Supplementary Materials

ACM Reference Format:
. 2022. Supplementary Materials. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 NETWORK STRUCTURE

1.1 Structure of AdaAT Module

Figure 1 illustrates the details of our AdaAT module. The basic blocks are as same in [2].

- Figure 1(a)~(d) show the basic blocks of the DownBlock, the ResBlock, the SameBlock and the Upblock.
- Figure 1(e) shows the details of the appearance encoder.
- Figure 1(f) shows the details of the transformation encoder.
- Figure 1(g) shows the details of the convolutional layers in feature alignment.
- Figure 1(h) shows the details of the appearance decoder.

1.2 Structure of Discriminator

The structure of discriminator is as same as in [2]. We set $D\_num\_blocks = 4$, $D\_block\_expansion = 64$ and $D\_max\_features = 256$.

2 IMPLEMENTATION DETAILS

In facial 3DMM reconstruction, we use scaled orthographic projection (weak perspective projection) to compute the 3DMM parameters of facial shape p_s, facial expression p_e, head rotation R, head translation T and scale s from facial landmarks. In cross-identity face reenactment, we only swap the p_e, R and T between source face and driving face.

In training stage, we use adam optimizer [1] with the default setting to optimize our network. The learning rate is set to 0.0001. On the HDTF dataset, we use a coarse-to-fine training strategy as same as in [3] to synthesize 512×512 resolution videos.

3 FEATURE MAPS VISUALIZATION

In Figure 2, we show more feature maps before and after 1_{st} AdaAT operation.

4 OVERFITTING ON IPER

In Figure 3, we show the over fitted results on iPER dataset. In the source/driving pose, the identity wears a black and white coat. In the synthetic pose, the identity is over fitted to the training image in coat texture.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

5 EXPLORATORY EXPERIMENT

5.1 How to improve the visual quality?

We reproduce previous methods [2, 5, 6] to explore the factors of improving visual quality in talkig face generation. In the training stage, we try two training strategies. The first one is the coarse-to-fine (256-to-512) strategy as in [4], i.e., we first train the network on 256×256 videos and then train the network on 512×512 videos. The second one is direct-training strategy, i.e., we directly train the network on 512×512 videos.

Figure 4 shows the qualitative results under several factors of training strategy, method, level of head motion and synthetic resolution. The rows (R) show the results of different resolutions and head motions. The columns (C) show the results of different methods and training strategies. In the comparison between training strategies, the coarse-to-fine strategy outperforms the direct-training strategy on all methods (see C1 *vs* C2, C3 *vs* C4, C6 *vs* C7). In [6], the visual quality largely depends on the level of head motion (see R1 vs R2, R3 vs R4). The large head motion causes blurry results on all resolutions. The main reason is that traditional CNN cannot realize misaligned image mapping.

We summarize three factors to improve the visual quality in talking face generation according to our experiments: (1) *High-resolution audio-visual dataset.* High-resolution audio-visual dataset is one necessary condition to improve the resolution of synthetic videos. (2) *Effective misaligned image mapping method* The method of misaligned facial image mapping is also very important if driving the head motions. (3) *Coarse-to-fine training strategy.* Coarse-to-fine training strategy outperforms direct-training strategy in our experiments.

5.2 If dense input necessary?

The work [5] proposes one viewpoint that the dense input is necessary for high resolution video generation according to their control variate experiments. Based on exploratory experiments, we propose another more reasonable explanation of their control variate experiments. We first review the experiments in their paper. They set two conditions in the experiments, including "sparse facial landmark input (sparse) *vs* approximate dense flow input (dense)" and "vanilla network (vanilla) *vs* explicit modeling in network (explicit)". The results show that the order of the visual quality is "dense + explicit" > "dense + vanilla" > "sparse + explicit" > "sparse + vanilla".

We propose another explanation of their results. In the experiment of "sparse + vanilla", the network can not realize misaligned facial image mapping, so this setting gets the worst visual quality. In the experiment of "sparse + explicit", the explicit modeling based network first compute the dense flow and then warp the feature map with the computed dense flow, so they can perform misaligned facial image mapping. However, the network is directly trained on 512×512 videos, so the visual quality deteriorates a lot. In the experiment of "dense + vanilla", the approximate dense flow warps the reference image to conduct misaligned image mapping. The network can be trained directly on 512×512 videos due to

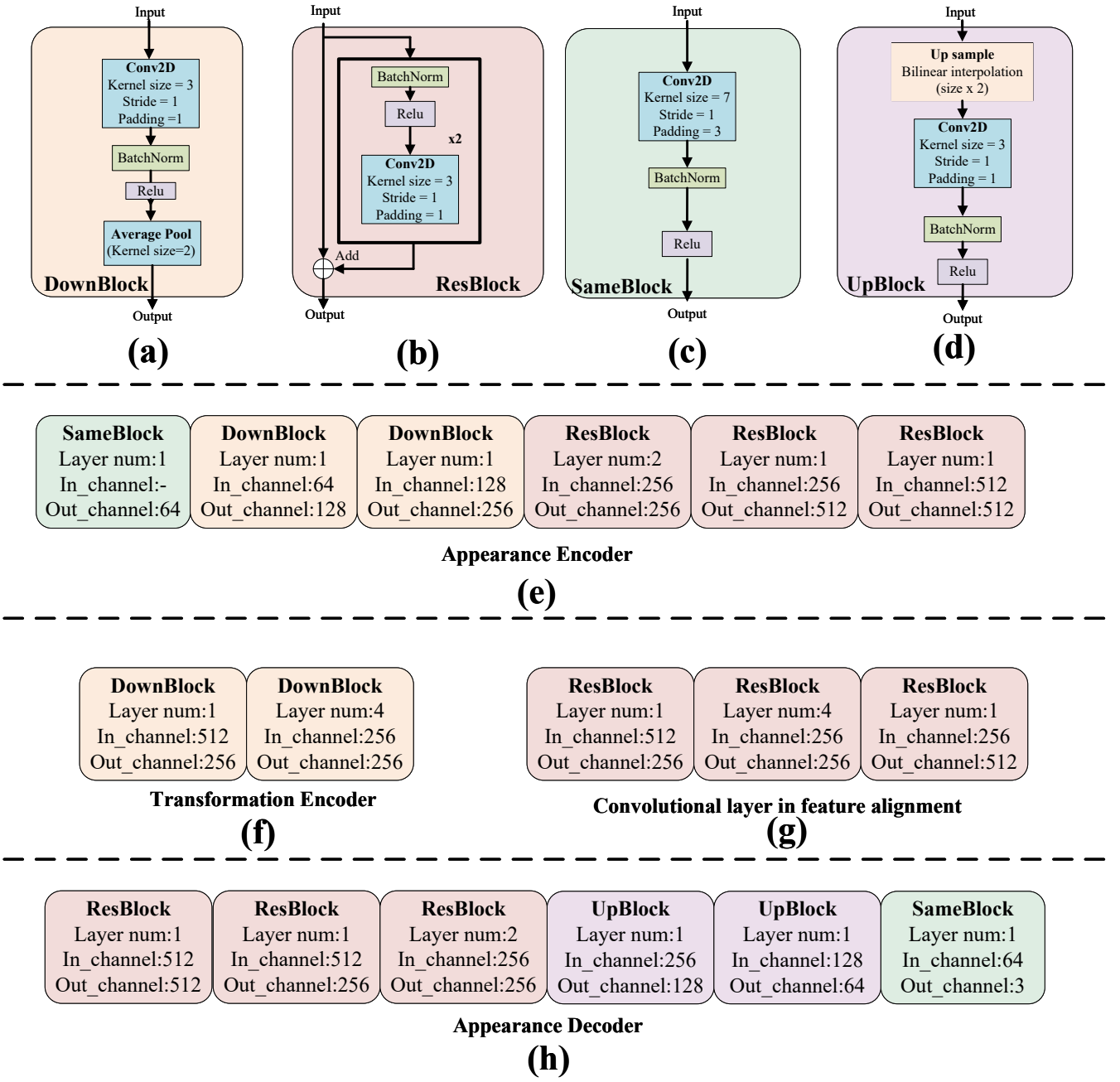

Figure 1: Structure of AdaAT Module.

the strong capacity of the input approximate dense flow, so the visual quality is better than "sparse + explicit". In the experiment of "dense + explicit", the explicit modeling based network refines the approximate dense flow to be more reliable, so this setting gets the best visual quality.

Our proposed explanation is more reasonable due to the following two reasons: (1) Our method proves the importance of misaligned image mapping method and coarse-to-fine strategy in improving the visual quality. (2) the sparse facial landmarks are proved effectively in our experiments.

6 DEMO VIDEO

A video demo is included in the supplementary material.

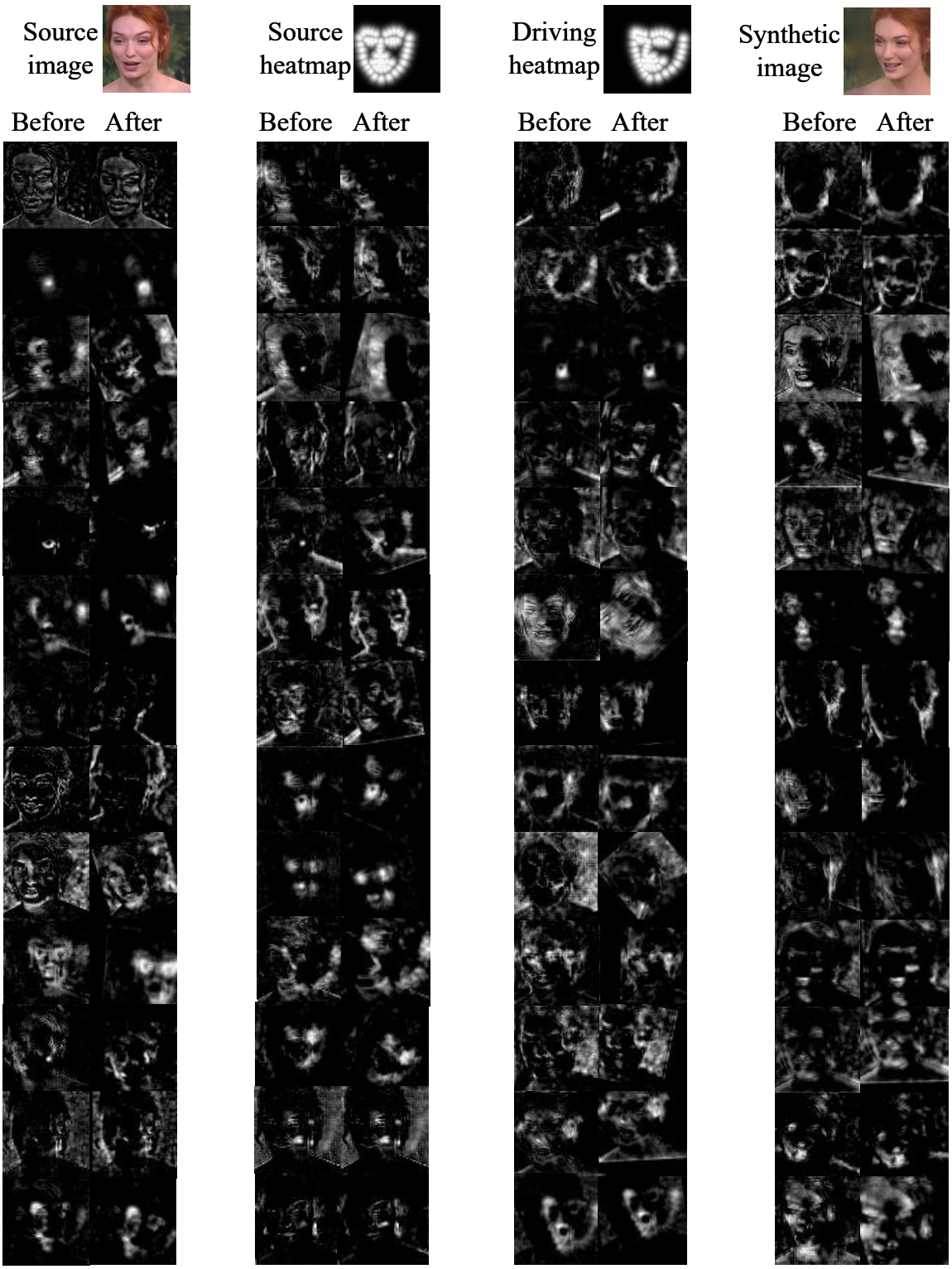

Figure 2: More feature maps before and after AdaAT.

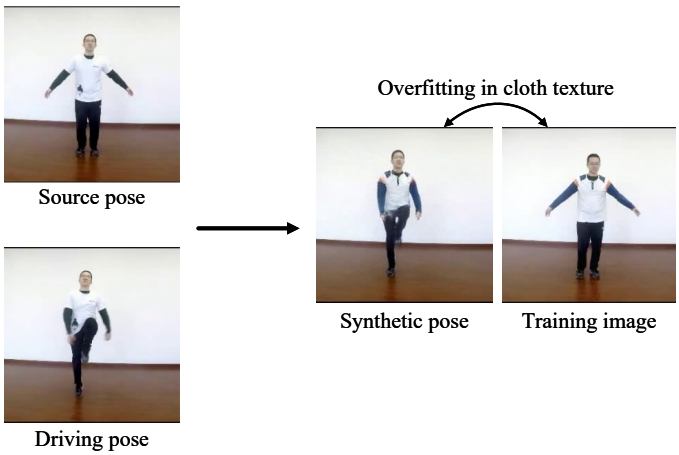

Figure 3: The over fitted identity in iPER dataset.

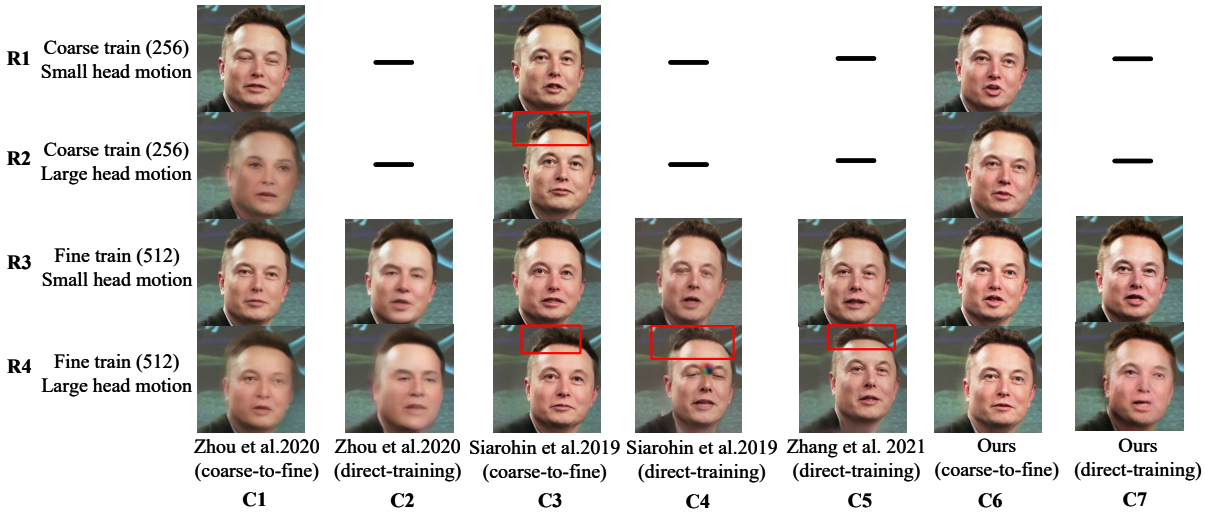

Figure 4: The results of exploratory experiment. Rows show the results of different resolutions and head motions. Columns show the results of different methods.

7 ETHICAL CONSIDERATION

We strongly advocate using our technology properly. To prevent the abuse of our method, anyone who employ our method to synthesize fake videos should mark with "fake video". Fortunately, detecting the synthetic and manipulated videos has got much attention and achieved much progress. As part of our responsibility, we are happy to promote the development of detection methodologies by sharing our dataset, source codes for their future research.

REFERENCES

[1] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[2] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems* 32 (2019).
[3] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10039–10049.
[4] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10039–10049.
[5] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3661–3670.
[6] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–15.