

Flow Guided Transformable Bottleneck Networks for Motion Retargeting — Supplemental Material

Jian Ren Menglei Chai Oliver J. Woodford* Kyle Olszewski Sergey Tulyakov
Snap Inc.

{jren, mchai, kolszewski, stulyakov}@snap.com

1. Architecture Details

Here we provide a detailed description of our network architecture, as shown in Figure 1. Details for the 3D Resampling in Figure 1 are illustrated Figure 2. The 2D encoding network Enc_{2D} (introduced in Section 3.2 of the main paper) includes three convolution layers (conv) and four Residual Blocks (ResBlocks) [2]. We provide the kernel size, padding size, stride, and the number of output channels in Figure 1. The 3D encoding network Enc_{3D} consists of three 3D conv layers (the first three layers shown in Figure 2). Similarly, the 3D decoding network Dec_{3D} includes three 3D transposed conv layers (the last three layers in Figure 2), and the 2D decoding network Dec_{2D} has four ResBlocks, two transposed 2D conv layers and one conv layer.

The skip connections between Enc_{3D} and Dec_{3D} are implemented via 3D flow, as shown in Figure 2. The flow network F also includes 2D encoding and 3D encoding, which have a similar architecture to Enc_{2D} and Enc_{3D} , except that the strides in Enc_{3D} are set to 1. We let the ResBlocks in the flow network share the same weights as Enc_{2D} to better learn the implicit 3D representation.

2. More Qualitative Results

We portray more results in Figure 3, in which the driving videos are from the iPER [1] dataset and use different subjects than the source reference images. Here we visualize the animation results of each source subject’s body being controlled by the pose from the corresponding driving video. The full video is provided in the supplementary material, in the file named iPER_driving.mp4.

We show more human body animation results in Figure 4, in which the driving videos are from the Youtube Dancing dataset. The full video is also provided, in the file named Youtube_Dancing_driving.mp4.

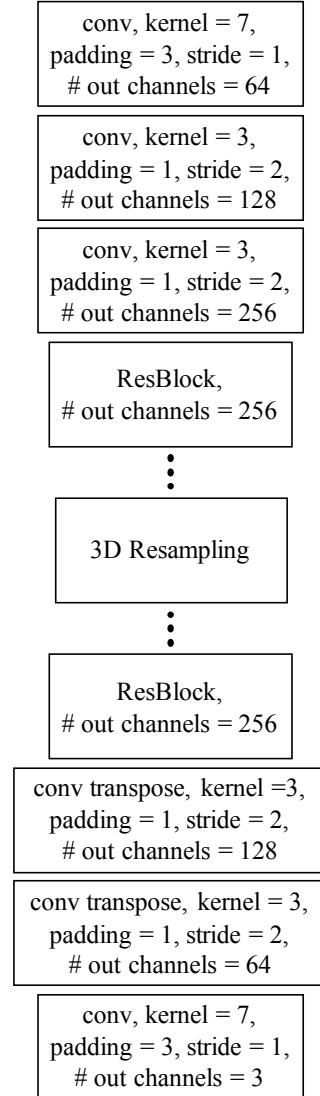
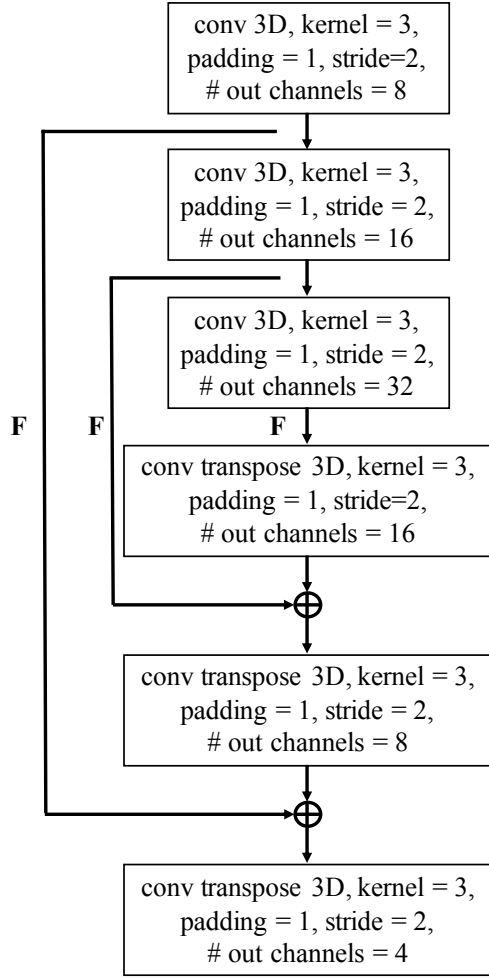


Figure 1. The overall architecture. We provide the kernel size, padding, stride, and the number of output channels for the convolution layers (conv), and the number of output channels for the ResBlocks.

*Work done while at Snap Inc.



F: Flow Guided Sampling

⊕: Concatenation

Figure 2. The architecture for the 3D flow-guided skip connections.

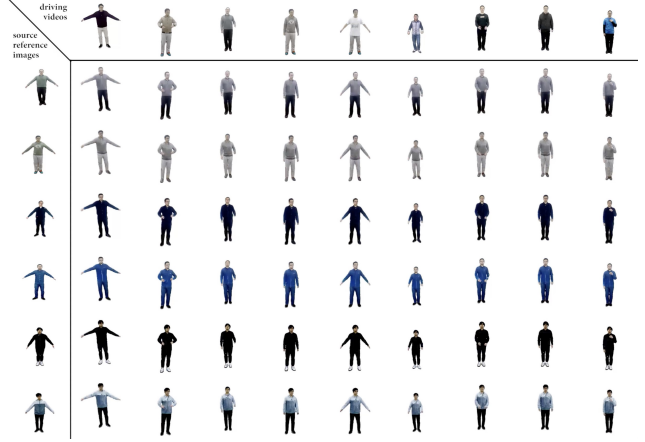


Figure 3. Qualitative results. **Left Column:** one source reference image from each subject. **First Row:** the driving videos are from the iPER dataset. Please see the corresponding supplementary video for a better illustration.

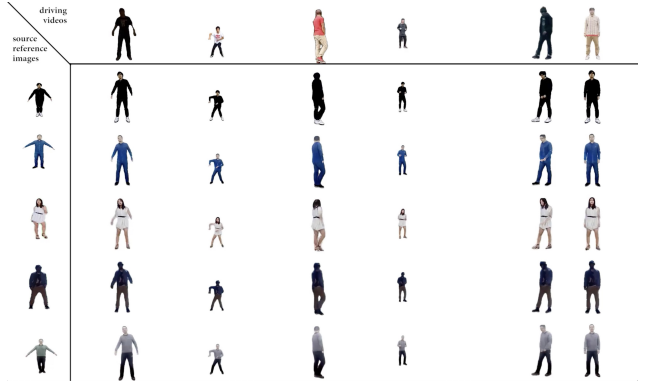


Figure 4. Qualitative results. **Left Column:** one source reference image from each subject. **First Row:** the driving videos from the Youtube Dancing dataset. Please see the corresponding supplementary video for a better illustration.

References

- [1] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Int. Conf. Comput. Vis.*, pages 5904–5913, 2019.
- [2] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.