# Supplementary of "Towards Accurate Facial Motion Retargeting with Identity-consistent and Expression-exclusive Constraints"

In the supplementary materials, we provide more details and show more experimental results of the proposed method.

## 1 Discussion on Pseudo Identity Label

In the proposed identity-consistent constraint, we use the average of the predicted identity coefficients of the same person as the pseudo label. Record that our goal is to enforce a consistent prediction of identity coefficients for the same person. For that, we have tried anther two choices. One is using the median of predicted identity coefficients as the pseudo identity label, while another is to enforce the identity coefficients of the same person remain the same for all pairs of frames. We conduct experiments to compare these three choices on the facial motion retargeting task. As shown in Table 1, all three choices contribute to good expression estimation performance, which verifies the insight in our paper that explicitly enforcing the consistent prediction of identity plays an important role in capturing accurate expressions. Besides, using the average or median of predicted identities as pseudo label performs better than the "pair" choice. As a result, we use the average of predicted identity coefficients as pseudo label in our identity-consistent loss in the paper.

## 2 Landmark Combination Strategy

We calculate the landmark loss by measuring the difference between ground-truth 2D face landmarks $\mathbf{Q}$ and the corresponding projected landmarks $\hat{\mathbf{Q}}$ of the reconstructed 3D face. Numerous 3D face reconstruction methods use facial landmarks detected from a 3D face alignment method FAN (Bulat and Tzimiropoulos 2017) as ground-truth landmarks, since the 3D landmarks can describe the face contour correctly. However, the detected landmarks of FAN are not as accurate as 2D methods especially when the eyes are closed. To alleviate this issue, we use an additional offline 2D face alignment method Dlib (King 2009), which can detect more accurate landmarks when the eyes are closed.

To acquire more accurate facial landmarks as ground truth, we propose a landmark combination strategy to combine the facial landmarks from FAN and Dlib. Specifically, We run a head pose estimation algorithm (Yang et al. 2019) to estimate the yaw angle of the face image. When the estimated yaw angle $\leq 15^{\circ}$, we use the facial landmarks from Dlib for the inner face region (*i.e.*, eye, eye brow, nose and mouth) and facial landmarks from FAN for the face contour, and set the landmark weight of the eyes to 10. Otherwise, we only use the facial landmarks from FAN and set the landmark weight of the eyes to 1. Through this combination strategy, we obtain more accurate facial landmarks, which is helpful to estimate accurate expressions especially in the case of eyes closed.

## 3 Contradictory Expression Pair Set

We use the delta blendshapes (*i.e.*, displacements from the rest pose) taken from the FaceWarehouse database (Cao et al. 2013) as the expression blendshapes $\mathbf{D}_{exp}$. This blendshape model contains 46 expression units, most of which are described in Ekman's Facial Action Coding System (FACS) (Friesen and Ekman 1978). To avoid the co-occurrence of contradictory expression units, we define a contradictory expression pair set $\mathcal{O}$ from the blendshape model, and propose an expression-exclusive constraint to guide the model to suppress the contradictory expression units that shouldn't appear. Each pair $(i, j) \in \mathcal{O}$ is the subscript of $\boldsymbol{\beta}$. Here, $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ are the expression coefficients of a contradictory expression unit pair. We collect the elements of the contradictory expression pair set by selecting those expression units that exist in the same face region but cannot appear simultaneously. The contradictory expression pair set $\mathcal{O}$ is shown in Table 2.

Table 1: Comparisons of different choices of identity label on the facial motion retargeting in terms of MAE.

| Choice | Average | Median | Pair |
|---|---|---|---|
| Avg | **0.269** | 0.270 | 0.299 |

Table 2: The contradictory pairs in our defined contradictory expression pair set from the expression blendshape model.

| Pair Index | Contradictory Expression Unit Name |
|---|---|
| (1, 9) | (Right Eye Close, Right Eye Wide) |
| (3, 9) | (Right Lower Lid Raise, Right Eye Wide) |
| (2, 10) | (Left Eye Close, Left Eye Wide) |
| (4, 10) | (Left Lower Lid Raise, Left Eye Wide) |
| (5, 13) | (Right Upper Lid Droop, Right Upper Lid Raise) |
| (6, 14) | (Left Upper Lid Droop, Left Upper Lid Raise) |
| (7, 11) | (Right Eye Slide Left, Right Eye Slide Right) |
| (8, 12) | (Left Eye Slide Right, Left Eye Slide Left) |
| (15, 18) | (Right Brow Lower, Right Brow Raise) |
| (16, 19) | (Left Brow Lower, Left Brow Raise) |
| (15, 17) | (Right Brow Lower, Brow Raise) |
| (16, 17) | (Left Brow Lower, Brow Raise) |
| (21, 23) | (Jaw Slide Right, Jaw Slide Left) |
| (24, 25) | (Mouth slide Right, Mouth slide Left) |
| (26, 28) | (Right Lip Corner Lower, Right Lip Corner Pull (strong)) |
| (26, 30) | (Right Lip Corner Lower, Right Lip Corner Pull (weak)) |
| (26, 32) | (Right Lip Corner Lower, Right Lip Corner Stretch) |
| (27, 29) | (Left Lip Corner Lower, Left Lip Corner Pull (strong)) |
| (27, 31) | (Left Lip Corner Lower, Left Lip Corner Pull (weak)) |
| (27, 33) | (Left Lip Corner Lower, Left Lip Corner Stretch) |
| (34, 36) | (Upper Lip Suck, Upper Lip Raise) |
| (34, 42) | (Upper Lip Suck, Upper Lip Open) |
| (35, 37) | (Lower Lip Suck, Lower Lip Depress) |
| (35, 38) | (Lower Lip Suck, Lower Lip Open) |
| (41, 37) | (Lower Lip Close, Lower Lip Depress) |
| (41, 38) | (Lower Lip Close, Lower Lip Open) |

## 4 Expression Test Set of FEAFA

FEAFA (Yan et al. 2019) is a facial expression dataset, which contains 123 facial videos of 122 identities with about 10K frames. We divide the videos into two parts, 103 videos as the training set and the remaining 20 videos as the test set. Note that FEAFA presents well-annotated Action Unit(AU) intensity labels for each video frame based on the Facial Action Coding System(FACS) (Friesen and Ekman 1978), which is

Table 3: Comparisons of facial motion retargeting accuracy (measured by Mean Absolute Error (MAE)) on FEAFA test set with different methods **under lightweight network**. A Lower error means the method performs better for capturing expressions.

| Method | Eye Close | Brow Lower | Brow Raise | Mouth Open | Lip Suck | Lip R/D | Kiss | Nose Wrinkle | Lip Corner Pull | Lip Corner Stretch | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MS-SFN | 0.204 | 0.585 | 0.358 | 0.430 | 0.396 | 0.381 | 0.687 | 0.909 | 0.509 | 0.502 | 0.496 |
| RingNet | 0.247 | 0.380 | 0.370 | 0.327 | 0.714 | 0.315 | 0.509 | 0.805 | 0.470 | 0.391 | 0.453 |
| Personalized | 0.268 | 0.369 | 0.364 | 0.200 | 0.590 | 0.334 | 0.521 | 0.738 | 0.442 | 0.353 | 0.418 |
| Ours (w/o $\mathcal{L}_{idc}$ + $\mathcal{L}_{exp}$) | 0.424 | 0.279 | 0.504 | 0.398 | 0.828 | 0.389 | 0.564 | 0.857 | 0.501 | 0.422 | 0.516 |
| Ours (w/ $\mathcal{L}_{idc}$) | **0.163** | 0.243 | **0.077** | 0.164 | **0.423** | 0.305 | 0.320 | 0.498 | 0.415 | 0.301 | 0.291 |
| Ours (w/ $\mathcal{L}_{idc}$ + $\mathcal{L}_{exp}$) | **0.163** | **0.222** | 0.127 | **0.154** | 0.455 | **0.274** | **0.263** | **0.436** | **0.379** | **0.278** | **0.275** |

consistent with our expression blendshapes. To measure the expression accuracy, we collect an expression test set from the first 20 videos (test set) of FEAFA by selecting some extreme expression frames according to the annotated AU labels. Specifically, we select the frames with annotated AU value greater than 0.8, and manually filter some redundant and blurred frames. The number of images in each of the expression categories are: eye close: 187, brow lower: 78, brow raise: 99, mouth open: 173, lip corner pull: 123, lip corner stretch: 41, lip suck: 142, lip raise/depress: 137, kiss: 138, nose wrinkle: 55, total: 1173 images.

## 5 Reimplementation Details about Baselines

We compare our method with MS-SFN (Chaudhuri, Vesdapunt, and Wang 2019) and Personalized (Chaudhuri et al. 2020) on the facial motion retargeting task. MS-SFN proposes a multi-scale single face network to regress the identity and expression coefficients of a bilinear 3DMM model (Cao et al. 2013). Personalized proposes to jointly learn a personalized face blendshape model and estimate the tracking parameters in a multi-frame framework. To better learn the personalized blendshape model, it aggregates the identity features from multiple images by average pooling.

We reimplement the above two methods since the code and models are unavailable. For a fair comparison, we use the same training datasets and keep the same settings as ours in these two methods. Note that Personalized uses their own 3D face model which is not publicly released and focuses on constructing personalized expression blendshapes, while we focus on accurate facial motion retargeting in our paper. For these reasons, we mainly reimplement Personalized with average aggregation mechanism using our face model and neglect the construction of personalized expression blendshape. For a fair comparison, we use the same backbone as ours for Personalized in the experiments.

## 6 More Experiment Results using MobileNet-V2 Backbone

In facial motion retargeting applications, it is also important to achieve real-time performance. Therefore, we conducted experiments on a light-weight backbone MobileNet-V2 (Sandler et al. 2018) to verify the effectiveness of our method. Specifically, Personalized (Chaudhuri et al. 2020), RingNet (Sanyal et al. 2019) and our CPEM use MobileNet-V2 backbone, while MS-SFN (Chaudhuri, Vesdapunt, and Wang 2019) keeps its specially designed multi-scale network structure. The results on facial motion retargeting are shown in Table 3. Our method consistently outperforms the baseline methods thanks to the effectiveness of our proposed identity-consistent and expression-exclusive constraints. Compared with the ResNet-50 backbone used in the paper, the average MAE only increased by 0.031 (0.275 vs 0.244), which shows that our method can achieve good performance under the lightweight network framework. Moreover, compared with other methods using the ResNet-50 backbone, our CPEM using MobileNet-V2 backbone still achieves the lowest MAEs (see Table 1 in the paper), which further verifies the effectiveness of the proposed method.

Table 4: Effect of frame number $T$ on the facial motion retargeting in terms of MAE.

| $T$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| Avg | 0.290 | **0.244** | 0.255 | 0.251 |

## 7 Influence of Frame Number $T$

The number of frames is an important hyper-parameter in our multi-frame training framework. Thus, we conduct experiments on different number of frames to investigate its effect. The results are shown in Table 4. With fewer frames (*i.e.*, $T = 2$) from a video clip, it is not sufficient to acquire a stable and accurate pseudo identity label to compute the identity-consistent loss, which slightly hampers the performance of the proposed method. As the number of frames increases (*i.e.*, from 4 to 8), the performance of the proposed method is stable, which suggests when giving sufficient frames, our approach is insensitive to the changes of frame number. Overall speaking, the proposed method is relatively robust to the changes of frame number when the number of frames is sufficient. Considering the training cost and performance gain, we set $T = 4$ in all our experiments.

## 8 Failure Case Analysis

We show some examples of side profiles and other challenging cases (*i.e.*, extreme occlusion and illumination) of our proposed method on AFLW2000-3D in Figure 1. We observe satisfactory results on images of side profiles, which is also verified by the face alignment results in Table 2 of

the paper. However, for more extreme cases (**occlusion or illumination**), the 3D face reconstruction and facial motion retargeting performance may drop a bit. In the future, we will attempt to address these issues and develop a more robust model.

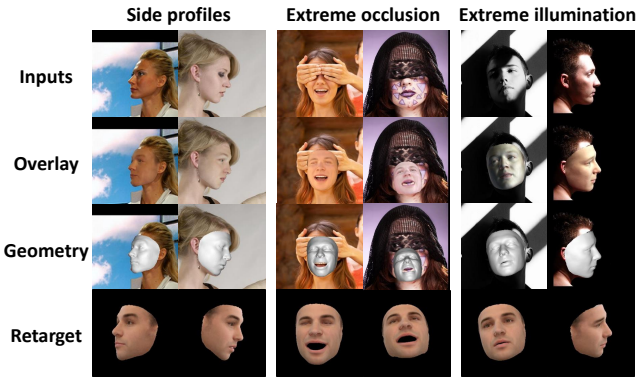


Figure 1: Examples of some challenging cases with side profiles, extreme occlusion and illumination conditions.

## References


Bulat, A.; and Tzimiropoulos, G. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*.

Cao, C.; Weng, Y.; Zhou, S.; Tong, Y.; and Zhou, K. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3): 413–425.

Chaudhuri, B.; Vesdapunt, N.; Shapiro, L.; and Wang, B. 2020. Personalized Face Modeling for Improved Face Reconstruction and Motion Retargeting. In *Proceedings of the European Conference on Computer Vision*.

Chaudhuri, B.; Vesdapunt, N.; and Wang, B. 2019. Joint face detection and facial motion retargeting for multiple faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9719–9728.

Friesen, E.; and Ekman, P. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2): 5.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10: 1755–1758.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Sanyal, S.; Bolkart, T.; Feng, H.; and Black, M. J. 2019. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7763–7772.

Yan, Y.; Lu, K.; Xue, J.; Gao, P.; and Lyu, J. 2019. Feafa: A well-annotated dataset for facial expression analysis and 3d facial animation. In *IEEE International Conference on Multimedia & Expo Workshops*, 96–101. IEEE.

Yang, T.-Y.; Chen, Y.-T.; Lin, Y.-Y.; and Chuang, Y.-Y. 2019. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1087–1096.