Integrating Intuitive Interaction and Large Language Model Guidance for Efficient 3D Annotation

Haruya Ishigami†‡, Kenji Iwata† and Yutaka Satoh†‡
†: National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan
‡: University of Tsukuba, Tsukuba, Japan
e-mail: {ishigami.h2002, kenji.iwata, yu.satou} @aist.go.jp

Abstract— **In this paper, we propose a method to intuitively annotate and analyze 3D models of disaster scenes by integrating 3D Gaussian Splatting (3DGS) and a Large Language Models (LLM). However, most existing systems are limited to visual reproduction, and it remains challenging to sufficiently utilize the generated data for practical problem-solving in the real world. Using aerial data captured by drones, we construct a two-layer structure that combines high-quality visual representation through 3DGS with surface mesh geometry. This framework enables the intuitive and flexible placement of annotations, such as panels and lines, at intended positions in 3D space in response to user operations. Furthermore, we developed a method where a LLM analyzes disaster risks based on user instructions and visual information, and subsequently displays the analysis results at arbitrary locations within the 3D environment. By fusing intuitive information sharing with AI-driven analysis support, the proposed system enhances the efficiency and quality of decision-making in disaster management.**

Keywords-3D Gaussian Splatting; Annotation System; Large Language Model.

I. INTRODUCTION

Recently, the emergence of innovative 3D scene reconstruction technologies, exemplified by 3DGS [1], has enabled the rapid generation of high-quality and high-density 3D models for large-scale and complex environments. Building on this development, numerous efforts are being made across various fields to high-fidelity reconstruct and visualize real-world spaces using images acquired by drones and other platforms [2]. However, most existing systems primarily focus on visual reproduction and viewing, and it remains challenging to sufficiently utilize the generated 3D data for practical decision-making and problem-solving in the real world [3].

Data utilization is particularly indispensable in disaster response scenarios to rapidly assess damage and consider rescue plans or evacuation routes. While the use of 3D models via drones has proven effective for gaining a bird's-eye view of an entire site [4], simple visualization is insufficient to support complex decision-making in environments characterized by high uncertainty. To achieve rapid and accurate information sharing, a mechanism is required to directly annotate crucial information—such as hazardous locations, navigable paths, and damage status—onto the 3D model. Furthermore, to supplement limited human resources on-site and ensure reliable rescue operations, it is considered highly effective to combine objective situational analysis and decision-making support provided by AI [5].

3DGS is well-suited for the purpose of this study due to its ability to generate models with high speed and excellent visual quality. However, 3DGS is a volumetric rendering method that represents each point as 3D Gaussian distributions, which are then projected into screen space to synthesize images. Consequently, it lacks explicit geometric boundary information, such as distinct surfaces and edges, presenting a technical challenge in that depth information cannot be directly obtained. To associate and display manual user input or AI analysis results at accurate positions in 3D space, obtaining reliable depth information is essential.

In this study, we constructed a system that integrates 3D model generation technology with a LLM to improve operational efficiency and realize smooth information sharing on-site. Specifically, the process begins with capturing site images and constructing a data foundation. Next, mesh information is generated using Surface-Aligned Gaussian Splatting (SuGaR) [6] after initial processing with COLMAP [7]. By leveraging accurate depth information derived from this mesh, we enable intuitive 3D spatial annotation by the user. Furthermore, we implemented advanced analysis support functions by integrating an LLM. This allows the LLM to verbalize damage situations and propose countermeasures based on user-provided information, visual features of the site, and instruction prompts.

The objective of this system is to enhance the quality of decision-making in disaster scenarios by fusing intuitive information sharing with AI-driven analysis support. This approach, which allows for rapid information annotation while seamlessly navigating between bird’s-eye and immersive perspectives, is expected to drastically improve information transmission efficiency among rescue teams and related organizations, thereby contributing to more reliable rescue operations.

II. RELATED WORK

3D Gaussian Splatting (3DGS) is a method that takes multi-view images as input, represents and optimizes a 3D point cloud as Gaussian distributions, and generates novel view images with high speed and high precision. It acquires a 3D point cloud and camera parameters from RGB images using COLMAP's Structure-from-Motion [7]. Using this point

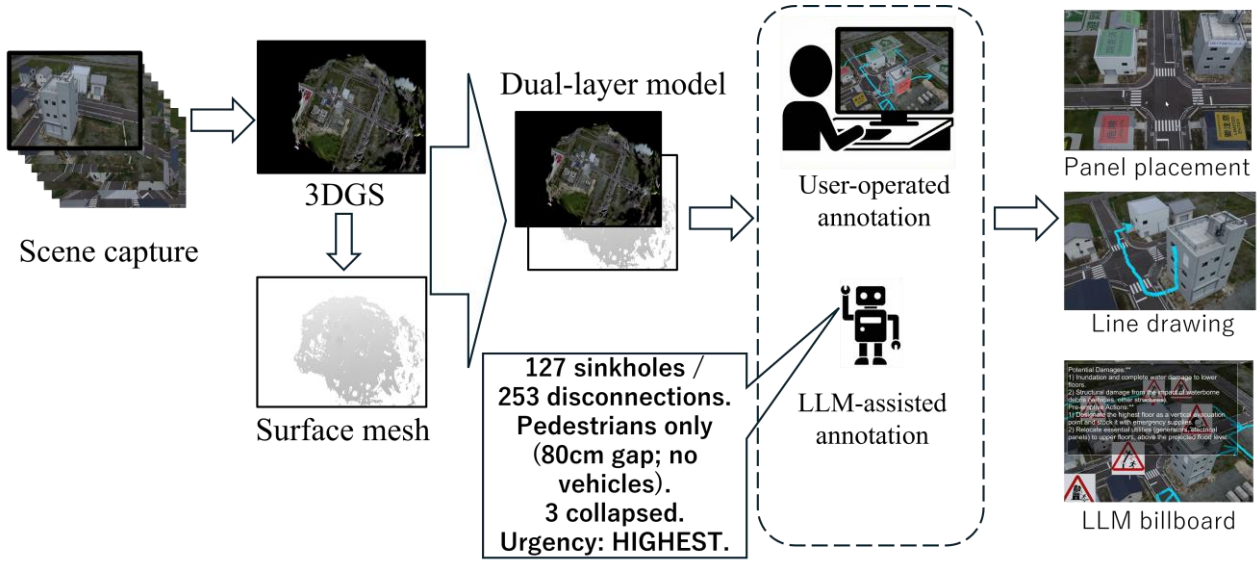

Figure 1. Overview of our proposed LLM-assisted 3D annotation system.
Drone-captured imagery is used to generate a Dual-layer model (integrating 3DGS and SuGaR surface mesh). The system combines user-driven operations (panels/lines) with LLM-assisted analysis (disaster risk/mitigation) to embed spatial information. The resulting annotations are viewable from arbitrary perspectives.

cloud as an initial Gaussian distribution, it optimizes various parameters (3D position, opacity, covariance, and SH coefficients representing color) through a machine learning approach. During optimization, Gaussians are also added or removed. Numerous presentations have been made regarding this method, including applications to diverse fields, such as robotics and VR, as well as speed and accuracy improvements [2].

Generally, 3DGS requires images captured from several dozen different viewpoints. If there are few input images, the shooting and calculation costs are reduced, but there is a problem that rendering quality drops significantly. In response to this, Few-shot View Gaussian Splatting [9] estimates monocular depth from rendered images and uses that information to complement the shape of unobserved regions, achieving high-quality view synthesis even from about three images.

Furthermore, Instant Splat [10] is a method that generates an initial point cloud from a large-scale pre-trained geometric model without relying on SfM. It reconstructs both camera parameters and the scene with high speed and high precision through self-supervised optimization. By using Gaussian-based bundle adjustment, it achieves a speedup of more than 30 times compared to conventional 3DGS while realizing high rendering quality (SSIM) even in sparse-view environments.

On the other hand, while 3DGS enables high-quality image generation, there is a challenge that the optimized Gaussian distribution is not structurally organized, making it difficult to extract a clear mesh structure. Therefore, it is difficult to use the generation results directly for editing or annotation, and additional processing is required to apply it to operations where users specify arbitrary positions in 3D space to add information. To address this problem, SuGaR, proposed by Guédon et al., introduces a regularization term to align Gaussian distributions with the scene surface. Using this alignment, it achieves fast and high-precision mesh extraction via Poisson reconstruction. Furthermore, through a mechanism that simultaneously optimizes the mesh and Gaussians, it enables high-quality rendering in a shorter time compared to conventional Neural SDF-based methods, as well as flexible operations, such as editing, animation, and lighting adjustment.

In recent years, the utilization of drone technology in disaster response has attracted attention, and comprehensive organization of these trends is proceeding. A survey [4] investigated 52 research papers published between 2009 and 2020, classifying drone applications in disasters into four categories: Mapping/Disaster Management, Search and Rescue, Transportation, and Training [8]. The contribution to the mapping field is particularly notable, and its effectiveness for situational awareness and decision-making support at disaster sites has been confirmed. On the other hand, discussions regarding use in post-disaster areas, such as victim identification and medical support, are not yet sufficient and are pointed out as future challenges.

Applications that realize an interactive 3D manipulation environment using pen input and touch operations on mobile devices include Feather [11] and Cozy Blanket [12]. While these applications possess high operability and convenience, they are primarily aimed at 3D content creation and do not support annotation uses for adding and sharing information on 3D models of real spaces.

Research on the systematization of interaction design in Virtual Reality (VR) and Augmented Reality (AR) environments is also progressing [13]. Research is being conducted on immersive systems that allow interactive manipulation [14], such as deforming 3D content generated by 3D Gaussian Splatting, and frameworks that can perform shape changes, color adjustments, and style transfers on 4D scenes [15]. However, many of these focus on model

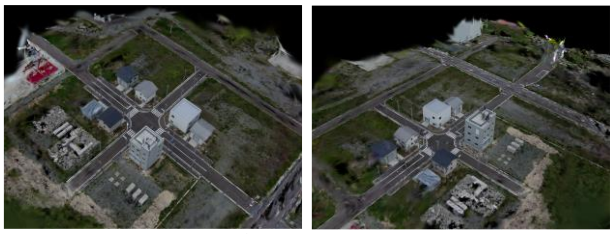

Figure 2. 3DGS

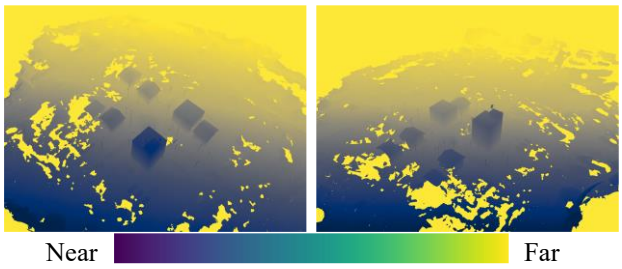

Figure 3. Example of Depth Information Computation (Color map)

generation or interaction design, and there is room for consideration regarding mechanisms to support information addition and sharing on generated 3D models. Therefore, in this study, we propose a method to realize intuitive and accurate annotation operations using 3D model generation technology to address such issues.

III. PROPOSED METHOD

A. System Overview

This system is an interactive annotation tool, and its overview is illustrated in Figure 1. From the input images, a 3DGS model is generated for visual display (Figure 2), and a surface mesh is constructed using SuGaR. Figure 3 shows the result of computing depth values from the surface mesh and assigning a color map to them.Users can freely navigate the 3D scene from arbitrary viewpoints, place pictograms representing disaster risks analyzed by the LLM (Google Gemini), and draw line objects through mouse-drag operations.

B. Determination of Target Coordinates

Since 3DGS models do not explicitly represent surface geometry, a surface mesh generated by SuGaR is introduced to identify the 3D coordinates corresponding to a click position on the screen.

When a user performs an annotation operation, the depth value d corresponding to the input screen coordinates (x, y) is determined based on the surface mesh. The screen coordinates (x, y) and the depth value d are then converted into normalized device coordinates (NDC). The conversion is defined as follows:

$$x_{ndc} = \frac{2x}{ScreenWidth} - 1\,,$$

$$y_{ndc} = \frac{2y}{ScreenHeight} - 1\,, \tag{1}$$

$$z_{ndc} = 2d - 1\,,$$

Here, d is a normalized depth value where 0 is the nearest and 1 is the farthest within the range from the camera's Near Clip to Far Clip.Next, using the camera's projection matrix P, the coordinates (x_c, y_c, z_c) in the camera coordinate system are obtained from the NDC coordinates $(x_{ndc}, y_{ndc}, z_{ndc})$ as follows:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ w \end{bmatrix} = P^{-1} \begin{bmatrix} x_{ndc} \\ y_{ndc} \\ z_{ndc} \\ 1 \end{bmatrix} \tag{2}$$

As a correction for perspective transformation, the obtained camera coordinates are normalized using w:

$$x'_c = x_c / w, \qquad y'_c = y_c / w, \qquad z'_c = z_c / w \tag{3}$$

Finally, by using the view matrix V, the world coordinates $(X_w, Y_w, Z_w)$are obtained:

$$\begin{bmatrix} X_w \\ Y_w \\ Z_w \\ w \end{bmatrix} = V^{-1} \begin{bmatrix} x'_c \\ y'_c \\ z'_c \\ 1 \end{bmatrix} \tag{4}$$

Based on the computed world coordinates, annotations are placed at positions in the 3D space corresponding to the user's operations. During this process, to prevent misdetections caused by noise inherent to 3DGS or occlusions such as power lines, offset processing is applied. In addition, for line drawing operations, instead of directly using the depth value corresponding to the simple click position, the median depth value within a neighboring region is used. This approach enables stable target coordinate specification even in the presence of outliers.

C. Panel Placement Method

In the panel placement process (Figure 3), the user first selects the type of panel to be placed, and the panel is then positioned at the user-specified location via a click operation. Since the orientation of the panel must be adjusted for each specified location, the proposed system determines the orientation of the panel object based on the local structure of the 3D scene around the target coordinates and the user's viewpoint at the time of annotation input.

To uniquely define an orientation in 3D space, two axes are required: the Z-axis (forward vector) and the Y-axis (upward vector). In this system, the Z-axis direction is estimated by applying principal component analysis (PCA) to infer the orientation of the surface in the vicinity of the target coordinates, while the Y-axis direction is determined using the upward vector of the camera at the time of annotation input.

By defining the Z- and Y-axis directions in this manner, the panel object is naturally aligned with the surface geometry, adapting to local surface irregularities and inclinations, thereby improving visual consistency. In addition, using the camera's upward direction for the Y-axis preserves the expected notion of the "top" of the panel relative to the user's

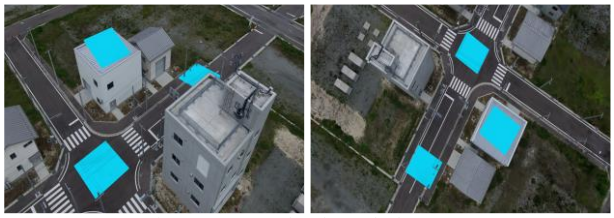

Figure 4. Example of Panel Placement

Figure 5. Example of Line Drawing

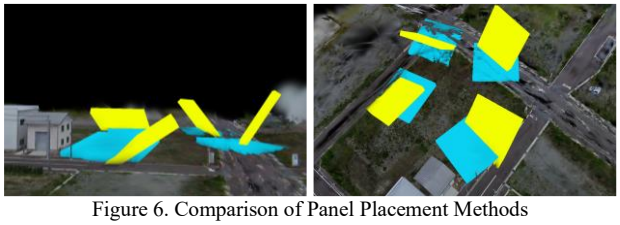

Figure 6. Comparison of Panel Placement Methods

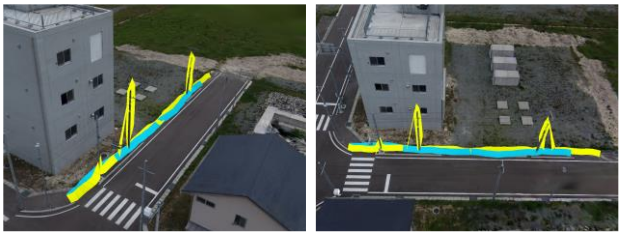

Figure 7. Comparison of Line Drawing Methods

viewpoint at the time of annotation, enabling panel placement that is consistent with the user's operational intent.

To obtain the surface normal of the 3D scene around the target coordinates, the user specifies a single point in screen space. Screen-space coordinates within an n-pixel neighborhood around the selected point are then collected. The corresponding depth values are retrieved and converted into 3D coordinates. Finally, principal component analysis (PCA) is applied to these local 3D points to estimate a normal vector representing the local surface orientation.

D. Line Drawing Method

In the line drawing process (Figure 5), coordinates sampled during the user's drag operation are connected by cylindrical segments to form a polyline. If input coordinates are sampled every frame, excessive sampling density causes the system to overly react to minor surface irregularities (e.g., small bumps on the ground), resulting in unintended zigzag trajectories. Conversely, excessive smoothing may lead to the loss of important terrain features such as steps or steep slopes.

To address this trade-off, sampling is triggered based on a fixed movement distance threshold. This approach preserves essential terrain characteristics while reducing noise caused by small surface irregularities, enabling smooth line drawing. In addition, to prevent drawn points from being embedded within the model surface, a constant offset is applied in the direction toward the camera.

When drawing lines, objects located in the foreground may interfere with the intended placement on the target surface. To mitigate this issue, in addition to the depth value corresponding to the user-specified screen coordinate (x, y), depth values of neighboring screen coordinates are also computed. By using the median of these depth values, the system enables stable line placement on the intended target surface even when small foreground objects, such as utility poles, are present.

E. Integration of LLM Analysis and Human Operation

In the LLM-assisted functionality, risk factors and mitigation measures related to structures displayed on the LLM interface are generated at arbitrary timings specified by the user. Although LLMs are capable of advanced situational reasoning, they are not well-suited for precise 3D localization. Therefore, the proposed system adopts a process in which the LLM generates semantic information and the user explicitly specifies its placement location.

The generated information is placed as billboard-style pictograms in the 3D space. When a pictogram is clicked, detailed analysis results are displayed in a pop-up window, allowing the user to interactively review the generated content.

IV. EXPERIMENT

We conducted an evaluation using disaster simulation data obtained at the Fukushima Robot Test Field [8]. In the comparison of panel placement methods (Figure 6), principal component analysis (PCA) was employed to determine the surface normal direction, and panel placement was performed while varying the number of points used for normal estimation. Specifically, 9 points (yellow panels) and 225 points (light blue panels) were used. Compared to the yellow panels, the light blue panels were less affected by local surface irregularities in the surface mesh at the placement location and were more consistently aligned with the ground plane, indicating more stable placement.

In the comparison of line drawing methods (Figure 7), using the click position coordinates directly (yellow line in Figure 7) resulted in interference with obstacles such as utility poles. In contrast, when using the median of depth values (light blue line in Figure 7), stable line drawing was achieved along the ground context without interference. These results confirm that lines can be accurately drawn along object surfaces, such as the ground and building walls.

Regarding the evaluation of the LLM-assisted functionality, the objective of this experiment was to integrate and visualize analysis results obtained from Gemini within the 3D space, and to verify its operation. For each target structure,

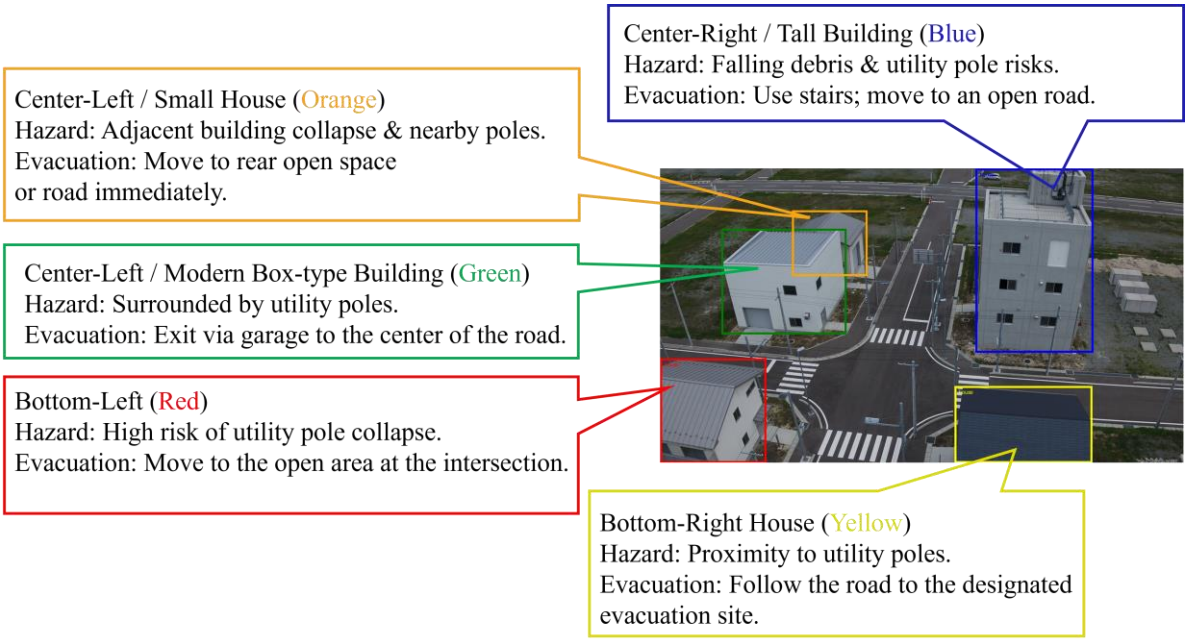

Figure 8. LLM output result

Figure 9. Use Case of the System

the LLM analyzes potential risks that may arise during a disaster and provides guidance on evacuation actions.

Figure 8 shows an example of output generated by the LLM. For each detected building, the results are categorized into risk information and evacuation guidance. Even when buildings are spatially close to each other, the results are distinguished by color-coded representations, and positional information and simple structural features are provided, making it easy to confirm the correspondence between the scene and the generated textual descriptions. For example, the result shown in the center-left (Orange) recognizes that the building is adjacent to a large house and recommends evacuating to a vacant lot behind the building rather than to an intersection. The result shown in the center-right (Blue) generates statements such as “tall structure” and “caution regarding falling rooftop equipment,” confirming that relevant structural characteristics are identified. In addition, relatively small objects such as utility poles are also recognized, and appropriate warnings indicating danger or caution during evacuation are generated. These results demonstrate that the system can assess risks and propose evacuation actions based on the identified risks.

The generated results can be placed as pictograms at appropriate locations in the 3D space through user interaction. Detailed analysis results output by the LLM can also be reviewed via pop-up displays (Figure 9, left). These findings demonstrate that users can rapidly and comprehensively add information by simply reviewing the LLM’s proposals and specifying placement locations, without the need to compose textual descriptions from scratch. Figure 9 also shows the results of annotation operations performed in the 3D space. As demonstrated, user-generated annotations can be reviewed from various viewpoints, enabling flexible and comprehensive spatial understanding.

V. CONCLUSION

In this study, we proposed a system that enables accurate annotation placement by combining the high-quality visual rendering of 3D Gaussian Splatting (3DGS) with depth information derived from a mesh model. By incorporating a surface mesh, the proposed approach overcomes the inherent

limitation of 3DGS in depth acquisition. Furthermore, the use of median depth values and offset processing allows for noise-robust and highly visible annotations.

The proposed mechanism, in which AI and humans collaboratively enrich information in a 3D space, provides a foundational framework for future disaster response systems. As for future work, we plan to conduct usability evaluations through questionnaires and user studies assuming real-world field operations, to assess the operability of the system and the effectiveness of information presentation. In addition, we will explore the integration of LLMs fine-tuned with disaster-related data to improve the reliability of generated proposals and to further automate expert-level situational assessment.

By seamlessly connecting the information organized within this system to on-site operational support and enabling end-to-end assistance from analysis to field activities, we aim to contribute to faster and more effective disaster response.

REFERENCES

[1] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," ACM Trans. Graph., vol. 42, no. 4, pp. 139:1-139:14, 2023.

[2] B. Fei, J. Xu, R. Zhang, Q. Zhou, W. Yang, and Y. He, "3D Gaussian as a New Vision Era: A Survey," arXiv preprint arXiv:2402.07181, 2024.

[3] F. Nex, D. Roca, J. Fritsch, M. Gerke, and N. Kerle, "Seismic Damage Semantics on Post-Earthquake LOD3 Building Models Generated by UAS," ISPRS Int. J. Geo-Inf., vol. 10, no. 5, p. 345, 2021.

[4] S. M. S. M. Daud, M. Y. P. M. Yusof, C. C. Heo, L. S. Khoo, M. K. C. Singh, and M. S. Mahmood, "Applications of drone in disaster management: A scoping review," Science & Justice, vol. 62, no. 1, pp. 30-42, 2022.

[5] T. Seidl, C. Heckman, and S. Nikolaidis, "3D Gaussian Splatting for Human-Robot Interaction," in Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), pp. 986-990, 2024.

[6] A. Guédon and V. Lepetit, "Sugar: Surface-aligned gaussian splatting for efficient 3D mesh reconstruction and high-quality mesh rendering," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5354-5363, 2024.

[7] J. L. Schönberger and J. M. Frahm, "Structure-from-Motion Revisited," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4104-4113, 2016.

[8] Fukushima Robot Test Field, https://www.fipo.or.jp/robot/en/ 2025.12.01.

[9] Z. Zhu, Z. Fan, Y. Jiang, and Z. Wang, "FSGS: Real-Time Few-shot View Synthesis using Gaussian Splatting," in European conference on computer vision, Cham: Springer Nature Switzerland, pp. 145-160, 2024.

[10] Z. Fan et al., "InstantSplat: Sparse-view SfM-free Gaussian Splatting in Seconds," arXiv preprint arXiv:2403.20309, 2024.

[11] Feather, https://www.feather.art/ 2025.12.01.

[12] CozyBlanket, https://sparseal.com/cozyblanket/ 2025.12.01.

[13] M.-X. Chen, H. Hu, R. Yao, L. Qiu, and D. Li, "A Survey on the Design of Virtual Reality Interaction Interfaces," Sensors, vol. 24, no. 19, pp. 6204, 2024.

[14] Y. Jiang et al., "VR-GS: A Physical Dynamics-Aware Interactive Gaussian Splatting System in Virtual Reality," arXiv preprint arXiv:2401.16663, 2024.

[15] L. Liu, C. Wang, Z. Chen, and D. Xu, "4DGS-Craft: Consistent and Interactive 4D Gaussian Splatting Editing," arXiv preprint arXiv:2510.01991, 2025.