

# 3D Morphable Models and Beyond

James Gardner<sup>\*1</sup>, Bernhard Egger<sup>†2</sup>, William Smith<sup>†3</sup>,  
Christian Theobalt<sup>†4</sup>, and Stefanie Wuhler<sup>†5</sup>

- 1 University of York, GB. [james.gardner@york.ac.uk](mailto:james.gardner@york.ac.uk)
- 2 Friedrich-Alexander-Universität Erlangen-Nürnberg, DE. [bernhard.egger@fau.de](mailto:bernhard.egger@fau.de)
- 3 University of York, GB. [william.smith@york.ac.uk](mailto:william.smith@york.ac.uk)
- 4 MPI für Informatik – Saarbrücken, DE. [theobalt@mpi-inf.mpg.de](mailto:theobalt@mpi-inf.mpg.de)
- 5 INRIA – Grenoble, FR. [stefanie.wuhler@inria.fr](mailto:stefanie.wuhler@inria.fr)

---

## Abstract

3D Morphable Models are models separating shape from appearance variation. Typically, they are used as a statistical prior in computer graphics and vision. Recent success with neural representations have caused a resurgence of interest in visual computing problems, leading to more accurate, higher fidelity, more expressive, and memory-efficient solutions. This report documents the program and the outcomes of Dagstuhl Seminar 22121, “3D Morphable Models and Beyond”. This meeting of 39 researchers covered various topics, including 3D morphable models, implicit neural representations, physics-inspired approaches, and more. We summarise the discussions, presentations and results of this workshop.

**Seminar** March 20–25, 2022 – <http://www.dagstuhl.de/22121>

**2012 ACM Subject Classification** Computing methodologies → Computer graphics; Computing methodologies → Image-based rendering; Computing methodologies → Shape modeling; Computing methodologies → Animation; Computing methodologies → Computer vision; Computing methodologies → 3D imaging

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

**Digital Object Identifier** 10.4230/DagRep.12.3.97

## 1 Executive Summary

*James Gardner (University of York, GB)*

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

*William Smith (University of York, GB)*

*Christian Theobalt (MPI für Informatik – Saarbrücken, DE)*

*Stefanie Wuhler (INRIA – Grenoble, FR)*

**License**  Creative Commons BY 4.0 International license  
© James Gardner, Bernhard Egger, William Smith, Christian Theobalt, Stefanie Wuhler

A total of 63 people were invited to the seminar in the first round of invitations. 39 people attended, with 15 of those attending the seminar virtually. Participants came from both academia and industry and at varying stages of their careers. As this seminar took place at the trailing end of the Covid-19 pandemic, it ran in a hybrid format, and for many attendees, Dagstuhl was the first in-person seminar in several years. Due to the fantastic facilities of the Dagstuhl campus, the hybrid format was a great success, enabling accessible and inclusive

---

\* Editorial Assistant / Collector

† Editor / Organizer



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

3D Morphable Models and Beyond, *Dagstuhl Reports*, Vol. 12, Issue 3, pp. 97–116

Editors: James Gardner, Bernhard Egger, William Smith, Christian Theobalt, and Stefanie Wuhler



Dagstuhl Reports

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

communication with remote participants. Daily Covid testing for those in-person ensured that everyone remained safe throughout the week. Eighteen presented their work in around 15-30 minute presentations; an abstract for each talk is included in this report.

Alongside traditional presentations, many sessions were left available for activities suggested by the seminar participants. These could involve workshops, discussions, presentations, or any other suggested format. During the week, participants could propose plans for the flexible sessions and the structure of the seminar became fixed as activities and topics for the sessions were provided. Summaries for the results of these flexible sessions are contained in this report. One slot was reserved for a joint group discussion on the ethical concerns of the research we are developing. This resulted in a vivid discussion on the steps we as a community should be taking to encourage the ethical use of the technology we are developing. One of the discussions that received broad support was the design of a cheap, open-source method for collecting camera calibrated illumination environments. This resulted in a Slack channel for the group of interested researchers and the pursuit of an early prototype design. We started the seminar with a short introduction from all participants. Everyone was given one slide to introduce themselves and asked to prepare a question, challenge or goal to discuss during the seminar.

## 2 Table of Contents

### Executive Summary

*James Gardner, Bernhard Egger, William Smith, Christian Theobalt, Stefanie Wuhrer* 97

### Overview of Talks

Learning to Fit Morphable Models <i>Federica Bogo</i> . . . . .	101
From Pixels to Expressive 3D Bodies <i>Timo Bolkart</i> . . . . .	101
Plausible (Neural) Rendering of Bodies & Garments in Motion <i>Duygu Ceylan</i> . . . . .	102
Inferring people’s anatomic skeleton from their external appearance <i>Marilyn Keller</i> . . . . .	102
Deep Signatures – Learning Invariants of Planar Curves <i>Ron Kimmel and Roy Velich</i> . . . . .	102
Computer Vision does not generalize – 3DMMs and beyond can help <i>Adam Kortylewski</i> . . . . .	103
NeRF for View Synthesis <i>Ben Mildenhall</i> . . . . .	104
Deep Relighting of 3D Faces <i>Shunsuke Saito</i> . . . . .	104
Digital Humans in Motion <i>Justus Thies</i> . . . . .	105
A Structured Latent Space for Human Body Motion Generation <i>Stefanie Wuhrer</i> . . . . .	105
Implicit 3DMMs for Full Heads Including Hair <i>Tarun Yenamandra</i> . . . . .	106
Towards Precise Completion of Deformable Shapes <i>Oshri Halimi</i> . . . . .	106
Do We Still Need to Detect Faces and Facial Landmarks? Do We Need to Estimate 3D Face Shapes? <i>Tal Hassner</i> . . . . .	107
On Implicit Avatars, Racial Bias, and the Light/Albedo Ambiguity <i>Victoria Fernández Abrevaya</i> . . . . .	107

### Working groups

What are the ‘Killer Applications’ of 3D Implicit Representations <i>Ben Mildenhall</i> . . . . .	108
Are Neural Implicit Representations the Future of Morphable Models? <i>Christian Theobalt</i> . . . . .	109
How Much 3D is Needed? <i>Duygu Ceylan</i> . . . . .	109

Metrical 3D Reconstruction of the Human Face	
<i>Justus Thies</i> . . . . .	110
Synthetic Data Generation Using Morphable Models	
<i>Federica Bogo</i> . . . . .	110
Morphable Models from Physics	
<i>Dan Casas</i> . . . . .	111
Ethics	
<i>Bernhard Egger, William Smith, Christian Theobalt, Stefanie Wuhrer</i> . . . . .	112
<b>Participants</b> . . . . .	115
<b>Remote Participants</b> . . . . .	116

## 3 Overview of Talks

### 3.1 Learning to Fit Morphable Models

*Federica Bogo (Meta Reality Labs Research – Zürich, CH)*

**License** © Creative Commons BY 4.0 International license  
© Federica Bogo

**Joint work of** Federica Bogo, Vasileios Choutas, Jingjing Shen, Julien Valentin

Fitting parametric models of human bodies, hands or faces to image data is an important problem in computer vision. Many recent approaches leverage deep neural networks to regress the parameters of the model directly from the input. These methods are fast and robust, but require large amounts of annotated data and may fail to tightly fit the observations. Therefore, their output is often leveraged as a starting point for an iterative, optimization-based algorithm minimizing an energy function. These functions typically involve a data term, plus priors encoding knowledge of the problem’s structure; unfortunately they are difficult to both formulate and tune. In this talk, I will discuss how learning-based continuous optimization can capture the best of both deep-learning-based regression and classic optimization. I will discuss recent advances in the field and introduce a novel, learning-based approach for human body fitting, inspired by the Levenberg-Marquardt algorithm. Finally, I will identify some limitations of current state-of-the-art approaches and outline a few directions for future research.

### 3.2 From Pixels to Expressive 3D Bodies

*Timo Bolkart (MPI für Intelligente Systeme – Tübingen, DE)*

**License** © Creative Commons BY 4.0 International license  
© Timo Bolkart

**Joint work of** Radek Daneczek, Michael J. Black, Timo Bolkart, Yao Feng, Vasileios Choutas, Dimitrios Tzionas

**Main reference** Radek Daneczek, Michael J. Black, Timo Bolkart: “EMOCA: Emotion Driven Monocular Face Capture and Animation”, CoRR, Vol. abs/2204.11312, 2022.

**URL** <https://doi.org/10.48550/arXiv.2204.11312>

**Main reference** Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, Michael J. Black: “Collaborative Regression of Expressive Bodies using Moderation”, in Proc. of the International Conference on 3D Vision, 3DV 2021, London, United Kingdom, December 1-3, 2021, pp. 792–804, IEEE, 2021.

**URL** <https://doi.org/10.1109/3DV53792.2021.00088>

Recovering expressive humans from images is essential for understanding human behavior. Faces and their emotional expressions provide an important source of information about a person’s internal emotional state. Unfortunately, the best recent 3D face regression methods from monocular images are unable to capture the full spectrum of facial expression, such as subtle or extreme emotions. We address this problem with EMOCA, by introducing a novel deep perceptual emotion consistency loss during training, which helps ensure that the reconstructed 3D expression matches the expression depicted in the input image. Reasoning about humans in images requires estimating not only the face, but the full expressive body. To that end, we present PIXIE. PIXIE combines a body-driven attention scheme with a moderator that merges features of body-part experts to reconstruct 3D bodies with articulated hands and expressive faces directly from images.

### 3.3 Plausible (Neural) Rendering of Bodies & Garments in Motion

*Duygu Ceylan (Adobe Research – London, GB)*


**License**  Creative Commons BY 4.0 International license  
© Duygu Ceylan

**Joint work of** Duygu Ceylan, Meng Zhang, Niloy J. Mitra, Tuanfeng Wang, Jae Shin Yoon, Cynthia Lu, Jimei Yang, Hyun Soo Park, Zhixin Shu

While there has been a lot of work on capturing 3D human body pose from single images and learning to generate pose-conditioned human models either in 2D or 3D, a relatively less explored area is to model motion dependent deformations. In this talk, I will discuss some of my recent work in utilizing motion features to synthesize plausible garments in 2D or 3D. I will specifically point out the main challenges and speculate on potential directions.

### 3.4 Inferring people’s anatomic skeleton from their external appearance

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

**License**  Creative Commons BY 4.0 International license  
© Marilyn Keller

**Joint work of** Marilyn Keller, Silvia Zuffi, Michael J. Black, Sergi Pujades  
**Main reference** Marilyn Keller, Silvia Zuffi, Michael J. Black, Sergi Pujades: “OSSO: Obtaining Skeletal Shape from Outside”, CoRR, Vol. abs/2204.10129, 2022.  
**URL** <https://doi.org/10.48550/arXiv.2204.10129>

Modeling the human internal anatomy is key in medicine and biomechanics. While many statistical models of the human being have been developed, those mainly describe their external appearance or individual bones. In this talk, I present how we learn a statistical model of the whole skeleton and its correlation with the body shape.

We do so using 1000 male and 1000 female dual-energy X-ray absorptiometry (DXA) scans. To these, we fit a parametric 3D body shape model (STAR) to capture the body surface and a novel part-based 3D skeleton model to capture the bones. This provides inside/outside training pairs. We model the statistical variation of full skeletons using PCA in a pose-normalized space. We then train a regressor from body shape parameters to skeleton shape parameters and refine the skeleton to satisfy constraints on physical plausibility. We name our inference tool OSSO, for “Obtaining Skeletal Shape from Outside”. Given an arbitrary 3D body shape and pose, OSSO predicts a realistic skeleton inside.

### 3.5 Deep Signatures – Learning Invariants of Planar Curves

*Ron Kimmel (Technion – Haifa, IL) and Roy Velich (Technion – Haifa, IL)*

**License**  Creative Commons BY 4.0 International license  
© Ron Kimmel and Roy Velich

According to an important theorem by É. Cartan [1, 2], two planar curves are related by a group action, if and only if their signature curves, with respect to a given transformation group, are identical. Signature curves are parametrized by the group’s differential invariants. Therefore, differential invariants provide a fundamental building block for the solution of the equivalence problem of planar curves, and geometric structures in general. We propose a learning paradigm for numerical approximation of differential invariants of planar curves.


Deep neural-networks' (DNNs) universal approximation properties are utilized to estimate geometric measures. The proposed framework is shown to be a preferable alternative to axiomatic constructions. Specifically, we show that DNNs can learn to overcome instabilities and sampling artifacts and produce numerically-stable signatures for curves subject to a given group of transformations in the plane. We compare the proposed schemes to alternative state-of-the-art axiomatic constructions of group invariant arc-lengths and curvatures. We evaluate our models qualitatively and quantitatively and propose a benchmark dataset to evaluate approximation models of differential invariants of planar curves.

### References

- 1 E. Cartan. *La méthode du repère mobile, la théorie des groupes continus et les espaces généralisés*. The Mathematical Gazette, 1935
- 2 Olver, Peter J. *Classical Invariant Theory*. Cambridge University Press, 1999

## 3.6 Computer Vision does not generalize – 3DMMs and beyond can help

Adam Kortylewski (*MPI für Informatik – Saarbrücken, DE*)

License  Creative Commons BY 4.0 International license  
© Adam Kortylewski

Main reference Angtian Wang, Adam Kortylewski, Alan L. Yuille: “NeMo: Neural Mesh Models of Contrastive Features for Robust 3D Pose Estimation”, CoRR, Vol. abs/2101.12378, 2021.

URL <https://arxiv.org/abs/2101.12378>

In this talk, I pointed out a fundamental issue in computer vision research, namely that there is a large gap between the performance of vision models on academic benchmarks and their generalization ability in real-world applications. As an illustrative example, I contrasted the outstanding performance of vision models on popular and challenging benchmarks with the still unsolved problem of detecting simple STOP signs with self-driving cars. The fundamental issue is that we assume in academic benchmarks that the training and test data are very similar (i.e. i.i.d. distributed), while autonomous systems that interact with the real-world are often confronted with data that is, in some aspect, different from what has been observed at training time (e.g. unseen illumination, context, occlusion, texture, etc.). In the second half of this talk, I discussed how advances in statistical generative models and neural rendering could potentially help to close the generalization gap. I referred to some of our recent work on integrating deep neural networks with statistical generative models in 2D [2] and 3D-aware architectures [1], which enabled machines to generalize to unseen occlusion, to perform amodal segmentation [5], and to reason about occlusion ordering [3]. I also discussed recent work where we used generative models to benchmark vision systems through adversarial examination [4], and efforts to design new datasets that focus on capturing real-world generalization [6].

### References

- 1 Wang, Angtian, Adam Kortylewski, and Alan Yuille. “NeMo: Neural Mesh Models of Contrastive Features for Robust 3D Pose Estimation.” International Conference on Learning Representations. 2021.
- 2 Kortylewski, Adam, et al. “Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

- 3 Yuan, Xiaoding, et al. “Robust instance segmentation through reasoning about multi-object occlusion.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
- 4 Ruiz, N., Kortylewski, A., Qiu, W., Xie, C., Bargal, S. A., Yuille, A., and Sclaroff, S. (2021). Simulated Adversarial Testing of Face Recognition Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
- 5 Sun, Yihong, Adam Kortylewski, and Alan Yuille. “Amodal Segmentation through Out-of-Task and Out-of-Distribution Generalization with a Bayesian Model”. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
- 6 Zhao, Bingchen, et al. “ROBIN: A Benchmark for Robustness to Individual Nuisances in Real-World Out-of-Distribution Shifts.” arXiv preprint arXiv:2111.14341 (2021).

### 3.7 NeRF for View Synthesis

*Ben Mildenhall (Google – London, GB)*

**License** © Creative Commons BY 4.0 International license  
© Ben Mildenhall

**Joint work of** Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, Jonathan Barron, Peter Hedman, Dor Verbin, Todd Zickler

**Main reference** Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng: “NeRF: representing scenes as neural radiance fields for view synthesis”, Commun. ACM, Vol. 65(1), pp. 99–106, 2022.

**URL** <https://doi.org/10.1145/3503250>

Recent years have seen a massive jump in quality for the task of photorealistic novel view synthesis, mainly driven by hybrid neural rendering pipelines in which part or all of the scene representation or rendering process are optimized for final image quality using gradient descent. Our group at Google has focused on pushing further towards higher resolution, bigger scenes, and more physically accurate view synthesis, with the hopes that progress on representations and rendering methods for this underpinning task can be fruitfully transferred to many other problems in 3D vision. I will give a brief overview of our recent work on extending NeRF to perform better on large scenes, shiny objects, and noisy camera data.

### 3.8 Deep Relighting of 3D Faces

*Shunsuke Saito (Reality Labs – Pittsburgh, US)*

**License** © Creative Commons BY 4.0 International license  
© Shunsuke Saito

**Joint work of** Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, Jason M. Saragih

**Main reference** Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, Jason M. Saragih: “Deep relightable appearance models for animatable faces”, ACM Trans. Graph., Vol. 40(4), pp. 89:1–89:15, 2021.

**URL** <https://doi.org/10.1145/3450626.3459829>

We present a method for building high-fidelity animatable 3D face models that can be posed and rendered with novel lighting environments in real-time. Our main insight is that relightable models trained to produce an image lit from a single light direction can generalize to natural illumination conditions but are computationally expensive to render. On the other hand, efficient, high-fidelity face models trained with point-light data do not generalize to novel lighting conditions. We leverage the strengths of each of these two approaches. We first



train an expensive but generalizable model on point-light illuminations, and use it to generate a training set of high-quality synthetic face images under natural illumination conditions. We then train an efficient model on this augmented dataset, reducing the generalization ability requirements. As the efficacy of this approach hinges on the quality of the synthetic data we can generate, we present a study of lighting pattern combinations for dynamic captures and evaluate their suitability for learning generalizable relightable models. Towards achieving the best possible quality, we present a novel approach for generating dynamic relightable faces that exceeds state-of-the-art performance. Our method is capable of capturing subtle lighting effects and can even generate compelling near-field relighting despite being trained exclusively with far-field lighting data.

### 3.9 Digital Humans in Motion

*Justus Thies (MPI für Intelligente Systeme – Tübingen, DE)*

**License** © Creative Commons BY 4.0 International license  
© Justus Thies

The main theme of my work is to capture and to (re-)synthesize the real world using commodity hardware. It includes the modeling of the human body, tracking, as well as the reconstruction and interaction with the environment. The digitization is needed for various applications in AR/VR as well as in movie (post-)production. Teleconferencing and remote collaborative working in VR is of high interest since it is the next evolution step of how people communicate. A realistic reproduction of appearances and motions is key for such applications. Capturing natural motions and expressions as well as the photorealistic reproduction of images under novel views are challenging. With the rise of deep learning methods and, especially, neural rendering, we see immense progress to succeed in these challenges. In this talk, I will focus on the image synthesis of humans, the underlying representation of appearance, geometry, and motion to allow for explicit and implicit control over the synthesis process.

### 3.10 A Structured Latent Space for Human Body Motion Generation

*Stefanie Wuhler (INRIA – Grenoble, FR)*

**License** © Creative Commons BY 4.0 International license  
© Stefanie Wuhler

**Joint work of** Mathieu Marsot, Stefanie Wuhler, Jean-Sebastien Franco, Stephane Durocher  
**Main reference** Mathieu Marsot, Stefanie Wuhler, Jean-Sebastien Franco, Stephane Durocher: “A structured latent space for human body motion generation”, arXiv, 2021.  
**URL** <https://doi.org/10.48550/ARXIV.2106.04387>

We study learning a structured latent space to represent and generate temporally and spatially dense 4D human body motion. Once trained, the proposed model generates a multi-frame sequence of dense 3D meshes based on a single point in a low-dimensional latent space. This latent motion representation can be learned in a data-driven framework that builds upon two existing lines of works. The first analyzes temporally dense skeletal data to capture the global displacement, poses and temporal evolution of the motion, while the second analyzes static densely captured human scans in 3D to represent realistic 3D human body surfaces in a low-dimensional space. Building upon the respective advantages of these

two concepts allows the model to simultaneously represent temporal motion information for sequences of varying duration and detailed 3D geometry at every time instant of the motion. Experiments demonstrate that the resulting latent space is structured in the sense that similar motions form clusters in this space, and that the latent space allows to generate plausible interpolations between different actions.

### 3.11 Implicit 3DMMs for Full Heads Including Hair

*Tarun Yenamandra (TU München, DE)*

**License** © Creative Commons BY 4.0 International license  
© Tarun Yenamandra

**Joint work of** Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, Christian Theobalt

**Main reference** Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, Christian Theobalt: “i3DMM: Deep Implicit 3D Morphable Model of Human Heads”, in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pp. 12803–12813, Computer Vision Foundation / IEEE, 2021.

**URL** [https://openaccess.thecvf.com/content/CVPR2021/html/Yenamandra\\_i3DMM\\_Deep\\_Implicit\\_3D\\_Morphable\\_Model\\_of\\_Human\\_Heads\\_CVPR\\_2021\\_paper.html](https://openaccess.thecvf.com/content/CVPR2021/html/Yenamandra_i3DMM_Deep_Implicit_3D_Morphable_Model_of_Human_Heads_CVPR_2021_paper.html)

3DMMs are morphable models of human faces. Existing mesh-based 3DMMs consider only a part of the human head, commonly the face region. While some also model the shape of the head, no existing 3DMM can model hair along with other features of human heads. This is partly due to the unavailability of full head data and due to the challenges in modeling hair with mesh-based representations. Can an implicit representation-based 3DMM help solve some of the limitations? What are the challenges of such models?

### 3.12 Towards Precise Completion of Deformable Shapes

*Oshri Halimi (Technion – Haifa, IL)*

**License** © Creative Commons BY 4.0 International license  
© Oshri Halimi

**Joint work of** Oshri Halimi, Ido Imanuel, Or Litany, Giovanni Trappolini, Emanuele Rodolà, Leonidas J. Guibas, Ron Kimmel

**Main reference** Oshri Halimi, Ido Imanuel, Or Litany, Giovanni Trappolini, Emanuele Rodolà, Leonidas J. Guibas, Ron Kimmel: “Towards Precise Completion of Deformable Shapes”, in Proc. of the Computer Vision – ECCV 2020 – 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIV, Lecture Notes in Computer Science, Vol. 12369, pp. 359–377, Springer, 2020.

**URL** [https://doi.org/10.1007/978-3-030-58586-0\\_22](https://doi.org/10.1007/978-3-030-58586-0_22)

According to Aristotle, “the whole is greater than the sum of its parts”. This statement was adopted to explain human perception by the Gestalt psychology school of thought in the twentieth century. Here, we claim that when observing a part of an object which was previously acquired as a whole, one could deal with both partial correspondence and shape completion in a holistic manner. More specifically, given the geometry of a full, articulated object in a given pose, as well as a partial scan of the same object in a different pose, we address the new problem of matching the part to the whole while simultaneously reconstructing the new pose from its partial observation. Our approach is data-driven and takes the form of a Siamese autoencoder without the requirement of a consistent vertex labeling at inference time; as such, it can be used on unorganized point clouds as well as on triangle meshes. We demonstrate the practical effectiveness of our model in the applications of single-view deformable shape completion and dense shape correspondence, both on synthetic and real-world geometric data, where we outperform prior work by a large margin.

### 3.13 Do We Still Need to Detect Faces and Facial Landmarks? Do We Need to Estimate 3D Face Shapes?

Tal Hassner (*Facebook AI, California, US*)

License  Creative Commons BY 4.0 International license  
© Tal Hassner

This talk aims to challenge long held and widely popular best practices for designing digital face processing pipelines. Specifically, nearly all face processing pipelines begin with face detection. Many systems then continue by localizing 2D facial landmarks for each detected face, a step typically used for face alignment and often also for 3D face reconstruction. Finally, depending on the application, some systems also estimate the 3D shape of each face. My talk proposes that these steps may be remnants of legacy designs from a time before effective deep learning was available, and are no longer required for many practical use cases. In fact, not only are these steps redundant, they also add unnecessary compute while introducing noise. As alternatives to these steps, I will share memory and compute efficient solutions for face detection, face alignment, and 3D face rendering. *\*The talk represents work done in academia, prior to joining Facebook / Meta AI and so does not represent that company in any way.*

### 3.14 On Implicit Avatars, Racial Bias, and the Light/Albedo Ambiguity

Victoria Fernández Abrevaya (*MPI für Intelligente Systeme – Tübingen, DE*)

License  Creative Commons BY 4.0 International license  
© Victoria Fernández Abrevaya

Joint work of Victoria Fernández Abrevaya, Yufeng Zheng, Marcel C. Bühler, Xu Chen, Michael J. Black, Otmar Hilliges

We discuss two works covering two different aspects of 3DMMs.

In the first part we present IMAvatar [1], a new method for learning implicit morphable head avatars from videos. Traditional morphable face models provide fine-grained control over expression and pose, but cannot easily capture geometric and appearance details. Neural volumetric representations approach photorealism but are hard to animate, and do not generalize well to unseen expressions. To address this gap we introduce IMAvatar, a novel method for learning head avatars with an implicit representation, directly from monocular videos. Inspired by conventional 3DMMs, IMAvatar represents the expression- and pose-related deformations via learned blendshapes and skinning fields. We employ ray tracing and iterative root-finding to locate the canonical surface intersection for each pixel. The experimental results show that our method improves geometry and covers a more complete expression space, compared to state-of-the-art methods.

In the second part we shift the focus to face appearance. We find that current diffuse albedo estimation methods are biased towards light skin tones due to (1) strongly biased priors that prefer light skin, and (2) algorithmic approaches that do not address the light/albedo ambiguity. We discuss here a solution for the latter that builds on a key observation: the full scene image contains important information about lighting that can be used for disambiguation. Our experimental results show significant improvement compared to state-of-the-art methods on albedo estimation, both in terms of accuracy as well as fairness.

## References

- 1 Y. Zheng, V. F. Abrevaya, M. C. Bühler, X. Chen, M. J. Black, O. Hilliges, *I M Avatar: Implicit Morphable Head Avatars from Videos*. arXiv, 2021.

## 4 Working groups

### 4.1 What are the 'Killer Applications' of 3D Implicit Representations

*Ben Mildenhall (Google Research – London, UK)*

License  Creative Commons BY 4.0 International license  
© Ben Mildenhall

Our first discussion topic, proposed by Ben Mildenhall, asked 'what are the 'killer applications' of 3D implicit representations?'. It was thought that applications in 3D vision are progressing slowly compared to 2D and that perhaps this was due to not yet currently having a killer application. Some felt that judging the slow development of 3D compared to current 2D technology was unfair due to 2D having more than 100 years of development and use in both technology and culture. Potentially, the slow development of 3D is similar to the original slow development of 2D over the past 100 years. It was agreed that we need to convince the general population that 3D computer vision is valuable and that cost of 3D capture technology is a significant factor. The only number the market understands is cost.


It was suggested that whilst one view is on what is tech feasible. Another discussion point is to imagine every consumer with this technology and what they would do with it? What universal adoption would look like? One interesting thing is the tendency of technology and games toward crafting a space to share with others. Animal Crossing, where users build and share islands or VR Chat, is a social VR game with no other objective than creating, sharing, and communicating. These have minimal to no game mechanics, just sharing. People could make an artistically pleasing space where others can gather in Augmented Reality (AR). Others suggested that some of the most exciting applications will be the democratisation of content creation. People either have 3D content creation skills or must go to a special effects company and ask for the content to be created for a price. However, neural rendering will enable massive reductions in the cost and skill required to create models. A simple spoken query to the network and a model will be generated for you.

Some felt that the current limiting factor in creating these experiences is difficulty in controlling or parameterising implicit representations. The killer applications enable users to control the neural representation, ask the game to look the way they like, and the network renders it for them. Along a similar line, it was generally agreed that we need to design and build improved ways of interacting with 3D content easily. 3D content on a 2D interface is not taking full advantage. In 2D, it is difficult for most users to interact with 3D content it is the user interfaces at the moment are challenging for many people. Similar to when 3D printers first arrived, it was the poor user experience that prevented widespread adoption.

Overall it was agreed by all that the future of 3D vision and neural rendering is inspiring and that the next decade will see an explosion of creative and exciting uses of these technologies. There are numerous directions to explore, but the combined progress of research and engineering in hardware and software will enable widespread adoption.

## 4.2 Are Neural Implicit Representations the Future of Morphable Models?

*Christian Theobalt (MPI für Informatik – Saarbrücken, DE)*

License  Creative Commons BY 4.0 International license  
© Christian Theobalt

In the second flexible session, we tested a different format. Splitting into three groups, two physical and one remote, each group discussed the proposed topic for 30 minutes. We then regrouped and shared the results of the discussions with each other.

We concluded that in recent years, the research community had been excited to see that the implicit representation-based methods are achieving good performance in many aspects and are addressing many problems of morphable models, for example, in dealing with self-occlusion problems. They improve the rendering quality of scenes, faces, and objects and are also more efficient in the sense that only several images from different views are required for training. The trained network now being able to interpolate pose and generate images of almost the same quality. Compared to the morphable models using meshes, these methods do not need dense correspondences for supervision. Some methods do not even require information regarding the pose, volume or camera parameters.

On the other hand, we encourage further explorations of the generative ability of the implicit representation-based methods. We know that many questions for the morphable models or the mesh renderers still hold for the implicit models. For example, to build models with good generative ability, achieve relighting, render different materials, separate the albedo with shadow and highlight without losing any details, etc. It is also important to think about having smooth control over these properties. The morphable models solve this with dense correspondences, and the deformation space is well modelled so that we can have parametric control over specific properties. We encourage further thinking on how such correspondences should be encoded and how we can have smooth control over the deformations for the implicit models. Combining the Generative Adversarial Networks (GANs) is promising in this direction, yet we also expect other researchers to explore if a statistical model based on implicit representations is feasible or not. Implicit representations solve many problems that the morphable models have, yet we encourage further thinking on how to reach the same generative ability and smooth control over certain properties.

## 4.3 How Much 3D is Needed?

*Duygu Ceylan (Adobe Research – London, GB)*

License  Creative Commons BY 4.0 International license  
© Duygu Ceylan

In this workshop session, several questions were proposed for discussion. What is the best representation? Are 2D workflows good enough? Do we need explicit 3D? Is course 3D good enough? How much 3D do we need for 2D synthesis? Furthermore, how do we scale 3D methods to “in-the-wild” robustly? The group felt this was an old discussion that had been going on for many years.

The choice of needing 3D over 2D is very much task-dependent. Traditionally cartoons were created in 2D. Now, they are 3D even though they sometimes represent 2D as it is a more accessible representation to work with. Arguing over the coordinate system should

be decided by the problem. For example, people used to argue over what edge detection is the best, but it was shown that all edge detections could be viewed as the same algorithm and depend on the task. It depends on the objective function. There is no right coordinate system for all problems. Depending on the cost function, if the cost function enforces 3D consistency, it was suggested that networks would learn to use 3D implicitly and learn from data as its the lowest energy state. Generally, it was agreed that one could do everything using just 2D, but 3D might need orders of magnitude less data, and it will be easier to learn.

The discussion amongst the group naturally flowed into a debate about how much humans learn from data and how much we use prior knowledge. For a lot of things we are now approximating using neural networks, we have physical theories; could we find these theories within the networks. Alternatively, could we use the networks to infer new theories about physical systems. It was argued that we would not be able to find physical laws from the weights of networks, but perhaps discovering laws that govern human behaviour is possible. The core issue is having the network communicate this information to us. For example, something like MuZero can play Go at a very high level, but we cannot understand the meaning behind its actions.

#### 4.4 Metrical 3D Reconstruction of the Human Face

*Justus Thies (MPI für Intelligente Systeme – Tübingen, DE)*

License  Creative Commons BY 4.0 International license  
© Justus Thies

Recent publications argue that a 3D reconstruction is need for various applications in AR/VR. These applications often rely on a metrically plausible reconstruction, since the face/human is displayed in a metrical environment (i.e., objects have a known scale). However, benchmarks for 3D face reconstruction like NoW [1] use a scale-invariant evaluation scheme (they search for an optimal scale to align the prediction and the reference). Measuring actual metrical reconstruction errors (i.e., only allowing for a rigid alignment without scaling) leads to a significantly different benchmark results and rankings. This discrepancy has to be discussed more prominently in the literature, since ideally we like to reconstruct metrical faces.

#### References

- 1 S. Sanyal, T. Bolkart, H. Feng, M. Black, *Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision*, Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019.

#### 4.5 Synthetic Data Generation Using Morphable Models

*Federica Bogo (Meta Reality Labs Research – Zürich, CH)*

License  Creative Commons BY 4.0 International license  
© Federica Bogo

Morphable models have been successfully used to generate synthetic data. Recent examples are the FaceSynthetics [1] and AGORA [2] datasets. These works show how synthetic data generation is a flexible, powerful tool to train machine learning models, which can be used on real image data captured in the wild. This session focused on analysing the strengths and limitations of current synthetic pipelines and identifying how we can better leverage morphable models to generate higher-quality data.

It was suggested that an important question is how much one should focus on photo-realism. Minor discrepancies in local pixel intensity statistics between generated images and real ones can severely harm inference. However, rendering all the components realistically in an image (people, their interactions with the scene, background etc.) is very challenging in practice; this suggests one might need to invest significant effort (including manual work from technical artists) to obtain high-quality data.

Moreover, even with high-quality synthetic data, we might not be able to adequately capture the long-tailed distributions commonly found in the real world. A potential way to identify limitations in the data could be to enable a feedback loop from the trained model, tested on new examples, back to the training data itself to highlight problematic cases. Currently, the research community lacks the tooling to generate realistic synthetic data at scale quickly. We see some efforts in this direction [3], but there is still a lot to do for humans.


In the end, morphable models can play a crucial role in enabling a “virtuous cycle” from data collection to inference – which allows us to understand the world through our models – to synthetic data generation: better inference can enable the generation of higher-quality data; in turn, higher-quality generated data can enable the development of more accurate and robust models to perceive the world.

## References

- 1 E. Wood et al. *Fake it till you make it: Face analysis in the wild using synthetic data alone*. ICCV 2021.
- 2 P. Patel et al. *AGORA: Avatars in geography optimized for regression analysis*. CVPR 2021.
- 3 K. Greff et al. *Kubric: A scalable dataset generator*. CVPR 2022.

## 4.6 Morphable Models from Physics

Dan Casas (*Universidad Rey Juan Carlos – Madrid, ES*)

License  Creative Commons BY 4.0 International license  
© Dan Casas

State-of-the-art data-driven approaches to model 3D garment deformations are trained using supervised strategies that require large datasets, usually obtained by expensive physics-based simulation methods or professional multi-camera capture setups. An alternative is to use physics-based deformation models. Formulating the problem as a set of physics-based loss terms that can be used to train neural networks without precomputing ground-truth data. Can this approach be applied to more areas of morphable models and computer vision in general?

The group liked this idea seeing it as a variant of thinking about what explicit real-world knowledge we have that can be used to simplify the situation. However, it gets more complicated when attempting to include all other physical properties, such as material parameters and friction in garment modelling. No physical simulation can do that at the moment; parameters of materials that are entirely unknown, the mass of yarns, for example, and other material properties are not fully modelled. This massive gap between the real and simulation world limits the application of these approaches. This simulation gap resulted in the suggestion that these models should really be called ‘physics inspired’ models as you are making a physical assumption about the world in the simulation, using a subspace of the physics.

Physics models are simplifications, so they are limited to the space that the model can represent; perhaps combining this first-principles approach with a data-driven refinement phase could be a way to go. Any explicit knowledge we can give the model aids in finding solutions with a good prior who needs the data. These approaches could help with explainability, as it is now possible to see what the model is understanding and doing to produce the effects.

Some expressed concern over how we combine physics models with neural scene representations. Understanding physics could help generalisation but the group see a challenge in combining these two. One of the benefits of implicit neural representations is that it can be done without correspondence.


## 4.7 Ethics

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

*William Smith (University of York, GB)*

*Christian Theobalt (MPI für Informatik – Saarbrücken, DE)*

*Stefanie Wuhrer (INRIA – Grenoble, FR)*

**License**  Creative Commons BY 4.0 International license  
 © Bernhard Egger, William Smith, Christian Theobalt, Stefanie Wuhrer

Our final workshop session was reserved for an extended discussion focusing on ethics. So often, as researchers, we are focused on new functionalities and challenges. However, what we are starting to develop now, for example, technologies that identify an individual's emotions, actions and identities, whilst enabling great opportunities, can have use cases that are extremely negative and not in the broader social interest. As leaders in this field, we know more about the possible issues than the general public and politicians. It is our responsibility to identify and predict any negative use cases and conceive of possible options to address these. Predicting negative use cases is a challenging problem, but one we need to have some answers for. To begin to address this, we split this discussion into four sections:

- Threats and ethical concerns.
- What should we do? What is our role?
- What more can our institutions do?
- Essential questions to ask ourselves when starting a project.

### Threats and Ethical Concerns

It was generally agreed that advances in neural rendering whilst having enormous positive use cases could enable some very concerning applications. The rise of Deep Fakes has shown the power of these approaches at producing misinformation, and the level of skill required to use these methods to create fake content is decreasing yearly. Further advances in these approaches to 3D computer vision will only enhance their realism and increase the difficulty of identifying real from fake. Similar unethical use cases involve any non-consensual use of a person's likeness; for example, in pornographic applications or as impersonation when using a telepresence system, it might not be possible to be confident that the person speaking is that person. Whilst privacy concerns over facial recognition systems are well documented. Morphable models are likely to be used to analyse people, emotions, attitudes and intents.



The group also discussed bias in the datasets and models we are developing. All datasets and models are biased, some to the point that the tools they form are potentially useless, offensive or upsetting. Do we try to understand the biases, or must we compensate for the biases somehow? In terms of facial recognition, it is now potentially the most important criterion, more so than accuracy. Research into ways to fix or quantify the bias is essential. However, it was also raised that it is sometimes a conscious decision to increase bias; for example, removing children from datasets might be the right thing to do. Ethical bias is not just due to data bias; an unbiased dataset can still have a model or data that has intrinsic bias. For example, women's faces might naturally have less variation in appearance, making recognition more challenging.

### **What should we do? What is our role?**

The question was raised as to whether we should consider banning a publication. Is it about preventing the research or preventing the use case? This is also challenging as there are many suitable applications of these technologies. It was generally agreed that it is likely we can not keep these algorithms out of the public sphere. That focus should be exclusively on public education and detection and preventative measures. There will always be bad actors, and it will be a cat and mouse game to combat them; we should be focusing on being ahead of these actors, not trying to prevent them from occurring, as that is impossible. Therefore, should we explicitly be working on counter models for the harmful applications of the technology we are developing? This could be an explicit part of the papers and research we do, not just considering.

Some felt that it would be necessary to create an algorithm equivalent to the FDA (Food and Drug Administration) that monitors public use cases of algorithms. Any algorithm released to the public would have to go through this administration. It could also be the focus of such an institution to educate the public. As researchers, we must push policymakers to address issues in these publications and institutions. We should be attempting to have regular meetings with policymakers that involve conversations about the impacts (positive and negative) and keeping people informed. Higher frequency of communication between researchers, the public and policy makers. Many of us have not had an opportunity to speak to policymakers. As academics, we are given by society many resources to do amazing things, and we must give something back. It is up to us to create these connections and start these dialogues. One can gain much from speaking with these groups and discussing these topics outside of one's research group.

### **What more can our institutions do?**

We discussed the different rules for different publications, i.e. CVPR requires approval from an institutional ethical review board, and NeurIPS asks reviewers to flag papers that concern them. Many felt that these rules were not sufficient to self enforce us as a community to think about these aspects. Some think it has to be on the reviewer's side, as asking the others to police themselves is not a practical task. Should it potentially be an ethical review board rather than the responsibility of the reviewers? Alternatively, should there be a multi-disciplinary and consistent board of scientists operating across conferences and checking flagged papers? Others felt ethical statements required by these papers were too

generic to the point they are not helpful, raising concerns that if researchers are only thinking about the ethical question only at the point of writing the paper or at the end of writing the paper, it is already too late. Potentially we should be educating every researcher as a reviewer and give training on ethical considerations such that during the research, they are able to think more deeply about these issues.

### Essential questions before starting a project

Whilst the three previous discussions were fascinating and beneficial we felt it would be of use, especially for junior scientists, to collect a list of questions enabling an actionable output to support researchers in the field. Here we split into five groups, and each group was tasked with thinking of a set of questions one should ask themselves before starting a research project. These were then collated into a single document that could be a helpful starting guide for members of the seminar in future projects. The collated questions are listed below:

- What are potential misuses?
- What companies/institutions might be interested as well?
- What's the worst use case?
- Are people suffering from your research?
- What is the dataset you need to ensure there is no bias in your research? Is bias important for your research/use case?
- Would it be harmful in a democratic country?
- Who is funding the project?
- What is the field of application? Medical, entertainment, military?
- Privacy and consent of the data you would require or collect?
- Are people negatively affected by data collection? (Categorising, terrorist content online) Or obtained via nefarious means?
- Would any contracts or work provided by the project be ethical, provide a good standard of living?
- Will the world be a better place if this research is done?
- Am I using data for the purpose it was intended?
- When we pay participants to have their face or bodies captured, do they truly understand what they are giving away? Do they really understand how it will be used and how it could be used in future?

## Participants

- Thabo Beeler  
Google – Zürich, CH
- Volker Blanz  
Universität Siegen, DE
- Timo Bolkart  
MPI für Intelligente Systeme –  
Tübingen, DE
- Dan Casas  
Universidad Rey Juan Carlos –  
Móstoles, ES
- Bernhard Egger  
Friedrich-Alexander-Universität  
Erlangen-Nürnberg, DE
- Victoria Fernandez Abrevaya  
MPI für Intelligente Systeme –  
Tübingen, DE
- James Gardner  
University of York, GB
- Oshri Halimi  
Technion – Haifa, IL
- Patrik Huber  
University of York, GB
- Marilyn Keller  
MPI für Intelligente Systeme –  
Tübingen, DE
- Ron Kimmel  
Technion – Haifa, IL
- Adam Kortylewski  
MPI für Informatik –  
Saarbrücken, DE
- Chunlu Li  
Universität Basel, CH
- Ben Mildenhall  
Google – London, GB
- Nick Pears  
University of York, GB
- Sami Romdhani  
IDEMIA, FR
- Vincent Sitzmann  
MIT – Cambridge, US
- Ayush Tewari  
MIT – Cambridge, US
- Christian Theobalt  
MPI für Informatik –  
Saarbrücken, DE
- Roy Velich  
Technion – Haifa, IL
- Thomas Vetter  
Universität Basel, CH
- Vanessa Wirth  
Universität Erlangen-  
Nürnberg, DE
- Stefanie Wuhrer  
INRIA – Grenoble, FR
- Tarun Yenamandra  
TU München, DE





### Remote Participants

- Michael J. Black  
MPI für Intelligente Systeme –  
Tübingen, DE
- Federica Bogo  
Meta Reality Labs Research –  
Zürich, CH
- Duygu Ceylan  
Adobe Research – London, GB
- Andrew Fitzgibbon  
Graphcore – Cambridge, GB
- Tal Hassner  
Facebook – Menlo Park, US
- Hao Li  
Pinscreen – Los Angeles and UC  
Berkeley, US
- Xiaoming Liu  
Michigan State University –  
East Lansing, US
- Arianna Rampini  
Sapienza University of Rome, IT
- Emanuele Rodolà  
Sapienza University of Rome, IT
- Shunsuke Saito  
Reality Labs – Pittsburgh, US
- William Smith  
University of York, GB
- Justus Thies  
MPI für Intelligente Systeme –  
Tübingen, DE
- Maximilian Weiherer  
Universität Erlangen-  
Nürnberg, DE
- Jiajun Wu  
Stanford University, US
- Stefanos Zafeiriou  
Imperial College London, GB