StatueDiff: Regionally Enhanced Cross-Scale Diffusion for Single-View 3D Reconstruction of Classical Statues

Yiming Du[1], Nicola Lercari[1]

[1] LMU Munich, Institute for Digital Cultural Heritage Studies, 80539 Munich, Germany -
yiming.du@campus.lmu.de, n.lercari@lmu.de

Keywords: Deep Learning, Computer Vision, 3D Reconstruction, Digital Preservation, Museum Exhibition.

Abstract

Promoted by techniques such as deep learning, automated 3D reconstruction of cultural artefacts and historical sites has made significant progress in many application fields in recent years, such as the creative industry. In museums, digital technologies are driving the transition from static display forms to more immersive and dynamic exhibition experiences. However, most artefacts in museum environments are still displayed in traditional showcases, with only a limited number of institutions adopting digital twins and presenting 3D models through their online platforms. This situation is particularly evident in the case of artefacts such as statues, with an emphasis on passive viewing and limited interaction with the audience. The development of permanent exhibitions that move beyond static visual presentation thus remains a largely unexplored frontier in interactive technology in museums. To address these challenges, this paper proposes a diffusion-based framework for Single-View 3D Reconstruction, using sculptural heritage objects as a case study. This deep learning approach takes a colour photograph as input and aims to generate a photorealistic 3D reconstruction with detail suitable for digital exhibitions in museums. Our preliminary results show that this approach has the potential to create a novel integration of art and technology, contributing to new possibilities for audience participation and immersive experiences within cultural heritage exhibitions.

1. Introduction

With the advancement of deep learning techniques, virtually reconstructed places and artefacts of the past have increasingly become an established practice citation (Lercari and Busacca, 2020). This technology facilitates the accurate recording and long-term preservation of cultural artefacts, creating new opportunities for digital exhibitions and interactive experiences in museums and cultural institutions. However, most museum artefacts are currently displayed in traditional formats, especially statues, which are typically limited to static displays lacking interactive and immersive experiences for visitors. This limitation is mainly due to the relatively high cost and time-consuming process of existing 3D reconstruction techniques, which makes it challenging for museums to expand the use of digital twin technology more widely. Achieving efficient, low-cost, and highly interactive digital exhibitions has therefore become an important issue in the field of cultural heritage.

Current techniques for 3D reconstruction, such as Triangulation Laser Scanning, Structured Light Scanning, and Silhouette-based Reconstruction, can produce high-precision 3D models and adapt to various cultural heritage scenarios. However, these techniques often require complex and expensive equipment and professional operation, making them economically unfeasible and inflexible when dealing with a large number of museum artefacts. Furthermore, existing research has shown that exhibitions with higher interactivity positively impact visitors’ attention and experience (Bollo and Dal Pozzolo, 2005). The development of lightweight and easily scalable interactive digital exhibition technologies has thus become an important approach to enhancing the attraction of museums.

The recent emerging diffusion models provide a new solution to these challenges. As a novel probabilistic generative model, diffusion models achieve high-quality outputs by simulating a gradual denoising process. These models have made breakthrough achievements in tasks such as image generation, video synthesis, and 3D data generation. In the field of 3D reconstruction specifically, diffusion models have demonstrated strong generalization capabilities, enabling the inference of complete 3D structures from limited visual information, such as reconstructing complete 3D models from a single or few images.

This paper uses Greek statues as a case study and proposes a single-view 3D reconstruction framework based on diffusion models. We aim to efficiently generate detailed and photorealistic 3D models from a single input image to support digital interactive exhibitions in museums. More specifically, we constructed a high-quality dataset of Classical and Hellenistic sculptures. We supplemented it with high-precision human body scan data to train diffusion models for cross-scale 3D shape generation. Additionally, cross-scale diffusion, local enhancement, and texture fusion are used to further refine the preliminary 3D structures into high-fidelity mesh models, integrating texture information to enhance visual realism.

Preliminary results demonstrate that the diffusion-based framework shows significant advantages in single-view reconstruction tasks for statues. It effectively captures overall structures and accurately presents local details. This approach has the potential to overcome the current limitations in museum digital exhibitions, providing visitors with a more interactive and immersive viewing experience.

2. Related Work and Background

2.1 3D Reconstruction for Digital Preservation of Cultural Heritage

The digital 3D reconstruction of cultural heritage has increasingly demonstrated its importance in contemporary society

(Gomes et al., 2014). This technology not only accurately documents and preserves the appearance of artefacts, providing valuable digital archives for future generations, but also plays a key role in creating high-precision replicas, identifying artefact forgeries, and promoting the widespread dissemination of virtual museums and digital collections. Currently, various 3D reconstruction techniques are widely applied in the field of cultural heritage, each based on different principles and equipment suitable for different scenarios. The following are some important technologies used in this field (Gomes et al., 2014):

2.1.1 Triangulation Laser Scanning: Known for its high precision, this technique involves a device emitting a laser pattern onto an object, and calibrated optical sensors capture their positions, calculating depth through simple triangulation principles to build 3D models (França et al., 2005).

2.1.2 Structured Light Scanning: The system projects a predefined light pattern onto an object and captures the light as it interacts with the scene. By analyzing the information from the pattern distorted by the scene, the 3D geometric shape is reconstructed (Bell et al., 1999).

2.1.3 Shape from Stereo: This method employs a dual optical sensor system to capture two images simultaneously. It determines depth by calculating distances between corresponding points in the images and is often used for real-time 3D data acquisition (Scharstein and Szeliski, 2003).

2.1.4 Shape from Silhouette: By collecting multi-perspective images, this method deduces an object's 3D geometry from its silhouette information. In the process of building the model, texture information is frequently integrated into the model's construction (Laurentini, 1994).

2.1.5 Shape from Shading and Shape from Photometry: Both methods utilise a collection of images of an object taken under different illumination conditions to determine its geometry. To improve the accuracy of reconstruction, shape from photometry often incorporates calibrated reference objects or light sources (Gomes et al., 2014).

2.1.6 Shape from Focus: This method is commonly applied in microscopy to observe small objects such as insects. It infers depth by analyzing images captured from the same viewpoint with varying focal adjustments (Schechner and Kiryati, 2000).

2.1.7 Topographic Methods: This method comprises a geodetic station suitable for collecting 3D data from large sites or vast areas.

Although these methods are widely popular and effective across archaeological and cultural heritage contexts, they also share some limitations. From a broader perspective, acquiring high-fidelity digital twins often requires significant time and financial investments. A challenging trade-off therefore exists between high data accuracy and lower economic and temporal costs.

In the context of museums, applying these traditional 3D scanning techniques to many exhibits would create significant economic and time pressures, thus restricting the capability of museums for large-scale digitization. Museum visitor behaviour has been analysed, highlighting that it is stimulated and guided by a complex system of signs and objects, with exhibit interactivity significantly influencing visitor engagement and attention (Bollo and Dal Pozzolo, 2005). Therefore, balancing digitization costs with visitor attraction is an essential consideration. Currently, there are some studies on museum interaction aiming to enhance their interactivity and attraction. For example, a gamified application was developed using deepfake technology to animate the heads of sculptures, enabling the statues to speak in sync with text or audio files through lip movements, thereby enhancing visitor engagement (Zaramella et al., 2023). However, interactive research specifically focused on full-body classical statues remains relatively limited and requires further exploration.

2.2 Diffusion Models

Diffusion Models are an emerging class of probabilistic generative models (Croitoru et al., 2023). In recent years, they have significantly changed the landscape of generative artificial intelligence, particularly making breakthrough progress in image generation tasks and even surpassing the longstanding dominance of Generative Adversarial Networks (GANs) in this challenging area (Goodfellow et al., 2014). The core idea of diffusion models comes from non-equilibrium thermodynamics, simulating a special "diffusion" process to generate data.

There are two stages in this process. The first stage is the *forward diffusion process*, which gradually adds noise to complex data distributions until the data becomes a simple, easily sampled noise distribution, typically Gaussian. The second stage is the *reverse denoising process*. During this process, the model starts with pure noise and iteratively refines it step-by-step, gradually recovering the original data's structure and details from the noise. With each iteration, the model gradually guides the data closer to the actual data distribution by making predictions and eliminating a tiny amount of noise from the current state. Taking image generation tasks as an example, diffusion models repeatedly add noise to an image until it becomes a pure noise image and then gradually remove the noise, ultimately producing a clear new image closely resembling the original.

Diffusion models are fundamentally based on three main mathematical frameworks: denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020), score-based generative models (SGMs) (Song and Ermon, 2019), and stochastic differential equations (Score SDEs) (Karras et al., 2022). In computer vision, diffusion models not only handle direct image generation but also demonstrate strong generalization capabilities and are applied to more complex tasks, such as semantic segmentation, video generation, point cloud completion and generation, anomaly detection, and more.

Diffusion models are also widely used for generating 3D data with various forms of 3D representation, including explicit representations such as meshes and voxels and implicit representations such as point clouds and implicit functions. By leveraging the 3D priors data within diffusion models, existing research has already achieved the reconstruction of objects from limited viewpoints. For example, NeRDi uses diffusion models to guide the enhancement of a single image to 360° view synthesis (Deng et al., 2023). It applies a pre-trained Stable Diffusion model to denoise NeRF-rendered results, ensuring alignment with the input image. DreamSparse utilizes a pre-trained diffusion model to synthesize views under sparse viewpoints (Yoo et al., 2023). It extracts geometric features from input views through a 3D geometry module. It trains a spatial guidance model to condition the pre-trained diffusion model, ensuring that synthesized images align consistently with the viewpoints of the input object. Zero1to3 also fine-tunes a pre-trained Stable

Diffusion model using a synthetic dataset containing paired images and their relative camera parameters, enabling the generation of multiview images from a single input image (Liu et al., 2023).

Applying diffusion models to the 3D modelling of museum exhibits offers an efficient and economical new path for digitally exhibiting cultural artefacts by reconstructing their 3D models from limited image inputs. This innovative method not only significantly reduces the time and resources required by museums while maintaining the basic accuracy of heritage artefacts, but more importantly, these high-efficiency generated 3D digital twins can be integrated into museum interactive exhibits, greatly enhancing visitors' interactive experience.

3. Methodology

3.1 DataSet

The statues of the Classical and Hellenistic periods are known for their realistic and naturalistic styles, featuring the structure of muscles and bones and the aesthetics of the human body in a realistic and naturalistic manner (Fowler, 1989). However, as deep learning models require a large amount of data to train on, the open-source 3D dataset specifically tailored for statues artefacts in this period is limited.

To address this problem, we have constructed a dataset tailored explicitly for statues from the Classical and Hellenistic periods. This collection is designed to facilitate advanced 3D reconstruction research. To further enhance the model's understanding and generalization capability of realistic human body geometry, we supplement our primary dataset with high-quality human body 3D scanning datasets widely recognized in the research community (e.g., THuman2.0 (Yu et al., 2021)). For a quantitative evaluation of the reconstruction performance of the model in realistic contexts, we also used the existing datasets containing multiple temporal sequences of human body scans (e.g., CAPE dataset (Ma et al., 2020)).

Our core dataset, the Sculptural Heritage Dataset, contains high-quality multiview images and corresponding high-precision 3D point cloud data derived from a substantial collection of sculptural replicas of Museum für Abgüsse Klassischer Bildwerke (MfA) in Munich, Germany. The museum holds an extensive collection of 1:1 scale gypsum replicas of well-known Classical and Hellenistic statues. We systematically collected numerous high-resolution multiview images using a Digital Single-Lens Reflex camera (DSLR) and the corresponding 3D scans through a structured-light laser scanner. This comprehensive collection of 2D images and 3D geometric data forms the foundational basis for our research.

For 3D data collection, we used the Artec Leo hand-held structured light scanner to create a high-resolution 3D model of the sculpture(Figure 1). During scanning, the Artec Leo consistently operated within a range of approximately 0.3–0.5 metres from the sculpture's surface, scanning up and down and covering the surface comprehensively from multiple angles and overlapping areas to ensure the capture of fine geometric details. We paid particular attention to the complex details of the sculpture, especially its more elaborate features such as the face, clothing, and headdress. At the same time, we visually inspected the acquired data and used on-site comparisons between the scan shown in the Artec Leo display and the real object to ensure data integrity and reduce noise. We then processed the raw scan data using Artec Studio V.19 software and produced polygonal meshes to be used to train the model.

Figure 1. High-resolution 3D scanning model of Classical and Hellenistic period sculptures using Artec Leo (author shown).

3.2 Data preprocessing

To ensure the consistency of training data and robustness of the algorithmic model, all 3D models used for training undergo a series of detailed preprocessing steps. These include quality enhancement of the original scanned data, such as noise correction, registration, hole filling, and texture improvements as a means to achieve complete and high-fidelity geometries. Subsequently, precise geometric alignment and pose normalization are performed, which are crucial for improving the learning efficiency of the model.

For datasets like THuman2.0, predefined parameters can be utilized directly. For our self-constructed Sculptural Heritage Dataset, we employ a semi-automatic process to generate or infer the necessary structured parameters, ensuring compatibility for algorithm training. Finally, we render multiple-view 2D images from these processed 3D models and pair them with corresponding 3D shapes, creating a rich dataset for training 3D reconstruction algorithms.

3.3 Pipeline for Single-View 3D Reconstruction

Our automated 3D reconstruction pipeline, StatueDiff, integrates two core components designed to efficiently generate high-quality 3D models of sculptures from a single image (Figure 2). We selected 80% of the data as the training set and performed multiple rounds of iterative optimisation on this dataset. The remained 20% of the data was used as an independent validation set to verify the robustness and generalisation of the model.

3.3.1 Cross-Scale Diffusion Model: Our method uses a conditional diffusion model that progressively denoises latent representations to achieve the target 3D shape. Specifically, the diffusion model takes a single input image as a conditional signal, which is initially encoded into a latent space to guide the

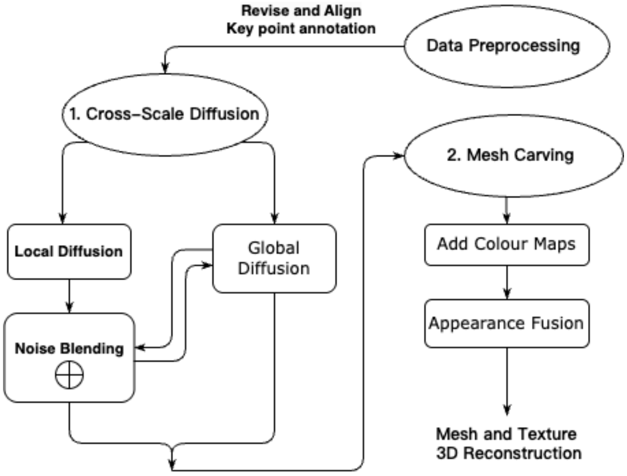

Figure 2. The automated StatueDiff pipeline integrates two core components.

denoising process. The diffusion process effectively learns to restore clear 3D structures from noisy representations. When handling details at different scales within complex 3D structures, we adopt a cross-scale diffusion mechanism, enabling the model to generate overall morphology and finer local features simultaneously. After generating preliminary 3D results via the cross-scale diffusion model, we perform local enhancement targeting the features of complex textures, headdresses, and other critical details of the sculpture. We then model separately these local details using specific feature processing techniques to improve the network’s ability to capture complex patterns. This ensures that these local improvements remain consistent with the 3D structure and avoid the introduction of unnatural artefacts.

3.3.2 Mesh Carving and Texture Fusion: After obtaining refined 3D geometric information, we proceed to the mesh carving and texture fusion stages. In this stage, we initialize the mesh by using prior geometric information estimated from images, laying the foundation for subsequent detailed optimization. We then refine the model’s geometric details through an iterative optimization process, progressively modifying the mesh structure to fill in minor missing areas while ensuring surface continuity. Finally, we integrate texture information into the model’s surface to generate a realistic appearance, ensuring high visual realism in digital presentations.

4. Preliminary Results

Our preliminary results indicate that the StatueDiff framework effectively handles the inherent complexity of varying poses and occlusions found in single images. The generated 3D models maintain the overall structural accuracy of sculptures and successfully capture rich geometric details, including intricate facial features and clothing folds (Figure 3 and Figure 4). These high-quality 3D reconstructions can be integrated into interactive museum facilities, such as immersive information kiosks or augmented reality applications, creating educationally meaningful and visually striking interactive experiences for visitors. This innovative display format goes beyond traditional static exhibitions by significantly improving visitors’ more profound understanding and intuitive experience of cultural artefacts. Moreover, through convenient digital access methods, our approach supports museums in broadening their educational missions and promotes new pathways for public engagement in cultural heritage.

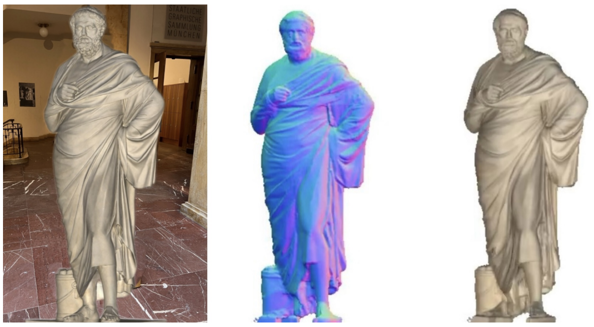

Figure 3. Input Image(Left), Voxel Grid(Middle) and Textured Model(Right) generated by StatueDiff.

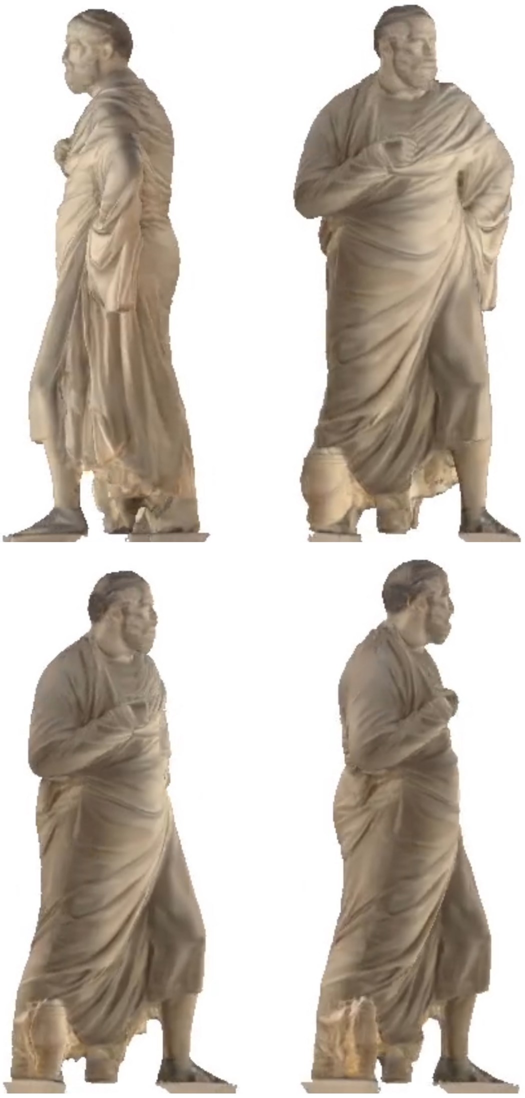

Figure 4. Four-view of the reconstructed 3D sculpture generated by StatueDiff.

5. Conclusions and Future Work

This study proposes StatueDiff, an Enhanced Cross-Scale diffusion for a single-view 3D reconstruction framework for Classical Statues aimed at enhancing interactive museum exhibitions. The approach enhances user interaction by automatically generating detailed 3D models from a photographic input. Through the process, including cross-scale diffusion, mesh carving, and texture fusion, StatueDiff ensures the generated 3D models accurately represent overall structures while fully capturing crucial artistic features such as faces, headdresses, and clothing details. This research is an interdisciplinary exploration of integrating cultural heritage with deep learning techniques. The preliminary results highlight the significant potential of diffusion models in advancing digital cultural heritage applications. In addition to its technical contributions, this framework provides a lightweight and scalable curatorial approach for developing more engaging and immersive museum exhibitions.

Future research will focus on enhancing the material generalization and model robustness to further expand the applicability of this framework across various types of sculptural artefacts. Regarding the material, although our dataset provides high-quality data, it is primarily sourced from the collections and replicas of a single museum, the Museum für Abgüsse Klassischer Bildwerke (MfA) in Munich. This could potentially impact the generalisability of the model algorithm, such as the ability to generalise across different materials, sizes, or preservation conditions of cultural artefacts. In the future, our research will explore expanding the diversity of the dataset to include artefacts with different styles from different periods beyond Classical and Hellenistic, with the aim of further enhancing the robustness and generalization capabilities of the model algorithm.

Regarding the model’s accuracy and robustness, we will introduce objective evaluation methods and metrics to verify the effectiveness of this approach. We plan to use a dual approach combining qualitative and quantitative assessments to evaluate the framework we proposed. For the qualitative assessment, we plan to conduct surveys of visitors through questionnaires and semi-structured interviews. Participants will be asked to provide an intuitive assessment of the realism and accuracy of the reconstructed models. Our questionnaire will guide participants to evaluate the 3D models based on perceived realism and accuracy, specifically focusing on overall appearance, detail fidelity, and structural consistency. Additionally, we plan to include questions regarding the impact of the 3D models generated by StatueDiff on the MfA’s exhibition appeal and engagement to measure the system’s overall integration within its exhibition environment. For the quantitative evaluation, we plan to use established measures such as chamfer distance and normal consistency. These measures will allow us to perform rigorous quantitative analysis of the geometric accuracy and detail fidelity between the reconstructed models and the corresponding 3D scanned data of the statues. With this approach, with aim is to effectively quantify the structural similarity between our generated 3D output and the actual artefacts.

Acknowledgements

The authors would like to express their sincere gratitude to the Museum für Abgüsse Klassischer Bildwerke (MfA) in Munich, Germany, for granting access to their sculpture collection and permitting data acquisition for this research. Special thanks to Dr. Nele Schröder-Griebel for her kind assistance during the data collection process.

References

Bell, T., Li, B., Zhang, S., 1999. Structured light techniques and applications. *Wiley encyclopedia of electrical and electronics engineering*, 1–24.

Bollo, A., Dal Pozzolo, L., 2005. Analysis of visitor behaviour inside the museum: An empirical study. *Proceedings of the 8th international conference on arts and cultural management*, 2, Citeseer, 1–13.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., Shah, M., 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9), 10850–10869.

Deng, C., Jiang, C., Qi, C. R., Yan, X., Zhou, Y., Guibas, L., Anguelov, D. et al., 2023. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20637–20647.

Fowler, B. H., 1989. *The Hellenistic Aesthetic*. Univ of Wisconsin Press.

França, J. G. D., Gazziro, M. A., Ide, A. N., Saito, J. H., 2005. A 3d scanning system based on laser triangulation and variable field of view. *IEEE International Conference on Image Processing 2005*, 1, IEEE, I–425.

Gomes, L., Bellon, O. R. P., Silva, L., 2014. 3D reconstruction methods for digital preservation of cultural heritage: A survey. *Pattern Recognition Letters*, 50, 3–14.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840–6851.

Karras, T., Aittala, M., Aila, T., Laine, S., 2022. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35, 26565–26577.

Laurentini, A., 1994. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 16(2), 150–162.

Lercari, N., Busacca, G., 2020. A glimpse through time and space: Visualizing spatial continuity and history making at Çatalhöyük, Turkey. *Journal of Eastern Mediterranean Archaeology & Heritage Studies*, 8(2), 99–122.

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C., 2023. Zero-1-to-3: Zero-shot one image to 3d object. *Proceedings of the IEEE/CVF international conference on computer vision*, 9298–9309.

Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M. J., 2020. Learning to dress 3d people in generative clothing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6469–6478.

Scharstein, D., Szeliski, R., 2003. High-accuracy stereo depth maps using structured light. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, 1, IEEE, I–I.

Schechner, Y. Y., Kiryati, N., 2000. Depth from defocus vs. stereo: How different really are they? *International Journal of Computer Vision*, 39, 141–162.

Song, Y., Ermon, S., 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Yoo, P., Guo, J., Matsuo, Y., Gu, S. S., 2023. Dreamsparse: Escaping from plato’s cave with 2d diffusion model given sparse views. *Advances in Neural Information Processing Systems*, 36, 3307–3324.

Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y., 2021. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5746–5756.

Zaramella, M., Amerini, I., Russo, P., 2023. Why don’t you speak?: A smartphone application to engage museum visitors through deepfakes creation. *Proceedings of the 5th Workshop on analySis, Understanding and proMotion of heritAge Contents*, 29–37.