

Generative Models for 3D Vision

Laura Neschen^{*1}, Bernhard Egger^{†2}, Adam Kortylewski^{†3},
William Smith^{†4}, and Stefanie Wuhler^{†5}

1 INRIA Rhône-Alpes, FR. laura.neschen@inria.fr

2 Friedrich-Alexander-Universität Erlangen-Nürnberg, DE.
egger.bernhard@gmail.com

3 Universität Freiburg, DE and MPI für Informatik – Saarbrücken, DE.
akortyle@mpi-inf.mpg.de

4 University of York, GB. william.smith@york.ac.uk

5 INRIA – Grenoble, FR. stefanie.wuhler@inria.fr

Abstract

Generative models that allow synthesis of realistic 3D models have been of interest in computer vision and graphics for over 2 decades. While traditional methods use morphable models for this task, more recent works have adopted powerful tools from the 2D image domain such as generative adversarial networks, neural fields and diffusion models, and have achieved impressive results. The question of which tools are most suitable for applications such as reconstructing 3D geometry from partial data, and creating digital 3D content remains open. This report documents the program and outcomes of Dagstuhl Seminar 25202 titled “Generative Models for 3D Vision”. This meeting of 25 researchers covered a variety of topics such as generative models and priors for 2D tasks, medical applications, and digital representations of humans, including how to evaluate and benchmark different methods. We summarise the discussions, presentations, and results of this seminar.

Seminar May 11–16, 2025 – <http://www.dagstuhl.de/25202>

2012 ACM Subject Classification Computing methodologies → 3D imaging; Computing methodologies → Computer vision; Computing methodologies → Computer graphics

Keywords and phrases 3D Computer Vision, Computer Graphics, Generative Models, Implicit Representations, Neural Rendering, Statistical Modelling

Digital Object Identifier 10.4230/DagRep.15.5.96

1 Executive Summary

Bernhard Egger (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE)

Adam Kortylewski (Universität Freiburg, DE and MPI für Informatik – Saarbrücken, DE)

Laura Neschen (INRIA Rhône-Alpes, FR)

William Smith (University of York, GB)

Stefanie Wuhler (INRIA – Grenoble, FR)

License  Creative Commons BY 4.0 International license

© Bernhard Egger, Adam Kortylewski, Laura Neschen, William Smith, and Stefanie Wuhler

The rise of purely data-driven generative models, in particular generative adversarial networks, auto-regressive models, neural fields and diffusion models, has led to a step change in image synthesis quality. It is now possible to create photorealistic images with high level semantic control and solve many desirable use cases such as 2D inpainting. Whilst prior models were

* Editorial Assistant / Collector

† Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Generative Models for 3D Vision, *Dagstuhl Reports*, Vol. 15, Issue 5, pp. 96–113

Editors: Laura Neschen, Bernhard Egger, Adam Kortylewski, William Smith, and Stefanie Wuhler



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

object specific (e.g. 3D Morphable Models of Faces), we now have generative models for images and videos that can represent various object classes and generate a huge variety of objects and scenes, even in different styles. The drawback of purely data-driven approaches is that the control and explainability provided by 3D and physically-based parameters is lost. It is also difficult (and perhaps prohibitively inefficient) to learn 3D consistent representations without prior models purely from 2D data alone.

For this seminar, a total of 58 researchers were invited, and 25 of them attended. Participants came from both academia and industry and at varying stages of their careers. Thirteen participants presented their work in around 15-30 minute presentations, and an abstract of each presentation is included in this report. We started the seminar with a short introduction of each participant. Everyone was given one slide to introduce themselves and asked to prepare a question, challenge or goal to discuss during the seminar.

In addition to traditional presentations, multiple slots were left for research discussions with the full group or sub-groups of the participants. The first set of these slots was filled with topics that participants proposed before the start of the seminar. Five participants led research discussions of about 1 hour each about a topic or a problem they considered important. Some of these discussions were led with the full group, while others were discussed in sub-groups, and the resulting conclusions were shared with the full group afterwards. Additionally, two 2 hour discussion slots were initially reserved to be filled with suggestions that came up during the seminar. These two long discussions concerned research questions that were identified as being important for the research community in the course of the seminar, namely the topics of metrics and capture, and hard problems in the research community that merit being studied more. All proposed topics led to lively discussions about various problems around generative models. Summaries of the results of these flexible sessions are contained in this report. In addition to these organized discussions, there were numerous informal discussions during both the Wednesday outing and free time slots that are not summarized in this report.

2 Table of Contents

Executive Summary

<i>Bernhard Egger, Adam Kortylewski, Laura Neschen, William Smith, and Stefanie Wuhler</i>	96
--	----

Overview of Talks

Synthetic Data for Generative AI <i>Thabo Beeler</i>	100
Learning to Infer Parametric Representations for 3D Plants from Scans <i>Samara Ghrer</i>	100
Statistical Approaches to Internal Anatomy Prediction <i>Marilyn Keller</i>	101
On Learning to Reconstruct Shape using World Priors, Handling Topological Inconsistencies, and Edge Integration as a Model of Choice <i>Ron Kimmel</i>	101
Towards Photorealistic 3D Head Avatars <i>Tobias Kirschstein</i>	103
Understanding Self-Supervised Learning <i>Adam Kortylewski</i>	103
Leveraging 3D Representations for Multi-view Synthesis and Editing with 2D Generative Models <i>Or Patashnik</i>	104
Scaling Digital Humans <i>Shunsuke Saito</i>	104
Clever Data Curation for Smart Supervision <i>William Smith</i>	104
Generating 3D Human Motion with Language <i>Gül Varol</i>	106
3D Foundation Models for Enhanced Geometry in Gaussian Splatting <i>Yaniv Wolf</i>	106
D-Garment: Physics-Conditioned Latent Diffusion for Dynamic Garment Deformations <i>Stefanie Wuhler</i>	107

Open problems

Interpolation of Shapes with Topological and Structural Variability <i>Andreea Ardelean and Samara Ghrer</i>	107
Exploring problems with infinite data <i>Timotei Ardelean</i>	109
Benchmarking Monocular Reconstruction Discussion <i>James Gardner</i>	109
Capturing useful datasets for 3D vision tasks <i>Samara Ghrer</i>	111

Believe Propagation over Spatial Domains with Generative Models
Jan Eric Lenssen 111

Hard problems in computer vision and the necessity of modeling
Yaniv Wolf. 112

Participants 113

3 Overview of Talks

3.1 Synthetic Data for Generative AI

Thabo Beeler (Google Research – Zürich, CH)

License © Creative Commons BY 4.0 International license
© Thabo Beeler

Inspired by the seminal “Fake It Til You Make It” paper published in 2021, people have successfully trained ML models on synthetic data for research and product alike, demonstrating that the infamous domain gap can be overcome, even though the generated data still falls short in visual quality. Research has produced models trained on synthetic data that not only achieve parity with models trained on real data, but by now even surpass models trained on real data due to the full control (i.e. multiview data), additional modalities (i.e. depth and normals) and pixel perfect annotations (i.e. 3D keypoints). Research such as the seminal “Sapiens” work has further demonstrated success by finetuning foundational models trained on real data with synthetic data, to enable generalization despite limited amounts of synthetic assets.

All of those models are discriminative – they analyze real images to derive higher order semantics from it, such as keypoints, segmentations, or surface normals. On the generative side people have been considerably less successful, due to the remaining domain gap in the data generated, causing domain shifts as visible for example in “Rodin”. Recent works, such as “Cafca” or “SynShot”, have proposed to overcome this domain shift by fine-tuning on real data, typically following a three-step approach: 1) pretrain a prior on synthetic data, 2) invert sparse views into that prior, 3) employ a fine-tuning strategy to go off model. While this can produce results without domain shift, it’s of course not a fully generative model anymore, and as such has its own limitations – for example one cannot unconditionally sample from it. Looking forward we predict the development to trend towards fusing high quality 3D synthetic data with the emerging large scale image or video foundational models to overcome the domain gap for generative 3D models.

3.2 Learning to Infer Parametric Representations for 3D Plants from Scans

Samara Gherrer (University of Grenoble, FR)

License © Creative Commons BY 4.0 International license
© Samara Gherrer

Joint work of Samara Gherrer, Christophe Godin, Stefanie Wuhrer
Main reference Samara Gherrer, Christophe Godin, Stefanie Wuhrer: “Learning to Infer Parameterized Representations of Plants from 3D Scans”, CoRR, Vol. abs/2505.22337, 2025.
URL <https://doi.org/10.48550/ARXIV.2505.22337>

Reconstructing faithfully the 3D architecture of plants from unstructured observations is a challenging task. Plants frequently contain numerous organs, organized in branching systems in more or less complex spatial networks, leading to specific computational issues due to self-occlusion or spatial proximity between organs. Existing works either consider inverse modeling where the aim is to recover the procedural rules that allow to simulate virtual plants, or focus on specific tasks such as segmentation or skeletonization. We propose a unified approach that, given a 3D scan of a plant, allows to infer a parameterized representation of the plant. This representation describes the plant’s branching structure, contains parametric information for each plant organ, and can therefore be used directly in a variety of tasks.

3.3 Statistical Approaches to Internal Anatomy Prediction

Marilyn Keller (MPI für Intelligente Systeme – Tübingen, DE)

- License** © Creative Commons BY 4.0 International license
© Marilyn Keller
- Joint work of** Marilyn Keller, Keenon Werling, Soyong Shin, Scott L. Delp, Sergi Pujades, C. Karen Liu, Michael J. Black, Silvia Zuffi, Vaibhav Arora, Abdelmouttaleb Dakri, Shivam Chandhok, Jürgen Machann, Andreas Fritsche
- Main reference** Marilyn Keller, Keenon Werling, Soyong Shin, Scott L. Delp, Sergi Pujades, C. Karen Liu, Michael J. Black: “From Skin to Skeleton: Towards Biomechanically Accurate 3D Digital Humans”, ACM Trans. Graph., Vol. 42(6), pp. 253:1–253:12, 2023.
URL <https://doi.org/10.1145/3618381>
- Main reference** Marilyn Keller, Silvia Zuffi, Michael J. Black, Sergi Pujades: “OSSO: Obtaining Skeletal Shape from Outside”, in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022, pp. 20460–20469, IEEE, 2022.
URL <https://doi.org/10.1109/CVPR52688.2022.01984>
- Main reference** Marilyn Keller, Vaibhav Arora, Abdelmouttaleb Dakri, Shivam Chandhok, Jürgen Machann, Andreas Fritsche, Michael J. Black, Sergi Pujades: “HIT: Estimating Internal Human Implicit Tissues from the Body Surface”, in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16–22, 2024, pp. 3480–3490, IEEE, 2024.
URL <https://doi.org/10.1109/CVPR52733.2024.00334>

Personalized anatomical digital twins are essential for medicine, computer graphics, and biomechanics, but serving internal anatomy usually requires expensive medical imaging. Instead, we can leverage the correlation between external body shape and internal structures to predict the anatomy directly from a subject’s appearance. Learning this correlation raises three key challenges: building datasets with paired observations of body shapes and internal anatomy, annotating these datasets, and learning models that capture the correlation between external and internal features. In this talk, I will showcase how we became able to predict skeleton geometry, bone location, and soft tissue distribution solely from external body shape.

3.4 On Learning to Reconstruct Shape using World Priors, Handling Topological Inconsistencies, and Edge Integration as a Model of Choice

Ron Kimmel (Technion – Haifa, IL)

- License** © Creative Commons BY 4.0 International license
© Ron Kimmel

Deep learning is a disruptive line of research that continues to reshape how computational problems are formulated and solved. While remarkable advances have been made in tasks involving classification, segmentation, and reconstruction, certain fundamental limitations remain – especially when handling geometric data not suited to traditional convolutional structures. This note outlines our ongoing efforts to address shape reconstruction and analysis in such scenarios, drawing inspiration from differential geometry, spectral theory, and human vision.

Deep learning architectures are designed around the notion of shift-invariance and low-dimensional latent structures. These assumptions empower convolutional neural networks (CNNs) to generalize well across a range of problems. However, for geometric structures where no natural shift-invariant coordinate system exists, standard methods often fall short. In our group, we have been exploring new avenues that combine classical geometry with modern machine learning to reconstruct and analyze shapes – even when those shapes are partially occluded, topologically noisy, or nonrigid. These include tools for shape matching, geodesic measurement, and the design of semi-supervised learning techniques rooted in axiomatic geometric principles.

One of geometry’s grand challenges is explaining the world we live in. Shape comparison, particularly of nonrigid objects, exemplifies this. Unlike rigid objects, nonrigid shapes lack a universal parameterization, complicating comparison. Over the past two decades, we have developed methods for analyzing and matching such shapes using the Gromov Hausdorff (GH) distance, which quantifies discrepancies between metric spaces. We have advanced this from theoretical abstraction to practical approximation using spectral embeddings.

Traditionally, computer vision (CV) aimed to extract geometry from images, while computer graphics (CG) sought to generate images from geometry. Today, these domains converge. A fundamental question we address is: Can neural networks help discover invariants or derive parameterizations of geometric data? To this end, we showed that the Gromov distance between metric spaces offers a powerful framework for non-rigid shape analysis.

Classical computer vision focused on “shape from X” problems: shape from shading (often solved using eikonal solvers), shape from stereo (requiring correspondence resolution), and shape from auto-stereograms. A compelling example comes from stereoscopic vision. Consider looking through a cardboard tube with one eye, while the other eye is blocked by one hand – this creates a visual illusion of a “hole” in the hand, as the brain attempts to force a false correspondence between views. Such hallucination phenomena motivated the study of auto-stereograms generation, see [2], and geometric reconstruction from such data using methods described in [1]. It motivates the concept of prior based geometry reconstruction as recently adopted and adapted in [3]. It demonstrates how prior knowledge of image formation models captured by neural network trained to solve the shape from stereo problem, can significantly improve reconstruction accuracy integrating Gaussian splatting with synthetic stereo pairs.

Recently, our attention has turned to shape invariants, especially under partial observation. We introduced the *Wormhole Loss* [4], designed for robust partial matching by emphasizing topological consistency. This follows our earlier work on partial shape correspondence using functional maps [5], which outperforms previous efforts such as [6, 7] by incorporating both local and global topology sensitive invariants into the learning objective.

Finally, another classical idea that should be re-examined through the lens of deep learning is edge detection. The variational model for optimal edge integration via regularized Laplacian zero-crossings [8] raises an intriguing question: can such elegant mathematical formulations be embedded into a neural network framework? If so, which applications would benefit most, and how could these formulations serve as loss functions or regularizers in modern computer vision applications?

References

- 1 Ron Kimmel. 3D Shape Reconstruction from Autostereograms and Stereo. *Journal of Visual Communication and Image Representation*, 13:324-333, 2002.
- 2 A. M. Bruckstein, R. Onn, and T. J. Richardson. Improving the vision of magic eyes: A guide to better autostereograms. In *Advances in Image Understanding, A Festschrift for Azriel Rosenfeld*, IEEE Comput. Sci., 1996.
- 3 Y. Wolf, A. Bracha, and R. Kimmel. Gs2mesh: Surface reconstruction from Gaussian splatting via novel stereo views. *European Conference on Computer Vision*, 2024.
- 4 A. Bracha, T. Dagès, and R. Kimmel, “Wormhole Loss for Partial Shape Matching,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [Online]. Available: <https://arxiv.org/abs/2410.22899>
- 5 A. Bracha, T. Dagès, and R. Kimmel, “On Partial Shape Correspondence and Functional Maps,” in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. [Online]. Available: <https://arxiv.org/abs/2310.14692>

- 6 G. Rosman, M. M. Bronstein, A. M. Bronstein, and R. Kimmel. Nonlinear dimensionality reduction by topologically constrained isometric embedding. *International Journal of Computer Vision*, 2010.
- 7 A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Topology-invariant similarity of nonrigid shapes. *International Journal of Computer Vision*, 2009.
- 8 R. Kimmel and A. M. Bruckstein. Regularized Laplacian zero crossings as optimal edge integrators. *International Journal of Computer Vision*, 2003.

3.5 Towards Photorealistic 3D Head Avatars

Tobias Kirschstein (TU München – Garching, DE)

License © Creative Commons BY 4.0 International license
© Tobias Kirschstein

In recent years, 3D head avatars have become ever more realistic, leading to their integration into first commercial products. In this talk, we will explore some of the key developments that drive research in this area, including datasets, methods, and efforts to standardize evaluation. Studio-level multi-view video datasets such as the NeRSemble dataset enable creating high quality animatable reconstructions of human heads. For example, the current state-of-the-art method, Neural Parametric Gaussian Avatars, combines multi-view supervision with a detailed 3D face model to produce lifelike digital doubles. In the future, the field is moving toward learning larger generative 3D priors which could help tackle under-constrained scenarios such as reconstructing 3D head avatars from single images. Ultimately, we will have to answer the question whether the future lies in recording more data in controlled studio environments or exploring new paradigms, like 3D GANs, that can infer 3D priors from 2D data alone.

3.6 Understanding Self-Supervised Learning


Adam Kortylewski (Universität Freiburg, DE and MPI für Informatik – Saarbrücken, DE)

License © Creative Commons BY 4.0 International license
© Adam Kortylewski

SSL methods gave us general-purpose visual features like DinoV2 that perform well as backbone for a variety of tasks, like classification, segmentation and even low-level 3D tasks like depth or normal estimation. Interestingly, these features emerge despite being trained with pre-text objectives that are rather unrelated to actual (3D) vision tasks, like making image crops of the same image similar to each other. In this talk, I will discuss some initial ideas about the connection between SSL of visual features and visual correspondence tasks. I will discuss experiments that show how current SSL features are not just good at 2D vision but also 3D vision tasks, and how we could possibly develop pre-text objectives that explicitly optimize semantic correspondence capabilities of visual features, ultimately leading to the emergence of more advanced visual features.

3.7 Leveraging 3D Representations for Multi-view Synthesis and Editing with 2D Generative Models

Or Patashnik (Tel Aviv University, IL)

License  Creative Commons BY 4.0 International license
© Or Patashnik

While 2D generative models have achieved remarkable success in image synthesis and editing, extending their power to multi-view and 3D tasks often leads to inconsistencies and limited control. In this talk, I will discuss a line of work showing how incorporating 3D representations, ranging from explicit 3D shapes to emergent 3D feature fields, can significantly enhance the capabilities of 2D generative models for multi-view generation and editing. I will first demonstrate how explicit geometric structures, such as rough meshes or low-fidelity 3D models, can guide and stabilize the generation process, enabling better control and consistency. I will then discuss how, even without an explicit shape, constructing a 3D-aware feature field during generation can substantially improve multi-view coherence. Together, these approaches show that introducing 3D representations, whether provided upfront or learned during the process, fundamentally amplifies the capabilities of 2D diffusion models and enables high-quality, controllable 3D content creation.

3.8 Scaling Digital Humans

Shunsuke Saito (Codec Avatars Lab – Pittsburgh, US)

License  Creative Commons BY 4.0 International license
© Shunsuke Saito

In this talk, we discuss how to put the modeling of 3D Digital Humans on the table of large-scale training. In particular, we share our recent experiments to see the potential benefit and risk of pre/post-training regime. Self-supervised learning with large scale data and then finetuning on small but clean data leads to generalization and high-quality at scale. This shows surprising generalization to the domain outside the post-training data distribution. We also discuss the importance of 3D in the era of powerful 2D generative models. By presenting our recent work on efficient 3D Gaussian avatar decoding, we illustrate how efficiency plays a critical role in content creation with 3D-based approaches.

3.9 Clever Data Curation for Smart Supervision

William Smith (University of York, GB)

License  Creative Commons BY 4.0 International license
© William Smith

This talk summarised work from the past 10 years that opportunistically repurposed data collected for one purpose to provide supervision for a 3D vision task. The purpose of the talk was to: 1. highlight that there are likely many more datasets available that have not yet been tapped for such purposes, 2. to try to categorise previous methods and 3. to stimulate discussion around ideas for problems or datasets that could benefit from or be used in this way. The proposed initial categorisation along with representative methods was:

- Multiview curation at training time for single view task supervision [1, 2, 3, 4]
- Two view curation at training time for two view task supervision [5, 6, 7]
- Compositing [8]
- Wide FOV to supervise beyond cropped narrow FOV [9]
- Diffusion-based editing [10]

In the discussion that followed, many sources of interesting supervision were proposed. Aging or fitness journey timelapse videos can be used to understand changes in appearance with a fixed identity. Zipline or rollercoaster videos provide smooth fly throughs that are repeated in many different conditions with almost the same camera trajectory. Ice bucket challenge videos provide examples of the same clothes dry and wet. Flash photography, for example at red carpet events, provides photometric stereo cues. Google streetview provides many lighting conditions for the same scene. Object press videos show how different materials deform. Slow TV such as railway journeys provide revisits of slowly changing landscapes.

References

- 1 Zhengqi Li and Noah Snavely. *MegaDepth: Learning Single-View Depth Prediction from Internet Photos*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2041-2050, 2018.
- 2 Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. *Learning the Depths of Moving People by Watching Frozen People*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4521-4530, 2019.
- 3 Zhengqi Li and Noah Snavely. *Learning Intrinsic Image Decomposition from Watching the World*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9039-9048, 2018.
- 4 Ye Yu and William A. P. Smith. *InverseRenderNet: Learning single image inverse rendering*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3155-3164, 2019.
- 5 Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. *Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow*. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.
- 6 Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. *DUST3R: Geometric 3D Vision Made Easy*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697-20709, 2024.
- 7 Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. *Stereo4D: Learning How Things Move in 3D from Internet Stereo Videos*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- 8 Finlay G. C. Hudson and William A. P. Smith. *Track Anything Behind Everything: Zero-Shot Amodal Video Object Segmentation*. arXiv preprint arXiv:2411.19210, 2024.
- 9 Rundong Luo, Matthew Wallingford, Ali Farhadi, Noah Snavely, and Wei-Chiu Ma. *Beyond the Frame: Generating 360° Panoramic Videos from Perspective Videos*. arXiv preprint arXiv:2504.07940, 2025.
- 10 Alex Trevithick, Ruihan Yang, Rundi Wu, Ruiqi Gao, Ben Poole, Aleksander Holynski, Jonathan T. Barron, Noah Snavely, and Angjoo Kanazawa. *SimVS: Simulating World Inconsistencies for Robust View Synthesis*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

3.10 Generating 3D Human Motion with Language

Gül Varol (ENPC – Marne-la-Vallée, FR)

License © Creative Commons BY 4.0 International license
© Gül Varol

Text-driven 3D human motion generation methods have recently evolved to synthesize language-controllable avatars. My talk briefly shows results of motion generation with various levels of granularity and presents recent extensions to motion editing and hand motion generation. An obvious challenge in this domain is the scarcity of data. We will discuss ways to scale up the data to train these models, with a particular focus on noise vs scale trade-off.

3.11 3D Foundation Models for Enhanced Geometry in Gaussian Splatting

Yaniv Wolf (Technion – Haifa, IL)

License © Creative Commons BY 4.0 International license
© Yaniv Wolf

Joint work of Yaniv Wolf, Amit Bracha, Ron Kimmel

Main reference Yaniv Wolf, Amit Bracha, Ron Kimmel: “GS2Mesh: Surface Reconstruction from Gaussian Splatting via Novel Stereo Views”, in Proc. of the Computer Vision – ECCV 2024 – 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXIX, Lecture Notes in Computer Science, Vol. 15147, pp. 207–224, Springer, 2024.

URL https://doi.org/10.1007/978-3-031-73024-5_13

In recent years, Gaussian Splatting (GS) has emerged as an efficient and accurate method for scene representation. However, while GS excels in novel view synthesis, it struggles with accurately representing geometry, as appearance and geometry are often conflicting objectives for the Gaussian representation itself (consider a colorful flat wall, or a textureless round surface). Attempts to simultaneously optimize both appearance and geometry often lead to suboptimal compromises. To bridge the gap between appearance and geometry, we introduce GS2Mesh, a novel pipeline that extracts accurate surface geometry from any optimized GS scene without degrading its visual quality. Rather than directly optimizing the Gaussians for both appearance and geometry, we leave the Gaussians optimized for appearance, and leverage a robust 3D data-driven prior, in the form of a pre-trained stereo matching foundation model, for the geometry extraction. Specifically, we render stereo-aligned image pairs from the optimized GS scene and feed them to the pre-trained stereo model, resulting in high-quality disparity maps. These maps are then scaled according to the camera parameters to form multi-view consistent depth maps, which are fused together to reconstruct accurate 3D surfaces. Our approach achieves state-of-the-art performance on popular 3D reconstruction benchmarks, as well as robust performance on in-the-wild scenes captured via smartphones. Due to the modularity of the approach, performance is expected to improve as better GS and stereo matching models emerge.

3.12 D-Garment: Physics-Conditioned Latent Diffusion for Dynamic Garment Deformations

Stefanie Wuhrer (INRIA – Grenoble, FR)

License © Creative Commons BY 4.0 International license

© Stefanie Wuhrer

Joint work of Antoine Dumoulin, Adnane Boukhayma, Laurence Boissieux, Bharath Bhushan Damodaran, Pierre Hellier, Stefanie Wuhrer

Main reference Antoine Dumoulin, Adnane Boukhayma, Laurence Boissieux, Bharath Bhushan Damodaran, Pierre Hellier, Stefanie Wuhrer: “D-Garment: Physics-Conditioned Latent Diffusion for Dynamic Garment Deformations”, CoRR, Vol. abs/2504.03468, 2025.

URL <https://doi.org/10.48550/ARXIV.2504.03468>

Adjusting and deforming 3D garments to body shapes, body motion, and cloth material is an important problem in virtual and augmented reality. Applications are numerous, ranging from virtual change rooms to the entertainment and gaming industry. This problem is challenging as garment dynamics influence geometric details such as wrinkling patterns, which depend on physical input including the wearer’s body shape and motion, as well as cloth material features. Existing work studies learning-based modeling techniques to generate garment deformations from example data, and physics-inspired simulators to generate realistic garment dynamics. The presentation covered our recent learning-based approach trained on data generated with a physics-based simulator. Compared to prior work, our 3D generative model learns garment deformations for loose cloth geometry, especially for large deformations and dynamic wrinkles driven by body motion and cloth material. Furthermore, the model can be efficiently fitted to observations captured using vision sensors. We propose to leverage the capability of diffusion models to learn fine-scale detail: we model the 3D garment in a 2D parameter space, and learn a latent diffusion model using this representation independent from the mesh resolution. This allows to condition global and local geometric information with body and material information. The presentation concluded with some open questions that benefitted from discussions in the diverse group of the seminar.

4 Open problems

4.1 Interpolation of Shapes with Topological and Structural Variability

Andreea Ardelean (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE) and Samara Gherrer (University of Grenoble, FR)

License © Creative Commons BY 4.0 International license

© Andreea Ardelean and Samara Gherrer

This open discussion centers on the problem of representing and understanding the structural topology of growing plants, raising foundational questions about how plant morphology evolves over time and how it can be modeled computationally. Participants explored whether plant topology – particularly the branching structure – changes during growth, and if so, how to model this dynamic nature in a consistent and analyzable way. Analogies to human anatomy (e.g., variability in tooth roots or spinal structure) highlighted similar topological ambiguity in biological systems.

A range of modeling strategies were discussed. Neural implicit representations such as NeuralSDF offer a powerful tool for interpolating between shapes, after which a skeleton can be extracted. However, challenges arise when interpolations generate non-physical or

biologically invalid intermediate states – for example, a plant structure with a fractional number of branches, or a tooth with an unrealistic root configuration. In this context, skeleton extraction and branching modeling must be grounded in biological plausibility.

Stochastic modeling was proposed as an effective alternative to rule-based procedural methods. While procedural modeling requires the system to strictly follow hand-crafted rules, stochastic models can learn from data, making them more flexible for capturing the natural variability in plant structures. Parameters of branching processes (e.g., angles, lengths, probabilities of bifurcation) can be learned directly from real plant data, enabling more realistic simulations.

There was a shift in perspective proposed: rather than thinking in purely topological terms, plant structures may be better understood as graph-like entities, where branches and junctions form a structured hierarchy. This aligns better with biological reality and allows for more intuitive comparisons between different plant forms.

To compare plant structures across stages of growth or between individuals, several mathematical tools were proposed. Diffusion distances were suggested as a more robust alternative to geodesic distances, offering smoother and more meaningful metrics for shape comparison. In cases where the structure contains holes or discontinuities, Fourier-based approaches might offer better representations. Still, the core challenge remains: how to define meaningful correspondences between topologically diverse structures, especially when interpolations between them don't represent real physical states (e.g., morphing between different glyph forms of the letter “g”).

From an application-driven standpoint, several motivations for understanding plant topology were discussed. These include:

- Monitoring and diagnosing plant health (e.g., detecting disease via structural anomalies).
- Understanding growth patterns of individual plants versus species-level trends (nature vs. nurture).
- Predicting future states or motions of plants based on current observations.

The discussion emphasized the importance of discrete vs. continuous modeling, particularly in cases like the transition from flower to fruit, which represents a discrete structural change atop a generally continuous growth process. Mixture models or hybrid representations (continuous shapes with discrete categorical states) were considered useful for capturing this dual nature.

Drawing parallels to 3D morphable models (3DMMs) used in face reconstruction, the conversation highlighted the trade-off between model complexity and usability. While complex models may better capture biological variance, they are harder to fit and interpret. 3DMMs benefit from disentanglement and controllability, features that plant models currently lack due to the high diversity and complexity of plant morphologies.

Throughout the discussion, there was a recurring theme: the purpose of the model should drive its design. For instance, determining a person's eye color is more efficiently done through observation than genetic analysis – a metaphor for focusing modeling efforts on efficient, goal-aligned representations. Similarly, there's a risk in over-parameterizing models, leading to poor generalization, or under-parameterizing, resulting in loss of essential structure.

Ultimately, effective plant topology modeling must balance generative and discriminative approaches, provide biologically grounded constraints, and respect the discrete-continuous duality of natural structures. The goal is not just to simulate growth but to create meaningful, applicable models that align with real-world use cases such as agriculture, ecology, and evolutionary biology.

4.2 Exploring problems with infinite data

Timotei Ardelean (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE)

License © Creative Commons BY 4.0 International license
© Timotei Ardelean

In order to explore the interplay between data, compute, and algorithms, it was proposed to analyze our current methods and training methodologies on constrained, well-defined tasks where perfect training data can be easily generated. By generating synthetic data without restrictions, the effect of scaling w.r.t to data and training time could more easily be disentangled, and algorithmic limitations could be identified. The discussions produced several suggestions of Computer Vision problems where shortcomings of the current models could be investigated, including maze solving, the n-queens puzzle, and Sudoku (as inspired by the previous talk of this seminar: Believe Propagation over Spatial Domains with Generative Models). A very simple example we considered is Conway's Game of Life, which we analyzed with a quick hands-on experiment, by training a model to predict the update rule of the game. The model quickly learned the rules of the game using synthetic randomly generated boards, supervised by the true update rule. We also tested for generalization by excluding certain patterns from training. For the simple cases tested, the model was able to generalize to these unseen patterns, suggesting the network learns more than a simple look-up table operation.

4.3 Benchmarking Monocular Reconstruction Discussion

James Gardner (University of York, GB)

License © Creative Commons BY 4.0 International license
© James Gardner

The Limitations of Current Benchmarking

It was noted that existing metrics, especially for generative tasks like single-view 3D reconstruction, are fundamentally flawed. An analogy was drawn to image compression, where metrics such as PSNR failed to capture perceptual artifacts, this led to the development of new perceptual metrics, which themselves became targets for optimization, not always resulting in genuinely better qualitative outcomes. A similar pattern is evident in LLM evaluation, where comprehensive benchmark suites are still critiqued for fostering optimization towards specific metrics over broader user satisfaction. This phenomenon, often described by Goodhart's Law – where a metric ceases to be effective once it becomes a target – is a significant point of concern. Consequently, reviewers often find it necessary to look beyond numerical scores, emphasizing qualitative aspects and the soundness of the proposed methods.

Proposed Directions for Better Benchmarks

Several approaches for enhancing benchmarks were put forward:

Focus on Failure Modes and Use Cases: It was proposed that new metrics should not be designed merely to show a method's superiority, but should instead arise from an understanding of how and why current methods fail. If a particular artifact, such as object interpenetration, is identified as problematic, metrics should be formulated to quantify it. This, however, introduces complexities, including the definition of object separability (e.g.,

determining if cushions are distinct from a sofa). The specific application or use case was highlighted as critically important, metrics must align with the intended outcome, whether for applications like 3D printing (where plausibility might be prioritized over exact scale) or interactive scene manipulation (where object decomposition is vital).

Plausibility and Perceptual Evaluation: For ill-posed problems that involve substantial hallucination or completion, the plausibility of the generated 3D geometry or novel views is of paramount importance. Proposals included:

- More extensive user studies.
 - The leveraging of other sophisticated models, such as Video Models or Vision-Language Models (VLMs), to evaluate the consistency and “common sense” of reconstructed scenes (e.g., how objects might interact or react to physical phenomena like an earthquake). The underlying principle is that if novel views are plausible and internally consistent, the geometry might be deemed adequate for many applications, even if precise geometric accuracy remains elusive.
- Embracing Ambiguity and Distributions:** Given that many 3D vision tasks are ill-posed and possess multiple valid solutions (e.g., completing an occluded part of an object), it was argued that models should predict a distribution of potential solutions. Evaluation methodologies would then need to assess the likelihood of the ground truth under this predicted distribution. This approach necessitates considerably larger and more varied test datasets to adequately probe ambiguous scenarios.

Identifying the “Next Frontier”: It was stated that effective benchmarks should delineate the “next significant problem.” When current methodologies begin to saturate existing benchmarks, this is considered a signal to identify new, more demanding tasks that will cause the performance of existing state-of-the-art methods to decline, thereby stimulating further advancements.

Contentious Metrics and Community Influence

The discussion also briefly addressed alternative, non-traditional indicators of a method’s utility, such as the number of GitHub stars. It was argued by some that this reflects real-world adoption, the quality of the accompanying code, and overall impact. However, strong counter arguments were presented, noting that such indicators lack scientific rigor, can be disproportionately influenced by “hype,” may not accurately reflect true scientific merit or superior performance (particularly for less mainstream research areas), and could disadvantage robust methods that have less polished codebases or are not aligned with current popular trends.

In summary, there was an agreement that current benchmarking practices within 3D vision are insufficient. The proposed way forward involves the development of metrics that are more intimately aligned with specific use cases, effectively capture perceptual plausibility, appropriately account for the inherent ambiguity present in many 3D tasks, and are derived from an understanding of current model limitations rather than focusing on simple numerical rankings. The intrinsic difficulty is in the creation of metrics that are simultaneously robust, meaningful, and resistant to superficial optimization.

4.4 Capturing useful datasets for 3D vision tasks

Samara Ghrer (University of Grenoble, FR)

License © Creative Commons BY 4.0 International license
© Samara Ghrer

Dataset propositions and ideas to be collaboratively captured:

- 4D human motion dataset with a large number of markers to capture motions that cannot be captured by cameras, or by other motion capture datasets with lower number of markers that lead to smoothed motions.
- Large volumetric dataset of humans in motion with multi-people, self and object interactions.
- Multi view face dataset in both indoor and outdoor lighting conditions with marker alignment (invisible markers on the back of the head, or using sunscreen).
- MRI + volumetric dynamic or multi-pose captures.
- Large and diverse dataset that is captured across different labs with an agreed setting.
- DagMark: A 3D vision (different tasks) benchmarking challenge where people can submit datasets, evaluation schemes and metrics. Once submitted, the benchmarks are automatically compared to existing ones and added to the collection. This challenge can be hosted on a website for worldwide contributions.

Raised questions:

- Is it possible to simultaneously capture MRI scans and real-life images using a combination of MRI scanners and cameras?
- Plausibility vs accuracy: Can we design a dataset that supports both ends?

4.5 Believe Propagation over Spatial Domains with Generative Models

Jan Eric Lenssen (MPI für Informatik – Saarbrücken, DE)

License © Creative Commons BY 4.0 International license
© Jan Eric Lenssen

Joint work of Christopher Wewer, Bartłomiej Pogodzinski, Bernt Schiele, Jan Eric Lenssen

Main reference Christopher Wewer, Bartłomiej Pogodzinski, Bernt Schiele, Jan Eric Lenssen: “Spatial Reasoning with Denoising Models”, in Proc. of the International Conference on Machine Learning (ICML), 2025.

URL <https://geometric-rl.mpi-inf.mpg.de/srm/>

Video and Multi-View Generation Models are taking over (3D) Computer Vision. We are expecting them to model extremely complex, high-dimensional distributions. However, in many cases they spectacularly fail and collapse to hallucinations. I’d like to pose the question: is scaling up data and compute the best or only solution to solve this problem? And is it a good idea to keep up with exponential growth of distribution complexity? Bayesian inference came up with intelligent ways to simplify distribution complexity in the past. It might be a good time to explore similar ideas for continuous generative models in spatial domains. In this talk, I would introduce these discussion points and present some preliminary results we obtained in recent investigations.

4.6 Hard problems in computer vision and the necessity of modeling

Yaniv Wolf (Technion – Haifa, IL)

License  Creative Commons BY 4.0 International license
© Yaniv Wolf

The group discussed what defines a “hard problem” in computer vision and how models should be designed to solve them. While some problems benefit from large-scale “big hammer” solutions, others require specialized tools. In that context, neural networks were recognized as convenient, but also problematic in post-data regimes due to heavy data dependence, limited explainability, and robustness issues.

The discussion focused on two main issues:

- The first issue discussed was whether a “hard problem” is defined by the amount of data required to train a model to solve it, and whether any problem can be solved without the need for generalization if enough data is available. Several examples were presented, such as training NeRFs with synthetic/real data, learning geometric consistency in video models, understanding grammar in LLMs, and solving complex puzzles such as 3D mazes and Sudoku.
- The second issue discussed was the impact of explainability in models. Some participants argued that explainability is essential for humans, as the designers of these models, to understand what they are designing. Others believed that introducing parametric models might negatively affect model performance. An example discussed was self-driving cars, and whether they should adhere strictly to theoretical traffic laws or learn from actual human driving behavior.

Finally, caution was expressed regarding overly complex solutions (“big hammers”) that risk memorization instead of genuine generalization.

Participants

- Andreea Ardelean
Friedrich-Alexander-Universität
Erlangen-Nürnberg, DE
- Timotei Ardelean
Friedrich-Alexander-Universität
Erlangen-Nürnberg, DE
- Thabo Beeler
Google Research – Zürich, CH
- Timo Bolkart
Google Research – Zürich, CH
- Neill Campbell
University of Bath, GB
- Rishabh Dabral
MPI für Informatik –
Saarbrücken, DE
- Olaf Dünkel
MPI für Informatik –
Saarbrücken, DE
- Bernhard Egger
Friedrich-Alexander-Universität
Erlangen-Nürnberg, DE
- James Gardner
University of York, GB
- Samara Ghrer
University of Grenoble, FR
- Marilyn Keller
MPI für Intelligente Systeme –
Tübingen, DE
- Ron Kimmel
Technion – Haifa, IL
- Tobias Kirschstein
TU München – Garching, DE
- Adam Kortylewski
Universität Freiburg, DE & MPI
für Informatik – Saarbrücken, DE
- Jan Eric Lenssen
MPI für Informatik –
Saarbrücken, DE
- Ruoshi Liu
Columbia University –
New York, US
- Laura Neschen
INRIA Rhône-Alpes, FR
- Or Patashnik
Tel Aviv University, IL
- Ryan Po
Stanford University, US
- Shunsuke Saito
Codec Avatars Lab –
Pittsburgh, US
- William Smith
University of York, GB
- Christian Theobalt
MPI für Informatik –
Saarbrücken, DE
- Gül Varol
ENPC – Marne-la-Vallée, FR
- Yaniv Wolf
Technion – Haifa, IL
- Stefanie Wuhrer
INRIA – Grenoble, FR

