DRCmpVis: Visual Comparison of Physical Targets in Mobile Diminished and Mixed Reality

Richen Liu, Shunlong Ye, Zhifei Ding, Guang Yang, Shenghui Cheng, and Klaus Mueller, *Fellow, IEEE*

***Abstract*—Numerous physical objects in our daily lives are grouped or ranked according to a stereotyped presentation style. For example, in a library, books are typically grouped and ranked based on classification numbers. However, for better comparison, we often need to re-group or re-rank the books using additional attributes such as ratings, publishers, comments, publication years, keywords, prices, etc., or a combination of these factors. In this article, we propose a novel mobile DR/MR-based application framework named DRCmpVis to achieve in-context multi-attribute comparisons of physical objects with text labels or textual information. The physical objects are scanned in the real world using mobile cameras. All scanned objects are then segmented and labeled by a convolutional neural network and replaced (diminished) by their virtual avatars in a DR environment. We formulate three visual comparison strategies, including filtering, re-grouping, and re-ranking, which can be intuitively, flexibly, and seamlessly performed on their avatars. This approach avoids breaking the original layouts of the physical objects. The computation resources in virtual space can be fully utilized to support efficient object searching and multi-attribute visual comparisons. We demonstrate the usability, expressiveness, and efficiency of DRCmpVis through a user study, NASA TLX assessment, quantitative evaluation, and case studies involving different scenarios.**

***Index Terms*—Diminished reality, visual comparison, mixed reality, mobile environment.**

I. Introduction

IN OUR everyday life, we often spend a great amount of time searching for a specific target from numerous candidates (e.g., searching for algorithm-related books in a library or bookstore). In this case, we may get limited information about the objects from the appearances of the physical objects. For example, the books' spine side in libraries just provide limited information, while users often require to know much more about the books, including the topics, ratings, comments, sales volume/borrowing rate, most relevant books, authors' other series of books, etc. Similarly, it would take us too much time to reorganize objects' information including their multi-attributes for better comparison. Considering a usage scenario inside a library or bookstore that consists - (1) filtering & highlighting: users are likely to search for a book according to the fuzzy book name or the author's name (a nominal variable) when they enter a library or a large bookstore, as shown in Fig. 1(a), and then they would browse all the books and filter them to get a smaller number of candidate books such as the keyword "Algorithm" (nominal) for further comparison. There are two subsequent actions they would take: (2) re-grouping: re-group the candidates according to the topics (such as "dynamic programming", nominal), publishers (e.g., "ACM", "Springer" or "MIT Press", nominal), or even more additional attributes, as shown in Fig. 1(b). (3) re-ranking: choose the candidates according to their ratings (ordinal), prices (quantitative), sales volume/borrowing rate (quantitative), or even more additional attributes, as shown in Fig. 1(c). Besides, users may want to know extra information about the books by mobile devices, if they could not be found from the book covers. However, it is time-consuming to search the extra information for all candidates, and it is also tedious to re-group them and write down the key information by juxtaposed comparison. Additionally, we also frequently encounter the situations where individuals struggle to differentiate between goods (such as coffee, food, or other beverages) or face challenges when choosing a particular item from a multitude of options due to an inability to identify or recall the significant distinctions among them. Such scenarios involving visual comparisons of numerous physical objects are prevalent in our daily lives. For example, it is neither easy for us to remember all the ingredient differences of multiple coffees, nor convenient to compare them with multi-attributes, when we are in a cafe.

Stolte et al. [1] introduced that the overall data flow across multi-dimensional data queries, visualizations, and analyses consists of "selecting subsets of the data for analysis, then *filter*, *sort*, and *group* the results" [1]. Furthermore, according to Jacques Bertin's book "The Semiology of Graphics" [2], data types of visual variables include *nominal*, *ordinal*, and

Manuscript received 6 September 2023; revised 10 January 2024; accepted 22 January 2024. Date of publication 25 January 2024; date of current version 29 October 2024. This work was supported in part by the National Natural Science Foundation of China the NSFC under Grant 62372241, and in part by the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University under Grant VRLAB2023B05. Recommended for acceptance by D. Keefe. *(Corresponding authors: Shunlong Ye; Shenghui Cheng.)*

Richen Liu is with the School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University, Nanjing 210098, China, and also with State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: richen@pku.edu.cn).

Shunlong Ye, Zhifei Ding, and Guang Yang are with the School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University, Nanjing 210098, China (e-mail: 19180306@njnu.edu.cn; 222202011@njnu.edu.cn; 19180113@njnu.edu.cn).

Shenghui Cheng is with the School of Engineering, Westlake University, Hangzhou 310024, China (e-mail: SheCheng@cs.stonybrook.edu).

Klaus Mueller is with Computer Science Department, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: mueller@cs.stonybrook.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2024.3358419, provided by the authors.

Digital Object Identifier 10.1109/TVCG.2024.3358419

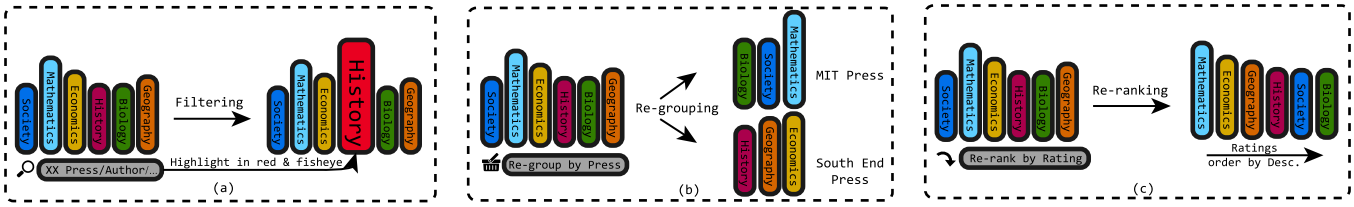

Fig. 1. Three types of data flow tasks within the DR/MR-based computational framework: (a) filtering, (b) re-grouping, and (c) re-ranking in DR environment. We take a library scenario as an example.

quantitative. Considering these two principles [1], [2], we further structure the data flow space into three families:

- *Filtering*: highlight the filtered results with fisheye deformation to provide visual cues about their physical positions (available attributes: *nominal*, *ordinal*).
- *Re-grouping*: re-group the objects according to one/multiple attributes via breaking the original physical layouts in the DR/MR environment (available attributes: *nominal*, *ordinal*, *quantitative*).
- *Re-ranking*: sort the objects according to one/multiple attributes via reorganizing the original physical layouts in the DR/MR environment (available attributes: *ordinal*, *quantitative*).

Extended reality (XR) techniques have great potential to facilitate the resolution of the above-mentioned issues. Various combinations of searching, re-grouping and re-ranking can be achieved in the virtual space of mobile XR. These actions can be taken flexibly on the virtual avatars of physical objects. XR typically consists of virtual reality (VR), augmented reality (AR), and mixed reality (MR). MR is strictly defined by Milgram and Kishino [3], which was considered as a mixture of real and virtual objects within a single display. The distinctions between AR and MR are fuzzy [4]. To the best of our knowledge, there is no literature that strictly defines their differences due to the overlaps. Situated analytics (SA) is another concept which considers AR as one of its four primary elements, including situated information, abstract information, augmented reality interaction, and analytical interaction [5]. SA is capable of supporting visual analytics' analytical reasoning by embedding the visual representations and interaction of the resulting data in the physical environment using AR. ElSayed et al. [5] recognized SA as an emerging research field at the intersection of visual analytics and AR. A recent addition to this domain is the introduction of the concept of diminished reality (DR) [6], [7]. DR pertains to the manipulation of a perceived environment in real-time, involving actions like concealing, eliminating, or revealing objects [6], [7]. According to the survey on DR [6] summarized by Mori et al. DR examples include four types: diminishing, seeing through, replacing, and inpainting real objects. Specifically, DR can degrade visual functions for a certain purpose (diminishing). For example, the color information of a visual field can be thinned out or distorted [6]. It can be used to cover real objects with images of their occluded background to make the objects virtually invisible in our vision (seeing through), or delete the undesired objects and replace them with virtual objects (replacing) [7]. It can also be used to generate plausible background images based on the context (inpainting).

In this paper, we propose an interactive application framework named DRCmpVis, enabling users to compare numerous physical objects with text labels/information in mobile DR. The physical objects are captured from the camera of personal mobile devices (mobile phones or tablets) in real-time, then the text information can be extracted to recognize different objects. We mainly use the feature "replacing" of DR in DRCmpVis. It replaces the real objects with virtual avatars. Multi-dimensional comparisons can be completed by filtering, re-grouping, re-ranking, and their combinations on their avatars in the DR context. The additional augmented information of the objects can be also encoded into some simple visual comparisons.

Why do we mainly use DR instead of traditional database search to reorganize the additional information before decision making? We take the library/bookstore case as an example. One scheme to show additional information about physical objects is to query them directly from the database of a library/bookstore. However, there are several limitations of this scheme due to the inconsistency between the physical space and the virtual space in a database system: (1) the books in a library/bookstore often would be put in a wrong position by a librarian or readers, which may be inconsistent with the information in its database. (2) Users might frequently read unborrowed books on tables, making it challenging for others to fetch these books through database queries. Additionally, users might forget the precise positions where they picked up the books they were reading. (3) The books on a best-seller bookshelf in a bookstore are often updated in the physical world while it is tedious for a librarian or a bookstore attendant to update the database frequently. (4) Last but not least, users often have limited permission to access the database of a shop. Overall, DR is an optimal solution to keep information consistent between the physical space and the virtual space while significantly reducing visual clutter.

Another question is why we use DR + MR instead of just using AR (or MR). We use it because the physical objects can be replaced by their virtual avatars in DR, which allows various comparisons performed in the virtual space flexibly and seamlessly. For example, the re-layouts of virtual avatars in DR can be flexibly performed in virtual space while avoiding breaking the original physical layouts. While in AR, it is difficult to conduct re-grouping and re-ranking because both physical objects and virtual information are in an identical AR environment. Besides, DR saves visualization space and helps to reduce visual clutter and operational ambiguity. DR is also capable of building an

information bridge between the changing physical space and the virtual space seamlessly. Actually, we use both DR and MR in DRCmpVis. The augmented information is visualized in an MR context. MR in this paper is similar to AR, because the distinctions between AR and MR are fuzzy [4].

There are two device choices for DRCmpVis, e.g., the mobile devices and the HMDs like Microsoft HoloLens 2 or Meta Quest 3. We suggest prioritizing the use of mobile devices for DRCmpVis due to three reasons. First, the current HMDs may cause cybersickness [8], [9], [10], such as dizziness, nausea, vertigo, or sweating. Second, DR may replace the physical objects into virtual avatars. For example, the physical books on the current bookshelf will be replaced with virtual 3D books in a library. The HMDs are more prone to cause users to bump into the bookshelf due to its lower ability to perceive the outside physical environment which is less likely with handheld mobile devices. Third, the mobile devices are portable and easier to carry in public places.

Technically, we use a trained convolutional neural network (CNN) named PaddleSeg [11] to segment and label all the objects. Furthermore, we extract the text information by an OCR-based neural network. In the experiments, we evaluate the proposed DRCmpVis using four usage scenarios, a user study, a performance evaluation, and a NASA-TLX measurement, compared with two traditional methods. The contributions of this work are summarized as follows:

- We propose a novel DR/MR-based computational framework to compare physical objects with text labels or text information. The framework enables users to fully utilize the efficient computation resources in virtual space and the in-context interactions in physical space in real-time.
- We classify the multidimensional comparison tasks in DR in terms of all three types of attributes (nominal, ordinal, and quantitative), and then integrate commonly-used visualizations into the DR/MR context to achieve flexible object comparisons.
- We design three DR-based visual comparison strategies for physical object multi-attribute comparisons, i.e., filtering, re-grouping, and re-ranking, avoiding breaking the original physical layouts of the physical objects.

II. Related Work

Visual comparison aims at providing visual support for the understanding of underlying abstract data sets [12]. The visual comparison tasks in this paper are a little different from the traditional ones because the compared items in DRCmpVis are physical objects.

A. DR-Based Data Presentation

There is little literature that strictly defines the differences between VR, AR, and DR, while XR is often considered as consisting of VR, AR, MR, and DR. In a narrow sense, DR is different from AR, which shows the physical reality of the world. AR-based data presentation [13], [14] allows developers to create AR applications that overlay digital virtual information into the reality, while DR makes objects disappear from the physical world environment and their virtual avatars can be used to replace their positions and provide flexible information visualization in virtual world.

We find few recent related works focused on DR-based applications, especially for data presentations. For example, Kawai et al. [15] found that the background geometry has few constraints, where the reality can be removed. In order to simulate the geometric shape of a similar background, they proposed it can be achieved by combining local planes and using the perspective distortion technology of correcting the texture. A method [16] of blending and replacing textures was further proposed. The texture of the remaining part of the video and the mixed texture of the target area is blended and replaced, and then the blended results are used in the next frame of the video to be played. The key idea of their approach is that the texture image of the target area can be updated in real-time according to the changes in lighting so that the overall video appears natural. Hashiguch et al. [17] combined AR and DR to examine how the cross-modal effects of AR and DR are achieved, and why people's sense of weight is changed by continuous visual changes between AR and DR. In practical applications, Herling et al. [18] designed a real-time reduction of reality method that can achieve high-quality video. However, most of the existing methods are based on texture synthesis or replacement, which are difficult to implement when the background is complex or has any shape. Li et al. [19] proposed a new system-level framework for reducing reality. This method uses online photo collections to provide appearance and 3D information to achieve 3D structure acquisition in an offline process.

B. MR/AR Visualization Applications

MR/AR can be used to realize data visualization in the physical space to promote certain visual explorations and combine presentations with personal ideas and preferences [20]. Liu et al. [21] categorized the work on XR + InfoVis into four types according to abstract data types, i.e., graph/network data visualization [22], [23], [24], high-dimensional data visualization (e.g., immersive parallel coordinate plots [25] and IATK [26]), time-series data visualization [27], [28], and text data visualization [29]. MR/AR techniques shed new light on the visualizations of large scale graph data, high-dimensional data and time-series data, because they can virtualize as many views as possible theoretically. For example, a virtual tiled display wall [30] in an immersive environment is capable of breaking the limitations of physical screens. Besides, MR/AR can be used to offer expressive data visualizations [31], [32]. For example, a tool named MARVisT [33] was designed to allow users to realize expressive AR glyph-based visualizations. PapARVis [34] is capable of designing an environment that can debug both static and virtual content simultaneously. Another tool named ARShopping [31] was proposed to help consumers make purchasing decisions. However, it is difficult to conduct object re-layout in an AR environment. Besides, ARShopping needs to print QR-code markers to help detect the corresponding

TABLE I
COMPARISON TO THE MOST RELATED RECENT WORK ABOUT DATA PRESENTATION TOOLS TOWARDS VR, AR, OR DR

	Task					Visual Presentation			Work-flow (Single/Collab)
	Virtual Space	Augmented Information	Searching	Re-grouping (Multi-attrib)	Re-ranking (Multi-attrib)	Glyph Vis	Small Multiples	Fish Eye Highlight	
DXR		■				■	■		Sin(PV)
AVT	■	■							Collab
VRIA	■	■				■	■	■	Collab
VR Visc	■	■				■	■		Sin(PV)
VR Collab Vis	■	■							Collab
IATK	■	■				■	■	■	Collab
SA Vis		■					■		Sin(PV)
PapARVis		■					■		Sin(PV)
MarVisT		■				■			Sin(PV)
Our Work		■	■	■	■	■	■	■	Sin(PV)

DXR [38], Augmented Virtual Teleportation (AVT) [39], Situated Analytics (SA Vis) [40], Data Visceralization (VR Visc) [41], Shared Surfaces and Spaces (VR Collab Vis) [42], IATK [26], VRIA [43], PapARVis [34], MARVisT [33]. The workflow can be categorized into PV (single user in a personal data presentations), single user (Sin) or collaborative users (Collab).

products, which may introduce too much extra efforts to keep consistency between products and their markers, because the products may often be moved or purchased by customers.

C. Position in XR-Based Visual Comparison Tools

Some library tools were designed to help users better explore books, including xexes [35] and Bohemian Bookshelf [36]. Hieraxes integrated the power of hierarchical book browsing into a 2D visualization, which preserves the overview of search results and enables users to rapidly comprehend them. Bohemian Bookshelf helps users explore how information visualization supports serendipitous book discoveries. The adjacencies between books can be highlighted and further explored. Besides, a visualization tool named HORUS EYE [37] was further designed to simulate bird and snake vision to highlight data of interest, e.g., the book titles. Both Hieraxes and Bohemian Bookshelf are non-immersive book exploration tools, while HORUS EYE is a visualization tool which does not support visual comparisons on multi-attributes of physical objects. In contrast, DRCmpVis is an immersive application framework that enables multiple objects' multi-attribute comparisons in an interactive mobile environment.

Regarding the existing AR-based visual comparison tools, augmented information can be overlapped on the physical objects. Bach et al. [32] discussed the design space in an AR context, including a visualization design example in library scenarios. We summarize and discuss the differences between DRCmpVis and the most related tools as shown in Table I, according to the tasks (augmented information, searching, re-grouping, re-ranking), visual presentations (glyph, small multiples, fish eye highlight), and workflow (personal or collaborative). Specifically, we mainly focus on the DR environment while most of the related tools [31], [33], [34], [38], [39] focus on AR and some [26], [42], [43] are more close to VR. Compared with them, DRCmpVis can provide flexible and intuitive interaction on the virtual avatars themselves, including object re-layout such as re-grouping, re-ranking and filtering. DRCmpVis can achieve object re-layouts without breaking the original physical layouts. It can also save visualization space because the physical objects will be replaced by their virtual avatars in real time in the DR environment.

III. DESIGN RATIONALE

We illustrate the design goals, design considerations and design details of DRCmpVis in this section.

A. Design Goals

We summarize four design goals for the applications built on DRCmpVis.

- G1: enable to filter/search physical objects for better comparison, and then highlight the results to indicate their positions in reality (using nominal attributes).
- G2: enable to re-group the physical objects for comprehensive comparison (using nominal, ordinal, or quantitative attributes).
- G3: provide functionality to re-rank or sort physical objects, enhancing the interactive visual comparisons (using ordinal or quantitative attributes).
- G4: achieve multi-attribute object comparison by using simple visual comparisons in MR space.

B. Design Considerations

In this paper, we choose multiple usage scenarios to demonstrate that the proposed approach is not ad-hoc, including the scenarios in a library/bookstore, a coffee shop, an eyeshadow shop, and a restaurant (Shaxian County cuisine). The latter two scenarios are moved to the Appendix file due to the page limit, available online.

We summarize the design considerations and design details of DRCmpVis towards the design goals (G1-G4):

First, these applications should be designed to enable filtering the numerous physical objects for better comparison by one or multiple fuzzy keywords (G1). The filtering keywords can be input by voice, as suggested by the participants in the pre-study of the work, because voice input is simple to use in the public's personal context. However, the provision of text input through a virtual keyboard is also incorporated for situations where vocal input might not be feasible. The search results should be highlighted by visual cues to indicate their positions in reality.

Second, these applications should be designed to re-group the physical objects in terms of one or multiple attributes of the target objects (G2), e.g., re-grouping them according to their nominal, ordinal or quantitative attributes, which can help users better compare target candidates.

Third, these applications should be designed to enable re-ranking the disordered physical objects for visual comparison in terms of one or multiple ordinal or quantitative attributes (G3). For example, books in a library are usually sorted by classification or index number, which might not align with users' diverse sorting requirements. Sorting them by the rating, price, publisher, or publish year is helpful in target comparisons. Similarly, the books in a bookstore are often sorted by user groups, and more information like ratings and prices are ignored.

Consequently, readers might save substantial time in searching for an ideal book amidst the shelves.

Fourth, in people's daily life, the visible information alongside an object is usually not enough (G4). For example, we can see the title and name of a book on a bookshelf, and can see the price of a cup of coffee in a menu. However, the rating of coffees and books, the ingredients of drinks, foods and fruits are often neither shown directly nor is it feasible to make comparisons in terms of attributes. Therefore, the tool should be designed to display additional information which is often hidden from users or tedious for them to compare.

C. Design Details: System Workflow Design

DRCmpVis consists of two parts. The first part is the mobile client, which is used to take panoramic photos or record a real-time video and then render objects in DR. The second part is the server, which is employed to process almost all of the data. The overall processing is described as follows: the mobile client constantly takes pictures or records a real-time video of numerous objects and sends them to the server. The remote server processes those pictures or key frames, recognizing objects in them in real time, and sends the objects' data back to the mobile client, which displays them in new layouts. The implementation has two considerations:

Separate Heavy Computing and DR/MR Presentation. Unlike traditional applications, DRCmpVis shifts most of the computationally intensive tasks to the server. The mobile client only needs to send the requests in multi-thread to ensure real-time object recognition. This enables DRCmpVis to handle a large amount of data without adding a heavy burden to the user's mobile device or influencing the user's interaction experience. In the library/bookstore scenario, for example, more than a thousand books can be recognized in DR/MR with panoramic pictures.

Separate Processing of Text and Texture. The text and texture in one picture usually contain most of our desired information. We apply different neural networks to process these two kinds of data. This makes our model not only suitable for situations where information is expressed more in text, such as a book or a menu, but also for texture which contains more information.

Separate DR Mode and Non-Immersive Mode. The visual comparison tasks in this paper consist of object re-layout and augmented information comparison. The former includes filtering, re-grouping, and re-ranking for virtual avatars, while the latter includes comparative visualizations using charts, plots, glyph, etc. On the one hand, DRCmpVis shows object re-layout results in DR mode, because it is more intuitive to conduct such interactions on the virtual avatars themselves. On the other hand, it displays the visual comparison plots and charts in non-immersive mode to get a larger presentation space. We note that the comparison stage occurs often after filtering or when the number of candidate objects is relatively small. The visualization plots and charts in this stage are more significant than the avatar presentation.

IV. Implementation

Some technical challenges that we have addressed in DRCmpVis are summarized as follows:

- *Challenge I: Building the Application Framework.* Image segmentation, image labelling, OCR-based text extraction, and image recognition are the significant modules of the framework. We have integrated two latest deep neural networks into the framework. All of them are encapsulated as the APIs of the framework.
- *Challenge II: Coordinate Transformation Between Physical Space and Virtual Space*. We should keep the coordinates consistent between virtuality and reality. This step is to build the virtual avatars mapped to the physical objects and then mix them seamlessly in an identically calibrated coordinated system. We have developed and encapsulated the related functions into the APIs of the framework.
- *Challenge III: Integrating Comparative Visualizations Into the DR/MR Context*. We have integrated some commonly-used visualization components/techniques into the framework, e.g., bar charts, line charts, word cloud, ingredient glyph, small multiples, F+C techniques, etc. One of the most important criteria to select the visualization types is whether they are general-purposed, and whether they are simple or advanced. All the related functions are encapsulated into the APIs of the framework.
- *Challenge IV: Database Construction of Augmented Information of Target Objects*. The details about how to build the database of augmented information are described in Section 1.4 of the Appendix, available online.
- *Challenge V: Enhancing the Lighting Environment in the Real World*. In practical applications, it is important to reduce the interference of reflect light on the physical objects, which would probably decrease the OCR recognition rate. The solution is to capture multiple frames with a time interval (e.g., 0.5 seconds), when the camera is scanning, then synthesize the captured images to restore the reflect regions.

The workflow of DRCmpVis is shown in Fig. 2. For detailed information about the implementation, please refer to the Appendix file, available online.

A. Technical Implementation

(1) The Front-End Development Platform: To make the implementation more scalable, we have encapsulated the device-dependent APIs of DR/AR/MR for different mobile devices. For example, either ARKit [44] or ARCore [45] is employed to encapsulate the APIs for different mobile device platforms. The device-dependent APIs include:

Device Positioning: ARKit/ARCore provides the APIs for achieving the real-time position M of the mobile device in the physical space.

Distance Measurement: the platform can provide real-time distances between the mobile device. The position of the device and the distance can be used to build a coordinate system in the

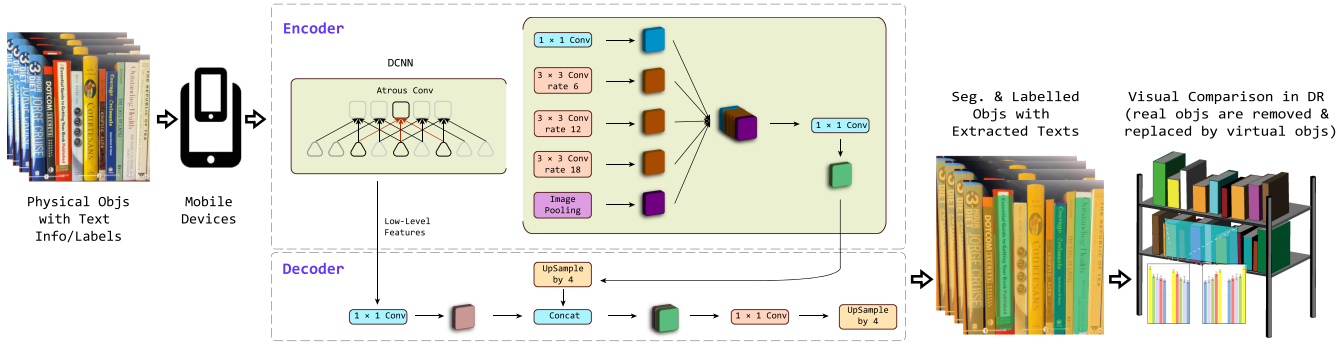

Fig. 2. Workflow of the proposed DRCmpVis. We illustrate it using the library/bookstore case. Regarding the deep neural network used in image segmentation and text recognition, the encoder module encodes multi-scale contextual information by applying atrous convolution at multiple scales, whereas the simple yet effective decoder module refines the segmentation results along object boundaries.

physical space. The distance can be measured by a camera with LiDAR scanner [44].

Object positioning: the APIs can be used to achieve the real-time positions of an object in the physical space, if it did appear in the captured image. In short, we use two types of device APIs for positioning in the physical space, including device positioning and object positioning.

(2) Breadth-First Search and Two CNN Platforms: Image Segmentation CNN and Optical Character Recognition CNN: We use image segmentation deployed on the server to recognize objects in the images sent from the mobile devices. The segmented object image is labeled and sent back to the mobile devices, facilitating object presentations within the DR/MR space. In fact, we initially used the breadth-first search (BFS) algorithm to finish image segmentation and recognition. However, the BFS algorithm is based on RGB values, it shows high constraints in the actual scenarios, including lighting, spine design, etc. In addition, the assumption itself has a strong limitation: many objects do not have regular color separation. This means that the same algorithm is difficult to apply to various scenarios. Finally, we adopted deep neural networks to achieve automatic image segmentation and labelling and text recognition, aiming to support various scenarios.

To get a better result in various scenarios, we apply a trained CNN-based open-source platform named PaddleSeg [11] to do image segmentation and labelling. PaddleSeg is one of the state-of-the-art deep learning models for semantic image segmentation, whose goal is to assign semantic labels to every pixel in the input image. In PaddleSeg, DeepLab [46] is one of its key modules. Therefore, we take DeepLab as an example to illustrate how PaddleSeg is integrated into DRCmpVis, as shown in Fig. 2. The encoder module encodes multi-scale contextual information by applying atrous convolution at multiple scales, whereas the simple yet effective decoder module refines the segmentation results along object boundaries.

The image (or panoramic image) we capture or the real-time video we record is input into the first network (the top left of Fig. 2), while the labeled samples are input into the second network (the top right of Fig. 2). Regarding the text extraction, we use the traditional CNN-based optical character recognition approach, following a language adaptive design [47], to recognize a large amount of text characters over numerous objects in reality.

(3) Real-Time Position Update: In scenarios such as libraries or bookstores, where hundreds or even thousands of objects are involved, updating all the objects' positions for each frame is challenging. In our implementation, we track the positions of the target objects in real-time, because the processed objects may be moved in the physical space. For example, the coffee menus would probably be moved in a cafe, or the mobile device is often moved when in use. Real-time tracking facilitates the positions of virtual objects to be updated accordingly.

In our implementation, we segment the captured images into multiple blocks by CNNs, and then track the objects in blocks by the image detection algorithms provided by the encapsulated APIs. The real-time tracking animation of the objects (such as the coffee menu) can be viewed in the supplemental video of the submission.

B. Database Construction of Augmented Information

We created a large database on the server for all application scenarios that require real-time information feedback [48], [49]. The database contains additional information on different attributes of the objects. For example: (1) Global book database. More than two million books were created on the server of DRCmpVis, making it easy to quickly find the ISBN, title, author, author introduction, abstract, publisher, cover image, pages, tags, etc. The book dataset was downloaded from the open data website "Amazon product data" [48], [49], [50], containing product reviews and metadata from Amazon, including 142.8 million reviews for their products and 22.5 million reviews for books. The Amazon database was last updated in 2018. (2) Coffee database. The coffee database was created by the Web crawler, which crawled collections from well-known coffee websites. For example, coffee data comes from Starbucks, including the coffee's name, description, ingredients list, preview image, and process introduction. The details about how to build the database of augmented information for all application scenarios are described in the Appendix, available online.

Fig. 3. Usage scenario in a library: a user searches and compares candidate books progressively in a library by DRCmpVis. (a) Scan the original physical bookshelves with 778 books. (b) DRCmpVis shows 118 books in the DR/MR environment after fuzzy searching "economic" via voice input. (c) Re-group them by publisher and search "Chicago". Books from "University of Chicago Press" and other publishers are placed on different layers. The user browses those books with a fisheye effect. (d) Further search with the keyword "social" in each publisher group, results are highlighted in red. (e) Re-rank those books by ratings. Books sorted in descending order are placed from the left to the right. (f) Select several candidate books, which are moved to a reserved layer of the bookshelves automatically. (g) DRCmpVis shows candidates by a word cloud of abstracts, introductions or comments. (h) Compare candidates by rating and price via bar chart. (i) Choose the target and restore all books to their original physical layout, search the target by its book name, and the target book is highlighted (glittered) in red. (j) Approach the target book and fetch it according to its location on the screen.

C. Integrating Visual Comparison Components Into a DR/MR Context

Regarding the visual comparisons of the additional attributes (augmented information), the related data is sent to the server and the client receives the processed data from the server. We designed several visual comparison components like*bar chart*, *line chart*, *word cloud*, *ingredient glyph*, etc., which can be chosen and composed by users in different example scenarios. We also employ *small multiples* to gain juxtapositions from the comparative data presentations, which are appreciated by the participants in the user study. Besides, we adopt a *focus+context* exploration scheme by using the *fisheye* algorithm, which scales the size of objects according to its distance to the focus one. It helps to magnify the target object among numerous objects, e.g., a candidate book among hundreds of books. Furthermore, we create a virtual translucent screen in the DR environment to show those additional attributes.

V. Example Scenarios

We illustrate how DRCmpVis facilitates visual comparisons for physical objects in a DR environment in this section.

A. Library/Bookstore Scenario

Suppose Zelda is a student majoring in economics. She prefers books from the "University of Chicago Press", which is recognized as having been publishing high-quality books. She comes to the social science area in a library/bookstore, facing several bookshelves with around a thousand books, as shown in Fig. 3(a).

(1) *Fuzzy filtering*: she scans the bookshelves by the panoramic camera of her tablet with DRCmpVis installed. 778 books are scanned and recognized in total. She then filters unrelated books by saying "economic" via voice input of the mobile devices. DRCmpVis deals with the input voice and filters those books by fuzzy search. Seeing that only 118 economic books remain, Zelda chooses to visualize those books in the DR/MR space and browses them as shown in Fig. 3(b). She finds that only one book nearby is from "University of Chicago Press", then she wants to find more books on "economics" and published by "University of Chicago Press".

(2) *Re-grouping*: she re-groups those 118 books by publisher and searches by saying "Chicago" or enter it by the virtual keyboard of her tablet. This time, seven books from the "University of Chicago Press" are highlighted and placed on a bookshelf in front of her with a fisheye effect (Fig. 3(c)). Books from other publishers are also grouped and placed on the other layers of the shelf, so she chooses a book from them.

(3) *Fuzzy re-filtering*: she wants to re-filter the books with the fuzzy keyword "social", there are 21 books highlighted in red (Fig. 3(d)). She uses the fisheye to view each book's details including titles or authors similar to Fig. 3(c). But she finds these social books not highly rated or the authors are not on her favorite author list. Consequently, she shifts her approach and decides to re-rank the books based on their ratings.

(4) *Re-ranking*: she sorts all of the books which are placed from left to right on the same layer of the shelf in descending order (Fig. 3(e)). Then she selects four books that seem suitable, those selected books are moved to a reserved layer of the virtual bookshelf which is designed to place the candidate books (Fig. 3(f)), just like a virtual shopping cart.

(5) *Comparing by word cloud in small multiples*: she views and compares the word cloud of each book's keywords. Among those four books, one book has keywords "story" and "understand", other books' keywords are "city", "environmentalist"

Fig. 4. Usage scenario in a cafe: a user builds visual comparisons for a coffee menu. (a) Scan the coffee menu. (b) Search "Latte". Three coffees are found and highlighted. (c) View the results by the fisheye. The focused coffee is magnified, with its augmented information shown beside it. (d) Re-group all the coffees by sugar content intervals. (e) Select four candidate coffees. They are moved to the right side of the menu. (f) Compare candidate coffees by their ingredient graphs in small multiples. (g) Re-group coffees by fat. (h) Re-rank coffees by calories. Coffees with more calories are moved to the left side, while those with fewer calories are moved to the right. (i) Compare the word cloud of the candidate coffees. (j) View coffees on the right side to choose one with fewer calories.

and "empire" (Fig. 3(g)). Zelda is interested in the "story" and the "empire" one, but she is also concerned with the prices if she is going to buy the book in a bookstore.

(6) *Comparing with diagrams in small multiples*: so she decides to compare both the ratings and the prices of these books via bar charts (Fig. 3(h)). She discovers that the "story" (the first) rated just as highly as the "empire" one (the fourth) but is slightly less expensive. Consequently, she opts for the "story" edition and restores those books to their original layout.

(7) *Title precise searching*: finally, she searches for books with the title "human resources management" by voice input or text input. The book is magnified and highlighted on the left upper side (Fig. 3(i)) by flashing. She walks and locates the book in the physical reality space according to its position shown on the screen (Fig. 3(j)).

B. Cafe Scenario

To demonstrate that the proposed framework can support different scenarios where the objects are labeled with texts or presented as texts, in this section we demonstrate another example scenario: a coffee shop.

A new coffee shop opens on Zelda's campus. She does not know much about coffee, but she is willing to try several in the new coffee shop. She walks into the coffee shop and takes a picture of the coffee menu using DRCmpVis. Soon she scans 40 different drinks, and DRCmpVis recognizes them and shows them on a virtual menu in the DR/MR context.

The virtual menu consists of 40 virtual objects which are presented as texts (e.g., coffee names) and the background texture of the original menu, which can be obtained by image segmentation, image labeling, and text extraction using neural networks DeepLab [46] and PaddleSeg [11]. The original menu in the physical world is replaced by the virtual menu, whose positions can be updated in real-time along with the original one. The real-time tracking animation of the coffee menu can be explored in the supplementary video of the submission.

Zelda remembers that she ordered a cup of espresso once before, which she thinks is rather bitter, so she wants to see the ingredients. She first voice inputs "Latte" and finds that it is highlighted in the menu (Fig. 4(b)). She checks the detailed ingredients of the latte and learns that most of the lattes contain too much milk. She further explores the menu by ingredient glyphs and finds "Espresso" is typically bitter, as no sugar is added to it (Fig. 4(c)).

Zelda then re-groups the coffees according to sugar (Fig. 4(d)–(e)). She browses and selects several drinks with high ratings in the "medium sweet" and "sweet" groups, as shown in Fig. 4(f). Then she compares those drinks' ingredients in small multiples, and finds that Cappuccino has a balance among sugar, milk, and caffeine, which may suit her taste, as shown in Fig. 4(g). However, her fitness coach's advice crosses her mind emphasizing that she needs to limit her calorie intake to 1,300 calories every day. However the coffee summary shows that Cappuccino has 140 calories per cup. So she re-ranks all the drinks by calorie content. This time, the coffees are sorted from left to right by calories, as shown in Fig. 4(h). She begins browsing on the right hand side, where coffees with relatively low calories are located. She finds several coffees that she has never tasted. To have a quick grasp of them, she views their word cloud (Fig. 4(i)). She learns that Blonde Roast is regarded to be "mellow" in the word cloud, Iced Coffee is "rich", and Caffee Americano has the keyword "espresso", which may be too bitter for her. She browses Blonde Roast's summary, which confirms that it only contains five calories per cup (Fig. 4(j)). Finally, she chooses Blonde Roast and enjoys its "soft and mellow flavor" described in the summary. We note that DRCmpVis can also handle larger menus like a big poster hanging on the wall outside the coffee shop, as shown in Fig. 5.

VI. Evaluation: User Study and Performance

In the evaluation, we aimed to assess DRCmpVis regarding the following aspects: (a) whether visual searching/filtering of

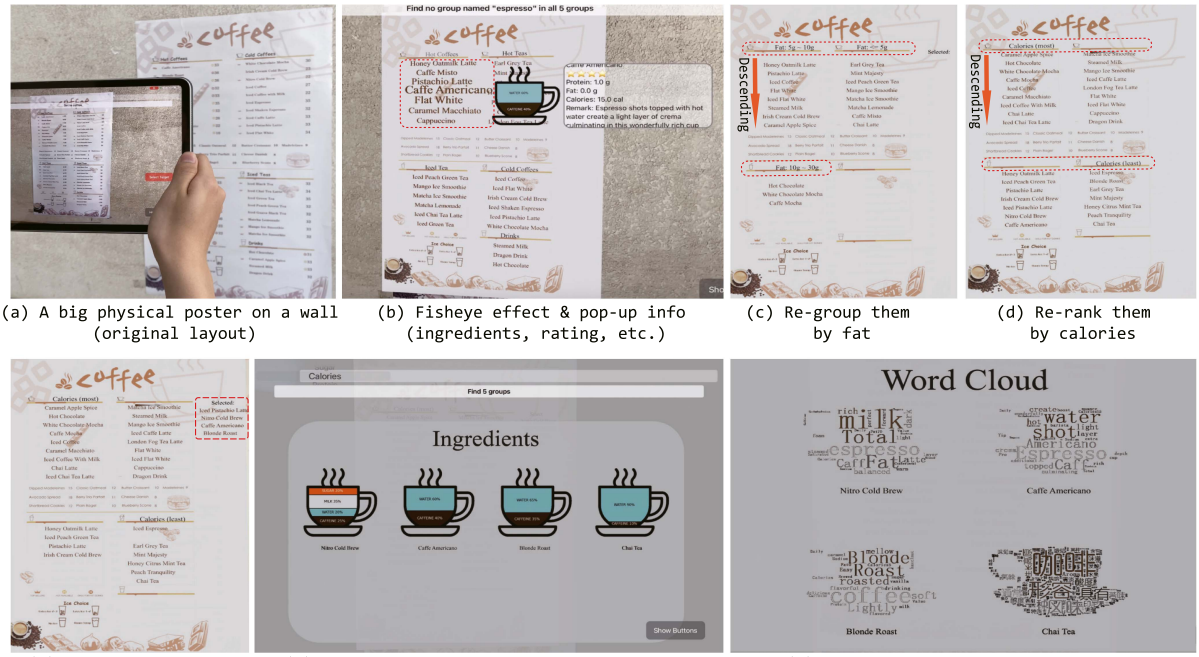

Fig. 5. Usage scenario outside a cafe: a user compares candidate coffees by augmented information from a big poster. (a) The original poster hanging on a wall outside a coffee shop. (b) View coffee's augmented information with a fisheye. (c) Re-group coffees by fat content intervals. (d) Re-rank coffees by calories. (e) Select four candidate coffees with relatively lower calories. (f-g) Compare candidate coffees by ingredients (f) or word cloud (g) and choose one.

DRCmpVis is helpful for users to compare and locate targets (G1); (b) whether visual re-grouping and re-ranking satisfy users' requirements on object comparison (G2, G3); (c) whether the augmented information provided in MR is useful and expressive (G4).

We conducted four measures, including subjective measures and objective measures:

- *Questionnaire Study:* a 5-point Likert scale was utilized to gauge and assess the comprehensive functionality of DRCmpVis. The questionnaire study is task-driven. We use $T\,i$ to name the task that happens in the i th scenario, i.e., *T1* is designed to evaluate the library usage scenario, while *T2* is for the coffee shop scenario.
- *NASA-TLX:* a 21-point Likert scale used to measure mental demands, physical demands, temporal demands, effort, performance, and participant's level of frustration by comparing DRCmpVis with two traditional methods.
- *Open Questions:* solicited to gain a general assessment of the technology proposed by us, intuitiveness, practicality, suggestions for improvement, and comparisons with traditional methods.
- *Quantitative Evaluation:* performance and accuracy measurements of each module of DRCmpVis, including the modules for scanning, image segmentation, and processing. The baseline methods of the quantitative evaluation are two control group methods, including a method for blind finding (without any tools), and a method using a database (DB) retrieval tool. In the library usage scenarios, DB-based tools can be often provided by a library, thus in our evaluation, we use the tool provided by our university library.

A. Study Design

Questionnaire. The questionnaire comprised a series of questions meticulously crafted with a 5-point Likert scale, spanning from 1 (indicating strong disagreement) to 5 (indicating strong agreement). We recruited 22 participants to take part in this study through a volunteer recruitment platform (10 males and 12 females) from 18 to 26 years old, they are from ten different majors of the university.

Procedures. Before starting the tasks, participants were required to fill in the pre-study questionnaires. We discovered that the majority of participants were not familiar with XR technologies, but most of them had experience choosing coffee at coffee shops and searching for books in libraries. Frequently, individuals encounter chaotic situations in their daily lives, such as dealing with a substantial quantity of disordered or unorganized books. In such cases, locating a specific target book proves to be a challenging endeavor. Most of our participants had prior experiences where they had trouble finding books in libraries where books are sorted by traditional index numbers.

Regarding the coffee scenario, most of the participants also, at some time, had experienced confusion when choosing different coffees. It was difficult for them to distinguish different coffees according to their approximate ingredient information. Also, recalling whether a particular coffee variety includes milk, cream, and sugar, as well as comparing the caloric content of

two distinct cups of coffee, had proven to be arduous for them. The pre-study questionnaire results also indicate that 98.7% of participants disagree with the notion that coffee shop staff will offer a retrieval system for use. Additionally, 86.7% of participants report that coffee shop staff are unlikely to provide specific ingredient information for comparison.

In the pre-study survey before the questionnaire step, we had one question to survey how many users liked simple visualization tasks or advanced visualization tasks in the DR applications. The survey result indicates that 98.8% of participants prefer a DR app with simple and easy to use visualizations instead of advanced visualizations. Thus, regarding the example apps built by DRCmpVis, we only integrated some commonly-used simple visualization components/techniques into the framework, e.g., bar charts, line charts, word cloud, ingredient glyph, small multiples, F+C techniques, etc. All the related functions are encapsulated into the APIs of the framework.

Following, the investigators introduced the capabilities and usage of DRCmpVis. In *T1* and *T2*, the investigators showed a simple example to the participants first and then released the specific task. After all the tasks were completed, the participants were asked to complete the post-study questionnaires. All participants received compensation of equal value regardless of their performance.

Free exploration. Participants were encouraged to explore DRCmpVis freely before the study. They could use search functions to filter the available books, regrouping them according to different attributes, such as the publisher or the range of publishing years. Additionally, participants had options to utilize re-ranking techniques to facilitate their comparison of ratings and prices. A free exploration step was designed to help participants get familiar with the UI and the functions of DRCmpVis.

B. User Study Tasks

T1 is divided into three subtasks. In *T1*, participants are required to find four different books in four bookshelves. In *T1-1*, participants need to search for the first book without using any tools. In *T1-2*, the task continues with three additional subtasks. In this task, participants can use the library retrieval system. In *T1-2-1*, the second book is placed in the correct position recorded in the library's database system. While in *T1-2-2*, the third book is inserted in a wrong position by other readers or librarians accidentally. *T1-2-3* involves searching for the fourth book, which is the last copy of this book in the library's inventory, however, it is read by someone else in the library. It means it is impossible for participants to find the fourth book. In *T1-3*, participants use DRCmpVis to find the four books in sequence, which are randomly inserted in different bookshelves after *T1-2*. All of the timing results are recorded. After completing *T1-3*, the participants are suggested to use the re-grouping facility of DRCmpVis to find other books with identical keywords (G1) and publishers (G2) and re-rank them by sorting the ratings or prices of the result books (G3). Finally, they can use DRCmpVis to find the books they wish to read.

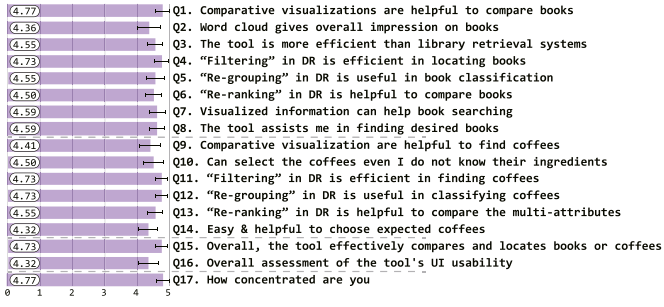

Fig. 6. Post-study result: most of participants react positively to DRCmpVis.

T2 requires participants to search for different types of lattes from the physical coffee menu, and next they need to re-group all the lattes by sugar content (G2), re-rank them by calories (G3), and then find the one with the least calories according to the ingredients (G4) visualized by DRCmpVis, as shown in Fig. 4. After that, participants can also check the menu and select other coffees they are unfamiliar with. They can compare them in the MR context using ingredient glyphs and word cloud, as shown in Fig. 5(f).

C. User Study Results

We analyze the collected quantitative and qualitative results. The questionnaires can be divided into four parts, i.e., the library case, the cafe scenario, overall evaluation and UI, and the involvement, as shown in Fig. 6. From the evaluation point of view, the questionnaires can be divided into usability, expressiveness, effectiveness, involvement, and suggestions from the participants.

a. Usability: According to our study, most participants gave positive feedback on the overall evaluation of DRCmpVis (Q15: $\mu = 4.73,\ 95\%\ CI = [4.53, 4.92]$, G1). In particular, apart from the UI design itself (Q16: $\mu = 4.32,\ 95\%\ CI = [4.00, 4.64]$, G4), the participants also appreciated the voice input, fisheye effect, and result highlighting. They said that these designs made the interactions smooth and intuitive. From the questionnaire results, we can find that the participants could readily search targets and compare candidate targets by using the VR design and comparative data presentations, respectively.

Regarding the usability evaluation about the two scenarios, i.e., the library/bookstore scenario (Q1: $\mu = 4.77,\ 95\%\ CI = [4.54, 5.01]$, G4, and Q3: $\mu = 4.55,\ 95\%\ CI = [4.32, 4.77]$, G1) and the cafe scenario (Q14: $\mu = 4.31,\ 95\%\ CI = [4.03, 4.60]$, G4), the participants gave high praise, because they thought DRCmpVis was intuitive to use in DR scenarios.

b. Expressiveness: According to the cafe scenario (Q9: $\mu = 4.41,\ 95\%\ CI = [4.08, 4.73]$, G4) and the library/bookstore scenario (Q2: $\mu = 4.36,\ 95\%\ CI = [4.01, 4.71]$, G4, and Q7: $\mu = 4.59,\ 95\%\ CI = [4.33, 4.85]$, G4) bar charts, word cloud, small multiples efficiently aided participants in developing a comprehensive understanding of physical objects. The participants also noted that the comparative ingredient glyphs significantly contributed to forming comprehensive impressions of the distinctions among the various types of coffees and books.

TABLE II
PERFORMANCE AND ACCURACY EVALUATION OF DRCMPVIS (SECONDS)

Scenario	Scanning Time	Segmentation Time	Processing Time	Segmentation Rate
Library Scenario	3.873	6.439	7.577	95.13%
Cafe Scenario	0.693	4.215	5.546	100.00%

The"Scanning Time" is the average time to scan a bookshelf in the library scenario and a menu in the cafe scenario, respectively. The "Segmentation Time" is the average time to segment images within one server request. The "Processing Time" is the total time of segmentation and labeling provided by the cloud service provider, while the "Segmentation Rate" is the average accuracy of image segmentation. We obtain the average results are computed by more than 10 tests.

c. Effectiveness: The participants responded positively and confirmed the effectiveness of filtering (Q4: $\mu = 4.73$, $95\%\ CI = [4.53, 4.93]$, G1), re-grouping (Q5: $\mu = 4.55$, $95\%\ CI = [4.28, 4.81]$, G2) and re-ranking (Q6: $\mu = 4.50$, $95\%\, CI = [4.24, 4.76]$, G3) books in a DR environment.

Compared with blind finding, the time cost was reduced from an average of 4.56 to 0.45 minutes for each book with the help of DRCmpVis. One notable exception came from a participant, who is a temporary librarian where the tasks took place. He spent only 5 seconds finding one of the target books in the physical library. We revisited him and he said "I happen to be familiar with this bookshelf and DRCmpVis is indeed useful for the public, which can significantly reduce my workload as a librarian".

Participants could select the coffees even if they did not know their ingredients (Q10: $\mu = 4.50$, $95\%\, CI = [4.20, 4.80]$, G1). Similarly, they also responded positively on filtering (Q11: $\mu = 4.73$, $95\%\, CI = [4.54, 4.91]$, G1), re-grouping (Q12: $\mu = 4.73$, $95\%\ CI = [4.54, 4.91]$, G2) and re-ranking (Q13: $\mu = 4.55$, $95\%\, CI = [4.30, 4.79]$, G3) coffees in a DR context. Most participants found the subsequent visual comparisons helpful for them as they did not know much about the ingredients of coffees on the menu. "It helps a lot especially when someone cares about fat intake and obesity" said one participant.

d. Involvement: As indicated by Q17 ($\mu = 4.77$, $95\%\ CI = [4.58, 4.96]$), almost all participants felt engaged when carrying on the tasks. They all believed that the tasks were quite smooth and interesting.

D. NASA-TLX Measures

We further evaluate DRCmpVis by comparing it with two traditional methods as control groups based on NASA-TLX measurements, i.e., target blind finding without any tools, and target finding by database retrieval system (DB finding). We recruited another 22 participants to take part in this study through the same volunteer recruitment platform, who were randomly selected from ten different majors of the university.

A repeated measure analysis of variance (ANOVA) on the NASA-TLX questionnaire demonstrated significant main effects for the three technologies in terms of physical demand ($F_{2,63} = 98.7303, p < 0.001, \eta^2 = 0.758$), effort ($F_{2,63} = 97.07, p < 0.001, \eta^2 = 0.755$), and frustration ($F_{2,63} = 61.78, p < 0.001, \eta^2 = 0.662$), as shown in Fig. 7. It is worth mentioning that the mental demand of blind finding is significantly lower than for the other two methods, because blind finding is the simplest approach which has the smoothest learning curve. In contrast, it requires some learning to master the DB-based searching tool and DRCmpVis, with DRCmpVis requiring the highest mental effort. However, for all other metrics, this trend is reversed. The physical demand in DRCmpVis is significantly lower than in the other two methods (all $p < 0.001$). The temporal demand of the two traditional methods (all $p < 0.001$) are significantly higher than that of DRCmpVis ($p < 0.001$). Because the timing results of DRCmpVis are much better than the other two, as shown in Tables II and III. The physical demand of DRCmpVis is also significantly lower than DB finding ($p = 1.9E - 09$). Similarly, a diminishing pattern in users' temporal demand is evident in the three techniques: "Blind finding"-"DB finding", $p = 5.8E - 06$; "DB finding"-"DRCmpVis finding" $p = 3.1E - 12$; "Blind finding"-"DRCmpVis finding", $p = 2.7E - 19$. The frustration demand follows the same pattern ("Blind finding"-"DB finding", $p = 1.6E - 02$; "DB finding"-"DRCmpVis finding", $p = 3.0E - 09$; "Blind finding"-"DRCmpVis finding", $p = 2.7E - 17$).

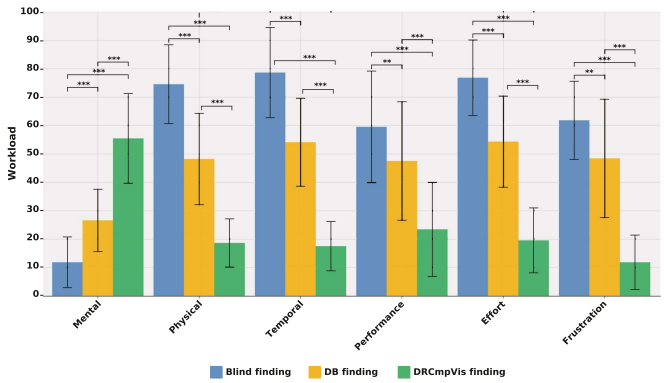

Fig. 7. Scores of NASA-TLX evaluation for two control groups of methods and the proposed DRCmpVis. Error bars indicate standard errors. Statistical significant differences are denoted by ** (p <0.01), *** (p <0.001).

Suggestions From Open Questions: Feedback and suggestions were collected from the evaluation, which are listed as follows:

Several participants thought that shifting from virtual space of DR to physical space was quite useful for them to find candidate targets. P6 noted: "The fisheye deformation of books makes them overlapped and cluttered". Considering the density of books on the shelves in the library/bookstore, a possible alternative is pushing away nearby books to enhance the current fisheye deformation. Moreover, some participants suggested that it would be beneficial to include visual cues indicating the physical directions of the target book when searching for multiple books from different bookshelves.

Overall, most participants expressed a strong preference for DRCmpVis compared with the other two traditional methods. There were also participants who commented in the open-ended responses that DRCmpVis, was convenient, efficient, and relatively easy to learn, requiring less effort to locate target objects.

TABLE III
TASK-DRIVEN QUANTITATIVE EVALUATION RESULTS (MINUTES)

Methods	Book Searching Time	Latte Search Time	Augmented Info	Target Comparison
Blind finding (without any tools)	4.56	NA	No	No
With a DB retrieval system (targets are available on the correct shelf)	2.53	NA	No	Retrieve books with given keywords
With a DB retrieval system (targets are inserted in wrong positions)	5.34	NA	No	Retrieve books with given keywords
With a DB retrieval system (available but being viewed by other borrowers)	unlimited	NA	No	Retrieve books with given keywords
The proposed DRCmpVis	0.45	0.13	Color highlight Fisheye highlight Pop-up glyphs	Re-grouping Re-ranking Visual comparison

We compare the proposed DRCmpVis with the traditional two methods. The results show the participants' time costs in a task involving finding a target (e.g., a book with a given keyword) using different tools/methods. We recruited 22 participants to participate the experiments. All retrieval times represented the average time taken to find the target object. "Latte Search Time" refers to the average time to search for the keyword "Latte" (eight Lattes in total in the menu). Note: most coffee shops do not provide a retrieval system for users, thus they are marked as Not Available (NA).

E. Quantitative Evaluation

For the sake of achieving a dependable and consistent server service, we chose to deploy the back-end server on a non-free cloud platform in our experimental setup. The virtual cloud resources are limited in our experiment due to their expensive charges. The configuration of the cloud service we paid for is Intel Xeon Platinum 6271 (dual-core) running at 2.60 GHz and 4 GB memory. The mobile device of all the experiments of this paper was an iPad Pro, with DRCmpVis installed. It is worth noting that the hardware configuration can be improved for more expensive cloud service packages.

An important module of DRCmpVis is the image segmentation and labeling, which is provided by a trained CNN platform named PaddleSeg [11]. In our experiments, both the PaddleSeg and the database with augmented information are built on the cloud server. We found that the average image segmentation rates of DRCmpVis for all the example scenarios were larger than 95.0% (the additional example scenarios are moved to the Appendix due to the page limit of the paper), available online. The quantitative evaluation results are shown in Table II.

The timing performance of DRCmpVis was also obtained in the study. It just included the library scenario because most coffee shops do not provide a retrieval system for users. In the experiment, all the participants were allocated to different time slots. They filled in their personal information on the questionnaire first on a library table, and then they were informed how to use the control group methods and DRCmpVis. After that, they could try all the tools freely until they were familiar with them. Each participant used DRCmpVis and the control group methods to finish the book finding tasks in a random order. The detailed following sub-task steps were the same as that in task **T1**, as described in Section VI-B.

It is worth mentioning that the target books are different in all sub-tasks of **T1** for each participant, as described in Section VI-B. Additionally, the distances between the target books and the starting position (a library table) of the participants are similar. We got the timing results when the sub-tasks were performed. The test results are shown in Table III. We can find the task *T1-3* assisted by DRCmpVis takes about 0.45 minutes on average to find a book, which is only 9.9% of the search time used in the blind finding method.

We also summarize some feature comparisons in Table III. For example, DRCmpVis can provide much more augmented info by highlighting in the MR space and offering pop-up glyph displays adjacent to the corresponding objects in the MR space. The candidate targets can be compared by visual comparison components and small multiples in MR space, according to their additional nominal, ordinal, and quantitative attributes.

VII. DISCUSSION AND FUTURE WORK

We summarize the scalability issues, alternative designs, and some limitations of DRCmpVis as follows:

The Scope of the Application Scenarios of DRCmpVis: In addition to the illustrative scenarios outlined in the paper, the current iteration of DRCmpVis accommodates a range of diverse application scenarios. These scenarios involve objects with textual labels or textual information, such as menus encompassing items like coffee, beverages, food items, and so forth. Additionally, the tool caters to use cases like supermarket goods featuring labels denoting names and prices or utilizing QR codes, among other possibilities. We have tested DRCmpVis on drinking menus and food menus in restaurants and found it also works well there. Besides, we find DRCmpVis can be easily extended to objects with colors such as eye shadows, colored balls in a large amusement park, colored goods in supermarkets, etc. For more details about the image-based case (eye shadow), please refer to the Appendix file, available online. The usage environment of DRCmpVis includes public places like a library, a bookstore, a cafe, etc.

In addition to voice input, we also provide text input by using a virtual keyboard integrated into the DR/MR interface to support the scenarios where users are inconvenient to make a sound, e.g., a public place that needs to be quiet or a noisy environment. Besides, it is difficult for users to capture real-time videos when users are in some crowded setting. In some cases, libraries will be influenced by the crowded environment, but in other cases, such as the cafe case, this factor is irrelevant.

DRCmpVis is not feasible for libraries or bookstores when the book information is unavailable to fetch, or it is hard to download or crawl from the Internet, e.g., the libraries with ancient books, because the framework will query additional augmented information from the constructed database according to the information scanned from the physical objects. Another application issue is that DR may cause users to bump into bookshelves because it replaces the physical objects with their virtual avatars. Fortunately, the mobile devices are typically not

large enough to obscure a user's view. A potential solution is to add visual cues about the boundaries of certain physical objects in future work.

Scalability Issue on Image Segmentation: It is worth noting that the image segmentation components of DRCmpVis are scalable and not limited by the number of objects, because the CNN and the OCR algorithm are run on the server which can handle even thousands of books in the library scenario in our experiments. More importantly, unlike a mobile device, the computation resources of the server are scalable enough and could be easily upgraded. As a result, whereas DRCmpVis recognizes almost all of the books scanned by the user, we recommend the user to first filter out unrelated books by fuzzy searching before actually visualizing those books in the DR/MR space in order to narrow down the data space.

The Limitation of Text Recognition: DRCmpVis recognizes objects by images taken from mobile devices. Ideally, the user only needs to take one panoramic image or animation that contains all the objects. However, objects are often not recognized if they are too small in the image, which occurs when users are standing too far away from numerous objects. For example, in the library/bookstore scenario, the recognition results are often not reasonable, even if the latest open source version of OCR deep neural networks [11], [46] were used. A potential solution can be for users to walk closer to the bookshelves and scan one layer at a time by panoramic stitching. A similar approach is also advised when there are poor light conditions or limited imaging quality.

The image segmentation and labeling service needs to be requested once due to an image recognition module on the client app of DRCmpVis, when the positions of the objects are not changed. We plan to use a buffer strategy and a front-end image recognition module to accelerate the text recognition processes from the panoramic images or the captured videos. The image recognition module will verify whether the newly captured panoramic image is saved in the buffer. If yes, the segmentation and labeling records in the buffer can be reused without requesting the server twice. However, it may take some time for us to construct the record buffers when DRCmpVis is first used in a scenario environment. Thus, the tool would be much more efficient after the first-time buffer construction in a new scenario environment.

Possible Performance Improvement: To get a stable and reliable service of the server, we deploy the server part on a non-free cloud in our experiment, as described in Section VI. The hardware configuration can be improved for more expensive service packages. Thus maybe the performance especially for the segmentation and labeling could be further improved.

In the future, we plan to apply DRCmpVis in more general usage scenarios in our daily lives, and extend the usage of DRCmpVis to other scenarios, such as choosing cups, fruits or flowers. This seems feasible since objects with text on them or in different colors and shapes can be well recognized by trained neural networks. However, objects with irregular 3D shapes and without textual information on them are difficult to be recognized by current algorithms including state-of-the-art neural networks. Finally, we also plan to extend DRCmpVis to both mobile devices and HMDs, if the cybersickness problems are significantly alleviated.

VIII. Conclusion

In this article, we introduce a novel DR/MR-based application framework called DRCmpVis. This framework is specifically designed to facilitate visual comparisons among multiple physical objects featuring text labels or textual information. The framework seamlessly integrates efficient data computation in virtual space with in-context interaction in the physical space. It enables multidimensional comparisons of candidate objects, leveraging their nominal, ordinal, or quantitative attributes.

Users initiate the process by capturing panoramic photos in the real world using the cameras of mobile devices. To streamline the search, users can input a fuzzy search keyword into the nominal attributes of objects through voice or text commands, tailoring the results to their preferences. The search outcomes are highlighted in the DR environment, using color and deformation cues to indicate their real-world positions. Moreover, users can dynamically re-group or re-rank candidates based on their multifaceted attributes. To enhance the comparative experience, additional augmented information about the objects can be seamlessly integrated into the identical MR context.

Acknowledgments

The authors want to thank all the reviewers for their valuable comments. We would like to thank Chufan Lai, Jiacheng Zhang, Shiyuan Hong, Zhongyuan Mao, Rongxin Cang, Shuyu Bao, Ayush Kumar and the user study participants for their assistance in implementation or evaluation.

References

[1] C. Stolte, D. Tang, and P. Hanrahan, "Polaris: A system for query, analysis, and visualization of multidimensional relational databases," *IEEE Trans. Vis. Comput. Graph.*, vol. 8, no. 1, pp. 52–65, First Quarter 2002.

[2] J. Bertin, *Semiology of Graphics: Diagrams, Networks, Maps*. Madison, WI, USA: Univ. Wisconsin Press, 1967.

[3] P. Milgram and F. Kishino, "A taxonomy of mixed reality visual displays," *IEICE Trans. Inf. Syst.*, vol. E77-D, no. 12, pp. 1321–1329, Dec. 1994.

[4] M. Takemura and Y. Ohta, "Diminishing head-mounted display for shared mixed reality," in *Proc. Int. Symp. Mixed Augmented Reality*, Darmstadt, Germany, 2003, pp. 1–8.

[5] N. A. M. ElSayed, B. H. Thomas, R. T. Smith, and K. Marriott, "Using augmented reality to support situated analytics," in *Proc. IEEE Virtual Reality*, Arles, France, 2015, pp. 175–176.

[6] S. Mori, S. Ikeda, and H. Saito, "A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects," *IPSJ Trans. Comput. Vis. Appl.*, vol. 9, no. 17, pp. 1–14, 2017.

[7] G. Queguiner, M. Fradet, and M. Rouhani, "Towards mobile diminished reality," in *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct*, Munich, Germany, 2018, pp. 226–231.

[8] J. Mayor, L. Raya, and A. Sanchez, "A comparative study of virtual reality methods of interaction and locomotion based on presence, cybersickness, and usability," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 3, pp. 1542–1553, Third Quarter 2021.

[9] P. Caserman, A. Garcia-Agundez, A. G. Zerban, and S. Göbel, "Cybersickness in current-generation virtual reality head-mounted displays: Systematic review and outlook," *Virtual Reality*, vol. 25, pp. 1153–1170, 2021.

[10] A. Singla, S. Göring, D. Keller, R. R. R. Rao, S. Fremerey, and A. Raake, "Assessment of the simulator sickness questionnaire for omnidirectional videos," in *Proc. IEEE Virtual Reality 3D User Interfaces*, 2021, pp. 198–206.

[11] Y. Liu et al., "PaddleSeg, end-to-end image segmentation kit based on paddlepaddle," 2019. [Online]. Available: https://github.com/PaddlePaddle/PaddleSeg
[12] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts, "Visual comparison for information visualization," *Inf. Visual.*, vol. 10, no. 4, pp. 289–309, 2011.
[13] S. Zollmann, T. Langlotz, R. Grasset, W. H. Lo, S. Mori, and H. Regenbrecht, "Visualization techniques in augmented reality: A taxonomy, methods and patterns," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 9, pp. 3808–3825, Sep. 2021.
[14] D. Kalkofen, C. Sandor, S. White, and D. Schmalstieg, "Visualization techniques for augmented reality," in *Handbook of Augmented Reality*, Berlin, Germany: Springer, 2011, pp. 65–98.
[15] N. Kawai, T. Sato, and N. Yokoya, "Diminished reality based on image inpainting considering background geometry," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 3, pp. 1236–1247, Mar. 2016.
[16] S. H. Said, M. Tamaazousti, and A. Bartoli, "Image-based models for specularity propagation in diminished reality," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 7, pp. 2140–2152, Jul. 2018.
[17] S. Hashiguchi, S. Mori, M. Tanaka, F. Shibata, and A. Kimura, "Perceived weight of a rod under augmented and diminished reality visual effects," in *Proc. 24th ACM Symp. Virtual Reality Softw. Technol.*, New York, NY, USA, 2018, pp. 1–6.
[18] J. Herling and W. Broll, "PixMix: A real-time approach to high-quality diminished reality," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, 2012, pp. 141–150.
[19] Z. Li, Y. Wang, J. Guo, L.-F. Cheong, and S. Z. Zhou, "Diminished reality using appearance and 3D geometry of internet photo collections," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, 2013, pp. 11–19.
[20] S. Butscher, S. Hubenschmid, J. Müller, J. Fuchs, and H. Reiterer, "Clusters, trends, and outliers," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, Art. no. 90.
[21] R. Liu et al., "Interactive extended reality techniques in information visualization," *IEEE Trans. Human-Mach. Syst.*, vol. 52, no. 6, pp. 1338–1351, Dec. 2022.
[22] M. Sorokin et al., "Ring graphs in VR: Exploring a new and novel method for node placement and link visibility in VR-based graph analysis," in Proc. SIGGRAPH Asia Posters, 2018, pp. 1–2.
[23] M.-J. Zhang, J. Li, and K. Zhang, "An immersive approach to the visual exploration of geospatial network datasets," in *Proc. ACM SIGGRAPH Conf. Virtual-Reality Continuum Appl. Ind.*, 2016, pp. 381–390.
[24] O. Kwon, C. Muelder, K. Lee, and K.-L. Ma, "A study of layout, rendering, and interaction methods for immersive graph visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 7, pp. 1802–1815, Jul. 2016.
[25] S. K. Tadeja, T. Kipouros, and P. O. Kristensson, "Exploring parallel coordinates plots in virtual reality," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, 2019, Art. no. LBW2617.
[26] M. Cordeil et al., "IATK: An immersive analytics toolkit," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces*, 2019, pp. 200–209.
[27] N. Reski, A. Alissandrakis, and A. Kerren, "Exploration of time-oriented data in immersive virtual reality using a 3D radar chart approach," in *Proc. Nordic Conf. Hum.-Comput. Interact.: Shaping Exp. Shaping Soc.*, 2020, Art. no. 33.
[28] J. A. W. Filho, C. M. D. S. Freitas, and L. Nedel, "Comfortable immersive analytics with the VirtualDesk metaphor," *IEEE Comput. Graph. Appl.*, vol. 39, no. 3, pp. 41–53, May/Jun. 2019.
[29] J. I. Maletic, J. Leigh, and A. Marcus, "Visualizing software in an immersive virtual reality environment," in *Proc. Int. Conf. Softw. Eng.*, 2001, pp. 12–13.
[30] H. Wang, X. Chen, Z. Xia, H. Wang, X. Wang, and R. Liu, "iTDW: Immersive tiled display wall with clustering-driven layout," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, 2022, pp. 1–6.
[31] B. Xu, S. Guo, E. Koh, J. Hoffswell, R. Rossi, and F. Du, "ARShopping: In-store shopping decision support through augmented reality and immersive visualization," in *Proc. IEEE Visual. Vis. Analytics*, 2022, pp. 120–124.
[32] B. Bach, R. Sicat, H. Pfister, and A. Quigley, "Drawing into the AR-Canvas: Designing embedded visualizations for augmented reality," in *Proc. Workshop Immersive Analytics*, 2017.
[33] Z. Chen, Y. Su, Y. Wang, Q. Wang, H. Qu, and Y. Wu, "MARVisT: Authoring glyph-based visualization in mobile augmented reality," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 8, pp. 2645–2658, Aug. 2020.
[34] Z. Chen, W. Tong, Q. Wang, B. Bach, and H. Qu, "Augmenting static visualizations with PapARVis designer," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–12.
[35] B. Shneiderman, D. Feldman, A. Rose, and X. F. Grau, "Visualizing digital library search results with categorical and hierarchical axes," in *Proc. 5th ACM Conf. Digit. Libraries*, 2000, pp. 57–66.
[36] A. Thudt, U. Hinrichs, and S. Carpendale, "The bohemian bookshelf: Supporting serendipitous book discoveries through information visualization," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2012, pp. 1461–1470.
[37] N. A. M. ElSayed, R. T. Smith, and B. H. Thomas, "HORUS EYE: See the invisible bird and snake vision for augmented reality information visualization," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, Merida, Mexico, 2016, pp. 203–208.
[38] R. Sicat et al., "DXR: A toolkit for building immersive data visualizations," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 715–725, Jan. 2019.
[39] T. Rhee, S. Thompson, D. Medeiros, R. dos Anjos, and A. Chalmers, "Augmented virtual teleportation for high-fidelity telecollaboration," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 5, pp. 1923–1933, May 2020.
[40] N. ElSayed, B. Thomas, K. Marriott, J. Piantadosi, and R. Smith, "Situated analytics," in *Proc. Big Data Vis. Analytics*, 2015, pp. 1–8.
[41] B. Lee, D. Brown, B. Lee, C. Hurter, S. Drucker, and T. Dwyer, "Data visceralization: Enabling deeper understanding of data using virtual reality," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 1095–1105, Feb. 2021.
[42] B. Lee, X. Hu, M. Cordeil, A. Prouzeau, B. Jenny, and T. Dwyer, "Shared surfaces and spaces: Collaborative data visualisation in a co-located immersive environment," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 1171–1181, Feb. 2021.
[43] P. W. S. Butcher, N. W. John, and P. D. Ritsos, "VRIA: A web-based framework for creating immersive analytics experiences," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 7, pp. 3213–3225, Jul. 2021.
[44] ARKit, "Arkit," 2018. [Online]. Available: https://developer.apple.com/cn/augmented-reality/arkit/
[45] ArCore, "Arcore," 2018. [Online]. Available: https://developers.google.cn/ar/
[46] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.
[47] O. Y. Ling, L. B. Theng, A. Chai, and C. McCarthy, "A model for automatic recognition of vertical texts in natural scene images," in *Proc. IEEE 8th Int. Conf. Control Syst. Comput. Eng.*, 2018, pp. 170–175.
[48] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2015, pp. 43–52.
[49] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. Int. Conf. World Wide Web*, 2016, pp. 507–517.
[50] J. McAuley, "Amazon product data," 2018. [Online]. Available: https://jmcauley.ucsd.edu/data/amazon/

Richen Liu is an Associate Professor of Nanjing Normal University in China. He has published more than 30 papers, including ACM CHI, IEEE VIS, EuroVis, PacificVis, and the journals like *IEEE Transactions on Visualization and Computer Graphics*, *IEEE Transactions on Human-Machine Systems*, *IEEE Transactions on Big Data*, *Bioinformatics*, SPE, etc. His current research interests include XR + visualization, AI + metaverse, and HCI. He served as multiple committees and reviewers. For more information, please visit https://dabigtou.github.io/richenliu/.

Shunlong Ye was a student with Nanjing Normal University and a research assistant with DtXR research group. His research focus lies in the field of XR + visualization and medical image processing. Now he is with the Nanjing University of Science and Technology. He has published several papers on the conferences including ACM CHI and IEEE VIS (poster).

Zhifei Ding is currently working toward the graduate degree with Nanjing Normal University. He has published two papers on *IEEE Transactions on Big Data* and VINCI, respectively. His current research interests include XR + visualization and Big Data visualization.

Guang Yang received the undergraduate degree in computer science from Nanjing Normal University, in 2022. Now he is working toward the master's degree with Sun Yat-sen University. His research interests include XR + visualization and computer graphics.

Shenghui Cheng received the PhD degree in computer science from Stony Brook University, and conducted research with Brookhaven National Laboratory and Harvard Medical School. He is a westlake fellow and the director of the Intelligent Visualization Lab, WestLake University. He also served as a consultant for the World Bank and Cedar Sinai Medical Center, US. His research interests include visualization, visual analytics, AI and metaverse.

Klaus Mueller (Fellow, IEEE) is currently a professor of computer science with Stony Brook University and a senior scientist with Brookhaven National Lab. His research interests include explainable AI, visual analytics, data science, and medical imaging. He is won the US NSF Early CAREER Award, the SUNY Chancellor's Award for Excellence in Scholarship and Creative Activity, and the IEEE CS Meritorious Service Certificate. To date, his more than 300 papers have been cited more than 13,500 times.