A BRIEF STORY OF GENERATIVE MODELLING FOR DIGITAL HUMANS

By MAXIME RAAFAT
with SERGEY PROKUDIN's supervision

19.10.2022

OUTLINE

A Short History of Computer Vision

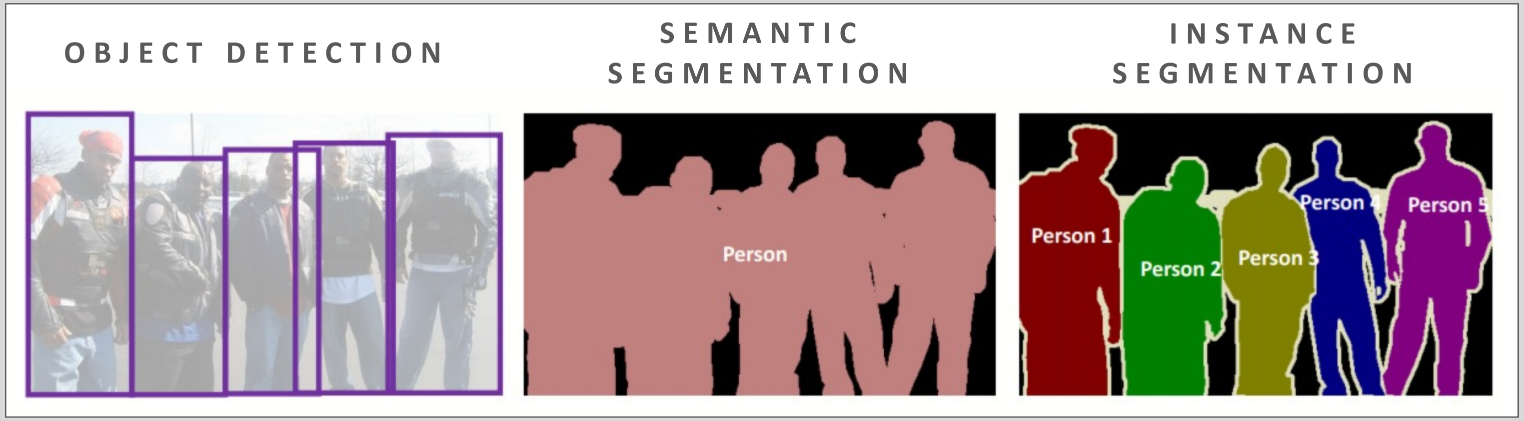
Representation of Digital Human Avatars

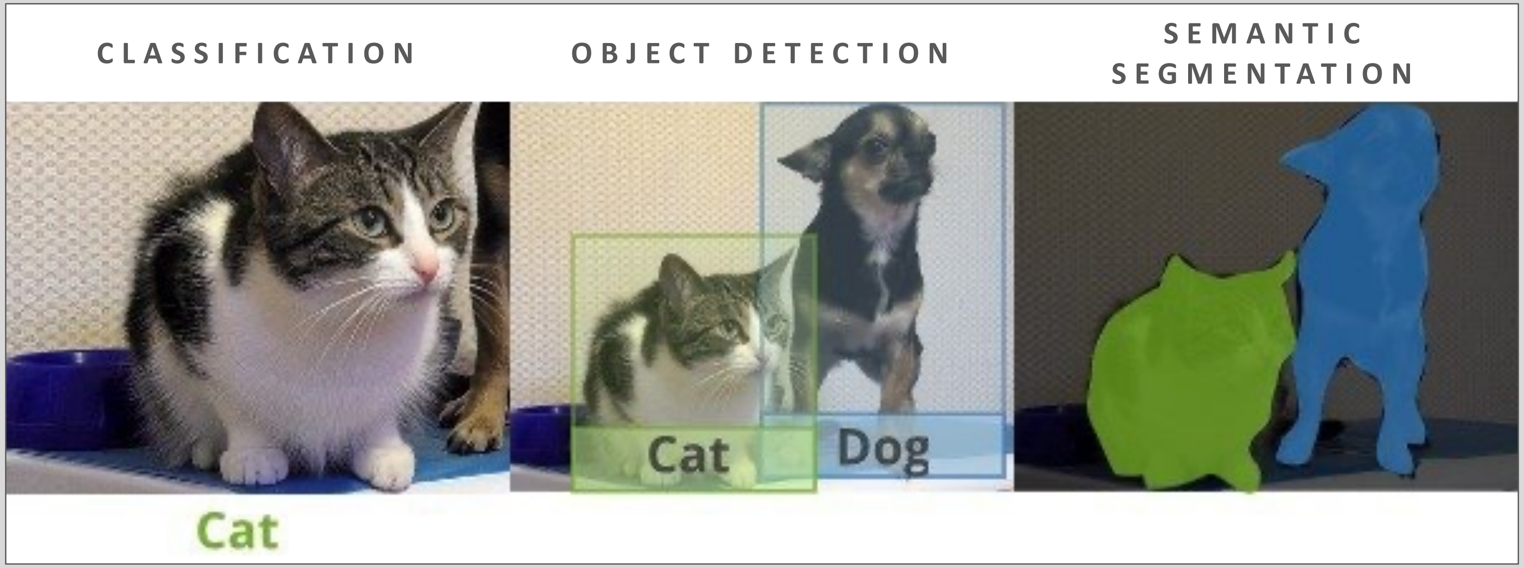
Synthesis of Digital Human Avatars

INTRODUCTION

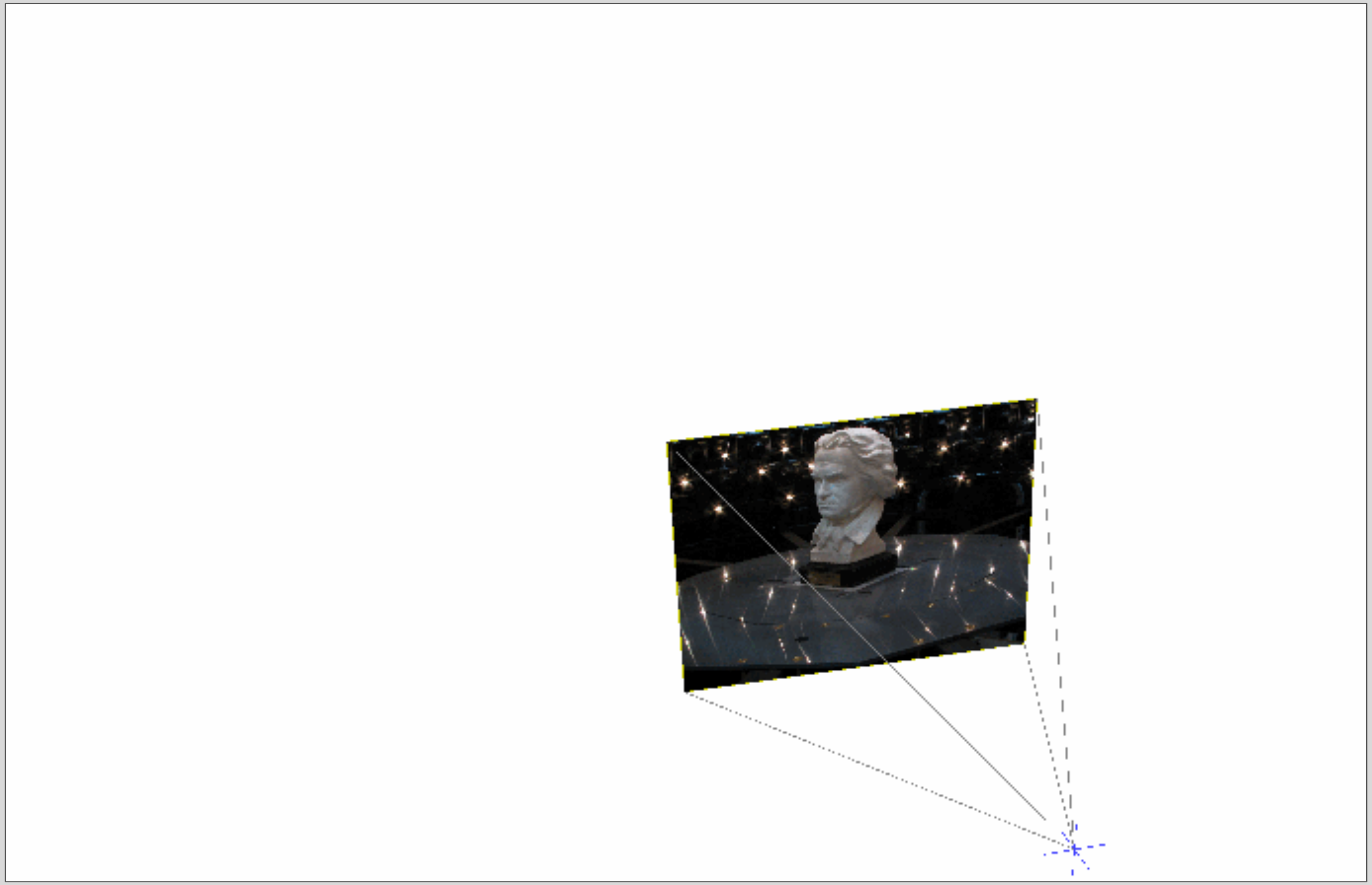
A SHORT HISTORY OF COMPUTER VISION

UNDERSTANDING THE WORLD





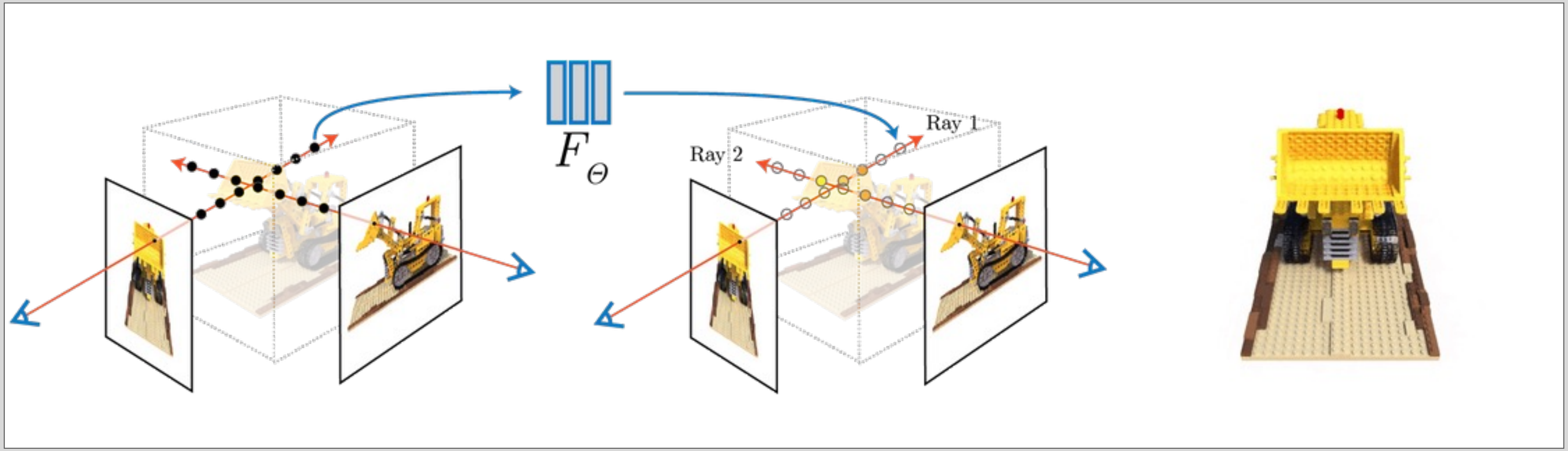
CAPTURING THE WORLD



CAPTURING THE WORLD

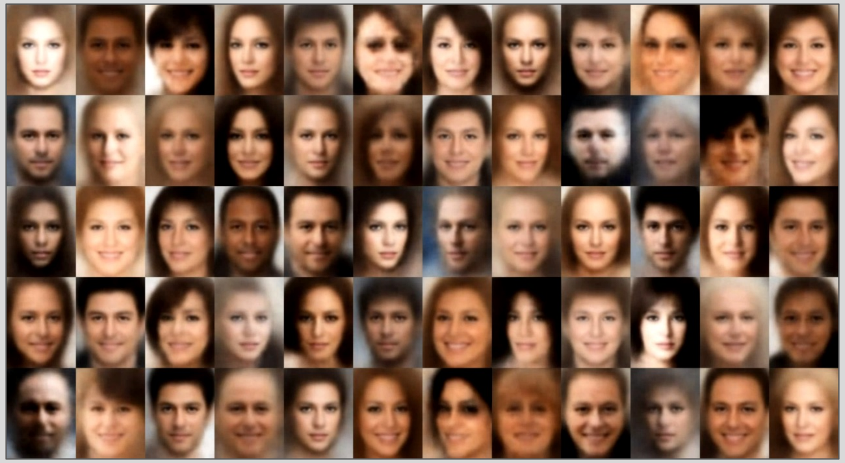
Nowadays, everything is *deep*

Neural Radiance Fields[1]



CREATING NEW WORLDS

VAEs[2]



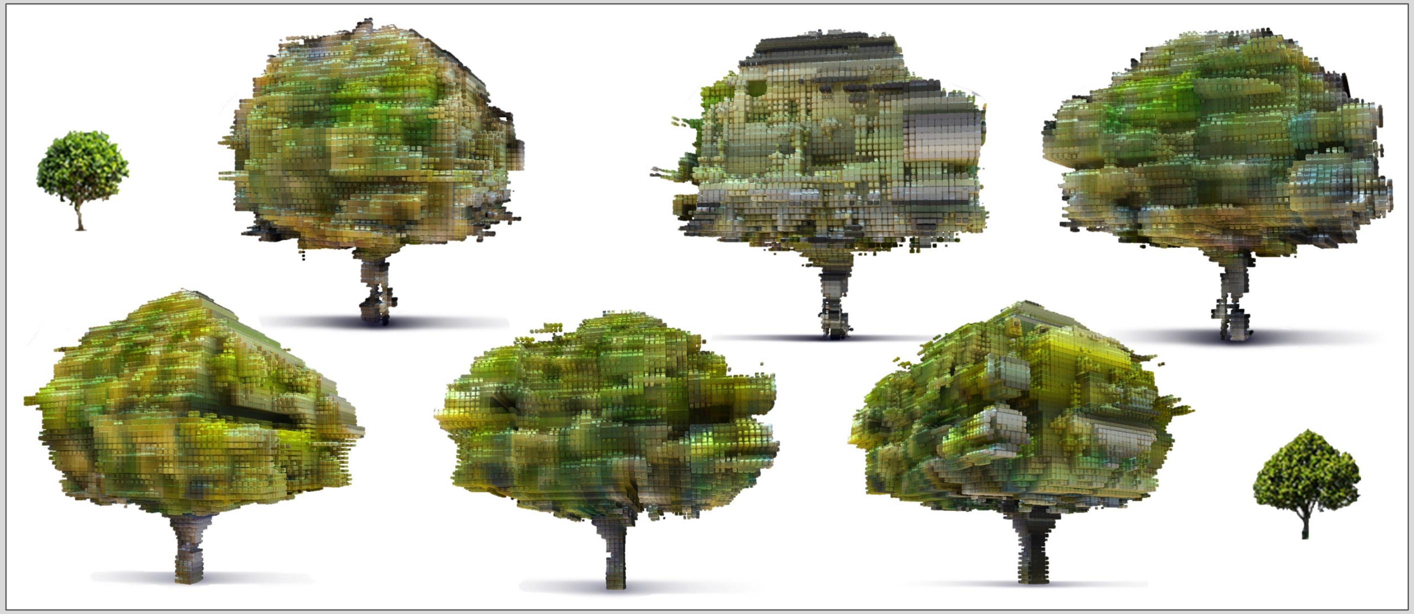
GANs (StyleGAN[3])

DIFFUSION[4]

CREATING NEW WORLDS

Naïve (explicit) CNN-based extensions to 3D

Modern implicit or hybrid efficient 3D generators



PlatonicGAN[5]

EG3D[6]


MICROSOFT
HOLOLENS


GHOST
©2020 SONY INTERACTIVE ENTERTAINMENT LLC
EMRE EKMEKC


META QUEST PRO

CHAPTER I

REPRESENTATION OF DIGITAL HUMAN AVATARS

EXISITING WORKS

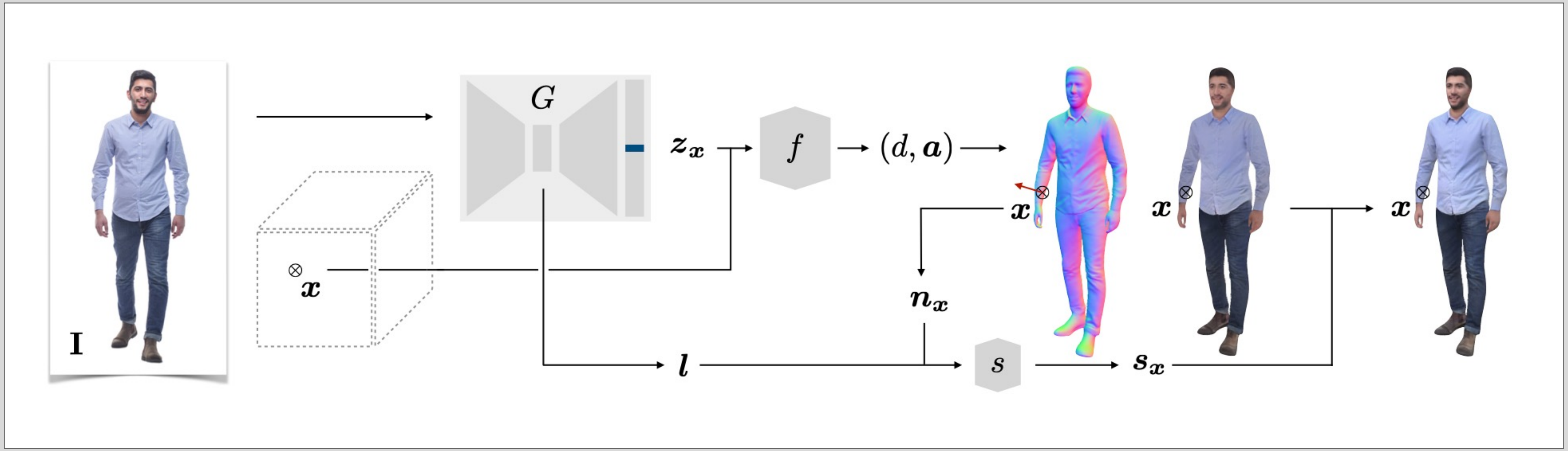
Mesh-based optimization method

Detailed human avatars from monocular video[7] : vertices $V = V_{SMPL} + \Delta V$

EXISITING WORKS

Mesh-based regression method

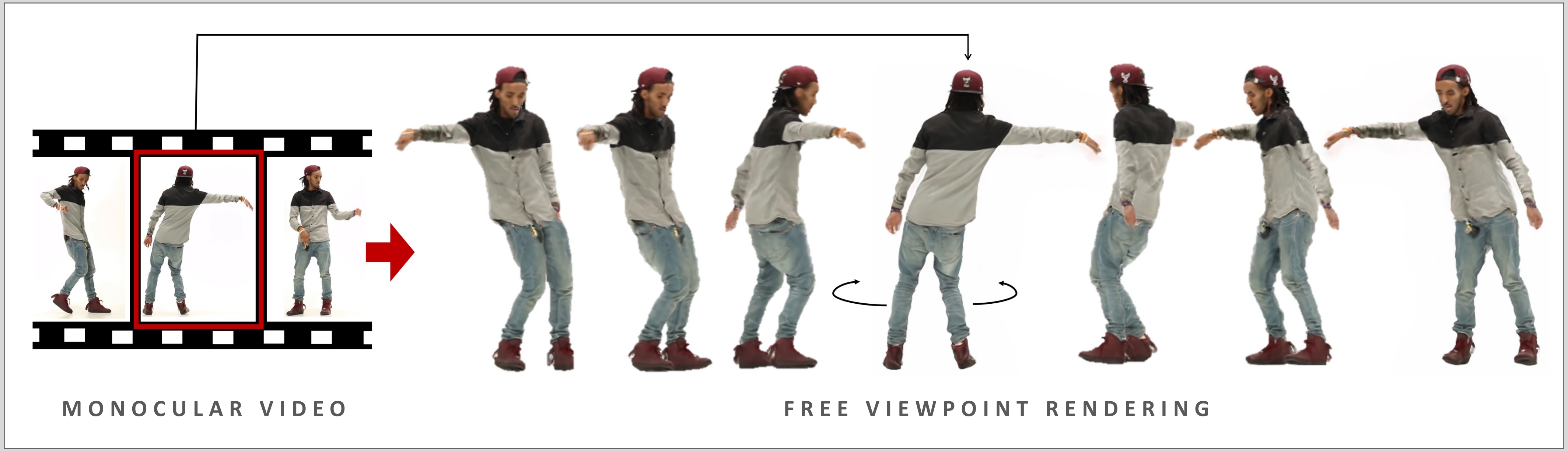
PHORHUM[8] : surface point x shading $s_x = s(a, n_x, l)$



EXISITING WORKS

Implicit volumetric-based method

HumanNeRF[9] : canonical volume + skeletal and non-rigid motions



OUR APPROACH

BY SERGEY PROKUDIN

Expressive *point-based* appearance

UV field $A = [A_{rgb}, A_{\delta}] \in \mathbb{R}^{w_a \times h_a \times 4}$

Point cloud $X = \{x = (x_{xyz}, x_{rgb}) \in \mathbb{R}^6\}$
formed by $x^m_{xyz} \sim$ SMPL-X[10,11] surface

$$x_{xyz} = x^m_{xyz} + \delta \cdot n_{xyz}$$

$$\delta = A_{\delta}[u, v]$$

$$x_{rgb} = A_{rgb}[u, v]$$

CHAPTER II

SYNTHESIS OF DIGITAL HUMAN AVATARS

EXISITING WORKS

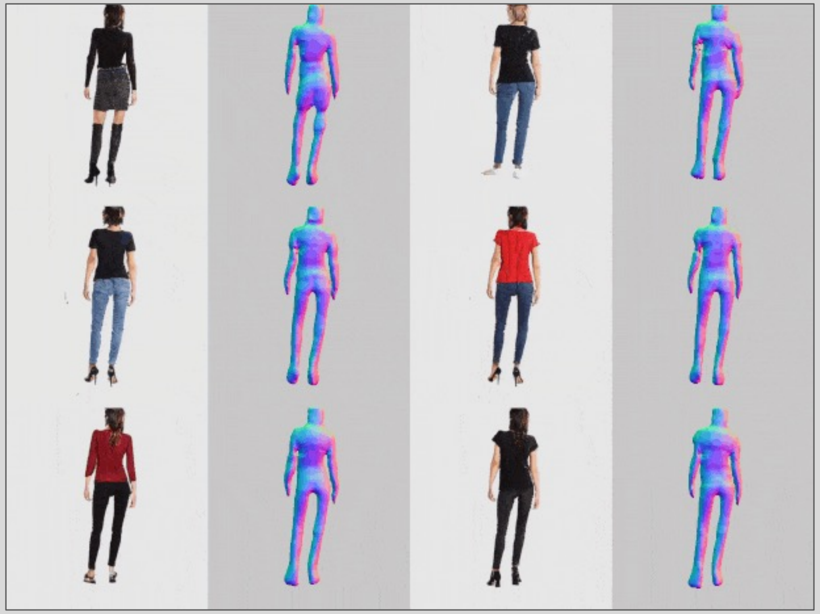
StylePeople[12]

SMPL-X with learned deep features texture + deferred neural renderer

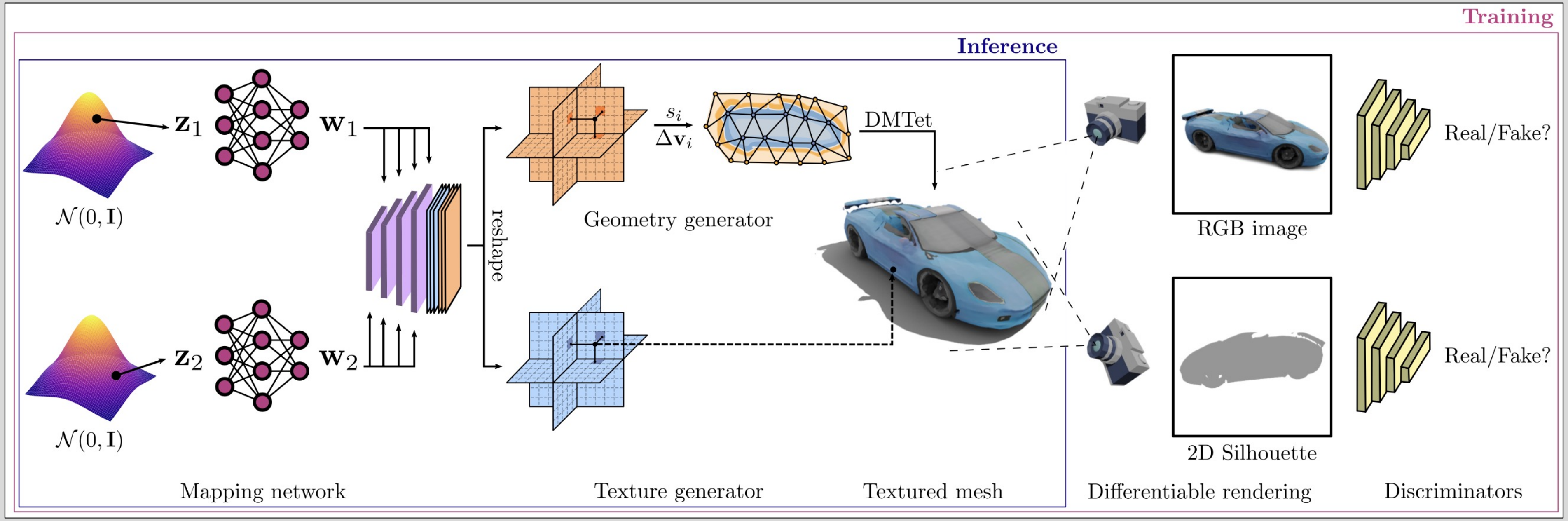
EXISITING WORKS

AvatarGen[13]

EG3D’s tri-plane representation with canonical generation and mapping



EXISITING WORKS



GET3D[14]

EXISITING WORKS

StyleGAN-Human[15]

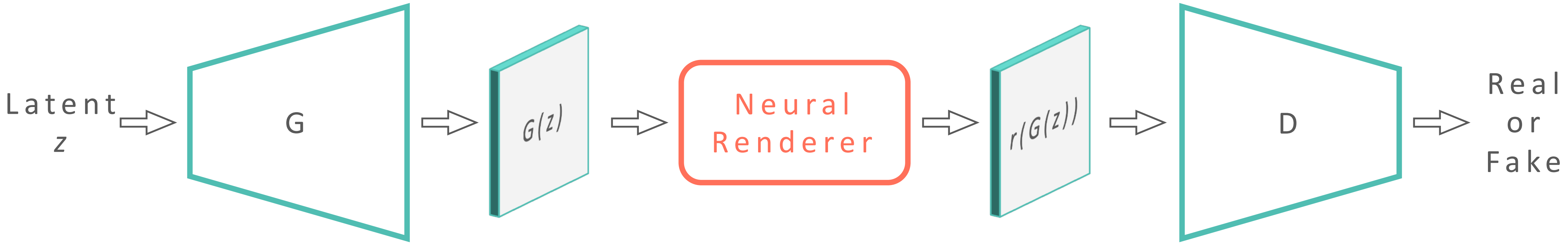
Vanilla StyleGAN trained on SHHQ dataset

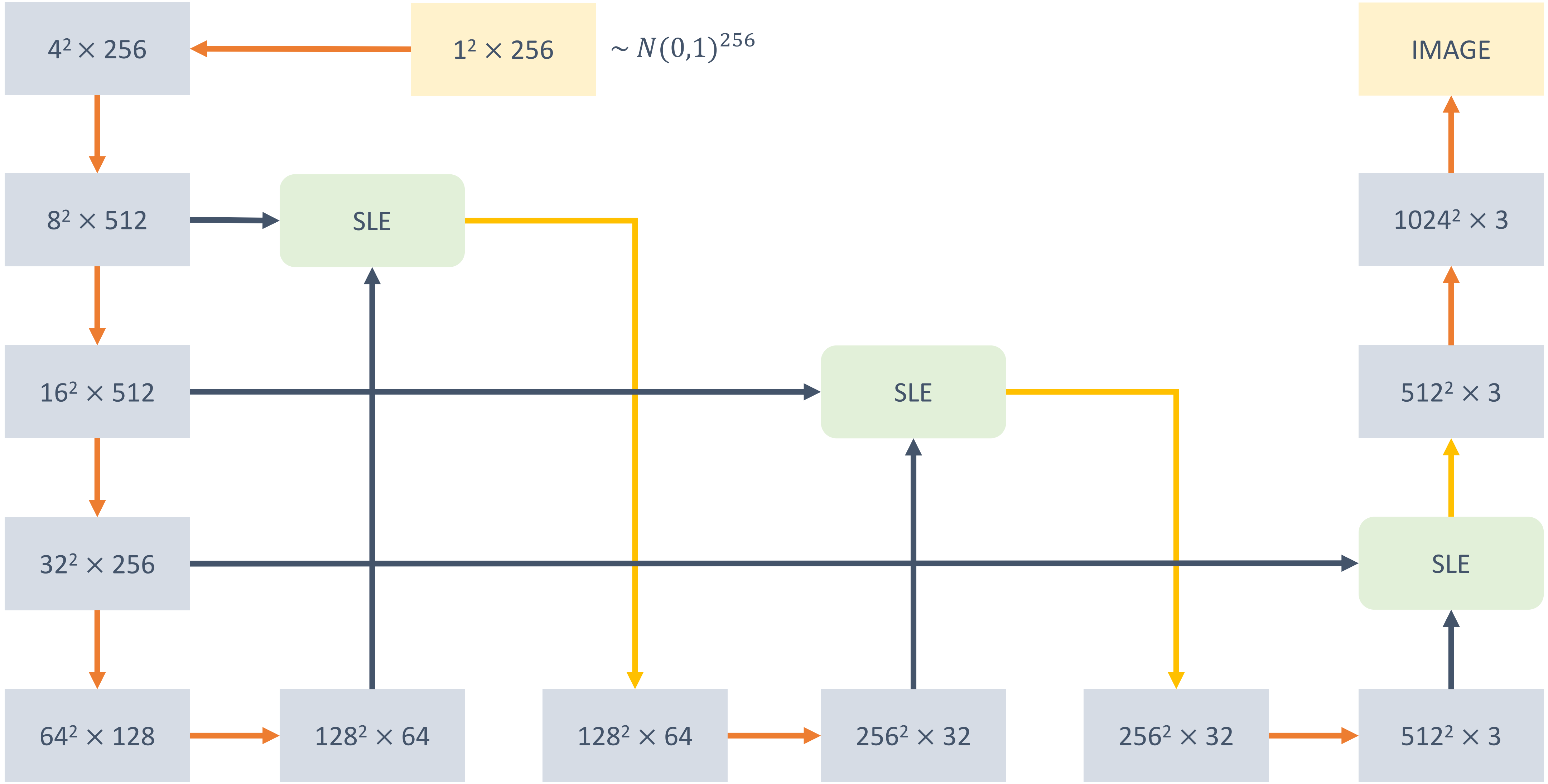
OUR METHOD

3DiGAN pipeline

Lightweight 3D aware *implicit* GAN, operating in an implicit UV state

Feed rendered generated textures to the discriminator, rather than the generated raw outputs





FastGAN[16] GENERATOR

SHHQ × PIXIE[17]

EVALUATIONS

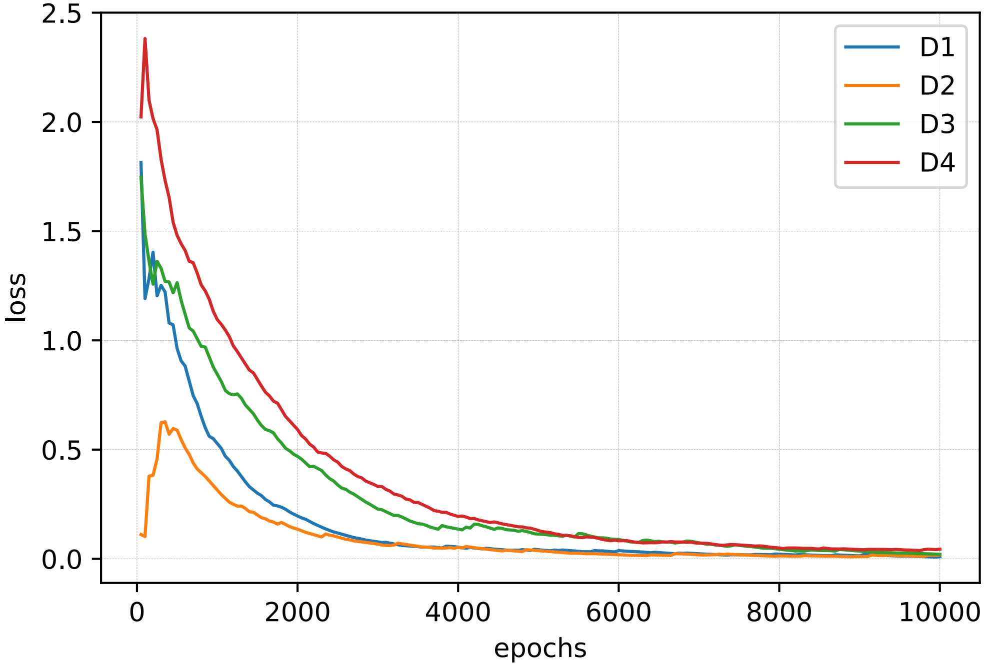
Multi-view single scene, RGB only with true provided geometry

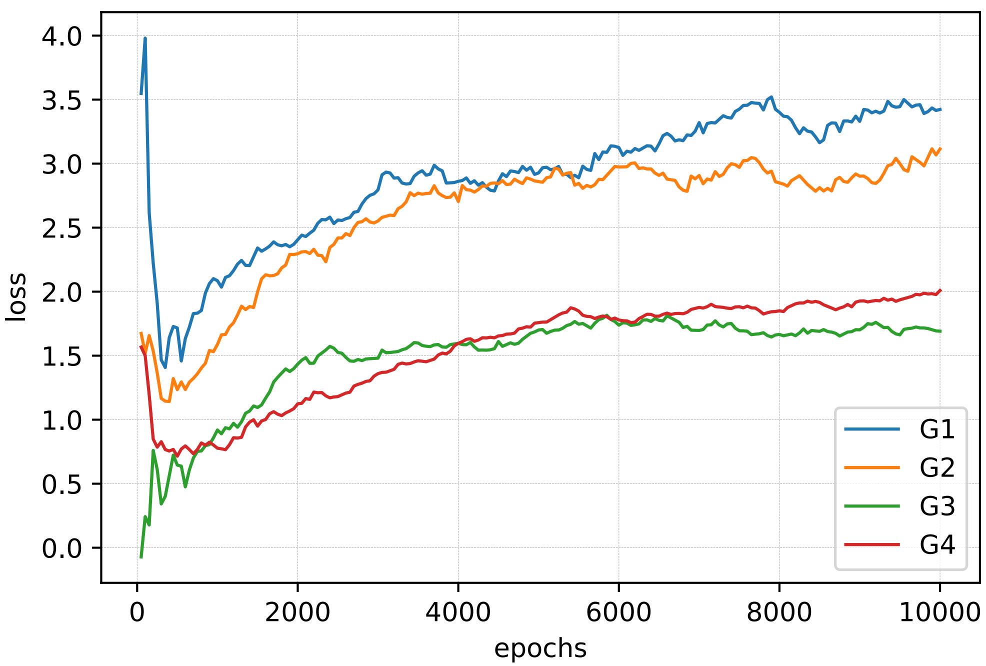
Pulsar[18] sphere rendering at 256×256, radius 0.01 and 10^5 points

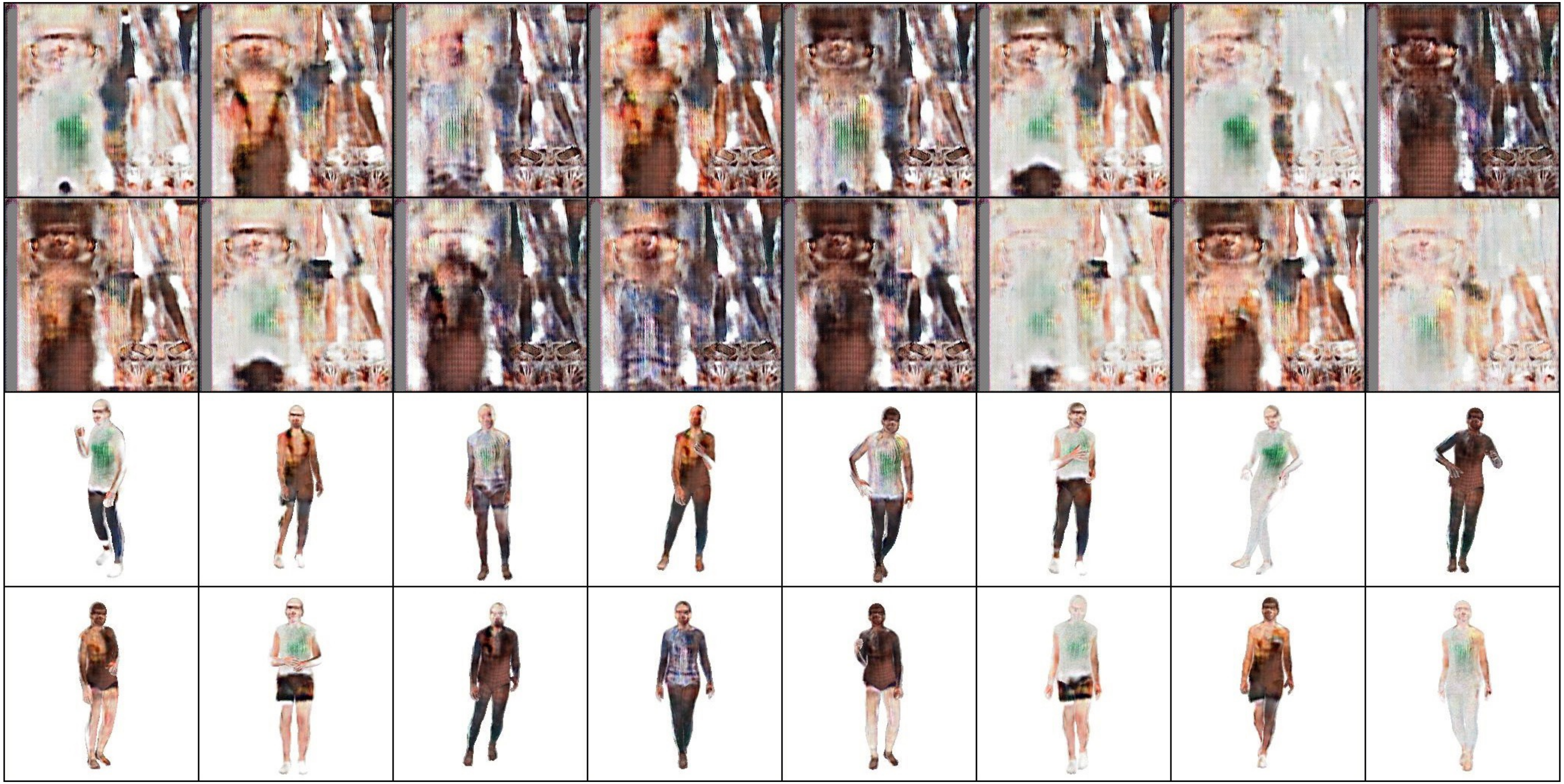
EVALUATIONS

DEFAULT (1) + SMOOTHING (2) + TTUR 0.1 (3) + DROPOUT (4)

EVALUATIONS







SINGLE-VIEW SCENES WITH PSEUDO GROUND TRUTH PIXIE GEOMETRY

LIMITATIONS

Displacement learning currently fails

Joint learning of RGB and geometry is difficult for CNNs

GANs are unstable during training

The UV projection step weakens gradients from D to G

For reconstruction, map Gaussian to single spike distribution with $\sigma = 0$

Pose and illumination dependencies are baked into the UVs

Potential solution? Diffusion Models

REFERENCES

1. Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." Communications of the ACM 65.1 (2021): 99-106.
2. Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
3. Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
4. Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020): 6840-6851.
5. Henzler, Philipp, Niloy J. Mitra, and Tobias Ritschel. "Escaping plato's cave: 3d shape from adversarial rendering." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
6. Chan, Eric R., et al. "Efficient geometry-aware 3D generative adversarial networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
7. Alldieck, Thiemo, et al. "Detailed human avatars from monocular video." 2018 International Conference on 3D Vision (3DV). IEEE, 2018.
8. Alldieck, Thiemo, Mihai Zanfir, and Cristian Sminchisescu. "Photorealistic Monocular 3D Reconstruction of Humans Wearing Clothing." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
9. Weng, Chung-Yi, et al. "Humannerf: Free-viewpoint rendering of moving people from monocular video." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
10. Loper, Matthew, et al. "SMPL: A skinned multi-person linear model." ACM transactions on graphics (TOG) 34.6 (2015): 1-16.
11. Pavlakos, Georgios, et al. "Expressive body capture: 3d hands, face, and body from a single image." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
12. Grigorev, Artur, et al. "Stylepeople: A generative model of fullbody human avatars." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
13. Zhang, Jianfeng, et al. "AvatarGen: a 3D Generative Model for Animatable Human Avatars." arXiv preprint arXiv:2208.00561(2022).
14. Gao, Jun, et al. "GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images." arXiv preprint arXiv:2209.11163 (2022).
15. Fu, Jianglin, et al. "StyleGAN-Human: A Data-Centric Odyssey of Human Generation." arXiv preprint arXiv:2204.11823(2022).
16. Liu, Bingchen, et al. "Towards faster and stabilized gan training for high-fidelity few-shot image synthesis." International Conference on Learning Representations. 2020.
17. Feng, Yao, et al. "Collaborative regression of expressive bodies using moderation." 2021 International Conference on 3D Vision (3DV). IEEE, 2021.

THANK YOU