FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning Supplementary Document

Chenxu Zhang[1], Yifan Zhao[2], Yifei Huang[3], Ming Zeng[4], Saifeng Ni[5]
Madhukar Budagavi[5], Xiaohu Guo[1]
[1]The University of Texas at Dallas [2]Beihang University [3]East China Normal University
[4]Xiamen University [5]Samsung Research America
{chenxu.zhang, xguo}@utdallas.edu, zhaoyf@buaa.edu.cn, yifeihuang17@gmail.com
zengming@xmu.edu.cn, {saifeng.ni, m.budagavi}@samsung.com

Here we elaborate more technical details of our proposed FACIAL-GAN, including the temporal network $\mathbf{G}^{tem}$ and the local phonetic network $\mathbf{G}^{loc}$, and present personalization metric in detail.

1. Temporal Correlation Generator

We employ the temporal network $\mathbf{G}^{tem}$ to extract the temporal correlations of the T frames sequence. Here we set T as 128. As shown in Table 1, $\mathbf{G}^{tem}$ takes the DeepSpeech feature $\mathbf{a} \in \mathbb{R}^{128\times29}$ and initial state $\mathbf{s} \in \mathbb{R}^{71}$ as inputs, generating temporal features $\mathbf{z} \in \mathbb{R}^{128\times32}$. To be more specific, each convolution layer is followed by a leakyReLU activation and batch normalization [1].

2. Local Phonetic Generator

We employ the local phonetic network $\mathbf{G}^{loc}$ to generate local features $\mathbf{c_t} \in \mathbb{R}^{32}$ for the t-th frame. As shown in Table 2, $\mathbf{G}^{tem}$ takes the DeepSpeech feature $\mathbf{a_t} \in \mathbb{R}^{16\times29}$ as input and generates local features $\mathbf{c_t}$. Similarly, each convolution layer is followed by a ReLU activation and batch normalization. The Tanh activation is applied after the fully connected layer (FC).

3. Personalization Metric

Here, we discuss the details of our personalization metric which includes head pose classifier and eye blink classifier. To evaluate this personalization capability, we train a typical N-way pose classification network of N identities by matching their input head poses. Our motif is that the generated personalized information (*i.e.*, head pose and eye blink) of the kth identity should also be classified as the kth category.

Our proposed classifier is optimized on our training set with N identities in order to evaluate the head pose results obtained by different methods. Typical convolutional layers followed by an average pooling layer and one dense layer are used to implement the head pose classifier, where the output of the last fully connected layer is set to $N = 5$. The input is the head pose motion of 128 frames, resulting in $\mathbb{R}^{128\times6}$ dimensions. For the classification problem, the cross-entropy loss is applied in this task, and Adam [2] optimizer is used with the learning rate $lr = 0.0001$.

Table 1. Detailed architecture of Temporal Correlation Generator $\mathbf{G}^{tem}$.

Type	Kernel	Stride	Output
DeepSpeech	-	-	$1 \times 128 \times 29$
Conv 2D	4×4	2×2	$32 \times 64 \times 13$
Conv 2D	4×4	2×2	$128 \times 32 \times 5$
Conv 2D	4×4	1×1	$256 \times 32 \times 2$
Conv 2D	3×1	1×1	$256 \times 32 \times 2$
Reshape	-	-	256×64
Bilinear	-	-	256×128
Concat $\mathbf{s}$	-	-	327×128
Conv 1D	3	1	256×128
Conv 1D	3	1	256×128
Conv 1D	3	1	32×128

Table 2. Detailed architecture of Local Phonetic Generator $\mathbf{G}^{loc}$.

Type	Kernel	Stride	Output
DeepSpeech	-	-	$1 \times 16 \times 29$
Reshape	-	-	$29 \times 16 \times 1$
Conv 2D	3×1	2×1	$32 \times 8 \times 1$
Conv 2D	3×1	2×1	$32 \times 4 \times 1$
Conv 2D	3×1	2×1	$64 \times 2 \times 1$
Conv 2D	3×1	2×1	$64 \times 1 \times 1$
Reshape	-	-	64
FC	-	-	32

Similarly, we build an eye blink classifier to evaluate the personalization of generated videos. The eye blink classifier uses the same structure and training setting as the head pose classifier. We choose the best performance on the validation set, which has an accuracy of 90% for head pose classifier and 75% for eye blink classifier.

4. Ablation Study: Self-reenactment

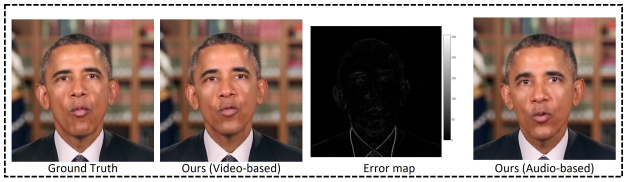

Figure 1. Ablation study for self-reenactment.

As shown in Fig. 1, we conduct two kinds of self-reenactment experiments. 1) Audio-based self-reenactment to evaluate face attribute generation: from LMD results in the main paper, our lip motion result is better than others quantitatively compared with GT sequences. 2) Video-based self-reenactment to verify the effectiveness of rendering network: the result of our face generation is very close to GT (see error map). Note that the error regions of clothes are caused by the slight offset of the body, which do not affect the quality of our result.

References

[1] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37, pages 448–456, 2015. 1

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 1