Visual Cue Effects on a Classification Accuracy Estimation Task 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 two 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. With Bayesian multilevel modeling, we capture the intricate visual effects and find that no cue 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. Using an HMD, adding visual motion cues, providing a third data dimension, or showing a more complicated dataset leads to longer response times. We speculate that most visual cues may not strongly affect perception in immersive analytics unless they change people's mental model about data. In summary, by studying participants as they interpret the output from a complicated machine learning model, we advance our understanding of how to use the visual cues in 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 challenge here is assessing the effects of *visual cues*—global or local properties that perceptually prioritize objects or regions [2], [3]. There are two fundamental classes of visual cues: *primary cues* providing physiological percepts (e.g., stereopsis and accommodation), and *pictorial cues* used to depict 3D depth in 2D pictures (e.g., occlusion, perspective, texture, and shading) [2]. Different visual cues affect many aspects of visual perception and cognition, such as depth perception [4]–[6], spatial judgments [7], [8], and shape understanding [9]–[11]. As such, understanding visual cue effects in immersive analytics could facilitate the broader and more appropriate use of VR/AR techniques.

Prior work has not focused on visual cue effects in immersive analytics [12]–[14]. For example, some research used inconsistent colors and shading across the desktop monitor and VR [15], [16], and neglected to consider how pictorial cues affect depth and spatial perception. To further complicate matters, similar studies from other domains may not directly apply to information visualization. These studies relied on tasks like navigating a simulated 3D world or comparing a handful of virtual objects in an elementary task [17], which is categorically different from visualization tasks performed on thousands of visual elements representing different data points [8].

To this end, we explicitly study visual cue effects on people's task performance in immersive analytics. Specifically, we examined four visual cues: visual motion, perspective (graphic projection), shading, and dimensionality. We chose these cues because they added minimal visual clutter but showed strong effects on perception in the literature [1], [5], [7]. We compared these cues on both an immersive VR environment using a head-mounted display (HMD) and a non-VR environment using a desktop monitor, and on two sets of data (called "data models," see Sec. 2.2). These visual cues and datasets cover typical and common parameters of immersive visualizations.

We used scatterplots as the central visualization. Scatterplots support both low-level object-centric tasks [18] and high-level visual aggregation [18], [19]. They show data features such as correlation [20], anomalies [21], clusters [22], [23], and dimension-reduction results [24]–[27]. Scatterplots encode data points into spatial coordinates [28]–[30], making them suitable to examine visual cues where surrounding spatial coordinates can affect the perception of a data point. Scatterplots can also be shown across different devices such as a desktop monitor [31]–[34], VR [35]–[40], and AR [12], [41]. Thus, using scatterplots offer insights into a variety of visualization tasks and visual cue effects.

We designed a task in which participants assessed a neural network's classification performance from a scatterplot of the last hidden layer's outputs (see Fig. 1) [42]–[44]. These high-dimensional outputs contain necessary information about classification performance and were reduced to 2D or 3D space for interpretation. This task requires an interpretation of visualization and is a mid-level task:

- *Fumeng Yang is currently with Northwestern University, Evanston, IL. She conducted this study while she was at Brown University. E-mail: fy@northwestern.edu.*
- *James Tompkin and David H. Laidlaw are with the Department of Computer Science, Brown University, Providence, RI. E-mails: {james_tompkin, david_laidlaw}@brown.edu.*
- *Lane Harrison is with the Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA. Email: lharrison@wpi.edu.*

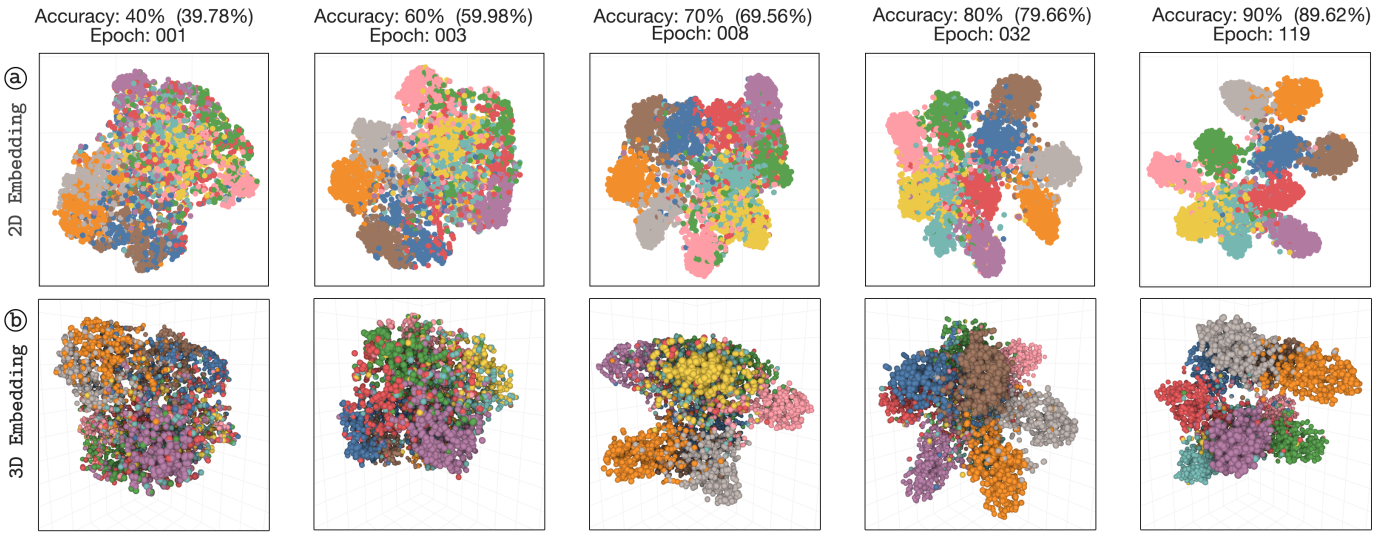

Figure 1. **Examples of last hidden layer outputs and classification accuracy**. In training the neural network, the visual properties of the last hidden layer outputs change systematically with the classification accuracy. ⓐ The top row shows examples of the experimental condition `Orthographic • Flat Shading • Image Data • 2D Embedding`. ⓑ The bottom row shows examples of the experimental condition `Perspective • Ambient Occlusion • Image Data • 3D Embedding`. Each column shows 2D or 3D embedding results for the same high-dimensional dataset.

it includes low-level perceptual processes (e.g., reading a value, judging groupings) that are possible to generalize, and also reflects real-world practitioners's analytical processes that are possible to be evaluated. Participants were simply asked to provide a number—classification accuracy—as their final assessment. This task connects the huge gamut of choices in visual-cue design to their effects on visualization interpretation.

There are two main research contributions in this paper:

Contribution 1 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.

Contribution 2 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 [45]. Visual cues interacted with each other in complicated and subtle ways, especially for estimation error.

Our experimental materials, data, and analysis scripts are posted at `https://doi.org/10.17605/OSF.IO/PKUVZ`.

2 Background and Motivation

2.1 Past studies on visual cues

A number of studies compared stereoscopic and non-stereoscopic devices. These reported mixed results for different tasks. Several surveys [46]–[49] agreed that a stereoscopic display may improve participants' performance over a desktop (e.g., [50]–[52]), especially for more difficult tasks [53]. Others reported mixed [54]–[58] or negative results [59], and the effects were subject to individual differences [60]–[62] and choice of tasks [46]–[48]. Yet few studies focused on visual cue effects.

Perceptual science also investigated visual cue effects extensively, but did not concentrate on visualization. Visual cues were found to affect many aspects of visual perception and cognition, like 2D visual comparison [63], [64], depth perception [4]–[6], 3D length [65], spatial relationships [2], [8], [66], spatial judgments [7], [8], and shape understanding [9]–[11]. Most of these studies examined a small number of objects (e.g., one mesh object). On the contrary, studies from computer graphics often cope with complex scenes, and improving shading is a general theme for showing more details and enhancing user experience [28]–[30], [67], [68]. These studies were conducted without regard to an analytical or reasoning task with users.

On visualization and visual analytics, previous studies compared different devices and reported moderate effects of visual cues. The studies on graph visualization showed that a VR environment might have positive [69], [70] or neutral effects [13], [71], [72] on task performance and completion time compared to a desktop monitor. Similar studies compared scatterplots across a desktop monitor and an HMD on tasks of selecting a cluster and identifying outliers, reporting mixed results [15], [16], [73]. These studies used inconsistent colors and shading across the desktop monitor and VR [15], [16]. Other studies have similar issues of using different visualization forms, like presenting 2D scatterplot matrices and 3D scatterplots for different devices [14]. A few other studies examined visual motion (interaction techniques) [12], [40], [74] and different encoding channels [58] in VR. However, these studies have not systematically investigated the impacts of visual cues, particularly pictorial cues applying to all data points, in a task of making sense of complex data [75].

2.2 Current study on visual cues

Given the literature, we selected two viewing devices, four visual cues, and two sets of data for our experiment. These variables and their manipulations typify commonly-seen visual cues.

Device Previous studies and surveys reported mixed results for comparison between stereoscopic and non-stereoscopic displays. It is necessary to test different devices to disen-

Figure 2. **Task overview**. In each trial, participants estimated classification performance from a scatterplot of the last hidden layer's output of a neural network.

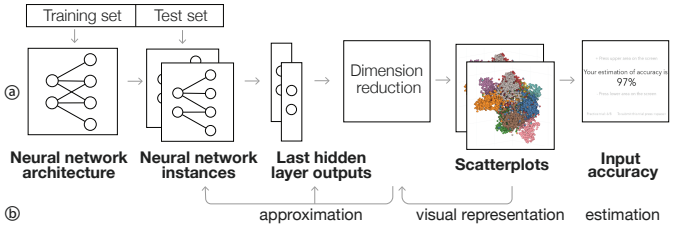

Figure 3. **Stimuli generation**. We first trained a neural network on the training set and saved the intermediate neural network after each epoch. For each intermediate neural network, we used the test set as the input and calculated its last hidden layer output, which was then dimension reduced and used to render a scatterplot. Participants were presented with the scatterplot and assessed the classification accuracy of the neural network on the test set. They press the mouse/controller to input a percentage as their answer, which was compared to the ground truth accuracy to derive their task performance (see Sec. 6.7).

tangle the intrinsic properties of an immersive environment (e.g., stereopsis).

Visual motion Motion alone shows stronger effects than stereoscopy alone [76], [77], but it is less effective than stereoscopy with head-tracking [51], [78]; motion and stereoscopy might be of the same effectiveness [79]. Motion allows people to vary their vantage point to gain more information, which may help make sense of dimension reduction results orienting differently in the low-dimensional space.

Shading Shading affects perception of shape, depth, and spatial relationships [4] in a 3D space. In an immersive environment, illumination improved completion time for graph path tracing [70], while visual realism might hurt task performance [80], [81]. Information visualization often shows abstract data [82] and commonly uses solid untextured colors. However, if data is in 3D space, varying shading may improve depth perception [83] and then task performance.

Graphical projection Previous studies also show the impact of different graphical projection methods [7], [84]. Perspective projection alters an object's depicted size with its distance from the projection center, often used in VR to generate stereoscopy and immersiveness to improve spatial judgment [2] and distance perception [7], [84]. Orthographic projection preserves size and uses 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 [22] for more precise presentation. Less studied in VR, the effect of orthographic projection on immersive analytics is unclear.

Dimensionality Reducing a high-dimensional dataset to a 3D space is necessary to generate an immersive visualization, and a 3D data projection is also often more precise than a 2D data projection (e.g., the Kullback-Leibler divergence [85] is smaller). However, visualization in a 3D space often causes perception discrepancies (e.g., [42]–[44]). To precisely understand visual cue effects and to establish a fair baseline, the present study incorporated both 2D and 3D datasets.

Data model To understand the generalizability of the results, we considered different data properties. Data properties like the number of data points [86], [87], the complexity of a graph visualization (e.g., the number of nodes) [69], [88] or a scene [81], and the shape of clusters [89] may affect task performance. We termed the collection of these properties "data model" to encompass the differences in data distribution, complexity, cluster shape, and other data properties (e.g., Figs. 1 and 6).

3 Task Design

To derive a measurable task to provide insights into immersive analytics, we drew inspiration from recent successes that examined the last hidden layer output of a neural network [42] to identify misclassified instances, training effects, and hidden structures [43], [44]. Visualizing the last hidden layer's output conveys the model's internal representations and its likely classification performance. As such, the task we chose was for participants to estimate a model's classification accuracy based on a scatterplot of the last hidden layer's output (see Figs. 2 and 3). This task combines low-level perception with more open-ended high-level model assessments, and lands on a mid level.

3.1 Justification

The task supports internal, construct, external, and ecological validity [90], [91] in the following ways.

This task reflects how people perceive classification results, and therefore helps establish a relationship between visual cues and classification perception. To measure the perceived classification performance, we sought a quantitative metric (e.g., accuracy, precision, and loss [92]) and used accuracy for simplicity; accuracy is defined as the percentage of correctly classified instances.

To ensure construct validity, we conducted a pilot study and confirmed that participants could estimate accuracies near the true classification accuracy. Further, we requested participants have at least passing knowledge of machine learning and visualization (Sec. 6.6) and included both training and practice sessions to ascertain an association between classification accuracy and scatterplots (Sec. 6.3).

This task supports external validity because it resembles cluster perception in multi-class scatterplots. Assessing classification performance visually can be considered cluster detection under uncertainty. Factors such as data properties (e.g., outliers and cluster distance), cognition (e.g., expertise), and primarily perceptual components (e.g., visual cues) could systematically affect participants' estimates.

This task also partly supports ecological validity because it could be an example of similar tasks that users plausibly undertake in real-world visual analytics. While estimating classification accuracy is a manufactured task, inspecting the last hidden layers' outputs facilitates the understanding

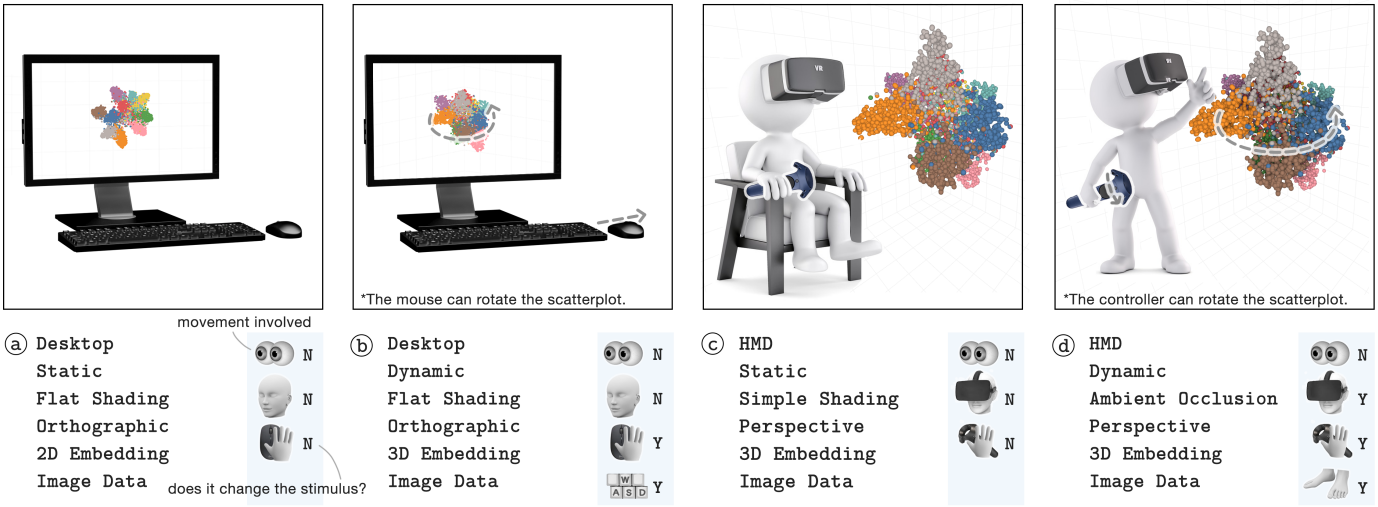

Figure 4. **Device, visual motion, and movements involved**. We designed both `Desktop` (ⓐⓑ) and `HMD` (ⓒⓓ) conditions. Participants saw fixed scatterplots in `Static` conditions (ⓐⓒ); and in `Dynamic` conditions, they interacted and rotated the scatterplots via a mouse and a keyboard (`Desktop`, ⓑ)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

of the neural network's performance and assists in model diagnostics [42]–[44].

3.2 Task limitations

Designing psychophysical tasks that elicit measurable and separable responses is an ongoing challenge for immersive analytics [93], which we aim to address in part through this study. We reduce a realistic task combining multiple facets of perceiving clusters. However, this realism also brings costs to the experimental framework in terms of indirect or difficult-to-control factors. These indirections may span from data generation to participants' interpretation. Approximation errors in each step may propagate along the visual inspection pipeline: in the present task, the model's internal representation is approximated by the high-dimensional last hidden layer's output, which is further approximated by dimension-reduction result, which is then visually encoded into a scatterplot and presented to error-prone participants (Fig. 3b). These intermediate steps combined with visual cues under question are likely to affect participants' estimation accuracy in complex ways. While it is difficult, if not impossible, to completely separate and gauge these indirect effects, the task can measure participants' performance with approximation error.

4 Visual Cues

As described in Sec. 2.2, we manipulated six variables in total: (1) `device`, (2) `visual motion`, (3) `graphical projection`, (4) `shading`, (5) `dimensionality`, and (6) `data model`. We also discussed all possible visual cues collected from the literature as a table in Appx. C.

4.1 Device

We used two `devices` (see Fig. 4): a desktop with a monitor, and the same desktop with an HMD (HTC Vive Pro), denoted as `Desktop` and `HMD`, respectively. Certain visual cues only present in an immersive environment (e.g., immersiveness and presence [94], [95]); and we regarded them as contextual cues for the immersive environment.

4.2 Visual motion

We had two visual motion levels: the first level used static stimuli, denoted by `Static Stimuli` (or `Static`); and the second level allowed participants to update the stimuli via movement and rotation, denoted by `Dynamic Stimuli` (or `Dynamic`). The difference between `Static` and `Dynamic` is if participants can alter a stimulus on the display. In `Static` conditions, participants saw static images on both `Desktop` and `HMD`. They sat in a chair when using the HMD (Fig. 4c). This discouraged moving and avoided further issues arising from motion sickness or nausea. In `Dynamic` conditions, participants varied their camera position using a keyboard and a mouse (`Desktop`, Fig. 4b) or via head tracking and walking in the room (`HMD`, Fig. 4d). Participants could also rotate a scatterplot around its center by dragging a mouse (`Desktop`, Fig. 4b) or moving a VR controller (`HMD`, Fig. 4d).

Discussion Our experiment focused on visual motion cues that could trigger a change in the visualization. Motion as a sensory cue can be detected beyond the human vision system; for example, walking and moving one's head may activate the proprioceptive and vestibular systems. Proprioceptive and vestibular systems were unlikely to alter the stimuli nor consequently affect their appearance. We ignored these *non-visual motion effects* (e.g., proprioceptive and vestibular) which are not directly related to a visualization task. It might be impossible to eliminate these intrinsic properties of an HMD or to approximate them precisely on a desktop. However, we could still quantify the plausible effects of these additional non-visual cues by measuring each combination of `device` and `visual motion`, and their effects were captured as interaction effects between the two variables. Also, motion as a depth cue can be generated from *motion parallax*, *motion perspective*, or *kinetic effect* [96], and both `Desktop` and `HMD` can provide all three.

4.3 Shading

Our experiment had three `shading` levels: the first level used solid colors, denoted as `Flat Shading` (Fig. 5c). The second

level utilized the commonly-used Phong lighting model (the ambient and diffuse terms), denoted as `Simple Shading` (Fig. 5d). The third level used the Phong lighting model with ambient occlusion, a more advanced rendering technique [17], denoted as `Ambient Occlusion` or `A.O.` (Fig. 5e). Here `Ambient Occlusion` conditions implemented *screen-space ambient occlusion* (SSAO) [30] and normal reconstruction from the depth buffer. Many techniques have been developed to compute ambient occlusion, especially for scatterplots (or particle visualization) [30], 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].

4.4 Graphical projection

Our experiment explored both orthographic and perspective projections, denoted by `Orthographic` and `Perspective`, respectively (see Fig. 6). The implementation of perspective projection was straightforward. The implementation of orthographic was conducted with preservation of binocular disparity in VR, including an additional perspective projection. Further explanations for this implementation are available in Appx. A.

4.5 Dimensionality

We used the same t-Distributed Stochastic Neighbor Embedding (t-SNE) procedure to reduce high-dimensional hidden layer outputs to `2D Embedding` and `3D Embedding` (see Sec. 5.1). The `2D Embedding` results are semantically similar to the `3D Embedding` results (Figs. 1a-b, 6a-b), sharing the same visual properties such as cluster shape and local structure.

Discussion This comparison across 2D and 3D embeddings were similar to the study by Sedlmair et al. [22], and it improved compatibility between the two devices, compared to the approach of using scatterplot matrices as a baseline by Kraus et al. [14]. Additionally, `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. 1a and 6a). A 2D dataset does not have the third dimension for computing depth. It was possible to add a synthetic third dimension to a 2D dataset (e.g., setting z to 0), yet this caused serious z-buffer fighting or strongly hinted the drawing order as a third, non-existent dimension. Thus, we ruled out these other alternatives to avoid confusing participants.

4.6 Data model

Within this study, a data model consists of a training set and a neural network architecture; optimizing such a pair produces a set of hidden-layer embeddings with various classification accuracies (Fig. 3).

We used two data models: (1) `Text Data`, defined in training the text dataset bAbI [99] on a memory neural network (MemNN) [100], and (2) `Image Data`, defined in training the image dataset CIFAR-10 [101] on a residual neural network [102]. For `Text Data`, we used the training set with 10,000 samples (called "single supporting fact") and 200 epochs; we used the test set of 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. The statistics of classification accuracies are available in Figs. 7b and c.

Discussion 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 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. 7a).

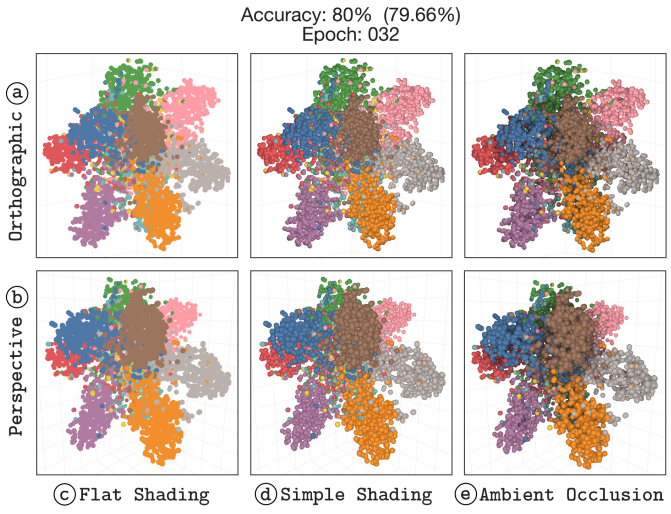

Figure 5. **Shading and graphic projection**. ⓐ The top row shows examples of `Orthographic` projection for three different `shading` levels: `Flat Shading`, `Simple Shading`, and `Ambient Occlusion`. ⓑ The bottom row shows examples of `Perspective` projection. All six scatterplots show the same underlying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

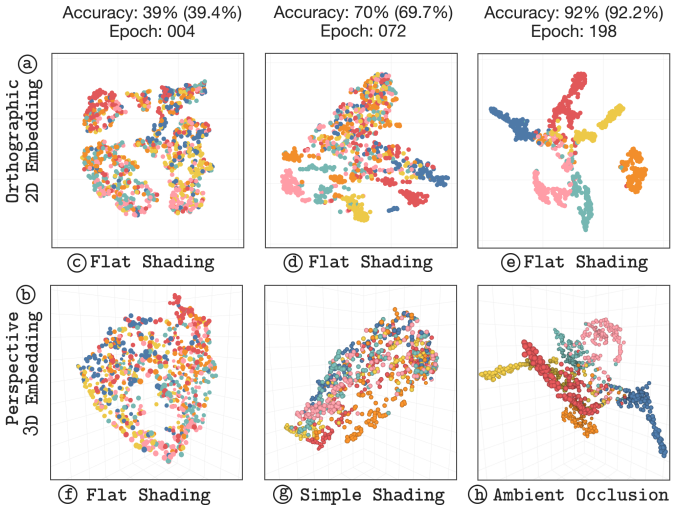

Figure 6. **Examples of `Text Data`**. ⓐ The top row shows examples of 2D `Embedding` across different accuracy levels. ⓑ The bottom row shows examples of 3D `Embedding` of the same hidden layer output across different accuracy and shading levels.

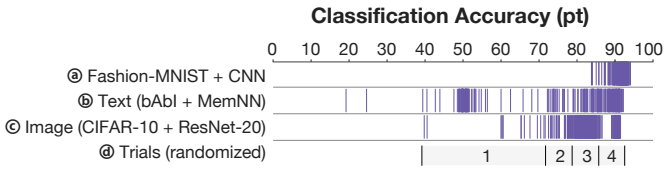

Figure 7. **Classification accuracy ranges of different data models**. ⓐ The first row shows that a common dataset and neural network architecture generate only very high classification accuracies. ⓑ-ⓒ The second and third rows show that the classification accuracy ranges of the two data models used in this study, and ⓓ the last row shows our sampling process to select four trials for each condition.

5 Control Variables

Other differences between a desktop monitor and an immersive environment may also affect participant's task performance. We inspected and controlled a set of *control variables*. Here we discuss three: (1) `dimension-reduction process`, (2) `visual content`, and (3) `interactions`. We believe these three variables were most important for our experiment and would be interesting to readers.

5.1 Dimension-reduction process

As described in Sec. 4.5, we used the prevalent t-SNE algorithm to reduce the dimensionality of the last hidden layers' outputs. t-SNE preserves cluster information and the local structure of the original dataset. The results of 2D and 3D t-SNE on the same data are the *optimal* approximations in their own low-dimensional spaces constrained by the same distance function. We generated all the datasets using the same parameter settings (perplexity $= 30$, learning rate $= 150$, steps $= 600$, and fixed random seeds) that guarantee convergence. We also manually reviewed all the resulting datasets for convergence and confirmed that the layout of a 2D dataset can be empirically viewed as the corresponding flattened 3D dataset. All the datasets and images are available in supplementary materials.

We had evaluated other dimension reduction techniques, such as Uniform Manifold Approximation and Projection (UMAP) [105] and Principal Component Analysis (PCA), but t-SNE was the best fit. UMAP preserves more global structure than t-SNE, but it arranges data points in the same class very close to each other, which causes cluttering. Similarly, PCA resulted in indistinguishable clusters for both data models.

5.2 Visual content

Our experimental system preserves 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 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.

5.3 Interactions

We also calibrated *interaction* techniques, important for understanding space and depth in an immersive environment [12], [106], [107]. Interaction techniques support visual motion cues (Sec. 4.2). In our experiment, rotation interaction provides *motion parallax*; head movement, physical, or virtual navigation provides *motion perspective*; each provides *kinetic depth perception*. Both devices ought to support these in a consistent manner, described as follows.

Metaphor The metaphor was the same in all conditions: participants moved the mouse on the xz-plane in `Desktop` conditions (Fig. 4b); or they moved the controller on the xy-plane in `HMD` conditions (Fig. 4d).

Rotation The rotation operations were implemented based on the Arcball technique [108] and movements. `Desktop` conditions allowed mouse movements (Fig. 4b), and `HMD` conditions used the movements projected on the 2D view plane (Fig. 4d) as the analogy. We chose a scale factor so that a long stroke drawn using the VR controller roughly matched a long stroke drawn using the mouse for an average-sized participant.

Navigation We approximated head movement and physical navigation in VR for a desktop monitor. We allowed virtual movement where participants could move along the xz-plane using a keyboard (the WASD keys, see Fig. 4b), which was common in the research across modalities [109], [110]. One keystroke was mapped to a change of 0.02 units, resulting in a speed of 0.6 units per second if a participant was constantly holding the key ($\sim$30Hz); this corresponded to a speed of .60 meters per second in VR, similar to a slow walking speed for an average-sized participant.

6 Study Design

With the task design (Sec. 3) and the experimental variables (Secs. 4 and 5), our research question is

> **Research question**: What are the effects of visual cues and the relationships between them in a task of visually assessing classification performance?

We reflected this research question in our experimental design as follows.

6.1 Experimental design

We used a mixed factorial design with repeated measures. A mixed design is one of the basic and the most widely used designs in human-subjects experiments [111]. The between-subjects variable was `data model`, and each participant finished `Text Data` or `Image Data`. The other four variables were within-subjects and assigned in the following order: `device`, `dimensionality`, `graphical projection`, `shading`, and all participants finished all valid combinations of them.

All the combinations of the visual cues and devices were compatible with `3D Embedding`, resulting in 48 different experimental conditions $= 2$ `devices` $\times 2$ `motion` levels $\times$ 3 `shading` levels $\times 2$ `graphical projection` methods $\times 2$ `data models`. There were four `2D Embedding` conditions for different devices and datasets, resulting in $52 = (48 + 4)$ conditions (combinations). For example combinations, see Fig. 1 or 4 above; for a list of all combinations, see Fig. 9 below. Each participant finished 26 combinations.

This mixed design was favored over 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. 7); (2) using the same `data model` throughout hundreds of trials may lead to strong learning and practice effects [112]. This design leveraged the within-subjects component to reduce the impacts of individual differences, allowed an inference of interaction effects and required fewer participants. It partially captured uncertainty and the relationships between visual cues.

6.2 Procedure

Participants started with a consent form and then took part in the introduction session with one and the same experimenter. In the experiment, they first saw an overview 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 to record their strategies (e.g., "Any visual features or patterns you looked for?") and additional comments. After completing all 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 took 60 to 90 minutes to accomplish the experiment.

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 adjusted a number (0 to 100) using the mouse or the controller to provide their answers. In `HMD` conditions, participants were asked to move back to the initial position before a new trial.

6.3 Training and practice

To ensure that participants understood the task and met our expectations, we included introduction and training sessions at the beginning of the experiment and 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...For each scatterplot, we ask you to estimate the classification accuracy for that dataset."* Participants were also presented with eight examples of four accuracy levels. The full instructions and the examples are available in Appx. D.

At the beginning of an estimation session, participants practiced each assigned condition once but twice for `2D Embedding` because there were fewer `2D Embedding` conditions. Practice trials were grouped and conduced in the same fashion as the main trials, sampling from the datasets similarly (Fig. 7d), but without the true classification accuracy at the end of the trial.

6.4 Experimental setup

The scale of the immersive environment was about 3.10 $\times$ 3.00 meters in a physical room of 6.00 $\times$ 4.40 meters; the room was constantly quiet and had an unobstructed floor. The near and far clipping planes were fixed to 0.018 and 75.0, respectively, so that all the visual elements were visible, of a similar size, and similar to an overview [74] when viewed at the default (initial) camera position. The default (initial) camera position and other camera parameters were determined based on the following rules. First, the scatterplots shown in the `HMD` • `Static` conditions roughly fell into central vision. Second, orthographic and perspective projections generated scatterplots of a similar size in pixels. Third, the camera position was slightly higher than the center of the scatterplot, matching the angle of viewing a monitor on a physical desk [73]. Fourth, participants could see the reference frames even in `Static` conditions.

All the scatterplots were placed surrounding the center of the space extending about 1 meter (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 0.65 meters. 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 1.626 meters between a participant and the center of the scatterplot. Each sphere in the scatterplot was about 0.0017 meters (7.5 pixels) for `Desktop` or 0.015 meters 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 0.30 meters below their eyes.

6.5 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 embeddings were pre-computed using Python 3.5 and the `scikit-learn` library. 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, an NVIDIA GeForce GTX 1080 Ti graphics card, a 32GB RAM, and an ASUS PA248 monitor (24", 1920 $\times$ 1200, 60 Hz). The same desktop drived the HMD (HTC Vive Pro, 2018 model, 2299 $\times$ 2554 pixels per eye, 90 Hz); only one handheld controller was used to interact with the scatterplots and control the flow of the experiment; and the two base stations (the Lighthouse tracking system) stayed in the same locations throughout the experiment.

6.6 Participants

We recruited 32 participants (16 female and 16 male) from our institution and others nearby. We paid participants $10 per hour as compensation for their time.

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. As a result, we

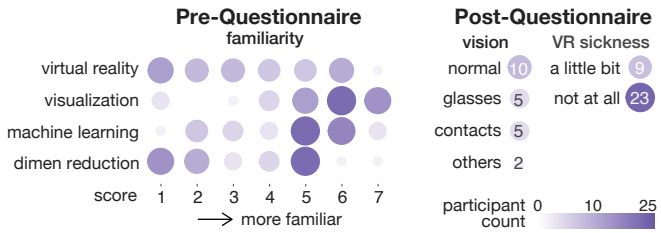

Figure 8. **Participants' background.** Most participants were familiar with visualization and machine learning, while their familiarity with virtual reality devices varied.

recruited 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 normal or corrected-to-normal vision, and not colorblind. The detailed demographics and self-assessments are reported in Fig. 8.

All the experimental sessions were finished before the COVID-19 pandemic and proctored by the same author and experimenter using the same experimental protocol (see Appx. B), setups, and the apparatus.

6.7 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 responded in the same way.

Response time is defined as the time interval between when a scatterplot was first shown and then removed. The scatterplot might be removed because of participants actively proceeding or a timeout (60 seconds).

Between these two measures, we are more interested in error magnitude. The response time was affected by the unexpected noise, especially in `HMD • Dynamic` conditions: the HTC Vive Pro headset might show a flashing screen occasionally [113], increasing wait times.

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 that participants claimed they skipped accidentally. We also manually excluded three error trials, of which the response was very close to the default answer, but the estimation error was extremely large (>50 points); we think that these trials were also skipped but not reported. As a result, we based our analysis on the remaining 3,320 trials.

7 Analysis Methods

To understand visual cue effects in immersive analytics, we use three questions to guide our analysis.

7.1 Guiding questions (GQs)

Our guiding questions [114] are as follows:

GQ1 How do the six manipulated variables affect participants' performance? This research question was inspired by previous studies on examining dominant pictorial cues in a 3D space [2], [10], but our context is immersive analytics.

GQ2 How do the experimental variables interact with each other? The literature suggests that depth cues interact in complicated ways [17], [40]. We wish to understand the interaction effects of the selected cues (e.g., across different devices [14]). In contrast to these studies, we used an analytical task and selected a broader range of cues.

GQ3 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 understand participants' interpretation based on their answers to the open-ended questions.

The results of **GQ 1** support **Contribution 1** declared in Sec. 1; the results of **GQs 2 and 3** support **Contribution 2**.

7.2 Quantitative analysis

We used Bayesian inference for GQs 1 and 2 and focused on reporting effect sizes. Bayesian inference is more suitable for our experiment than significance tests in the following ways. Our research question concerns the visual cue effects in the current experimental setup (see Sec. 6). Therefore, a method like Bayesian inference that focuses on quantifying the effect sizes is suitable for answering this question. Our experiment had a set of variables and conditions, raising concerns about multiple comparisons. Bayesian inference mitigates multiple comparison issues [115] by calibrating all assessments to prior distributions. The current movement to avoid dichotomous thinking [116] (i.e., significant or not based on if $p < 0.05$) further prompted our consideration of Bayesian inference. Last, the literature on sensory cues also suggests that a Bayesian approach to incorporate uncertainty and prior knowledge [45] is appropriate.

We built a Bayesian multilevel model for each measure using a gamma distribution as the likelihood. In particular, we used a hurdle gamma distribution for error magnitude because of the zero values (no error) [117]. As such, in `brm`'s extension of Wilkinson-Rogers-Pinheiro-Bates notation [118]–[120], the error magnitude model 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 $trial\_true\_accuracy$ is the ground truth classification accuracy in a trial. This model reflects our experimental design: the first two terms define that the logarithm of the mean of error magnitude in a joint linear function of all six experimental variables, based on possible interaction effects. We modeled participants and the true accuracy as the random intercepts to acknowledge the similarity in data from participant and an accuracy level. We also modeled `dimensionality` and `data model` as group-level effects (random slopes) to improve comparability across dimensionality and data models.

Similarly, the response time 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" [121]). This model is different from the first model in the group-level effects (random slopes) and the inclusion of a submodel. Response time is likely to be affected by different `device` and `visual motion` other than `dimensionality` and `data model`. The distributions of response time could be very different in shape across `device` and `motion` levels. We thus used a submodel with population-level effects for these shape parameters (adding group-level effects to this submodel will make the model not converge).

We used weakly informative priors that 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 were comparable. We also checked convergence, effective sample size, and posterior prediction to ensure that the models are appropriate. We implemented these using `R` packages `rstan` [122], `brms` [118], and `tidybayes` [123]. The details of modeling and diagnostics are available in supplementary materials.

It is likely that the imbalance between 2D and 3D trials may shift the results towards 3D trials. We will show the results of these different trials by conditioning on one and reporting the conditional probability.

7.3 Qualitative analysis

We followed *thematic analysis* to analyze participants' strategies and answer GQ3. This analysis was based on their answers to the open-ended question. Two experienced researchers (one author and one other coder) extracted the strategies separately, coded all answers independently, and merged the codings via discussion.

8 Results

Following our guiding questions, we extract posterior distributions for each variable and interaction terms to understand effect sizes. We interpret the results based on the probability of observing an increase in error magnitude (or response time) and how large the increase is. We report a summary of posterior medians and 95% quantile credible intervals (Bayesian analogy to confidence intervals) in Fig. 9. Overall, the interaction effects between variables very subtle and complicated for error magnitude, but `Dynamic` (e.g., having `visual motion`) or `HMD` may largely increase response time.

8.1 GQ1: The effects of each variable

To the first guiding question, we reported the model coefficients of each experimental variable (Fig. 10). These population-level coefficients tell the overall "weight" of each variable for an average participant. If the model coefficient is greater than 1, we know that this variable may increase error magnitude or response time.

8.1.1 Results

Error magnitude (Fig. 10a) ① Adding `visual motion` can surely reduce mean error magnitude by a factor in the range of [0.81x,0.94x]. ② Changing `shading`, `dimensionality`, or `data model` could affect mean error magnitude about 70% to 90% of the time by a factor around 1.10x or 0.90x (or about 10%), but the possible effect sizes for the latter two vary a lot.

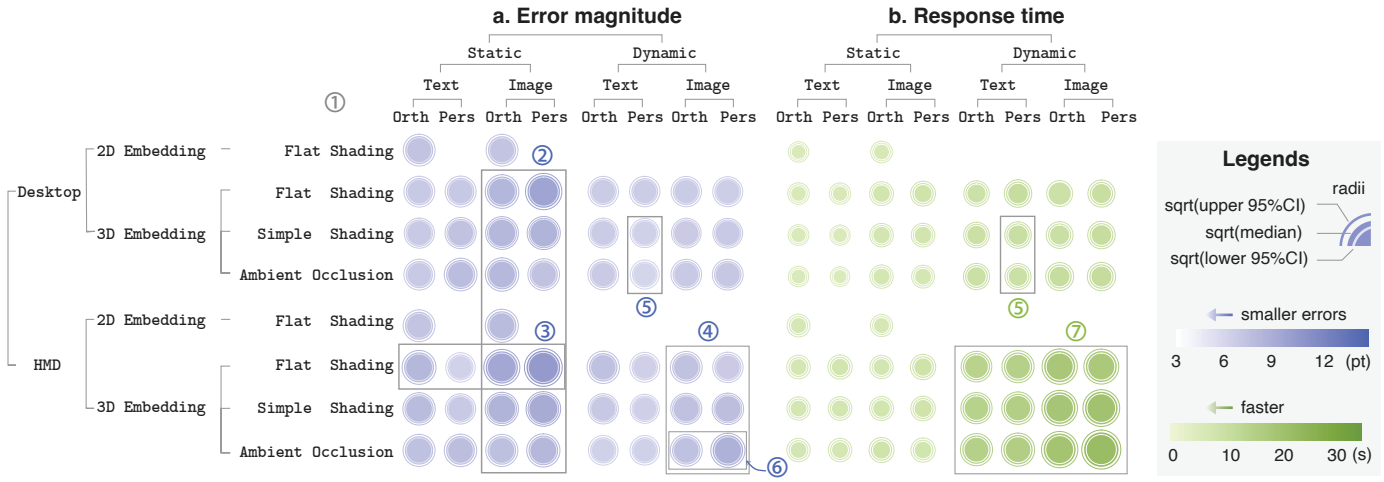

Figure 9. **This figure summarizes all the effects.** We dual-encode posterior median as both color and size and encode the lower and upper bounds of 95% credible intervals (CIs) as radii to show uncertainty in posterior estimates. The detailed numeric results are available in Appx. E. **Example interpretation**: ① Between the two measures, the differences in error magnitude are smaller; the interaction effects among error magnitude are also subtle and more complicated. `Static` scatterplots, different `shading`, `graphical projection`, and `data model` could have small effects on error magnitude. More advanced `shading` generally reduces errors, ② especially for a more complicated dataset like `Image Data`. ③ `Pers.` may show opposite effects on error magnitude. With `Dynamic` scatterplots, these subtle effects fade out, and ④ more advanced `shading` could slightly increase errors, but this effect is very small. ⑦ 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`)

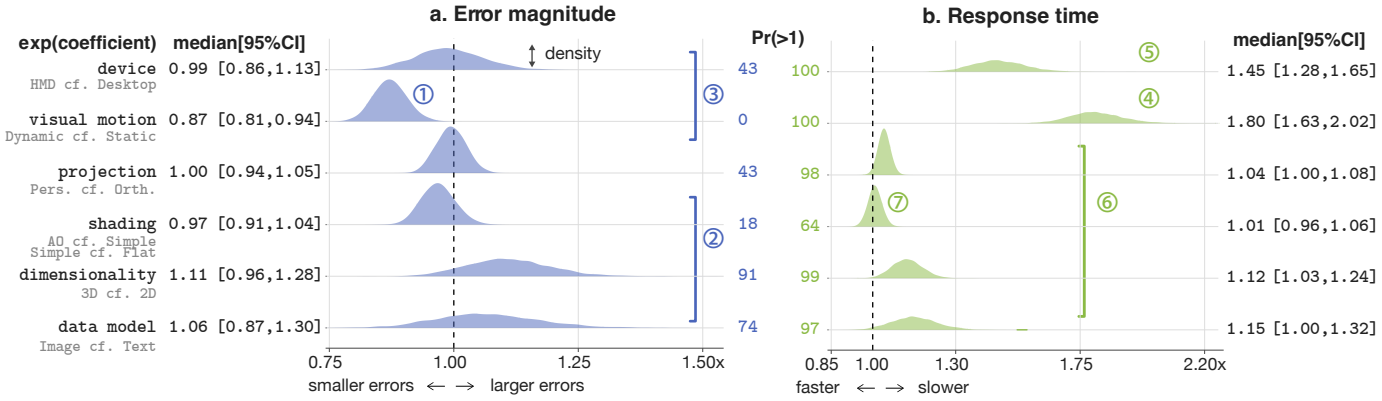

Figure 10. **GQ1: The population-level effects of each manipulated variable.** We show posterior distributions, 95% quantile credible intervals (CIs), median for the model coefficients, and the probability of a coefficient being greater than 1 (increasing). **Example interpretation**: when averaging other factors, changing from `Static` to `Dynamic` ① is very likely to reduce an average participant's estimation error by a factor in the range of [0.81x, 0.94x], ④ and also to let the participant respond more slowly by a factor of in the range of [1.63x, 2.02x], but the actual effect size varies a lot. (`Orth` is `Orthographic Projection`, `Pers` is `Perspective Projection`, and `AO` is `Ambient Occlusion`)

③ Switching `device` or `graphical projection`, on average, seems to only affect error magnitude by chance.

Response time (Fig. 10b) ④ ⑤ Varying `device` (`Desktop` → `HMD`) or `visual motion` (`Static` → `Dynamic`) could largely increase mean response time for sure (e.g., it may double mean response time in the worst case). ⑥ Manipulating `graphical projection`, `dimensionality`, or `data model` also increase response time almost for sure, by a smaller factor in ranges up to 1.32x. ⑦ Manipulating `shading` only affects response time by chance.

8.1.2 *Discussion: Individual cues*

There seems to be a tradeoff between error magnitude and response time for individual cues. This observation aligns with the findings from previous studies comparing a desktop monitor with a VR/AR device [15], [16], [46]–[48], [59], [73]. With two exceptions, we summarize the key findings and implication from GQ 1 as follows:

> **Key finding 1**: In this classification accuracy estimation task, visual cues in order from strongest to weakest effects are `visual motion`, `dimensionality`, `shading`, and `projection`, but `shading` may not affect response time.
>
> **Implication 1**: Manipulating *any* cue could show—sometimes nearly imperceptible—effects on estimation error and response time; this is possibly because that each cue could improve a part of people's perception in 3D space but also force 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 This is shown in Fig. 9 ③ as an increase in error magnitude from [7.68pt,12.00pt] to [8.43pt,13.12pt] for perspective projection. Perspective projection alters size to depict depth. According to *Gestalt psychology* [124], [125], people seek similarity in elements (e.g., size, movement) to form groups visually. However, if the spheres from the same group (e.g., the spheres with the same color) are of different sizes, this conflict may prevent people from grouping them. As such, perspective projection may sometimes have caused a decline in task performance, and therefore orthographic projection occasionally showed better performance. The second exception is that `3D embedding` can both increase and reduce errors, and increase response time. This is shown in Fig. 10 ② as errors changed by a factor in the range of [0.96x,1.28x], and in Fig. 10 ⑥ as response time increased by a factor in the range of [1.03x,1.24x]. `3D embedding` has the third dimension as a cue. While `3D embedding` better resembles the original high-dimensional data and provide more information than `2D embedding`, this extra information still has to be presented carefully.

There are a few explanations for the moderate effects and the tradeoff. 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 a scatterplot at different perspectives; participants were less familiar with an HMD; also, the HMD provides more pixels but fewer pixels per degree. Visual motion causes more time for examining the stimuli. The chosen dimension-reduction technique (t-SNE) preserves cluster structure well, and this may have dominated over the tested visual cues [22].

8.2 GQ2: The interaction effects

For our second guiding question, we reported the model coefficients to examine the interaction effects between variables (see Fig. 11). We considered only two-way interaction effects for simplicity; more complicated interaction effects are implied in Fig. 9 above.

8.2.1 *Results*

Error magnitude (Fig. 11a) ① Most of the variable pairs suggest a two-way interaction effect happening about 70% to 90% of the time, except that ② `device` seems to only interact with `visual motion` or `shading` by chance. ③ For instance, changing `visual motion` and further changing `shading` increases errors 89% of the time by a factor in the range of [0.95x,1.24x].

Response time (Fig. 11b) We find three two-way interaction effects that can almost affect response time for sure. ④ ⑤ Changing `device` and then `visual motion` increase response time by a factor in the range of [1.10x,1.92x] or [1.00x,1.20x]; ⑥ changing `device` or `visual motion`, and further changing `shading` increase response time by a

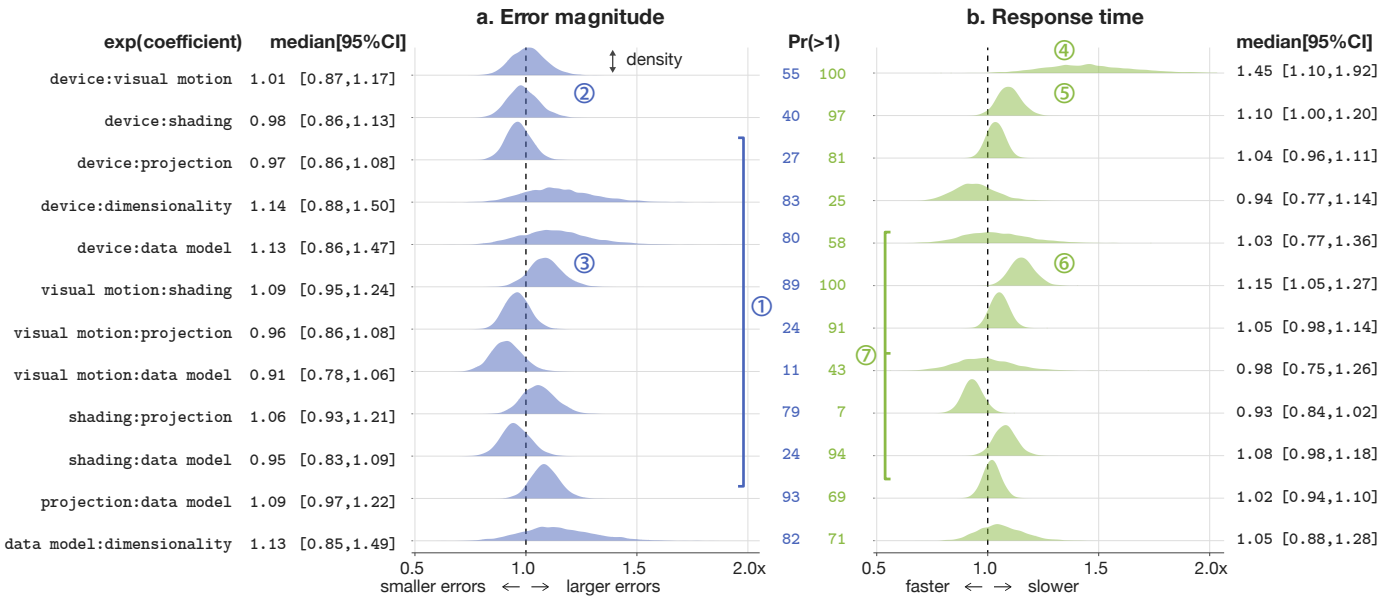

Figure 11. **GQ2: The two-way interaction effects.** We report posterior distributions, 95% quantile credible intervals, and medians of two-way interaction coefficients, and the probability of a coefficient being greater than 1 (increasing). For error magnitude, most variables interact with each other moderately. For response time, there is a strong interaction effect between `device` and `visual motion`, and several small to moderate interaction effects. **Example interpretation**: when averaging other factors, if we change from `Desktop` to `HMD`, further changing from `Flat Shading` to `Ambient Occlusion` ② affects an average participant's estimate error by chance; ⑤ however, it lets the average participant respond slightly more slowly, typically by a factor in the range of [1.00x, 1.20x].

factor in the range of [1.05x,1.27x]. ⑦ The other variable pairs indicate interaction effects happen about 70% to 90% of the time (e.g., `device:shading`) or by chance (e.g., `visual motion:data model`).

8.2.2 *Discussion: Cue integration*

Our results hint at a connection to cue-integration theory [7], [126]; that is, people can make more accurate estimates of environment properties by integrating multiple sources of information [45]. In our experiment, visual cues interact with each other in complicated ways, similar to the findings from prior AR studies [17]. Each cue may facilitate or impede part of the perception, and combining cues may cause conflicts. For example, perspective projection and ambient occlusion together generate a decent sense of depth, and combining them was very effective on a desktop monitor. However, perspective projection and ambient occlusion may not always be beneficial, and participants may have used more than one cue, and cue integration is occurring. We summarize the key findings and implications from GQ 2 as follows.

Key kinding 2: `Visual motion` and `shading`, and `projection` and `data model` show relevantly strong interaction effects on error magnitude; but these visual cues (e.g., `visual motion` and others) show much stronger interaction effects on response time.

Implication 2: Besides the differences in devices and physical movements, manipulating *multiple* visual cues may aid in assessing classification performance.

For example, advanced shading (e.g., ambient occlusion) on both devices generally improves participants' performance; they reduced estimation error without causing a substantially longer response time. This is shown in Fig. 9 ⑤ as error magnitude reduced from [4.82pt, 7.79pt] to [5.24pt, 7.81pt], and in Fig. 9 ⑤ as response time reduced from [7.54s, 13.33s] to [8.12s, 13.94s].

The intricate interaction effects between visual cues also lead to another implication.

Implication 3: 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` and `HMD`, participants' performance declined slightly if further adding perspective projection. This is shown in Fig. 9 ⑥ as error magnitude increased from [5.91pt, 9.63pt] to [6.81pt,10.89pt].

We have a few explanations. The first relates to *visual complexity*. Multiple visual cues together aid 3D perception, but they also increase visual complexity, preventing participants from using visual similarity to group elements. The second is that cues may conflict with each other [45]. For example, our shading models use darker colors to indicate a more distant position in 3D space, while perspective projection uses smaller sizes to indicate further. However, a more distant smaller sphere may not necessarily be darker than a near one, and this is likely to cause a conflict in perception. The third one is that adding more cues demands a precise implementation for each; 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 black edges and aliasing, which are salient in VR, possibly explaining that ambient occlusion is not always as effective as a simple shading model.

8.3 GQ3: Participants' strategies

Last, we present the results from thematic analysis and report participants' strategies in Fig. 12.

8.3.1 Results

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. 12b). 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 sessions, and the rest used different strategies in different sessions.

A few participants reported the interaction techniques they used (Fig. 12c), with rotation being the most helpful one (28.13%). In their anecdotal feedback (Fig. 12d), participants found occlusion harmful and reported that VR could be distracting or causing physical discomfort (37.50%), but VR also could be entertaining or helpful for the task (37.50%). In sum, the key finding from this qualitative study is as follows.

> **Key finding 4:** Participants use different strategies when different visual cues are presented, even with the same device or dataset.

8.3.2 Discussion: Strategy consistency

All these strategies appear to be feasible for assessing classification performance and estimating accuracy from the scatterplots. The two dominating strategies suggest that the task was correctly interpreted as visually distinguishing classes. The presence of less commonly used strategies suggests that other factors (e.g., visual cues and outliers) may have affected participants' strategies and understandings. The results also suggest that half of the participants interpreted the task consistently across different combinations of visual motion and device, and others used different strategies across these combinations. Different visual cues may aid or hinder the use of a certain strategy, and more effective combinations of cues might assist in a more commonly used strategy. Manipulating cues to facilitate the dominating strategies may improve participants' performance on average, but the optimal may be to query individual's strategies on different devices and datasets and then personalize the visualizations.

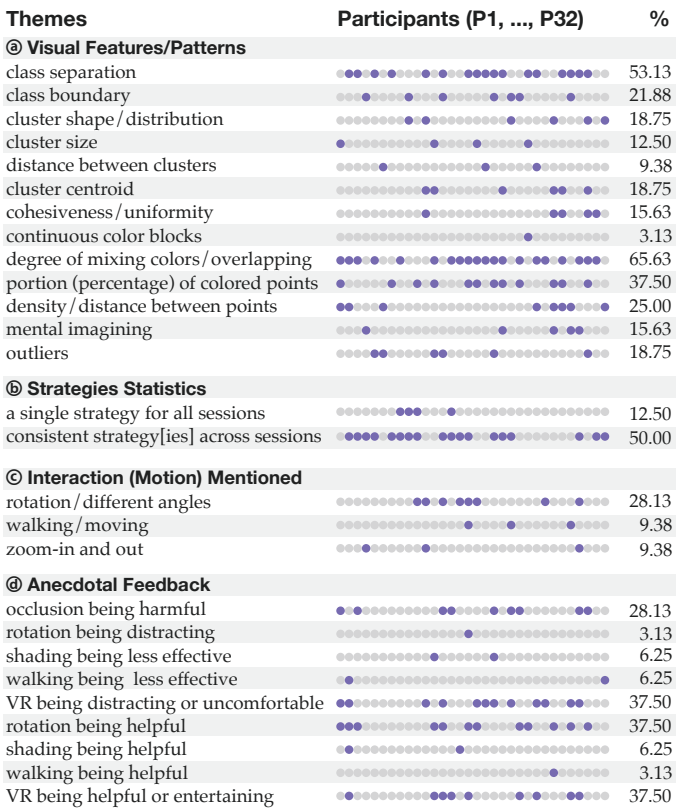

Figure 12. **GQ3: Participants' strategies.** Each dot represents a participant, and a darker dot indicates that a theme appeared in that participant's answers.

9 General Discussion

In this study, we find that most of visual cues have subtle effects on classification accuracy estimation, and combining them can result in perceptual improvements, conflicts and various participant strategies. Following this line, we speculated that perceiving classification accuracy may be regarded as a combination of multiple low-level abstract tasks. The literature supports that multiple low-level tasks could constitute a mid-level perception task, similar to using low-level visual proxies to summarize mean values in bar charts [127] and processing multiple cues in parallel in visual search [128]. To prove this speculation requires systematic manipulating each subtask and knowing which subtask dominates perception, which could be a promising extension of the present work.

Given the conjecture above, our results might be specific to this task related to machine learning. For example, this task may rely on three subtasks: class separation, proportion estimation, and outlier detection. Visual motion is likely to aid all three, shading may facilitate class separation, graphical projection and dimensionality may not have a clear effect on any. Therefore, we observed the corresponding effect sizes that range from strong to weak. Also, while the student participants were still relative novices in machine learning, their behavioral data demonstrated the potential of using virtual reality for understanding and debugging a machine learning model; their feedback will help enhance the design and use of visualizations for similar tasks.

This mid-level task with a complex experimental design may not produce clear results in an absolute sense. However, this complexity likely resembles the randomness in the real-world. **Our results suggest that visual cues might *not* strongly affect perception in a realistic analytical task, unless they change how people *think* about the data**. Visual motion is an example of this. It shows people different perspectives of a dataset and alters people's mental model about the classification performance.

We tested a subset of important visual cues and controlled as many factors as possible, but there is still room for improvement. For example, in our designing process, we

noticed that the direction of light could be another important factor because it affected shading, but we did not find enough literature to suffice the speculation. We designed for an averaged-size participant, and part of our configuration can be further adjusted for different participants. In addition, other non-visual cues, such as haptic vibration feedback [40], 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, as reflected in the manuscript, a comparison across modalities for compound visual cues is innately difficult. We endeavored to disentangle the predicament by broadly consulting the literature and cautiously designing the experiment. Though imperfect, there are valuable lessons from this study. While the delicate experimental design and the variance in a virtual reality system raise challenges in quantifying visual cue effects, we find Bayesian inference is particularly suitable and powerful.

10 Conclusion

We assessed the visual cue effects on the task of estimating the classification accuracy of a deep neural network based on its last hidden layer's 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 estimation error but increasing response time. Compared to a desktop monitor, an HMD can lead to better, worse, or similar performance depending on the combination of cues (e.g., whether visual motion is available). Improving shading reduces estimation error slightly, but this 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. We speculate that visual cues might not strongly affect perception in a realistic analytical task, unless they change people's mental model about data. Our work advances the understanding and modeling 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, by NSF IIS-2107409, and by the NSF 2127309 to the Computing Research Association for the CIFellows Project. We thank the anonymous reviewers and the editors for their thoughtful comments. We thank Shenghui Cheng, Loudon Cohen, Ailin Deng, Aaron Gokaslan, Mi Feng, Elaine Jiang, Benjamin Knorlein, Johannes Novotny, Emily Reif, Jing Qian, Kexin Qu, and Yalong Yang for their help with the 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] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

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

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

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

[59] 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.

[60] 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.

[61] 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.

[62] 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.

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

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

[65] 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.

[66] 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.

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

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

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

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

[83] 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.

[84] 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.

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

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

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

[92] 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.

[93] B. Ens, B. Bach, M. Cordeil, U. Engelke, M. Serrano, W. Willett, A. Prouzeau, C. Anthes, W. Büschel, C. Dunne *et al.*, "Grand challenges in immersive analytics," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–17.

[94] 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.

[95] 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.

[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] 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.
[116] P. Dragicevic, "Fair statistical communication in hci," in *Modern statistical methods for HCI*. Springer, 2016, pp. 291–330.
[117] 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.
[118] 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.
[119] 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.
[120] 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.
[121] J. P. Klein and M. L. Moeschberger, *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.
[122] Stan Development Team, "RStan: the R interface to Stan," 2018, r package version 2.18.2. [Online]. Available: http://mc-stan.org/
[123] M. Kay, *tidybayes: Tidy Data and Geoms for Bayesian Models*, 2019, r package version 1.0.4. [Online]. Available: http://mjskay.github.io/tidybayes/
[124] W. Köhler, "Gestalt psychology." 1929.
[125] K. Koffka, *Principles of Gestalt psychology*. Routledge, 2013.
[126] H. H. Bülthoff and H. A. Mallot, "Integration of depth modules: stereo and shading," *Josa a*, vol. 5, no. 10, pp. 1749–1758, 1988.
[127] 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.
[128] H. Wässle, "Parallel processing in the mammalian retina," *Nature Reviews Neuroscience*, vol. 5, no. 10, pp. 747–757, 2004.

Fumeng Yang is a postdoctoral fellow in the Department of Computer Science at Northwestern University. She obtained her PhD degree in computer science from Brown University. Her research interests include information visualization, virtual reality, predictive modeling, and human computer interaction.

James Tompkin is the John E. Savage Assistant Professor of Computer Science at Brown University. His research at the intersection of computer vision, computer graphics, and human-computer interaction helps develop new visual computing tools and experiences. His doctoral work at University College London on large-scale video processing and exploration techniques led to creative exhibition work in the Museum of the Moving Image in New York City. Postdoctoral work at Max-Planck-Institute for Informatics and Harvard University helped create new methods to edit content within images and videos. Recent research has developed new machine learning techniques for low-level scene reconstruction, view synthesis for VR, and content editing and generation.

Lane Harrison is an Associate Professor in the Department of Computer Science at Worcester Polytechnic Institute. Prior to joining WPI, he was a postdoctoral fellow in the Department of Computer Science at Tufts University. He obtained his Bachelor's and PhD degrees in computer science from the University of North Carolina at Charlotte. Lane directs the VIEW group at WPI, where he and his students leverage computational methods to understand and shape how people use visualizations and visual analytics tools.

David H. Laidlaw is a professor in the Computer Science Department at Brown University. He received his PhD degree in computer science from the California Institute of Technology, where he also did post-doctoral work in the Division of Biology. His research centers on applications of visualization, modeling, computer graphics, and computer science to other scientific disciplines. He is a fellow of the IEEE.