èkphrasis

46th International Conference
of Representation Disciplines Teachers
Congress of Unione Italiana per il Disegno

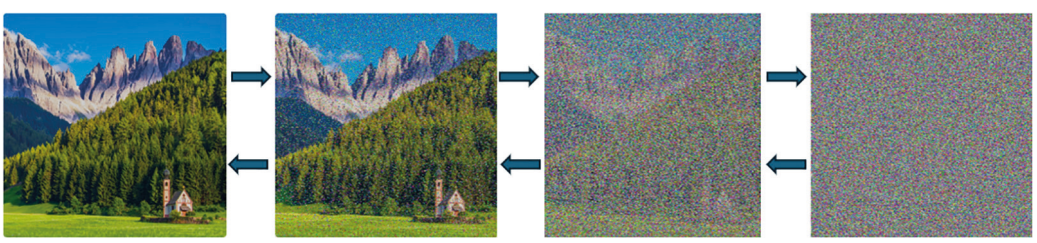
3D Models from Text Descriptions: Using Artificial Intelligence for Representation of Cultural Heritage

Francesca Condorelli

Abstract

This study explores the use of AI (Artificial Intelligence) to generate 3D models from textual descriptions and single-view image datasets, with the aim of digitally representing cultural heritage. The proposed approach combines advanced deep learning technologies, such as GAN and neural NeRF, to create 3D reconstructions from visual and textual data of limited quality and number. The use of textual descriptions overcomes the shortage of visual data and facilitates the reconstruction of architectural structures that are inaccessible because they no longer exist, have been destroyed or are only illustrated, offering a new way of digitally preserving architectural heritage. This innovative approach opens up new opportunities for the preservation and enhancement of cultural heritage, making it possible even in cases where historical structures are inaccessible.

Keywords
Text-to-3D, NeRF, Gaussian Splatting, Single Image dataset, Heritage representation.



Schematic diagram of the operation of Diffusion Models algorithms (elaboration by the author).

doi.org/10.3280/oa-1430-c893

Introduction

In the field of architectural representation, the use of three-dimensional models is widely adopted as a means of visualizing and understanding built environments. Despite technological advances, several challenges remain in creating these models, especially in complex scenarios. One of these arises when traditional surveying techniques fail due to insufficient data for 3D modeling. This problem is particularly evident in cases where photogrammetry is ineffective due to lack of available images or, more critically, when acquisition of new data is impossible because the architecture in question no longer exists, has been destroyed or is inaccessible. However, these lost or inaccessible structures often hold immense cultural and historical significance, making their documentation essential for heritage preservation. Traditional architectural documentation has relied heavily on fragmentary textual descriptions, drawings and photographic records that, while valuable, often lack the necessary spatial and volumetric accuracy required for complete digital reconstructions. The absence of complete data sets is a significant obstacle for researchers and conservators who wish to digitally recreate structures with high fidelity. In addition, the inability to conduct direct surveys of such sites necessitates the use of alternative methodologies to generate three-dimensional reconstructions.
This paper envisions a future in which historical architectural descriptions found in documents and treatises can be used as hints in text-to-image, Artificial Intelligence (AI)-based algorithms functioning as a modern form of *èkphrasis*. However, the current state of algorithm development is not yet sufficiently advanced to support this approach, as there are no trained datasets specifically designed for this type of input. Consequently, the first step of this research is to conduct preliminary tests using available open-source algorithms capable of generating 3D models by combining single images with short textual descriptions.
This work aims to assess the feasibility and limitations of current AI methodologies, while laying the groundwork for future applications that will fully integrate historical textual descriptions into AI-driven reconstruction processes. This approach not only redefines traditional means of architectural documentation, but also introduces a new path for the analysis of lost heritage through digital means. Using state-of-the-art artificial intelligence models, this study demonstrates how the synthesis of linguistic and visual data can improve historical understanding, enabling researchers to generate 3D reconstructions of architecture that would otherwise be lost over time.

Artificial intelligence-generated 3D reconstructions as modern *èkphrasis*

Advances in deep learning, particularly in Generative Adversarial Networks (GANs) and Neural Radiance Fields (NeRFs), make it possible to create realistic 3D models using both short prompts and datasets consisting of a small number of data or even single images. Recent developments in text-to-3D methodologies, such as Gaussian Splatting and NeRF-based reconstruction techniques, have shown promising results in generating architectural representations from minimal input data. The advantages of these new techniques include

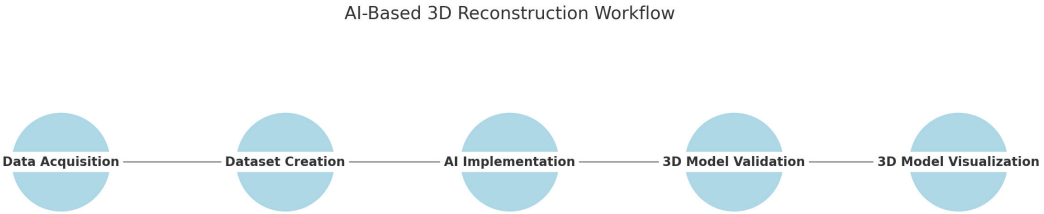


Fig. 1. Workflow of the proposed methodology (elaboration by the author).

the ability to generate high-quality 3D scenes with detailed geometries and photorealistic textures without the need for large datasets by exploiting pre-trained models on images. Referring to the latest text-to-image algorithms such as *Text2Nerf* [Zhang 2024] and *Magic3D* [Lin 2023] that exploit NeRF technology, and *GaussianDreamer* [Yi 2023] and *3DTopia* [Hong 2023] have the advantage of allowing rapid model generation, reducing creation time to only a few minutes in some cases, which is critical for applications from a computational perspective. Some approaches also offer the ability to customize 3D models efficiently, as in the case of *DreamBooth3D* [Raj 2023], which allows customization of specific objects from just a few examples. Techniques such as *DreamEditor* [Zhuang 2023] and *UniDream* [Li 2024] also solve consistency problems in lighting and geometry, improving visual quality and the ability to make changes with greater realism. From a perspective of applications of these algorithms in the field of architecture and cultural heritage, recent research has been concerned with using AI and text-to-image to create innovative architectural forms and new types of buildings, combining human and artificial creativity [Bono 2024; Horvath 2024], to increase the variety of art datasets and improve the understanding and description of works [Cioni, 2023], and for the restoration of damaged paintings by combining inpainting and text-to-image to reconstruct missing areas with artistic coherence [Nurmala Sari 2024], visually reconstruct destroyed cultural sites by integrating traditional preservation practices [Arzomand 2024].
However, despite these advances, several limitations still exist. Models continue to depend on high-resolution datasets that require a great deal of initial preparation work. Generating 3D models with complex geometries and higher diversity is still a challenge,

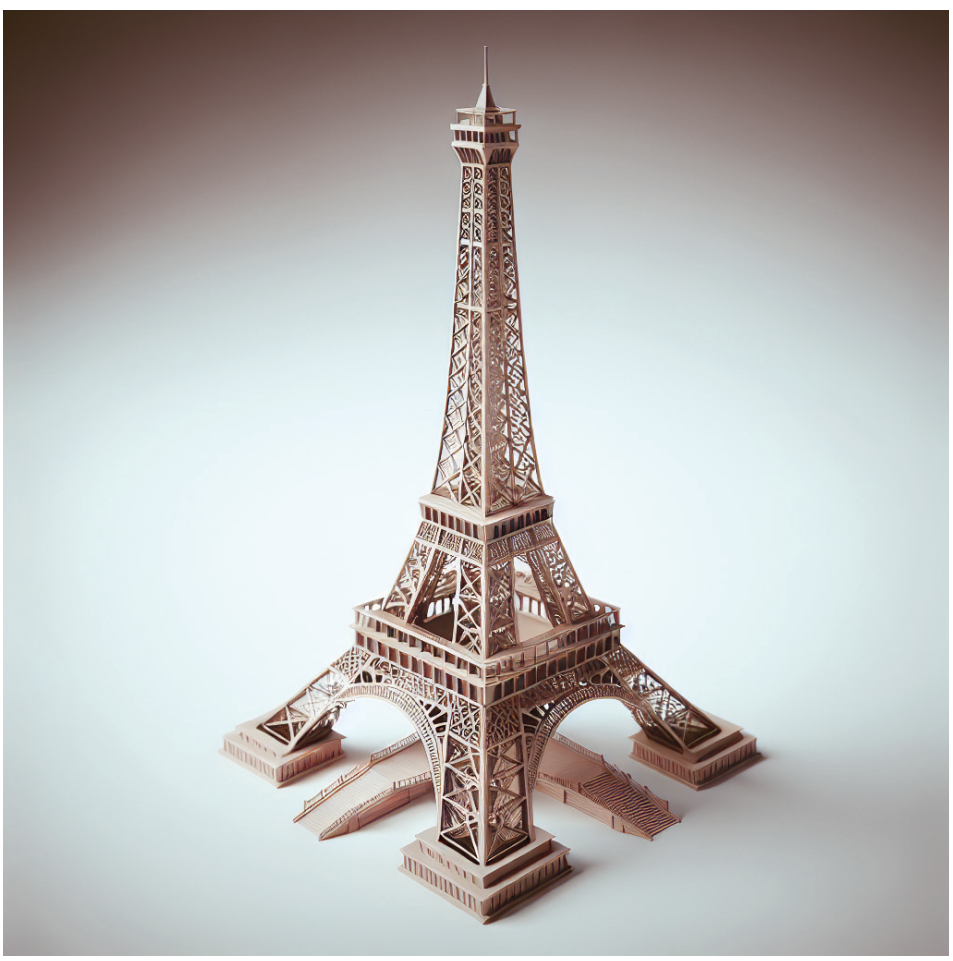


Fig. 2. Tex-to-3D result using the "Tour Eiffel" prompt (elaboration by the author).

especially when fully realistic and variable results are sought in complex scenarios. Despite the use of advanced 2D diffusion models, the lack of true "3D awareness" in generations remains a major limitation, leading to difficulties in retrieving detailed geometries. In addition, some approaches still fail to ensure consistent geometry and texture generation across different views or under different lighting conditions. The first major problem lies in the difficulty these tools have in understanding a language appropriate for architectural analysis; the results, understood as both the quality of architectural images and models, are therefore affected by this limitation [Alyidiz 2023; Albaghajati 2023]. The second problem is the lack of a dataset specifically created with artistic and architectural purposes, which makes it difficult to use prompts for creating 3D models

Fig. 3. 3D model result of Brixen Cathedral obtained from single photo (elaboration by the author).

referring to cultural heritage descriptions. For this reason, previous studies have been concerned with developing benchmarks to assess how text-to-image models represent diverse cultures, improving the generation of culturally accurate images [Liu 2024] and to assess cultural awareness and diversity in text-to-image models, highlighting shortcomings in the representation of global cultures [Kannen 2024].
Yet these approaches certainly offer new avenues for reconstructing architectural heritage that is in difficult accessibility conditions, and this paper is intended to be the beginning of a study that will be carried forward into future research to exploit the advantages of AI algorithms to aid 3D reconstructions in difficult situations where the available dataset is not adequate for standard applications because it consists of images of only one view of the object to be reconstructed and text. Initial tests carried out to achieve this goal are reported in the following sections.

Methodolgy and results

The methodology adopted in this research consists of a multi-step process that integrates artificial intelligence models based on deep learning, textual descriptions, and available visual data to generate 3D architectural reconstructions. The process is developed as described next and shown in figure (fig. 1).
First, extensive data acquisition work was carried out. Individual photographs, illustrations and visual references of the architectural object under consideration were collected. At

Fig. 4. Result tex-to-3D using the prompt "cathedral of Brixen" (elaboration by the author).

the same time, textual descriptions were extracted from historical documents, architectural treatises and academic articles to enrich the dataset. In addition, metadata such as material properties, dimensions, and the environmental context were integrated where possible to provide a more complete and accurate view.
Next, a data pre-processing phase was performed. Image enhancement techniques were applied to optimize clarity, remove noise and improve contrast in order to make the images more interpretable by AI models. Data normalization ensured compatibility between different input sources, allowing AI models to consistently interpret heterogeneous formats.
For 3D reconstruction, NeRF-based models were used, which generate volumetric representations of the structure by estimating missing spatial details from single image inputs. Gaussian platting techniques [Chen 2024] were employed to refine the estimation of depth and texture of materials, improving realism and spatial coherence. In addition, GAN models [Li 2023] were used to fill gaps in the data by synthesizing missing architectural elements based on contextual knowledge. Specifically, the implementation was performed by leveraging the Threestudio platform [Guo 2023], which integrates different algorithms such as *Dream3D* [Xu 2023].The next phase involved the validation and refinement of the AI-generated 3D models. These models were compared with existing designs and known architectural references to assess their accuracy. The validated 3D models can be integrated and visualized on interactive platforms, making the reconstructions accessible for academic, educational and heritage preservation applications [Condorelli 2023].
The following are the results of the first tests conducted on several case studies reflecting different situations in which cultural heritage may find itself. The first case concerns the Eiffel Tower (fig. 2), one of the world's most famous monuments, which, being a universal and already well-documented symbol, is certainly present in the datasets used for training artificial intelligence models. Its presence in numerous archives and digital resources makes it easier to create accurate and photorealistic 3D models, as detailed information about its structure, materials and architectural features can be drawn upon.
The second case is the Brixen Cathedral (fig. 3), an example of a monument that, although existing, certainly has fewer visual references than the Eiffel Tower. It was chosen to assess the differences between the two monuments.
Finally, the third case concerns the Tower of Babel (fig. 5), a monument that exists exclusively in the collective imagination of many cultures and that, while often depicted

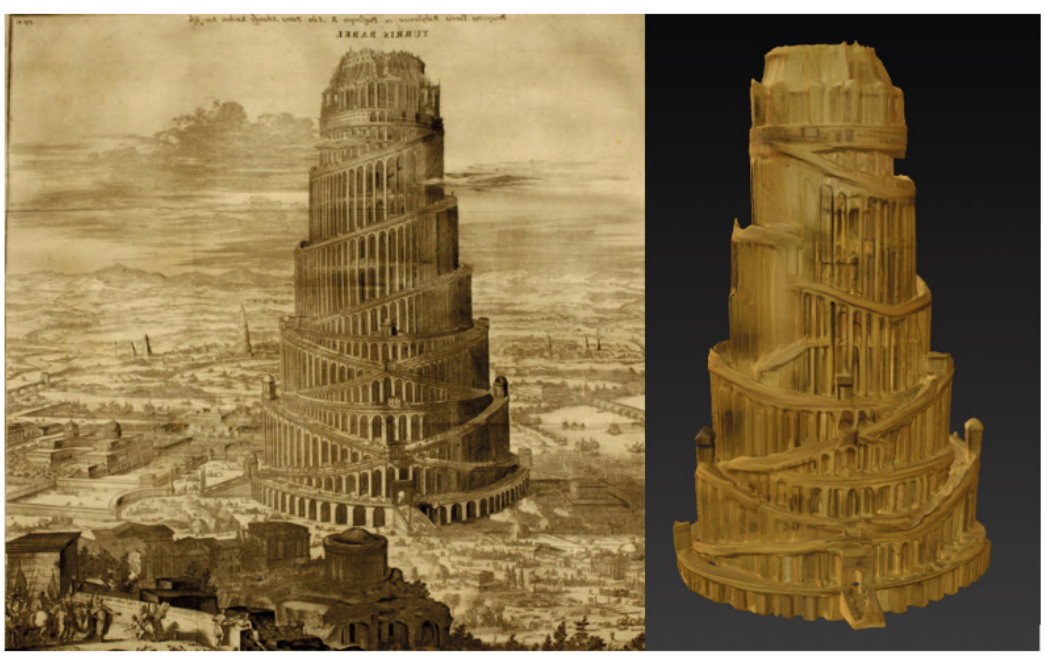


Fig. 5. Result of the 3D model of the Tower of Babel obtained from single photo (elaboration by the author).

in various artistic and literary forms, has never physically existed. The Tower of Babel presents one of the most complex challenges for digital reconstruction, as there is no concrete or physical data on its structure. However, artistic representations and historical descriptions offer cues for the creation of models that attempt to reflect different interpretations of its form, offering a visual representation that fits the many ways it has been imagined throughout history.
Results obtained from initial tests using simple textual descriptions combined with a single view image as a reference for 3D reconstruction led to the generation of models of varying formal accuracy. Although these models are affected by inaccuracies mainly due to the limitations of the initial dataset, the quality of the reconstructions is nevertheless satisfactory. The imperfections found are largely attributable to the lack of complete and accurate data, but, despite this, the generated models provide a useful and consistent visual representation of the studied object. Particularly in cases of monuments such as towers, which exhibit some structural symmetry, reconstruction is easier, as the regular geometric features allow the algorithm to infer missing details more accurately. This especially helps in complex scenarios such as historic towers, where, despite the paucity of data, symmetry can serve as a guide for plausible reconstruction. However, it is clear that the final quality depends very much on the type of monument and the availability of source data. For example, as hypothesized, the Eiffel Tower is certainly present in the datasets used by AI for training and in fact its reconstruction is quite faithful to reality.

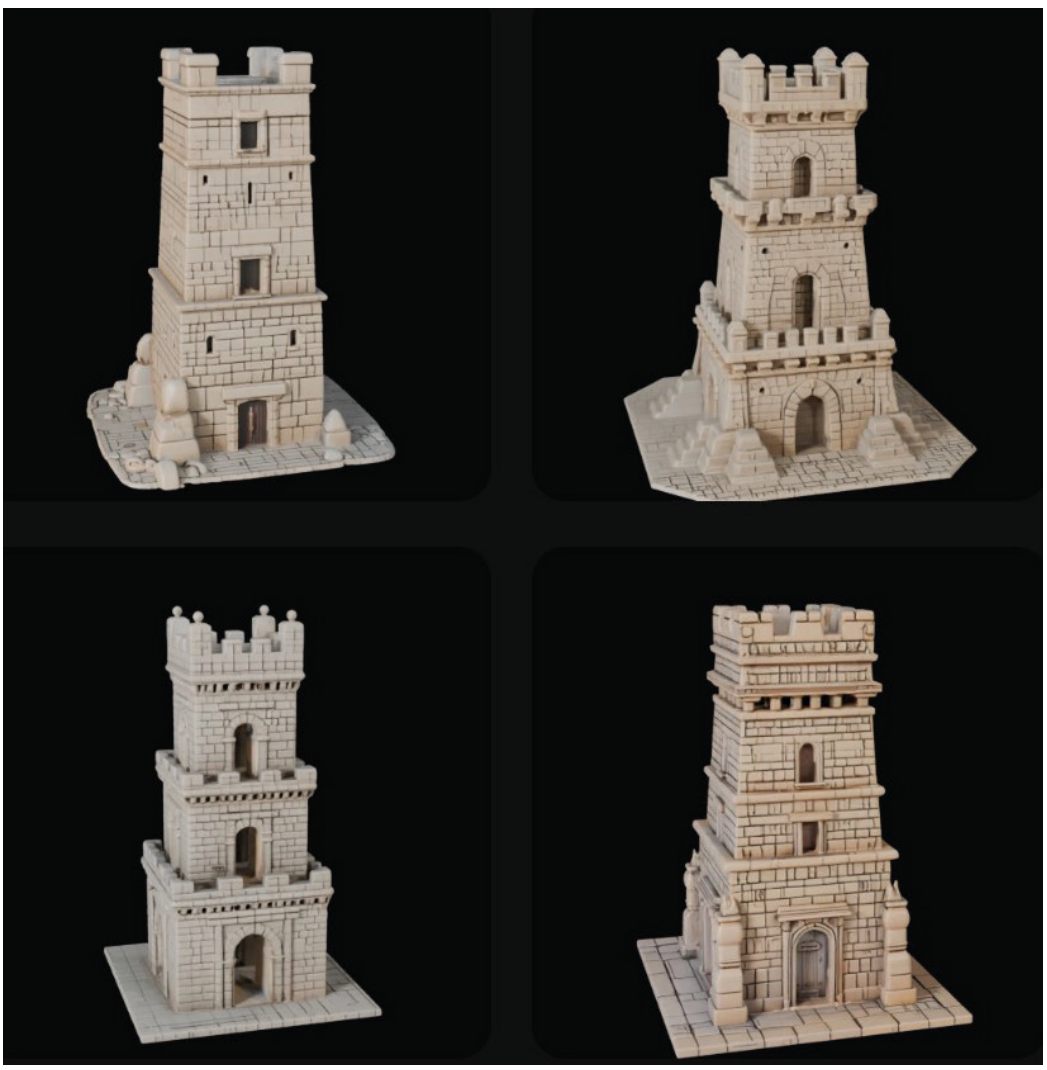


Fig. 6. Result tex-to-3D using the "Tower of Babel" prompt (elaboration by the author).

With regard to the Brixen Cathedral, it is evident that there are references in the datasets to churches that feature architecture typical of South Tyrol (fig. 4) albeit different from the original Cathedral, which is obviously unique. As for the Tower of Babel there are too many references and interpretations in the literature and it was difficult to obtain an unambiguous model (figs. 5, 6).
These are only the first tests, and it is emphasized that in cases where there are insufficient physical or photographic records, the generation of 3D models through the use of textual descriptions is the only possible way to document and preserve cultural heritage. It is essential to continue working on expanding and improving datasets, as well as optimizing algorithms, to achieve more precise and accurate results. Despite these challenges, the results obtained so far are very satisfactory, considering the inherent difficulties of this type of work and the highly challenging context.

Conclusions: the future of *èkphrasis* in digital reconstruction

In conclusion, this research demonstrates the potential of artificial intelligence techniques in generating 3D models from textual descriptions, opening up new possibilities for the documentation and preservation of cultural heritage.
The results obtained, although still partial and influenced by inaccuracies related to the datasets used, confirm that this methodology offers a valuable tool for reconstructing historical structures, even in the absence of direct data or complete documents. In particular, the combination of textual descriptions with reference images has resulted

Fig. 7. Result tex-to-3D using the prompt "tower of Babel with the bristling scaffolding at the top of the structure, and the houses for the workers built on the ramps lower down, to a procession of camels and other beasts being led towards the main entrance" (elaboration by the author).

in reconstructions of acceptable accuracy, especially for monuments with regular geometries such as towers. However, the generated models still require improvements, both in terms of formal accuracy and diversification of input data, to achieve optimal levels of accuracy. The symmetry of some structures certainly facilitated the reconstruction process, but in other cases the limitations of the dataset affected the final quality. Therefore, it is necessary to continue to refine both the algorithms and the datasets in order to overcome the challenges related to the scarcity of data and the variability of architectural forms.

Despite the challenges, the results achieved are promising and represent an important step toward the use of AI as a tool for cultural heritage preservation. Indeed, this methodology offers a valuable solution for documenting and preserving historic buildings that would otherwise be in danger of being forgotten, offering new possibilities for historic preservation and exploration. In the future, there is a need to focus on continuous expansion and improvement of datasets, as well as refinement of generation techniques, to achieve even more satisfactory results. As AI continues to evolve, its role in architectural representation will not only enrich our understanding of the past, but also help redefine the very concept of *èkphrasis* in the digital age, giving new shape and meaning to our historical perception.

Reference List

Albaghajati, Z.M., Bettaieb, D.M. & Malek, R.B. (2023). Exploring text-to-image application. In *Architectural Structures and Construction*, n. 3, pp. 475-497. https://doi.org/10.1007/s44150-023-00103-x.

Alyıldız, C. (2023). Generative text-to-image models in architectural design: A study on relationship of language, architectural quality and creativity. In *ICONTECH International Journal*, 7(3), 12-26.

Arzomand, K., Rustell, M., & Kalganova, T. (2024). From ruins to reconstruction: Harnessing text-to-image AI for restoring historical architectures. In *Challenge Journal of Structural Mechanics*, 10(2), pp. 69-85. http://dx.doi.org/10.20528/cjsmec.2024.02.004.

Bono, G. (2024). Text-to-building: experiments with AI-generated 3D geometry for building design and structure generation. *ARIN* 3, 24 (2024). https://doi.org/10.1007/s44223-024-00060-5.

Chen, Z., Wang, F., Wang, Y., & Liu, H. (2024). Text-to-3D using Gaussian Splatting. In *Proceedings Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, pp. 21401-21412. https://doi.org/10.1109/CVPR52733.2024.02022.

Cioni, D., Berlincioni, L., Becattini, F. (2023). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops,* pp. 1707-1716.

Condorelli, F., Luigini, A., Tramelli, B. (2023). Digital Turris Babel. Augmented release of Athanasius Kircher's Archontologia, In A. Giordano, M. Russo, R. Spallone, (Eds). *Beyond Digital Representation. Digital Innovations in Architecture, Engineering and Construction*. Cham: Springer. https://doi.org/10.1007/978-3-031-36155-5_6.

Guo, Y. C., Liu, Y.T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.-H., Zou, Z.-X., Wang, C., Cao, Y.-P., Zhang, S.-H. (2023). *Threestudio: A unified framework for 3D content generation*. https://github.com/threestudio-project/threestudio.

Hong, F., Tang, J., Cao, Z., Shi, M., Wu, T., Chen, Z., Yang, S., Wang, T., Pan, L., Lin, D., & Liu, Z. (2023). *3DTopia: Large text-to-3D generation model with hybrid diffusion priors*. arXiv. https://doi.org/10.48550/arXiv.2403.02234.

Horvath, A. S., & Pouliou, P. (2024). AI for conceptual architecture: Reflections on designing with text-to-text, text-to-image, and image-to-image generators. In *Frontiers of Architectural Research*, 13(3), pp. 593-612. https://doi.org/10.1016/j.foar.2024.02.006.

Kannen, N., Ahmad, A., Andreetto, M., Prabhakaran, V., Prabhu, U., Dieng, A. B., Bhattacharyya, P. (2024). *Beyond aesthetics: Cultural competence in text-to-image models*. arXiv. https://doi.org/10.48550/arXiv.2407.06863.

Li, C., Zhang, C., Cho, J., Waghwase, A., Lee, L.-H., Rameau, F., Yang, Y., Bae, S.-H., & Hong, C. S. (2023). *Generative AI meets 3D: A survey on text-to-3D in the AIGC era*. arXiv. https://doi.org/10.48550/arXiv.2305.06131.

Li, H., Shi, H., Zhang, W., Wu, W., Liao, Y., Wang, L., Lee, L.-H., & Zhou, P.Y. (2024). DreamScene: 3D Gaussian-based text-to-3D scene generation via formation pattern sampling. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer Vision ECCV 2024. Lecture Notes in Computer Science*, n. 15132, pp. 214-230. Cham: Springer. https://doi.org/10.1007/978-3-031-72904-1_13.

Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., & Liu, M.-Y. (2023). Magic3D: High-Resolution Text-to-3D Content Creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 300-309. https://doi.org/10.48550/arXiv.2211.10440.

Liu, B., Wang, L., Lyu, C., Zhang, Y., Su, J., Shi, S., et al. (2024). On the cultural gap in text-to-image generation. In *Frontiers in Artificial Intelligence and Applications*, 392, pp. 930-937. https://doi.org/10.3233/FAIA240581.

Nurmala Sari, I., Sugahara, R., & Du, W. (2024). Artistic outpainting through adaptive image-to-text and text-to-image generation. In *Proceedings of the 2024 10th International Conference on Computing and Artificial Intelligence*, pp. 20-25. Association for Computing Machinery, New York.. https://doi.org/10.1145/3669754.3669758.

Raj, A., Kaza, S., Poole, B., Niemeyer, M., Ruiz, N., Mildenhall, B., Zada, S., Aberman, K., Rubinstein, M., Barron, J., Li, Y., Jampani, V. (2023). DreamBooth3D: Subject-Driven Text-to-3D Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris (France) 2-6 October 2023, pp. 2349-2359. https://doi.org/10.1109/ICCV51070.2023.00223

Xu, J., Wang, X., Cheng, W., Cao, Y.-P., Shan, Y., Qie, X., Gao, S. (2023). Dream3D: Zero-Shot Text-to-3D Synthesis Using 3D Shape Prior and Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver (Canada) 17-24 June 2023, pp. 20908-20918. https://doi.org/10.1109/CVPR52729.2023.02003.

Yi, T., Fang, J., Wang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X. (2023). *GaussianDreamer: Fast generation from text to 3D Gaussians by bridging 2D and 3D diffusion models*. https://doi.org/10.48550/arXiv.2310.08529

Zhang, J., Li, X., Wan, Z., Wang, C., Liao, J. (2024). Text2NeRF: Text-driven 3D scene generation with neural radiance fields. In *IEEE Transactions on Visualization and Computer Graphics*, 30(12), pp. 7749-7762. https://doi.org/10.1109/TVCG.2024.3361502.

Zhuang, J., Wang, C., Lin, L., Liu, L., Li, G. (2023). DreamEditor: Text-driven 3D scene editing with neural fields. In *SIGGRAPH Asia 2023 Conference Papers* (SA '23). Association for Computing Machinery, New York. https://doi.org/10.1145/3610548.3618190.

Author
Francesca Condorelli, Free University of Bozen, francesca.condorelli@unibz.it

To cite this chapter: Francesca Condorelli (2025). 3D Models from Text Descriptions: Using Artificial Intelligence for Representation of Cultural Heritage. In L. Carlevaris *et al.* (Eds.). *èkphrasis. Descrizioni nello spazio della rappresentazione/èkphrasis. Descriptions in the space of representation.* Proceedings of the 46th International Conference of Representation Disciplines Teachers. Milano: FrancoAngeli, pp. 2669-2678. DOI: 10.3280/oa-1430-c893.

ISBN 9788835182412