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

Journal:	*Transactions on Visualization and Computer Graphics*
Manuscript ID	TVCG-2023-09-0532
Manuscript Type:	Regular
Keywords:	Visual comparison, Diminished reality, I.3.6.d Interaction techniques < I.3.6 Methodology and Techniques < I.3 Computer Graphics < I Computing Methodologies, Immersive visualization
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
01.DRCmpVis_Final_IEEEVIS23.mp4	

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

Abstract— Numerous physical objects in our daily lives are grouped or ranked according to stereotyped presentation style. For example, in a library, the books are normally grouped and ranked based on the classification number. However, for better comparison, we commonly need to re-group or re-rank the books with other attributes like their ratings, presses, comments, published years, keywords, prices, etc, or a combination of them. In this paper, 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 physical 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. It 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 user study, NASA TLX assessment, quantitative evaluation, and case studies using different scenarios.

Index Terms—Diminished reality, visual comparison, virtual avatars, mixed reality

1 Introduction

The development and popularity of extended reality (XR) devices and the techniques have led to an increasing number of studies designing new application tools. XR generally consists of virtual reality (VR), augmented reality (AR), and mixed reality (MR). MR is strictly defined by Milgram and Kishino [38], which was considered as a mixture of real and virtual objects within a single display. The distinctions between AR and MR are fuzzy [46]. 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 [17]. 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. [17] think SA is a new area of research at the intersection of visual analytics and AR. Besides, a new concept diminished reality (DR) [39,40] was further introduced recently. DR pertains to the manipulation of a perceived environment in real-time, involving actions like concealing, eliminating, or revealing objects [39,40]. According to the survey on DR [39] summarized by Mori et al., DR examples include four types: diminishing, seeing through, replacing, and inpainting real objects.

Stolte et al. [45] have summarized 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" [45]. Furthermore, according to Jacques Bertin's book "The Semiology of Graphics" [5], data types of visual variables include *nominal*, *ordinal*, and *quantitative*. To enable users in finding and comparing physical objects with multi-dimensional attributes. Considering these two principles [5,45], we further structure the data flow space in DR/MR context 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 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 DR/MR environment (available attributes: **ordinal**, **quantitative**).

In our everyday life, we often spend a great amount of time searching for a specific object from numerous candidates (e.g., searching for algorithm-related books in a library or a 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 a 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 Figure 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 probably 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 Figure 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 Figure 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.

Except for the example of finding/comparing targets from numerous candidates, 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 in a cafe.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

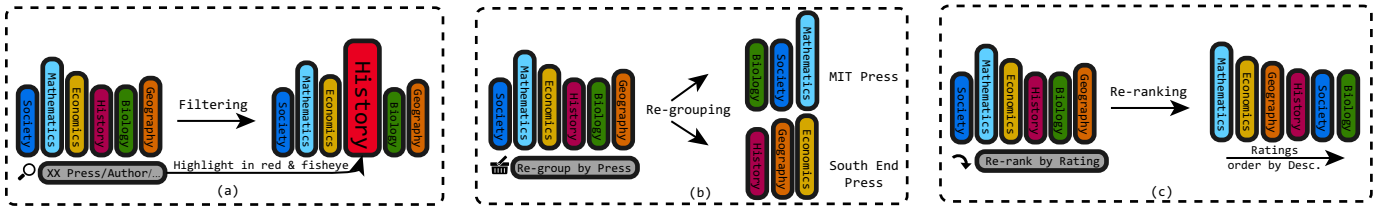


Fig. 1: Three types of data flow tasks within the DR/MR-based computational framework: (a) filtering, (b) re-grouping, and (c) re-ranking in DR environment. We take a library scenario as an example.

The tasks mentioned above in our daily lives present three main challenges. First, it is tedious for us to find the target objects from numerous candidates, especially when we only know some fuzzy information/keywords of the targets. Second, the object information visually presented on the physical objects is limited to help us compare the candidates progressively and then find the final targets. Third, the original physical layouts of the objects are often in a stereotyped presentation style and of little use in object comparison, e.g., the books on the bookshelves are often sorted by the classification number in libraries/bookstores while we often need to compare them using multi-dimensional attributes (publishers, rated scores, topics, keywords, prices, publication year, etc.).

To address the issues, we propose an interactive application framework named *DRCmpVis,* enabling visually compare numerous physical objects with text labels/information in mobile diminished reality. It builds multidimensional comparisons avoiding breaking the original physical layouts and provides additional augmented information by comparative data presentations in an identical context. 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.

In our work, *DRCmpVis* replaces the real objects with virtual objects, then we mainly used the term DR in this paper. Strictly speaking, plenty of virtual information of targets is also provided in the reality environment, thus we also use the term MR. With *DRCmpVis*, multi-dimensional comparisons can be completed by filtering, re-grouping, re-ranking, and their combinations in DR context. The additional augmented information of the objects can be encoded into some simple visual comparisons in MR context.

We use a trained convolutional neural network (CNN) named PaddleSeg [34] to segment and label all the objects. Furthermore, we extract the text information by an OCR-based neural network. In the experiment, 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 the three different types of attributes (nominal, ordinal, and quantitative), and then integrate commonly-used visualizations into 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.

2 Related Work

Visual comparison aims at providing visual support for the understanding of underlying abstract data sets [19]. The visual comparison tasks in this paper are a little different from the traditional ones because the compared items in *DRCmpVis* are physical objects.

2.1 XR-Based Data Visualization

There is not many literature that strictly defines the differences between VR, AR, and DR, while XR is often considered as consisting of VR, AR, MR, and DR. DR refers to the removal of physical objects from real-time video [39, 40]. In a narrow sense, it is different from AR, which shows the physical reality of the world. AR-based visualizations [27, 54] 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.

Embedded data representations are capable of linking systems to physical things [51]. As a significant method to connect digital data with physical world, XR can realize data presentation in the physical space to promote certain visual explorations and combine presentations with personal ideas and preferences [7]. When integrating ubiquitous data into everyday life, spatial immersion issues like depth perception, data localization, and object relations become relevant. Works concerned with XR nowadays can be roughly classified as mobile (or tablets) handhelds [15], and head-mounted displays (HMD) systems [24] according to the computing paradigms. The Hololens device consists of a depth sensing camera that roughly calculates the distance of each pixel in view and pieces together a mesh or spatial map of the environment [18]. Google Tango [35] and Intel Realsense [29] offer similar technologies. The software development kit (SDK) [2] provides programmers with more freedom and flexibility to design excellent immersive applications with their own inspiration, such as ARToolkit [25, 49], Vuforia [13], and ARCore for Android [3]. A-Frame [1] enables the public to create immersive scenes in the browser integrating by WebVR [50] content within HTML.

XR-based data presentations have been applied to many fields. The CityViewAR [20] provides information about destroyed buildings and historical sites that are destroyed by the earthquakes. Focus and context information can also be separated by well-designed AR techniques [26]. Then, XR in the interpretation of terrain relief [8] shows great usability, which functions as a motivational tool for 3D data presentations. Applications in the lately emerging field of Augmented Reality Art show the canonical potential of XR as a new artistic intermediary [48].

We find few recent related works focused on DR-based applications, especially for data presentations. For example, Kawai et al. [28] find 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 new method [42] of blending and replacing textures is 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 use the blended results into 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. [21] 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. [23] design 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. [32] 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.

2.2 Interactive Immersive Building Tools

There are usually more technical challenges in immersive authoring tools compared with the pure desktop PC environment due to two gaps [10]. The first one is the steep learning curve of programming on the embedded immersive devices such as HMDs. The second one is the tedious offline workflow where users are required to debug and program frequently between immersive devices and desktop PCs [10].

Many tools have been proposed that allow interactively building and exploring data in an immersive environment. For example, MARVisT [10] allows users without background expertise to bind data on physical-world objects to realize expressive AR glyph-based visualizations. DXR [44] further provides a GUI for easy and quick edits and previews of data presentations immersed in the virtual world. IATK [12] allows for easy assembly of data presentations through a grammar of graphics that a user can allocate in a GUI, in addition to a dedicated API. PapARVis [11] is capable of designing an environment that can debug both static and virtual content simultaneously. Automated Window/Icon/Menu/Pointing Device User Interface (WIMP-UI) [52] generation has been thought about a promising technology for over two decades. iVisDesigner [14] achieves high level of interaction by means of conceptual modularity, covering a vast information presentation design space. A mixed-initiative system Voyager [53] that supports faceted browsing of recommended charts chosen according to statistical and perceptual measures.

2.3 Relationship with The Most Related Work

Some library tools were designed to help users better explore books, including Hieraxes [43] and Bohemian Bookshelf [47]. Hieraxes integrates 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 help 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 [16] is 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.

We note that there are several related XR-based data presentation tools [10, 11]. We summarize and discuss the differences between *DRCmpVis* and the most related ones as shown in Table 1 according to the data scale, tasks (augmented information, searching, re-grouping, re-ranking), visual presentations (glyph, small multiples, fish eye highlight), workflow (personal, single, or collaborative).

First, one of the differences between our work and the existing XR-based data presentation tools like MARVisT [10] are the data scale and the tasks, we focus on numerous objects, especially for the case that the number is tens, hundreds, or even thousands. Actually, *DRCmpVis* can handle more than 1,000 physical objects or even much more like books in a library/bookstore due to the efficient client-server design and the high rates of image segmentation and recognition of the backend on the server, whereas most of the XR-based related tools just focus on physical objects with the number smaller than 30 [10], e.g., PapARVis [11] (≤ 5), Situated Analytics [15] (≤ 5), MarVisT [10], etc. The large data scale of this paper poses a new challenge in image segmentation, object labeling, text information recognition and the XR-based data presentation.

Second, we mainly focus on the DR environment while most of the existing related tools focus on AR or even more close to VR [6, 12, 31, 41, 44]. DR can link the data computation in virtual space with the interaction in physical space and provide information re-organization to get a comprehensive and better target comparison.

	Data Scale (Physical Target Number)	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	virtual	■	■				■	■		Sin(PV)
AVT	virtual	■	■							Collab
VRIA	virtual	■	■				■	■	■	Collab
VR Visc	virtual	■	■				■	■		Sin(PV)
VR Collab Vis	virtual	■	■							Collab
IATK	virtual	■	■				■	■	■	Collab
SA Vis	<5		■					■		Sin(PV)
PapARVis	<5		■					■		Sin(PV)
MarVisT	<30		■				■			Sin(PV)
Our Work	40~1000+		■	■	■	■		■	■	Sin(PV)

Table 1: Comparison to the most related recent work about data presentation tools towards VR, AR, or DR. DXR [44], Augmented Virtual Teleportation (AVT) [41], Situated Analytics (SA Vis) [15], Data Visceralization (VR Visc) [30], Shared Surfaces and Spaces (VR Collab Vis) [31], IATK [12], VRIA [6], PapARVis [11], MARVisT [10]. The workflow can be categorized into PV (single user in a personal data presentations), single user (Sin) or collaborative users (Collab).

Third, we focus on filtering, re-grouping, and re-ranking according to the extra attributes of numerous physical objects, instead of augmenting the existing static presentations like in PapARVis [11]. The personal tasks are different from the most related work due to the larger data scale of *DRCmpVis*.

3 Design Rationale

We illustrate the design goal, design considerations and design details of *DRCmpVis* in this section. Before the descriptions of design goals, we need to answer a question: *why do we need DR in daily lives to reorganize the additional information before decision making?* We take the library/bookstore case of this paper as an example. One scheme to show additional information about physical objects is to query them directly from the database of the 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 is inconsistent with the information in the 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 world space and the virtual data space while significantly reducing visual clutter*. Objects are data in the physical space of the DR environment. The physical objects (targets) can be replaced by their virtual avatars, 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 on the objects. Furthermore, it saves visualization space and helps to reduce visual clutter and operational ambiguity caused by showing the physical objects and their avatars simultaneously. Besides, DR is also capable of building an information bridge between the changing physical space and the virtual space seamlessly. Regarding AR or SA (AR is one of its four primary elements as mentioned above), however, a new visualization space should be brought to the information presentation [17], then the contextual information can be provided around the physical targets.

3.1 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.

3.2 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 page limit.

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-rank 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 number 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, 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 the name of a book in a book shelve, 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 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.

3.3 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.

4 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, 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 identical calibrated coordinated system. We have developed and encapsulated the related functions into the APIs of the framework.
- **Challenge III: integrating comparative visualizations into 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, 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**.
- **Challenge V: enhancing the lighting environment in the reality world**. In the 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 synthesizing the captured images to restore the reflect regions.

For detailed information about the implementation, please refer to the Appendix file.

4.1 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 [4] or ARCore [3] 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 physical space. The distance can be measured by the camera with LiDAR scanner [4].

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

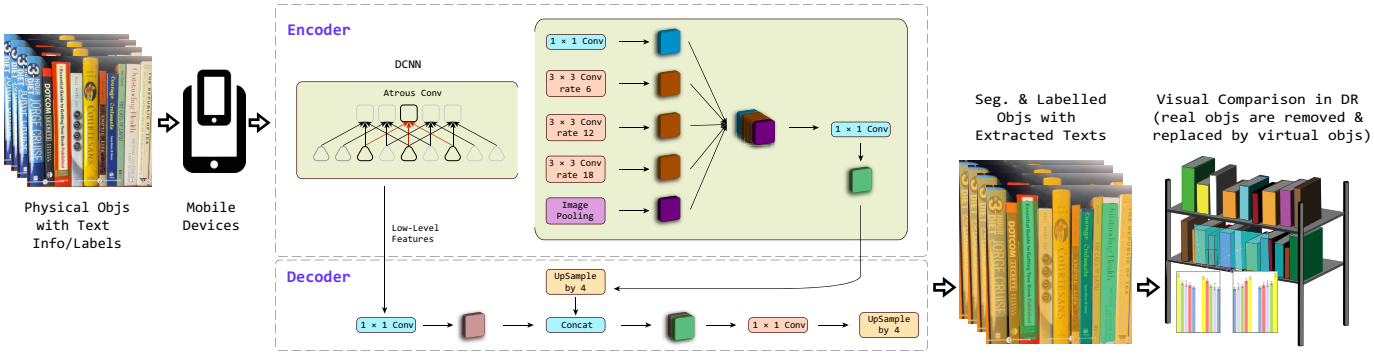


Fig. 2: The workflow of the proposed *DRCmpVis*. We illustrate it using one of the application cases, the library/bookstore case. Regarding the deep neural network used in image segmentations and text recognitions, the encoder module encodes multi-scale contextual information by applying atrous convolution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

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. Actually, we initially use 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 & 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 [34] 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 [9] is one of its key modules. Therefore, we take DeepLab as an example to illustrate how PaddleSeg is integrated into *DRCmpVis*, as shown in Figure 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 panoramic image we captured or the real-time video we recorded is input into the first network (the top left of Figure 2), while the labeled samples are input into the second network (the top right of Figure 2). Regarding the text extraction, we use the traditional CNN-based optical character recognition approach, following a language adaptive design [33], to recognize a large amount of the 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 the 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 the 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.

4.2 Database Construction of Augmented Information

We create a large database on the server for two application scenarios that require real-time information feedback [22, 37]. The database contains additional information on different attributes of the objects. In order to make the data updated periodically and improve the scalability of the framework, we design a data synchronizer with a pattern matching algorithm and regular expression matching algorithm, which can be used to download the open data automatically and fetch the data attributes to update them in the database.

(1) Global book database. More than two million books are 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 is downloaded from the open data website "Amazon product data" [22, 36, 37], containing product reviews and metadata from Amazon, including 142.8 million reviews for their products and 22.5 million reviews for books. It includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs). The Amazon database was last updated in 2018.

(2) Coffee database. The coffee database is 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, process introduction.

4.3 Integrating Visual Comparison Components into DR/MR Context

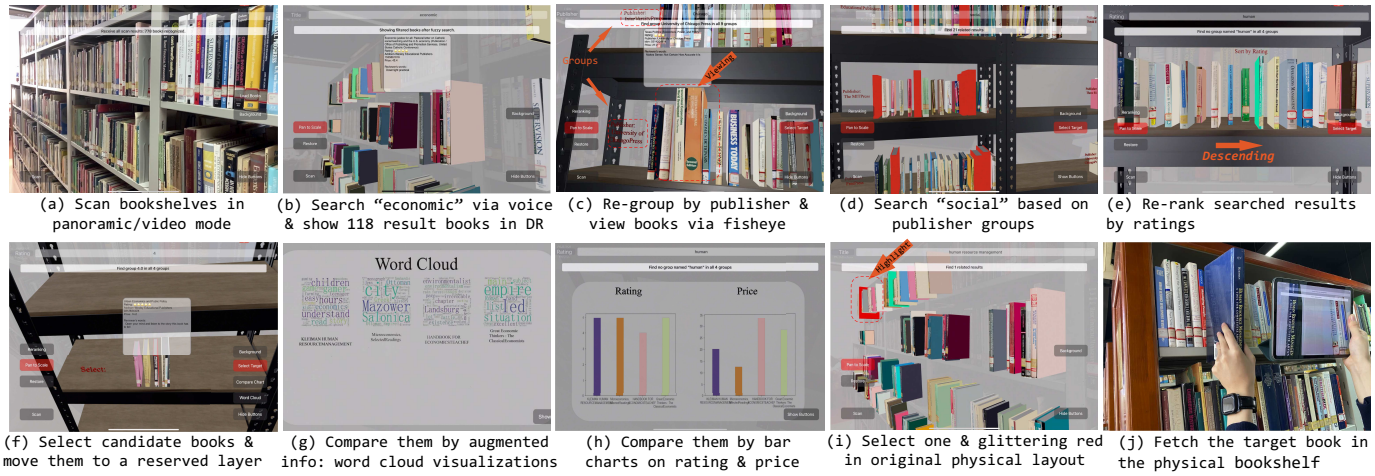
Regarding the visual comparisons of the additional attributes (augmented information), the related data is sent to the server and the client receives the processed data from the server. We design several visual comparison components like *bar chart*, *line chart*, *word cloud*, *ingredient glyph*, etc., 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 *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.

5 Example Scenarios

To illustrate how *DRCmpVis* facilitates visual comparisons for physical objects with text labels in DR environments and demonstrate the robustness of the proposed framework.

5.1 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



(a) Scan bookshelves in panoramic/video mode
(b) Search "economic" via voice & show 118 result books in DR
(c) Re-group by publisher & view books via fisheye
(d) Search "social" based on publisher groups
(e) Re-rank searched results by ratings
(f) Select candidate books & move them to a reserved layer
(g) Compare them by augmented info: word cloud visualizations
(h) Compare them by bar charts on rating & price
(i) Select one & glittering red in original physical layout
(j) Fetch the target book in the physical bookshelf

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 presses are placed on different layers. The user browses those books with a fisheye effect. (d) Further search with a 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 word cloud of abstracts, introduction 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.

area in a library/bookstore, facing several bookshelves with around a thousand books, as shown in Figure 3 (a).

(1) **Fuzzy filtering**: she scans the bookshelves by the panoramic camera of her tablet with *DRCmpVis* installed. There are 778 books that 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 Figure 3 (b). She finds that only one book nearby is from "University of Chicago Press", then she wants to find more books on "economic" and published by "University of Chicago Press".

(2) **Re-grouping**: she re-groups those 118 books by publisher and searches by saying "Chicago" or input 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 (Figure 3 (c)). Books from other presses 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 fuzzy keyword "social" , there are 21 books highlighted in red (Figure 3 (d)). She uses fisheye to view each book's details including titles or authors similar to Figure 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 either 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 by descending order (Figure 3 (e)). Then she selects four books that seem suitable, those selected books are moved to a reserved layer of the virtual bookshelf which are designed to place the candidate books (Figure 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" and "empire" (Figure 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 kinds of diagrams in small multiples**: so she compares both the ratings and the prices of these books via bar charts (Figure 3 (h)). She finds that the "story" (the first) rated as high as the "empire" one (the fourth), but is a little cheaper than the "empire" one. So she chooses the "story" one and restores those books to their original layout.

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

5.2 Cafe Scenario

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

A new coffee shop opens on Zelda's campus. She doesn't 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 by *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 achieved by the image segmentation, image labeling, and text extraction using neural networks DeepLab [9] and PaddleSeg [34]. 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 firstly voice inputs "Latte" and finds that it's highlighted in the menu (Figure 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 surely bitter, as no sugar is added to it (Figure 4 (c)).

Zelda then re-groups those coffees according to sugar (Figure 4 (d-e)). She browses and selects several drinks with high ratings in the "medium sweet" and "sweet" group, as shown in Figure 4 (f). Then she compares those drinks' ingredients in small multiples, and finds that Cappuccino has a balance among sugar, milk, and caffeine, which

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 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.

may suit her taste, as shown in Figure 4 (g). However, her fitness coach's advice crosses her mind that she needs to limit her calorie intake to 1300 calories every day, whereas the coffee summary shows that Cappuccino has 140 calories per cup. So she re-ranks all the drinks by calorie content. This time, coffees are sorted from left to right by calorie, as shown in Figure 4 (h). She begins browsing on the right side, where coffees with relatively low calories are located. She finds several coffees that she hasn't drunk. To have a quick grasp of them, she views their word cloud (Figure 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 (Figure 4 (j)). Finally, she chooses Blonde Roast and enjoys its "soft and mellow flavor" described in the summary. In addition, *DRCmpVis* can also handle larger menus like a big poster hanging on the wall outside the coffee shop, as shown in Figure 5.

6 Evaluation: User Study and Framework Performance

In the evaluation, we aimed to assess *DRCmpVis* regarding the following aspects: (a) whether visual searching/filtering of *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 have conducted four measures, including subjective measures and objective measures:

- **User Study:** a 5-point Likert scale was utilized to gauge and assess the comprehensive functionality of *DRCmpVis*.
- **NASA-TLX:** 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:** regarding 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 modules of *DRCmpVis*, including the modules of scanning, image segmentation & labelling, overall processing, etc.

6.1 Study Design

User Study 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.

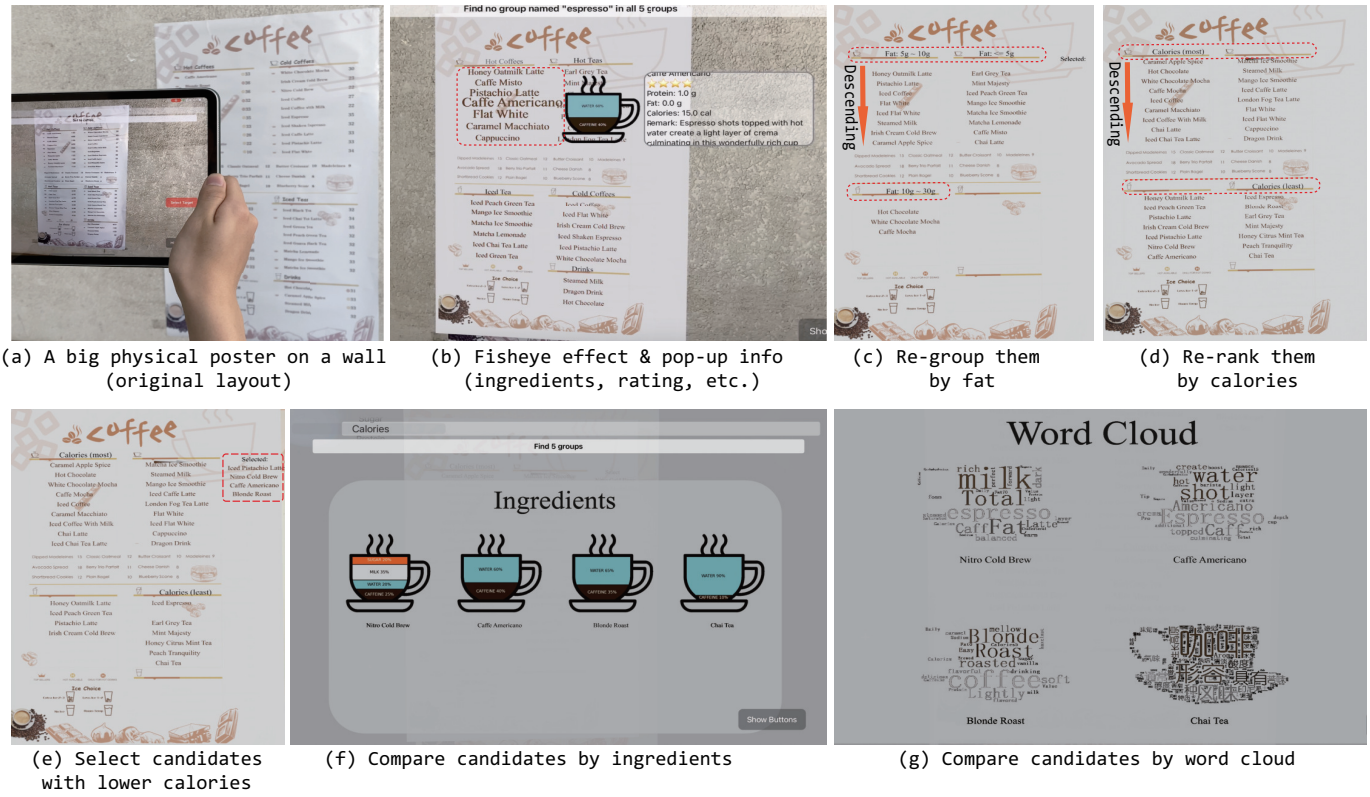
Procedures. *T1* was performed in a library, while *T2* was performed in a coffee shop. Before starting the tasks, participants were required to fill in the pre-study questionnaires.

We discovered that the majority of participants were not familiar with XR technology, 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 them had trouble finding books in libraries where books are sorted by traditional index numbers.

Regarding the coffee scenario, most of them also felt confused when choosing different coffees. It is 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, proves to be arduous for them. 98.7% of participants do not agree that coffee shop staff will provide a retrieval system for you to use, while 86.7% of participants indicate that coffee shop staff will not provide specific ingredient information for comparison.

In the pre-study survey before the questionnaire step, we have one question to survey how many users like simple visualizations 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 just 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.

Then, the investigators introduced the capability and usage of *DRCmpVis*. In *T1* and *T2*, the investigators showed a simple example to the users first and then released the specific task. After all the tasks were finished, the participants were asked to complete the post-study questionnaires. All the participants got gifts of equal value regardless of their performance.



(a) A big physical poster on a wall (original layout)
(b) Fisheye effect & pop-up info (ingredients, rating, etc.)
(c) Re-group them by fat
(d) Re-rank them by calories
(e) Select candidates with lower calories
(f) Compare candidates by ingredients
(g) Compare candidates by word cloud

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.

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

6.2 User Study Tasks

We use *Ti* to name the task that happens in the *i*-th scenario. The first task *T1* is about the library case, while the second task *T2* is about the cafe menu case.

T1 is divided into three subtasks. In *T1*, participants are required to locate four different books. In *T1-1*, participants 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 to 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 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 the proposed tool, *DRCmpVis*, to find the four books from the aforementioned tasks. The timing results are recorded in all of the tasks in *T1*. After completing *T1-3*, the participants are suggested to use re-grouping of *DRCmpVis* to find other books with the 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 are desired to read.

T2 requires participants to search for different types of lattes from the physical coffee menu. *T2* has only two tasks, because coffee shops do not provide users with a coffee retrieval system unless the manager or the waiters. In *T2-1*, participants search for lattes from the physical menu, while in *T2-2*, they can use *DRCmpVis* to highlight all the candidates that satisfied the task requirements. The timing results of *T2-1* and *T2-2* are also recorded in each task. After these two timing experiments, participants are required 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 Figure 4. After that, participants could also check the menu and select other coffees that they are unfamiliar with. They could compare them in the MR context using ingredient glyphs and word cloud, as shown in Figure 5 (f).

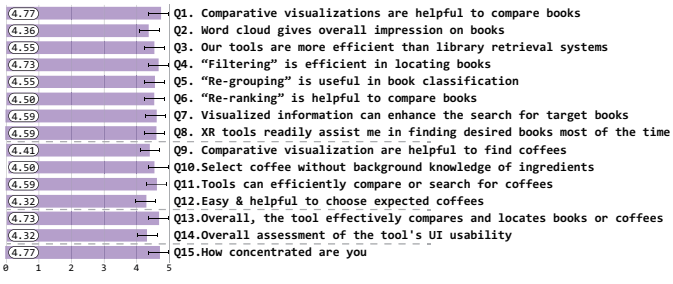


Fig. 6: Post-study result: most of participants react positively to *DRCmpVis*.

6.3 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 Figure 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 with the *DRCmpVis* (Q13: $\mu = 4.73$, 95% $CI = [4.53, 4.92]$ G1). In particular, UI design (Q14: $\mu = 4.32$, 95% $CI = [4.00, 4.64]$ G4), besides, the participants also appreciated the voice input, fisheye effect, and result highlighting. They said these designs make the interactions smooth and intuitive. From the questionnaire results, we can find that they can 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 (Q3 ($\mu = 4.55$, 95% $CI = [4.32, 4.77]$ G1) and Q1 ($\mu = 4.77$, 95% $CI = [4.54, 5.01]$ G4)) and the cafe scenario (Q12 ($\mu = 4.31$, 95% $CI = [4.03, 4.60]$ G4)), the participants gave high praise, because they thought *DRCmpVis* is intuitive to use in 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 aid participants in developing a comprehensive understanding of physical objects. The participants also noted that the comparative ingredient glyphs significantly contribute to forming comprehensive impressions of the distinctions among various types of coffees and books.

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) of books in libraries.

Compared with blindly finding, the time cost is reduced from an average of 4.56 minutes to 0.65 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". In the cafe scenario, the time cost of finding eight lattes from the menu is reduced from 0.45 minutes to 0.13 minutes with the help of *DRCmpVis*.

In response to selecting coffee without background knowledge of ingredients (Q10: $\mu = 4.50$, 95% $CI = [4.20, 4.80]$ G1) aiming to efficiently compare or search for different coffees (Q11: $\mu = 4.59$, 95% $CI = [4.36, 4.81]$ G1), most participants found the subsequent visual comparisons helpful for them as they didn't 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 Q15 ($\mu = 4.77$, 95% $CI = [4.58, 4.96]$), almost all participants felt concentrated when carrying on the tasks. They all believed that the tasks were quite smooth and interesting.

6.4 NASA-TLX Measures

We further evaluate the proposed *DRCmpVis* by comparing it with two traditional methods as control groups based on NASA-TLX measurements, i.e., target blinding finding without any tools (Blinding finding), 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 are randomly from ten different majors of the university.

A repeated measures 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 Figure 7. It is worth mentioning that the mental demand of blind finding is significantly lower than the other two methods, because the blind finding is the simplest approach which just has the smoothest learning curve. It requires some learning to master DB-based searching tool and *DRCmpVis*. The physical demand in *DRCmpVis* is significantly lower than in

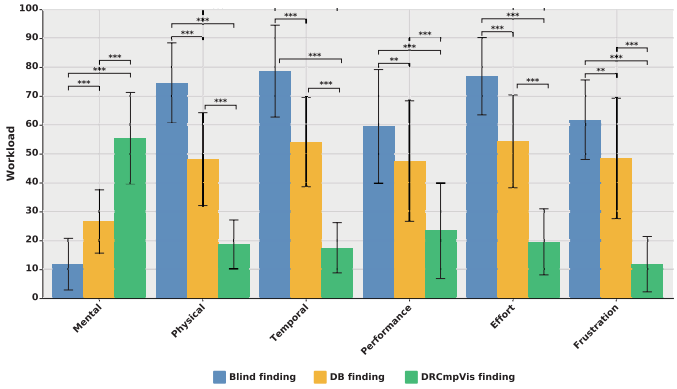


Fig. 7: The scores of NASA-TLX evaluation for two control groups of methods and the proposed *DRCmpVis*. Error bars indicate standard errors. Statistical significant differences are denoted by ** (p <0.01) , *** (p <0.001).

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 Table 2 and Table 3. 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", $p = 2.7E-17$).

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 is 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. Besides, some participants suggested that it would be better if we add visual cues about the physical directions of the target book when they search 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*, is convenient, efficient, and relatively easy to learn, requiring less effort to locate target objects.

6.5 Quantitative Evaluation

For the sake of achieving a dependable and consistent server service, we choose 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 is 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 & labeling, which is provided by a trained CNN platform named PaddleSeg [34]. In our experiments, both the PaddleSeg and the database with augmented information are built on the cloud server. We can find that the average image segmentation & labeling rates of *DRCmpVis* for all the example scenarios are larger than 95.0% (the additional example scenarios are moved to the Appendix due to page limit of the paper). The quantitative evaluation results are shown in Table 2.

Scenario	Scanning Time	Segmentation Time	Processing Time	Seg. & Labelling Rate
Library Scenario	0.321	0.631	0.952	95.13%
Cafe Scenario	0.454	3.891	4.345	100.00%

Table 2: Performance and accuracy evaluation of *DRCmpVis* (seconds). The "Scanning Time" is the average time to scan a panoramic photo 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 each back-end server request, while the "Seg. & Labeling Rate" is the average accuracy. We obtain the average results based on 16 tests.

Methods	Book Searching Time	Latte Searching Time	Augmented Info	Target Comparison
Blind finding (without any tools)	4.56	0.45	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*	0.65	0.13	Color highlight Fisheye highlight Pop-up glyphs	Re-grouping Re-ranking Visual comparison

Table 3: Task-driven quantitative evaluation results (in minutes). We compare the proposed *DRCmpVis* with the traditional two methods. The results shows 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 taken by all participants to search for the keyword "Latte". Note: most coffee shops do not provide a retrieval system for users, thus they are marked as Not Available (NA).

Furthermore, we have also conducted some quantitative and qualitative tests for the two tasks without/with *DRCmpVis*. There are 22 participants involved in the tasks. The task is to ask the participants to find all types of "Latte" from the physical coffee menu, as shown in Figure 4 (a). The results searched by *DRCmpVis* are shown in Figure 4 (b). There are eight different lattes in total. The test results are shown in Table 3. We can find the task *T1-3* finished by *DRCmpVis* takes about 0.65 minutes on average to find the target book from 1238 books, which is only 14.3% of the search time used in the blind finding method (without any tools). Similarly, the task *T2-2* finished by *DRCmpVis* takes about 0.13 minutes on an average to find all lattes, which is 28.7% of the searching time in the coffee shop without *DRCmpVis*. By comparing the proportion, it can be seen that the more target objects searched, the greater the advantage of *DRCmpVis*.

We also summarize some feature comparisons in Table 3. 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.

7 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. Besides, we find *DRCmpVis* can be easily extended to the 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. 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 they are in some crowded setting. In some cases, libraries will be influenced by the crowded environment, but in other cases, such as the cafe menu case, are irrelevant.

However, *DRCmpVis* is not feasible for the libraries or bookstores when the book information is unavailable to fetch, or it is hard to download or crawl from Internet, e.g., the ancient book libraries, etc., because the framework will query additional augmented information from the constructed database according to the information scanned from the physical objects.

Scalability issue on image segmentation and image labeling. It is worth noting that the image segmentation components of *DRCmpVis* are *scalable* and not limited by the object number, because the CNN and the OCR algorithm are run on the server which can even handle thousands of books in the library scenario in our experiments. More importantly, unlike the mobile device, the computation resources of the server are *scalable* enough and could be easily upgraded. As a result, whereas *DRCmpVis* recognizes almost all 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.

Why do we mainly use DR/MR instead of VR, or why not use a fully database-based VR as an alternative design? First, information should be updated periodically between the virtual world and the physical world. The object data in the virtual space should be consistent with that in the physical world in *DRCmpVis*. Because in the library/bookstore case, the book positions would be often changed due to the previews by buyers/borrowers, the books are also often inserted into the wrong positions or even wrong bookshelves by buyers/borrowers. In the cafe scenario, the menus are also often moved in a coffee shop (as shown in the supplemental video). All such scenarios need to involve the real-time physical world information into *DRCmpVis*, which makes *DRCmpVis* should include DR/MR instead of VR. In *DRCmpVis*, actually, the image recognition module on the front-end mobile device will initially and periodically detect whether the object information in the physical world is changed. If yes, all the changed positions of the objects will be updated by the deep network deployed on the cloud server.

Second, users often need to go back to the reality to "highlight" the targets after the searching or the comparing steps to help users find them. For example, the target books/the target coffee items will be highlighted in the real background after users' searching or visual

comparisons (as shown in the supplemental video). The in-context highlighting in reality requires DR/MR instead of VR.

Third, it is impossible or time-consuming for bookstore salesmen/librarians to update the database of the books' new positions immediately, if we choose VR instead of DR/MR.

The limitation of text recognition. *DRCmpVis* recognizes objects by images taken from mobile devices. Ideally, the user only needs to take one panoramic picture that contains all the objects. However, objects' details may not be recognized if they are too small in the picture, that is the user is standing too far away from the numerous objects. For example, in the library/bookstore scenario, instead of scanning all layers of the bookshelves, the user may walk closer to the bookshelves and scan one layer at one time by panoramic stitching due to lack of light or limited imaging quality.

The image segmentation & labeling service needs to request once due to an image recognition module on the client app of *DRCmpVis*, when the positions of the objects are not changed. Because the coffee menu in a cafe is often unchanged. Actually, we 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. In our strategy, the latest captured panoramic image will be saved to the buffer of the client app. The image recognition module will verify whether the newly captured panoramic image is saved in the buffer. If yes, the segmentation & labeling records in the buffer can be reused without requesting the server twice. This strategy is *useful and efficient* in almost all the usage scenarios due to the quick response by the client app. 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 is 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 6. The hardware configuration can be improved for more expensive service packages. Thus maybe the performance especially for the segmentation & labeling could be further improved. We plan to make *DRCmpVis* to be applied in more general usage scenarios in our daily lives. In the future, we plan to extend the usage scenario of *DRCmpVis* to others like choosing cups, fruits or flowers, and other more general scenarios in our daily life. Because objects with text on them or in different colors and shapes can be well recognized by trained neural networks. However, objects with different irregular 3D shapes and without textual information on them are difficult to be recognized by current algorithms including the-state-of-the-art neural networks.

8 Conclusion

In this paper, we propose a novel DR/MR-based application framework named *DRCmpVis*, which is designed to build visual comparisons towards multiple physical objects with text labels or text information. The efficient data computation in virtual space is linked with the in-context interaction in physical space in the framework. The framework can provide multidimensional comparisons for candidate objects, exploiting all their three types of attributes, i.e., nominal, ordinal, or quantitative attributes. Users first take panoramic photos from the real world by the cameras of mobile devices. They can input a fuzzy searching keyword in objects' nominal attributes by voice or text (according to users' environment) to narrow down the number of candidate targets. The search results will be highlighted by color and deformation in the DR environment to indicate their positions in the reality; Furthermore, users possess the capability to regroup or re-rank candidates based on their multifaceted attributes. Additional comparative augmented information of the objects can be integrated in an identical MR context.

References

[1] AFrame. aframe. https://aframe.io/.

[2] D. Amin and S. Govilkar. Comparative study of augmented reality sdks. *International Journal on Computational Science & Applications*, 5(1):11–26, 2015.

[3] ArCore. Arcore. https://developers.google.cn/ar/.

[4] ARKit. Arkit. https://developer.apple.com/cn/augmented-reality/arkit/.

[5] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, 1967.

[6] P. W. S. Butcher, N. W. John, and P. D. Ritsos. Vria: A web-based framework for creating immersive analytics experiences. *IEEE Transactions on Visualization and Computer Graphics*, 27(7):3213–3225, 2021.

[7] S. Butscher, S. Hubenschmid, J. Müller, J. Fuchs, and H. Reiterer. Clusters, trends, and outliers. In *CHI Conference on Human Factors in Computing Systems*. ACM, apr 2018.

[8] C. C. Carrera, J. L. S. Perez, and J. de la Torre Cantero. Teaching with ar as a tool for relief visualization: usability and motivation study. *International Research in Geographical and Environmental Education*, 27(1):69–84, 2018.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.

[10] Z. Chen, Y. Su, Y. Wang, Q. Wang, H. Qu, and Y. Wu. Marvist: Authoring glyph-based visualization in mobile augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 26(8):2645–2658, 2020.

[11] Z. Chen, W. Tong, Q. Wang, B. Bach, and H. Qu. Augmenting static visualizations with paparvis designer. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2020.

[12] M. Cordeil, A. Cunningham, B. Bach, C. Hurter, B. H. Thomas, K. Marriott, and T. Dwyer. IATK: An immersive analytics toolkit. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 200–209. IEEE, 2019.

[13] D. Cushnan and H. E. Habbak. *Developing ar games for ios and android*. Packt Publishing Ltd, 2013.

[14] Donghao Ren, Tobias Höllerer, and Xiaoru Yuan. ivisdesigner: Expressive interactive design of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2092–2101, 2014.

[15] N. ElSayed, B. Thomas, K. Marriott, J. Piantadosi, and R. Smith. *Situated analytics*. 2015.

[16] 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 *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pp. 203–208. Merida, Mexico, may 2016.

[17] N. A. M. ElSayed, B. H. Thomas, R. T. Smith, and K. Marriott. Using augmented reality to support situated analytics. In *IEEE Virtual Reality*. Arles, France, Mar. 2015.

[18] G. Evans, J. Miller, M. I. Pena, A. MacAllister, and E. Winer. Evaluating the microsoft hololens through an augmented reality assembly application. In *Degraded Environments: Sensing, Processing, and Display 2017*, vol. 10197, p. 101970V. International Society for Optics and Photonics, 2017.

[19] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[20] Gun Lee, Andreas Dünser, Seungwon Kim, and Mark Billinghurst. Cityviewar: A mobile outdoor ar application for city visualization. In *2012 IEEE International Symposium on Mixed and Augmented Reality - Arts, Media, and Humanities (ISMAR-AMH)*, pp. 57–64, 2012.

[21] S. Hashiguchi, S. Mori, M. Tanaka, F. Shibata, and A. Kimura. Perceived weight of a rod under augmented and diminished reality visual effects. In *The 24th ACM Symposium on Virtual Reality Software and Technology*, VRST '18, pp. 1–6. ACM, New York, NY, USA, 2018.

[22] R. He and J. McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *International Conference on World Wide Web*, pp. 507–517, 02 2016.

[23] J. Herling and W. Broll. PixMix: A real-time approach to high-quality diminished reality. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 141–150. IEEE, Nov 2012.

[24] D. Herr, J. Reinhardt, R. Krüger, G. Reina, and T. Ertl. Immersive visual analytics for modular factory layout planning. In *Proc. IEEE VIS Workshop Immersive Analytics*, 2017.

[25] K. Hirokazu. Artoolkit: Library for vision-based augmented reality. *Technical Report of Ieice Prmu*, 101:79–86, 2002.

[26] D. Kalkofen, E. Mendez, and D. Schmalstieg. Interactive focus and context visualization for augmented reality. In *IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 191–201, 2007. doi: 10.1109/ISMAR.2007.4538846

[27] D. Kalkofen, C. Sandor, S. White, and D. Schmalstieg. Visualization

techniques for augmented reality. In *Handbook of augmented reality*, pp. 65–98. Springer, 2011.

[28] N. Kawai, T. Sato, and N. Yokoya. Diminished reality based on image inpainting considering background geometry. *IEEE Transactions on Visualization and Computer Graphics*, 22(3):1236–1247, March 2016.

[29] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik. Intel realsense stereoscopic depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–10, 2017.

[30] B. Lee, D. Brown, B. Lee, C. Hurter, S. Drucker, and T. Dwyer. Data visceralization: Enabling deeper understanding of data using virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1095–1105, feb 2021.

[31] 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 Transactions on Visualization and Computer Graphics*, 27(2):1171–1181, feb 2021.

[32] 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 *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 11–19. IEEE, Oct 2013.

[33] O. Y. Ling, L. B. Theng, A. Chai, and C. McCarthy. A model for automatic recognition of vertical texts in natural scene images. In *2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, pp. 170–175, 2018.

[34] Y. Liu, L. Chu, G. Chen, Z. Wu, Z. Chen, B. Lai, and Y. Hao. Paddleseg, end-to-end image segmentation kit based on paddlepaddle. `https://github.com/PaddlePaddle/PaddleSeg`, 2019.

[35] Marder-Eppstein and Eitan. Project tango. In *ACM SIGGRAPH 2016 Real-Time Live!*, pp. 25–25. 2016.

[36] J. McAuley. Amazon product data. https://jmcauley.ucsd.edu/data/amazon/, 2018.

[37] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel. Image-based recommendations on styles and substitutes. *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52, 2015.

[38] P. Milgram and F. Kishino. A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems*, vol. E77-D(12):1321–1329, 12 1994.

[39] S. Mori, S. Ikeda, and H. Saito. A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. *IPSJ Transactions on Computer Vision and Applications*, 9(17):1–14, 2017.

[40] G. Queguiner, M. Fradet, and M. Rouhani. Towards mobile diminished reality. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct*, pp. 226–231. IEEE, Munich, Germany, 2018.

[41] T. Rhee, S. Thompson, D. Medeiros, R. dos Anjos, and A. Chalmers. Augmented virtual teleportation for high-fidelity telecollaboration. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1923–1933, 2020.

[42] S. H. Said, M. Tamaazousti, and A. Bartoli. Image-based models for specularity propagation in diminished reality. *IEEE Transactions on Visualization and Computer Graphics*, 24(7):2140–2152, July 2018.

[43] B. Shneiderman, D. Feldman, A. Rose, and X. F. Grau. Visualizing digital library search results with categorical and hierarchical axes. In *Proceedings of the fifth ACM conference on Digital libraries*, pp. 57–66.

[44] R. Sicat, J. Li, J. Choi, M. Cordeil, W.-K. Jeong, B. Bach, and H. Pfister. Dxr: A toolkit for building immersive data visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):715–725, 2019.

[45] C. Stolte, D. Tang, and P. Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52—65, jan 2002.

[46] M. Takemura and Y. Ohta. Diminishing head-mounted display for shared mixed reality. In *International Symposium on Mixed and Augmented Reality*, pp. 1–8. IEEE, Darmstadt, Germany, 2003.

[47] A. Thudt, U. Hinrichs, and S. Carpendale. The bohemian bookshelf: supporting serendipitous book discoveries through information visualization. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1461–1470, 2012.

[48] Vladimir Geroimenko. Augmented reality technology and art: The analysis and visualization of evolving conceptual models. In *2012 16th International Conference on Information Visualisation*, pp. 445–453, 2012.

[49] D. Wagner and D. Schmalstieg. Artoolkit on the pocketpc platform. In *2003 IEEE International Augmented Reality Toolkit Workshop*, pp. 14–15, 2003.

[50] WebVR. Webvr. `https://webvr.info/`.

[51] W. Willett, Y. Jansen, and P. Dragicevic. Embedded data representations. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):461–470, jan 2017.

[52] WIMP. Wimp. `https://www.interaction-design.org/literature/book/the-glossary-of-human-computer-interaction/wimp`.

[53] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):649–658, 2016.

[54] 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 Transactions on Visualization and Computer Graphics*, 27(9):3808–3825, 2021.

Dear Editors/Reviewers,

We wish to retain reviewer continuity and request a double-blind review as in the initial submission (VIS'23 ID: 1263) of this work.

We sincerely appreciate the valuable feedback received in the previous round of review. They think the paper is interesting and is potential make good technical contribution. For example:
---------- (Reviews of VIS'23) ----------
Overall, the reviewers expressed positivity regarding the technical contributions of this paper, acknowledging the effort and technical pipeline involved in building the system and its applications:

The basic concept of the framework is interesting, and the scenarios illustrate a number of seemingly well-thought-out techniques. (R4, R3)

The presented system shows potential for future data visualizations. (R2, R3)

The presented system demonstrates a level of technical credibility, and the reviewers applaud the authors for their effort in building the system. (R1, R2, R3, R4).

The video is well-made and shows an interesting combination of techniques. (R2, R4)

I believe this paper has some nice technical ingredients that can lead to a really nice paper in the near future once the issues are addressed and new insights/designs/techniques are presented! (R1)

---------- (Reviews of VIS'23) ----------

We also appreciate the reviewers recommend us resubmit the revised paper to TVCG:
"*Considering the feedback from the reviewers, it is recommended that the authors conduct completely new studies, undertake significant revisions, and resubmit the paper to the IEEE Transactions on Visualization and Computer Graphics (TVCG).*" (Summary Review)

All the reviewers' comments are helpful for us to revise the paper, aiming to help us improve the paper substantially. We have conducted a significant major revisions in the whole three months after we received the reviews. **The biggest concern raised by the reviewers are the evaluation**. In summary, the evaluation part is revised based on the following points:

1) **[New] A completely new task-driven user study**: a formal task-driven evaluation, including the evaluation about the designed visualizations;
2) **[New] Get statistical significant differences by comparing with two control group methods**: achieve the significance of difference ($p < 0.01$) of the proposed tool compared with two control groups;
3) **[New] Repeated-measures ANOVA on the NASA TLX questionnaire**: demonstrate whether the tool significantly better differences on the visualization tasks in XR, compared with two existing methods;

4) **[New] Quantitative evaluation:** include a performance evaluation on scanning time, segmentation time, processing time, segmentation and labelling rate, target finding time (compared with two control group methods), etc.

Besides, we have taken all the reviewers' concerns very seriously and have conducted a comprehensive revision of our paper point by point. In addition to the authors ourselves (one is from Stony Brook University), we also asked some native speakers to proofread the paper carefully. We have included the final draft and a document highlighting the differences between our initial submission and this final version. Thank you for your consideration.

Best Regards,

Authors
Sept. 2023

Category I - Technical Details and Performance

- Comment #01: The implementation section offers little information about challenges and their Resolution. (R4)

Response to Comment #01: thanks for the review. We have presented more details about the implementation challenges and the corresponding solutions in Section 4 (Section Implementation), the details about the implementation are moved to the **Appendix** file:

(1) Building the application framework. Image segmentation, image labelling, OCR-based text extraction, image recognition are the significant modules of the framework. We have integrated several latest deep neural networks into the framework. All of them are encapsulated as the APIs of the framework.

(2) Coordinate transformation between physical space and virtual space. This step is to build the virtual avatars mapped to the physical objects and then mix them seamlessly in an identical calibrated coordinated system. We have developed and encapsulated the related functions into the **APIs of the framework.**

(3) Integrating comparative visualizations into DR/MR context. We have integrated some commonly-used visualization components/techniques into the framework, e.g., bar charts, line charts, word clouds, 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, whether they are simple or advanced. All the related functions are encapsulated into the APIs of the framework.

Fig. 1 We design algorithms to help reduce the interference of the reflect light on the objects

(4) Enhancing the lighting environment in the reality world. In the practical applications, it is important to reduce the interference of reflect light (Fig. 1), illumination compensation when the lighting is weak.

The solution to the former issue is to capture multiple frames with a time interval (e.g., 0.5 seconds), when the camera is scanning, then synthesizing the captured images to restore the reflect regions. The solution to the latter one is to integrate the corresponding image processing algorithms into the system. Please in the Para. 1, Section 4 (Section Implementation) and the details can be found in the **Appendix** file.

- Comment #02: "How accurate is the scanning and detection of books? In what cases does it not work? In Table 2, how were the segmentation rates and times measured?" "It is essential to evaluate the system performance of the presented frameworks, which should include a proper assessment of factors such as scanning, detection, segmentation rates, and processing times." (R1)

Response to Comment #02: thank you, we have conducted new evaluation studies and experiments, and added the timing results and accuracy evaluation results, as shown in Fig. 2 (Table 2 in the paper) and Fig. 3 (Table 3 in the paper).

Scenario	Scanning Time	Segmentation Time	Processing Time	Seg. & Labelling Rate
Library Scenario	0.321	0.631	0.952	95.13%
Cafe Scenario	0.454	3.891	4.345	100.00%

Fig. 2 Timing results (seconds) and the accuracy performance

Methods	Book Searching Time
Blind finding (without any tools)	4.56
With a DB retrieval system (targets are available on the correct shelf)	2.53
With a DB retrieval system (targets are inserted in wrong positions)	5.34
With a DB retrieval system (available but being viewed by other borrowers)	unlimited
The proposed *DRCmpVis*	0.65

Fig. 3 Time cost comparison about the three methods (minutes)

- Comment #03: The authors don't explain the preprocessing process, the steps they take to build the application using a "framework". I would like to know if this is feasible for any library or bookstore to scan and accurately identify books. In addition, it is unclear whether the application has the ability to scan books by identifying each item individually. (R3)

Response to Comment #03: thank you for your comments. The framework supports different libraries and bookstores scenarios because the entire processing steps are independent on the scenario itself. The scanning, segmentation & labelling, text recognition are achieved by corresponding models and algorithms. We use deep learning platforms (PaddlePaddle), OCR, iOCR, and other algorithms for target object segmentation and recognition. In our experiments, actually, we have tested the proposed *DRCmpVis* in different libraries and different books with different languages (English and Chinese).

However, *DRCmpVis* is not feasible for the libraries or bookstores when the book information is unavailable to fetch, or it is hard to download or crawl from Internet, e.g., the ancient book libraries, etc., because the framework will query additional augmented information from the constructed database according to the information scanned from the physical objects. The application scope of the work is discussed in Section 7 (Section "Discussion and Future Work").

Category II - Justification for Vague Descriptions

- Comment #04: "What are the benefits of diminished versus augmented reality?" (R1)

Response to Comment #04: thanks, diminished reality offers the advantage of enhancing focus and attention by selectively removing or reducing distracting elements from real-world scenes. Multiple target objects can be flexibly grouped, sorted, searched (interacting with the real environment), and located (returning to the real environment and updating in real-time to ensure that the objects in the virtual world are as consistent as possible with the real objects) in the corresponding virtual environment. All these physical re-layout operations can be easily performed in a virtual environment with their virtual avatars, while **AR** can't provide the virtual avatars of the physical objects because the physical objects are not diminished, their augmented information are just displayed in a pop-up widget surrounding them. For more details about the justification, please check them in Para.3 in Section 3.

- Comment #05: In abstract: rephrase "stereotyped strategies" into something clearer - what do the authors mean by this? (R1)

Response to Comment #05: thanks for the suggestion. We have rephrased "stereotyped strategies" into "stereotyped presentation style". It means the books in libraries/bookstores are all grouped and sorted according to a single given rule, which restrict the usability and flexibility for some users, e.g., regrouping, re-ranking according to various keywords (attributions).

- **Comment #06:** The author initially presented only two examples, but in Section 6.4, they mentioned "three examples." (R3)

Response to Comment #06: thanks, we had three examples in the previous submission, two are in the paper, and the third one is moved to the Appendix file due to page limit. We have clarified in the paper, please check it.

In the revision, actually, we have added a new example scenario (the fourth), i.e., a restaurant scenario (an example in Shaxian County cuisine), the example is also presented in the Appendix file.

- **Comment #07:** I believe the techniques presented in this paper are really interesting and I praise the authors for the approach. However, where it not for the video, I would have missed how interesting it is. I would suggest to also look into the definitions of methods and really argument along the right lines: diminished reality, mixed reality, or situated analytics. (R4)

Response to Comment #07: thank you for your comments. We have added the definitions about DR, MR and SA into the paper, and cited the corresponding literatures:

"DR pertains to the manipulation of a perceived environment in real-time, involving actions like concealing, eliminating, or revealing objects [22]. The objects to be diminished/removed in DR should be detected and tracked while the camera is freely moving [39, 44]. Mixed reality (MR) is strictly defined by Milgram and Kishino [37]. They think MR is a superset of AR in terms of a mix of real and virtual objects within a single display. The distinctions between augmented reality (AR), augmented virtuality (AV), MR [42], and diminished reality are fuzzy. To the best of our knowledge, there are not many pieces of 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 [16], including situated information, abstract information, augmented reality interaction, and analytical interaction. 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. [16] think SA is a new area of research at the intersection of visual analytics and AR."

We also described why we used DR and MR instead of other XR technologies. Please check them in Para. 3 of Section 3:

DR is an optimal solution to keep information consistent between the physical world space and the virtual data space while significantly reducing visual clutter. Objects are data in the physical space of the DR

environment. The physical objects (targets) can be replaced by their virtual avatars, which allows various interaction 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 on the objects. It saves visualization space and help to reduce the visual clutter and operational ambiguity caused by showing the physical targets and their avatars simultaneously. Besides, DR is also capable of building an information bridge between the changing physical space and the virtual space seamlessly. Regarding AR or SA (AR is considered as one of its four primary elements [16]), however, a new visualization space should be brought to the information presentation [16], then the contextual information can be provided around the physical targets.

Category III - Design Details

- Comment #08: the design goals are interesting for an immersive analytics application. Design considerations offers important information but it is not well presented. (R4)

Response to Comment #08: thanks, we have re-written Section 3.2 (Section Design Considerations), the presentation is re-organized, please check them in the paper.

- Comment #09: The manual scanning required by DRCmpVis would be tedious for large collections of books, i.e., in order for DRCmpVis to work, a user will have to scan potentially a huge area, depending on how comprehensive the user wants to be in his/her search. (R1)

Response to Comment #09: thanks for the comments. Users just need to scan several shelves where the target books are potentially placed, instead of scanning all the shelves. We have conducted a complete new user study to evaluate the tool, including the average timing results of scanning time for a user to find a target book. Most of the participants just need to scan one or two shelves to find it, because the books are grouped according to book topics (e.g., classification number).

Scenario	Scanning Time
Library Scenario	0.321
Cafe Scenario	0.454

Fig. 4 Timing and accuracy performance of DRCmpVis (seconds). The "Scanning Time" is the average time to scan a panoramic photo in the library scenario and a menu in the cafe scenario, respectively. We obtain the average results based on 16 tests.

Methods	Book Searching Time
Blind finding (without any tools)	4.56
With a DB retrieval system (targets are available on the correct shelf)	2.53
With a DB retrieval system (targets are inserted in wrong positions)	5.34
With a DB retrieval system (available but being viewed by other borrowers)	unlimited
The proposed *DRCmpVis*	0.65

Table 5 (Table 3 in the paper) A NEW study is conducted in the revision for comparative evaluation (minutes). It shows the task-driven quantitative results (the average time costs from 22 participants' tests) and some qualitative comparisons regarding *DRCmpVis* and two traditional methods. The evaluation include finding a target with given keywords, e.g., a book with given keyword in its title.

- Comment #10: In the applications, the real world is entirely blocked out and replaced with 3D replicas. Thus, interaction with the real world or real-time updates after the initial query is very limited.

Response to Comment #10: thank you for the reviews. The back-end scanning module of the tool will update the information between virtual world and physical world during the explorations of the virtual targets, users can set the update time interval. Besides, users can restore the layout to the physical world anytime if they want. In the study, we found multiple participants thought the restore function is useful for them to locate the positions of candidate books.

Category IV - Evaluation: Method Comparison and Discussion

- Comment #11: What are the advantages/disadvantages of *DRCmpVis* compared to browser-based (non XR) visual analytics tools for exploring books in the library (R1). At least the 3D presentation should be compared with a normal interactive ranking populated through from the database results to the initial query. (R4)

Response to Comment #11: thanks for the comments.

First, we have justified why we need DR to reorganize the additional information in the paper: we take the library/bookstore case of this paper as an example. One scheme to show additional information about

physical objects is to query them directly from the database of the library/bookstore. However, there are several limitations of this scheme due to the inconsistency between the physical world space and the virtual data space in a database, e.g., book position in a library/bookstore:

1) The books in a library/bookstore often would be put in the wrong position by a librarian or numerous users, which is inconsistent with the information in the database.
2) Users might frequently read books on desks or tables, making it challenging for others to locate 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 are often updated in the physical world while it is tedious for a librarian or a bookstore attendant to update the database frequently.

Last but not least, common users often have limited permission to access the database of a cafe or an eyeshadow shop. DR can get the nearly latest information from the real world, which makes the information between virtual space and physical space as consistent as possible.

Second, we have added a new task-driven quantitative evaluation results to compare with two traditional methods: blind finding (without any tools) and finding with DB-based retrieval system. We found the proposed *DRCmpVis* is much better than the traditional methods. The average time costs of book finding are 14.25%, 25.69%, and 12.17% of those in "blind finding", "DB-based retrieval system (the book is available on the correct shelf)", "DB-based retrieval system (wrong position)", respectively.

Methods	Book Searching Time	Latte Searching Time	Augmented Info	Target Comparison
Blind finding (without any tools)	4.56	0.45	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*	0.65	0.13	Color highlight Fisheye highlight Pop-up glyphs	Re-grouping Re-ranking Visual comparison

Fig. 6 Task-driven quantitative evaluation results (new experiments in the revision). We compare the proposed *DRCmpVis* with the traditional two methods. The results shows 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 taken by all participants to search for the keyword "Latte". Note: most coffee shops do not provide a retrieval system for users, thus they are marked as Not Available (NA).

- Comment #12: "Clear limitations of the framework and the applications should be discussed" "However, only the advantages were discussed. A clear disadvantage that has to be discussed". (R1)

Response to Comment #12: thanks for the suggestions! We have discussed the limitations in Section 7, please check them:

(1) 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. Besides, we find *DRCmpVis* can be easily extended to the 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.

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 they are in some crowded setting. In some cases, libraries will be influenced by the crowded environment, but in other cases, such as the cafe menu case, are irrelevant. We think *DRCmpVis* can benefit from scenarios where DR/MR visualization offers unique information unavailable to a user.

(2) Scalability issue on image segmentation and image labeling. It is worth noting that the image segmentation components of *DRCmpVis* are scalable and not limited by the object number, because the CNN and the OCR algorithm are run on the server which can even handle thousands of books in the library scenario in our experiments. More importantly, unlike the mobile device, the computation resources of the server are scalable enough and could be easily upgraded. As a result, whereas *DRCmpVis* recognizes almost all 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.

(3) The limitation of text recognition. *DRCmpVis* recognizes objects by images taken from mobile devices. Ideally, the user only needs to take one panoramic picture that contains all the objects. However, objects' details may not be recognized if they are too small in the image, because the user may stand too far away from the numerous objects. For example, in the library/bookstore scenario, instead of scanning all layers of the bookshelves, the user may walk closer to the bookshelves and scan one layer at one time by panoramic stitching due to lack of light or limited imaging quality.

The image segmentation & labeling service needs to request once due to an image recognition module on the client app of *DRCmpVis*, when the positions of the objects are not changed. Because the coffee menu in a cafe is often unchanged. Actually, we 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. In our strategy, the latest captured panoramic image will be saved to the buffer of the client app. The image recognition module will verify whether the newly captured panoramic image is saved in the buffer. If yes, the segmentation& labeling records in the buffer can be reused without requesting the server twice. This strategy is useful and efficient in almost all the usage scenarios due to the quick response by the client app. 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 is much more efficient after the first-time buffer construction in a new scenario environment.

(4) 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 6. The hardware configuration can be improved for more expensive service packages. Thus maybe the performance especially for the segmentation & labeling could be further improved. We plan to make *DRCmpVis* to be applied in more general usage scenarios in our daily lives. In the future, we plan to extend the usage scenario of DRCmpVis to others like choosing cups, fruits or flowers, and other more general scenarios in our daily life. Because objects with text on them or in different colors and shapes can be well recognized by trained neural networks. However, objects with different irregular 3D shapes and without textual information on them are difficult to be recognized by current algorithms including the-state-of-the-art neural networks.

- Comment #13: This approach differs from general AR apps that superimpose overlays onto the targeted physical objects in the real view. Compared to these existing AR approaches, what would be the advantage of *DRCmpVis*? I hope the authors discuss this issue more carefully. (R3)

Response to Comment #13: thank you for your reviews! Diminished reality offers the advantage of enhancing focus and attention by selectively removing or reducing distracting elements from real-world scenes. Multiple target objects can be flexibly grouped, sorted, searched (interacting with the real environment), and located (returning to the real environment and updating in real-time to ensure that the objects in the virtual world are as consistent as possible with the real objects) in the corresponding virtual environment. All these physical re-layout operations can be easily performed in a virtual environment with their virtual avatars, while AR can't provide the virtual avatars of the physical objects because the physical

objects are not diminished, their augmented information are just displayed in a pop-up widget surrounding them. We also further discussed *why we used DR and MR instead of other XR technologies*. Please check them in Para. 3 of Section 3:

DR is an optimal solution to keep information consistent between the physical world space and the virtual data space while significantly reducing visual clutter. Objects are data in the physical space of the DR environment. The physical objects (targets) can be replaced by their virtual avatars, which allows various interaction 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 on the objects. It saves visualization space and help to reduce the visual clutter and operational ambiguity caused by showing the physical targets and their avatars simultaneously. Besides, DR is also capable of building an information bridge between the changing physical space and the virtual space seamlessly. Regarding AR or SA (AR is considered as one of its four primary elements [16]), however, a new visualization space should be brought to the information presentation [16], then the contextual information can be provided around the physical targets.

- Comment #14: I think the system will also not generalize well to different scenarios/environments or at least this was not convincingly evaluated or proven in the paper. (R1) The paper lacks a thorough description of how visualization researchers can leverage this framework to create and develop their own applications. (R3)

Response to Comment #14: thank you for the comments.

First, there are four showcase example applications in the paper, as shown in Fig. 7, the fourth is a DIY DR tool created by a visualization researches. The latter two are moved to the Appendix file due to page limits, as shown in Fig. 8, please check the thorough descriptions about the two scenarios in Appendix.

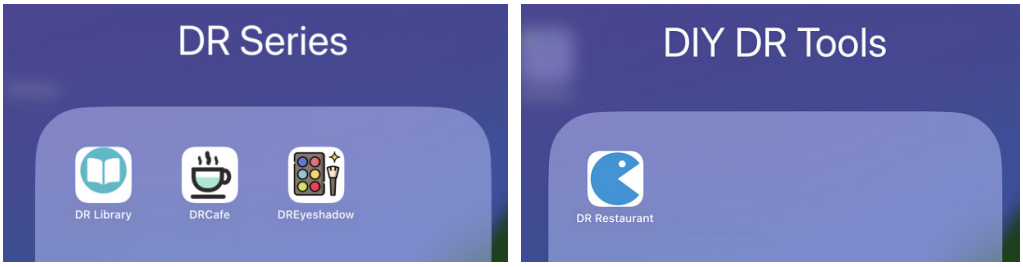


Fig. 7. Four showcase example apps in the paper. The last one is a DIY DR tool created by a visualization researcher, which is a case in a restaurant scenario (Shaxian County cuisine, in Chinese), the latter two are moved to the Appendix file due to page limits

(a) Scan various eyeshadows in the original physical space

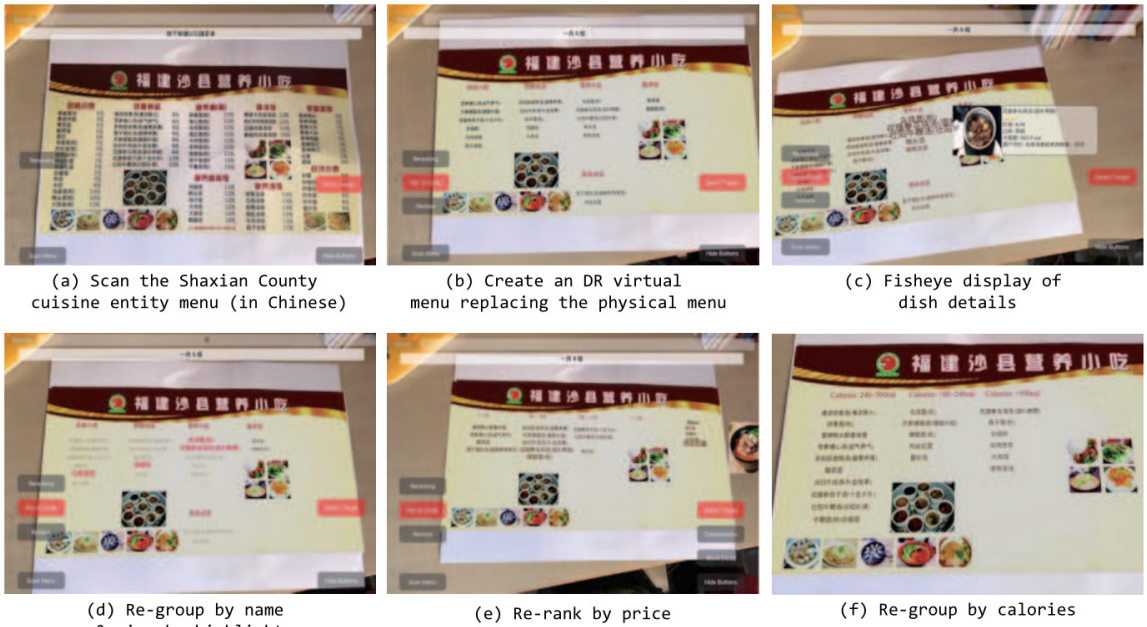
(b) Re-group by eyetypes & view them by fisheye

(c) 3-D virtual makeup try-on

(d) Re-group by scheme numbers

(e) Search by scheme number & highlight in the original layout

Fig. 8 [in Appendix] an eye-shadow scenario. This scenario extracts image/texture information from the target objects, it shows a user compares eye-shadows using the framework of *DRCmpVis*. (a) Scan various eye-shadows displayed on a cosmetic table. (b) Re-group eye-shadows by eyetypes and view the augmented information of the focus one as well as an eye image showing places to apply it on. (c) View the effects of candidate eye-shadows via 3D virtual makeup try-on. (d) Re-group eye-shadows by scheme numbers. Different scheme numbers have different features like "Deep Blue" or "Soft Smokey". (e) Choose "Scheme 13" in their original physical layout, eyeshaodws belongs to this scheme number are highlighted.



(a) Scan the Shaxian County cuisine entity menu (in Chinese)

(b) Create an DR virtual menu replacing the physical menu

(c) Fisheye display of dish details

(d) Re-group by name & view by highlight

(e) Re-rank by price

(f) Re-group by calories

Fig. 9 [new in Appendix] a restaurant scenario (Shaxian County cuisine, in Chinese). In this scenario, we extend the usage scenario to Chinese texts recognition. The framework is independent on language environment, because the deep network on the server supports cross-language. It enables compare more additional information about

different Chinese dishes. (a) Scan the Shaxian County cuisine entity menu (in Chinese). (b) Create an DR virtual menu replacing the physical menu. (c) Fisheye visualization shows dish details (price, taste, reviews, ingredients, etc). (d) Re-group dishes by name keywords. Highlight the dishes with the keyword Chicken in the virtual menu. (e) Re-rank dishes based on different price ranges after re-grouping. (f) Re-group dishes by calories.

Second, how visualization researchers can leverage this framework to create and develop their own applications. The steps are list as follows:

(1) Database construction of augmented information. The augmented information can be added by visualization researchers. For example, the book dataset is downloaded from the open data website "Amazon product data" [21, 35, 36], containing product reviews and metadata from Amazon, and including 142.8 million reviews for their products and 22.5 million reviews for books. The coffee database is created by the web crawler, which crawled collections from Starbucks, including the coffee's name, description, ingredients list, preview image, process introduction.

(2) Coordinate transformation between physical space and virtual space. This step is easy and simple to be done by visualization researchers, because it is mainly achieved by the camera with LiDAR scanner, we have encapsulated the related functions into the APIs of the framework.

(3) Choose or design new visual comparison components into DR/MR context. We have designed several commonly used visual presentation 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 visualizations, which are appreciated by the participants in the user study. Besides, we adopt a Focus+Context exploration scheme by using fisheye algorithm, which scaling 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. All the related functions are encapsulated into the APIs of the framework.

Note: some steps are automatically finished by the deep network of the framework, e.g., the image/texture segmentation, image labelling, OCR-based text extraction, image recognition. Please check them in Appendix file.

- Comment #15: The framework and applications support only very simple visualization tasks (select, filter, group) and it is unclear how the DR framework can be applied to other advanced visualization tasks. (R1, R2, R3)

Response to Comment #15: thanks! In the pre-study survey before the questionnaire step, we have one question to survey how many users like simple visualizations tasks or advanced visualization tasks in the

DR applications. The survey result indicates that 98.8% of participants chose an app with simple visualizations. Thus, we just integrated some commonly-used simple visualization components/techniques into the framework, e.g., bar charts, line charts, word clouds, ingredient glyph, small multiples, F+C techniques, etc. All the related functions are encapsulated into the APIs of the framework.

Furthermore, advanced visualizations can also be integrated into the framework. One of the most important criteria to select the visualization types is whether they are general-purposed, whether they are simple or advanced. Please check them in Section 6.3 in the paper.

Category VI - Evaluation: User Study

- Comment #16: The evaluation primarily relies on users' questionnaire responses, without clear confirmation from the authors on whether there were significant differences in quantitative results, such as task time, between the two groups. (R3) The study could have used other tools in such evaluation and would provide some more objective scoring along already validated dimensions (R4).

Response to Comment #16: we appreciate the comments for improving the evaluation of the paper. We have conducted several new experiments on quantitative evaluation in the revisions.

(1) The first experiment is to achieve the timing results of *DRCmpVis*, including the scanning time, segmentation time, processing time and the accuracy of segmentation and labelling, as shown in Fig. 10.

Scenario	Scanning Time	Segmentation Time	Processing Time	Seg. & Labelling Rate
Library Scenario	0.321	0.631	0.952	95.13%
Cafe Scenario	0.454	3.891	4.345	100.00%

Fig. 10 Timing and accuracy performance of DRCmpVis (seconds). The "Scanning Time" is the average time to scan a panoramic photo 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 each back-end server request, while the "Seg. & Labeling Rate" is the average accuracy. We obtain the average results based on 16 tests.

(2) The second experiment is designed to compare *DRCmpVis* with two traditional methods: blind finding (without any tools) and finding with DB-based retrieval system. We found the proposed *DRCmpVis* is much better than the traditional methods. The average time costs of book finding are 14.25%, 25.69%, and 12.17% of those in "blind finding", "DB-based retrieval system (the book is available on the correct shelf)", "DB-based retrieval system (wrong position)", respectively, as shown in Fig. 11.

Methods	Book Searching Time	Latte Searching Time
Blind finding (without any tools)	4.56	0.45
With a DB retrieval system (targets are available on the correct shelf)	2.53	NA
With a DB retrieval system (targets are inserted in wrong positions)	5.34	NA
With a DB retrieval system (available but being viewed by other borrowers)	unlimited	NA
The proposed *DRCmpVis*	0.65	0.13

Fig. 11 We compare the proposed *DRCmpVis* with the traditional two methods. The results shows 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 taken by all participants to search for the keyword "Latte" in eight instances. Note: most coffee shops do not provide a retrieval system for users, thus the time is marked as Not Available (NA).

(3) We also conducted a NASA-TLX evaluation to evaluate the performance of the proposed *DRCmpVis* with two traditional methods, as shown in Fig. 12. We find the proposed *DRCmpVis* is statistical significantly better than the two traditional methods in terms of the latter five demands (Fig. 12). However, the mental demand of "Blind finding" is significantly better than the other two methods, because the "Blind finding" is the simplest approach which just has the smoothest learning curve.

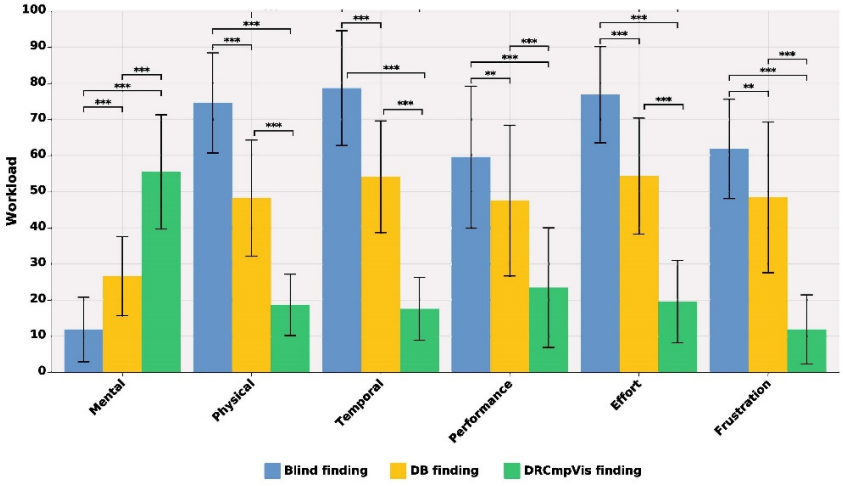


Fig. 12 (A NEW study conducted in the revision) the scores of NASA-TLX evaluation for two control groups of methods and the proposed *DRCmpVis*. Error bars indicate standard errors. Statistical significant differences are denoted by ** ($p < 0.01$) , *** ($p < 0.001$).

Category VII - Reference

- Comment #17: "A revision should incorporate additional papers related to their study" (R1, R4). "Add references and comparison (either descriptive or in user study) with visual analytics tools for libraries as well as related immersive analytics tools". (R1)

Response to Comment #17: thanks for the suggestions. We have added four related literatures (the first three are mentioned by R1) on visual analytics tools or immersive tools in Section 2 (Section Related Work). Besides, we have added the description about the relationship with the proposed *DRCmpVis*.

+ B. Shneiderman, D. Feldman, A. Rose, and X. F. Grau. Visualizing digital library search results with categorical and hierarchical axes. In Proceedings of the fifth ACM conference on Digital libraries, pp. 57–66, 2000.

+ A. Thudt, U. Hinrichs, and S. Carpendale. The bohemian bookshelf: supporting serendipitous book discoveries through information visualization. In Proceedings of the SIGCHI conference on human factors in computing systems, pp. 1461–1470, 2012

+ 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," 2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct), Merida, Mexico, pp. 203-208, 2016

+ 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 Transactions on Visualization and Computer Graphics, 27(9):3808–3825, 20

Category VIII - Presentation and Writing

- Comment #18: This is minor, but for the example library scenario in Sec. 5.1., it would really be helpful for readers if Fig. 4 was on the same page so that the reader can easily follow the text with the many references to the figure. The same applies for the coffee shop scenario. (R1)

Response to Comment #18: thank you, we have moved the figures to the texts where first cited them as near as possible. Please check them.

- Comment #19: I highly recommend that the authors ask a native English speaker to help with revision. (R1, R4)

Response to Comment #19: thank you for your suggestions. We asked a non-free native speaker in **Harvard Medical School** to proofread the paper carefully, who is also an expert in data visualization and human-computer interaction. Besides, the authors of the paper at Stony Brook University also have participated in the revision of the paper.

(The End of "Revision Response")

Review Comments of IEEE VIS'23

Reviewer 1
The Review
The paper presents DRCmpVis - a system for visual comparison of physical objects using mobile diminished reality. The system helps users find a target object by providing filtering, re-grouping, and re-ranking of objects in virtual space using digital information related to the objects, followed by highlighting of final selected object in real-world space, i.e., during physical retrieval. The filtering, re-grouping, and re-ranking are presented using diminished reality, i.e., the real-world scenario is replaced/overlaid with the virtual scenario.

On the positive side, I think the presented system has a lot of technical credibility, i.e., it is feasible with existing technology, and obviously the authors did a fantastic job and must have put in quite a lot of effort into building the system. I also think that the ingredients are there for a potentially very nice paper, but probably not at its current state and maybe not for the VIS conference.

On the negative side, I think the paper is not suited for VIS, but probably better for either IEEE VR or ISMAR after some polishing. The reason for this is that in terms of data visualization, I did not gain any new insights and felt that as a whole, the paper still seems to be half-baked and not ready for publication. For these reasons, I gave the paper a rating of 2 for reject. I provide more detail on my reasons, as well as some questions and suggestions, below.

Reasons for score:
-While the implemented system itself is very impressive, I felt that the novelty in terms of visualization design/techniques is lacking. I believe there is great potential in this work, but more has to be done in innovating in terms of maximizing the technical features combined with the concept of diminished reality. For instance, what new visualization designs and/or user interactions are possible given the technical capabilities? What can be done in diminished reality and not in augmented reality within the presented scenarios? I was expecting diminished reality to be a key feature in the paper especially given the title, however, I felt that its use was not well motivated and its benefits are not well explained and demonstrated.
-I think the system will also not generalize well to different scenarios/environments or at least this was not convincingly evaluated or proven in the paper. Will the system just work if a user tries DRCmpVis in an arbitrary library/bookstore/cafe? What does it take to make it work for a given new scenario?
-The writing and exposition needs improvement. There are some sentences that are just confusing - I highly recommend that the authors ask a native English speaker to help with revision. Furthermore, I think the motivation has to be improved to further establish the need for their system especially the use of diminished reality techniques. I also think some questions have to be addressed (listed below).

Questions to address:
-What are the advantages/disadvantages of DRCmpVis compared to browser-based (non XR) visual analytics tools for exploring books in the library (some listed below)? There was a brief mention of this in Sec. 3, however, only the advantages were discussed. A clear disadvantage that has to be discussed, in my opinion, is that manual scanning required by DRCmpVis would be tedious for large collections of books, i.e., in order for DRCmpVis to work, a user will have to scan potentially a huge area, depending on how comprehensive the user wants to be in his/her search.
-One can argue that a user can also use a browser-based tool for searching for a book (e.g., digital library or library websites, or applications like Bohemian Bookshelf [Thudt et al. @ CHI 2012]) and once a book is found, then they can use the search and highlight feature of DRCmpVis to perform the final physical search

and retrieval of the target book. I mention this not to discourage the authors but to try and think of ways to improve DRCmpVis to differentiate it and add features that will be unique and necessary for the search/exploration task. Maybe a study or comparison between the two approaches can even be made to gain new insights.
-How accurate is the scanning and detection of books? In what cases does it not work? In Table 2, how were the segmentation rates and times measured (e.g., for how many objects or images/frames)?
-What are the benefits of diminished versus augmented reality?

Suggestions:
-In abstract: rephrase "stereotyped strategies" into something clearer - what do the authors mean by this?
-Ask a native English speaker to go through the paper and improve grammar and exposition.
-This is minor, but for the example library scenario in Sec. 5.1., it would really be helpful for readers if Fig. 4 was on the same page so that the reader can easily follow the text with the many references to the figure. The same applies for the coffee shop scenario.
-Add a figure that shows an example of the actual image of the shelf and the recreated virtual version to demonstrate the accuracy of the scanning and detection.
-Add references and comparison (either descriptive or in user study) with visual analytics tools for libraries as well as related immersive analytics tools, some of which are listed below.

The bohemian bookshelf: supporting serendipitous book discoveries through information visualization: https://dl.acm.org/doi/abs/10.1145/2207676.2208607

Visualizing Digital Library Search Results with Categorical and Hierarchical Axes: https://dl.acm.org/doi/pdf/10.1145/336597.336637

HORUS EYE: See the Invisible Bird and Snake Vision for Augmented Reality Information Visualization: https://ieeexplore.ieee.org/document/7836498

I believe this paper has some nice technical ingredients that can lead to a really nice paper in the near future once the issues are addressed and new insights/designs/techniques are presented!
Marked-up Copy of the Paper

Reviewer 2
The Review
This paper proposes to use augmented reality on a tablet to present a situated analytics application using a form of diminished reality (DR). The systems scans the environment, detects objects and then presents selections of the virtual replicas of the object arranged in 3D. These replicas can be presented over a video background faded to grey or masked out to obtain a DR effect. Users can analyze the data by manipulating the replicas. A qualitative usability study (conduce) documents strongly positive reactions of the users.
Watching the video makes it clear that the strong point of the paper is the in-place search, filtering and highlighting (the "DR" mode). The user can immediately see the actual book (or its digital twin) rather than having to refer to an indirect database entry.
The situated analytics application itself has rather straight forward functions (select, filter, group etc.). The strength of the paper lies in its data preparation stage. The technical pipeline (scan, segment, text recognition, cross-referencing with database) for building the digital very impressive. It probably does not generalize trivially to other settings than the shown ones (library, coffee menu), since a lot of prior knowledge about the scenario is required, but generalized digital twins can be seen as out of scope of this work.
The views showing replicas out-of-place (e.g., the virtual empty shelf in the library scenario) seem less powerful, since they are essentially a 3D (or VR) version of a conventional table visualization. However,

the use of 3D replicas means there is a persistent representation of the objects throughout the entire experience.
This brings us to the question about the research contribution and the evaluation. The work is presented as a system, and the user study focuses on the usability of the system. While evaluation results are clearly positive on the aspect of usability, no comparison of the "DR visualization" to any alternative user interfaces is performed. This is disappointing, since the real core question is where a DR-style interface is preferred over a more conventional one. It could very well be that showing the replicas in-place works extremely well, while showing replicas out-of-place has little or no advantage over a conventional interface (e.g., a 2D table on the tablet). To obtain robust design guideline, the effect of the data preparation for the digital twin, the in-place DR visualization and the out-of-place DR visualization would have to be examined separately and compared to baseline conditions.
Since I am rather excited about the idea and the technical implementation, and given that this work is, in a way, a first of its kind, I am still leaning towards acceptance. VIS will be well off giving preference to innovative work such as this over the many more incremental topics that come up in the field.
Marked-up Copy of the Paper
(no file)

Reviewer 3
The Review
This paper introduces DRCmpVis, an Augmented Reality (AR)/Diminished Reality (DR) application framework that enables three reorganization schemes: filtering, regrouping, and reranking physical objects (such as books) from the real world. The authors have developed two applications that allow for grouping and ranking of physical products (as well as text on a menu) within our physical environment. These applications consider various attributes associated with the objects (e.g., rating, publisher, year, keywords, and price). The author utilizes an AR-based approach, where scanned physical objects are virtually rearranged in the AR space.

However, this reviewer has several concerns and offers suggestions for improvement, as follows:

The authors have described DRCmpVis as a framework based on DR/MR applications. Although the specific details about the framework are not entirely clear, a framework is typically a collection of common functionalities and reusable code that developers can utilize to build software applications more efficiently. If we assume my understanding is accurate, DRCmpVis should offer various codes, classes, libraries, tools, etc. However, the paper lacks a thorough description of how visualization researchers can leverage this framework to create and develop their own applications.

The library scenario presented in the paper was very interesting; that said, I wasn't entirely convinced by the system's performance in this specific scenario. I'm curious to know if the system actually enables users to visit any library and scan books without preprocessing. The authors failed to explain the preprocessing steps they undertook using the "framework" to build the application. I wonder if this scenario would be feasible for any library or bookstore, as it requires scanning and accurately recognizing book spines. Additionally, it is still unclear whether the application has the capability to recognize every item solely through scanning, which I believe would be quite challenging. I also want to more fully understand how easy the framework is for building such applications, as well as what the workflow entails for developing these demonstrations. It would be helpful to have more information about the steps involved in creating these applications.

The presented techniques primarily focus on supporting basic search and filtering tasks. It is unclear how the DR framework can be applied to other advanced visualization tasks. For instance, including a discussion

section that explores the potential applications of DRCmpVis for both general and complex visualization tasks would be beneficial.

The two applications presented in the paper are quite fascinating. Recently, we have been witnessing a rise in AR-based applications. While many existing AR-enhanced shopping apps superimpose digital or virtual content onto a user's real-world view or physical products (e.g., ARkit's Bookshelf Demo), this particular work takes a different approach by partially blocking out the physical world and rearranging physical objects (the author refers to this as the concept of DR). In our physical reality, rearranging physical objects is a challenging task due to various limitations like space, weight, surface, and information. In comparison to other similar AR apps, DRCmpVis takes a different approach by blocking the real view and physical objects using associated virtual objects. This approach differs from general AR apps that superimpose overlays onto the targeted physical objects in the real view. Compared to these existing AR approaches, what would be the advantage of DRCmpVis? I hope the authors discuss this issue more carefully.

The evaluation and its results seem to be weak. The evaluation primarily relies on users' questionnaire responses, without clear confirmation from the authors on whether there were significant differences in quantitative results, such as task time, between the two groups. It is also unclear if formal hypothesis testing was conducted to obtain quantitative evaluation results. Additionally, there is a lack of clarity regarding the methodology employed for the quantitative evaluation in terms of study procedure, task assignments, participant groups, etc.

The author initially presented only two examples, but in Section 6.4, they mentioned "three examples."

In summary, while the presented applications show potential for future data visualizations, they do not seem to make significant contributions to the field of InfoVis and Visual Analytics. I found it challenging to fully assess how the proposed framework can be utilized to develop new DR (Dimensionality Reduction) data visualizations. The study method and results did not provide enough evidence to convince me that the presented approaches are sufficiently effective for practical data analysis. A more in-depth discussion on the generalizability of the framework to other visualization tasks and its specific use case could offer better insights into the work. Given these deficiencies, I would not recommend publishing this paper in its current state for VIS.

The Summary Review (Due by May 15)
Overall, the reviewers expressed positivity regarding the technical contributions of this paper, acknowledging the effort and technical pipeline involved in building the system and its applications. However, each reviewer raised concerns about the current evaluation and visualization tasks and outlined several revisions necessary for acceptance to VIS 2023. During our discussion, a consensus emerged among the majority of reviewers that the extent of these changes would surpass the time constraints of approximately 3.5 weeks and could warrant a re-review. Considering the feedback from the reviewers, it is recommended that the authors conduct completely new studies, undertake significant revisions, and resubmit the paper to the IEEE Transactions on Visualization and Computer Graphics (TVCG).

*Strengths:
+ The basic concept of the framework is interesting, and the scenarios illustrate a number of seemingly well-thought-out techniques. (R4, R3)
+ The presented system shows potential for future data visualizations. (R2, R3)
+ The presented system demonstrates a level of technical credibility, and the reviewers applaud the authors for their effort in building the system. (R1, R2, R3, R4).
+ The video is well-made and shows an interesting combination of techniques. (R2, R4)

* Weaknesses and Suggestions:
- The design rationale and considerations (Section 3) make little sense and/or are not well presented. (R4)
- The paper lacks a thorough description of how visualization researchers can leverage this framework to create and develop their own applications. (R3)
- In the applications, the real world is entirely blocked out and replaced with 3D replicas. Thus, interaction with the real world or real-time updates after the initial query is very limited. (R2, R4)
- The implementation section offers little information about challenges and their resolution. (R4)
- The study results based on a simple questionnaire lack sufficient evidence to convincingly demonstrate that the presented DR approaches are effectively practical for visual analysis (R1, R2, R3, R4)
- The study primarily relies on users' questionnaire responses without clear confirmation from the authors on whether there were significant differences in quantitative results. (R3)
- The study could have used other tools in such evaluation and would provide some more objective scoring along already validated dimensions. (R4)
- The current comparison between the two groups in the study is not feasible. It is recommended that the presented DR applications be compared with existing AR or other UI approaches for the search/exploration task to provide a more meaningful comparison. (R1, R2, R3, R4)
- The authors should conduct completely new studies to quantify and present more valuable insights regarding participants' interactions with the system and DR and the appropriate comparisons (see the previous comment). (R2, R4)
- It is not clear whether the present techniques and framework will be able to generalize well to different scenarios/environments. (R1, R2, R3)
- The framework and applications support only very simple visualization tasks (select, filter, group) and it is unclear how the DR framework can be applied to other advanced visualization tasks. (R1, R2, R3)
- The framework's design lacks strong motivation rooted in research problems and challenges for the use of Diminished Reality techniques in the context of visual analysis. (R1)
- The authors are encouraged to review and familiarize themselves with the definitions and concepts related to methods and arguments that align with the concepts of diminished reality, mixed reality, or situated analytics. (R4)
- Clear limitations of the framework and the applications should be discussed. (R1)
- It is essential to evaluate the system performance of the presented techniques/frameworks, which should include a proper assessment of factors such as scanning, detection, segmentation rates, and processing times. (R1, R3)
- A revision should incorporate additional papers related to their study. (R1, R4)
- Overall, the writing quality and the exposition's clarity should be improved significantly in the revised version. (R1, R3, R4)

Reviewer 4
The Review
This paper presents an interactive querying, filtering and ranking tool for decision making in the real world. The proposed solution starts by capturing the environment with a wide angle lens camera, recognizing objects and retrieving textual information from a database service. With the information available, diverse methods for grouping, filtering and and selecting items are presented for 2 scenarios: a library and a coffee shop. A user study is carried out in the context of both applications which compares performance of using the proposed tool or searching / selecting visually.

The technique is presented as a diminished or mixed reality technique. The presentation includes preliminary discussions about data attributes and what operations the application should enable. The video shows an interesting combination of techniques.

This paper is challenging to rate. The concept is interesting and the scenarios illustrate a number of seemingly well thought out techniques. But there are numerous misconceptions and a lack of depth in the text. With just the paper, it is difficult to appreciate the value of this work. The experiment could have presented a more challenging comparison. As it is, the experiment adds little value except that users find it useful. I will elaborate on this review.

The technique is presented as a diminished or mixed reality. It is actually neither. Unless I missed something, there is no interaction with the real world, or real-time update after the initial query. The technique starts by capturing the environment and uses computer vision implemented with two neural networks to segment objects and retrieve information from a database. With the data retrieved from the database, a new 3D representation is created. In both cases, a gray semi-transparent background hides the real-world and the new representation is shown on top. The visual comparison, grouping, filtering, etc operations take place in this 3D representation. There is no selective modification of the real-world perception. This is a technicality. It can be solved by removing references to diminished reality and mixed reality and referring the techniques to situated analytics. It would require some edition of the paper and related work, but would be possible in the time available.

Situated analytics: Demonstrating immersive analytical tools with augmented reality
NAM ElSayed, BH Thomas, K Marriott, J Piantadosi… - Journal of Visual Languages & Computing, 2016

The design rationale makes little sense, but the design goals are interesting for an immersive analytics application. Design considerations offers important information but it is not well presented. It presents the operations that are expected from any visualization application re-grouping, ranking on attributes, etc. But it is not clear how this is supposed to interact with the real world or with the physical referents of the information being analyzed.

The implementation section offers little information about challenges and how they are solved. Possibly, the most interesting section is the one describing the visual comparison components and the scenarios.
The study proposes two tasks, one for each scenario and allows users to perform a number of operations. Then, a questionnaire created by the authors is used to assess utility of the main features. There are a number of questionnaires that can be used to assess usability (SUS), cognitive load (NASA TLX) and also satisfaction or other aspects. These tools are commonly used in such evaluations and would provide some more objective scoring along already validated dimensions.
The tasks in the study compare the proposed techniques with performing a search visually with some quantitative measurements.

Once one realizes the technique presented is principally a query/search technique that uses images from the real world to fetch information that is then visualized and analyzed through grouping, filtering, etc. After this realization, it is clear that the comparison of this technique against no support is unfair. At least the 3D presentation should be compared with a normal interactive ranking populated through from the database results to the initial query.

The paper is difficult to read. I would suggest to get help from a native speaker. But also to think what is the information you are trying to convey in each section and to clearly, succinctly but with detail in depth to describe that. There is also quite some repetition in the text.

I believe the techniques presented in this paper are really interesting and I praise the authors for the approach. However, where it not for the video, I would have missed how interesting it is. I would suggest to also look

into the definitions of methods and really argument along the right lines: diminished reality, mixed reality, or situated analytics.

(The End of "Review Comments of VIS'23")

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

Abstract— Numerous physical objects in our daily lives are grouped or ranked according to stereotyped presentation style. For example, in a library, the books are normally grouped and ranked based on the classification number. However, for better comparison, we commonly need to re-group or re-rank the books with other attributes like their ratings, presses, comments, published years, keywords, prices, etc, or a combination of them. In this paper, 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 physical 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. It 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 user study, NASA TLX assessment, quantitative evaluation, and case studies using different scenarios.

Index Terms—Diminished reality, visual comparison, virtual avatars, mixed reality

1 Introduction

The development and popularity of extended reality (XR) devices and the techniques have led to an increasing number of studies designing new application tools. XR generally consists of virtual reality (VR), augmented reality (AR), and mixed reality (MR). MR is strictly defined by Milgram and Kishino [38], which was considered as a mixture of real and virtual objects within a single display. The distinctions between AR and MR are fuzzy [46]. 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 [17]. 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. [17] think SA is a new area of research at the intersection of visual analytics and AR. Besides, a new concept diminished reality (DR) [39,40] was further introduced recently. DR pertains to the manipulation of a perceived environment in real-time, involving actions like concealing, eliminating, or revealing objects [39,40]. According to the survey on DR [39] summarized by Mori et al., DR examples include four types: diminishing, seeing through, replacing, and inpainting real objects.

Stolte et al. [45] have summarized 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" [45]. ~~According~~Furthermore, according to Jacques Bertin's book "The Semiology of Graphics" [5], data types of visual variables include *nominal*, *ordinal*, and *quantitative*. To enable users in finding and comparing physical objects with multi-dimensional attributes. Considering these two principles [5,45], we further structure the data flow space in DR/MR context 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 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 DR/MR environment (available attributes: **ordinal**, **quantitative**).

In our everyday life, we often spend a great amount of time searching for a specific object from numerous candidates (e.g., searching for algorithm-related books in a library or a bookstore). In this case, we may get limited information about the objects from the ~~visual presentations~~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. ~~The used additional attributes could be nominal, ordinal, or quantitative.~~ Considering a ~~use case~~usage scenario inside a library or a 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 Figure 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 probably 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 Figure 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 Figure 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.

Except for the example of finding/comparing targets from numerous candidates, 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

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

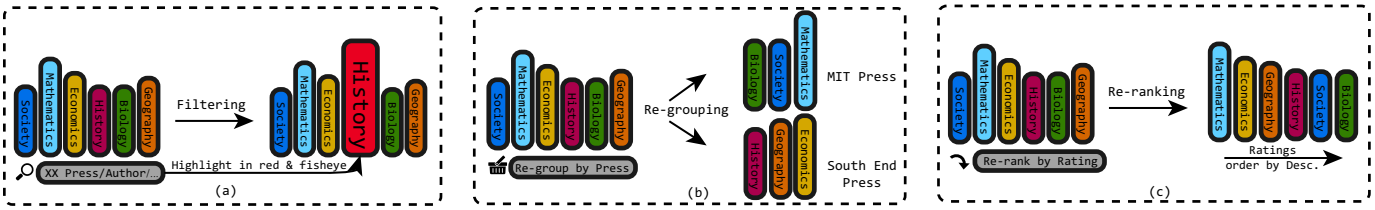


Fig. 1: Three types of data flow tasks within the DR/MR-based computational framework: (a) filtering, (b) re-grouping, and (c) re-ranking in DR environment. We take a library scenario as an example.

with multi-attributes, when we in a cafe.

The tasks mentioned above in our daily lives present three main challenges. First, it is tedious for us to find the target objects from numerous candidates, especially when we only know some fuzzy information/keywords of the targets. Second, the object information visually presented on the physical objects is limited to help us compare the candidates progressively and then find the final targets. Third, the original physical layouts of the objects are often in a stereotyped presentation style and of little use in object comparison, e.g., the books on the bookshelves are often sorted by the classification number in libraries/bookstores while we often need to compare them using multi-dimensional attributes (publishers, rated scores, topics, keywords, prices, publication year, etc.).

To address the issues, we propose an interactive application framework named *DRCmpVis,* ~~to~~enabling visually compare numerous physical objects with text labels/information in mobile diminished reality. It builds multidimensional comparisons ~~by~~avoiding breaking the original physical layouts and provides additional augmented information by comparative data presentations in an identical context. 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 ~~distinguish~~recognize different objects. ~~According to the survey on DR [39] summarized by Mori et al., DR examples include four types: diminishing, seeing through, replacing, and inpainting real objects.~~

In our work, *DRCmpVis* replaces the real objects with virtual objects, then we mainly used the term DR in this paper. Strictly speaking, plenty of virtual information of targets is also provided in the reality environment, thus we also use the term MR. With *DRCmpVis*, multi-dimensional comparisons can be completed by filtering, re-grouping, re-ranking, and their combinations in DR context. The additional augmented information of the objects can be encoded into some simple visual comparisons in MR context.

We use a trained convolutional neural network (CNN) named PaddleSeg [34] to segment and label all the objects. Furthermore, we extract the text information by an OCR-based neural network. In the experiment, we evaluate the proposed *DRCmpVis* using four ~~ease~~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 the three different types of attributes (nominal, ordinal, and quantitative), and then integrate ~~visual comparisons~~commonly-used visualizations into DR/MR context to achieve ~~different~~flexible object comparisons.
- We design three ~~reorganization~~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.

2 Related Work

Visual comparison aims at providing visual support for the understanding of underlying abstract data sets [19]. The visual comparison tasks in this paper are a little different from the traditional ones because the compared items in *DRCmpVis* are physical objects.

2.1 XR-Based Data Visualization

There is not many literature that strictly defines the differences between VR, AR, and DR, while XR is often considered as consisting of VR, AR, MR, and DR. DR refers to the removal of physical objects from real-time video [39, 40]. In a narrow sense, it is different from AR, which shows the physical reality of the world. AR-based visualizations [27, 54] 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.

Embedded data representations are capable of linking systems to physical things [51]. As a significant method to connect digital data with physical world, XR can realize data presentation in the physical space to promote certain visual explorations and combine presentations with personal ideas and preferences [7]. When integrating ubiquitous data into everyday life, spatial immersion issues like depth perception, data localization, and object relations become relevant. Works concerned with XR nowadays can be roughly classified as mobile (or tablets) handhelds [15], and head-mounted displays (HMD) systems [24] according to the computing paradigms. The Hololens device consists of a depth sensing camera that roughly calculates the distance of each pixel in view and pieces together a mesh or spatial map of the environment [18]. Google Tango [35] and Intel Realsense [29] offer similar technologies. The software development kit (SDK) [2] provides programmers with more freedom and flexibility to ~~use their own inspiration to~~ design excellent immersive applications with their own inspiration, such as ARToolkit [25, 49], Vuforia [13], and ARCore for Android [3]. A-Frame [1] enables the public to create immersive scenes in the browser integrating by WebVR [50] content within HTML.

XR-based data presentations have been applied to many fields. The CityViewAR [20] provides information about destroyed buildings and historical sites that are destroyed by the earthquakes. Focus and context information can also be separated by well-designed AR techniques [26]. Then, XR in the interpretation of terrain relief [8] shows great usability, which ~~serves~~functions as a motivational tool for 3D data presentations. Applications in the ~~newly~~lately emerging field of Augmented Reality Art show the ~~paradigmatic~~canonical potential of XR as a new artistic ~~medium~~intermediary [48].

We find few recent related works focused on DR-based applications, especially for data presentations. For example, Kawai et al. [28] find 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 new method [42] of blending and replacing textures is 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 use the blended results into 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. [21] 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. [23] design 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. [32] 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.

2.2 Interactive Immersive Building Tools

There are usually more technical challenges in immersive authoring tools compared with the pure desktop PC environment due to two gaps [10]. The first one is the steep learning curve of programming on the embedded immersive devices such as HMDs. The second one is the tedious offline workflow where users are required to debug and program frequently between immersive devices and desktop PCs [10].

Many tools have been proposed that allow interactively building and exploring data in an immersive environment. For example, MARVisT [10] allows users without background expertise to bind data on ~~real-world~~physical-world objects to ~~create~~realize expressive AR glyph-based visualizations. DXR [44] further provides a GUI for easy and quick edits and previews of data presentations immersed in the virtual world. IATK [12] allows for easy assembly of data presentations through a grammar of graphics that a user can ~~configure~~allocate in a GUI, in addition to a dedicated API. PapARVis [11] is capable of designing an environment that can debug both static and virtual content simultaneously. Automated Window/Icon/Menu/Pointing Device User Interface (WIMP-UI) [52] generation has been ~~considered~~thought about a promising technology for ~~at least~~over two decades. iVisDesigner [14] achieves ~~high interactive expressiveness~~ high level of interaction ~~through~~by means of conceptual modularity, covering a ~~broad~~vast information presentation design space. A mixed-initiative system Voyager [53] that supports faceted browsing of recommended charts chosen according to statistical and perceptual measures.

2.3 Relationship with The Most Related Work

Some library tools were designed to help users better explore books, including Hieraxes [43] and Bohemian Bookshelf [47]. Hieraxes integrates 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 help 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 [16] is 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.

We note that there are several related XR-based data presentation tools [10, 11]. We summarize and discuss the differences between *DRCmpVis* and the most related ones as shown in Table 1 according to the data scale, tasks (augmented information, searching, re-grouping, re-ranking), visual presentations (glyph, small multiples, fish eye highlight), workflow (personal, single, or collaborative).

First, one of the differences between our work and the existing XR-based data presentation tools like MARVisT [10] are the data scale and the tasks, we focus on numerous objects, especially for the case that the number is tens, hundreds, or even thousands. Actually, *DRCmpVis* can handle more than 1,000 physical objects or even much more like books in a library/bookstore due to the efficient client-server design and the high rates of image segmentation and recognition of the backend on the server, whereas most of the XR-based related tools just

	Data Scale (Physical Target Number)	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	virtual	■	■				■	■		Sin(PV)
AVT	virtual	■	■							Collab
VRIA	virtual	■	■				■	■	■	Collab
VR Visc	virtual	■	■				■	■		Sin(PV)
VR Collab Vis	virtual	■	■							Collab
IATK	virtual	■	■				■	■	■	Collab
SA Vis	<5		■					■		Sin(PV)
PapARVis	<5		■					■		Sin(PV)
MarVisT	<30		■				■			Sin(PV)
Our Work	40~1000+		■	■	■	■		■	■	Sin(PV)

Table 1: Comparison to the most related recent work about data presentation tools towards VR, AR, or DR. DXR [44], Augmented Virtual Teleportation (AVT) [41], Situated Analytics (SA Vis) [15], Data Visceralization (VR Visc) [30], Shared Surfaces and Spaces (VR Collab Vis) [31], IATK [12], VRIA [6], PapARVis [11], MARVisT [10]. The workflow can be categorized into PV (single user in a personal data presentations), single user (Sin) or collaborative users (Collab).

focus on physical objects with the number smaller than 30 [10], e.g., PapARVis [11] (≤ 5), Situated Analytics [15] (≤ 5), MarVisT [10], etc. The large data scale of this paper poses a new challenge in image segmentation, object labeling, text information recognition and the XR-based data presentation.

Second, we mainly focus on the DR environment while most of the existing related tools focus on AR or even more close to VR [6, 12, 31, 41, 44]. DR can link the data computation in virtual space with the interaction in physical space and provide information re-organization to get a comprehensive and better target comparison.

Third, we focus on filtering, re-grouping, and re-ranking according to the extra attributes of numerous physical objects, instead of augmenting the existing static presentations like in PapARVis [11]. The personal tasks are different from the most related work due to the larger data scale of *DRCmpVis*.

3 Design Rationale

We illustrate the design goal, design considerations and design details of *DRCmpVis* in this section. Before the descriptions of design goals, we need to answer a question: *why do we need DR in daily lives to reorganize the additional information before decision making?* We take the library/bookstore case of this paper as an example. One scheme to show additional information about physical objects is to query them directly from the database of the library/bookstore. However, there are several limitations of this scheme due to the inconsistency between the physical ~~world~~ space and the virtual ~~data~~ 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 is inconsistent with the information in the 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, ~~common~~ 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 world space and the virtual data space while significantly reducing visual clutter*. Objects are data in the physical space of the DR environment. The physical objects (targets) can be replaced by their virtual avatars, 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 on the objects. Furthermore, it saves visualization space and helps to reduce visual clutter and operational ambiguity caused by showing the physical objects and their avatars simultaneously. Besides, DR is also capable of

building an information bridge between the changing physical space and the virtual space seamlessly. Regarding AR or SA (AR is one of its four primary elements as mentioned above), however, a new visualization space should be brought to the information presentation [17], then the contextual information can be provided around the physical targets.

3.1 Design Goals

We summarize ~~the following~~ 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 ~~across their additional attributes~~ by using simple visual comparisons in MR space.

3.2 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 page limit. ~~Furthermore, different scenarios are used to evaluate different tasks, as shown in Fig. 2.~~

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. ~~Specifically, we use flash to highlight the search results and further provide a MR-based fisheye deformation design to highlight 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~~according to their experience in our daily life~~.

Third, these applications should be designed to enable re-rank the disordered physical objects for visual comparison ~~according to~~in terms of one or multiple ordinal or quantitative attributes (G3). For example, books in a library are usually sorted by classification number or index number, which might not align with users' diverse sorting requirements. ~~, e.g., sorting~~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, 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 the name of a book in a book shelve, 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 ~~not~~neither shown directly nor 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.

3.3 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.

4 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, 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 identical calibrated coordinated system. We have developed and encapsulated the related functions into the APIs of the framework.
- **Challenge III: integrating comparative visualizations into 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, 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**.
- **Challenge V: enhancing the lighting environment in the reality world**. In the 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 synthesizing the captured images to restore the reflect regions.

For detailed information about the implementation, please refer to the Appendix file.

4.1 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 [4]

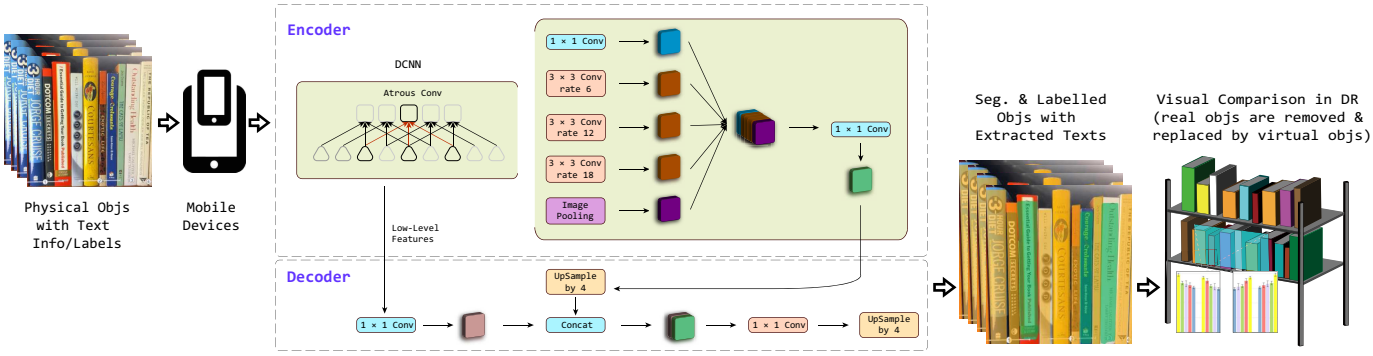


Fig. 2: The workflow of the proposed *DRCmpVis*. We illustrate it using one of the application cases, the library/bookstore case. Regarding the deep neural network used in image segmentations and text recognitions, the encoder module encodes multi-scale contextual information by applying atrous convolution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

or ARCore [3] 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 physical space. The distance can be measured by the camera with LiDAR scanner [4].

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. Actually, we initially use 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 ~~use of the~~ 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. ~~Therefore~~Finally, ~~in the current version, we adopted the method of automatic segmentation using neural networks to satisfy the needs of more scenarios.~~ we adopted deep neural networks to achieve automatic image segmentation & 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 [34] 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 [9] is one of its key modules. Therefore, we take DeepLab as an example to illustrate how PaddleSeg is integrated into *DRCmpVis*, as shown in Figure 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 panoramic image we captured or the real-time video we recorded is input into the first network (the top left of Figure 2), while the labeled samples are input into the second network (the top right of Figure 2). Regarding the text extraction, we use the traditional CNN-based optical character recognition approach, following a language adaptive design [33], to recognize a large amount of the 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 the 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. ~~, which makes~~Real-time tracking facilitates the positions of virtual objects to be updated accordingly.

In the implementation, we segment the captured images into multiple blocks by CNNs, ~~Then~~and then track the objects ~~especially for the~~in blocks by the image detection ~~algorithm~~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.

4.2 Database Construction of Augmented Information

We create a large database on the server for two application scenarios that require real-time information feedback [22, 37]. The database contains additional information on different attributes of the objects. In order to make the data updated periodically and improve the scalability of the framework, we design a data synchronizer with a pattern matching algorithm and regular expression matching algorithm, which can be used to download the open data automatically and fetch the data attributes to update them in the database.

(1) Global book database. More than two million books are 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 is downloaded from the open data website "Amazon product data" [22, 36, 37], containing product reviews and metadata from Amazon, including 142.8 million reviews for their products and 22.5 million reviews for books. It includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs). The Amazon database was last updated in 2018.

(2) Coffee database. The coffee database is 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, process introduction.

4.3 Integrating Visual Comparison Components into 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 design 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 *fisheye* algorithm, which scales the size of

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 presses are placed on different layers. The user browses those books with a fisheye effect. (d) Further search with a 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 word cloud of abstracts, introduction 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.

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.

5 Example Scenarios

To illustrate how *DRCmpVis* ~~makes~~facilitates visual comparisons for physical objects with text labels in DR environments and demonstrate ~~the approach is not ad-hoc~~the robustness of the proposed framework. ~~we show different example scenarios where users use *DRCmpVis* to search or filter objects to obtain their additional information, and locate the candidate targets, namely, the usage scenarios in a library/bookstore and a coffee shop.~~

5.1 Library/Bookstore Scenario

Suppose Zelda is a student majoring in economics. ~~who seeks to expand her knowledge by purchasing several books related to her field.~~ 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 Figure 3 (a).

(1) **Fuzzy filtering**: she scans the bookshelves by the panoramic camera of her tablet with *DRCmpVis* installed. There are 778 books that 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 Figure 3 (b). She finds that only one book nearby is from "University of Chicago Press", then she wants to find more books on "economic" and published by "University of Chicago Press".

(2) **Re-grouping**: she re-groups those 118 books by publisher and searches by saying "Chicago" or input 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 (Figure 3 (c)). Books from other presses 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 fuzzy keyword "social" , there are 21 books highlighted in red (Figure 3 (d)). She uses fisheye to view each book's details including titles or authors similar to Figure 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 either 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 by descending order (Figure 3 (e)). Then she selects four books that seem suitable, those selected books are moved to a reserved layer of the virtual bookshelf which are designed to place the candidate books (Figure 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" and "empire" (Figure 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 kinds of diagrams in small multiples**: so she compares both the ratings and the prices of these books via bar charts (Figure 3 (h)). She finds that the "story" (the first) rated as high as the "empire" one (the fourth), but is a little cheaper than the "empire" one. So she chooses the "story" one and restores those books to their original layout.

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

5.2 Cafe Scenario

To demonstrate the proposed framework can ~~be used in~~support different scenarios where the objects are labeled with texts or presented as texts, we show another example scenario in coffee shops in this section.

A new coffee shop opens on Zelda's campus. She doesn't 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 by *DRCmpVis*. Soon she scans 40 different drinks, and *DRCmpVis* recognizes them and shows them on a virtual menu in the DR/MR context.

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 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.

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 achieved by the image segmentation, image labeling, and text extraction using neural networks DeepLab [9] and PaddleSeg [34]. 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 firstly voice inputs "Latte" and finds that it's highlighted in the menu (Figure 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 surely bitter, as no sugar is added to it (Figure 4 (c)).

Zelda then re-groups those coffees according to sugar (Figure 4 (d-e)). She browses and selects several drinks with high ratings in the "medium sweet" and "sweet" group, as shown in Figure 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 Figure 4 (g). However, her fitness coach's advice crosses her mind that she needs to limit her calorie intake to 1300 calories every day, whereas the coffee summary shows that Cappuccino has 140 calories per cup. So she re-ranks all the drinks by calorie content. This time, coffees are sorted from left to right by calorie, as shown in Figure 4 (h). She begins browsing on the right side, where coffees with relatively low calories are located. She finds several coffees that she hasn't drunk. To have a quick grasp of them, she views their word cloud (Figure 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 (Figure 4 (j)). Finally, she chooses Blonde Roast and enjoys its "soft and mellow flavor" described in the summary. In addition, *DRCmpVis* can also handle larger menus like a big poster hanging on the wall outside the coffee shop, as shown in Figure 5.

6 Evaluation: User Study and Framework Performance

In the evaluation, we aimed to assess *DRCmpVis* regarding the following aspects: (a) whether visual searching/filtering of *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 have conducted four measures, including subjective measures and objective measures:

- **User Study:** a 5-point Likert scale was utilized to gauge and assess the comprehensive functionality of *DRCmpVis*.
- **NASA-TLX:** 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:** regarding 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 modules of *DRCmpVis*, including the modules of scanning, image segmentation & labelling, overall processing, etc.

6.1 Study Design

User Study 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. *T1* was performed in a library, while *T2* was performed in a coffee shop. 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 technology, 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 them had trouble finding books in libraries where books are sorted by traditional index numbers.

Regarding the coffee scenario, most of them also felt confused when choosing different coffees. It is difficult for them to distinguish different coffees according to their approximate ingredient information. Also, recalling whether a particular coffee variety includes milk, cream, and

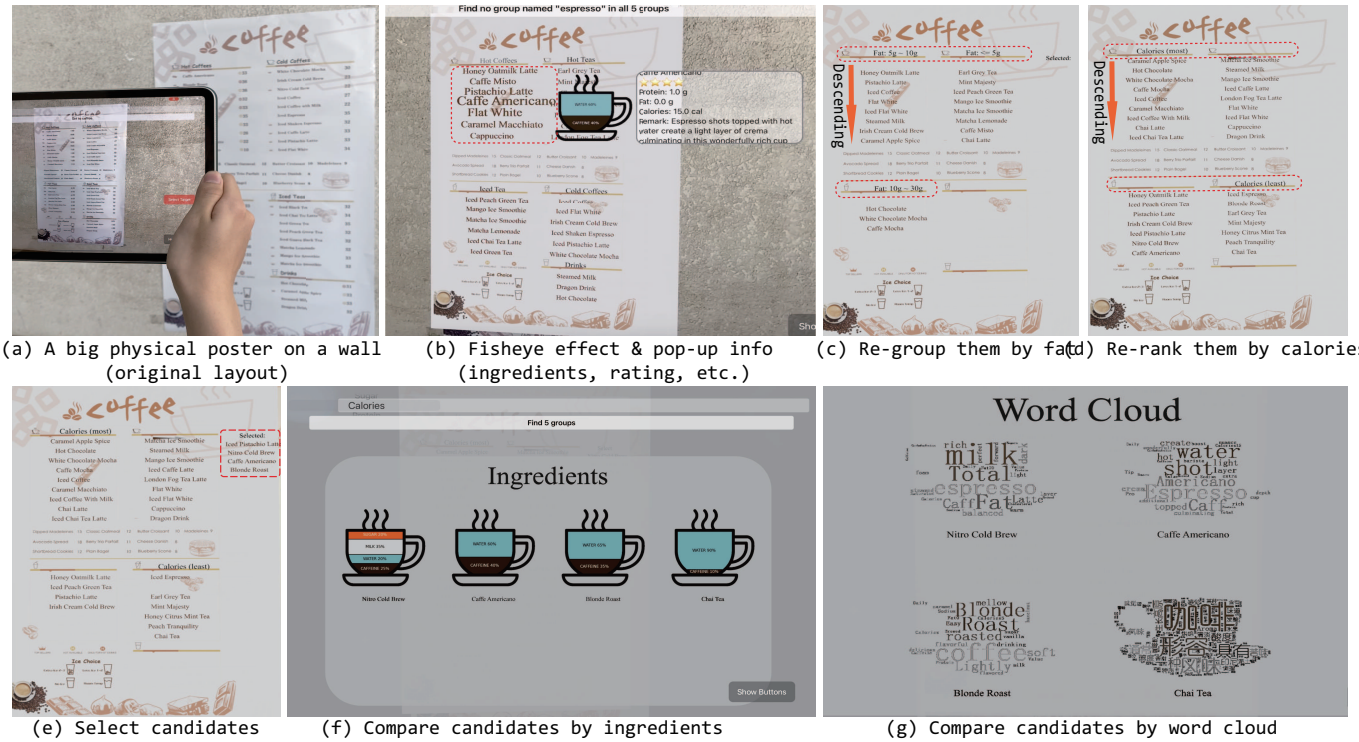


Fig. 5: Usage scenario outside a cafe: a user compares candidate coffees by augmented information from a big poster. (a) The original poster hanging on a wall outside a coffee shop. (b) View coffee's augmented information with a fisheye. (c) Re-group coffees by fat content intervals. (d) Re-rank coffees 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.

sugar, as well as comparing the caloric content of two distinct cups of coffee, proves to be arduous for them. 98.7% of participants do not agree that coffee shop staff will provide a retrieval system for you to use, while 86.7% of participants indicate that coffee shop staff will not provide specific ingredient information for comparison.

In the pre-study survey before the questionnaire step, we have one question to survey how many users like simple visualizations 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 just 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.

Then, the investigators introduced the capability and usage of *DRCmpVis*. In *T1* and *T2*, the investigators showed a simple example to the users first and then released the specific task. After all the tasks were finished, the participants were asked to complete the post-study questionnaires. All the participants got gifts of equal value regardless of their performance.

Free exploration. Participants are encouraged to explore *DRCmpVis* freely before the study. They can use search functions to filter the available books, regrouping them according to different attributes, such as the press or the range of publishing years. Additionally, participants have ~~the option~~options to utilize re-ranking techniques to facilitate their comparison of ratings and prices. Free exploration step is designed to help participants get familiar with the UI and the ~~interaction~~functions of *DRCmpVis*.

6.2 User Study Tasks

We use *Ti* to name the task that happens in the *i*-th scenario. The first task *T1* is about the library case, while the second task *T2* is about the cafe menu case.

T1 is divided into three subtasks. In *T1*, participants are required to locate four different books. In *T1-1*, participants 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 to 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 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 the proposed tool, *DRCmpVis*, to find the four books from the aforementioned tasks. The timing results are recorded in all of the tasks in *T1*. After completing *T1-3*, the participants are suggested to use re-grouping of *DRCmpVis* to find other books with the 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 are desired to read.

T2 requires participants to search for different types of lattes from the physical coffee menu. *T2* has only two tasks, because coffee shops do not provide users with a coffee retrieval system unless the manager or the waiters. In *T2-1*, participants search for lattes from the physical menu, while in *T2-2*, they can use *DRCmpVis* to highlight all the candidates that satisfied the task requirements. The timing results of *T2-1* and *T2-2* are also recorded in each task. After these two timing experiments, participants are required 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 Figure 4. After that, participants could also check the menu and select other coffees that they are unfamiliar with. They could compare them in the MR context using ingredient glyphs and word cloud, as shown in Figure 5 (f).

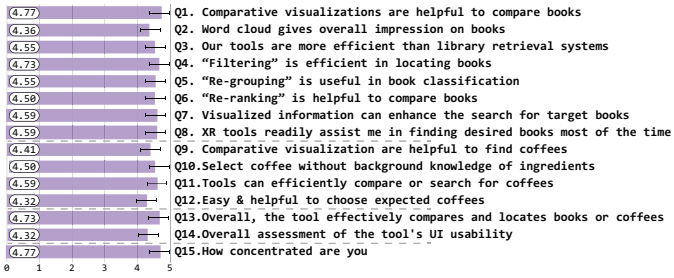


Fig. 6: Post-study result: most of participants react positively to *DRCmpVis*.

6.3 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 Figure 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 with the *DRCmpVis* (Q13: $\mu = 4.73$, 95% $CI = [4.53, 4.92]$ G1). In particular, UI design (Q14: $\mu = 4.32$, 95% $CI = [4.00, 4.64]$ G4), besides, the participants also appreciated the voice input, fisheye effect, and result highlighting. They said these designs make the interactions smooth and intuitive. From the questionnaire results, we can find that they can search targets and compare candidate targets by using the ~~virtuality-reality~~VR design and comparative data presentations, respectively.

Regarding the usability evaluation about the two scenarios, i.e., the library/bookstore scenario (Q3 ($\mu = 4.55$, 95% $CI = [4.32, 4.77]$ G1) and Q1 ($\mu = 4.77$, 95% $CI = [4.54, 5.01]$ G4)) and the cafe scenario (Q12 ($\mu = 4.31$, 95% $CI = [4.03, 4.60]$ G4)), the participants gave high praise, because they thought *DRCmpVis* is intuitive to use in 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 aid participants in developing a comprehensive understanding of physical objects. The participants also noted that the comparative ingredient glyphs significantly contribute to forming comprehensive impressions of the distinctions among various types of coffees and books.

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) of books in libraries.

Compared with blindly finding, the time cost is reduced from an average of 4.56 minutes to 0.65 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". In the cafe scenario, the time cost of finding eight lattes from the menu is reduced from 0.45 minutes to 0.13 minutes with the help of *DRCmpVis*.

In response to selecting coffee without background knowledge of ingredients (Q10: $\mu = 4.50$, 95% $CI = [4.20, 4.80]$ G1) aiming to efficiently compare or search for different coffees (Q11: $\mu = 4.59$, 95% $CI = [4.36, 4.81]$ G1), most participants found the subsequent visual comparisons helpful for them as they didn't 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 Q15 ($\mu = 4.77$, 95% $CI = [4.58, 4.96]$), almost all participants felt concentrated when carrying on the tasks. They all believed that the tasks were quite smooth and interesting.

6.4 NASA-TLX Measures

We further evaluate the proposed *DRCmpVis* by comparing it with two traditional methods as control groups based on NASA-TLX measurements, i.e., target blinding finding without any tools (Blinding finding), 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 are randomly from ten different majors of the university.

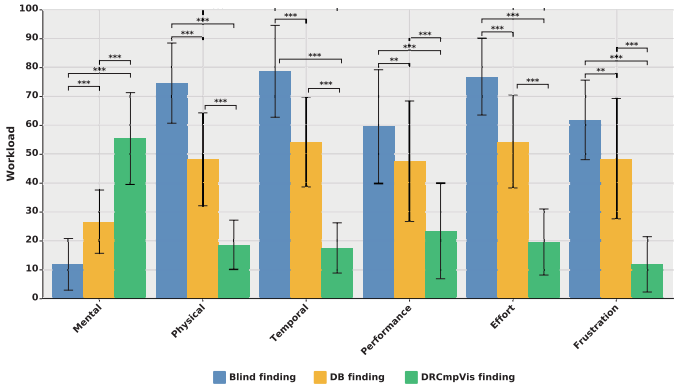


Fig. 7: The scores of NASA-TLX evaluation for two control groups of methods and the proposed *DRCmpVis*. Error bars indicate standard errors. Statistical significant differences are denoted by ** (p <0.01) , *** (p <0.001).

A repeated measures 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 Figure 7. It is worth mentioning that the mental demand of blind finding is significantly lower than the other two methods, because the blind finding is the simplest approach which just has the smoothest learning curve. It requires some learning to master DB-based searching tool and *DRCmpVis*. 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 Table 2 and Table 3. 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", $p = 2.7E-17$).

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 is 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. Besides, some participants suggested that it would be better if we add visual cues about the physical directions of the target book when they search 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*, is convenient, efficient, and relatively easy to learn, requiring less effort to locate target objects.

Scenario	Scanning Time	Segmentation Time	Processing Time	Seg. & Labelling Rate
Library Scenario	0.321	0.631	0.952	95.13%
Cafe Scenario	0.454	3.891	4.345	100.00%

Table 2: Performance and accuracy evaluation of *DRCmpVis* (seconds). The "Scanning Time" is the average time to scan a panoramic photo 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 each back-end server request, while the "Seg. & Labeling Rate" is the average accuracy. We obtain the average results based on 16 tests.

Methods	Book Searching Time	Latte Searching Time	Augmented Info	Target Comparison
Blind finding (without any tools)	4.56	0.45	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*	0.65	0.13	Color highlight Fisheye highlight Pop-up glyphs	Re-grouping Re-ranking Visual comparison

Table 3: Task-driven quantitative evaluation results (in minutes). We compare the proposed *DRCmpVis* with the traditional two methods. The results shows 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 taken by all participants to search for the keyword "Latte". Note: most coffee shops do not provide a retrieval system for users, thus they are marked as Not Available (NA).

6.5 Quantitative Evaluation

For the sake of achieving a dependable and consistent server service, we choose 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 is 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 & labeling, which is provided by a trained CNN platform named PaddleSeg [34]. In our experiments, both the PaddleSeg and the database with augmented information are built on the cloud server. We can find that the average image segmentation & labeling rates of *DRCmpVis* for all the ~~three~~example scenarios are larger than 95.0% (the additional example scenarios are moved to the Appendix due to page limit of the paper). The quantitative evaluation results are shown in Table 2.

Furthermore, we have also conducted some quantitative and qualitative tests for the two tasks without/with *DRCmpVis*. There are 22 participants involved in the tasks. The task is to ask the participants to find all types of "Latte" from the physical coffee menu, as shown in Figure 4 (a). The results searched by *DRCmpVis* are shown in Figure 4 (b). There are eight different lattes in total. The test results are shown in Table 3. We can find the task *T1-3* finished by *DRCmpVis* takes about 0.65 minutes on average to find the target book from 1238 books, which is only 14.3% of the search time used in the blind finding method (without any tools). Similarly, the task *T2-2* finished by *DRCmpVis* takes about 0.13 minutes on an average to find all lattes, which is 28.7% of the searching time in the coffee shop without *DRCmpVis*. By comparing the proportion, it can be seen that the more target objects searched, the greater the advantage of *DRCmpVis*.

We also summarize some feature comparisons in Table 3. ~~All the qualitative comparison results have been evaluated in the case study (see Section 5) and user study of the paper.~~ 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.

7 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. Besides, we find *DRCmpVis* can be easily extended to the 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. 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 they are in some crowded setting. In some cases, libraries will be influenced by the crowded environment, but in other cases, such as the cafe menu case, are irrelevant.

However, *DRCmpVis* is not feasible for the libraries or bookstores when the book information is unavailable to fetch, or it is hard to download or crawl from Internet, e.g., the ancient book libraries, etc., because the framework will query additional augmented information from the constructed database according to the information scanned from the physical objects.

Scalability issue on image segmentation and image labeling. It is worth noting that the image segmentation components of *DRCmpVis* are *scalable* and not limited by the object number, because the CNN and the OCR algorithm are run on the server which can even handle thousands of books in the library scenario in our experiments. More importantly, unlike the mobile device, the computation resources of the server are *scalable* enough and could be easily upgraded. As a result, whereas *DRCmpVis* recognizes almost all 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.

Why do we mainly use DR/MR instead of VR, or why not use a fully database-based VR as an alternative design? First, information should be updated periodically between the virtual world and the physical world. The object data in the virtual space should be consistent with that in the physical world in *DRCmpVis*. Because in the library/bookstore case, the book positions would be often changed due to the previews by buyers/borrowers, the books are also often inserted into the wrong positions or even wrong bookshelves by buyers/borrowers. In the cafe scenario, the menus are also often moved in a coffee shop (as shown in the supplemental video). All such scenarios need to involve the real-time physical world information into *DRCmpVis*, which makes *DRCmpVis* should include DR/MR instead of VR. In *DRCmpVis*, actually, the image recognition module on the front-end mobile device will initially and periodically detect whether the object information in the physical world is changed. If yes, all the changed positions of the objects will be updated by the deep network deployed on the cloud server.

Second, users often need to go back to the reality to "highlight" the targets after the searching or the comparing steps to help users find them. For example, the target books/the target coffee items will be highlighted in the real background after users' searching or visual comparisons (as shown in the supplemental video). The in-context highlighting in reality requires DR/MR instead of VR.

Third, it is impossible or time-consuming for bookstore salesmen/librarians to update the database of the books' new positions immediately, if we choose VR instead of DR/MR.

The limitation of text recognition. *DRCmpVis* recognizes objects by images taken from mobile devices. Ideally, the user only needs to take one panoramic picture that contains all the objects. However, objects' details may not be recognized if they are too small in the picture, that is the user is standing too far away from the numerous objects. For example, in the library/bookstore scenario, instead of scanning all layers of the bookshelves, the user may walk closer to the bookshelves and scan one layer at one time by panoramic stitching due to lack of light or limited imaging quality.

The image segmentation & labeling service needs to request once due to an image recognition module on the client app of *DRCmpVis*, when the positions of the objects are not changed. Because the coffee menu in a cafe is often unchanged. Actually, we 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. In our strategy, the latest captured panoramic image will be saved to the buffer of the client app. The image recognition module will verify whether the newly captured panoramic image is saved in the buffer. If yes, the segmentation & labeling records in the buffer can be reused without requesting the server twice. This strategy is *useful and efficient* in almost all the usage scenarios due to the quick response by the client app. 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 is 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 6. The hardware configuration can be improved for more expensive service packages. Thus maybe the performance especially for the segmentation & labeling could be further improved. We plan to make *DRCmpVis* to be applied in more general usage scenarios in our daily lives. In the future, we plan to extend the usage scenario of *DRCmpVis* to others like choosing cups, fruits or flowers, and other more general scenarios in our daily life. Because objects with text on them or in different colors and shapes can be well recognized by trained neural networks. However, objects with different irregular 3D shapes and without textual information on them are difficult to be recognized by current algorithms including the-state-of-the-art neural networks.

8 Conclusion

In this paper, we propose a novel DR/MR-based application framework named *DRCmpVis*, which is designed to build visual comparisons towards multiple physical objects with text labels or text information. The efficient data computation in virtual space is linked with the in-context interaction in physical space in the framework. The framework can provide multidimensional comparisons for candidate objects, exploiting all their three types of attributes, i.e., nominal, ordinal, or quantitative attributes. Users first take panoramic photos from the real world by the cameras of mobile devices. They can input a fuzzy searching keyword in objects' nominal attributes by voice or text (according to users' environment) to narrow down the number of candidate targets. The ~~searched~~search results will be highlighted by color and deformation in the DR environment to indicate their positions in the reality; Furthermore, users possess the capability to regroup or re-rank candidates based on their multifaceted attributes. Additional comparative augmented information of the objects can be integrated in an identical MR context.

References

[1] AFrame. aframe. `https://aframe.io/`.

[2] D. Amin and S. Govilkar. Comparative study of augmented reality sdks. *International Journal on Computational Science & Applications*, 5(1):11–26, 2015.

[3] ArCore. Arcore. `https://developers.google.cn/ar/`.

[4] ARKit. Arkit. `https://developer.apple.com/cn/augmented-reality/arkit/`.

[5] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, 1967.

[6] P. W. S. Butcher, N. W. John, and P. D. Ritsos. Vria: A web-based framework for creating immersive analytics experiences. *IEEE Transactions on Visualization and Computer Graphics*, 27(7):3213–3225, 2021.

[7] S. Butscher, S. Hubenschmid, J. Müller, J. Fuchs, and H. Reiterer. Clusters, trends, and outliers. In *CHI Conference on Human Factors in Computing Systems*. ACM, apr 2018.

[8] C. C. Carrera, J. L. S. Perez, and J. de la Torre Cantero. Teaching with ar as a tool for relief visualization: usability and motivation study. *International Research in Geographical and Environmental Education*, 27(1):69–84, 2018.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.

[10] Z. Chen, Y. Su, Y. Wang, Q. Wang, H. Qu, and Y. Wu. Marvist: Authoring glyph-based visualization in mobile augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 26(8):2645–2658, 2020.

[11] Z. Chen, W. Tong, Q. Wang, B. Bach, and H. Qu. Augmenting static visualizations with paparvis designer. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2020.

[12] M. Cordeil, A. Cunningham, B. Bach, C. Hurter, B. H. Thomas, K. Marriott, and T. Dwyer. IATK: An immersive analytics toolkit. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 200–209. IEEE, 2019.

[13] D. Cushnan and H. E. Habbak. *Developing ar games for ios and android*. Packt Publishing Ltd, 2013.

[14] Donghao Ren, Tobias Höllerer, and Xiaoru Yuan. ivisdesigner: Expressive interactive design of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2092–2101, 2014.

[15] N. ElSayed, B. Thomas, K. Marriott, J. Piantadosi, and R. Smith. *Situated analytics*. 2015.

[16] 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 *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pp. 203–208. Merida, Mexico, may 2016.

[17] N. A. M. ElSayed, B. H. Thomas, R. T. Smith, and K. Marriott. Using augmented reality to support situated analytics. In *IEEE Virtual Reality*. Arles, France, Mar. 2015.

[18] G. Evans, J. Miller, M. I. Pena, A. MacAllister, and E. Winer. Evaluating the microsoft hololens through an augmented reality assembly application. In *Degraded Environments: Sensing, Processing, and Display 2017*, vol. 10197, p. 101970V. International Society for Optics and Photonics, 2017.

[19] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[20] Gun Lee, Andreas Dünser, Seungwon Kim, and Mark Billinghurst. Cityviewar: A mobile outdoor ar application for city visualization. In *2012 IEEE International Symposium on Mixed and Augmented Reality - Arts, Media, and Humanities (ISMAR-AMH)*, pp. 57–64, 2012.

[21] S. Hashiguchi, S. Mori, M. Tanaka, F. Shibata, and A. Kimura. Perceived weight of a rod under augmented and diminished reality visual effects. In *The 24th ACM Symposium on Virtual Reality Software and Technology*, VRST '18, pp. 1–6. ACM, New York, NY, USA, 2018.

[22] R. He and J. McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *International Conference on World Wide Web*, pp. 507–517, 02 2016.

[23] J. Herling and W. Broll. PixMix: A real-time approach to high-quality diminished reality. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 141–150. IEEE, Nov 2012.

[24] D. Herr, J. Reinhardt, R. Krüger, G. Reina, and T. Ertl. Immersive visual analytics for modular factory layout planning. In *Proc. IEEE VIS Workshop Immersive Analytics*, 2017.

[25] K. Hirokazu. Artoolkit: Library for vision-based augmented reality. *Technical Report of Ieice Prmu*, 101:79–86, 2002.

[26] D. Kalkofen, E. Mendez, and D. Schmalstieg. Interactive focus and context visualization for augmented reality. In *IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 191–201, 2007. doi: 10.1109/ISMAR.2007.4538846

[27] D. Kalkofen, C. Sandor, S. White, and D. Schmalstieg. Visualization techniques for augmented reality. In *Handbook of augmented reality*, pp. 65–98. Springer, 2011.

[28] N. Kawai, T. Sato, and N. Yokoya. Diminished reality based on image inpainting considering background geometry. *IEEE Transactions on Visualization and Computer Graphics*, 22(3):1236–1247, March 2016.

[29] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik. Intel realsense stereoscopic depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–10, 2017.

[30] B. Lee, D. Brown, B. Lee, C. Hurter, S. Drucker, and T. Dwyer. Data visceralization: Enabling deeper understanding of data using virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1095–1105, feb 2021.

[31] 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 Transactions on Visualization and Computer Graphics*, 27(2):1171–1181, feb 2021.

[32] 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 *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 11–19. IEEE, Oct 2013.

[33] O. Y. Ling, L. B. Theng, A. Chai, and C. McCarthy. A model for automatic recognition of vertical texts in natural scene images. In *2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, pp. 170–175, 2018.

[34] Y. Liu, L. Chu, G. Chen, Z. Wu, Z. Chen, B. Lai, and Y. Hao. Paddleseg, end-to-end image segmentation kit based on paddlepaddle. `https://github.com/PaddlePaddle/PaddleSeg`, 2019.

[35] Marder-Eppstein and Eitan. Project tango. In *ACM SIGGRAPH 2016 Real-Time Live!*, pp. 25–25. 2016.

[36] J. McAuley. Amazon product data. https://jmcauley.ucsd.edu/data/amazon/, 2018.

[37] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel. Image-based recommendations on styles and substitutes. *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52, 2015.

[38] P. Milgram and F. Kishino. A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems*, vol. E77-D(12):1321–1329, 12 1994.

[39] S. Mori, S. Ikeda, and H. Saito. A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. *IPSJ Transactions on Computer Vision and Applications*, 9(17):1–14, 2017.

[40] G. Queguiner, M. Fradet, and M. Rouhani. Towards mobile diminished reality. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct*, pp. 226–231. IEEE, Munich, Germany, 2018.

[41] T. Rhee, S. Thompson, D. Medeiros, R. dos Anjos, and A. Chalmers. Augmented virtual teleportation for high-fidelity telecollaboration. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1923–1933, 2020.

[42] S. H. Said, M. Tamaazousti, and A. Bartoli. Image-based models for specularity propagation in diminished reality. *IEEE Transactions on Visualization and Computer Graphics*, 24(7):2140–2152, July 2018.

[43] B. Shneiderman, D. Feldman, A. Rose, and X. F. Grau. Visualizing digital library search results with categorical and hierarchical axes. In *Proceedings of the fifth ACM conference on Digital libraries*, pp. 57–66.

[44] R. Sicat, J. Li, J. Choi, M. Cordeil, W.-K. Jeong, B. Bach, and H. Pfister. Dxr: A toolkit for building immersive data visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):715–725, 2019.

[45] C. Stolte, D. Tang, and P. Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52—65, jan 2002.

[46] M. Takemura and Y. Ohta. Diminishing head-mounted display for shared mixed reality. In *International Symposium on Mixed and Augmented Reality*, pp. 1–8. IEEE, Darmstadt, Germany, 2003.

[47] A. Thudt, U. Hinrichs, and S. Carpendale. The bohemian bookshelf: supporting serendipitous book discoveries through information visualization. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1461–1470, 2012.

[48] Vladimir Geroimenko. Augmented reality technology and art: The analysis and visualization of evolving conceptual models. In *2012 16th International Conference on Information Visualisation*, pp. 445–453, 2012.

[49] D. Wagner and D. Schmalstieg. Artoolkit on the pocketpc platform. In *2003 IEEE International Augmented Reality Toolkit Workshop*, pp. 14–15, 2003.

[50] WebVR. Webvr. `https://webvr.info/`.

[51] W. Willett, Y. Jansen, and P. Dragicevic. Embedded data representations. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):461–470, jan 2017.

[52] WIMP. Wimp. `https://www.interaction-design.org/literature/book/the-glossary-of-human-computer-interaction/wimp`.

[53] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):649–658, 2016.

[54] 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 Transactions on Visualization and Computer Graphics*, 27(9):3808–3825, 2021.

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

Abstract— We illustrate the implementation of *DRCmpVis* in Section 1, and then showcase the third usage scenario (image recognition) and the fourth (Chinese recognition) to demonstrate the scalability of *DRCmpVis* in Section 2.

Index Terms—Diminished reality, visual comparison, virtual avatars, mixed reality

1 Implementation

There are some technical challenges that we have addressed in *DRCmpVis*:

- **Challenge I: building the application framework.** Image segmentation, image labelling, OCR-based text extraction, image recognition are the significant modules of the framework. We have integrated several 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 identical calibrated coordinated system. We have developed and encapsulated the related functions into the APIs of the framework. For more details, please refer to Sect. 1.3.
- **Challenge III: integrating comparative visualizations into DR/MR context**. We have integrated some commonly-used visualization components/techniques into the framework, e.g., bar charts, line charts, word clouds, 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, whether they are simple or advanced. All the related functions are encapsulated into the APIs of the framework. See details in Sect. 1.5.
- **Challenge IV: database construction of augmented information of target objects**. See details in Sect. 1.4.
- **Challenge V: enhancing the lighting environment in the reality world**. In the practical applications, it is important to reduce the interference of reflect light, illumination compensation when the lighting is weak. The solution to the former issue is to achieve multiple frames with time interval of 0.5 seconds when the camera is scanning, then synthesizing the captured images to restore the reflect regions. The solution to the latter one is to integrate the corresponding image processing algorithms into the system.

1.1 The Front-end Development Platform

We have encapsulated the device-dependent APIs of DR/MR for different mobile devices, to make the implementation more scalable. For example, either ARKit [2] or ARCore [1] is employed to encapsulate the APIs for different mobile device platforms. The device-dependent APIs can be summarized as follows:

- **Device positioning**: ARKit/ARCore provides the APIs for achieving the real-time position M of the mobile device in the physical space.
- **The distance between the target and the device**: ARKit provides the APIs for getting the real-time distances d between the mobile device, as shown in Fig. 2 (a). The position M of the device and the distance d can be used to build a coordinates system in the physical space. The distance can be measured by the camera with LiDAR scanner [2].
- **Target positioning**: ARKit provides the APIs for achieving the real-time position of a target in the physical space, if it did appear in the captured image.

In short, we use two types of APIs about positioning in the physical space, including device positioning and target positioning.

1.2 Image Segmentation and Optical Character Recognition

We use image segmentation deployed on the server to recognize targets that contains in the images sent from the mobile devices. The segmented target image is labelled and sent back to the mobile devices, which can be used in the target visualizations in DR.

Actually, we initially use breadth first search (BFS) algorithm to finish image segmentation and recognition. However, the BFS algorithm only based on RGB value, it shows high constraints in the actual use on the scenario, 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. Therefore, in the current version, we adopted the method of automatic segmentation of neural network to satisfy the needs of more scenarios.

To get a better result for various scenarios in real applications [9], we apply a trained CNN-based open-source platform named PaddleSeg [6] 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 (e.g., person, dog, cat and so on) to every pixel in the input image. In PaddleSeg, DeepLab [3] 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. 1. 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 panoramic image we captured or the real-time video we recorded is input into the first network (the top left of Fig. 1), while the labelled samples are input into the second network (the top right of Fig. 1).

Regarding textual information recognition, we use the traditional CNN-based optical character recognition approach, following a language adaptive design [5], to recognize a large amount of the text characters over numerous targets in reality. In our experiments, there are two server deployment methods that we can choose, the first one is

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

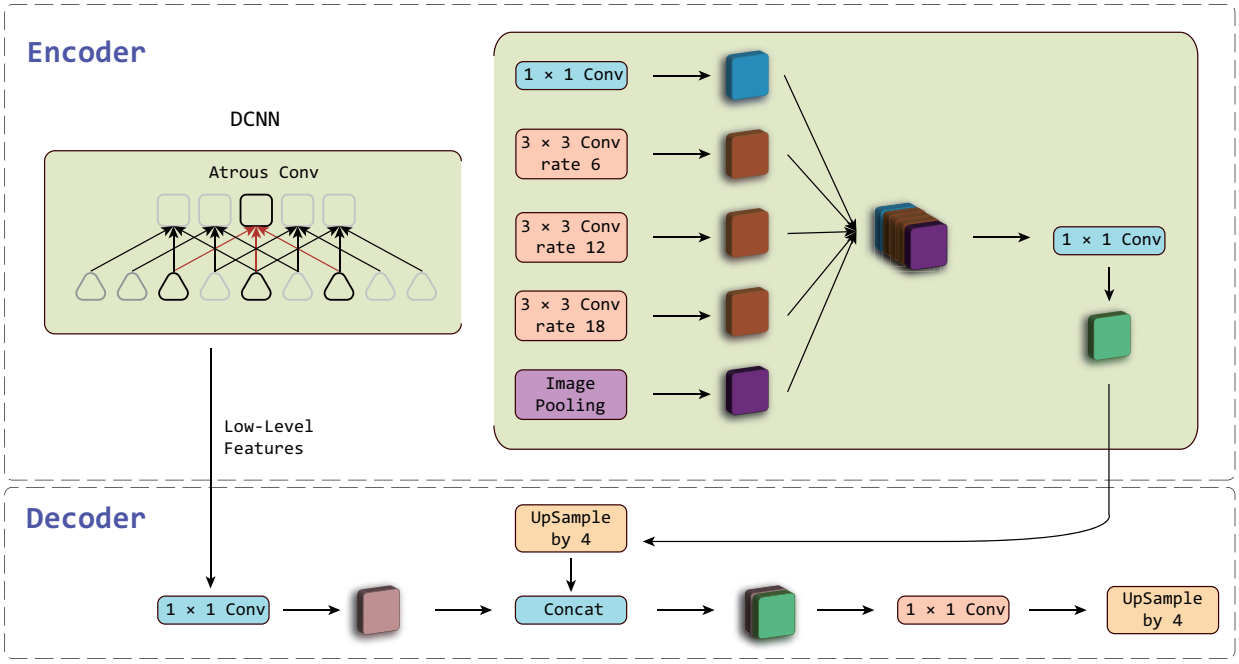


Fig. 1. The network illustration about how PaddleSeg [6] is integrated into *DRCmpVis*. We take DeepLab as an example, one of the key modules of PaddleSeg. The encoder module encodes multi-scale contextual information by applying atrous convolution at multiple scales, whereas the simple yet effective decoder module refines the segmentation results along object boundaries.

the non-free cloud service, the second one is to build the open source CNN platforms on our local machine.

1.3 Coordinate Transformation between Physical Space and Virtual Space

To guarantee that the positions of the targets in virtual space are consistent with those in their physical space, we design a coordinate transformation method to transform the 2D image coordinates into 3D physical coordinates. The image is segmented and labelled by the above-mentioned PaddleSeg [6].

On the mobile device (client), images taken by users are sent to the server for recognition. The server sends back JSON data indicating the 2D coordinates of targets in each image. If given an image with resolution of 3840×1880,, a book is recognized at (-640, 1280) with width 192 pixels and height 1344 pixels, as shown in Fig. 2 (a) and Fig. 2 (b). The position of the book in the physical space can be calculated by the coordinate transformation matrix between physical space and virtual space, which can be calculated as follows.

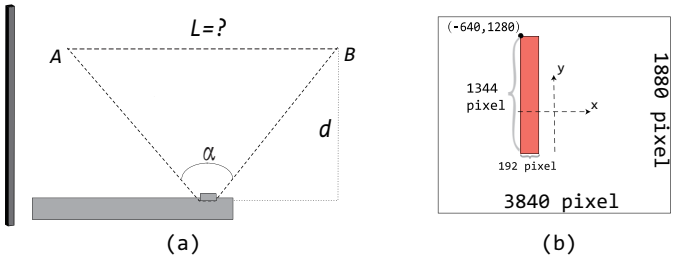


Fig. 2. The illustration of coordinate transformations. (b) The top view shows the scanning process with a field-of-view angle α, where d is the distance between the mobile device and the target books calculated by *DRCmpVis*.

Calculate the physical distance from the leftmost target object can be seen in the image to the rightmost one. We can achieve several variables by device APIs: camera view angle α, target distance d, and the total pixels t_p (3840 in this case) in the image space, as shown in Fig. 2 (a).

Assume L is the physical distance from the leftmost target object can be seen in the image to the rightmost one, where L can be calculated by the formula:

$$L = 2d \times \tan(\frac{\alpha}{2})$$

It is the distance between A and B in Fig. 2 (a). The physical distance unit l of within a single pixel can be calculated by:

$$l = \frac{L}{t_p} = \frac{2d}{t_p} \times \tan(\frac{\alpha}{2})$$

After that, we can achieve physical distance vector of the coordinate transformation matrix, which measures the offset from the coordinates center to the top-left corner of a target object, which is segmented and labelled by CNNs.

Real-time position update. We notice that hundreds of targets would be involved in the scenarios like library, bookstore or cosmetic store, which makes it difficulty in updating all the positions before rendering each frame. In the coordinate transformation algorithm, we track the positions of the target objects in real-time, because the target objects may be moved in the physical space. For example, the coffee menus and the eye-shadows would probably be moved in a cafe or a cosmetics shop, respectively.

We segment the captured image into multiple blocks by the CNNs and get some principal representative blocks. Then track the menu especially for the blocks by image detection algorithm provided by APIs of ARKit. The new information (e.g., the texts) rendered in the DR space will be transformed by projective geometry to achieve perspective effects. The real-time tracking animation of coffee menu can be seen in the supplementary video of the submission.

1.4 Database Construction of Augmented Information

We create a large database on the server for three application scenarios that require real-time information feedback [4, 8]. The database contains extra information of different attributes. The mobile device can

(a) Scan various eyeshadows in the original physical space

(b) Re-group by eyetypes & view them by fisheye

(c) 3-D virtual makeup try-on

(d) Re-group by scheme numbers

(e) Search by scheme number & highlight in the original layout

Fig. 3. Appendix of the 3rd case: an eye-shadow scenario. This scenario extracts image/texture information from the target objects, it shows a user compares eye-shadows using the framework of *DRCmpVis*. (a) Scan various eye-shadows displayed on a cosmetic table. (b) Re-group eye-shadows by eyetypes and view the augmented information of the focus one as well as an eye image showing places to apply it on. (c) View the effects of candidate eye-shadows via 3D virtual makeup try-on. (d) Re-group eye-shadows by scheme numbers. Different scheme numbers have different features like "Deep Blue" or "Soft Smokey". (e) Choose "Scheme 13" in their original physical layout, eyeshaodws belongs to this scheme number are highlighted.

access the database and provides visual presentations for augmented information in real-time.

In order to make the data updated periodically and improve the scalability of the tool, we implement a data synchronizer with pattern matching algorithm and regular expression matching algorithm, which can be used to download the open data automatically and fetch the data attributes to update them in the database.

Global Book Database More than two million books are 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 is downloaded from the open data website "Amazon product data" [4, 7, 8], containing product reviews and metadata from Amazon, including 142.8 million reviews from May 1996 to July 2014. It includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

Coffee/Eye-shadow/Shaxian County cuisine Database These two databases are created by the Web crawler, which crawls collections from well-known coffee and cosmetics brand websites. For example, coffee data comes from Starbucks, including the coffee's name, description, ingredients list, preview image, and process introduction, while eye-shadow data mainly comes from the website of Dior, including eye-shadow color, eye shape, location, usage steps, tips, and recommendations. The Shaxian County cuisine data from the official website of Shaxian County cuisine, including the price, taste, calories, and user reviews.

1.5 Choosing and Integrating Comparative Visualizations into DR Context

Regarding the visualizations for augmented information, the related data is sent to the server and the client receives the processed data from the server. We design several visual presentation 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 visualizations, which are appreciated by the participants in the user study. Besides, we adopt a *focus+context* exploration scheme by using *fisheye* algorithm, which scaling 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.

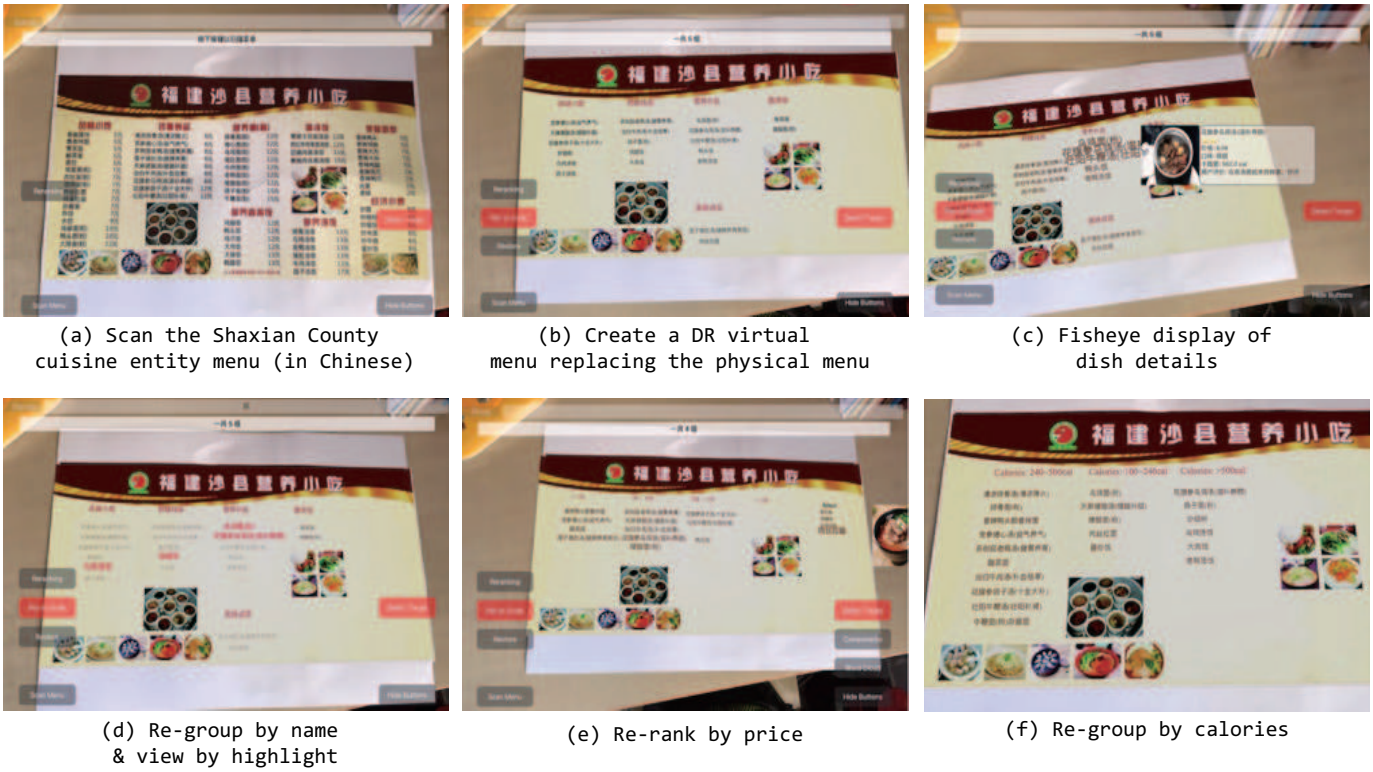
We create a virtual translucent screen in the DR environment to show those augmented information. Specifically, we create a virtual head model in the eye-shadow scenario to get a better makeup effect, stylizing the model's eyes with the selected eye-shadow color to show the 3D preview.

2 Appendix of Case Study

Except the textual information can be extracted from the target objects like books or menus, the framework is found to be applicable to scenarios involving image or texture information and is equally suitable for use in Chinese menu situations. For example, the textures/colors of different eye-shadows, quick comparison of ingredient content between different dishes on the menu (in Chinese).

2.1 The 3rd Case: Eye-shadow Scenario

One day Zelda is shopping in a cosmetics shop. She is not good at making up, especially eye makeup, because different eye-shadows may have unique effects, and sometimes several eye-shadows may be applied to different places to form a color combination. So she uses *DRCmpVis* and scans those eye-shadows displayed on the desk, as shown in Fig. 3 (a). Soon 15 different eye-shadows with 96 different colors (or textures) are recognized. She then re-groups them by eyetypes, and uses fisheye



(a) Scan the Shaxian County cuisine entity menu (in Chinese)

(b) Create a DR virtual menu replacing the physical menu

(c) Fisheye display of dish details

(d) Re-group by name & view by highlight

(e) Re-rank by price

(f) Re-group by calories

Fig. 4. Appendix of the 4th case: a restaurant scenario (Shaxian County cuisine, in Chinese). In this scenario, we extend the usage scenario to Chinese texts recognition. The framework is independent on language environment, because the deep network on the server supports cross-language. It enables compare more additional information about different Chinese dishes. (a) Scan the Shaxian County cuisine entity menu (in Chinese). (b) Create an DR virtual menu replacing the physical menu. (c) Fisheye visualization shows dish details (price, taste, reviews, ingredients, etc). (d) Re-group dishes by name keywords. Highlight the dishes with the keyword Chicken in the virtual menu. (e) Re-rank dishes based on different price ranges after re-grouping. (f) Re-group dishes by calories.

effect to view the details of the "protruding eye" group, as shown in Fig. 3 (b). A graph pops up on the side of the selected eye-shadow, showing the ideal position for users to apply it on. Zelda chooses a kind of golden brown eye-shadow, and previews it on a virtual 3D facial model, as shown in Fig. 3 (c). Zelda still finds it hard to choose several eye-shadows that matches each other, so she re-groups them by high rated schemes, this time, three recommended color schemes are lined up in front of her, as shown in Fig. 3 (d). Zelda views the details about each eye-shadow's effects and features, learns that eye-shadows in "Scheme 13" is suitable for simple day look. So she restores those colors to their original layout and search for "Scheme 13" using voice input or text input. Those eye-shadows in "Scheme 13" are flashing in red, as shown in Fig. 3 (e). As a result, she chooses an eye-shadow palette that contains several colors in "Scheme 13".

2.2 The 4th Case: Restaurant Scenario (Shaxian County Cuisine)

Zelda hears about a popular Shaxian County cuisine restaurant that has opened at her school and wants to try it out. She is eager to explore different dishes, but her knowledge of these dishes is limited, and she knows that the flavors can vary significantly. Simply looking at the menu doesn't provide her with enough information. Therefore, she decides to use *DRCmpVis* to scan the Chinese menu on her table, as shown in Fig. 4 (a). The *DRCmpVis* quickly recognizes 65 different dishes and creates an AR virtual menu overlaying the physical menu, as shown in Fig. 4 (b). Zelda clicks on "Pan to scal" to use a fisheye effect to examine the details of the various dishes, as depicted in Fig. 4 (c). She has a preference for chicken-based dishes, so she initially selects "Name" and inputs the keyword "Chicken" either through text or voice. *DRCmpVis* then filtered the groups, highlighting the dishes containing the keyword "Chicken" as shown in Fig. 4 (d). Zelda is particularly concerned about the prices and calorie content of the different dishes. Therefore, she chooses the "Price" option first and reorders the selected dishes by price. After clicking the "Reranking" again, the dishes are sorted from left to right in ascending order of price, as illustrated in Fig. 4 (e). Zelda examines the prices of the filtered dishes one by one by clicking the"Pan to scale", which displays detailed information, such as specific prices, portion sizes, and reviews using the fisheye effect. Finally, she selects "Calorie" and groups the dishes into three categories based on their calorie content, as shown in Fig. 4 (f). She continues to use the fisheye effect to explore additional information about specific types of dishes within each category. In the end, Zelda chooses her favorite dishes, and they live up to the descriptions and reviews she has read.

2.3 How Do Visualization Researchers Create Their Own Applications

How visualization researchers can leverage this framework to create and develop their own applications? We have summarized the steps as follows.

- **Database construction of augmented information**. The augmented information can be added by visualization researchers. For example, the book dataset is downloaded from the open data website "Amazon product data" [4, 7, 8], containing product reviews and metadata from Amazon, including 142.8 million reviews for their products and 22.5 million reviews for books. It includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs). The coffee database is created by the Web crawler, which crawled collections from well-known coffee websites. For example, coffee data

comes from Starbucks, including the coffees name, description, ingredients list, preview image, process introduction.

- **Coordinate transformation between physical space and virtual space**. This step is easy and simple to be done by visualization researchers, because it is mainly achieved by the camera with LiDAR scanner, we have encapsulated the related functions into the APIs of the framework.

- **Choose or design new visual comparison components into DR/MR context**. We have designed several commonly used visual presentation 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 visualizations, which are appreciated by the participants in the user study. Besides, we adopt a Focus+Context exploration scheme by using fisheye algorithm, which scaling 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.

REFERENCES

[1] ArCore. Arcore. `https://developers.google.cn/ar/`.

[2] ARKit. Arkit. `https://developer.apple.com/cn/augmented-reality/arkit/`.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.

[4] R. He and J. McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *International Conference on World Wide Web*, pp. 507–517, 02 2016.

[5] O. Y. Ling, L. B. Theng, A. Chai, and C. McCarthy. A model for automatic recognition of vertical texts in natural scene images. In *2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, pp. 170–175, 2018.

[6] Y. Liu, L. Chu, G. Chen, Z. Wu, Z. Chen, B. Lai, and Y. Hao. Paddleseg, end-to-end image segmentation kit based on paddlepaddle. `https://github.com/PaddlePaddle/PaddleSeg`, 2019.

[7] J. McAuley. Amazon product data. https://jmcauley.ucsd.edu/data/amazon/, 2018.

[8] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel. Image-based recommendations on styles and substitutes. *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52, 2015.

[9] Q. Zhu, B. Du, and P. Yan. Boundary-weighted domain adaptive neural network for prostate MR Image segmentation. *IEEE Transactions on Medical Imaging*, 39(3):753–763, 2020.