Toward High-Quality Diminished Reality Using 3D Gaussian Splatting

Ren Okada[1], Norihiko Kawai[1,*]

[1]Graduate School of Information Science, Osaka Institute of Technology, Kitayama 1-79-1, Hirakata, Osaka, Japan

Abstract

Diminished Reality (DR) is a technology that visually removes real-world objects from video in real time, making them appear as if they do not exist in that space. This is achieved by superimposing a hidden background image onto the unwanted object area in the video. However, it is difficult to acquire background images from all positions and orientations in a space, and thus, the generation of background images from arbitrary viewpoints is required. To address this issue, conventional methods have performed appearance correction of background images from different viewpoints using projective transformation, or background image generation using image inpainting. However, these methods may generate unnatural background images due to light reflections and limitations in the quality of image inpainting. In this study, we aim to achieve high-quality diminished reality by automatically detecting unwanted moving objects to be removed, virtually reconstructing the real space using 3D Gaussian Splatting and generating high-quality background images from arbitrary viewpoints. Through experiments in multiple scenes, we evaluated the DR results and processing time, revealing remaining challenges in pose-estimation robustness and computational efficiency toward achieving high-quality, real-time DR.

Keywords

Diminished reality, 3D Gaussian Splatting, Structure from Motion, Perspective-n-Point, COLMAP, YOLOv8

1. Introduction

In the field of television broadcasting, unwanted objects other than the target subject, such as pedestrians and cars, may unintentionally appear in frames. In such cases, the scene should not be broadcast as is from the perspective of portrait rights and privacy protection, requiring some form of real-time editing. To address this issue, Diminished Reality (DR) has been studied as a technique for visually removing unwanted objects from video in real time.

In this study, we propose a method to achieve diminished reality using 3D Gaussian Splatting (3DGS) [1], which can generate high-quality free-viewpoint images in real time. Specifically, we virtually reconstruct the real space using 3DGS from pre-captured images. Using this reconstruction, we generate a high-quality background image corresponding to the current camera pose, and composite it onto the regions of automatically detected unwanted moving objects, thereby visually erasing them from the video. In the following sections, we first provide an overview of related work, and then describe the proposed method.

2. Related Work

Diminished reality is broadly classified into three approaches. In the first approach, which uses cameras installed at multiple viewpoints [2, 3, 4], unwanted objects are removed using background images captured by cameras at different viewpoints. Although this method can obtain background images in real time from cameras at different viewpoints, it requires the installation of cameras at positions where background images can be acquired and synchronization between the cameras.

In the second approach, which estimates the background of the target object [5, 6, 7, 8, 9, 10, 11, 12], unwanted objects are removed by estimating the background of the object from surrounding areas

APMAR'26: The 18th Asia-Pacific Workshop on Mixed and Augmented Reality, Aug. 3-5, 2026, Taipei, Taiwan

*Corresponding author.

norihiko.kawai@oit.ac.jp (N. Kawai)

0000-0002-7859-8407 (N. Kawai)

using techniques such as inpainting. While this method enables diminished reality without restrictions from viewpoints or environments since prior camera installation and capturing are not required, there is a limit to the quality of background images produced by image inpainting, making it difficult to generate natural-looking background images depending on the complexity of the background.

In the third approach, which uses pre-captured background images [13, 14, 15, 16, 17], unwanted objects are removed by generating and compositing the background image of the target object using Image-Based Rendering (IBR). This method can acquire more accurate background images compared to image inpainting; however, the texture mapping approach adopted by many of these methods fails to reflect view-dependent information such as light reflection, and thus cannot perfectly reproduce the background region.

In our proposed method, we focus on the approach using pre-captured background images and utilize 3D Gaussian Splatting. As a conventional method using 3D Gaussian Splatting, a DR system for mobile devices called "DimSplat" has been proposed by Waldow et al. [17]. By acquiring background images from the virtual space, it is possible to generate background images that reflect view-dependent information. However, it requires users to manually perform initial camera alignment and specify the target for removal, making it unsuitable for dynamic environments or automation. In contrast, this study aims to achieve automatic and natural diminished reality by performing automatic detection of unwanted objects using YOLOv8 [18], along with highly accurate automatic estimation of the camera pose via feature point matching and the Perspective-n-Point (PnP) problem.

3. Proposed Method

3.1. Overview

The proposed method is divided into a preparation process and an online process, and the data obtained in the preparation process is used in the online process.

In the preparation process, first, images are captured from multiple viewpoints at the location where diminished reality will be applied. Next, Structure from Motion (SfM) estimates the camera poses of the captured images and the 3D point cloud corresponding to the feature points in the images. Using these as inputs, 3DGS training is performed to generate a virtual space.

In the online process, by performing feature point matching between the input frame obtained from a camera and the captured image with a similar camera pose, the 3D coordinates of the feature points in the input frame are acquired, and the camera pose is then estimated by solving the PnP problem. This estimation is reflected in the camera within the 3DGS virtual space to perform rendering. Next, unwanted objects are detected, and the rendered image is composited onto the detected region as a background image, thereby visually erasing the objects. In the following, each process is described in detail.

3.2. Preparation Process

3.2.1. Reference Image Capture and SfM Application

Images are captured from multiple different viewpoints at the location where users perform DR and then input into COLMAP [19]. COLMAP performs SfM, which matches feature points across multiple input images to determine the relative camera pose and the 3D positions of the feature points. We refer to the captured images as reference images in this paper.

Using the captured reference images, their camera poses, and the 3D point cloud, 3DGS training is performed to reconstruct the real space as a virtual space. 3DGS is characterized by its ability to accurately reproduce view-dependent information, such as light reflections. As a result, it is possible to generate photorealistic and high-quality background images from arbitrary viewpoints while preserving the fine details of the real world.

Table 1
PC specifications.

OS	Windows 11
CPU	Core Ultra 9 285K
Memory	32GB
GPU	GeForce RTX 5090 32GB

3.3. Online Process

3.3.1. Camera Pose Estimation

The video from the camera is processed frame-by-frame. Camera pose estimation is performed by solving the PnP problem using the coordinates of 3D points in 3D space and their corresponding 2D points on the frame. Here, the 3D coordinates of the 3D points corresponding to the 2D points on the reference images have already been obtained through the SfM performed during the 3DGS preparation process. Therefore, by extracting feature points from the input frame and matching them with those in the reference images, the 3D coordinates corresponding to the 2D points on the input frame are determined. For feature point matching, SIFT [20] is used to extract and describe feature points in the input frame, because COLMAP uses SIFT by default for feature extraction in the reference images.

Furthermore, to estimate the camera pose accurately with low computational cost, we select one of the reference images as a target image for feature matching. Here, a score representing the degree of difference in pose is calculated from the camera pose of each reference image obtained via SfM and those of the previous frame of the real-time camera video as follows:

$$score = \alpha\Delta\theta + \beta\Delta x, \tag{1}$$

where $\Delta\theta$ is the difference in the rotation angles of the cameras (in degrees), and Δx is the Euclidean distance between the camera positions, and α and β are weights. The reference image with the smallest score is selected and used as the matching target for the input frame.

3.3.2. Removal of Unwanted Objects Using Background Images

The camera pose estimated by solving the PnP problem is applied to the virtual camera in the 3DGS space, and rendering is performed to acquire an image with the same appearance as the input frame. By compositing this image as a background image onto the unwanted object regions, which are detected by YOLOv8 in the input frame, the unwanted objects are visually erased.

Here, if there is a difference in brightness or color between the rendered background image and the current frame, a difference in brightness or color will occur at the boundary of the composited region, causing an unnatural appearance. Therefore, we apply a luminance correction based on the method by Kawai et al. [9]. The input frame and the rendered image are each divided into grids, and the luminance of each grid cell is calculated as the average of its pixel values. For each grid cell outside the target regions, the ratio of the input-frame luminance to the rendered-image luminance is calculated. The luminance ratios in the target regions are then interpolated using Poisson interpolation. During compositing the background image, each pixel value of the background image is multiplied by the luminance ratio to adjust the luminance of the composited region and achieve a natural appearance.

4. Experiments

4.1. Experimental Overview

In experiments, we verify the effectiveness of the proposed method using three scenes: Scene A (a corridor in front of a laboratory), Scene B (an elevator hall), and Scene C (an outdoor area). Table 1 shows the PC specifications for the experiments. In the following, we describe the settings and results of the preparation and online processes.

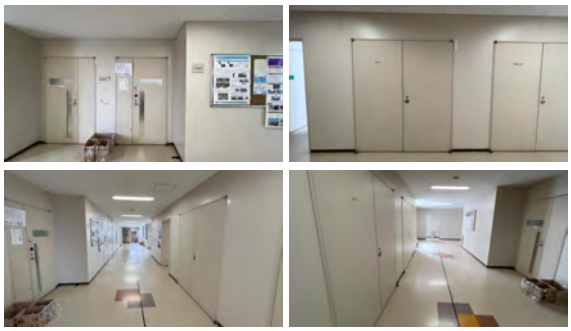


Figure 1: Scene A: Corridor in front of the laboratory.

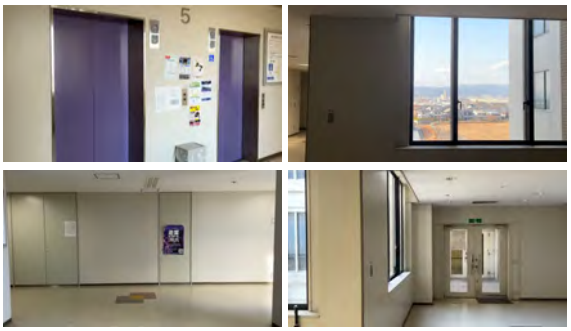


Figure 2: Scene B: Elevator hall.

Figure 3: Scene C: Outdoor area.

(a) Scene A

(b) Scene B

(c) Scene C

Figure 4: 3DGS rendering and SfM results (Left: Rendered image, Right: Camera poses and 3D point cloud).

4.2. Settings and Results of Preparation Process

For 3DGS, we used reference images captured from multiple viewpoints so as to cover the entire space as comprehensively as possible. We used 146 images for Scene A, 220 images for Scene B, and 208 images for Scene C. Figures 1, 2, and 3 show selected images from the captured image sets. Furthermore, the camera poses of the reference images and the correspondence data between the 2D and 3D coordinates of the feature points, which are intermediate outputs during the preparation process for 3DGS training, were stored in a text file.

Figure 4 visualizes the images rendered from free viewpoints with 3DGS, the estimated camera poses, and the reconstructed 3D point clouds after performing SfM with COLMAP. From Fig. 4, we can confirm that the camera poses and 3D point clouds are properly reconstructed, and the rendered images are of sufficient quality to serve as background images.

4.3. Settings and Results of Online Process

For rendering the background images used in diminished reality, we used Splatapult [21]. Splatapult is a 3DGS viewer implemented in C++ and OpenGL.

Although the goal of this study is diminished reality for real-time camera video, in this experiment, as a preliminary step, we used a video with a resolution of 1920 ×1080 pixels and a frame rate of 30 fps, which was recorded at the same location where the reference images were captured. In the method for this experiment, because the matching target image for the first frame of the input video cannot be automatically selected from the pose of the previous frame, the reference image as the matching target was manually specified only for the processing of the first frame. Furthermore, in calculating the score representing the degree of difference in pose to select the matching target, the weight α for the angle difference was set to 0.02, and the weight β for the distance was set to 0.001.

4.3.1. Results for Scene A

Figure 5 shows the input video images and the DR results at the 50th and 100th frames in the experiment for Scene A (a corridor in front of a laboratory). Figure 6 visualizes the results of the feature point matching in the same scene by connecting them with lines. Comparing the composited region and its surrounding region from Fig. 5, we can see that a background image with the same pose is rendered with almost no misalignment. Furthermore, from Fig. 6, roughly the same feature points are connected by lines, indicating that the matching is performed correctly.

Regarding luminance correction, Fig. 7 shows the images before and after luminance correction at the 56th and 95th frames. From Fig. 7(b), the unnatural appearance in the composited region after luminance correction disappeared compared to the surroundings, but as shown in Fig. 7(a), an unnatural appearance was still observed in the composited region even after luminance correction. Since the Poisson interpolation used for luminance correction in this study adjusts the overall luminance of the target region, it may result in unnatural luminance in scenes where the region is partially affected by shadows or direct sunlight.

4.3.2. Results for Scene B

Figure 8 shows the input video images and the DR results at the 2nd and 50th frames in the experiment for Scene B (an elevator hall). Fig. 9 visualizes the results of the feature point matching in the same scene. From the results at the 2nd frame in Figs. 8 and 9(a), background images are rendered with a roughly correct pose for the first few frames, but afterward, as shown in the results at the 50th frame in Figs. 8 and 9(b), almost no reliable feature correspondences are obtained, and the rendered background images are significantly misaligned. This suggests that the reference-image selection based only on the previous estimated pose is not sufficiently robust when the pose estimation error accumulates.

4.3.3. Results for Scene C

Figure 10 shows the input video images and the DR results at the 29th and 88th frames in the experiment for Scene C (an outdoor area). Figure 11 visualizes the results of the feature point matching. From Fig. 10(a), at the 29th frame, a background image with the same pose is rendered with almost no misalignment between the composited region and its surrounding region. Furthermore, from Fig. 11(a), roughly the same feature points are connected by lines, indicating that the matching is performed correctly. However, as shown in Fig. 10(b), from the 60th frame of the input video, a background image in which the composited region is misaligned with the surrounding region is composited, and from the matching results in Fig. 11(b), hardly any feature points are matched, meaning the correct camera pose is not obtained. This is considered to be because hardly any feature points can be matched from the reference images, as shown in Fig. 11(b), and the new camera pose cannot be calculated, which prevents the rendered background image from being updated from the previous frame.

(a) Frame 50

(b) Frame 100

Figure 5: DR results of Scene A
(Left: Input video, Right: DR result).

(a) Frame 50

(b) Frame 100

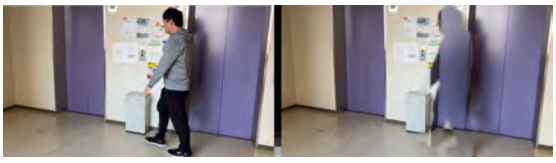
Figure 6: Matching results of Scene A
(Left: Input video, Right: Reference image).

(a) Frame 56

(b) Frame 95

Figure 7: Luminance adjustment comparison of Scene A
(Left: Input video, Center: Before adjustment, Right: After adjustment).

(a) Frame 2



(b) Frame 50

Figure 8: DR results of Scene B
(Left: Input video, Right: DR result).

(a) Frame 2

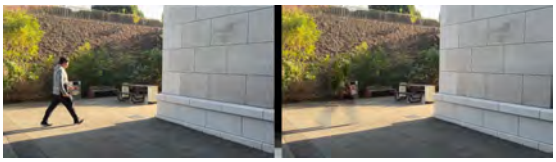
(b) Frame 50

Figure 9: Matching results of Scene B
(Left: Input video, Right: Reference image).

4.3.4. Processing Time

Regarding the processing time, the online process is broadly divided into matching, rendering, and object removal processes. In Scene A, the matching process took an average of 151.0 milliseconds per frame, the rendering process took an average of 39.9 milliseconds per frame, and the object removal with luminance adjustment process took an average of 167.3 milliseconds per frame, indicating that the matching and object removal processes required an overwhelmingly large amount of time. Because

(a) Frame 29



(b) Frame 88

Figure 10: DR results of Scene C
(Left: Input video, Right: DR result).

(a) Frame 29

(b) Frame 88

Figure 11: Matching results of Scene C
(Left: Input video, Right: Reference image).

the current processing speed is approximately 2 fps, improvements in feature point extraction and matching are necessary to achieve 30 fps, which is the frame rate of standard video.

5. Conclusion

In this study, we aimed to achieve diminished reality by virtually reconstructing the real space using 3DGS and generating high-quality background images from arbitrary viewpoints. In the proposed method, as a preparation process, images from multiple viewpoints are captured, the camera poses and 3D point clouds are estimated via SfM, and a 3DGS-based scene representation is constructed. In the online process, the camera pose of the input video is estimated by solving the PnP problem, and unwanted objects are removed by rendering background images from the virtual space and compositing them onto the regions of unwanted objects detected by YOLOv8.

The experimental results demonstrated that DR results could be generated by rendering the 3DGS scene from viewpoints corresponding to the input camera poses; however, depending on the viewpoint and scene of the input video, there were cases where background images with accurate poses were not output. Furthermore, the luminance correction using Poisson interpolation for the composited region still produced a slightly unnatural appearance, indicating the need for a localized luminance adjustment method that can handle shadows and sunlight. In addition, slight errors occurring in the frame-by-frame pose estimation and the fact that the processing time does not reach the frame rate of standard video were identified as remaining challenges. In future work, we will improve the luminance correction and the pose estimation accuracy as well as accelerate the processing speed.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP25H01154, JP25K15156, and Hoso Bunka Foundation.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini and ChatGPT in order to: Text translation, Grammar and spelling check, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. Kerbl, G. Kopanas, T. Leimakuhler, G. Drettakis, 3D Gaussian Splatting for real-time radiance field rendering, ACM Transactions on Graphics 42 (2023) 1–14.

[2] P. Barnum, Y. Sheikh, A. Datta, T. Kanade, Dynamic seethroughs: Synthesizing hidden views of moving objects, in: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, 2009, pp. 111–114.

[3] S. Jarusirisawad, T. Hosokawa, H. Saito, Diminished reality using plane-sweep algorithm with weakly-calibrated cameras, Progress in Informatics (2010) 11–20.

[4] F. Rameau, H. Ha, K. Joo, J. Choi, K. Park, I. S. Kweon, A real-time augmented reality system to see-through cars, IEEE Transactions on Visualization and Computer Graphics 22 (2016) 2395–2404.

[5] O. Korkalo, M. Aittala, S. Siltanen, Lightweight marker hiding for augmented reality, in: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, 2010, pp. 247–248.

[6] J. Herling, W. Broll, High-quality real-time video inpainting with PixMix, IEEE Transactions on Visualization and Computer Graphics 20 (2014) 866–879.

[7] M. Kari, T. Grosse-Puppendahl, L. F. Coelho, A. R. Fender, D. Bethge, R. Schütte, C. Holz, TransforMR: Pose-aware object substitution for composing alternate mixed realities, in: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, 2021, pp. 69–79.

[8] N. Kawai, M. Yamasaki, T. Sato, N. Yokoya, Diminished reality for AR marker hiding based on image inpainting with reflection of luminance changes, ITE Transactions on Media Technology and Applications 1 (2013) 343–353.

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

[10] S. Mori, J. Herling, W. Broll, N. Kawai, H. Saito, D. Schmalstieg, D. Kalkofen, 3D PixMix: Image inpainting in 3D environments, in: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), 2018, pp. 1–2.

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

[12] I. Narutomi, N. Kawai, Reproduction of specular reflection using 3D Gaussian Splatting in diminished reality for AR marker hiding, in: Proceedings of the 16th Asia-Pacific Workshop on Mixed and Augmented Reality (APMAR 24), 2024.

[13] F. L. Cosco, C. Garre, F. Bruno, M. Muzzupappa, M. A. Otaduy, Augmented touch without visual obtrusion, in: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, 2009, pp. 99–102.

[14] Z. Li, Y. Wang, J. Guo, L.-F. Cheong, S. Z. Zhou, Diminished reality using appearance and 3D geometry of internet photo collections, in: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, 2013, pp. 11–19.

[15] S. Mori, F. Shibata, A. Kimura, H. Tamura, Efficient use of textured 3D model for preobservation-based diminished reality, in: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality Workshops, 2015, pp. 32–39.

[16] N. Kawai, T. Sato, Y. Nakashima, N. Yokoya, Augmented reality marker hiding with texture deformation, IEEE Transactions on Visualization and Computer Graphics 23 (2017) 2288–2300.

[17] K. Waldow, J. Scholz, A. Fuhrmann, DimSplat: A real-time diminished reality system for revisiting environments using Gaussian splats in mobile WebXR, in: Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), 2025.

[18] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8, https://github.com/ultralytics/ultralytics, 2023. (Accessed Jun. 2, 2026).

[19] J. L. Schönberger, J.-M. Frahm, Structure-from-motion revisited, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[20] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.

[21] Hyperlogic, Splatapult – 3D Gaussian Splatting VR and viewer, https://

predictable-windows-368364.framer.app/splatapult-3dgs-vr-and-viewer, 2024. (Accessed Jun. 2, 2026).