Beyond human imagination: The art of creating prompt-driven 3D scenes with Generative AI

Giulio Federico, Fabio Carrara, Giuseppe Amato, and Marco Di Benedetto

CNR-ISTI, Pisa, Italy

Corresponding author: giulio.federico@isti.cnr.it

Keywords: Generative AI, Computer Graphics, Denoising Diffusion Probabilistic Model, Gaussian Splatting, NeRF, Signed Distance Field, Video Reconstruction, Deep Learning, Machine Learning, Artificial Intelligence, Text-to-3D, Image-to-3D, Urban Environment, Score Distillation Sampling

Extended Abstract

The reconstruction of large-scale real outdoor environments is crucial for promoting the adoption of Extended Reality (XR) in industrial and entertainment sectors. This task often requires significant resources such as depth cameras, LiDAR sensors, drones, and others, alongside traditional data processing pipelines like Structure-from-Motion (SfM), which demand extensive computational resources, thus preventing real-time processing. Additional constraints arise from the limited accessibility to the aforementioned resources. While 3D laser scanners (e.g., LiDAR) are precise and fast, they are expensive, often bulky – especially the high-quality models – and their effectiveness is contingent on the type of environment being scanned. Depth sensors offer a more affordable and compact alternative; however, due to their limited range, they are ideal only for indoor settings. Photogrammetry, while capable of producing high-quality results at a lower cost, can be time-consuming and computationally intensive. It also suffers from limited accuracy, strong dependence on lighting conditions, and the need for numerous photos from various angles that can be not always easily accessible.

To address these limitations, we initially proposed a Spatio-Temporal Diffusion neural architecture (Federico et al., 2024), a generative architecture based on diffusion models. This solution integrates simple and cost-effective temporal information (a brief temporally ordered sequence of photographs) with spatial information (a rough approximation of the environment to be reconstructed) to rapidly reconstruct complex 3D environments, filling in missing or noisy information. The use of a neural architecture stems from the need to achieve real-time processing, while the application of generative artificial intelligence serves to compensate for the lack of information, arising from the absence of access to costly resources or situations where certain data are unattainable (e.g., an unreachable viewpoint). We also introduced a novel 3D representation termed the Most Informative Part (SDF_MIP), a modification of the well-known Signed Distance Field (SDF) that aims to symmetrically distribute positive and negative voxels — a requirement we identified as essential during network training, particularly for

outdoor environments. Our model comprises a two-stage network: the first stage fuses temporal and spatial information to generate the missing data Figure P 29, left), while the second stage converts the SDF_MIP representation back to SDF (Figure P 29, right). An optimal trade-off between reconstruction quality and execution speed was achieved using the DDIM scheduler.

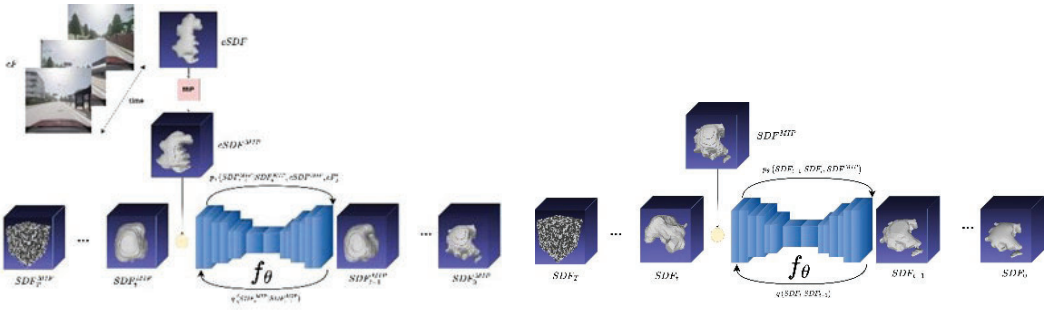


Figure P 29. Stage 1 (left) and Stage 2 (right) of our Spatio-Temporal Diffusion Neural Architecture.

Despite encouraging results, the reconstruction of complex environments required prohibitively long training times and high computational costs, failing to overcome the aforementioned limitations. Furthermore, the proposed solution does not scale well with the resolution and complexity of the target environment. We are currently developing a generative network architecture called Neural-Clipmap and an associated algorithm that alleviates training burdens, enables scalability, and minimizes computational and memory requirements. Specifically, the algorithm hallucinates and build a high quality version of a complex environment with a divide and conquer strategy by enhancing the structure of the underlying supporting octree, where each leaf is an atomic unit of computation (Figure P 30). Supported by the generative network (the diffusion one) (Ho et al., 2020), it determines whether a leaf node requires a coarsening operation (i.e., the input leaf is overly detailed and should be removed, refinement (i.e., the leaf requires further detailing), or no operation at all. The algorithm operates in two iterative phases (Figure P 30). In the first phase, for each leaf of the coarse octree, the generative network uses contextual information to modify it: a series of frames of color images of an hypothetical actor driving around a path, and the spatial neighbors of the leaf. The latter are taken at multiple levels. In particular, given a leaf, its spatial neighbors correspond to the nodes around it and, going up a level, those around the leaf's parent and so on up to the root. These two phases dynamically allow the octree levels to be pruned or increased. The generative network leverages the contextual information from our initial work (Federico et al., 2024), but it now employs a video vision encoder Arnab et al., 2021) to create a compact representation of the RGB frames. In this phase, we encountered challenges with training when using SDF or SDF_MIP, as the network struggled to appropriately correlate the contextual information. Consequently, we adopted the Triplane representation Chan et al, 2022), inspired by recent successes in neural representations, yielding more promising preliminary results.

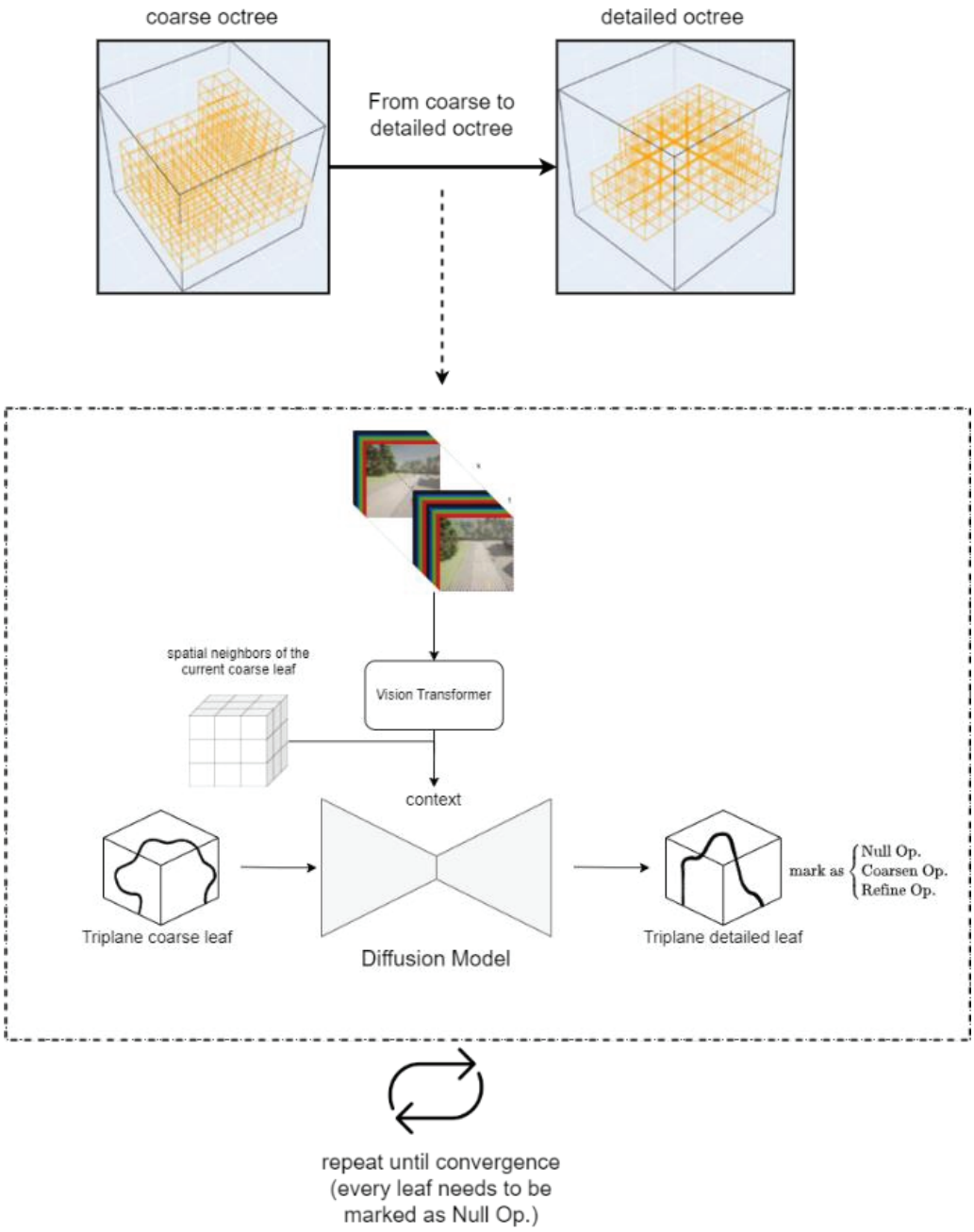
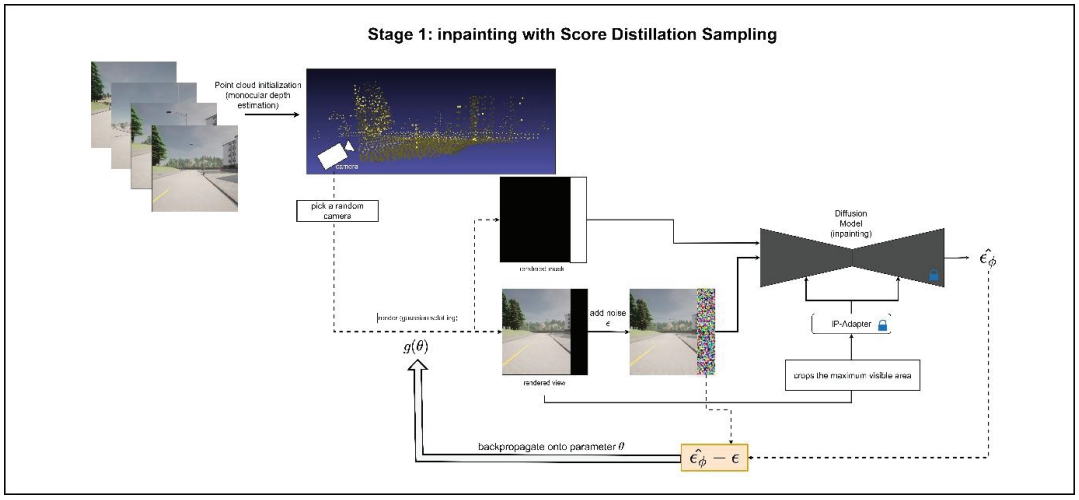


Figure P 30. The Neural-Clipmap algorithm progressively converts a coarse octree into a detailed octree.

However, generating the Triplane representation for every leaf of the octree and for each model in our dataset results in a frustrating delay. Additionally, increasing the model resolution exponentially raises the number of leaves, thereby contradicting our goal of easy scalability with resolution. Concurrently, we are exploring the use of 2D diffusion models (Ho et al., 2020) as priors for

reconstructing complex 3D environments. In this new attempt we minimize the input data by using only sequences of RGB images. Through a monocular depth estimator (Yang et al., 2024), we generate an initial point cloud of the environment, serving as the initialization for a 3D Gaussian Splatting representation (Kerbl et al., 2023). However, these points represent partial information, as the RGB image sequence covers only a portion of the environment. Thus, we employ 2D diffusion models as priors, utilizing a recent technique known as Score Distillation Sampling (SDS) (Poole et al., 2022) to reconstruct the missing information by moving the camera to points of interest. Unlike other works based on this technique (Lin et al., 2023; Tang et al., 2024; Liu et al., 2023) our challenge lies in starting from images rather than text and ensuring the consistency of the generated missing information with the existing data (termed "anchors"). SDS has been employed to generate simple 3D models and, despite this, suffers from various issues such as the Janus problem (Armandpour et al., 2023) difficulty in determining the appropriate guidance value, and overly saturated or blurred colors. To compensate for the lack of text guidance for the SDS, we used an image prompt adapter Ye et al., 2023). To tackle the issues of overly saturated or blurred colors, we devised a two-phase approach. The first phase uses the SDS (with an inpainting diffusion model) for initializing the missing parts, which may exhibit the aforementioned problems. The subsequent phase involves inpainting over the areas initialized by the SDS. For inpainting to work effectively, the missing parts require some initialization coherent with the real information available. Classical approaches include initializing the missing region with the average color of the real data, Perlin Noise, or, though slower, using an algorithm known as Patch Match (Connelly et al., 2009). We propose using SDS (Figure P 31, top) as a new method for initializing the missing areas and employing inpainting in the next phase (Figure P 31, below) to mitigate issues with saturated and blurred colors. Stage 1 is crucial not only for initializing the missing regions but also for performing multi-view inpainting in a manner consistent with other views, which would be unfeasible if starting directly from Stage 2. Furthermore, we could introduce a third Refine phase, where we add some noise to the rendered views and subsequently employ a standard diffusion model to denoise them, thereby eliminating any residual noise that may persist after Stage 2. For better quality results, the diffusion model for the refinement phase will likely be fine-tuned using the anchor images with the prior preservation technique (Ruiz et al., 2023)



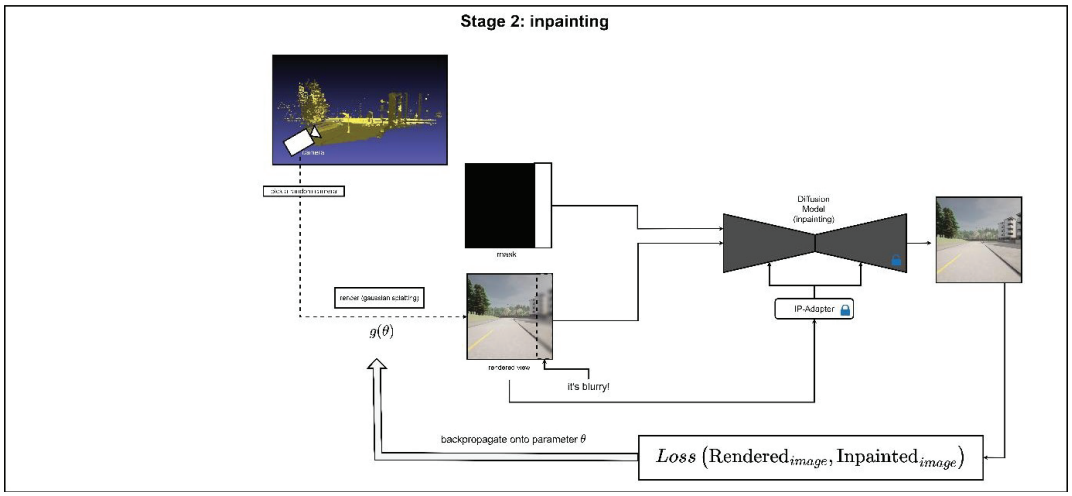


Figure P 31. Starting from a sequence of images (the anchors), a monocular depth estimator is used to initialize the point cloud and then gaussian splatting is used to render the views and optimize its parameters via Score Distillation Sampling (Stage 1, above). The inpainting model is then used to realistically fill the parts filled but blurred by Stage 1 (Stage 2, below)

References

Armandpour, M., Sadeghian, A., Zheng, H., Sadeghian, A., & Zhou, M. (2023). Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. arXiv preprint arXiv:2304.04968. https://arxiv.org/abs/2304.04968

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6836-6846).. https://arxiv.org/abs/2103.15691

Barnes, C., Shechtman, E., Finkelstein, A., & Goldman, D. B. (2009). PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph., 28(3), 24. 1–11. https://doi.org/10.1145/1576246.1531330

Chan, E. R., Lin, C. Z., Chan, M. A., Nagano, K., Pan, B., De Mello, S., ... & Wetzstein, G. (2022). Efficient geometry-aware 3d generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16123-16133). https://arxiv.org/abs/2112.07945

Federico, G., Carrara, F., Amato, G., & Di Benedetto, M. (2024, June). Spatio-Temporal 3D Reconstruction from Frame Sequences and Feature Points. In Proceedings of the 2024 ACM International Conference on Interactive Media Experiences Workshops (pp. 52-64). Association for Computing Machinery, New York, NY, USA, 52–64. https://doi.org/10.1145/3672406.3672415

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33, 6840-6851. https://arxiv.org/abs/2006.11239

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph., 42(4), 139-1. https://repo-sam.inria.fr/fungraph/3d- gaussian-splatting/

Lin, C. H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., ... & Lin, T. Y. (2023). Magic3d: High-resolution text-to-3d content creation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 300-309). https://arxiv.org/abs/2211.10440

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. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9298-9309). https://arxiv.org/abs/2303.11328

Nataniel Ruiz and Yuanzhen Li and Varun Jampani and Yael Pritch and Michael Rubinstein and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. https://arxiv.org/abs/2208.12242

Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. (2022). Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988. https://arxiv.org/abs/2209.14988

Tang, J., Ren, J., Zhou, H., Liu, Z., & Zeng, G. (2023). Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653. https://arxiv.org/abs/2309.16653

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10371-10381).https://arxiv.org/abs/2401.10891

Ye, H., Zhang, J., Liu, S., Han, X., & Yang, W. (2023). Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721. https://arxiv.org/abs/2308.06721