality of the last layers' outputs. t-SNE is prevalent in several research communities to show patterns in a high-dimensional dataset while preserving cluster information and local structure. The results of 2D and 3D t-SNE on the same data are considered the *optimal* approximations in their low-dimensional spaces (i.e., 2D or 3D space) constrained by the same distance function. We also manually checked all the datasets to ensure that the layout of a 2D dataset can be empirically considered the flattened of the corresponding 3D dataset; these datasets and images are available in supplementary materials.

We generated all the datasets using the same parameter settings (perplexity $= 30$, learning rate $= 150$, steps $= 600$, and fixed random seeds). These values ensured convergence,

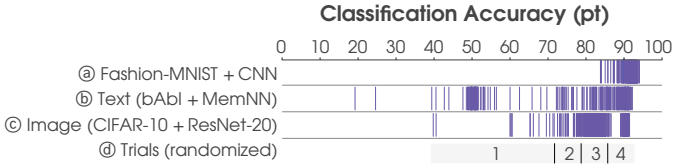

Figure 8. **Classification accuracy ranges of different data models**. ⓐ The first row shows that other common datasets and neural network architectures generate only very high classification accuracies. ⓑ-ⓒ The second and third rows show that the classification accuracy ranges of the two sets of trained neural networks, and ⓓ the last row shows our sampling process to select four trials for each condition.

which was manually checked for all the resulting datasets.

We had evaluated other dimension reduction techniques, but t-SNE was the best fit. We considered Uniform Manifold Approximation and Projection (UMAP) [105], known to preserve more global structure than t-SNE. However, UMAP requires more free parameters, and it was difficult to find a set of parameters that guaranteed reasonable results for all our datasets. UMAP also arranged data points of the same class very close to each other, causing cluttering and reducing legibility of the resulting scatterplots. Last, commonly-used Principal Component Analysis (PCA) resulted in nearly indistinguishable clusters for both data models (Text Data and Image Data).

Visual Content Our implementation intended to keep the consistency of visual content across different experimental conditions. All the conditions shared the same building and compiling processes and the same parameters for Vertex Array Objects (VAOs), Vertex Buffer Objects (VBOs), vertex and fragment shaders, as well as OpenGL context. We also used the same 3D hand mesh to indicate the mouse position on the screen and the spatial position of the controller in VR. We only varied the rendering process (e.g., *shading*), the target device (the screen or the VR compositor), and the camera position (e.g., fixed or changing) in different conditions.

Interactions We also carefully controlled *interaction* techniques, which are important factors for understanding space and depth in an immersive environment [106]. For example, prior studies show that a high interaction fidelity [16] may improve people's performance in tasks like gaming [107]. Interaction techniques support visual motion cues (remarked in Section 3.2 "Visual motion"). Specifically, rotation interaction provides *motion parallax*; head movement, physical, or virtual navigation provides *motion perspective*; each of these provides *kinetic depth perception*. Therefore, any use of motion or interaction triggers more than one cue that could be detected by human vision system. We aligned the interaction fidelity between Desktop and HMD conditions in the following three aspects.

- **Metaphor** In all conditions, the interaction metaphor was same: participants moved the mouse forward and backward in Desktop (Fig. 4b) or the controller up and down in HMD (Fig. 4d); both were on a 2D plane.
- **Rotation** In all Dynamic conditions, besides the input fidelity, the rotation operations between the two devices were the same. We implemented the Arcball technique [108] and calculated rotations based on movements. In Desktop conditions, we used mouse movements on the desktop monitor (Fig. 4b). In HMD conditions, we used the projected movements, which were the positions of the controller projected on the 2D view plane (Fig. 4d); this plane was the equivalent of a desktop monitor, and the projected movements were analogous to the mouse movements. As such, all the rotation interactions occurred in a 2D space. We chose a scale factor so that a long stroke drawn by holding the VR controller roughly matched a long stroke drawn by moving the wrist and the mouse for an average size participant.
- **Navigation** Nevertheless, it was impossible to find exact matches of head movement and physical navigation in VR for a desktop monitor. We approximated these by allowing virtual movement where participants could move along xz-axes using a keyboard (the WASD keys, see Fig. 4b). One key stroke was mapped to a 0.02 units change, which resulted in a speed of 0.6 units per second if constantly holding the key (∼30Hz); this corresponded to a speed of 60 centimeters per second in VR, similar to a

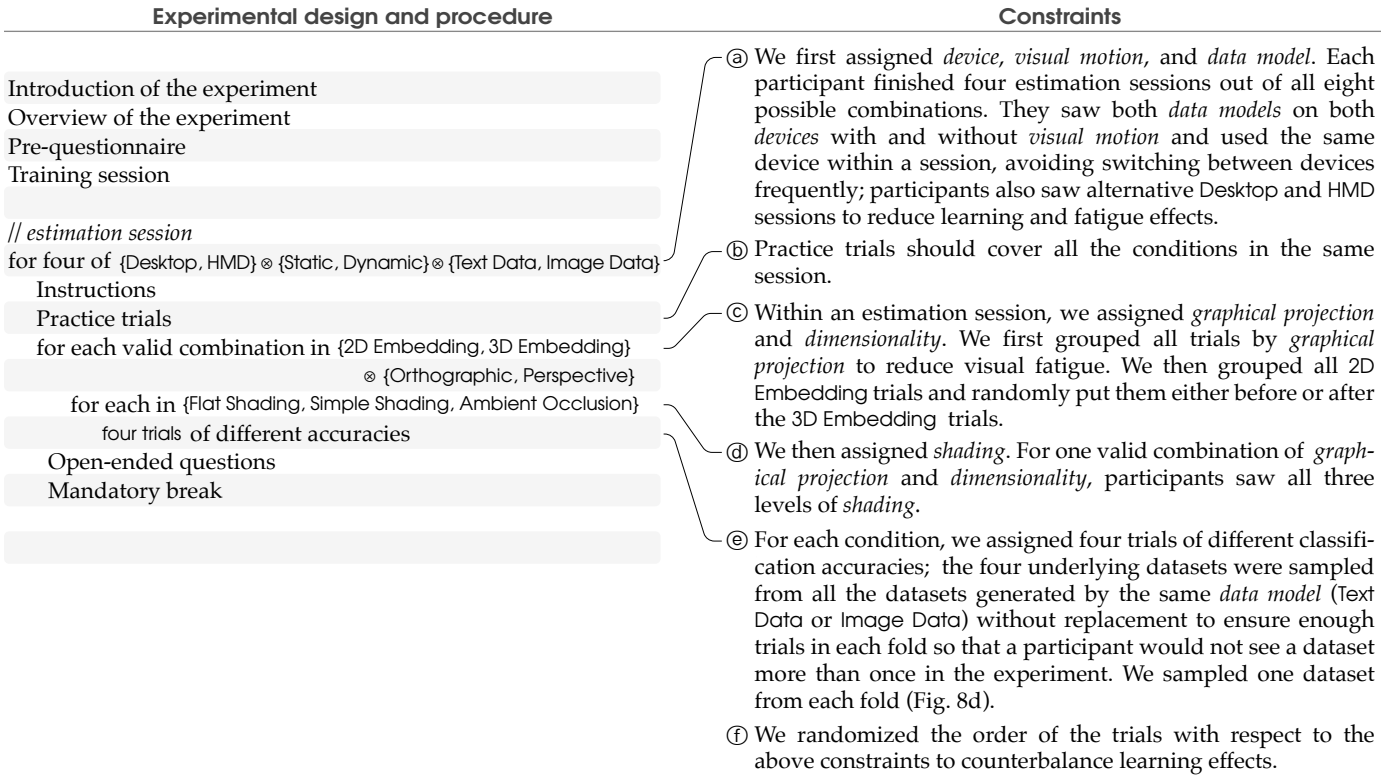

Figure 9. The experimental design, procedure, and constraints.

slow walking speed for an average size participant. Using keyboard input to approximate physical movement in VR is common in the research across modalities [109], [110].

4 Study Design

The research question of this study concerns **the effects of visual cues and the relationships between them in a task of assessing classification performance;** we leveraged the design for measuring vision and omitted other senses. We reflects this research question in our experiment design, of which overview is presented in Fig. 9.

4.1 Experimental Design

The number of experimental conditions depends on the cue combinations. All the combinations of the visual cues and devices were compatible with 3D Embedding, resulting in 48 different experimental conditions $= 2$ *devices* $\times$ 2 levels of *motion* $\times$ 3 levels of *shading* $\times$ 2 types of *graphical projection* $\times$ 2 *data models*. We also have four 2D Embedding conditions for different devices and datasets, resulting in $52 = (48 + 4)$ experimental conditions.

We used a mixed factorial design with repeated measures. Such a design is one of the basic and the most widely used designs in human-subjects experiments [111]. This design replaced a complete within-subjects experiment for two reasons: (1) both *data models* did not provide enough datasets of different accuracies to cover a full combination of visual cues (Fig. 8); (2) using the same *data model* throughout hundreds of trials may lead to strong learning and practice effects [112]. We therefore carefully assigned and balanced the conditions so that each participant finished half of all the combinations of experimental variables, and each experimental condition was assigned to the same number of participants. This design leveraged the within-subjects component to reduce the impacts of individual differences, allow us to infer interaction effects between variables, and require fewer participants.

4.2 Procedure

Before the experiment, participants signed a consent form and took part in the introduction session with the experimenter. They then started to use the experiment system. They first saw an overview of the experiment and filled in a pre-experiment questionnaire, including demographics questions and their self-assessment of familiarity with machine learning. They then finished four estimation sessions with a two-minute mandatory break between sessions; longer breaks were allowed. At the end of each session, they answered two open-ended questions: one to record their strategies (e.g., "Any visual features or patterns you looked for?") and the other for additional comments. After participants completed all four estimation sessions, they filled in a post-experiment questionnaire to briefly assess motion sickness in the experiment and to report their familiarity with the datasets. Participants typically spent 60 to 90 minutes in total.

In each trial, participants first saw a blank white screen. After one second, they saw the scatterplot and could explore the scatterplot for up to 60 seconds. When participants felt ready, they could proceed by pressing the space key (Desktop) or pulling the trigger on the controller (HMD). Participants then saw another blank screen for one second, immediately followed by the input interface, where they provided a number as their estimation of accuracy for this trial.

4.3 Training and Practice

To ensure that participants understood the task and expectation, we included introduction and training sessions at the beginning of the experiment. We also designed practice trials at the beginning of each estimation session.

In the introduction session, participants were shown how to wear the headset, use the controller, and move in the room. In the training session, participants learned the task and interaction techniques. They were given the following instructions to estimate classification accuracy from scatterplots: *"The scatterplots show the outputs of machine learning classifiers for test datasets. Each dot represents an instance. The color of a dot represents the ground truth class of that instance. The same color represents the same class. Different colors represent different classes...For each scatterplot, we ask you to estimate the classification accuracy for that dataset. You see one scatterplot each time. Then, you proceed to provide your estimation of accuracy, and the scatterplot will be removed."* Participants were also given eight examples of 4 accuracy levels. The full instructions and the examples are available in Appendix D.

At the beginning of each estimation session, participants practiced each assigned condition once and only once, except that they had two 2D Embedding trials because there were fewer 2D Embedding conditions to practice. We grouped practice trials in the same fashion as the main trials, sampled the datasets similarly (Fig. 8d), and excluded these datasets from the main trials. The procedure of a practice trial was similar to a main trial, except that participants saw the true classification accuracy at the end of the trial.

4.4 Experimental Setup

The scale of the immersive environment was about 310 $\times$ 300 centimeters in a physical room of 600 $\times$ 440 centimeters; the room was constantly quiet, and it had an unobstructed floor so that participants could move freely. The near and far clipping planes were fixed to 0.018 and 75.0, respectively, such that all the visual elements were visible, of a similar size, and similar to an overview [48] when viewed at the default (starting) camera position in all the conditions. The default (starting) camera position and other camera parameters were determined based on the following aspects. First, the scatterplots shown in the HMD $\circ$ Static conditions roughly fell into central vision so that participants were able to see color, shape, and shading. Second, orthographic and perspective projections generated scatterplots of a similar size. Third, the camera position was slightly higher than the center of the scatterplot, matching the angle of viewing a screen on a desktop and the metaphor of a physical desk [78]. Fourth, participants could see the reference frames even in Static conditions. All the scatterplots were placed surrounding the center of the space extending about 100 centimeters (one unit) in each of the xyz dimensions. Consequently, on average, a scatterplot was approximately 540 pixels subtending an angle of 12.64° (43.51 pixels per degree) for Desktop, assuming a viewing distance of 65 centimeters, and a scatterplot was

about 735 pixels subtending an angle of 34.19° (21.50 pixels per degree) for HMD assuming a default viewing distance of 162.6 centimeters between a participant and the center of the scatterplot. And each sphere in the scatterplots was about 0.17 centimeters (7.5 pixels) for Desktop or 15 centimeters for HMD in radius. The scatterplots, reference frames, and other visual elements were adjusted to match each participant's body height so that their center was 30 centimeters below participants' eyes.

4.5 Implementation and Apparatus

The two neural networks were implemented and trained using TensorFlow 1.12, Keras 2.2.4, and Python 3.5. All the 2D and 3D t-distributed stochastic neighbor embeddings were pre-computed using scikit-learn and Python 3.5. All the interfaces, visualizations, and interactions were implemented based on OpenGL 4.5, GSLS 4.50, Qt 5.12.1, C++14, OpenVR 1.2.10, and SteamVR 1.2.10; they were rendered using four sub-samples for anti-aliasing. The desktop was equipped with an AMD Ryzen 2700X 8-core processor, NVIDIA GeForce GTX 1080 Ti graphics card, a 32GB RAM, and an ASUS PA248 monitor (24", 1920 × 1200, 60 Hz). The same desktop was used to render to the HMD (HTC Vive Pro, 2018 model, 2299 × 2554 per eye, 90 Hz); only one handheld controller was used to interact with the scatterplots and control the flow of the experiment.

4.6 Participants

We recruited 32 participants (16 female and 16 male) from our institution and others nearby and finished the experiment before the COVID-19 pandemic. We paid participants $10 per hour as compensation for their time.

We expected participants to have passing knowledge in machine learning and to understand the concepts of classification and clustering. In the pilot study, three participants with a computer science background showed sufficient understanding of the concepts and task; the fourth and last participant with a design and art background claimed that she did not understand the relationship between the scatterplot and accuracy. Therefore, we used a recruiting criterion where participants claimed that they had taken or were taking at least one of the following graduate-level courses: machine learning, deep learning, computer vision, or data science, or that they used machine learning techniques in their research or at work.

We present participants' demographics and self-assessments in Fig. 10. We had undergraduates, masters, doctoral students, and postdoctoral researchers from the areas of computer science, data science, solid mechanics, applied mathematics, electrical engineering, engineering physics, brain science, neuroscience, biostatistics, economics, humanities and education, digital and media, and liberal arts. All the participants were between 18 and 65 years old ($\mu = 23.25$, $\sigma = 3.20$), having a normal or corrected-to-normal vision, and not colorblind.

All the experimental sessions were proctored by the same author as the experimenter following the same experimental protocol (see Appendix B). All the participants used the same setup and apparatus, and the two base stations (the Lighthouse tracking system) stayed in the same locations throughout the experiment.

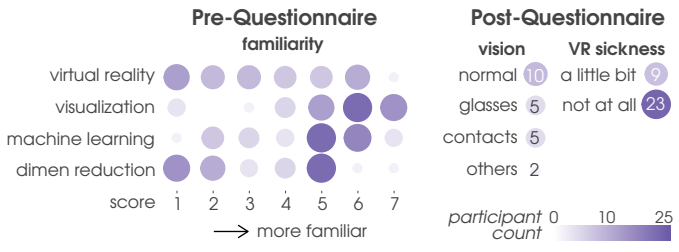

Figure 10. **Participants' background**. Most participants were familiar with visualization and machine learning, while their familiarity with virtual reality devices varied.

4.7 Dependent Variables (Measures)

We recorded two measures for each trial:

Error magnitude is defined as the amount of difference between participants' estimation and the actual classification accuracy. We used the difference between the two percentage numbers (integers) because the participants were trained, practiced, and responsed in the same way.

Response time is defined as the time interval between when the scatterplot was first shown and then removed. A scatterplot might be removed due to participants actively proceeding or a timeout.

Between the two measures we are more interested in error magnitude. The response time was affected by the unexpected noise, especially in HMD ∘ Motion conditions: the HTC Vive pro headset might show a flashing screen occasionally [113], causing wait times; participants sometimes tried to come back to the home position before they responded, which was counted in their response time.

In total, we collected 4,224 trials = 3,328 main + 896 practice trials = (26 × 4 + 28) trials per participant × 32 participants. We excluded practice trials and the five trials where participants claimed they skipped accidentally. We also manually excluded three error trials where the response was too close to the default answer, but the estimation error was extremely large (>50%); we think that these trials were also skipped but not reported. As a result, we based our analysis on the remaining 3,320 trials.

5 Analyses and Results

5.1 Guiding Questions (GQs)

Toward our goals of understanding the effects of visual cues and providing insights for immersive analytics, we framed five guiding questions (GQs) [114] to formalize the analyses.

GQ1 How do the six manipulated variables affect participants' performance? This research question was inspired by previous studies examining dominant pictorial cues in a 3D space [2], [14], but our context is immersive environment.

GQ2 How do the experimental variables interact with each other? The literature suggests that depth cues interact in complicated ways [21], [43]. We wish to understand the interaction effects of the selected cues (e.g., across different devices [18]). In contrast to these studies, we used a task for cluster perception and selected a broader range of cues.

GQ3 How does each manipulation affect participants' performance? This is similar to the hypothesis that the design of a visual object will influence the perceived depth [21]. Here we tested on thousands of virtual objects collectively and went beyond the perceived depth to broader cluster understanding and interpretation.

GQ4 Which resulting scatterplot designs are more effective for our task? Previous studies suggest that cue integration affects depth estimation in a complex manner, with certain combinations improving spatial perception [2], [21]. For our task and setup, we sought combinations of cues and devices that facilitate assessing classification performance.

GQ5 How do participants interpret classification accuracy from a scatterplot? Participants' perception of classification performance was measured in the unit of accuracy estimation, we aim to disambiguate aspects of participants' perceptual process based on their answers to the open-ended questions.

5.2 Analysis Methods

5.2.1 Quantitative Analysis

We used Bayesian inference for our quantitative analysis [115], [116]. Bayesian inference calibrates all assessments to prior distributions, provides inference given current evidence, and mitigates multiple comparison issues [117]. It is suitable for our experiment with a set of variables and conditions. The current research of sensory cue integration also suggests a Bayesian approach to incorporate uncertainty and prior knowledge of the environment for cue combinations [54]. Therefore, Bayesian modeling approaches appear to be a promising means for quantifying the impact of visual cues on participant performance for our experiment.

We built a Bayesian multilevel model for each of the two measures using gamma distributions as likelihood function and logarithm as link function. All the data are non-negative, and a gamma distribution captures the asymmetry in data; the literature also shows that time of waiting a response could follow a gamma distribution [118]; additionally, gamma distributions better describe our data than skewed normal and log-normal distributions. In particular, error magnitude contains zero (no error), and we therefore use a hurdle gamma distribution [119]; response time has an upper bound. Taking together, in brm's extension of Wilkinson-Rogers-Pinheiro-Bates notation [120], [121], [122], the first model of error magnitude is:

$$
\begin{aligned}
&error\_magnitude \sim hurdle\_gamma(\mu, shape, hu)\\
&log(\mu) = device * visual\_motion * shading * projection * data\_model\\
&\qquad + device * dimensionality * data\_model\\
&\qquad + (1 + dimensionality * data\_model \mid participantID)\\
&\qquad + (1 \mid trial\_true\_accuracy)
\end{aligned}
$$

where μ, $shape$ and hu are parameters of the gamma distribution, $projection$ is short for *graphical projection*, and $trial\_true\_accuracy$ represents the ground truth classification accuracy for a trial. This model was designed to closely follow our experimental design. The first two terms define a log of the mean of error magnitude as a joint linear function of all six experimental variables, where *device*, *visual motion*, *shading*, *graphical projection*, and *data model* could interact with each other, and *dimensionality* varies between *device* and *data model*. We grouped data by both participants and the true accuracy values (random intercepts). We modeled *dimensionality* and *data model* as group-level effects (random slope) to capture the differences in data properties from dimension reduction. Since the distributions of error magnitude are similar in shape across different experimental conditions, we did not use a submodel for the shape parameter; we assume both $shape$ and hu are consistent across different conditions. The second model is

$$
\begin{aligned}
&response\_time \mid cens(cen) \sim Gamma(\mu, shape)\\
&log(\mu) = device * visual\_motion * shading * projection * data\_model\\
&\qquad + device * dimensionality * data\_model\\
&\qquad + (1 + device * visual\_motion \mid participantID),\\
&shape = device * visual\_motion
\end{aligned}
$$

where $cens(cen)$ specifies which observations were beyond the 60 seconds upper boundary (called "left-censored" [123]). This model is very similar to the one above for error magnitude, but they are different in group-level effects (random slopes) and the choice of using a submodel. Compared with *dimensionality* and *data model*, response time is more likely to be affected by *device* and *visual motion*, which were random slopes. We also noticed that the distributions of response time could be very different in shape across device and motion levels, and therefore used a submodel with population-level effects for the shape parameter (adding group-level effects to this submodel would make the model not converge).

We used and centered weakly informed priors which can capture most of the observations within 2 standard deviations. We recoded each variable using orthogonal contrast coding (e.g., Desktop $\mapsto$ -0.5, HMD $\mapsto$ 0.5; Static $\mapsto$ -0.5, Dynamic $\mapsto$ 0.5) such that the model coefficients are compatible. We also performed convergence and posterior prediction checks to ensure that the models are appropriate for the data. We implemented these using R packages rstan [124], brms [120], and tidybayes [125]. The details of coding, modeling, and diagnostics are available in supplementary materials.

5.2.2 Qualitative Analysis

We followed *thematic analysis* to encode participants' strategies of accuracy estimation from their answers to the open-ended question (i.e., "Any visual features or patterns you looked for?"). Two experienced researchers (one author and one coder) extracted the strategies and encoded all answers separately; they then merged their encodings and strategies via discussion. We present the results in Section 5.3.5.

5.3 Results

5.3.1 GQ1: The Effects of Each Variable

Following the first guiding question, we looked at the model coefficients of each of the experimental variables. These coefficients are from population-level effects that tell us the overall "weight" of each variable when conditioning on that variable for an average participant; note that the sign and direction of an effect depends on how we coded the data. We present the results in Fig. 11.

Error magnitude Adding *visual motion* is very likely to affect error magnitude. Changing *shading*, *dimensionality*, or *data model* is likely to affect error magnitude a little; but the later two are less identifiable. Switching *device* or *graphical projection* methods is unlikely to affect error magnitude.

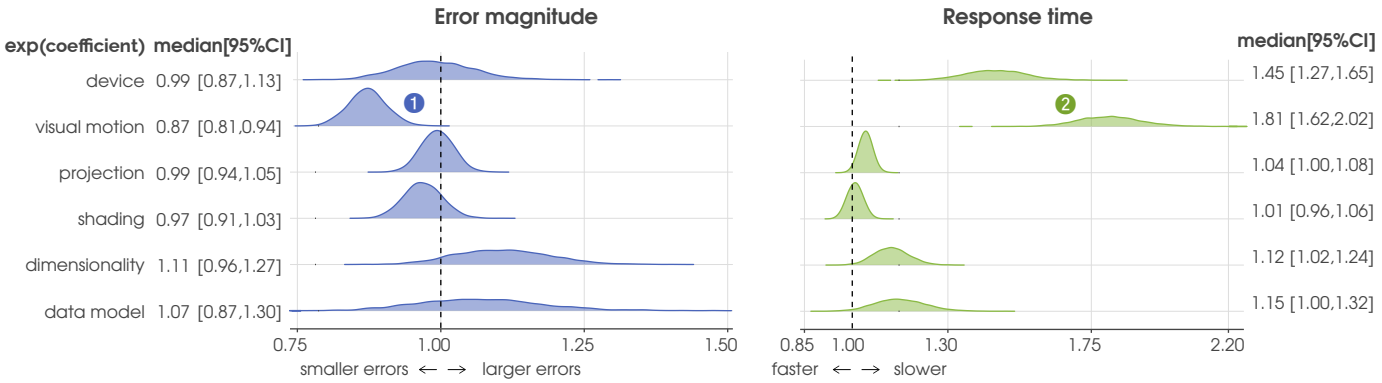

Figure 11. **GQ1: The population-level effects of each manipulated variable.** We show posterior distributions, 95% quantile credible intervals (CIs), and median for coefficients. **Example of interpretation**: when averaging other factors, changing from Static (-0.5) to Dynamic (0.5) ❶ is very likely to reduce an average participant's estimation error (0.87 times smaller in magnitude), ❷ and also to let the participant response more slowly (around 1.81 times longer), but this effect is less identifiable.

Response time Varying *device* or *visual motion* could largely affect response time (e.g., it may double response time in the worst case). Manipulating *graphical projection*, *dimensionality*, or *data model* is likely to affect response time a little. On average, manipulating *shading* did not seem to affect response time in this experiment.

5.3.2 GQ2: The Interaction Effects between Variables

Similar to our investigation for the first guiding question, we looked at the model coefficients to investigate the interactions effects between variables. We only considered two-way interaction effects for simplicity. More complicated interaction effects are implied by the analyses below in Section 5.3.4. We report the results in Fig. 12.

Error magnitude Most of the variable pairs suggest a small or moderate two-way interaction effect between them except that *device* does not seem to interact with *visual motion* or *shading*. For instance, we see a moderate interaction effect between *visual motion* and *shading*, which suggests that different levels of visual motion are likely to show different effects when combined with different levels of shading.

Response time We found a strong two-way interaction effect between *device* and *visual motion*, suggesting that different devices (Desktop and HMD) affects response time differently for different levels of motion. The other variable pairs suggest either small or moderate interaction effects (e.g., *device*:*shading*) or almost no interaction effect (e.g., *visual motion*:*data model*).

5.3.3 GQ3: The Effects of Manipulations

To compare different levels of manipulations, we draw samples from posterior distributions to get an estimate of mean for each measure. These posterior distributions show how an average participant does when conditioning on that level of manipulation. We present the results in Fig. 13.

Error magnitude We found that allowing Dynamic (cf. Static) is likely to decrease error of the average participant's estimate; and there is a similar effect of decreasing error when showing a Text Data dataset (cf. Image Data). Desktop (cf. HMD) is likely to slightly decrease the error of an average participant's estimate; and this is similar to those effects of Perspective Projection (cf. Orthographic Projection) and 3D Embedding (cf. 2D Embedding). Last, using more advanced

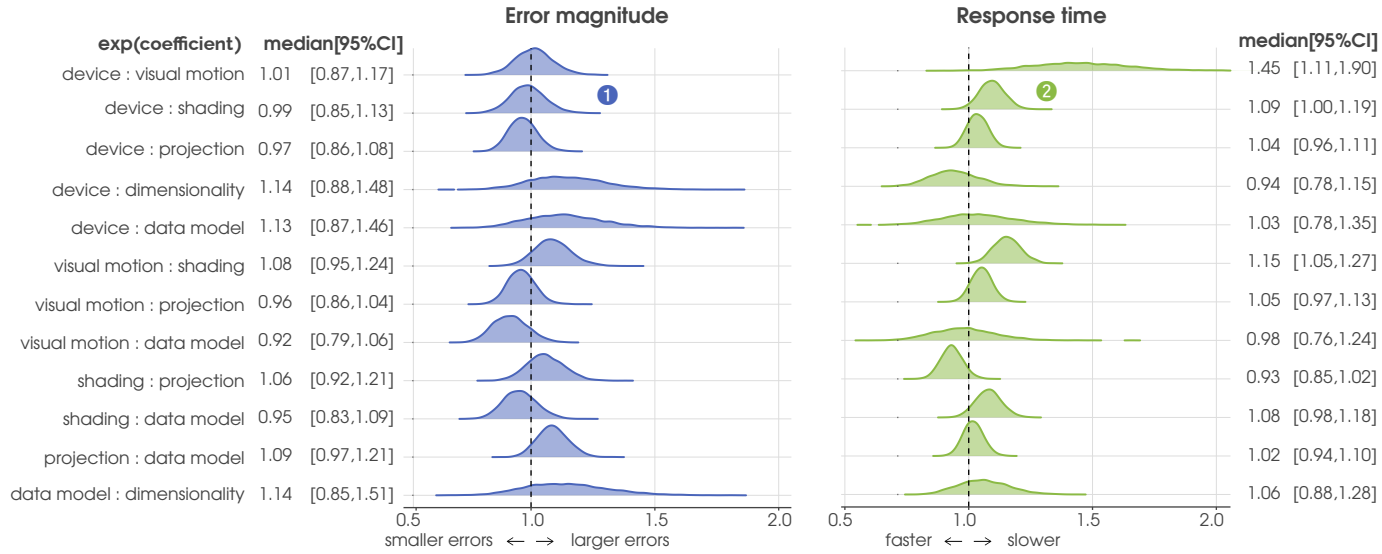

Figure 12. **GQ2: The two-way interaction effects.** We report posterior distributions, 95% quantile credible intervals, and median of two-way interaction coefficients. For error magnitude, most variables moderately interact with each other. For response time, there is a strong interaction effect between *device* and *visual motion*, and several small to moderate interaction effects. **Example of interpretation**: when averaging other factors, if we change from Desktop (-0.5) to HMD (0.5), further changing from Flat Shading (-0.5) to Ambient Occlusion (0.5) is ❶ unlikely to cause a strong difference in an average participant's estimate error; ❶ however, it is likely to let the average participant responds slightly more slowly, typically in the range of [1.00x, 1.19x] more slowly.

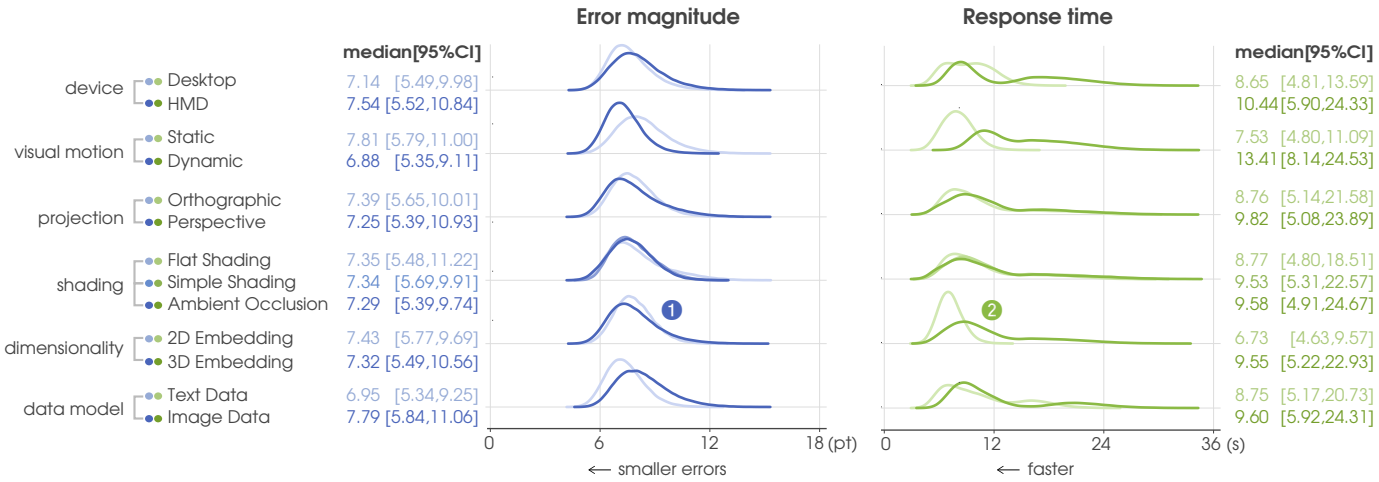

Figure 13. **GQ3: The effects of different manipulations.** We report posterior distributions, 95% quantile credible intervals, and median of mean for each measure. **Example of interpretation**: when averaging other factors, for a 3D Embedding dataset, ❶ an average participant typically estimates the accuracy 5.49 to 10.56 points aways from the true classification accuracy; these estimates are slightly smaller errors than the participant's estimate of a 2D Embedding dataset. ❷ The participant typically spends 5.22 to 22.93 seconds before responding to a 3D Embedding dataset, which are longer than the time of a 2D Embedding dataset.

shading shows a trend to decrease errors, but the difference seems very small and almost imperceptible.

Response time We found that using Static (cf. Dynamic) is very likely to largely decrease response time, and there is a similar and weaker effect of reducing response time when showing a 2D Embedding dataset (cf. 3D Embedding). On average, using Desktop (cf. HMD), switching to Orthographic Projection (cf. Perspective Projection), or showing a Text Data dataset (cf. Image Data) is likely to reduce response time a little. Again, degrading *shading* shows a trend to reduce response time, but this trend seems very small and nearly imperceptible. Last, the bimodal shape of several distributions is consistent with our previous observations of interaction effects; that is, different manipulations may show different effects depending on another variable (see Section 5.3.2).

5.3.4 GQ4: The Effectiveness of Each Combination

To investigate the effectiveness of each combination (corresponding to an experimental condition), we draw samples from the posterior distributions to get an estimate of mean. These posterior distributions show that how an average participant performs with one combination. We present a visual summary for all the 52 combinations in Fig. 14. Given our primary interest in error magnitude, we also form a rank list of performance based on median of mean and report the distributions of the best- and worst- performing combinations in Fig. 15; the full list is available in Appendix D.

Error magnitude Across all the tested combinations, we observed large differences between the best- and worst-performing combinations, yet many combinations were similar. The results indicate many interesting but small effects, and we note a few stronger effects here. Allowing Dynamic

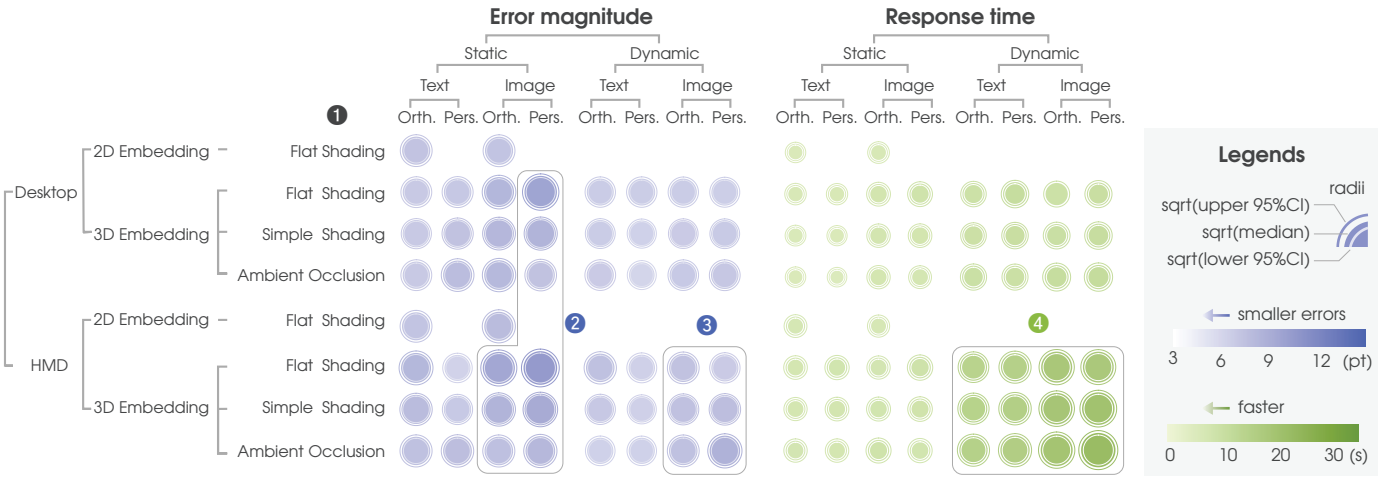

Figure 14. **GQ4: The effectiveness for each combination.** We dual-encode posterior median as both color and size and encode the lower and upper bounds of 95% CIs as radii to show uncertainty in posterior estimates. This figure visually summarizes all the effects. **Example of interpretation**: ❶ Between the two measures, the differences in error magnitude are milder; the interaction effects among error magnitude are also subtle and more complicated. With Static scatterplots, different *shading*, *graphical projection*, and *data model* could have small effects on error magnitude. More advanced *shading* generally decrease errors, ❷ especially for a more complicated dataset like Image Data. With Dynamic scatterplots, these subtle effects fade out, and ❸ more advanced *shading* could slightly increase errors, but this effect is very small. ❹ The differences in response time are dominated by the interaction effects between *device* and *visual motion*–using Dynamic or HMD is very likely to lead to a much longer response time, especially when doing both. *(Orth. is Orthographic Projection, and Pers. is Perspective Projection)*

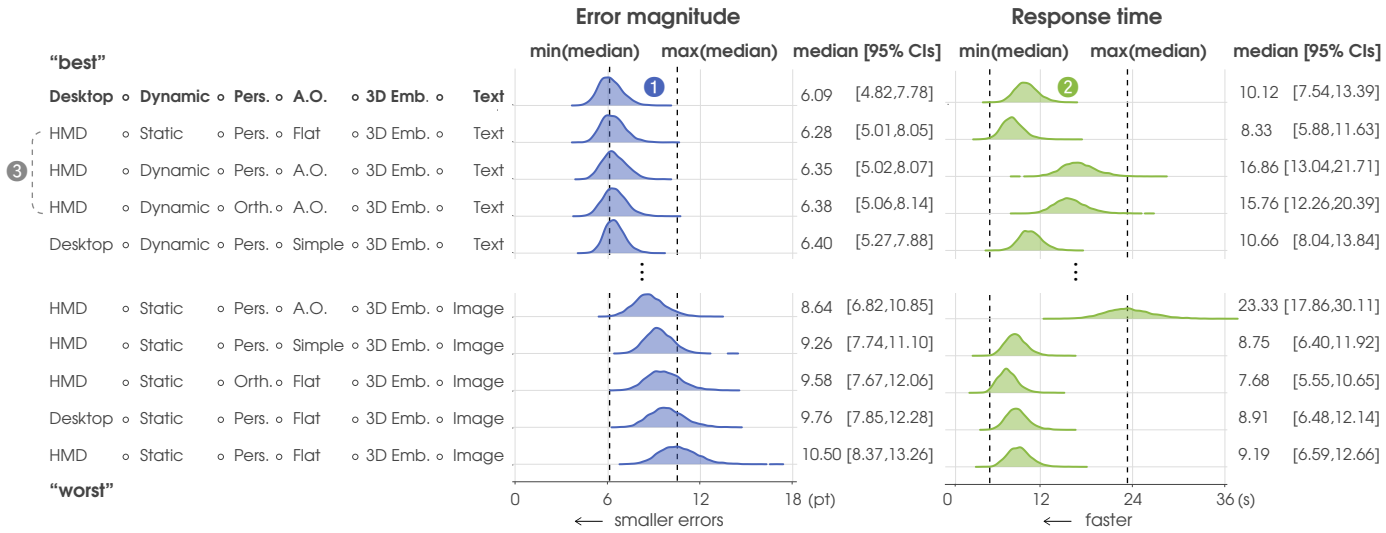

Figure 15. **GQ4: The effectiveness of the"best" and "worst" combinations,** sorted by posterior median of mean of error magnitude. This figure shows posterior distributions, medians, and 95% CIs of mean. The full rank list is available in Appendix D. **Example of interpretation**: for an average participant, the "best" combination that leads to the smallest error is Desktop ∘ Dynamic ∘ Perspective ∘ Ambient Occlusion ∘ 3D Embedding ∘ Text Data, ❶ where the estimate typically deviates [5.17,8.35] percentage points away from the true accuracy; ❷ this combination also has a relevantly shorter response time (typically in the range of [7.54, 13.39] seconds) among others. ❸ However, a few HMD combinations could be almost as "good" as this best combination, but they usually result in a longer response time. *(Emb. is Embedding, A.O. is Ambient Occlusion, Orth. is Orthographic Projection, and Pers. is Perspective Projection.)*

generally slightly decreases errors and mitigates the effects of other variables. Using Desktop (cf. HMD) generally decreases errors a little, but with a set of carefully chosen visual cues, an HMD performs nearly as well as the best-performing Desktop condition. Improving *shading* seems to decrease errors for Static scatterplots, but it slightly *increases* errors for some HMD ∘ Dynamic conditions. Additionally, the 2D Embedding conditions are similar to an average 3D Embedding condition.

Response time The results of response time are more consistent and salient. Using a Desktop or Static largely decrease response time, especially when doing both. A simpler dataset like Text Data is likely to slightly decrease response time, but this effect is subject to other variables. Last, 2D Embedding leads to a similar response time of an average Static ∘ 3D Embedding condition.

5.3.5 GQ5: Participants' Strategies

We present the results from thematic analysis and report participants' strategies in Fig. 16.

Strategies The most common two strategies are seeking "class separation" (53.13%) or looking for "degree of mixing colors/overlapping" (65.13%); other common strategies include estimating "portion (percentage) of colored points" (37.50%), inspecting "density/distance between points" (25.00%), and examining "class boundary" (21.88%). A unique strategy reported by only one participant is considering "continuous color blocks" (3.13%).

Participants' strategies varied across different sessions (Fig. 16b). Only four (12.50%) participants reported one single strategy, and the remaining twenty-eight (87.50%) developed more than one strategy in the experiment. Sixteen (50.00%) participants used consistent strategies across the four sessions, and the rest used different strategies in different sessions.

A few participants reported interaction techniques they used (Fig. 16c), with rotation being the most helpful one

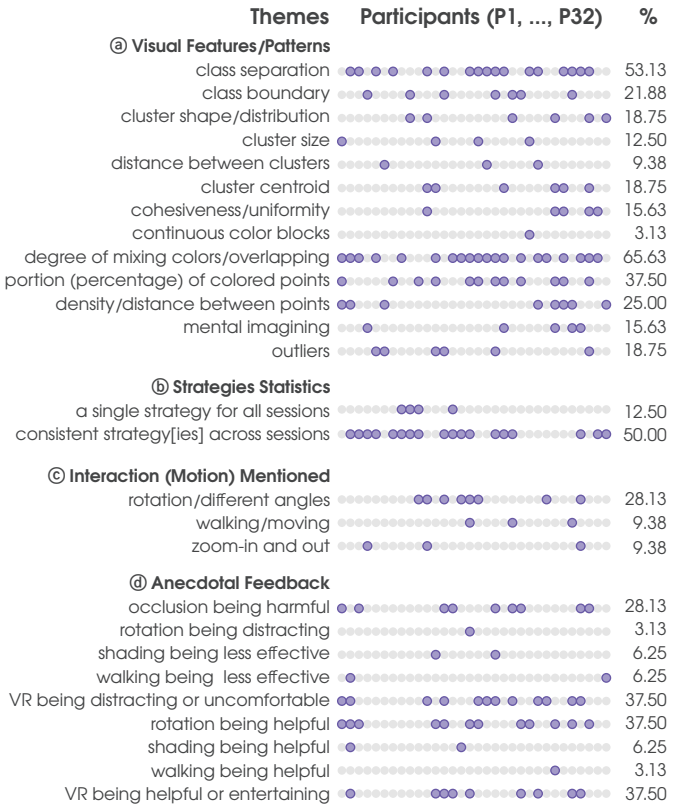

Figure 16. **GQ5: Participants' strategies.** Each dot represents a participant, and a purple dot indicates that a theme appeared in that participant's answers.

(28.13%). In the anecdotal feedback (Fig. 16d), participants found occlusion harmful, and that VR could be distracting or causing physical discomfort (37.50%), but that it also could be entertaining or helpful for the task (37.50%).

Discussion All these strategies appear to be feasible for assessing classification effectiveness and estimating classification accuracy. The two dominating strategies suggest that

the task was correctly interpreted as visually distinguishing classes. The presence of other indie strategies suggests that other factors (e.g., visual cues and outliers) may have affected participants' strategies and understandings. The consistence across different sessions may suggest that only half of the participants interpret the task consistently across different visual motion levels and devices; it implies that design for fixed and interactive visualizations should be considered separately. Different visual cues may aid or hinder the use of a particular strategy, and more effective combinations of cues might have assisted in a more commonly used strategy or more strategies. Manipulating cues to facilitate the dominating strategies could improve most participants' performance, but the optimal may be to test individual's strategies on different devices and datasets, and finally to personalize the visualizations.

6 General Discussion

6.1 Individual Cues

We find a tradeoff between error magnitude and response time for each individual cue. This observation aligns with the findings from previous studies comparing a desktop monitor with a VR/AR device [19], [20], [58], [59], [60], [70], [78]. The results from our study seem not fundamentally deviate from those studies using simple, abstract tasks. We speculated that perceiving classification accuracy may be considered a combination of multiple low-level abstract tasks, and the literature supports that multiple low-level tasks could form a high-level perception task; this is similar to using low-level visual proxies to summarize mean values in bar charts [126], the bottom-up model for visual attention [127], and parallelly processing multiple cues in visual search [128]. To prove this speculation requires systematic manipulating each subtask and knowing which subtask dominates perception, which is beyond the scope of this work.

With two exceptions, we summarize our reasoning and key messages from GQs 1 and 3 is as follows:

> Manipulating *any* cue could show certain—sometimes nearly imperceptible—effects on estimation error and response time; this is possibly because each cue could improve a part of people's perception in 3D space but also forces people to take a longer time to process and examine the cue.

The first exception is that the commonly-used perspective projection may slightly increase both error magnitude and response time; perspective projection utilizes size as a depth cue, but different sphere sizes may impede cluster perception where people visually group spheres to form a cluster. According to *Gestalt psychology* [129], [130], people seek similarity in elements (e.g., size, movement) to visually form groups, but if the elements from the same group (e.g., the spheres with the same color) are of different sizes, this conflict may prevent people from grouping the spheres of the same color. As such, perspective projection may sometimes have caused a decline in classification perception, and therefore orthographic projection sometimes surprisingly showed better performance. The second exception is that 3D embedding from t-SNE sometimes increases errors, although it generally reduces errors but increases response time. 3D embedding has the third dimension as a cue. While their projection results better represent the original high-dimensional data (e.g., the Kullback-Lerbler divergence is smaller) and provide more information than 2D embedding results, these extra information still have to be present carefully for participants to utilize them appropriately.

There are a few explanations for the moderate effects and the tradeoff. The dimension-reduction technique (t-SNE) we chose is well-established for preserving clusters, and this specific technique may have dominated cluster perception over the visual cues we were testing [28]. Given the dense experiment, carryover effects may have negatively affected participants. Further, using an HMD usually requires longer moving distances (e.g., walking vs. pressing keys, moving an arm vs. moving a mouse) to examine different perspectives of a scatterplot; participants were less familiar with performing an analytic task on such a device. Also, HMD provides more pixels but less pixels per degree. Allowing movements to enable visualmotion causes more time to be spent on examining the stimuli. Alternatively, these subtle effects may also explain the conflicting results in the literature.

6.2 Cue Integration

Our results hint at a connection to cue integration theory [12], [131], that is, people can make more accurate estimates of environment properties by integrating multiple sources of information [54]. In our experiment, visual cues interact with each other in complicated ways, similar to the findings from prior studies in AR [21]. Part of the reason is that participants had multiple strategies or changed their strategies. Each cue may facilitate or impede a strategy, and combining cues may cause conflicts in perception. Perspective projection and ambient occlusion together generate a good sense of depth, and combining them was very effective on a desktop monitor in our experiment. However, in some cases, perspective projection and ambient occlusion may not be beneficial; this suggests that participants may have used more than one cue, and cue integration is occurring.

We summarize our reasoning and key messages from GQs 2, 4, and 5 as follows.

> Besides the differences in devices and physical movements, manipulating *multiple* visual cues may aid assessing classification performance.

For example, more advanced shading (e.g., Phong lighting model, ambient occlusion) on both devices generally improves participants' performance; they reduced estimation error without causing a substantially longer response time.

> However, combining many individually-beneficial cues may cause a decline in performance: participants made worse estimates and spent more time, especially when they were working on a more complicated dataset.

For example, when assessing a dataset from Image Data with Dynamic in VR, participants' performance will decline slightly if further adding perspective projection.

We have three possible explanations for these obversations. The first relates to *visual complexity*; that is, multiple visual cues together aid 3D perception, but they also raise visual complexity that may prevent participants from using

visual similarity to group elements and seek patterns. The second is that cues may conflict with each other [54]. For example, our shading models use darker colors to indicate a further position in 3D space, while perspective projection uses smaller sizes to indicate a further position. However, a further small sphere may not necessarily be darker than a near one, and this could cause a conflict in perception. The third explanation is that adding more cues demands a precise implementation for each, because a small discrepancy in any implementation may hinder the overall perception. In our case, the screen-based ambient occlusion algorithm generates imperfect results such as noticeable black edges and aliasing, which are very salient in VR. These defects may explain that ambient occlusion may not always be as effective as the simple shading using the basic Phong lighting model.

6.3 Beyond the Effects of Visual Cues

Our experiment aimed to go beyond previous experiments for immersive analytics on two frontiers: one is to observe the effects of combining multiple cues suitable for immersive visualizations, and the other is to study a task that is more reflective of the real-world scenarios to improve ecological validity. These two aspects increased the complexity of experimental design and data analysis, but we offer insights, lessons, and protocols that future research could refer to.

We shed light on how to use virtual reality as an effective alternative to a traditional desktop monitor to understand complicated datasets and perform tasks like interpreting deep learning results. The third spatial dimension, interaction techniques, and stereopsis available in an immersive environment may reveal insights and strategies that are unique to such an environment; these properties also encourage an appropriate combination of visual cues, which disclose more visual encoding channels to express various data properties and to illustrate complex datasets.

We also aim to move beyond abstract data and tasks for visualization evaluation and therefore recruited participants with expertise in the application domain. While the student participants were still relatively novice to machine learning, their behavioral data demonstrate the potential of using virtual reality for understanding and debugging machine learning models; their feedback will help us enhance the design and use of visualizations for similar tasks.

We tested a subset of important visual cues and and controlled as many factors as possible; but there is still room for improvement. For example, in our design procedure, we noticed that the direction of light could be another important factor because it affects shading, but we did not find enough literature to support the speculation. We also designed for an average size participant, and we speculated that part of our configuration could be adaptive. In addition, other non-visual cues, such as haptic vibration feedback [43], body awareness (proprioception), and balance (the vestibular sense), may affect the task performance and could be incorporated with visual cues to improve the effectiveness of immersive analytics further.

Finally, while the elaborate experimental design and the variance in a virtual reality system raise challenges in data analysis, we find Bayesian inference is particularly suitable. Future research could look into Bayesian decision theory to mathematically compute cue integration.

7 Conclusion

We assessed the effects of visual cues on the task of estimating the classification accuracy of a deep neural network based on the last hidden layers' output and t-SNE. We found that participants' estimation was affected by the device used, the combination of cues shown, and the data they worked with. Among all of the cues, adding visual motion shows strong effects on reducing both estimation error and response time. An HMD can lead to better, worse, or similar performance to a desktop monitor depending on the combination of cues (e.g., especially whether visual motion is available). Improving shading generally slightly reduces estimation error, but the effect interacts with the choices of device and graphical projection. The relationships between cues are complicated and depend on data properties and participants' strategies, and our results provide weak evidence that using more cues may cause a decline in participants' performance. While we suggest using a few strong cues in design for immersive visualization, we anticipate that future studies could further explore the effects of cue integration. Our work advances the understanding of the effects of visual cues on visualization perception, provides insights for immersive analytics, and validates the use of visualizations for assessing the performance of a deep neural network.

Acknowledgments

This research was supported by hardware donations from NVIDIA. We thank the anonymous reviewers for their thoughtful comments. We thank Shenghui Cheng, Loudon Cohen, Ailin Deng, Aron Gokaslan, Mi Feng, Elaine Jiang, Benjamin Knorlein, Johannes Novotny, Emily Reif, Jing Qian, Kexin Qu, and Yalong Yang for their help with this research. We also thank Benjamin Fancy, Mi Feng, Jennifer Kim, and Jing Qian for their help with the manuscript.

References

[1] K. Marriott, J. Chen, M. Hlawatsch, T. Itoh, M. A. Nacenta, G. Reina, and W. Stuerzlinger, *Immersive Analytics: Time to Reconsider the Value of 3D for Information Visualisation*, 2018, pp. 25–55.

[2] L. C. Wanger, J. A. Ferwerda, and D. P. Greenberg, "Perceiving spatial relationships in computer-generated images," *IEEE CGA*, no. 3, pp. 44–51, 1992.

[3] M. M. Chun and Y. Jiang, "Contextual cueing: Implicit learning and memory of visual context guides spatial attention," *Cognitive psychology*, vol. 36, no. 1, pp. 28–71, 1998.

[4] S. Redmond, "Visual cues in estimation of part-to-whole comparison," 2019.

[5] R. Kosara, "Evidence for area as the primary visual cue in pie charts," in *IEEE VIS*, 2019.

[6] A. Gaggioli and R. Breining, "Perception and cognition in immersive virtual reality," *Communications through virtual technology: Identity community and technology in the internet age*, pp. 71–86, 2001.

[7] M. Livingstone and D. Hubel, "Segregation of form, color, movement, and depth: anatomy, physiology, and perception," *Science*, vol. 240, no. 4853, pp. 740–749, 1988.

[8] M. S. Langer and H. H. Bülthoff, "Depth discrimination from shading under diffuse lighting," *Perception*, vol. 29, no. 6, pp. 649–660, 2000.

[9] J. F. Norman, J. T. Todd, V. J. Perotti, and J. S. Tittle, "The visual perception of three-dimensional length," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, no. 1, p. 173, 1996.

[10] M. Luboschik, P. Berger, and O. Staadt, "On spatial perception issues in augmented reality based immersive analytics," in *Proc. the ACM Companion on Interactive Surfaces and Spaces*, ser. ISS Companion '16. ACM, 2016, pp. 47–53.

[11] L. Wanger, "The effect of shadow quality on the perception of spatial relationships in computer generated imagery," in *Proc. the symposium on Interactive 3D graphics*. ACM, 1992, pp. 39–42.

[12] J. E. Cutting and P. M. Vishton, "Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth," in *Perception of space and motion*. Elsevier, 1995, pp. 69–117.

[13] S. S. Georgieva, J. T. Todd, R. Peeters, and G. A. Orban, "The extraction of 3d shape from texture and shading in the human brain," *Cerebral cortex*, vol. 18, no. 10, pp. 2416–2438, 2008.

[14] A. E. Welchman, A. Deubelius, V. Conrad, H. H. Bulthoff, and Z. Kourtzi, "3d shape perception from combined depth cues in human visual cortex," *Nature neuroscience*, vol. 8, no. 6, p. 820, 2005.

[15] J. T. Todd and J. F. Norman, "The visual perception of 3d shape from multiple cues: Are observers capable of perceiving metric structure?" *Perception & Psychophysics*, vol. 65, no. 1, pp. 31–47, 2003.

[16] B. Bach, R. Sicat, J. Beyer, M. Cordeil, and H. Pfister, "The hologram in my hand: How effective is interactive exploration of 3d visualizations in immersive tangible augmented reality?" *IEEE TVCG*, vol. 24, no. 1, pp. 457–467, 2018.

[17] N. Greffard, F. Picarougne, and P. Kuntz, "Beyond the classical monoscopic 3d in graph analytics: An experimental study of the impact of stereoscopy," in *Workshop on 3DVis*, 2014, pp. 19–24.

[18] M. Kraus, N. Weiler, D. Oelke, J. Kehrer, D. A. Keim, and J. Fuchs, "The impact of immersion on cluster identification tasks," *IEEE TVCG*, 2019.

[19] J. A. Wagner Filho, M. F. Rey, C. M. Freitas, and L. Nedel, "Immersive visualization of abstract information: An evaluation on dimensionally-reduced data scatterplots," in *IEEE VR*, vol. 2, no. 3, 2018, p. 4.

[20] R. Etemadpour, E. Monson, and L. Linsen, "The effect of stereoscopic immersive environments on projection-based multi-dimensional data visualization," in *International Conference on Information Visualisation*, 2013, pp. 389–397.

[21] C. Diaz, M. Walker, D. A. Szafir, and D. Szafir, "Designing for depth perceptions in augmented reality," in *IEEE ISMAR*, 2017, pp. 111–122.

[22] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri, "Four types of ensemble coding in data visualizations," *Journal of vision*, vol. 16, no. 5, pp. 11–11, 2016.

[23] A. Sarikaya and M. Gleicher, "Scatterplots: Tasks, data, and designs," *IEEE TVCG*, vol. 24, no. 1, pp. 402–412, Jan 2018.

[24] R. A. Rensink and G. Baldridge, "The perception of correlation in scatterplots," in *Computer Graphics Forum*, vol. 29, no. 3. Wiley Online Library, 2010, pp. 1203–1210.

[25] L. Micallef, G. Palmas, A. Oulasvirta, and T. Weinkauf, "Towards perceptual optimization of the visual design of scatterplots," *IEEE TVCG*, vol. 23, no. 6, pp. 1588–1599, 2017.

[26] B. Saket, A. Endert, and C. Demiralp, "Task-based effectiveness of basic visualizations," *IEEE TVCG*, pp. 1–1, 2018.

[27] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory, "A taxonomy of visual cluster separation factors," *Computer Graphics Forum*, vol. 31, no. 3pt4, pp. 1335–1344, 2012.

[28] M. Sedlmair, T. Munzner, and M. Tory, "Empirical guidance on scatterplot and dimension reduction technique choices," *IEEE TVCG*, vol. 19, no. 12, pp. 2634–2643, 2013.

[29] Y. Wang, K. Feng, X. Chu, J. Zhang, C. Fu, M. Sedlmair, X. Yu, and B. Chen, "A perception-driven approach to supervised dimensionality reduction for visualization," *IEEE TVCG*, vol. 24, no. 5, pp. 1828–1840, 2018.

[30] M. Brehmer, M. Sedlmair, S. Ingram, and T. Munzner, "Visualizing dimensionally-reduced data: Interviews with analysts and a characterization of task sequences," in *Proc. BELIV*, ser. BELIV. ACM, 2014, pp. 1–8.

[31] R. Etemadpour, R. C. da Motta, J. G. de Souza Paiva, R. Minghim, M. C. F. de Oliveira, and L. Linsen, "Role of human perception in cluster-based visual analysis of multidimensional data projections," in *International Conference on Information Visualization Theory and Applications*, 2014, pp. 276–283.

[32] Y. Wang, X. Chen, T. Ge, C. Bao, M. Sedlmair, C. Fu, O. Deussen, and B. Chen, "Optimizing color assignment for perception of class separability in multiclass scatterplots," *IEEE TVCG*, pp. 1–1, 2018.

[33] D. Smilkov, N. Thorat, and C. Nicholson, "Embedding projector - visualization of high-dimensional data." [Online]. Available: https://projector.tensorflow.org/

[34] A. O. Artero and M. C. F. de Oliveira, "Viz3d: effective exploratory visualization of large multidimensional data sets," in *Proc. Brazilian Symposium on Computer Graphics and Image Processing*, 2004, pp. 340–347.

[35] J. Poco, R. Etemadpour, F. Paulovich, T. Long, P. Rosenthal, M. Oliveira, L. Linsen, and R. Minghim, "A framework for exploring multidimensional data with 3d projections," *Computer Graphics Forum*, vol. 30, no. 3, pp. 1111–1120, 2011.

[36] A. Gracia, S. González, V. Robles, E. Menasalvas, and T. von Landesberger, "New insights into the suitability of the third dimension for visualizing multivariate/multidimensional data: A study based on loss of quality quantification," *Information Visualization*, vol. 15, no. 1, pp. 3–30, 2016.

[37] B. Wang and K. Mueller, "Does 3d really make sense for visual cluster analysis? yes!" in *Workshop on 3DVis*, 2014, pp. 37–44.

[38] A. Batch, A. Cunningham, M. Cordeil, N. Elmqvist, T. Dwyer, B. H. Thomas, and K. Marriott, "There is no spoon: Evaluating performance, space use, and presence with expert domain users in immersive analytics," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 536–546, 2019.

[39] M. Simpson, J. Zhao, and A. Klippel, "Take a walk: Evaluating movement types for data visualization in immersive virtual reality," in *Immersive Workshop at IEEE VIS*, 2017.

[40] D. Raja, D. Bowman, J. Lucas, and C. North, "Exploring the benefits of immersion in abstract information visualization," in *Proc. Immersive Projection Technology Workshop*, 2004, pp. 61–69.

[41] D. Raja, "The effects of immersion on 3d information visualization," Master's thesis, Virginia Tech, 2006.

[42] R. Rosenbaum, J. Bottleson, Z. Liu, and B. Hamann, "Involve me and i will understand!–abstract data visualization in immersive environments," in *International Symposium on Visual Computing*. Springer, 2011, pp. 530–540.

[43] A. Prouzeau, M. Cordeil, C. Robin, B. Ens, B. H. Thomas, and T. Dwyer, "Scaptics and highlight-planes: Immersive interaction techniques for finding occluded features in 3d scatterplots," in *SIGCHI*. ACM, 2019, p. 325.

[44] R. Sicat, J. Li, J. Choi, M. Cordeil, W. Jeong, B. Bach, and H. Pfister, "Dxr: A toolkit for building immersive data visualizations," *IEEE TVCG*, vol. 25, no. 1, pp. 715–725, 2019.

[45] J. Staib, S. Grottel, and S. Gumhold, "Visualization of particle-based data with transparency and ambient occlusion," in *Computer Graphics Forum*, vol. 34, no. 3. Wiley Online Library, 2015, pp. 151–160.

[46] C. P. Gribble and S. G. Parker, "Enhancing interactive particle visualization with advanced shading models," in *Proceed of the symposium on Applied perception in graphics and visualization*. ACM, 2006, pp. 111–118.

[47] B. Kozlíková, M. Krone, M. Falk, N. Lindow, M. Baaden, D. Baum, I. Viola, J. Parulek, and H.-C. Hege, "Visualization of biomolecular structures: State of the art revisited," in *Computer Graphics Forum*, vol. 36, no. 8. Wiley Online Library, 2017, pp. 178–204.

[48] Y. Yang, M. Cordeil, J. Beyer, T. Dwyer, K. Marriott, and H. Pfister, "Embodied navigation in immersive abstract data visualization: Is overview+detail or zooming better for 3d scatterplots?" *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[49] M. Whitlock, S. Smart, and D. A. Szafir, "Graphical perception for immersive analytics," in *IEEE VR*, 2020.

[50] D. J. Chalmers, R. M. French, and D. R. Hofstadter, "High-level perception, representation, and analogy: A critique of artificial intelligence methodology," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 4, no. 3, pp. 185–211, 1992.

[51] P. E. Rauber, A. X. Falcão, and A. C. Telea, "Projections as visual aids for classification system design," *Information Visualization*, vol. 17, no. 4, pp. 282–305, 2018.

[52] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. P. Chau, "Activis: Visual exploration of industry-scale deep neural network models," *IEEE TVCG*, vol. 24, no. 1, pp. 88–97, 2018.

[53] P. E. Rauber, S. G. Fadel, A. X. Falcao, and A. C. Telea, "Visualizing the hidden activity of artificial neural networks," *IEEE TVCG*, vol. 23, no. 1, pp. 101–110, 2017.

[54] J. Trommershauser, K. Kording, and M. S. Landy, *Sensory cue integration*. Oxford University Press, 2011.

[55] C. Andrade, "Internal, external, and ecological validity in research design, conduct, and evaluation," *Indian journal of psychological medicine*, vol. 40, no. 5, pp. 498–499, 2018.

[56] R. McDermott, "Internal and external validity," *Cambridge handbook of experimental political science*, p. 27, 2011.

[57] J. Hernández-Orallo, P. Flach, and C. Ferri, "A unified view of performance metrics: translating threshold choice into expected classification loss," *Journal of Machine Learning Research*, vol. 13, no. Oct, pp. 2813–2869, 2012.

[58] J. P. McIntire, P. R. Havig, and E. E. Geiselman, "What is 3d good for? a review of human performance on stereoscopic 3d displays," *Proc. SPIE*, vol. 8383, pp. 8383 – 8383 – 13, 2012.

[59] J. P. McIntire and K. K. Liggett, "The (possible) utility of stereoscopic 3d displays for information visualization: The good, the bad, and the ugly," in *Workshop on 3DVis*, 2014, pp. 1–9.

[60] J. P. McIntire, P. R. Havig, and E. E. Geiselman, "Stereoscopic 3d displays and human performance: A comprehensive review," *Displays*, vol. 35, no. 1, pp. 18 – 26, 2014.

[61] M. H. van Beurden, G. Van Hoey, H. Hatzakis, and W. A. Ijsselsteijn, "Stereoscopic displays in medical domains: a review of perception and performance effects," in *Human Vision and Electronic Imaging XIV*, vol. 7240. International Society for Optics and Photonics, 2009, pp. 0–1.

[62] D. R. Melmoth and S. Grant, "Advantages of binocular vision for the control of reaching and grasping," *Experimental Brain Research*, vol. 171, no. 3, pp. 371–388, 2006.

[63] X. Luo, R. Kenyon, D. Kamper, D. Sandin, and T. DeFanti, "The effects of scene complexity, stereovision, and motion parallax on size constancy in a virtual environment," in *IEEE VR*. IEEE, 2007, pp. 59–66.

[64] A. Forsberg, M. Slater, K. Wharton, Prabhat, and M. Katzourin, "A comparative study of desktop, fishtank, and cave systems for the exploration of volume rendered confocal data sets," *IEEE TVCG*, vol. 14, pp. 551–563, 2007.

[65] M. H. Van Beurden, W. A. Ijsselsteijn, and Y. A. De Kort, "Evaluating stereoscopic displays: both efficiency measures and perceived workload sensitive to manipulations in binocular disparity," in *Stereoscopic Displays and Applications XXII*, vol. 7863. International Society for Optics and Photonics, 2011, p. 786316.

[66] Y. Bastanlar, D. Canturk, and H. Karacan, "Effects of color-multiplex stereoscopic view on memory and navigation," in *IEEE 3DTV*, 2007, pp. 1–4.

[67] C. A. Ntuen, M. Goings, M. Reddin, and K. Holmes, "Comparison between 2-d & 3-d using an autostereoscopic display: The effects of viewing field and illumination on performance and visual fatigue," *International Journal of Industrial Ergonomics*, vol. 39, no. 2, pp. 388–395, 2009.

[68] Y. Aitsiselmi and N. Holliman, "Using mental rotation to evaluate the benefits of stereoscopic displays," in *Stereoscopic Displays and Applications XX*, vol. 7237. International Society for Optics and Photonics, 2009, pp. 0–1.

[69] P. Willemsen, A. A. Gooch, W. B. Thompson, and S. H. Creem-Regehr, "Effects of stereo viewing conditions on distance perception in virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 17, no. 1, pp. 91–101, 2008.

[70] A. Price and H.-S. Lee, "The effect of two-dimensional and stereoscopic presentation on middle school students' performance of spatial cognition tasks," *Journal of Science Education and Technology*, vol. 19, no. 1, pp. 90–103, 2010.

[71] J. P. McIntire, P. R. Havig, L. K. Harrington, S. T. Wright, S. N. Watamaniuk, and E. L. Heft, "Clinically normal stereopsis does not ensure a performance benefit from stereoscopic 3d depth cues," *3D Research*, vol. 5, no. 3, p. 20, 2014.

[72] K. Kihara, H. Fujisaki, S. Ohtsuka, M. Miyao, J. Shimamura, H. Arai, and Y. Taniguchi, "Age differences in the use of binocular disparity and pictorial depth cues in 3d-graphics environments," in *SID Symposium Digest of Technical Papers*, vol. 44, no. 1. Wiley Online Library, 2013, pp. 501–504.

[73] H. Fujisaki, H. Yamashita, K. Kihara, and S. Ohtsuka, "Individual differences in the use of binocular and monocular depth cues in 3d-graphic environments," in *SID Symposium Digest of Technical Papers*, vol. 43, no. 1. Wiley Online Library, 2012, pp. 1190–1193.

[74] C. Ware and P. Mitchell, "Reevaluating stereo and motion cues for visualizing graphs in three dimensions," in *ACM APGV*, ser. APGV '05, 2005, pp. 51–58.

[75] O.-H. Kwon, C. Muelder, K. Lee, and K.-L. Ma, "A study of layout, rendering, and interaction methods for immersive graph visualization," *IEEE TVCG*, vol. 22, no. 7, pp. 1802–1815, 2016.

[76] B. Alper, T. Hollerer, J. Kuchera-Morin, and A. Forbes, "Stereoscopic highlighting: 2d graph visualization on stereo displays," *IEEE TVCG*, vol. 17, no. 12, pp. 2325–2333, 2011.

[77] D. Belcher, M. Billinghurst, S. Hayes, and R. Stiles, "Using augmented reality for visualizing complex graphs in three dimensions," in *Mixed and Augmented Reality, 2003. Proceedings. The Second IEEE and ACM International Symposium on*. IEEE, 2003, pp. 84–93.

[78] J. A. Wagner Filho, C. M. Freitas, and L. Nedel, "Virtualdesk: a comfortable and efficient immersive information visualization approach," in *Computer Graphics Forum*, vol. 37, no. 3. Wiley Online Library, 2018, pp. 415–426.

[79] M. Kraus, N. Weiler, D. A. Keim, A. Diehl, and B. Bach, "Visualization in the vr-canvas: How much reality is good for immersive analytics in virtual reality?" in *Workshop on the Creation, Curation, Critique and Conditioning of Principles and Guidelines in Visualization*, 2018.

[80] J. J. Cummings and J. N. Bailenson, "How immersive is enough? a meta-analysis of the effect of immersive technology on user presence," *Media Psychology*, vol. 19, no. 2, pp. 272–309, 2016.

[81] M. H. P. H. van Beurden, A. Kuijsters, and W. A. IJsselsteijn, "Performance of a path tracing task using stereoscopic and motion based depth cues," in *Workshop on Quality of Multimedia Experience*, 2010, pp. 176–181.

[82] B. W. van Schooten, E. M. A. G. van Dijk, E. Zudilova-Seinstra, A. Suinesiaputra, and J. H. C. Reiber, "The effect of stereoscopy and motion cues on 3d interpretation task performance," in *ACM AVI*, 2010, pp. 167–170.

[83] W. Barfield, C. Hendrix, and K.-E. Bystrom, "Effects of stereopsis and head tracking on performance using desktop virtual environment displays," *Presence: Teleoperators & Virtual Environments*, vol. 8, no. 2, pp. 237–240, 1999.

[84] A. E. Patla, E. Niechwiej, V. Racco, and M. A. Goodale, "Understanding the contribution of binocular vision to the control of adaptive locomotion," *Experimental Brain Research*, vol. 142, no. 4, pp. 551–561, 2002.

[85] S. Grottel, M. Krone, K. Scharnowski, and T. Ertl, "Object-space ambient occlusion for molecular dynamics," in *IEEE PacificVis*, 2012, pp. 209–216.

[86] N. Tatarchuk, "Advances in real-time rendering in 3d graphics and games i," in *ACM SIGGRAPH 2009 Courses*. ACM, 2009, p. 4.

[87] C. Lee, G. A. Rincon, G. Meyer, T. Höllerer, and D. A. Bowman, "The effects of visual realism on search tasks in mixed reality simulation," *IEEE TVCG*, vol. 19, no. 4, pp. 547–556, 2013.

[88] C. Stinson, R. Kopper, B. Scerbo, E. Ragan, and D. Bowman, "The effects of visual realism on training transfer in immersive virtual environments," in *Human Systems Integration Symposium*, 2011.

[89] N. Gershon and S. G. Eick, "Information visualization," *IEEE Computer Graphics and Applications*, no. 4, pp. 29–31, 1997.

[90] Q. C. Vuong, F. Domini, and C. Caudek, "Disparity and shading cues cooperate for surface interpolation," *Perception*, vol. 35, no. 2, pp. 145–155, 2006.

[91] S. V. Bemis, J. L. Leeds, and E. A. Winer, "Operator performance as a function of type of display: Conventional versus perspective," *Human Factors*, vol. 30, no. 2, pp. 163–169, 1988.

[92] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[93] M. Tory, D. Sprague, F. Wu, W. Y. So, and T. Munzner, "Spatialization design: Comparing points and landscapes," *IEEE TVCG*, vol. 13, no. 6, pp. 1262–1269, 2007.

[94] M. Tory, C. Swindells, and R. Dreezer, "Comparing dot and landscape spatializations for visual memory differences," *IEEE TVCG*, vol. 15, no. 6, pp. 1033–1040, 2009.

[95] Y. Yang, T. Dwyer, B. Jenny, K. Marriott, M. Cordeil, and H. Chen, "Origin-destination flow maps in immersive environments," *IEEE TVCG*, vol. 25, no. 1, pp. 693–703, 2019.

[96] D. Drascic and P. Milgram, "Perceptual issues in augmented reality," in *Stereoscopic displays and virtual reality systems III*, vol. 2653. International Society for Optics and Photonics, 1996, pp. 123–135.

[97] F. Scheer and M. Keutel, "Screen space ambient occlusion for virtual and mixed reality factory planning," 2010.

[98] N. Farahani, R. Post, J. Duboy, I. Ahmed, B. J. Kolowitz, T. Krinchai, S. E. Monaco, J. L. Fine, D. J. Hartman, and L. Pantanowitz,

"Exploring virtual reality technology and the oculus rift for the examination of digital pathology slides," *Journal of Pathology Informatics*, vol. 7, 2016.
[99] J. Weston, A. Bordes, S. Chopra, and T. Mikolov, "Towards ai-complete question answering: A set of prerequisite toy tasks," *CoRR*, vol. abs/1502.05698, 2015.
[100] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "Weakly supervised memory networks," *CoRR*, vol. abs/1503.08895, 2015. [Online]. Available: http://arxiv.org/abs/1503.08895
[101] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)." [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html
[102] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385
[103] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/
[104] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *CoRR*, vol. abs/1708.07747, 2017.
[105] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.
[106] M. Whitlock, E. Harnner, J. R. Brubaker, S. Kane, and D. A. Szafir, "Interacting with distant objects in augmented reality," in *IEEE VR*, 2018, pp. 41–48.
[107] R. B. Brady, D. J. Zielinski, D. A. Bowman, and R. P. McMahan, "Evaluating display fidelity and interaction fidelity in a virtual reality game," *IEEE TVCG*, vol. 18, pp. 626–633, 2012.
[108] K. Shoemake, "Arcball: a user interface for specifying three-dimensional orientation using a mouse," in *Graphics Interface*, vol. 92, 1992, pp. 151–156.
[109] J.-P. Huttner and R.-B. Susanne, "An immersive memory palace: supporting the method of loci with virtual reality," 2017.
[110] E. L. Legge, C. R. Madan, E. T. Ng, and J. B. Caplan, "Building a memory palace in minutes: Equivalent memory performance using virtual versus conventional environments with the method of loci," *Acta psychologica*, vol. 141, no. 3, pp. 380–390, 2012.
[111] R. E. Kirk, *Experimental design, 3rd Edition*. Wiley Online Library, 1995, pp. 512–515.
[112] M. G. Falleti, P. Maruff, A. Collie, and D. G. Darby, "Practice effects associated with the repeated assessment of cognitive function using the cogstate battery at 10-minute, one week and one month test-retest intervals," *Journal of Clinical and Experimental Neuropsychology*, vol. 28, no. 7, pp. 1095–1112, 2006.
[113] "Flashing grey screen," https://community.viveport.com/t5/Technical-Support/flashing-grey-screen/td-p/8607, accessed: 2019-03-25.
[114] J. McTighe and G. Wiggins, "What makes a question essential?" in *Essential questions: Opening doors to student understanding*. Ascd, 2013.
[115] E.-J. Wagenmakers, M. Marsman, T. Jamil, A. Ly, J. Verhagen, J. Love, R. Selker, Q. F. Gronau, M. Šmíra, S. Epskamp *et al.*, "Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications," *Psychonomic bulletin & review*, vol. 25, no. 1, pp. 35–57, 2018.
[116] M. Kay, G. L. Nelson, and E. B. Hekler, "Researcher-centered design of statistics: Why bayesian statistics better fit the culture and incentives of hci," in *SIGCHI*. ACM, 2016, pp. 4521–4532.
[117] A. Gelman, J. Hill, and M. Yajima, "Why we (usually) don't have to worry about multiple comparisons," *Journal of Research on Educational Effectiveness*, vol. 5, no. 2, pp. 189–211, 2012.
[118] B. Lambert, *A student's guide to Bayesian statistics*. Sage, 2018.
[119] D. J. Lewkowicz, "The concept of ecological validity: What are its limitations and is it bad to be invalid?" *Infancy*, vol. 2, no. 4, pp. 437–450, 2001.
[120] P.-C. Bürkner *et al.*, "brms: An r package for bayesian multilevel models using stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.
[121] G. Wilkinson and C. Rogers, "Symbolic description of factorial models for analysis of variance," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 22, no. 3, pp. 392–399, 1973.
[122] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, S. Heisterkamp, B. Van Willigen, and R. Maintainer, "Package 'nlme'," *Linear and nonlinear mixed effects models, version*, vol. 3, no. 1, 2017.
[123] J. P. Klein and M. L. Moeschberger, *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.
[124] Stan Development Team, "RStan: the R interface to Stan," 2018, r package version 2.18.2. [Online]. Available: http://mc-stan.org/
[125] M. Kay, *tidybayes: Tidy Data and Geoms for Bayesian Models*, 2019, r package version 1.0.4. [Online]. Available: http://mjskay.github.io/tidybayes/
[126] B. D. Ondov, F. Yang, M. Kay, N. Elmqvist, and S. Franconeri, "Revealing perceptual proxies with adversarial examples," *IEEE Transactions on Visualization and Computer Graphics*, 2020.
[127] L. Itti, "Models of bottom-up and top-down visual attention," Ph.D. dissertation, California Institute of Technology, 2000.
[128] H. Wässle, "Parallel processing in the mammalian retina," *Nature Reviews Neuroscience*, vol. 5, no. 10, pp. 747–757, 2004.
[129] W. Köhler, "Gestalt psychology." 1929.
[130] K. Koffka, *Principles of Gestalt psychology*. Routledge, 2013.
[131] H. H. Bülthoff and H. A. Mallot, "Integration of depth modules: stereo and shading," *Josa a*, vol. 5, no. 10, pp. 1749–1758, 1988.

PLACE PHOTO HERE

Fumeng Yang Biography omitted for review.

PLACE PHOTO HERE

James Tompkin Biography omitted for review.

PLACE PHOTO HERE

Lane Harrison Biography omitted for review.

PLACE PHOTO HERE

David H. Laidlaw Biography omitted for review.