

Appendix

Yuxiang Mao^{1,2}, Zhenfeng Fan¹, ZhiJie Zhang^{1,2}, Zhiheng Zhang^{1,2}, Shihong Xia^{1,2*}

¹Institute of Computing Technology, Chinese Academy of Sciences

²University of Chinese Academy of Sciences

1. Overview

The supplementary material for our paper includes this appendix and a supplementary video. The video provides a summary of our models along with additional visualized examples, particularly video demonstrations. Here, we provide more comprehensive explanations of quantitative experiments, additional qualitative results, and further analysis of our methods.

2. More Qualitative Results

Fig. 1 presents additional qualitative comparisons with existing detailed face reconstruction methods. Compared to previous works [3, 5, 18, 24], particularly those capable of producing animatable details [3, 5], our model shows improvements in detail richness and accuracy. Furthermore, our transferred person-specific models effectively capture unique wrinkle patterns specific to individuals while maintaining robustness to occlusions. Fig. 2 presents additional examples of the animation quality of our base model and person-specific models compared to state-of-the-art detail animation models [3, 5]. The details generated by our models are more feature-aligned and intuitive.

3. More Details about Quantitative Results

We provide additional details on the data processing for the quantitative experiments, the calculation methods for metrics, and the Cumulative Error Distribution (CED) curves (Fig. 3) for each metric.

300-W Dataset. 300-W [12, 13] is a publicly available facial landmark dataset that provides a training set consists of about 2000 usable images annotated with 68 facial landmarks. We first conducted data cleaning, discarding images with multiple individuals or with severe omissions of facial regions, as all models performed poorly on such images, failing to demonstrate differences in face alignment performance among the models. On 1424 cleaned images, we calculate the average RMSE error [13] for the 51 inner landmarks between the predictions of each model and



Figure 1. **More Comparison on detail shape reconstruction.** From left to right: Input image, FaceScape [24], FaceVerse [18], DECA [5], EMOCA-v2 [3], our base model, and our person-specific models.

the ground truth. Additionally, for DECA [5] and EMOCA (three versions) [3] that also use the FLAME model, we ex-

*corresponding author, xsh@ict.ac.cn

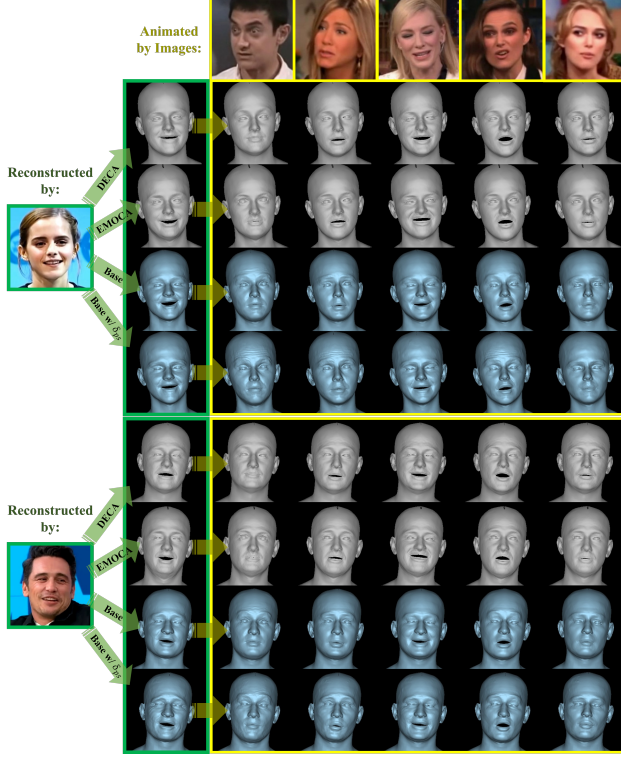


Figure 2. **Comparison on face animation.** Given a source image, DECA [5] (row 2, 6), EMOCA-v2 [3] (row 3, 7), and our base (row 4, 8) and person-specific (row 5, 9) models can respectively generate detailed 3D faces (green boxes). With a driving image (yellow boxes), these models can drive the face to exhibit corresponding expressions.

tract the predicted facial silhouette vertices, computing the average Euclidean distance from the 17 ground truth silhouette landmarks to the nearest silhouette vertices, normalized by the interocular distance, which is defined as the distance between the outer points of the eyes [13]. Given that there are 17 boundary and 51 inner landmarks in the 68-landmark annotation, we average the errors with a 1:3 weighting as the overall error.

300-VW Dataset. A protocol similar to 300-W is used for evaluation on 300-VW [14], in which we calculate the average RMSE error [13] for the 51 inner landmarks between the predictions of each model and the ground truth and additional boundary error for DECA [5] and EMOCA-v2 [3]. We average inner error and boundary error with a 1:3 weighting as the overall error.

FaceScape Dataset. After we complete the 68-landmark annotation (17 facial boundary and 8 inner mouth circle landmarks) with the results from the HRNet landmark detection [16]. The RMSE error between the predicted landmarks and the ground truth is then calculated [13]. Since SynergyNet [22] and 3DDFA-v2 [7] lack selecting methods for dynamic boundary landmarks, we separately calculate errors for the 51 static inner landmarks and all 68 land-

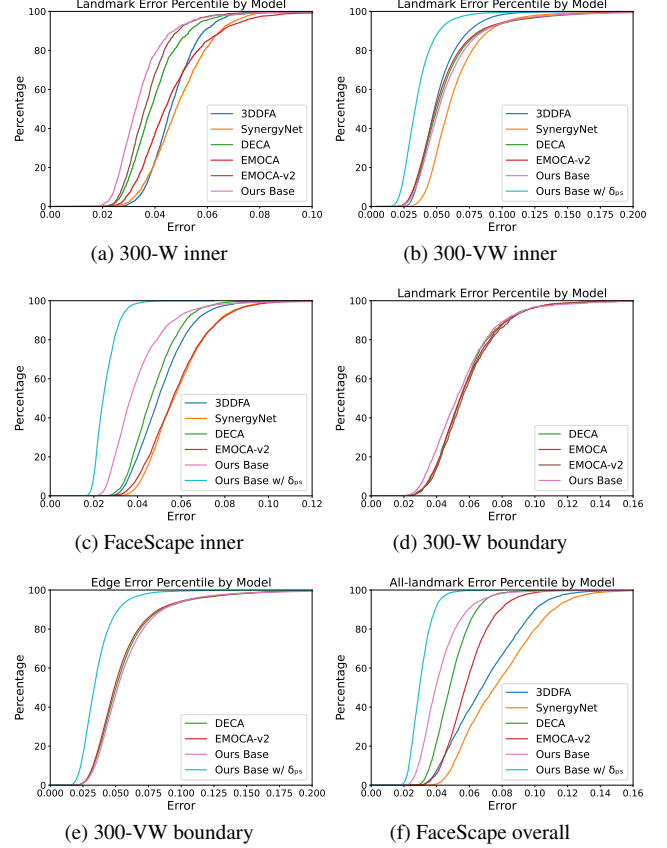


Figure 3. **Quantitative comparison to state-of-the-art.** The plots show the Cumulative Error Distribution (CED) curves with respect to the 51 inner facial landmarks and the boundary edge error (for 300-W [13] and 300-VW [14]), or the 51 inner facial landmarks and all 68 landmarks (for FaceScape [24]).

marks, calculating mean and variance for each identity per image, and then averaging across identities.

4. Further Explanation of Silhouette Loss

Fig. 4 shows the changes of our silhouette vertices (green) and the 2D landmarks provided by FLAME (red) as the model’s pose and expression vary. The FLAME model [10] provides a method to update the 17 3D boundary landmarks based on the face’s pose. FLAME uses a vector $\tau \in \mathbb{N}^{68}$ to represent the landmark indices. For boundary landmarks, FLAME pre-defines 78 sets of vertex indices $T \in \mathbb{N}^{78 \times 17}$, each corresponding to a certain pose. The model calculates rotation angles from the pose parameters to find the closest set of boundary landmarks. However, variations in facial shape and expression can also affect the vertex indices of these boundary landmarks. In contrast, we represent model silhouette edges using dense silhouette vertices, which are determined by the current vertex normal distribution. This approach more accurately depicts the 3D outer boundary

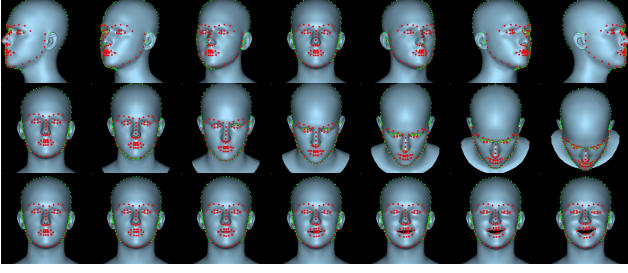


Figure 4. **Changes in silhouette vertices (green) and FLAME’s 2D landmarks (red) as pose and expression vary.** Row 1: Head rotation from 60° left to 60° right. Row 2: Head tilt from a frontal view to a 60° downward tilt. Row 3: Transition from a neutral expression to an increasingly happy expression.

edges and, due to its dense representation along the edges, naturally reduces tangential errors inherent in manual silhouette landmark annotations. Additionally, our method is compatible with arbitrary facial model topologies.

5. More Analysis about Teacher-Student Loss

The high capability of MAE allows it to better capture facial details compared to CNN-based architectures and provides a structural foundation for inserting person-specific adapters during subsequent person-specific transfer. However, during training, we observed that using only the shape-from-shading method (as in DECA [5]) to train the MAE resulted in displacement maps with significant artifacts, leading to numerous unnatural bumps on the detailed 3D face. This occurs because the shape and the rendered RGB image do not have a one-to-one correspondence. For the shape-from-shading optimization problem, there are numerous possible solutions, and these artifact-laden results can be one of them, as the rendered output appears artifact-free and closely matches the input image (as shown in Fig. 5). The Teacher-Student Loss leverages an important prior: smooth regions in the input image usually do not contain wrinkles, whereas areas with significant shading variations are typically caused by corresponding local facial details. The UNet structure naturally captures this prior, so we



Figure 5. **Detail reconstruction given by our base model, w/o or w/ our innovative teacher-student loss.** From left to right: input images; reconstructed detailed shape and its rendering produced by the base model w/o \mathcal{L}_{Tchr} ; displacement map from the base model w/o \mathcal{L}_{Tchr} ; reconstructed detailed shape and displacement map produced by the base model w/ \mathcal{L}_{Tchr} .

train a shallow UNet as the teacher to estimate the displacement map, guiding the MAE optimization towards producing more natural and intuitive results.

6. Loss Function

In this section, we provide a more detailed explanation of the loss functions used during training, excluding the silhouette vertex re-projection loss and teacher-student loss, which are already discussed in detail in the main paper.

6.1. Coarse Reconstruction Losses

Inner Landmark Re-Projection Loss. The landmark loss measures the L1 distance between annotated 2D ground-truth landmarks at the inner facial region $\mathbf{k}_i \in \mathbb{R}^2, i \in 18, \dots, 68$ and the projection of the corresponding landmarks $\tilde{\mathbf{k}}_i \in \mathbb{R}^3, i \in 18, \dots, 68$ on the FLAME mesh. The landmark loss is defined as:

$$\mathcal{L}_{inL} = \sum_{i=18}^{68} d(\mathbf{k}_i, \tilde{\mathbf{k}}_i) = \sum_{i=18}^{68} \left\| \mathbf{k}_i - s\Pi(\tilde{\mathbf{k}}_i) + \mathbf{t} \right\|_1. \quad (1)$$

Special Landmark Pairs Loss. The special landmark pairs loss is calculated on a set of landmark pairs (e.g., upper/lower eyelid or lips landmark pairs) S , by penalizing the relative positional differences between these landmarks to more effectively capture features such as the opening and closing of the eyes and mouth:

$$\mathcal{L}_{spL} = \sum_{(i,j) \in S} \left\| \mathbf{k}_i - \mathbf{k}_j - s\Pi(\tilde{\mathbf{k}}_i - \tilde{\mathbf{k}}_j) \right\|_1. \quad (2)$$

Photometric Loss. The photometric loss constrains the reconstructed image to closely resemble the input image and is calculated as:

$$\mathcal{L}_{pho} = \|R_I \odot (I - I_r)\|_1, \quad (3)$$

where R_I is a mask of the face skin region, with value 1 in the region and value 0 elsewhere, given by an open-source algorithm.

Perceptual Loss. The perceptual loss combines three perceptual losses to ensure high-level identity [5] and emotion consistency [3], as well as accurate lip movements [6].

$$\mathcal{L}_{per} = \mathcal{L}_{id} + \mathcal{L}_{emo} + \mathcal{L}_{lr} \quad (4)$$

Identity Loss. We introduce DECA’s identity loss to ensure that the reconstruction and input images are consistent at a high-level identity level. The loss measures the cosine similarity between the rendered images and input images in the feature embeddings of a pre-trained face recognition network and is calculated:

$$\mathcal{L}_{id} = 1 - \frac{f(I) \cdot f(I_r)}{\|f(I)\|_2 \cdot \|f(I_r)\|_2} \quad (5)$$

Emotion Consistency Loss. We introduce EMOCA’s emotion consistency loss to ensure the emotional content of the reconstructed 3D face aligns with that of the input image [3]. This loss measures the perceptual difference between the rendered images I_r and input images I using their respective emotion features $\epsilon_I = \text{Emo}(I)$ and $\epsilon_r = \text{Emo}(I_r)$:

$$\mathcal{L}_{emo} = \|\epsilon_I - \epsilon_{I_r}\|_2 \quad (6)$$

Perceptual Lip Movements Loss. We introduce SPEC-TRE’s perceptual lip movements loss to enhance the accuracy of mouth movement details in the reconstructed face [6]. This loss quantifies the perceptual discrepancy in speech-aware movements between the rendered images I_r and input images I through their respective “lipread” features $\mathbf{m}_I = LR(I)$ and $\mathbf{m}_r = LR(I_r)$, assessed via their cosine similarity:

$$\mathcal{L}_{lr} = 1 - \frac{\mathbf{m}_I \mathbf{m}_r}{\|\mathbf{m}_I\|_2 \cdot \|\mathbf{m}_r\|_2} \quad (7)$$

Regularization Loss. The regularization losses regularize the shape, expression, and albedo parameters, thereby preventing the model from overfitting too much noise from the training data:

$$\mathcal{L}_{reg} = \|\beta\|_2^2 + \|\psi\|_2^2 + \|\alpha\|_2^2 \quad (8)$$

Shape Consistency Loss. We introduce DECA’s shape consistency loss [5], which is based on the observation that for different images of the same individual, swapping their shape parameters should yield the same reconstruction. Therefore, when replacing the shape parameters β obtained from the encoding of the current image with the shape parameters β' from another image of the same individual, the new parameter set should still yield a good reconstruction:

$$\mathcal{L}_{sc} = \mathcal{L}_{coarse}(I, \beta', \psi, \theta, l). \quad (9)$$

Pretraining Phase. Before optimizing the coarse reconstruction branch of our base model, we conduct pretraining exclusively with landmark losses and regularization losses. This step ensures the initial alignment of the reconstructed face with the face in the image, as introducing additional losses before achieving this alignment would be ineffective. The loss functions employed during the pretraining phase are as follows:

$$\mathcal{L}_{pretraining} = \mathcal{L}_{inL} + \mathcal{L}_{spL} + \mathcal{L}_{reg}, \quad (10)$$

6.2. Detail Reconstruction Losses

Photometric Loss. Augmented by the displacement map, the detail photometric loss is calculated as:

$$\mathcal{L}_{phoD} = \mathcal{L}_{phoD}(I, I'_r) = \|R_I \odot (I - I'_r)\|_{1,1}, \quad (11)$$

where I'_r is rendered following the details differential rendering approach, and R_I is a mask of the face skin region.

D-MRF Loss. In accordance with DECA [5], we adopt an Implicit Diversified Markov Random Field (ID-MRF) loss [20] in our detail estimation, which extracts feature patches from different layers of a pre-trained network and minimizes the difference between corresponding nearest neighbor feature patches in the input image and detail rendering. This approach encourages the capture of high-frequency geometric details, making it superior to L1 losses in detail recovery. the loss is computed on layers *conv3.2* and *conv4.2* of VGG19 [15]:

$$\mathcal{L}_{mrf} = 2L_M(\text{conv4.2}) + L_M(\text{conv3.2}), \quad (12)$$

where $L_M(\cdot)$ denotes the ID-MRF loss that is applied to the feature patches extracted from I and I'_r .

Smoothness Loss. The smoothness loss serves to prevent overly sharp or high-frequency artifacts in the reconstructed details. During training, we apply regularization to both $D_{UNet} = \mathcal{D}_{UNet}(I)$ and \mathcal{H}'_{UV} :

$$\mathcal{L}_{smo} = \|\nabla D_{UNet}\|_{1,1}. \quad (13)$$

Soft Symmetry Loss. The soft symmetry loss is employed to regularize regions on the face outside the facial skin region R_I , to enhance the model’s robustness in occlusion regions and reduce boundary artifacts in obscured areas.

$$\mathcal{L}_{sym} = \|R_I \odot (D - \text{flip}(D))\|_{1,1}, \quad (14)$$

where *flip* is to *flip* the UV displacement map horizontally.

Detail Consistency Loss. Inspired by DECA [5], in each mini-batch, images share the same identity. Given two images I and I' , we can obtain reconstructed coarse FLAME geometries S and S' respectively, which allows us to calculate the corresponding tension maps T_{UV} and T'_{UV} . When reconstructing facial details using I and T'_{UV} , the resulting facial details should be consistent with image I' and the pseudo ground truth \mathcal{D}'_{Unet} from the teacher network. Therefore, the detail consistency loss is defined as:

$$\mathcal{L}_{dc} = \mathcal{L}_{animD}(I', \mathcal{D}'_{Unet}(I), \mathcal{D}(I, T'_{UV})). \quad (15)$$

7. More Implementation Details

Dataset Details. BUPT-Balancedface [19] comprises 1.3 million images of 28,000 celebrities, distributed fairly across different racial groups, with approximately 7,000 identities per race. Celeb-DF (v2) [11] includes 590 celebrity videos and 300 additional videos collected from

YouTube with subjects of different ages, ethnic groups and genders, and 5639 corresponding DeepFake videos. As we aim to learn the facial details that change with facial deformation in real human faces, we exclusively utilize the real videos. MEAD [17] is a video dataset of talking faces, featuring 60 actors and actresses who express 8 distinct emotions at 3 varying intensity levels.

Data Augmentation. For data augmentation, we randomly sample a scaling factor κ from a normal distribution characterized by a mean of 1 and a standard deviation of 0.3, constrained within the range of 1 to 1.2. Furthermore, we independently sample two bias coefficients η_x and η_y from a normal distribution with a mean of 0 and a standard deviation of 0.125, restricted to the interval of -0.1 to 0.1. Centering the image at an offset of $((\eta_x - \eta_x/\kappa) \times 100\%$ and $(\eta_y - \eta_y/\kappa) \times 100\%)$ along the x and y directions respectively, with 256κ as the side length of a square box, we extract the region within this square. This extracted portion is then resized again to 256×256 , which serves as the final input to the model.

Network Details. For coarse reconstruction, we use a Vision Transformer (ViT)[4] with a patch size of 16, an embedding dimension of 512, 8 attention heads, and a depth of 8. For detail reconstruction, the teacher model is a UNet consisting of one encoder block, one decoder block, and one bottleneck block that connects them. The encoder block contains two convolutional layers, the bottleneck block has one convolutional layer and one transposed convolutional layer, and the decoder block includes four convolutional layers. The student model is a masked autoencoder (MAE)[8] with a zero masking ratio, ultimately used for reconstructing animatable details. The encoder of the MAE has a patch size of 16, an embedding dimension of 768, 12 attention heads, and a depth of 12, while the decoder has an embedding dimension of 512, 16 attention heads, and a depth of 8. During person-specific transfer, the adapters are MLPs with a single hidden layer. The adapters placed in the ViT have a hidden layer size that is 1/32 of the input layer size, while those in the MAE have a hidden layer size that is 1/2 of the input layer size. This design allows for better capturing of person-specific detail features.

Loss Balancing Weights. For the coarse reconstruction, the total loss function with balancing weights is defined as:

$$\mathcal{L}_C = \lambda_{sil}\mathcal{L}_{sil} + \lambda_{inL}\mathcal{L}_{inL} + \lambda_{spL}\mathcal{L}_{spL} + \lambda_{pho}\mathcal{L}_{pho} + \mathcal{L}_{per} + \lambda_{reg}\mathcal{L}_{reg} + \mathcal{L}_{sc}, \quad (16)$$

where $\lambda_{sil} = 0.5$, $\lambda_{inL} = 0.5$, $\lambda_{spL} = 1$, $\lambda_{pho} = 2$, $\lambda_{reg} = 1e - 5$. The perceptual loss, \mathcal{L}_{per} , is given by:

$$\mathcal{L}_{per} = \lambda_{id}\mathcal{L}_{id} + \lambda_{emo}\mathcal{L}_{emo} + \lambda_{lr}\mathcal{L}_{lr}, \quad (17)$$

where $\lambda_{id} = 0.1$, $\lambda_{emo} = 1$, and $\lambda_{lr} = 0.05$. For the detail reconstruction, the total loss function with balancing



Figure 6. **Limitation when the face has tattoos.** Our transferred person-specific model (right) mistakenly interprets tattoos as intrinsic facial features during reconstruction. Our base model (middle) is also slightly affected, resulting in a depression in the cheek area.

weights is:

$$\mathcal{L}_{animD} = \mathcal{L}_{UNet} + \lambda_{Tchr}\mathcal{L}_{Tchr} + \lambda_{sym}\mathcal{L}_{sym} + \mathcal{L}_{dc}, \quad (18)$$

where $\lambda_{Tchr} = 4$, $\lambda_{sym} = 1e - 2$. \mathcal{L}_{UNet} includes the losses used for training the UNet teacher model:

$$\mathcal{L}_{UNet} = \lambda_{phoD}\mathcal{L}_{phoD} + \lambda_{mrf}\mathcal{L}_{mrf} + \lambda_{smo}\mathcal{L}_{smo} + \lambda_{regD}\mathcal{L}_{regD}, \quad (19)$$

where $\lambda_{phoD} = 1$, $\lambda_{mrf} = 5e - 3$, $\lambda_{smo} = 1e6$, and $\lambda_{regD} = 1e2$.

8. Limitations and Future Work

Facial Tattoos. The pre-trained facial skin region segmentation method [2] we utilize is incapable of excluding tattoos from the facial region. Consequently, our model erroneously learns them as part of the facial shape. Fig. 6 demonstrates that our base model and the transferred person-specific model mistakenly recognize tattoos as part of facial wrinkles, which is the inherent limitation of shape-from-shading. Utilizing more advanced segmentation models can exclude these interfering elements during training. Another available solution is to preprocess the images using facial tattoo removal networks [9] before performing face reconstruction.

Perspective Projection. Similar to recent learning-based generic models [1, 3, 5, 21], we employ a weak perspective camera model (or orthographic camera model). Nonetheless, in scenarios where the face is captured at a close distance (such as in smartphone selfies) or with a wide-angle lens, perspective distortion is not modeled. This oversight leads to the estimated face conforming directly to the distorted image, resulting in discrepancies between the estimated and the actual 3D face. Since the majority of facial images in large-scale face datasets for self-supervised training are either taken from a distance or cropped from larger images, the perspective effects are minimal, posing a challenge for the automatic regression of camera intrinsic parameters from these data (cf. [25]).

Future Work. Moving forward, we plan to train our model on the VFHQ dataset [23], which contains over

16,000 high-fidelity clips of interview scenarios. This will enable us to capture clearer and higher-resolution wrinkle details. Additionally, the face geometry estimated by our model, with its accurately aligned outer boundaries, serves as an ideal initial position for 3D Gaussian primitives, while the 3D Gaussian splatting technique provides our model with higher-fidelity rendering capabilities.

References

- [1] Zenghao Chai, Tianke Zhang, Tianyu He, Xu Tan, Tadas Baltrusaitis, HsiangTao Wu, Runnan Li, Sheng Zhao, Chun Yuan, and Jiang Bian. Hiface: High-fidelity 3d face reconstruction by learning static and dynamic details. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9087–9098, 2023. 5
- [2] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*, pages 1–4. IEEE, 2017. 5
- [3] Radek Daněček, Michael J Black, and Timo Bolkart. EMOCA: emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 1, 2, 3, 4, 5
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5
- [5] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4): 88:1–88:13, 2021. 1, 2, 3, 4, 5
- [6] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos. *arXiv preprint arXiv:2207.11094*, 2022. 3, 4
- [7] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision*, pages 152–168. Springer, 2020. 2
- [8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 5
- [9] Mathias Ibsen, Christian Rathgeb, Pawel Drozdowski, and Christoph Busch. Face beneath the ink: Synthetic data and tattoo removal with application to face recognition. *Applied Sciences*, 12(24):12969, 2022. 5
- [10] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2
- [11] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deep-fake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020. 4
- [12] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 1
- [13] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: database and results. *Image and Vision Computing*, 47:3–18, 2016. 1, 2
- [14] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossai, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 50–58, 2015. 2
- [15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4
- [16] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 2
- [17] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Proceedings of the European Conference on Computer Vision*, pages 700–717. Springer, 2020. 5
- [18] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20333–20342, 2022. 1
- [19] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 692–702, 2019. 4
- [20] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. *Advances in neural information processing systems*, 31, 2018. 4
- [21] Yandong Wen, Weiyang Liu, Bhiksha Raj, and Rita Singh. Self-supervised 3d face reconstruction via conditional estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13269–13278, 2021. 5
- [22] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *International Conference on 3D Vision (3DV)*, pages 453–463. IEEE, 2021. 2
- [23] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and bench-

- mark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. [5](#)
- [24] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 598–607, 2020. [1](#), [2](#)
- [25] Yajie Zhao, Zeng Huang, Tianye Li, Weikai Chen, Chloe LeGendre, Xinglei Ren, Ari Shapiro, and Hao Li. Learning perspective undistortion of portraits. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7849–7859, 2019. [5](#)