Supplementary Materials for *Self-supervised Representation Learning from Videos for Facial Action Unit Detection*

Yong Li[1,2], Jiabei Zeng[1], Shiguang Shan[1,2,3,4] , Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, 200031, China
[4]Peng Cheng Laboratory, Shenzhen, 518055, China
`yong.li@vipl.ict.ac.cn, {jiabei.zeng, sgshan, xlchen}@ict.ac.cn`

We provide additional details about the back propagation of TCAE in Section 1, and additional visualization results in Section 2.

1. The back propagation of the operation $\mathcal{T}$

Let $(\delta x, \delta y)$ denote the learned offsets in the displacement $\mathcal{T}$, $\mathcal{L}$ denotes the combination of losses in Eq. $(1) \sim (6)$ in the main paper. In the back-propagation, the gradient of the loss function $\mathcal{L}$ *w.r.t.* δx is computed as

$$\frac{\partial \mathcal{L}}{\partial \delta x} = \frac{\partial \mathcal{L}}{\partial \mathcal{I}(x,y)} \frac{\partial \mathcal{I}(x,y)}{\partial \hat{x}} \frac{\partial \hat{x}}{\partial \delta x} = -\frac{\partial \mathcal{L}}{\partial \mathcal{I}(x,y)} \frac{\partial \mathcal{I}(x,y)}{\partial \hat{x}},$$

where $\mathcal{I}(x,y)$ is the pixel at location (x,y) in the target image $\mathcal{I}$. Its corresponding pixel in the source image $\mathcal{S}$ is near location $(\hat{x}, \hat{y}) = (x - \delta x, y - \delta y)$.

Considering the bilinear interpolation, the $\mathcal{I}(x,y)$ in forward is computed as

$$\mathcal{I}(x,y) = \sum_{m=\lfloor \hat{x} \rfloor}^{\lceil \hat{x} \rceil} \sum_{n=\lfloor \hat{y} \rfloor}^{\lceil \hat{y} \rceil} \mathcal{I}(m,n)(1 - |\hat{x} - m|)(1 - |\hat{y} - n|).$$

Therefore, in backward, $\frac{\partial \mathcal{I}(x,y)}{\partial \hat{x}}$ can be computed as

$$\frac{\partial \mathcal{I}(x,y)}{\partial \hat{x}} = \sum_{m=\lfloor \hat{x} \rfloor}^{\lceil \hat{x} \rceil} \sum_{n=\lfloor \hat{y} \rfloor}^{\lceil \hat{y} \rceil} \mathcal{I}_s(m,n)(1 - |\hat{y} - n|)\text{sgn}(\hat{x}, m),$$

where $\text{sgn}(\hat{x}, m)$ is 1 if $\hat{x} \leq m$, otherwise -1. The gradient of the loss function $\mathcal{L}$ *w.r.t.* δy (i.e., $\frac{\partial \mathcal{L}}{\partial \delta y}$) is calculated similarly.

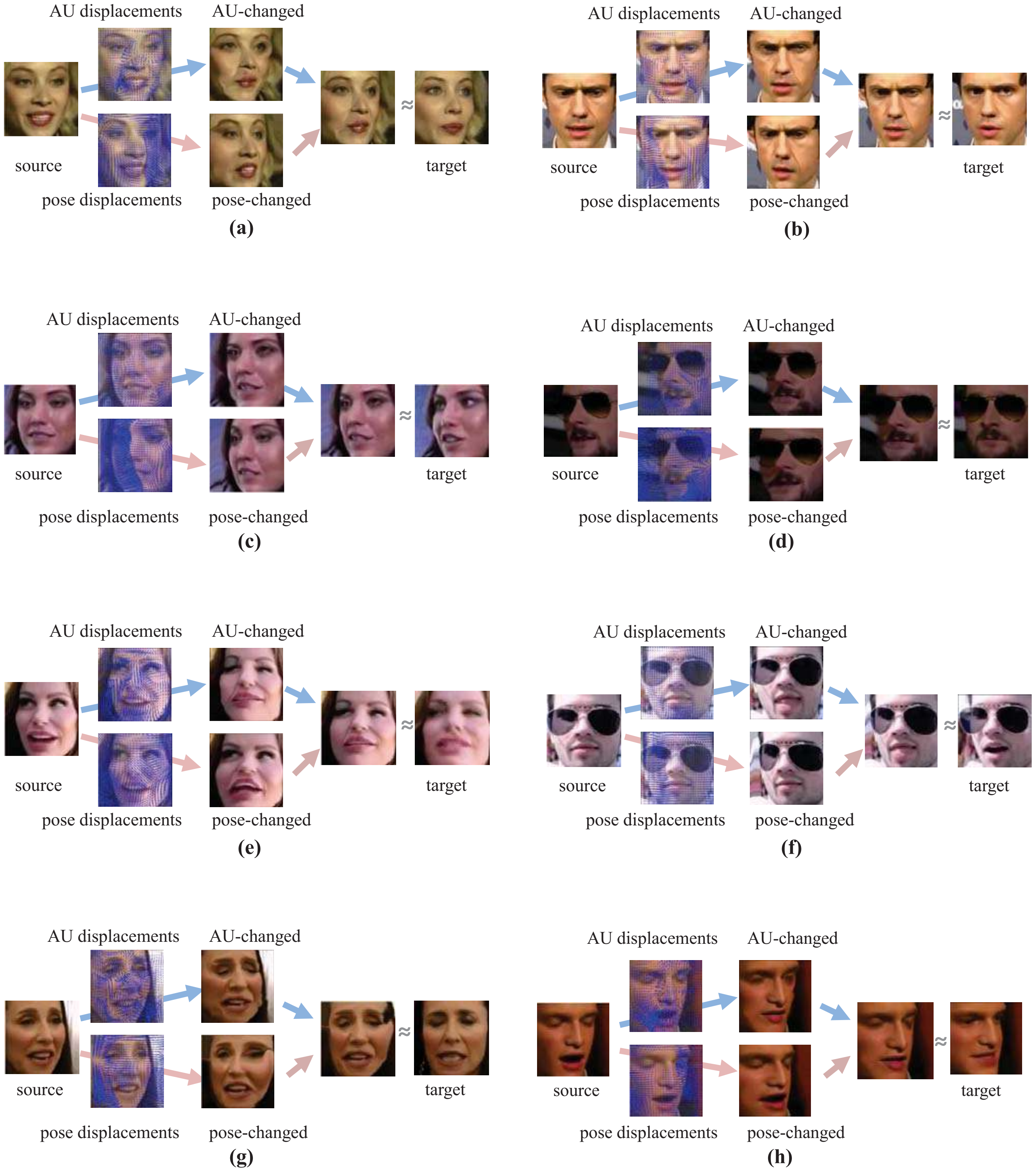
2. Additional visualization results

We have showed the pose and AU transforms for another 16 subjects in Fig. 1.

The generated AU-changed or the pose-changed face images in Fig. 1 are reasonable and quite realistic. The AU-changed face images preserve the poses in the source while have similar facial actions to the targets. The pose-changed face images preserve the same facial actions in the source but have similar head poses to the targets.

Compared with the pose displacements that display homogeneous directions, the offsets in AU displacements are sparse and show diverse orientations, which denote the local movements of facial muscles. The offsets in the AU displacements show various directions around eyebrows (e.g., Fig. 1 (e, g, j, k)), eyes (e.g., Fig. 1 (e, g, h, k, n, p)), mouth (e.g., Fig. 1 (a, c, d-o)), facial cheeks (e.g., Fig. 1 (a-e, g, j-p)).

Additionally, TCAE is capable of generating dense pose displacements when there is a big difference in the head pose of the source and target face images (e.g., Fig. 1 (c, d, g, i, k, n, p)). This is because the pixels in the source face should be moved further to compensate the gap caused by head pose.


AU displacements
AU-changed
source
≈
target
pose displacements
pose-changed
(a)
AU displacements
AU-changed
source
≈
target
pose displacements
pose-changed
(b)
AU displacements
AU-changed
source
≈
target
pose displacements
pose-changed
(c)
AU displacements
AU-changed
source
≈
target
pose displacements
pose-changed
(d)
AU displacements
AU-changed
source
≈
target
pose displacements
pose-changed
(e)
AU displacements
AU-changed
source
≈
target
pose displacements
pose-changed
(f)
AU displacements
AU-changed
source
≈
target
pose displacements
pose-changed
(g)
AU displacements
AU-changed
source
≈
target
pose displacements
pose-changed
(h)

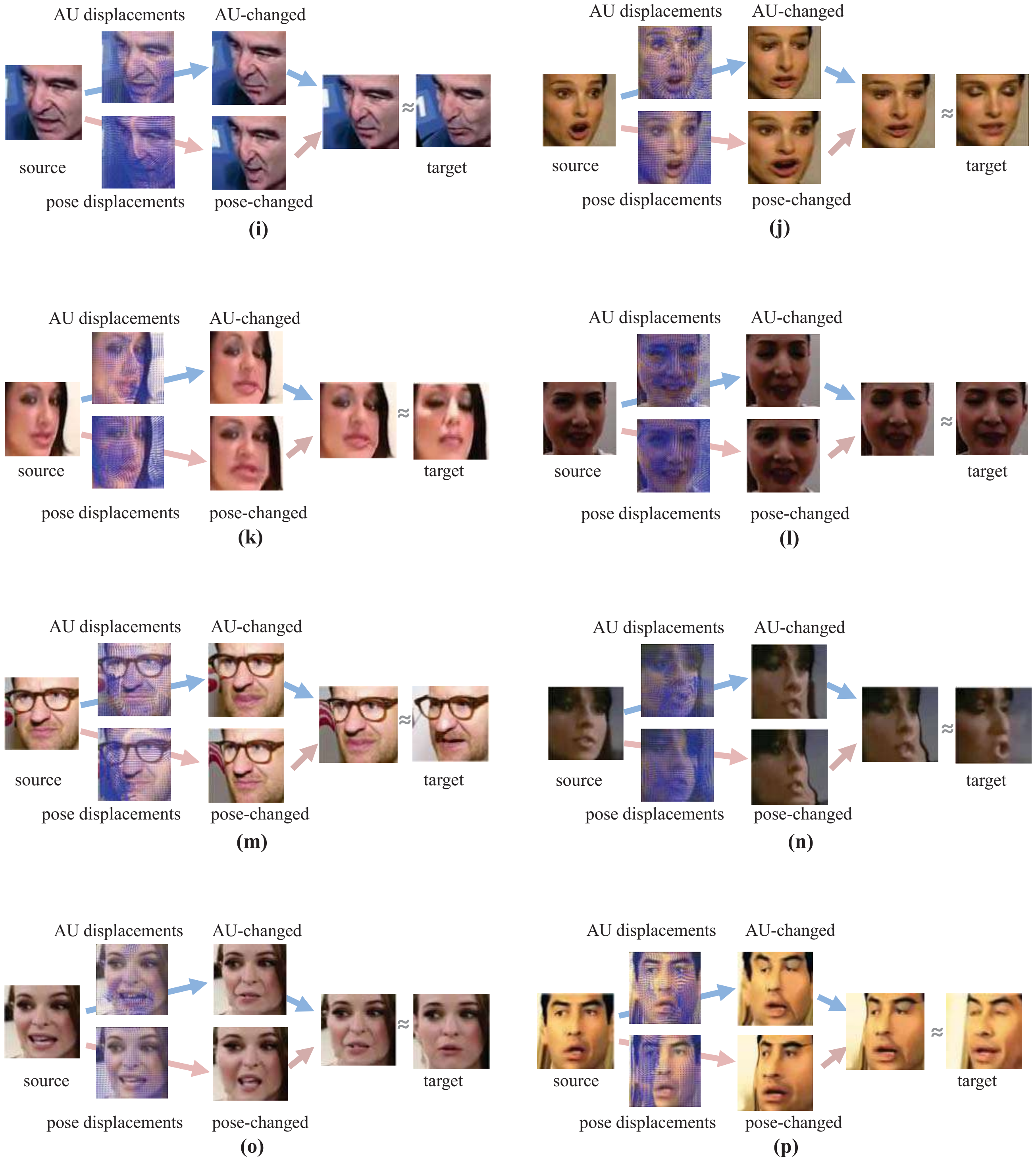


Figure 1. Visualizations of the pose and AU transforms for 16 subjects. The source is transformed to the AU-changed and pose-changed face images through the AU displacements and pose displacements respectively. AU-changed face image should has the same AUs as the target and the same pose as the source. Pose-changed face image should has the same pose as the target and the same AUs as the source. Better viewed in color and zoom in.