

ANR: Articulated Neural Rendering for Virtual Avatars

– Supplementary Document –

Amit Raj¹ Julian Tanke² James Hays¹
 Minh Vo³ Carsten Stoll⁴ Christoph Lassner³

¹Georgia Tech ²University of Bonn ³Facebook Reality Labs ⁴Epic Games

1. Overview

In this supplementary, we provide details about the implementation of our approach (Section 2), three experiments to further verify the design choices and model demonstration (Section 3), and lastly the comparison with other methods (Section 4). Importantly, we urge the reviewers to watch our supplementary video for the best demonstration of our method.

2. Implementation Details

Model Architecture: Our rendering networks \mathcal{R}_1 and \mathcal{R}_2 are variants of the UNet architecture as shown in Figure 1 and Figure 2. The discriminator for the adversarial loss uses a multi resolution PatchGAN over 3 spatial scales as in [2]. The networks are trained over 1500 frames per identity with a batch size of 4 at a resolution of 1024×1024 .

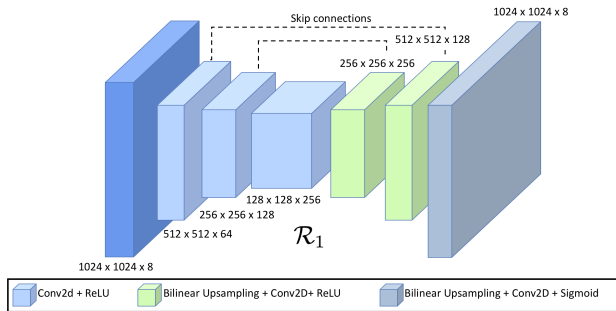


Figure 1. Architecture of the stage 1 network. It takes the rasterized 8-channel neural texture and produces an intermediate ‘static’ base RGB image and a 5-channels latent feature of the same resolution as the input.

Model Efficiency Metric (rIPFIP): As described in the manuscript we introduce a metric to measure the relative improvement of an approach that leverages 3D information over V2V (which is a purely 2D approach and has the largest number of model parameters of the models in our paper) scaled by the relative improvement in the number

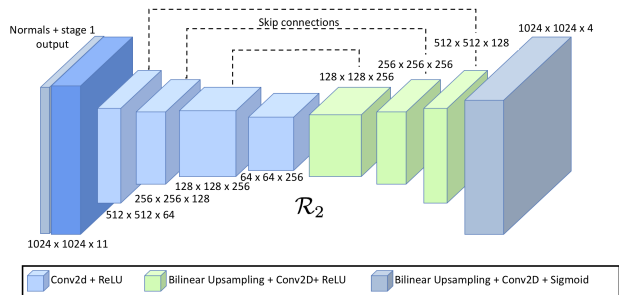


Figure 2. Architecture of the stage 2 network. It takes the intermediate features from state 1 and the rasterized body normals and produces the RGB image and a mask for the clothed human body at the same resolution as the input.

of parameters in the model. More specifically, this metric consists of 2 terms. The first term measures the relative improvement of the perceptual quality of a method (in LPIPS) over V2V and the second term measures the compactness of the model without taking into account the quality of the synthesized image. These two terms are expressed as

$$d_{\text{LPIPS}}(x) = \frac{(\text{LPIPS}_{V2V} - \text{LPIPS}_x)}{\text{LPIPS}_{V2V}}, \quad (1)$$

$$d_{\#p}(x) = \frac{\log(\#p_{V2V}) - \log(\#p_x)}{\log(\#p_{V2V})}, \quad (2)$$

where $\#p$ represents the number of parameters in a model. Both these terms lie in $(-\infty, 1]$ (\uparrow is better). Our model efficiency metric is defined as

$$\text{rIPFIP}(x) = d_{\text{LPIPS}}(x) * d_{\#p}(x), \quad (3)$$

which takes into account both the compactness of the model and the synthesized quality of the image of a particular method x .

3. Additional Verification Experiments

Normal Injection Ablation: We study the different ways in which normal information can be fused into the pipeline

to account for self-occlusions and pose dependent dynamic effects in Table 1. Our current design of injecting the normal in \mathcal{R}_2 produces the best avatars quantitatively. Furthermore, we notice that injecting normal information to \mathcal{R}_1 reduces the performance. This is because the \mathcal{R}_1 learns a ‘static’ base texture irrespective of the pose and having the the pose-conditioning hurts learning of this static base texture.

Table 1. Normal Injection Ablation. ‘early’ refers to concatenating the normals at the input of \mathcal{R}_1 , ‘late’ refers concatenating the normals at the input of \mathcal{R}_2 , and ‘both’ refers to adding the normals at input of both stages. ‘+so’ refers to models trained with the split optimization strategy.

	SSIM \uparrow	LPIPS \downarrow	FLIP \downarrow
early	0.962	0.0703	0.0363
late	0.968	0.0584	0.0321
both	0.966	0.0636	0.0332
early+so	0.965	0.0636	0.0342
late + so (ours)	0.973	0.0508	0.0289
both + so	0.966	0.0630	0.0338

Split optimization: Figure 4 shows the advantage of split optimization in terms of convergence of the avatar, as seen in the reconstruction error (in Table 2 in the main text). We also note that the split optimization improves performance regardless of the choice of fusion paradigm (see Table 1 in this supplementary file). This demonstrates the benefit of mesh geometric misalignment modeling, which leads to faster convergence and better results quantitatively rather than relying on purely data-driven optimization with well-tuned losses (GAN, feature loss, decaying ℓ_1 loss) and aggressive data augmentation.

Identity mixing: Figure 3 shows our avatars with clothing components replaced from other learned identities. Notice the ID specific and clothing specific deformations retained even after mixing identities.

4. Additional Comparison

Liquid Warping GAN (LWG): LWG [1] is a recent strong method that combines the benefits of explicit 3D modeling with image2image translation. Similar to us, LWG can synthesis the person appearance in novel view and poses. This method, however, does not have a persistent 3D texture that enables stable virtual rendering from arbitrary poses and viewpoints. We show a comparison with this method in Figure 6. In all aspects, realism of the reconstruction, face details and mask accuracy, our method notably outperforms LWG.

User study: As mentioned in the main text, we conducted 2 user studies with 80 participants and provide some more details about the studies here. In the first study, We gener-



Figure 3. Our avatars with identity and appearance mixing. The identities and pose are same along each row and the clothing is constant along each column. We observe that once we have learnt multiple identities, we can mix and match textures between them.

ated avatars using Vid2Vid (V2V), Textured Neural Avatar (TNA), and Deferred Neural renderer trained with added VGG loss (DNR), and our proposed ANR model. Each user was presented with 20 stimuli for 5s each, and asked to select the image which had the most realistic avatar. Avatars generated from all methods were posed in the same manner and composited into the same background. An example stimulus for this study is shown in Figure 7. ANR was preferred 81.6% of the time.

In the second user study, we conducted a 2-alternative forced choice study where users were presented a real image and an avatar generated from our method in different poses and asked to select the more realistic image. The images disappear after 10s after which the user has unlimited time to choose the best avatar. An avatar can be considered completely photo-realistic if we are able to fool the users 50% of the time (random choice). Our avatars were picked 30% of the time, showing strong realism in many cases.

References

[1] Wen Liu, Wenhan Luo Lin Ma Zhixin Piao, Min Jie, , and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 4

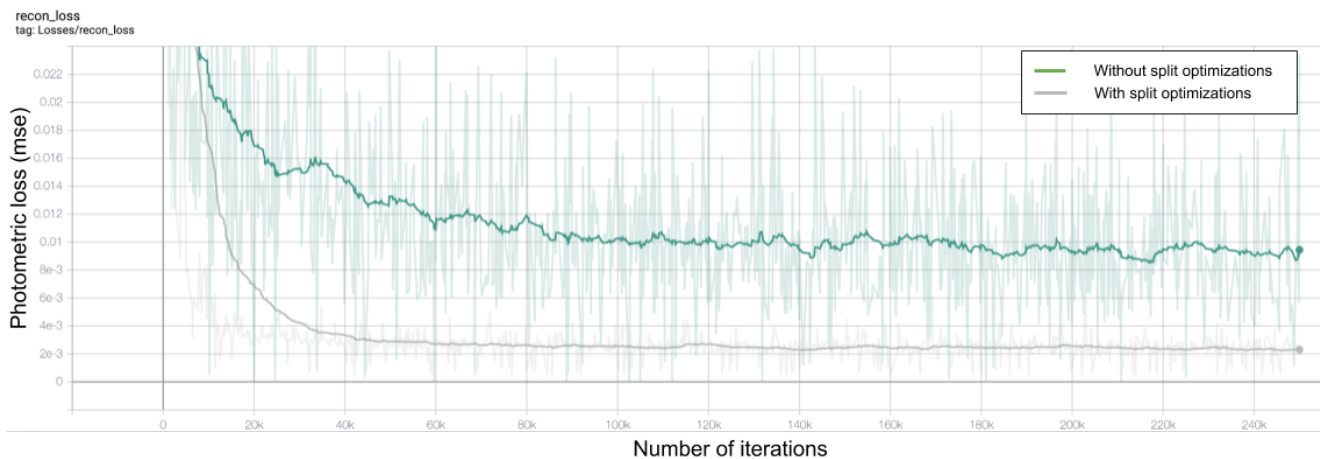


Figure 4. Reconstruction loss on validation poses of a learnt avatar. We observe that split optimization converges faster and achieves significantly lower reconstruction loss for the same training architecture.



Figure 5. Avatars learned using ANR in various novel poses and viewpoints



Figure 6. Comparison with Liquid warping GAN [1] on viewpoint synthesis. **Top two rows:** Liquid warping GAN; **bottom two rows:** our method. Our model is robust to a variety of poses and is able to preserve texture details, particularly visible on the white T-shirt. Figure is best viewed electronically at full magnification.

Please pick the best one

Question 1 / 22



Figure 7. Example stimulus in first user study. We generated avatars Vid2Vid (V2V), Textured Neural Avatar (TNA), and Deferred Neural renderer trained with added VGG loss (DNR), and our proposed ANR model. Each user was presented with 20 stimuli for only 5s each and asked to select the image which had the most realistic avatar. Avatars generated from all methods were posed in the same manner and composited into the same background.

malization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1