2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) | 979-8-3315-0691-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISMAR-Adjunct64951.2024.00018

A Way Out of the Replication Crisis in Diminished Reality Research

Shohei Mori*
Graz University of Technology

Dieter Schmalstieg†
University of Stuttgart

ABSTRACT

In this paper, we characterize the replication crisis observed in the current generation of diminished reality research and present a conceptual solution using a server-client system to frame existing diminished reality systems and their evaluations.

Index Terms: Replication, diminished reality, modular framework, dataset.

1 INTRODUCTION

Diminished reality (DR) is concerned with computationally removing physical objects from a given view [10]. A DR system is an interactive vision and graphics system that uses live inputs, such as video streams and ad hoc user inputs to identify the object of interest. We observe that researchers in this field have rarely open-sourced their solutions, with few exceptions [4, 2]. We assume that the primary reason for this situation comes from the spurious behaviors often faced in DR. For instance, unstable tracking systems and brittle initialization procedures make it difficult for non-experts to use DR successfully. Consequently, researchers are concerned about sharing such a difficult-to-use DR system [11].

This lack of baseline solutions for DR has been a roadblock to fair comparisons among existing solutions and the growth of the field. Every DR paper has performed its own set of evaluations. The problem is also clearly illustrated by the lack of user studies comparing different DR systems [1]. To overcome this problem, we require a standardized evaluation scheme.

Using conventional image metrics to evaluate the DR capabilities assumes that a ground truth is available [1]. Unfortunately, many DR scenarios do not have a unique ground truth. The background revealed after an object has been removed depends on the application and context. For example, "removing a manhole cover on the street" could mean generating plausible street textures over the cover, revealing the cavities underneath the cover, hiding the cover with virtual sand, etc. Since DR datasets consist of video pairs of a scene with and without the object of interest, preparing a ground truth image dataset is an ill-posed problem [8].

Yet, having a dataset, even if provides only a single reasonable version of a ground truth, would already enable researchers to benchmark various DR approaches. An image database consisting of carefully defined DR scenarios would also allow collecting human feedback on the compared DR systems. Together, both solutions can serve as a first step towards a DR benchmark.

In summary, the following three factors are missing in current DR research, leading to a kind of replication crisis.

- Missing open modular framework for future DR research
- Missing tool to evaluate DR systems implemented according to the framework
- Missing common dataset of input and reasonable ground-truth video frame pairs

*e-mail: s.mori.jp@ieee.org
†e-mail: dieter.schmalstieg@visus.uni-stuttgart.de

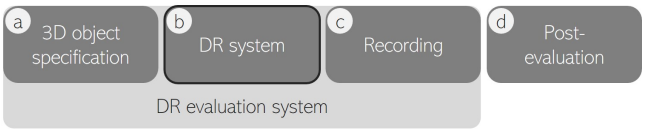

Figure 1: DR research workflow. We need two systems: one for a DR application or its core algorithms and another for evaluation. A DR evaluation system must allow researchers (a) to determine a target in the environment, (b) let users perform DR processing, and (c) record the results, including processed frames and metadata. Finally, researchers (d) perform post-evaluation, including collecting users' subjective feedback and analyzing the data. These two systems need not necessarily be implemented on a single computer system.

In this paper, we try to narrow the issues from a DR research workflow to a possible implementation using openly available development kits and software.

2 DR RESEARCH WORKFLOW

Figure 1 depicts a typical workflow of DR research. Given a DR system under assessment, an evaluation system is responsible for providing predetermined objects of interest and recording the results for a follow-up evaluation. Legacy approaches [6, 9] have used objects specified by authors and provided as many example results as possible for better (but still limited) validity [1].

Such DR prototype systems have usually been implemented on a desktop or laptop PC with a live camera. For mobility and testing in the wild, a client-server system is a more suitable choice, as a mobile device can be used as the client, while a server provides the necessary computational power for running computationally expensive operations such as deep neural network (DNN) evaluations [5].

3 MODULAR FRAMEWORK FOR DR SYSTEMS

The implementation of the DR system involves exploring the design choices for tracking, background modeling, core inpainting, region of interest (ROI) detection, background synthesis, and composition [10]. For brevity, we focus only on static scenes without any real background being revealed in the input frames.

A modular framework lends itself to iterative development because it integrates existing vision, graphics, and interface solutions (e.g., DNN-based image inpainting and semantic segmentation). Recent DR systems are designed for mobile applications with minimal computational load. A common choice is an approach that performs inpainting of images in only one or a few *keyframes* and warps the inpainted frames to match the current view [6, 9]. Another approach uses image streaming with lightweight inpainting performed in every frame [3, 5]. Core inpainting algorithms are typically implemented either using CPU multi-threading [6, 9] or GPU shaders [3, 5].

With a framework like ARFoundation in Unity, camera tracking and 3D object registration are easily available on mobile devices. Assuming a mobile device lacks the computational performance for DNN algorithms, the core inpainting algorithm can be implemented on the server, as suggested in Table 1. The only difference between the two approaches is the frequency of server-client communication.

2771-1110/24/$31.00 ©2024 IEEE
DOI 10.1109/ISMAR-Adjunct64951.2024.00018

Table 1: Baseline DR implementations for keyframe (e.g., InpaintFusion [9]) and streaming (e.g., DeepDR [3]) approaches.

Modules	Keyframe	Streaming
Tracking	Camera tracking	
Background modeling	3D textured mesh	Image frame
Core inpainting	RGB-D DNN-inpainting	RGB DNN-inpainting
ROI detection	3D bounding volume	
Background synthesis	Warping	Direct DNN output
Real–virtual composition	Alpha composition	

4 Ground Truth Collection for DR Systems

We plan to provide a DR dataset as a collection of input and ground-truth video frame pairs. The input image consists of (a) an object of interest in (b) the environment and (c) photometric interactions between them as illustrated in Figure 2. The ground truth frame contains only (b) the environment, making it a counterfactual image [13] (i.e., an image of (b) if there was no (a) and thus no (c)).

Table 2 summarizes potential approaches to collect DR datasets and their pros and cons. **Robo-arm** uses a robotic arm for repeated camera paths to record scenes with and without an object of interest [8]. All the other approaches synthesize a virtual object as an object of interest. **Video** records videos and composites a virtual object into the videos. With editing efforts, such compositions can be of high quality. An immersive (virtual relaity) setup enables researchers to revisit a specific scene. Immersive DR can even be interactive if video streaming is enabled [12]. An object of interest is inserted into the a digital twin of the real scene. Similarly, one may take images with and without a certain object and perform neural radiance field reconstruction [7] or create complete virtual replica of real scenarios[1]. **AR** composes an object of interest into a physical scene. Assuming the AR system is interactive, DR can be added to any AR system.

5 DR System Evaluation

A DR system is usually evaluated by measuring the performance of core modules in a single DR system or comparing it to baselines. Here, having a common framework and example implementations is pronounced.

Recorded video datasets in Section 4 enable us to perform **quantitative evaluation** of DR display results [1]. The AR evaluation system in Section 4 allows us to conduct **user studies** on overall usability, performance, and experiences. Similarly, **performance** is evaluated by collecting videos and framerates on the mobile platform and the server. Measuring network latency can capture another system aspect [5].

6 Conclusion

We described replication issues in DR and clarified possible design choices to address widely missing factors in current DR systems: an open-source modular framework for DR systems, an evaluation platform for DR systems, and datasets for benchmarking. Our future plan includes implementing the framework, recording the datasets, and opening a project page to access the code and datasets.

Acknowledgments

This work was supported in part by the Austrian Science Fund FWF (grant number P33634) and the German Alexander von Humboldt Foundation.

[1]DREAMING challenge https://dreaming.grand-challenge.org/

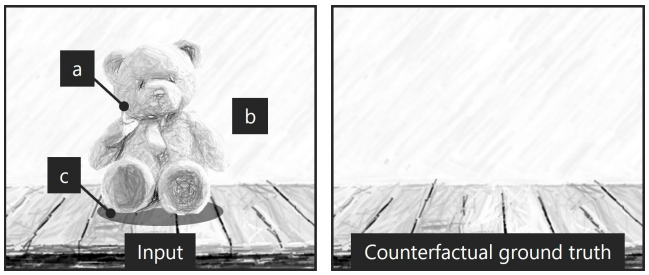

Figure 2: DR dataset. DR requires pairs of input and counterfactual ground truth videos.

Table 2: Approaches to collect ground truth in DR.

Factors	Robo-arm	Video	VR	AR
Experience	Offline	Offline	Off/Online	Online
Background	Real	Real	Virtual	Real
Object of Interest	Real	Virtual	Virtual	Virtual
Counterfactual lv.	Highest	High (sim.)	Mid. (sim.)	Low (sim.)
Replicability				
Scene-level	Yes (in lab)	Yes (in lab)	Yes	Yes (in lab)
Frame-level	Yes	Yes	Yes (on rec)	Yes (on rec)
Tracking-level	Simulated	Simulated	Simulated	Yes
System-level	No	No	Simulated	Yes

References

[1] S. Chen, L. Yu, Y. Liu, Z. Ding, J. Zhang, X. Wang, J. Han, and R. Liu. Diminished reality techniques for metaverse applications: A perspective from evaluation. *IEEE IoT-J*, 2024. (Early access).

[2] V. Gkitsas, V. Sterzentsenko, N. Zioulis, G. Albanis, and D. Zarpalas. Panodr: Spherical panorama diminished reality for indoor scenes. In *Proc. CVPR*, pp. 3716–3726, 2021.

[3] C. Gsaxner, S. Mori, D. Schmalstieg, J. Egger, G. Paar, W. Bailer, and D. Kalkofen. Deepdr: Deep structure-aware rgb-d inpainting for diminished reality. In *Proc. 3DV*, pp. 750–760. IEEE, 2024.

[4] D. Kalkofen, S. Mori, and M. Tatzgern. Visualization and graphics in mixed reality. In *Eurographics 2021 - Tutorials*, 2021.

[5] M. Kari, T. G.-Puppendahl, L. F. Coelho, A. R. Fender, D. Bethge, R. Schütte, and C. Holz. TransforMR: Pose-aware object substitution for composing alternate mixed realities. In *Proc. IEEE ISMAR*, pp. 69–79, 2021.

[6] N. Kawai, T. Sato, and N. Yokoya. Diminished reality based on image inpainting considering background geometry. *IEEE TVCG*, 22(3):1236–1247, 2016.

[7] A. Mirzaei, T. Aumentado-Armstrong, K. G. Derpanis, J. Kelly, M. A. Brubaker, I. Gilitschenski, and A. Levinshtein. SPIn-NeRF: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proc. CVPR*, 2023.

[8] S. Mori, Y. Eguchi, S. Ikeda, F. Shibata, A. Kimura, and H. Tamura. Design and construction of data acquisition facilities for diminished reality research. *ITE Trans. on MTA*, 4(3):259–268, 2016.

[9] S. Mori, O. Erat, W. Broll, H. Saito, D. Schmalstieg, and D. Kalkofen. InpaintFusion: Incremental rgb-d inpainting for 3d scenes. *IEEE TVCG*, 26(10):2994–3007, 2020.

[10] S. Mori, S. Ikeda, and H. Saito. A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. *IPSJ Trans. on CVA*, 9(17), 2017.

[11] S. Mori, D. Schmalstieg, and D. Kalkofen. Good keyframes to inpaint. *IEEE TVCG*, 29(9):3989–4000, 2022.

[12] W. Qiu, F. Zhong, Y. Zhang, S. Qiao, Z. Xiao, T. S. Kim, Y. Wang, and A. Yuille. Unrealcv: Virtual worlds for computer vision. *ACM MM Open Source Software Competition*, 2017.

[13] D. Winter, M. Cohen, S. Fruchter, Y. Pritch, A. Rav-Acha, and Y. Hoshen. ObjectDrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. *arXiv:2403.18818*, 2024.