Learning Facial Representations from the Cycle-consistency of Face

Supplementary Materials

1. Network Architectures

There are several networks used in our proposed model, including E_{exp}, E_{id}, D_{flow}, D_{exp}, D_{id}, and two MLPs (i.e. MLP_{de} and MLP_{re}). Figure 2 presents the network architecture of E_{id} and E_{exp}; Figure 3 presents the network architecture of D_{flow}; Figure 4 presents the network architecture of D_{exp}. Figure 5 presents the network architecture of D_{id} and two MLPs. More detailed descriptions are provided in the caption of each figure. Notice that we adopt leakyReLU with leakage 0.1 as our activation function used in all the networks, but we omit it in the figures for simplicity. Moreover, there is a channel attention module which is heavily used in the proposed networks, its visualization is provide in the Figure 1. This channel attention module is inspired from self-attention [3], but we basically use it to compute relations between channels instead of spatial pixels as in [3].

2. More Intermediate Network Outputs

We show more intermediate network outputs in Figure 6. Figure 6 presents intermediate network outputs from test dataset of Voxceleb2. The proposed method is able to generate consistent mean faces as well as the neutral faces with some important facial-attributes well preserved. It suggests the effectiveness of the proposed method.

3. More Image-to-Image Translation Results

We show more image-to-image translation results with different sources of motion sequences in Figure 7. For translation, we adopt different target images from test dataset of Voxceleb2. The results demonstrate that our proposed method can transfer the head pose and expression from the source to the target without noticeable artifacts.

4. Ablation Study

In this section, we perform ablation experiments for responding the advises from the anonymous reviewers. As shown in Table 1, the best result is achieved by using Vox1+2, 64×64 image size and adopt other frames in the video as transformation (T).

Dataset	Img Size	Transform (T)	RAF Acc(%)
Vox 1	64×64	other frames	66.63
Vox 1	64×64	horizontal flip	63.52
Vox 1	128×128	other frames	64.05
Vox 1+2	64×64	other frames	71.01

Table 1. The ablation study. We show the dataset size, image size, and transformation can affect the power of extracted representations.

5. Identity preservation in generated images

Regarding the concerns on the identity preservation in our frontalization and translation tasks of facial images from the anonymous reviewers, we perform the person verification on our frontalized/translated results with respect to their corresponding original images (1000 pairs in total) and achieve the accuracy of 69.2%, which is close to our performance (73.72%) of person recognition shown in Table 3, demonstrating that the identity is well-preserved during frontalization/translation by our proposed method.

References

[1] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Int. Conf. Comput. Vis.*, 2017. 3

[2] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 2

[3] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018. 1

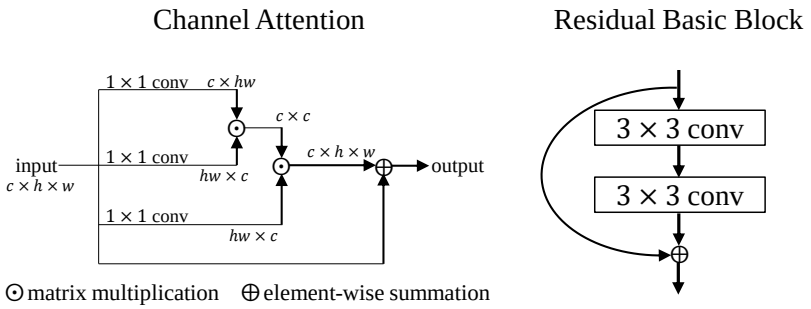

Figure 1. The channel attention module and residual basic block.

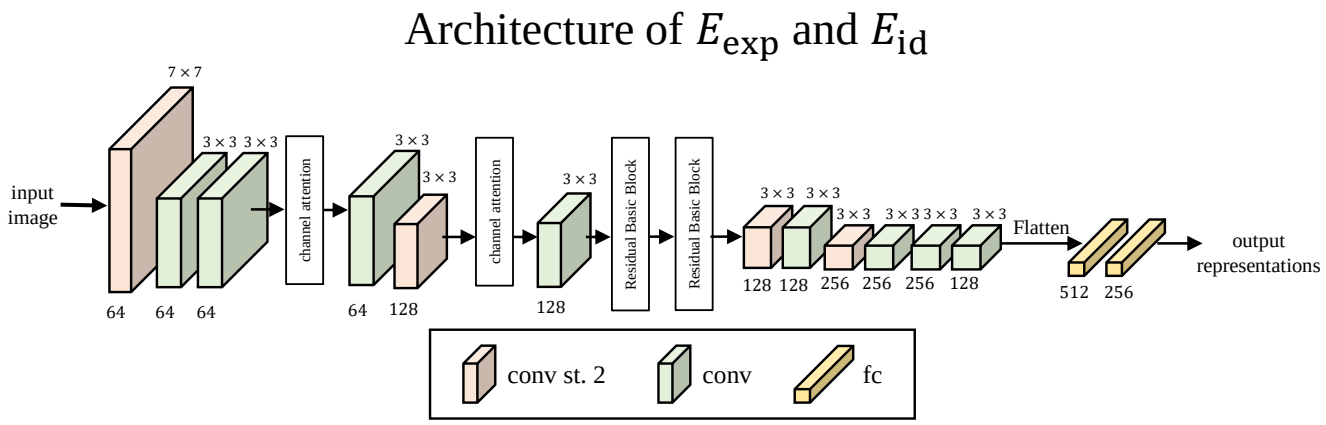

Figure 2. The network architecture of E_{id} and E_{exp}. We also present the size of convolution kernels and number of features in the figure. Each convolution/fully-connected layer is followed by a leakyReLU with leakage 0.1, except the last one. "conv st.2" denotes the convolution layer with stride 2.

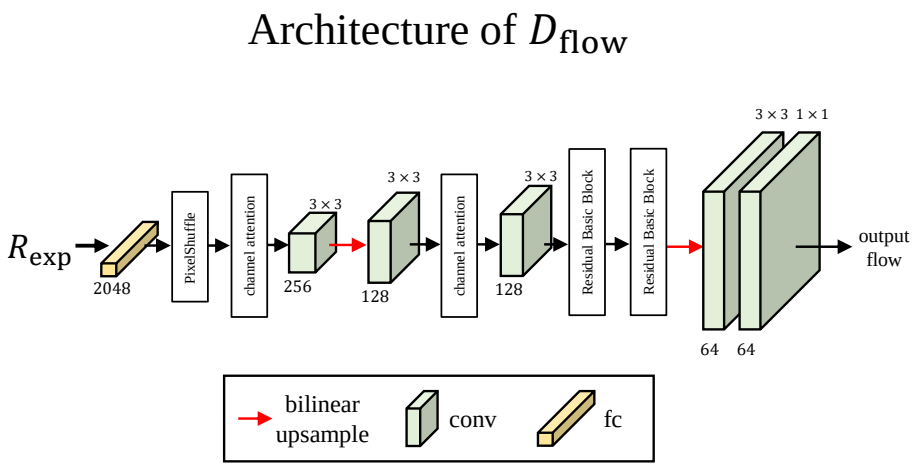

Figure 3. The network architecture of D_{flow}. We also present the size of convolution kernels and number of features in the figure. Each convolution/fully-connected layer is followed by a leakyReLU with leakage 0.1, except the last one which is followed by a Tanh. We use PixelShulffle [2] for increasing the spatial resolution of the features after the fully-connected layer.

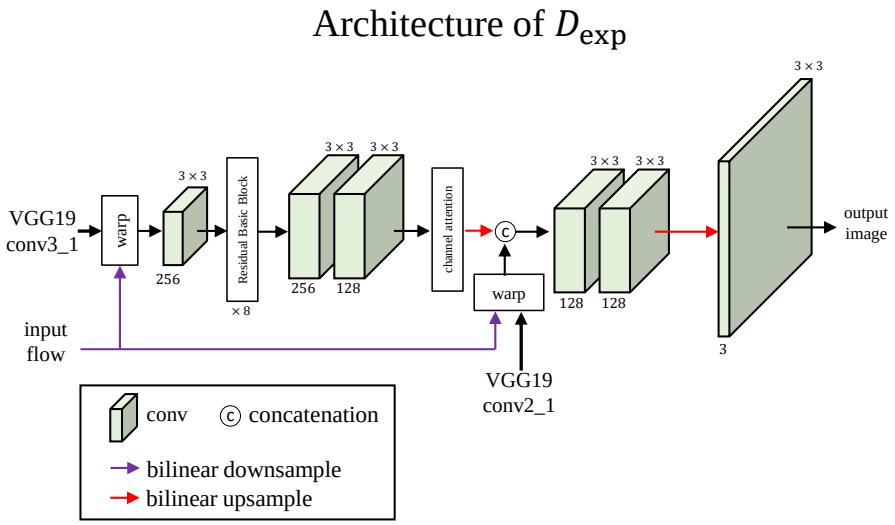

Figure 4. The network architecture of D_{exp}. We also present the size of convolution kernels and number of features in the figure. Each convolution layer is followed a leakyReLU with leakage 0.1 except the last one.

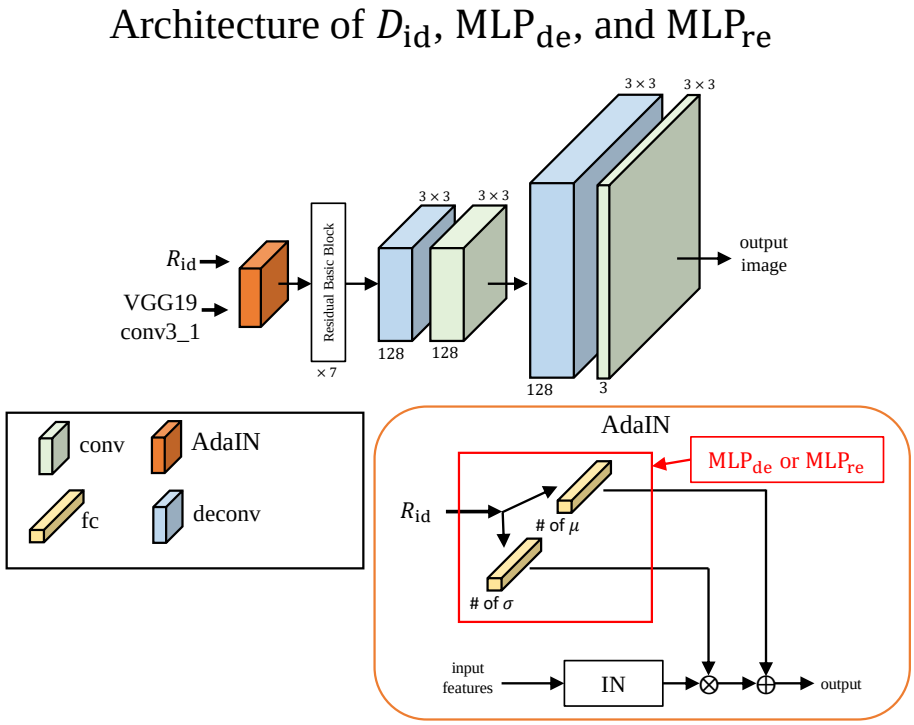

Figure 5. The network architecture of D_{id}, and the MLP. We also present the size of convolution kernels and number of features in the figure. Each convolution/fully-connected layer is followed by a leakyReLU with leakage 0.1 except the last one. The MLPs learn AdaIN [1] affine parameters (μ and σ). IN: instance normalization.

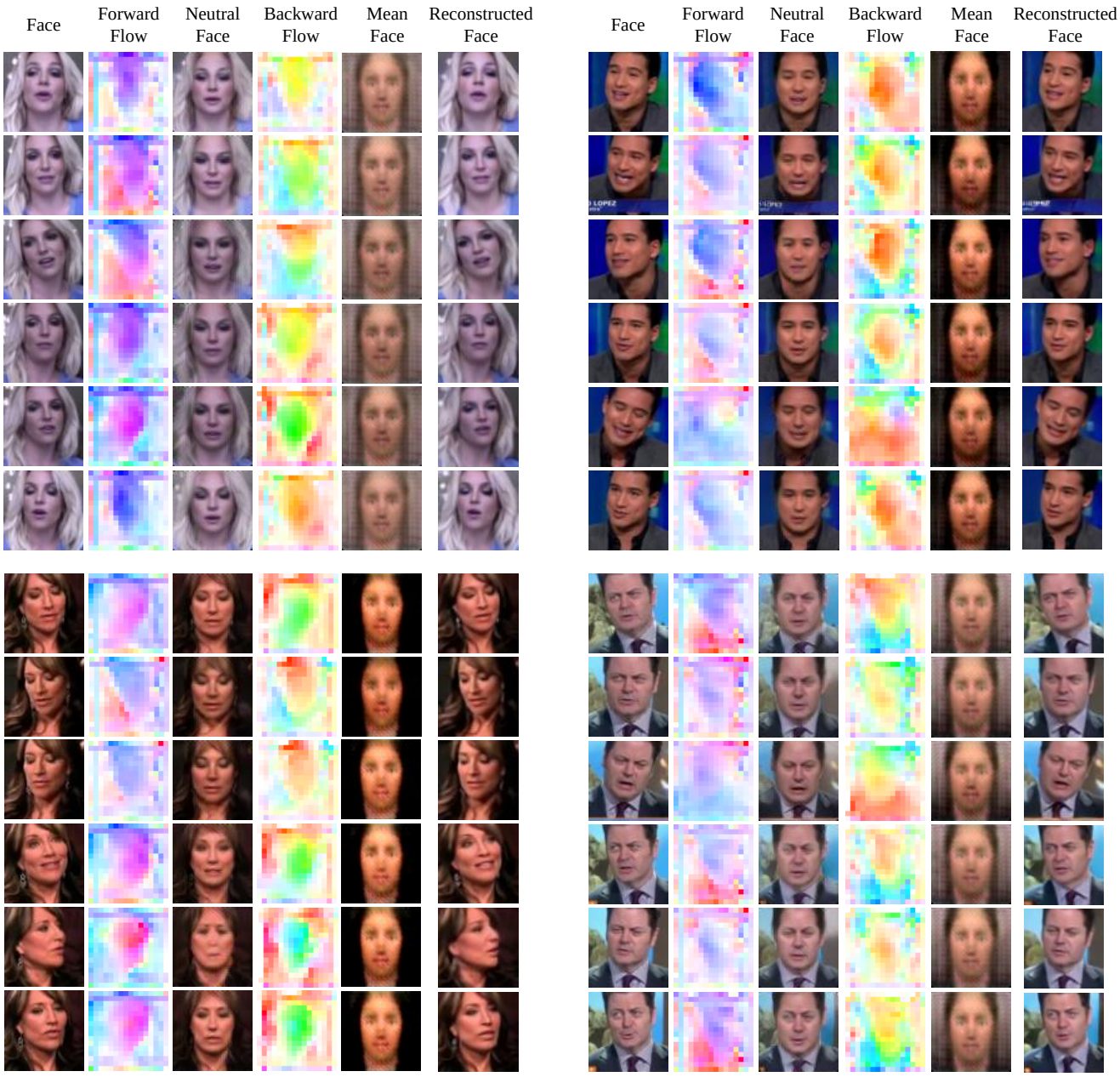

Figure 6. We show intermediate network outputs of different persons from test datatset of Voxceleb2.

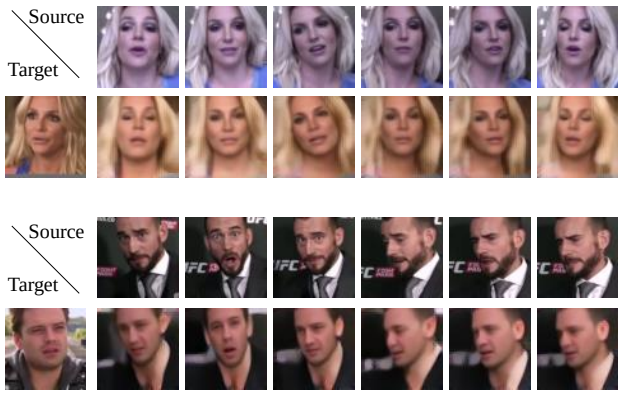

Figure 7. Image-to-image translation results on the facial motion sequence (as the source of face motion). We adopt different target images from test dataset of Voxceleb2.