ANR: Articulated Neural Rendering for Virtual Avatars – Supplementary Document –

Amit Raj [1] Julian Tanke[2] James Hays[1]
Minh Vo[3] Carsten Stoll[4] Christoph Lassner[3]

[1]Georgia Tech [2]University of Bonn [3]Facebook Reality Labs [4]Epic 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 $\times$ 1024.

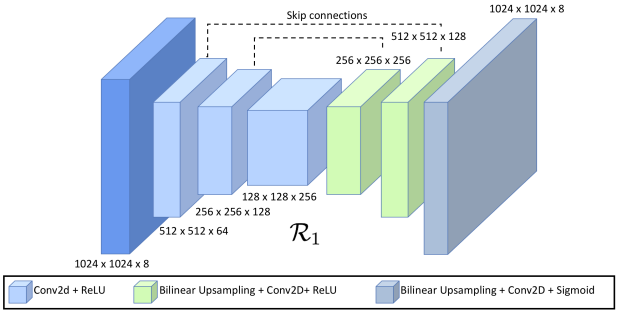

Figure 1. Architecture of the stage 1 network. It takes the rasterized 8-channel neural texture and produces an intermediate 'static' base RGB image and a 5-channels latent feature of the same resolution as the input.

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 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

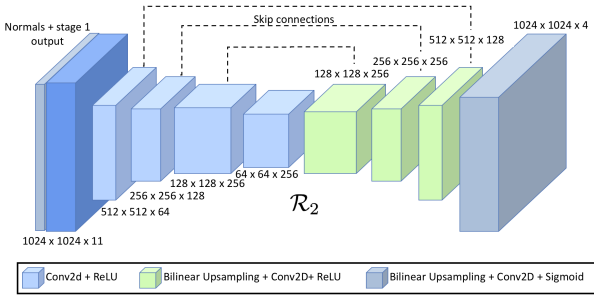

Figure 2. Architecture of the stage 2 network. It takes the intermediate features from state 1 and the rasterized body normals and produces the RGB image and a mask for the clothed human body at the same resolution as the input.

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

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

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

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

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 ↑	LPIPS ↓	FLIP↓
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 5 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 and Figure 9 shows our avatars with clothing components replaced from other learned identities. Notice the ID specific and clothing specific deformations retained even after mixing identities.

Generalization: Fig. 10 shows an example of an avatar for which only neural texture has been optimized on a new subject while keeping the pre-trained neural renderer fixed. We observe details of the T-shirt are also recovered correctly.This example indicates the strong generalization of our neural renderer despite being trained only on a few identities

Viewpoint variation: The examples in the paper were rendered from viewpoints to show the full body and representative for our capture. There is no restriction on the elevation or viewing distance (c.t. Fig. 4) and our model generalizes remarkably well to new elevation angles, given that no such data was present in the training set.

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.

Figure 4. Elevation changes, even though no such viewpoints are present in the training data, are handled robustly. Qualitatively, the model works best for ± 15 and produces acceptable results in ± 45. **This animated figure plays in Acrobat Reader**.

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

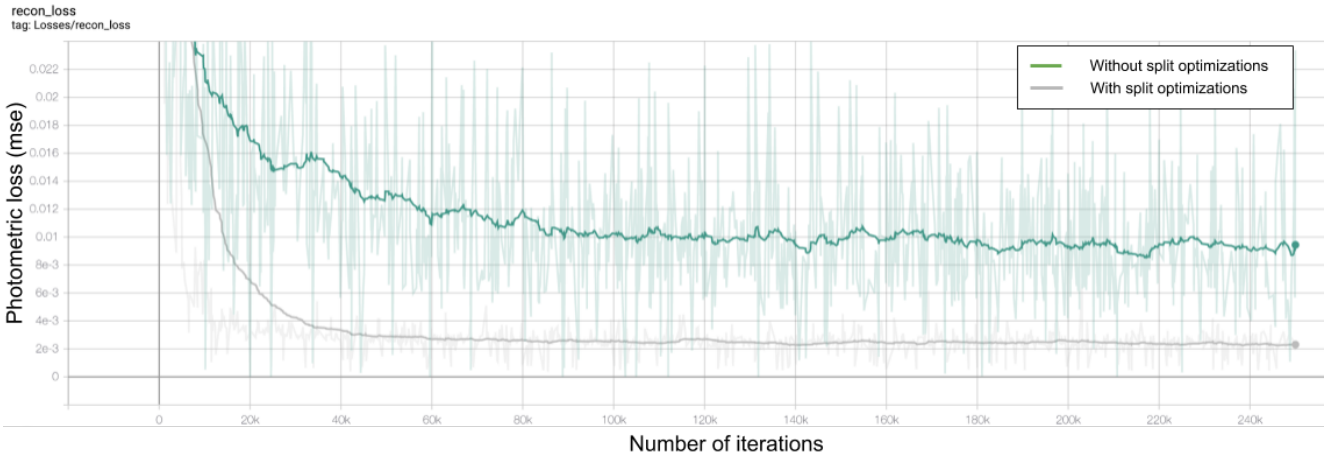

Figure 5. Reconstruction loss on validation poses of a learnt avatar. We observe that split optimization converges faster and achieves significantly lower reconstruction loss for the same training architecture.

Figure 6. Avatars learned using ANR in various novel poses and viewpoints

Figure 7. 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.

viewpoints. We show a comparison with this method in Figure 7. 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 generated 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 8. 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

[2] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

Figure 8. 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.

Figure 9. Virtual Try-On example. ANR enables texture mixing by swapping the regions of the neural texture. This example validates the disentanglement of appearance and neural shading network when ANR is trained on multiple identities. **This Animated figure plays in Adobe Reader**

Figure 10. A novel avatar *unseen* during neural renderer training. Only the neural texture is optimized for this identity. **This animated figure plays in Adobe Reader**