Neural Radiance Fields for Novel View and Human Pose Synthesis

Hannes Stärk, Philipp Reiser, Philipp Wolters, Matthias Nießner
Technical University of Munich

Abstract

While 2D generative adversarial networks have shown great performance in generating high-resolution images, they lack an understanding of the 3D world. In contrast, the recently proposed Neural Radiance Fields (NeRF) [4] allow to generate high quality images from novel viewpoints, while incorporating the physical image formation process into the pipeline. We adapt NeRF to render not only novel viewpoints but also unseen human poses. We do so by evaluating a fully-connected deep network on a 3D position, direction to the viewpoint and a human pose representation to produce RGB values. We optimize the dynamic Neural Radiance Field on three synthetic sequences and quantitatively and qualitatively outperform a Pix2Pix [1] baseline.

Keywords: neural rendering, novel view synthesis, SMPL, full body animation, 3D deep learning

1. Introduction

Recent advances in neural rendering methods have led to a high level of realism for synthetically generated images of humans and the generation of novel views. Most of these approaches are only able to change either the camera viewpoint or the human pose. The ability to flexibly control both of those dimensions would be useful in practice. Neural Radiance Fields (NeRF) have shown a great capability to synthesize photorealistic novel views of static scenes by modeling the radiance emitted at each 3D position in every direction. At the same time, there are deformable human body models like SMPL [2] which provide full control over the human pose and shape. We propose to adapt NeRF to use SMPL pose information as additional parameter for modeling the radiance at each point in space. This enables us to render unobserved views and human poses. With this, a user could, for example, generate a smooth video with a desired camera path and human motion while only having snapshots of the dynamic scene with corresponding pose estimation available.

2. Method

The "Skinned Multi-Person Linear Model" represents a human by a set of shape and pose parameters $m \in \mathbb{R}^{69}$. Let $x \in \mathbb{R}^3$ be a 3D location and $d \in \mathbb{S}^2$ be the direction from the location to the camera position (the viewpoint). With positional encoding, we map x and d to higher-dimensional feature representations:

$$\gamma(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \sin(2^1\pi p), \cos(2^1\pi p), ...) \tag{1}$$

Next, we approximate the continuous 5D scene representation (implicit function) with a MLP $f_\phi(\cdot)$ with parameters ϕ that maps the features and the corresponding SMPL parameters to a color c and a volume density $\sigma \in \mathbb{R}^+$:

$$\begin{gathered} f_\phi : \mathbb{R}^{69} \times \mathbb{R}^{L_x} \times \mathbb{R}^{L_d} \to \mathbb{R}^3 \times \mathbb{R}^+ \\ (m, \gamma(x), \gamma(d)) \mapsto (c, \sigma) \end{gathered} \tag{2}$$

To render a 2D image from the radiance field f_ϕ we approximate the intractable volume rendering integral by using numeric integration. Let $r(t) = o + td$ denote a camera ray and let $\{(c_r^i, \sigma_r^i)\}_{i=1}^N$ be the color and volume density values of N random samples along r. We obtain the RGB color value c_r via alpha composition

$$\begin{gathered} c_r = \sum_{i=1}^{N} T_r^i \alpha_r^i c_r^i \quad T_r^i = \prod_{j=1}^{i-1}(1 - \alpha_r^j) \\ \alpha_r^i = 1 - \exp(-\sigma_r^i \delta_r^i) \end{gathered} \tag{3}$$

where T_r^i and α_r^i denote transmittance and alpha value and δ_r^i is the distance between two neighboring sample points. The rendering operator $\pi(\cdot)$ summarizes the mapping from ray samples to RGB values:

$$\pi : (\mathbb{R}^3 \times \mathbb{R}^+)^N \to \mathbb{R}^3 \qquad \{(c_r^i, \sigma_r^i)\} \to c_r \tag{4}$$

Given a set of static 2D images, the parameters ϕ of the neural radiance field f_ϕ are optimized by minimizing the photometric loss between observations and predictions. We use "fine sampling" where additional 3D positions are sampled for each ray based on the density σ outputs of a first "coarse" network. These additional locations are evaluated by a "fine" network that is optimized jointly with the coarse net.

3. Data

We show experimental results for three synthetically rendered dynamic scenes with given ground truth for camera motion and human pose. The sequences of SMPL parameters that describe the human motion of a specific scene are taken from the AMASS [3] motion capture dataset. The textures of the human models are retrieved from the SURREAL [5] dataset which contains maps specifically for the SMPL body model.

We train on three different textures and movement sequences of a person (1) with waving arms (2) arms swinging on the side (3) walking. The images are rendered from 9-12 different viewpoints on a circular trajectory in front of the human. In total there are on average 225 training images for each sequence with a resolution of 256×256. As test data we use interpolations between the observed camera positions and human motion so that the model has to show capability to generalize to those unseen views and body poses.

4. Results

We quantitatively and qualitatively demonstrate that our proposed approach outperforms prior work in all scenarios. Additional ablation studies are used to justify our design choices. We compare our method against a Pix2Pix [1] baseline that is conditioned on a depth map for each image of the scene. This additional depth information is not available to our model and the corresponding information that is learned by our approach would be the 3D points at which the volume density is the highest (which should be the objects surfaces).

The quantitative metrics include the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) and learned perceptual image patch similarity (LPIPS) to compare the renderings with corresponding ground truth images [6].

In Table 1 we report the results on all three sequences, showing that our model excels on all tasks on every metric. The qualitative evaluation (see Figure 2) shows that our method enables more-fine grained control over the human pose parameters as can be seen at the pose of the head. Table 2 shows the quantitative ablation comparisons on one of the synthetic sequences. Row 1 refers to a basic model without positional encoding (PE), view-direction input (VDI) and fine sampling (FS). In row 2,3 and 4 we leave out each component respectively. Row 5 represents the complete model as a reference. Every element poses a benefit to the pipeline, as performance is decreasing, when leaving them out. The best result is achieved by the complete pipeline. Figure 1 shows that the positional encoding enables the network to fit data with high-frequency variations.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Pix2Pix Seq 1	29.734	0.975	0.029
Ours Seq 1	**37.107**	**0.991**	**0.014**
Pix2Pix Seq 2	29.795	0.976	0.025
Ours Seq 2	**36.830**	**0.993**	**0.012**
Pix2Pix Seq 3	28.500	0.972	0.024
Ours Seq 3	**36.158**	**0.994**	**0.013**
Pix2Pix Avg	29.343	0.974	0.026
Ours Avg	**36.698**	**0.9927**	**0.013**

Table 1: Quantitative comparison of metrics with Pix2Pix on three sequences that differ in the amount of animation of the human.

	PSNR ↑	SSIM ↑	LPIPS ↓
1) No PE, VDI, FS	30.731	0.978	0.036
2) No Positional Encoding	32.351	0.984	0.035
3) No View Direction Input	34.450	0.990	0.020
4) No Fine Sampling	34.539	0.991	0.019
5) Complete Model	**36.158**	**0.994**	**0.013**

Table 2: Ablation of our model evaluated on sequence 3 with the highest amount of animation.

Figure 1: Ablation comparison of single frames with and without parts of the pipeline. Top left: Ground Truth. Top right: Full pipeline. Bottom left: No directional input. Bottom right: No Positional Encoding.

Figure 2: The first row shows the ground truth images (subset of a walking movement). Below you can see the rendered sequence generated by the proposed approach. Row three compares directly the rendering frames to the Pix2Pix baseline.

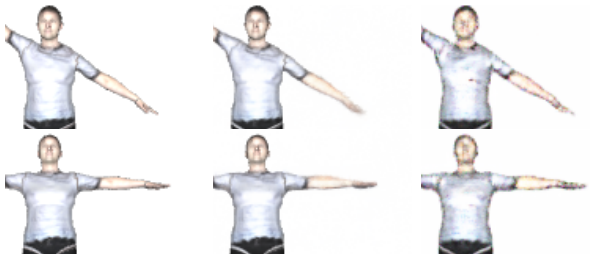

Figure 3: Rows correspond to the same viewpoint. First column: ground truth. Second column: renderings of proposed method. Third column: viewpoint renderings of Pix2Pix.

5. Conclusion

We presented a method to combine neural rendering of deformable human 3D models with novel view synthesis. This combination gives the user a degree of control that is much closer to classical rendering pipelines than in most neural approaches. We can interpolate between sparsely captured motion and large gaps in the seen viewpoints. While the degree to which the method generalizes is impressive, it has to be considered, that the method assumes ground truth SMPL parameters. For real-world data, only estimates of the human body pose would be available. It would be interesting to adapt the approach to account for this offset.

References

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5967–5976. IEEE Computer Society, 2017. 1, 2

[2] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 1

[3] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019. 2

[4] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *CoRR*, abs/2003.08934, 2020. 1

[5] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2

[6] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924, 2018. 2