Neural Rendering Pipeline for Face Inversion and Reconstruction

Alex J. Champandard
creative.ai Research Lab
alex@creative.ai

Abstract

There is a notable gap between generative face models—e.g. VAE or GAN—which mostly operate as a black box with high-level controls yet provide (near) photo-quality outputs, and computer graphics techniques that offer convenient interfaces for detail control but requiring more effort for realistic outputs. In this paper we present the world's first GDPR-compliant face generator that provides data subjects with *"meaningful information about the logic involved,"* and demonstrate an approach that can bridge the aforementioned gap—while almost reaching the precision of graphics pipelines and the realism of generative images.

1 Overview

Our hybrid approach combines research in deep face reconstruction [2] and neural rendering [5]. We propose a differentiable graphics pipeline that's conceptually inspired by ray tracing—except no rays are actually traced and there's no explicit geometry: it's entirely done within a deep learning model.

- Each "asset" (texture, model) is itself learned as a coordinate-conditional generative model.
- Key components of the pipeline (ray tracer, lighting) are approximated by residual networks.
- Shading equations are hand-written based on a physically plausible model of light.

At a high-level, our renderer takes a scene description (e.g. head pose, and facial expression) and pixel coordinate, then returns pixel colors to be assembled into an image. Internally, the renderer relies on multiple components shown in Figure 2:

1. **Projection** — Converts the screen-space pixel coordinates (2D) into model-space rays.
2. **Ray Tracing** — Computes the surface properties for each ray (i.e. depth, normal, UVs).
3. **Texturing** — Looks up the corresponding colors based on UV: specular and diffuse albedo.
4. **Lighting** — Approximates the amount of light (ambient, self-shadow) for each ray cast.
5. **Shading** — Calculates rendering equations per-pixel based on Blinn-Phong equations.

Figure 2 shows our neural rendering pipeline and its various components. The architecture intentionally includes components of graphics pipelines in order to be compatible with standard artist toolchains and methodologies—while remaining fully differentiable. The inclusion of expert graphics knowledge also acts as a strong prior, and our system is able to create representations of faces (i.e. 3D models, 2D textures) directly from few images, when it typically require precise scans.

2 Training

Self-Supervised Learning On the surface, our training method could be considered self-supervised: it requires a small collection of images and is able to learn models based on pixel colors. Specifically, we extract the face pixels of 12-15 photographs depending on the subject—which corresponds to

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Figure 1: When trained on 15-image set and given a input image (top left), our system is able to infer the 3D normals and 2D albedo texture, to reproduce the pixel colors (right).

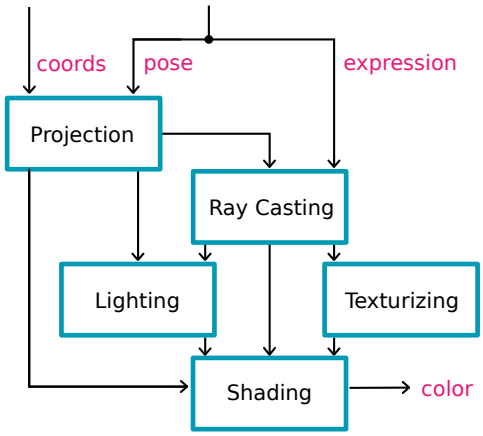

Figure 2: Architecture diagram of the pipeline. Certain components are differentiable but fixed methods (S), others are fully learned (RC+T+L), and some are hybrid (P).

a dataset of up-to 2,068,393 pixels. The inputs to the projection component (pose) and ray-tracing (expression) are encoded from 3D face landmarks extracted once from the image [1].

Reinforcement Learning While the image reconstructions are generally good quality, self-supervision does not consistently produce representations that are compatible with traditional graphics pipelines. Since the problem is under-constrained, we design a simple "curriculum" that incrementally builds up in complexity—effectively a form of macro-level regularization by controlling which components are active at which stage. Each for 100 epochs, we learn to generate: normals, textures, lighting, and fine-tune with all components active.

Optimization-Based As we require few images (but many pixels), our training can also include aspects of optimization-based face reconstruction. Specifically, we jointly optimize the configuration of a single directional light for each image (i.e. position, color, diffuse and specular coefficient) using the RAdam optimizer [3] with batch-size of 32768 pixels. We also adopt a coarse-to-fine approach that starts by approximating a blurred image with MSE loss, then adds more detail incrementally.

3 Evaluation

Since the system is trained only on a dozen images, the quality of the results primarily depends on whether the system was fine-tuned for the person in question; it is less reliable on arbitrary images than [2]. However, we find that it's possible to fit a model and re-synthesize specific images at surprisingly high-quality as shown in 1. Examples on celebrities where no specialized dataset exist requires pre-training, as shown in the Appendix. Compared to 3D morphable models based on polygon meshes, our solution is able to provide approximations for all pixels of the face and does an acceptable job for eyes and mouth.

However, our interest in neural rendering pipelines is more subjective:

1. Does the data within the pipeline conform to expectations of artists? For example, is it acceptable for albedo texture to contain some lighting information (e.g. ambient occlusion)?
2. How easy is it to manipulate the internal representations to obtain different results, but equally as photo-realistic as the original?

For future work, we intend to work more closely with photographers to enable them to use generative tools to make better portraits—and get closer to answering these questions!

Figure 3: Four different photos are reconstructed simultaneously via inverse rendering using a neural rendering pipeline. From left to right: a) fit of the pre-trained morphable model with lighting solved, b) optimization of landmarks to change position of face parts, c) fine-tuning of the neural texture to learn face details, d) the normals that were found as solution, and e) the target image.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[2] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. *ArXiv*, abs/1903.08527, 2019.

[3] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

[4] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2009, 2-4 September 2009, Genova, Italy*, pages 296–301, 2009.

[5] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: image synthesis using neural textures. *ACM Trans. Graph.*, 38:66:1–66:12, 2019.

4 Appendix

When fewer photos are available per person, or the collection of images was not specifically chosen to provide a good variation of lighting and pose, we pre-train the components using off-the-shelf 3D deformable models and their 2d textures [4].

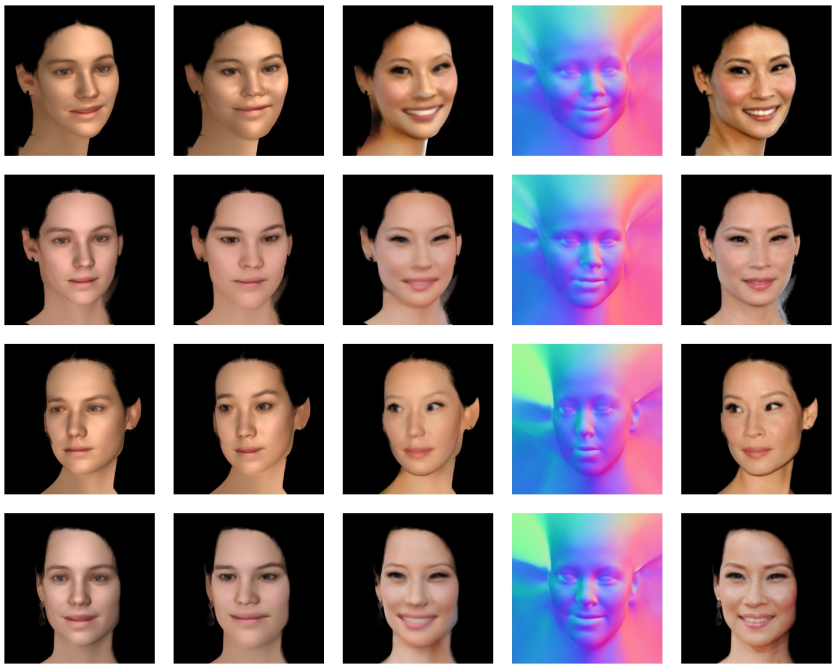

Figure 4: Identical process as Figure 3 for a different celebrity.

Figure 5: Failure case on the bottom row as the jaw is not accurately reproduced.