One-shot Human Motion Transfer via Occlusion-Robust Flow Prediction and Neural Texturing - Supplementary Document

Yuzhu Ji, Chuanxia Zheng, and Tat-Jen Cham

A. Details of Model Structure

A. Motion Network

We developed a FlowNet-like [1] structure with an encoder to produce the translation signals for the 2.5D branch, and a decoder with iterative refinement modules for producing multi-scale dense motion flow and occlusion map. The specific network structure is illustrated in Fig. A.1. In the encoder, We adopted a global average pooling (GAP) layer to get the latent representation of motion residual. Then, two separate fully connected layers (FC_1 and FC_2) are adopted to get translation signals for neural texture atlas and DensePose. The decoder part is with an iteratively upsampled refinement module to produce multi-scale dense motion flow, which is similar to FlowNet [1]. The blue lines denote the prediction heads of dense motion flow and occlusion maps.

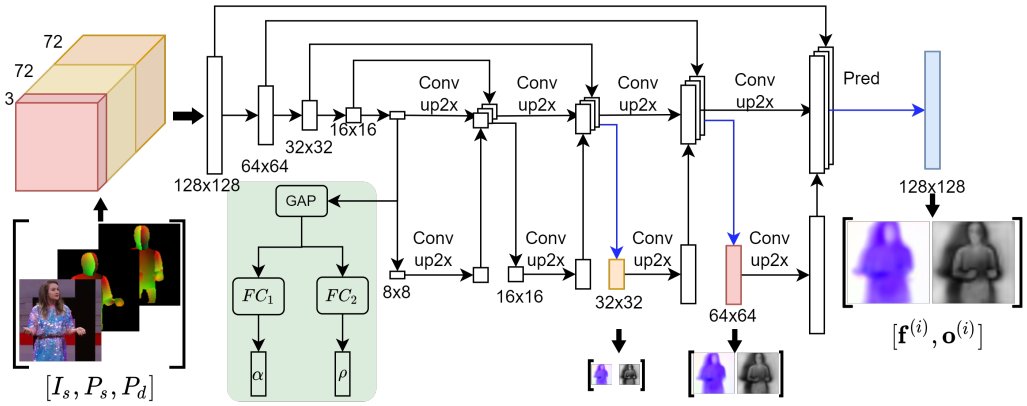

Fig. A.1. Details of motion network structure.

B. Feature Warping network

As mentioned in the paper, we present a multi-scale feature warping network, which is inspired by [2], to capture a more accurate appearance feature. Fig. A.2 illustrates the specific structure of the network. It shows that the warped source feature will be partially propagated to the decoder part to preserve the source appearance. To stabilize the estimation of motion flows and occlusion maps w.r.t different scales, we further applied flow warping on the source and generated images by using multi-scale dense motion flows and occlusion maps.

C. Neural Texture Mapping Network

The neural texture mapping (NTM) module aims to implicitly disentangle the appearance and geometry in 2.5D UV space. To achieve this, we developed a two-stream pipeline in our implementation. Fig. A.3 shows the detailed network structure of the NTM module. More specifically, the NTM module consists of two main branches, i.e., the neural texture atlas translation and the DensePose translation branch. Then the translated neural texture atlas and DensePose will be utilized to produce the neural texture feature by a differentiable texture mapping layer. Then the neural texture feature is rendered into an image.

Yuzhu Ji is with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, P. R. China; Chuanxia Zheng is with the Department of Engineering Science, University of Oxford, UK; Tat-Jen Cham is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (E-mail: andrewchiyz@gmail.com; cxzheng@robots.ox.ac.uk; astjcham@ntu.edu.sg).

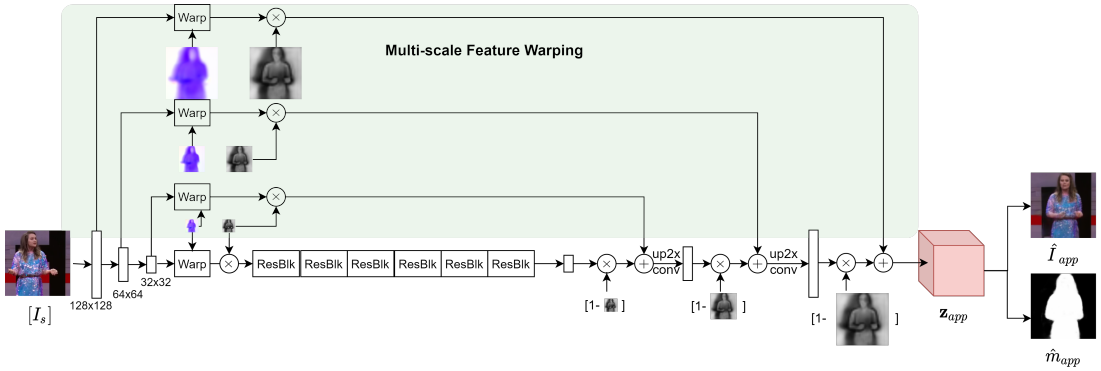

Fig. A.2. Details of feature warping network.

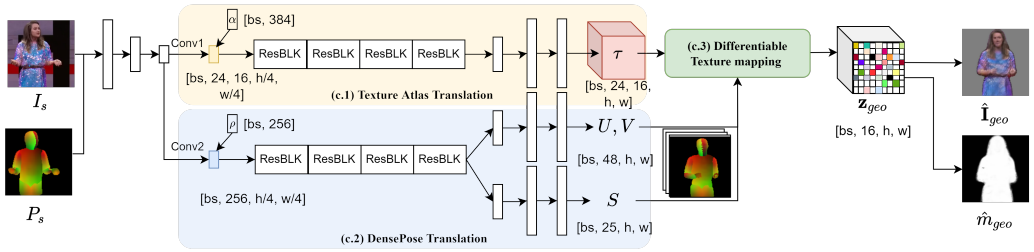

Fig. A.3. Details of neural texture mapping network.

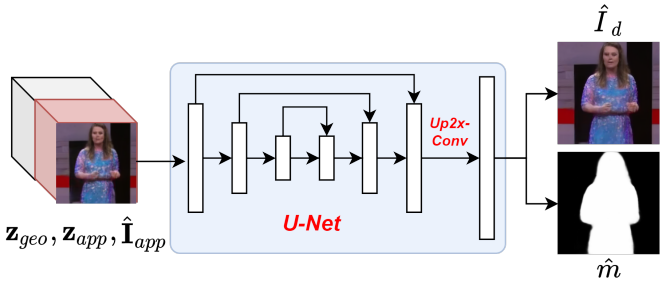

Fig. A.4. Details of BlenderNet.

D. Blending Module

In our model, we use a Blending Module to integrate the 2D flow-warping feature and 2.5D neural texture feature for better-recovering appearance and geometry. To achieve this, we adopted a shallow U-Net structure with three scales of downsampling blocks in our implementation. To generate images with higher resolution, we applied an upsampling and convolutional layers as a super-resolution block (see Fig. A.4).

REFERENCES

[1] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*. IEEE Computer Society, 2017, pp. 1647–1655.
[2] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion representations for articulated animation," in *CVPR*, 2021, pp. 13 653–13 662.