RAM-Avatar: Real-time Photo-Realistic Avatar from Monocular Videos with Full-body Control

Supplementary Material

In this supplementary document, we will introduce the additional implementation details, network structure, and training strategy.

1. Additional Implementation Details

The whole network is implemented using PyTorch and the value of hyper-parameters of the network are shown in Table 1. Adam is used as the optimizer and weight decay is 10^{-4}. The input images are cropped and resized into 1024 × 1024. We train 150,000 steps in total with an initial learning rate of 1e-4 and decrease it to 0 using cosine decay. The training stage takes about 10 15 hours on a single RTX3090 GPU. The real-time live system contains two parts, which are responsible for full-body tracking based on [64,76] and high-fidelity rendering respectively. Our system is able to run at 25 fps on a PC with two RTX 3090 GPUs.

Table 1. Hyperparameters for network training and evaluation.

Parameter Name	Value
H (the height of all feature maps)	1024
W (the weight of all feature maps)	1024
C_1 (the channel of the face feature map)	16
C_2 (the channel of the hand feature map)	16
C_3 (the channel of the body feature map)	48
C_4 (the channel of the attention feature map)	24

2. Network Structure

To achieve full-body control, we first generate 1024×1024×48 body feature map through standard rasterization based on the learned neural texture. As introduced in Section 3.2, we propose a dual attention module to augment the temporal consistency and enrich details. The distance function d(i,u) in Equation 1 is as follows:

$$d(i,u) = \frac{1}{M(i,u)^p + 1} \qquad (1)$$

where M(i,u) represents the Manhattan Distance between the *i*-th element and position *u*, p equal to 0.8. As described in Section 3.1, the network we adopted to extract distinctive features of the face region is a shallow U-net, which consists of 5 layers in total. The network responsible for extracting hand gesture features has a similar structure since both face and hands exhibit limited shape variations yet contain numerous subtle details. The architecture of such U-net is shown in Figure 1 and Figure 2. After the preparation for conditional feature maps is finished, we adopt a lightweight StyleUnet for high-fidelity rendering, which is illustrated in Figure 4. The motion distribution module contains one encoder and two decoders. The pose encoder consists of three fully connected layers sandwiched between two batch normalization (BN) layers and generates the latent representation of body poses. The pose decoders consist of two fully connected layers aim at translating the latent representation to the original pose, where one decoder takes the current pose as input and another decoder takes the previous four frames as input.

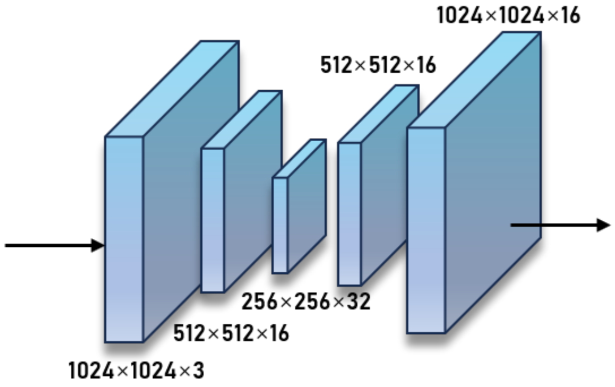

Figure 1. Detail structure of face feature extraction networks.

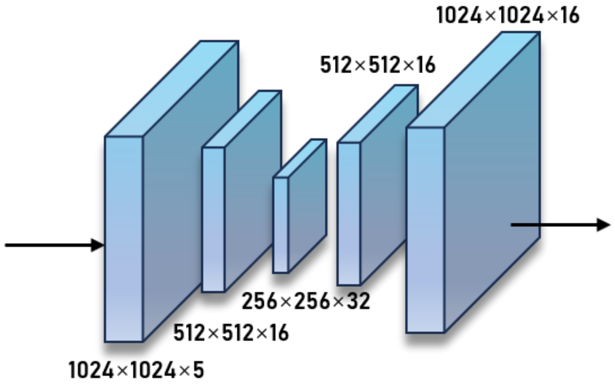

Figure 2. Detail structure of hand feature extraction networks.

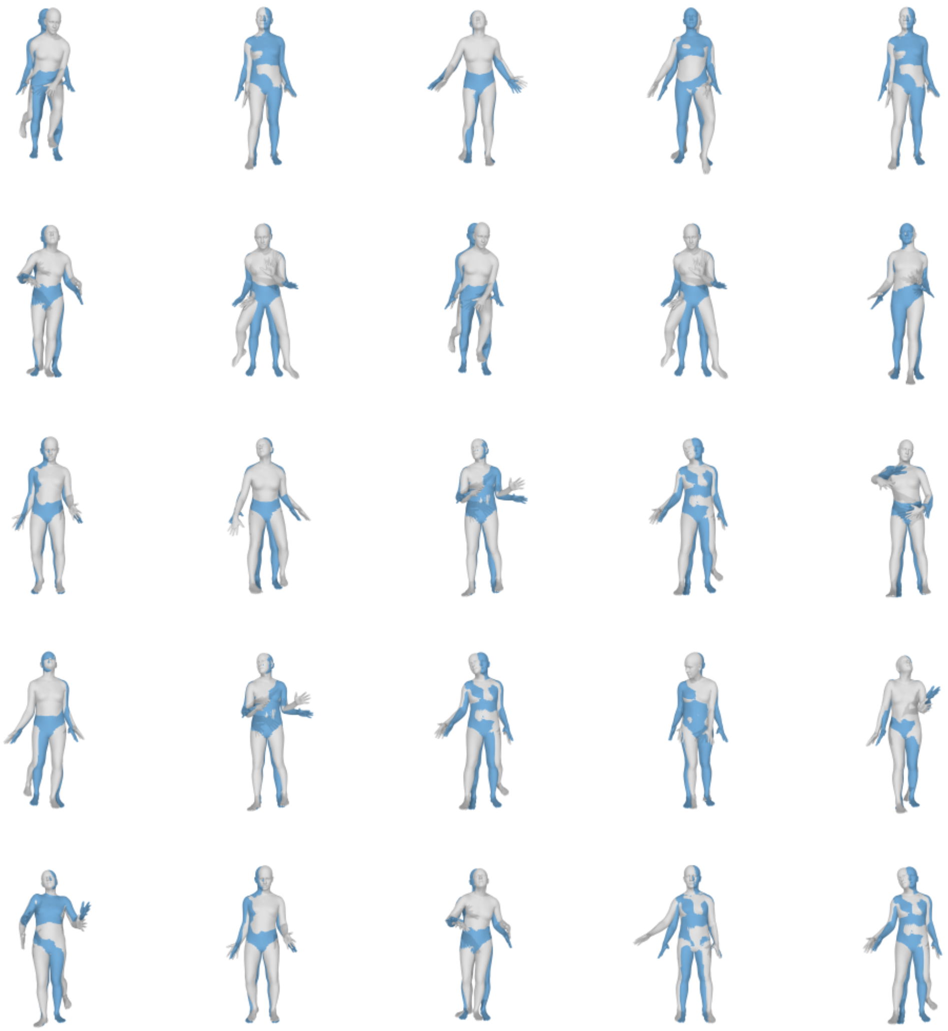

Figure 3. Visualization results of motion distribution align module on randomly generated motions. Given a training video mainly contains stand motions, *e.g.* speech video, our motion distribution align module maps the out-of-distribution poses into in-distribution poses. The white meshes represent the original motion, while the blue meshes represent the transferred motion.

3. Training Strategy

During training, we first train the whole network without dual attention module for 10,000 steps. The neural texture plane is frozen in the last 20,000 steps for better rendering. As illustrated in Section 3.3, we first train VAE on a collection of large-scale human motion datasets to derive a powerful latent representation and a robust encoder for 50 epochs in total. After that, we freeze the encoder and train another conditional decoder on the target avatar pose dataset for 200 epochs in total. We show some additional motion transfer results in Figure 3.

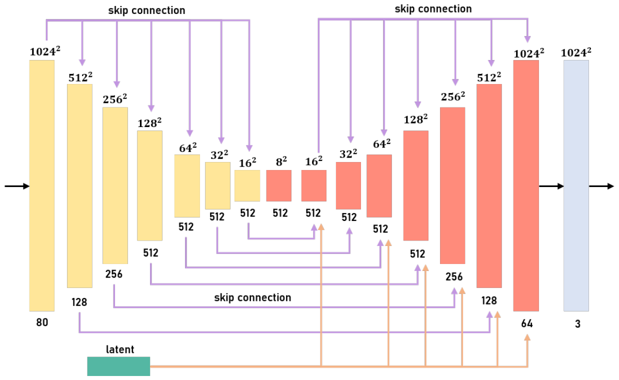

Figure 4. Detail structure of rendering network.

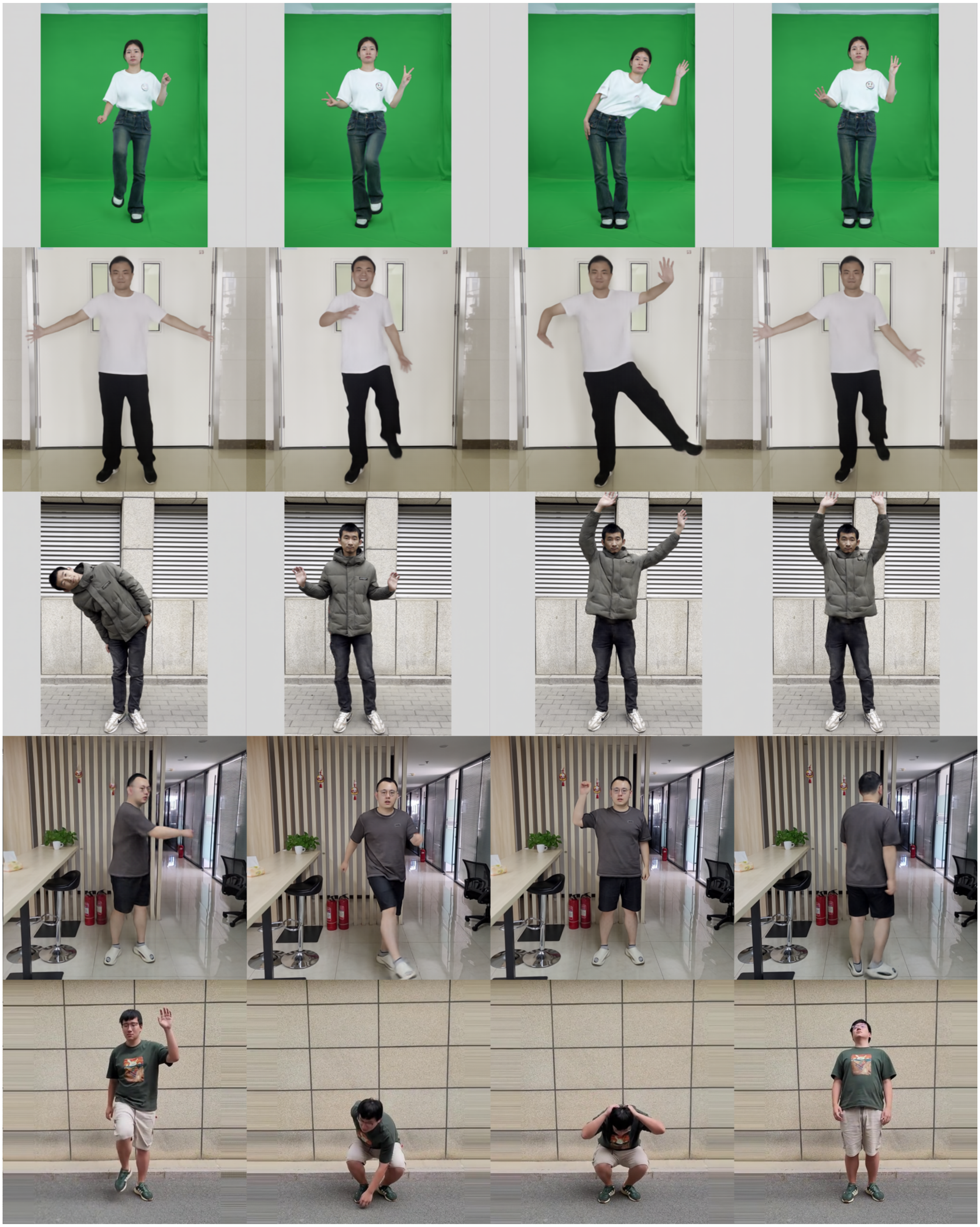

Figure 5. Avatars learned based on our method in various body poses, hand gestures, and facial expressions.