AU-Aware Dynamic 3D Face Reconstruction from Videos with Transformer Supplementary Material

Chenyi Kuang[1], Jeffrey O. Kephart[2], Qiang Ji[1]
[1] Rensselaer Polytechnic Institute, [2] IBM Thomas J. Watson Research Ctr.
{kuangc2,jiq}@rpi.edu, kephart@us.ibm.com

1. AU Spatial Correlations

We design an AU Spatial Correlation Module that takes geometry-based AU tokens as input, which are generated by applying multiple pre-defined vertex masks $\{\boldsymbol{M}_k^{au}\}_{k=1}^N$ to the input 3D mesh sequences $\{S_j\}_{j=1}^L$. The geometry-based AU-tokens $\boldsymbol{v}_k$ are expressed by

$$\boldsymbol{v}_k = \text{MLP}(\boldsymbol{M}_k^{au} \otimes (S_j - \bar{S})), k = 1, \cdots, N \quad (1)$$

where N is the total number of AUs. For each AU mask, a relevant face region consisting of a subset of vertices on 3D mesh is pre-defined. Elements in $\boldsymbol{M}_k$ corresponding to identified active vertices are assigned to "1", and "0" otherwise. Then we perform a smoothing of the weights in $\boldsymbol{M}_k$ for those boundary vertices to ensure the transition between "active" vertices and "inactive" vertices are smooth. In Fig. 1, we provide the visualization of $\boldsymbol{M}_k$ for 12 AUs that are involved in our experiments. The different active regions for different AUs, as an integrated prior knowledge in the model, reflects AU spatial correlations in terms of AU locations. For example, *AU1 (Inner Brow Raiser)* and *AU2 (Outer Brow Raiser)* are highly correlated as they have large number of overlapped active vertices. This is a general prior knowledge that can be applied universally, as the two AUs are controlled by the same facial muscles anatomically. Another kind of AU spatial correlations relates to the AU activation level, which will be learned by our model during the training. For example, a highly activated *AU12 (Lip Corner Puller)* will result in a subsequent activation of *AU6 (Cheek Raiser)*, but not for the lower-intensity case. The $\boldsymbol{M}_k$ are not updated during training or testing.

2. More Quantitative Evaluation

2.1. AU recognition

In addition to BP4D [6] and DISFA [4], we also show the performance of our model on Aff-Wild2 [3], which is a challenging dataset for expression recognition or AU detection due to various head pose, illumination and occlusion. We use 60% of Aff-Wild2 training data to train the model, for improving the model generalization ability under different environments. In Table. 1, we compare with two most recently published papers [1,5] performing multi-modal (image/video and audio) AU detection on Aff-Wild2 dataset. As we only use partial training data, in this paper we provide evaluations on the official validation set of Aff-Wild2. It shows that with the temporal model applied directly on videos, our performance is slightly better than [5]. But only using the mesh or combining the video and mesh will not contribute to performance improvement. We analyze the results carefully and identify the causes. As we have no access to any ground-truth 3D mesh data, we use a pre-trained reconstruction model to generate the input 3D mesh (coarse mesh) for every frame. The coarse mesh can be poorly aligned to the image by inaccurate head pose (error propagated from the coarse reconstruction model). However, in our model, only expression parameters are updated by the output embeddings of temporal module and spatial module and the final refined mesh sequences predicted by the transformer may remain to be incorrectly aligned. However, on BP4D and DISFA, our results prove to be better than SOTA methods, since the coarse

Table 1. Performances comparison with SOTA AU detection models. We show the average F1 score on the official validation set of Aff-Wild2. The CM is short for "correlation module". The last row represents our final model and the video input represents a temporal model and image input indicates a spatial model.

Method	Input Data	F1 score(in %)
Competetion baseline [2]	Image	39%
[5]	Audio	32.3%
[5]	Audio + CM	34.4%
[5]	Video	39.8%
[5]	Image + CM	47.9%
[5]	Video + CM	50.1%
[5]	Video + Audio + CM	52.3%
Transformer [1]	Image + Audio	52.5%
ours: $\mathcal{T}^S$ only	3D Mesh	35.2 %
ours: $\mathcal{T}^t$ only	Video	41.2%
ours: $\mathcal{T}^t + \mathcal{T}^S$	Video + 3D Mesh	40.8%

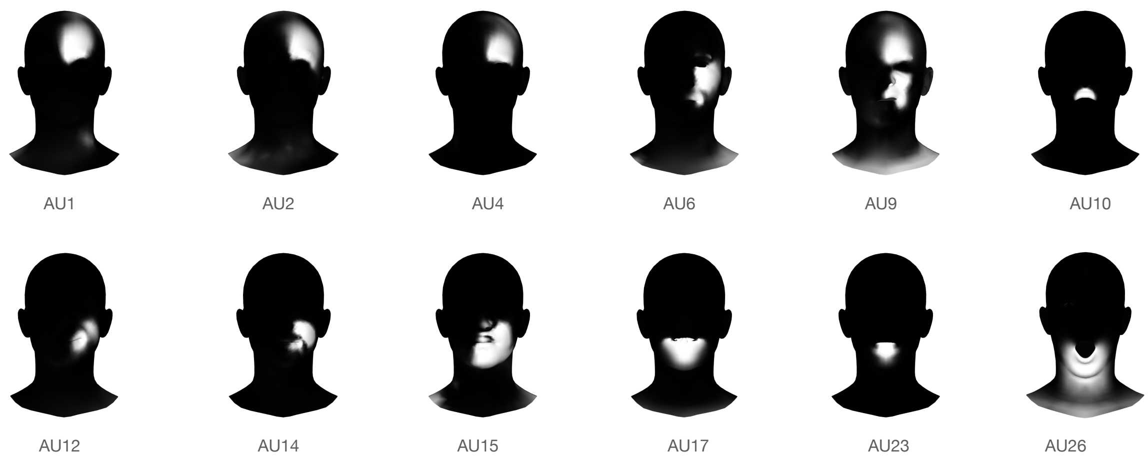

Figure 1. 3D AU vertex mask $\{M_k\}_{k=1}^N$, which will be applied on 3D face mesh to generate geometry-based AU tokens. Examples of 12 AUs are displayed.

reconstruction model perform well on these two datasets. This inspires us a promising direction to further improve our model, which is to refine the head pose alignment on Aff-Wild2 using our Temporal Module.

2.2. Inference time

We compare the inference time on image sequences and compare with SOTA 3D reconstruction model. On the evaluation set of multiface, we compare the average running time on a 20-frame sequence using different models in Table. 2, including DECA, EMOCA and DFNRMVS. Compare to EMOCA, which is also built on top of DECA basic model, our model has faster inference speed.

3. More Qualitative Results

In addition to quantitative AU detection results, we also show that our model can generate smooth and stable dynamic 3D face reconstruction which is also AU-aware. We provide more qualitative results in Fig. 2 and Fig. 3. The activated AUs inferred based on the geometry are also shown in the figure.

-	DECA	EMOCA	DFNRMVS	Ours
Test time (20 frames)	0.2105s	0.3052s	0.2642s	0.2794s

Table 2. Inference time comparison

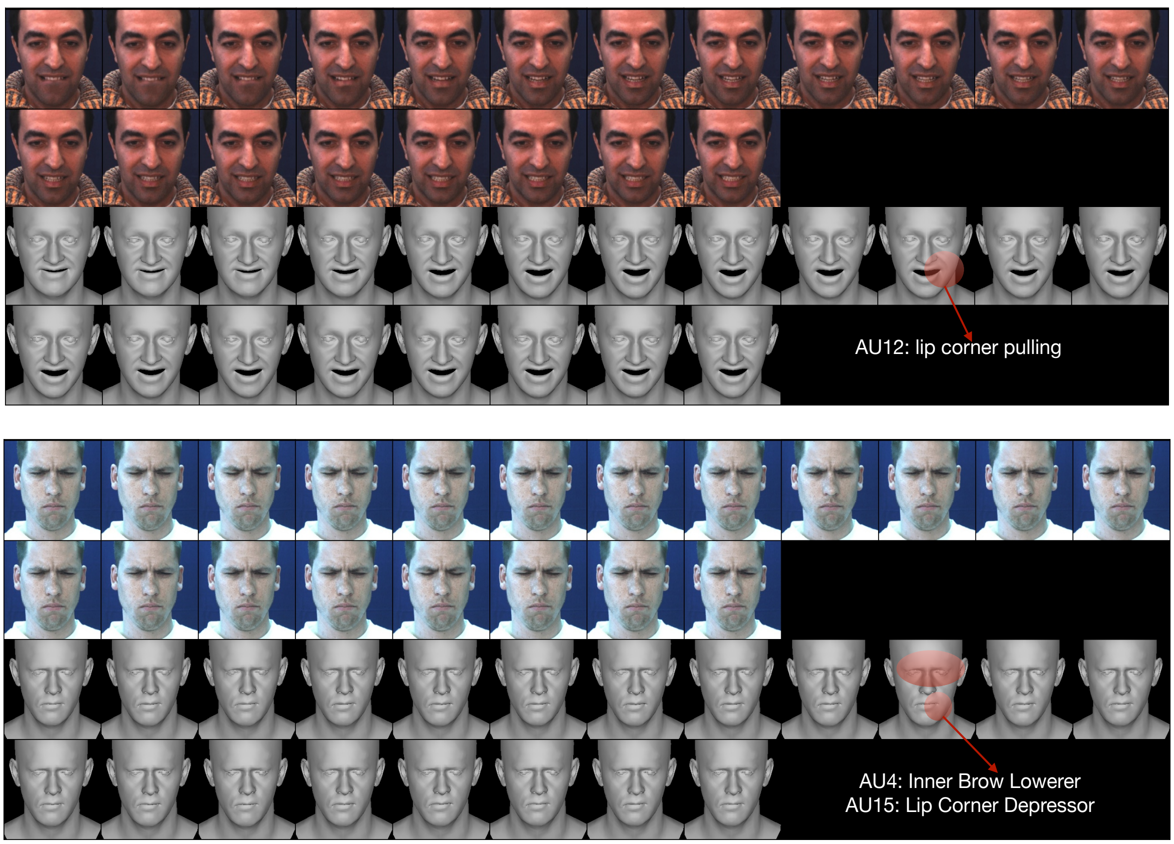

Figure 2. Dynamic 3D face reconstruction on validation data of DISFA.

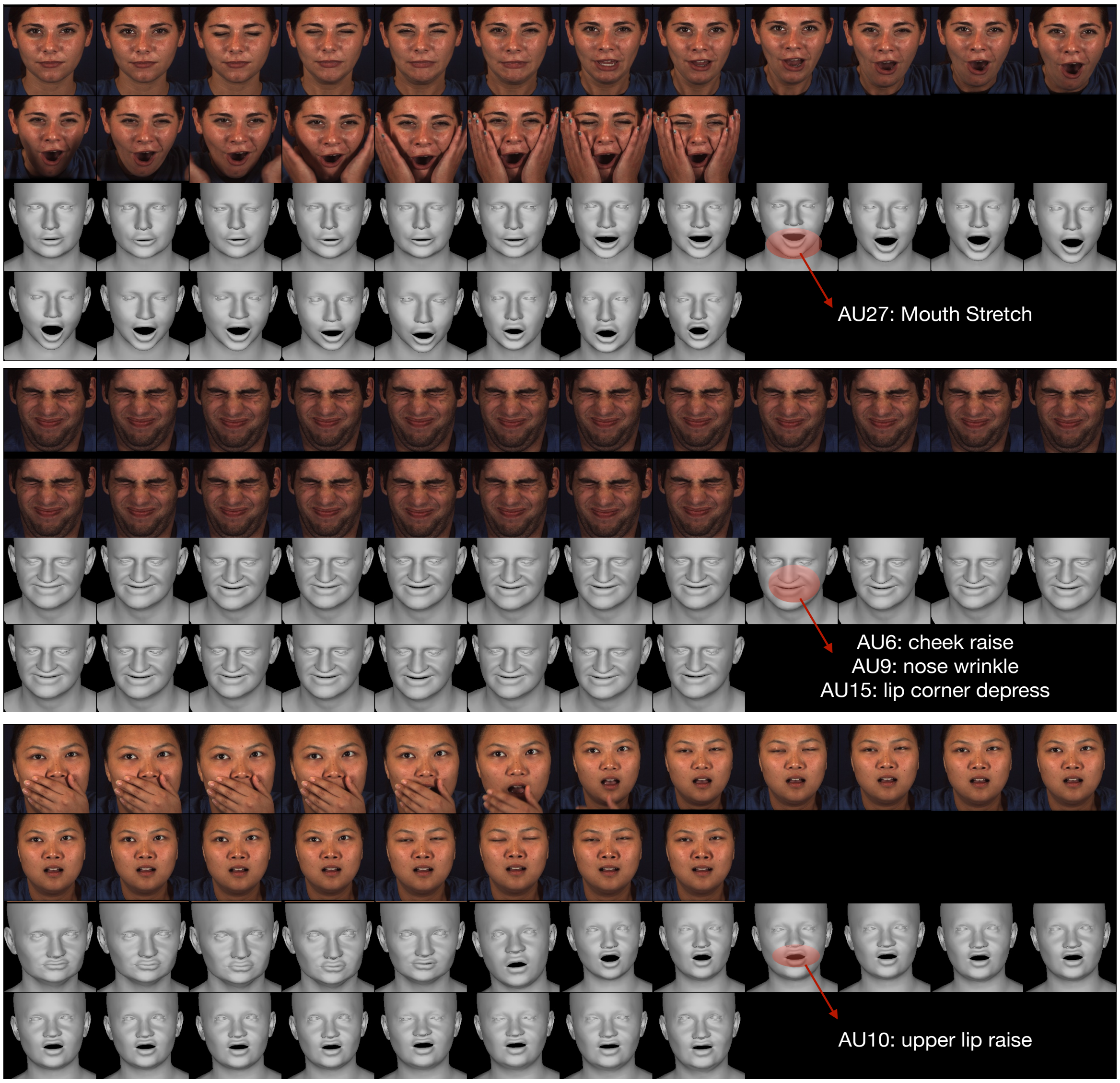

Figure 3. Dynamic 3D face reconstruction on validation data of BP4D.

References

[1] Yuanyuan Deng, Xiaolong Liu, Liyu Meng, Wenqiang Jiang, Youqiang Dong, and Chuanhe Liu. Multi-modal information fusion for action unit detection in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5854–5861, 2023. 1

[2] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2328–2336, 2022. 1

[3] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770, 2018. 1

[4] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing, 4(2):151–160, 2013. 1

[5] Lingfeng Wang, Jin Qi, Jian Cheng, and Kenji Suzuki. Action unit detection by exploiting spatial-temporal and label-wise attention with transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2470–2475, 2022. 1

[6] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing, 32(10):692–706, 2014. 1