Towards Renderable Representations using Foundation Models

(Masters 2 or Pre-Doc internship, possible extension to a Ph.D.)

George Drettakis and Linus Franke, GRAPHDECO, Inria Sophia Antipolis (France)

http://team.inria.fr/graphdeco

George.Drettakis@inria.fr, http://www-sop.inria.fr/members/George.Drettakis/
Linus.Franke@inria.fr, https://lfranke.github.io

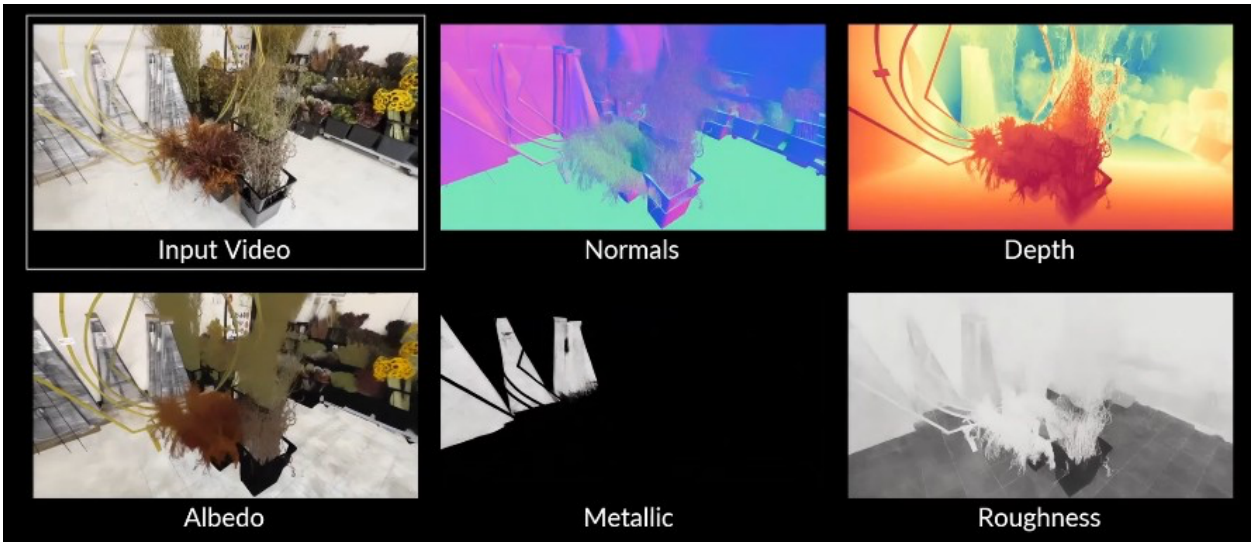

Figure 1: Diffusion Renderer [Liang et al. 2025] can predict per-pixel attributes such as normals, albedo, and shading directly from images or short video clips, but typically remain in image space and lacks a persistent and multi-view consistent 3D representation. Source: https://research.nvidia.com/labs/toronto-ai/DiffusionRenderer/

Context and goal

Creating 3D assets that can be rendered in graphics pipelines with full control over lighting, is a long-standing goal of Computer Graphics. The information of material properties and lighting needs to be disentangled from photographs, and then a 3D representation needs to be created to allow the captured asset to be used in downstream applications. This project is in this context.

Recent diffusion-based methods can infer rendering parameters such as normals or colors directly from images and short video clips. While techniques like Diffusion Renderer [Liang et al. 2025] or RGB<->X [Zeng et al. 2024] achieve strong image-space predictions, they do not reconstruct a persistent, multi-view consistent 3D representation and typically operate only on limited temporal windows.

This results in frame-wise inconsistencies and prevents producing 3D scenes that can be rendered. In contrast, geometry-aware approaches such as VGGT [Wang et al 2025] or world-consistent video diffusion [Zhang et al. 2025] demonstrate how multi-view or video supervision can yield point maps and depth estimates that are coherent across frames, but they lack material property information.

One major open challenge remains how to fuse per-frame diffusion outputs into a consistent 3D model that is usable for novel view synthesis and downstream rendering.

Approach

This project aims to design a hybrid pipeline that merges diffusion-based image/video predictions with geometry-grounded reconstruction. The intern will have access to local, Inria-wide and national GPU compute infrastructure.

One approach we will investigate would start with extracting per-frame outputs (normals, colors, or depth) using a diffusion prior and then unprojected into 3D using camera information. By leveraging VGGT-style point maps and multi-view constraints, these predictions could then be fused into a global representation, e.g., either as dense point clouds or volumetric radiance fields.

The project will address two major challenges of such an approach. First, to overcome video length limitations, the pipeline will explore temporal chunking and cross-frame harmonization in 3D space. Second, new rendering algorithms will be developed with the final goal to create a renderable 3D scene representation that combines the photometric fidelity of diffusion models with the geometric consistency of explicit reconstructions.

Work environment and requirements

The internship will take place at Inria Sophia Antipolis in the GRAPHDECO group (http://team.inria.fr/graphdeco) (the inventors of 3D Gaussian Splatting). The intern will have access to local, Inria-wide and national GPU compute infrastructure.

Candidates should be passionate about computer graphics and neural rendering methods, and have strong programming and mathematical skills. Knowledge in one or more of computer graphics, geometry processing and machine learning, experience in python, pytorch, cuda, C++, real-time rendering techniques, path-tracing (knowledge of mitsuba3 is a plus), OpenGL and GLSL on the graphics side are desirable.

How to apply

Applicants should either be Masters (5th year) students for an internship, or if applying for a pre-doc, they should already have a Masters degree in Computer Science, specialized in Computer Graphics and/or Computer Vision. Please email George.Drettakis@inria.fr with your CV, motivation letter and your transcripts for the last 2-3 years of study.

References

[Liang et al. 2025] Liang, R., Gojcic, Z., Ling, H., Munkberg, J., Hasselgren, J., Lin, Z.-H., Gao, J., Keller, A., Vijaykumar, N., Fidler, S., & Wang, Z. (2025, Juni). DiffusionRenderer: Neural Inverse and Forward Rendering with Video Diffusion Models. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) .

[Wang et al. 2025] Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., & Novotny, D. (2025). VGGT: Visual Geometry Grounded Transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition .

[Zeng et al 2024] Zeng, Z., Deschaintre, V., Georgiev, I., Hold-Geoffroy, Y., Hu, Y., Luan, F., Yan, L.-Q., & Hašan, M. (2024). RGB↔X: Image decomposition and synthesis using material- and lighting-aware diffusion models. ACM SIGGRAPH 2024 Conference Papers . https://doi.org/10.1145/3641519.3657445

[Zhang et al 2024] Zhang, Q., Zhai, S., Bautista, M. Á., Miao, K., Toshev, A., Susskind, J., & Gu, J. (2024). World-consistent Video Diffusion with Explicit 3D Modeling. arXiv .