HiFace: High-Fidelity 3D Face Reconstruction by Learning Static and Dynamic Details
——————————Appendix——————————

In this supplementary, we provide additional results and discussions to make our paper self-contained. Specifically, we present 1). more details on experiments and implementation in Sec. A, 2). details of loss functions in Sec. B, 3). more experimental comparisons and results in Sec. C, 4). discussions on limitations and future work in Sec. D, and 5). discussions on social impact in Sec. E.

A. Experimental & Implementation Details

A.1. Details of Dataset

We follow the synthetic data pipeline [13, 17] to synthesize $200k$ images, consisting of $40k$ identities with 5 frames. For each identity, we assign different expressions, viewpoints, illumination, accessories, and backgrounds to improve the model's robustness for training.

In addition, in the synthetic data pipeline, the ground-truth albedo, neutral displacement maps, stretched displacement maps, and compressed displacement maps are sampled from the 332 captured scans [13, 17] and recorded. In this paper, to ensure the efficiency of our model, we resize the 4096×4096 resolution assets into 512×512 as the ground-truth labels for training. For the real-world data [11, 12], we split them into the train ($400k$) and valid ($30k$). During training, common data augmentation techniques (*i.e.*, random shift, scale, rotation, and flip) are adopted to improve the robustness of our model.

A.2. Details of Ablation Studies

In the ablation on datasets, we only remove the real-world data while keeping other settings the same as our full HiFace. In the ablation on loss functions, we only remove specific loss functions while training the models that follow the same settings as our full HiFace.

For the ablation studies on SD-DeTail, in SD-1, instead of obtaining the dynamic detail by interpolating the compressed and stretched displacement maps as in SD-DeTail, we directly synthesize the dynamic detail by a learnable network end-to-end. The ground-truth labels for the synthetic dataset follow the dynamic composition in Eq. 9. More specifically, we use an MLP layer to map expression-aware $\boldsymbol{\phi}$ into 128-dim latent code $\mathbf{z_1}$, and follow [6] to use a U-Net decoder [14] to synthesize corresponding dynamic displacement map in 512×512 resolution.

In SD-2, instead of generating compressed and stretched displacement maps using the PCA bases as in SD-DeTail, we employ two learnable networks to synthesize the compressed and stretched displacement maps in parallel, and compound the dynamic details by interpolating the two displacement maps using Eq. 9. Similar to SD-1, we use two MLP layers to map $\boldsymbol{\phi}$ into 128-dim latent codes $\mathbf{z_2}$ and $\mathbf{z_2'}$, and use two U-Net decoders to synthesize corresponding polarized displacement maps in 512×512 resolution.

In SD-3, instead of generating the static displacement map using the PCA basis as in SD-DeTail, we directly synthesize the static detail by a learnable network end-to-end. More specifically, we use an MLP layer to map age-aware $\boldsymbol{\varphi}$ into 128-dim latent code $\mathbf{z_3}$, and follow [6] to use a U-Net decoder [14] to synthesize corresponding static displacement map in 512×512 resolution.

B. Details of Loss functions

In our main paper, we propose several simple yet effective loss functions to train our model end-to-end using both synthetic and real-world data. As we have justified through the detailed ablation studies, each loss function contributes to the coarse shape and details. In this section, we introduce details about self-supervised losses $\mathcal{L}_{\text{self}}$ and knowledge distillation $\mathcal{L}_{\text{kd}}$.

B.1. Details of Self-Supervised Losses

Following [6, 5], we leverage the differentiable renderer [7] to obtain the rendered image $\hat{\mathbf{I}}^r$, then use photo loss $\mathcal{L}_{\text{pho}}$ and identity loss $\mathcal{L}_{\text{id}}$ to compute the error between the input image $\mathbf{I}$ and $\hat{\mathbf{I}}^r$. We also follow [18] to use dense landmark loss $\mathcal{L}_{\text{lmk}}$ to calculate the error between the detected landmarks from $\mathbf{I}$ and projected landmarks from $\hat{\mathbf{S}}$, as self-supervised loss is still crucial to ensure satisfactory generalization to real-world images.

Photo loss $\mathcal{L}_{\text{pho}}$ computes the l_2 error between $\mathbf{I}$ and $\hat{\mathbf{I}}^r$:

$$\mathcal{L}_{\text{pho}} = \left\| \mathbf{M_I} \odot \left(\mathbf{I} - \hat{\mathbf{I}}^r \right) \right\|_2, \tag{A16}$$

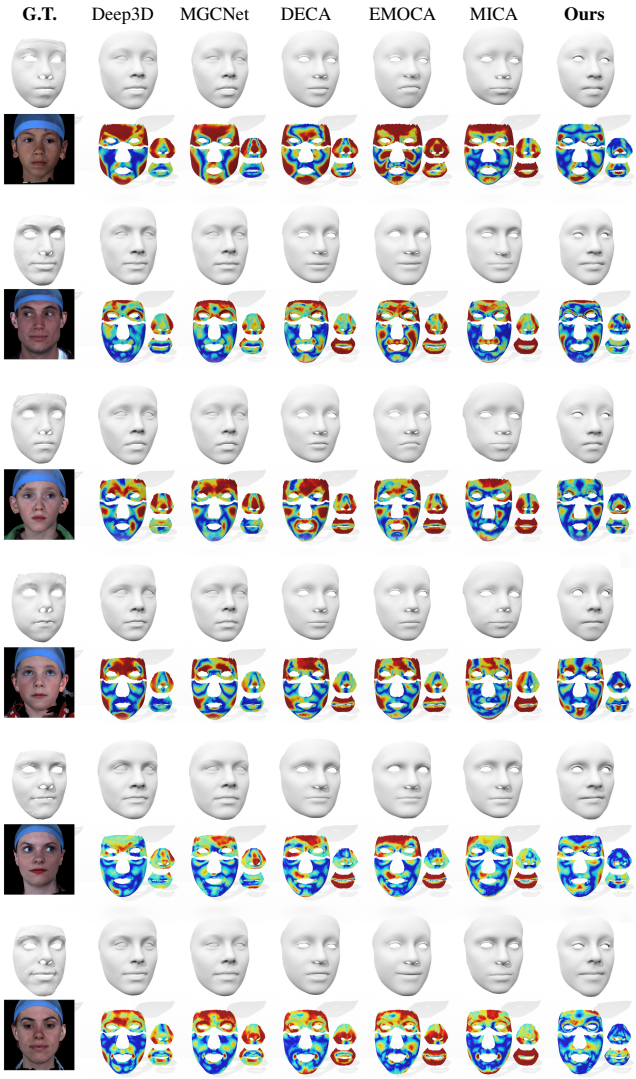

Figure A10. **Error map on REALY benchmark.** We visualize and compare the reconstruction error of HiFace to previous methods. From left to right: Input image & ground-truth, Deep3D [5], MGCNet [15], DECA [6], EMOCA [3], MICA [23]), and HiFace (Ours), where large (small) errors are colored in red (blue). The proposed method presents the best reconstruction quality.

where $\mathbf{M_I}$ is the region-of-interest mask [22] of image $\mathbf{I}$, which only considers facial skins and removes occlusions.

Identity loss $\mathcal{L}_{\text{id}}$ leverages the pretrained face recognition network $\mathbf{\Gamma}$ [4] to estimate the cosine similarity between high-level features from $\mathbf{I}$ and $\hat{\mathbf{I}}^r$:

$$\mathcal{L}_{\text{id}} = \frac{\mathbf{\Gamma}(\mathbf{I}) \cdot \mathbf{\Gamma}(\hat{\mathbf{I}}^r)}{\|\mathbf{\Gamma}(\mathbf{I})\|_2 \cdot \|\mathbf{\Gamma}(\hat{\mathbf{I}}^r)\|_2}, \tag{A17}$$

Dense landmark loss $\mathcal{L}_{\text{lmk}}$ leverages the landmark detector [18] to detect 669 dense landmarks of given 2D images, and estimate the distance between the detected and projected 2D points from the reconstructed shape $\hat{\mathbf{S}}$:

$$\mathcal{L}_{\text{lmk}} = \sum_{i=1}^{669} \frac{\|\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i\|_2}{2\sigma_i^2}, \tag{A18}$$

where $\boldsymbol{\mu}_i$ and σ_i are the coordinates and uncertainty of the i-th detected landmark from $\mathbf{I}$, respectively. $\hat{\boldsymbol{\mu}}_i$ denotes the i-th projected 2D landmark from the reconstructed shape $\hat{\mathbf{S}}$.

B.2. Details of Knowledge Distillation

The age prediction model $\mathbf{\Gamma}_{\text{age}}$ [9] predicts the age of given images into 9 categories: $0-2$, $3-9$, $10-19$, $20-29$, $30-39$, $40-49$, $50-59$, $60-69$, and $70+$. Therefore, we leverage 3-layer MLP to transform the static coefficient φ into a 9-dim vector, and use *softmax* to map into a probability distribution $\hat{\mathbf{p}}_{\text{age}}$.

C. Additional Experimental Results

In this section, we provide additional results to strengthen the superiority of HiFace in reconstructing 3D shapes with animatable details. Specifically, we present 1). error maps on REALY [1] in Fig. A10, 2). additional reconstruction comparisons of coarse shape and details in Fig. A11 and Fig. A12, respectively, 3). additional experiments on flexibility for both images and video sequences in Fig. A13 and Fig. A14, Fig. A15, Fig. A16, respectively, and 4). additional animation comparisons in Fig. A17 and Fig. A18, 5). user studies about the reconstruction and animation quality in Tab. A2 and Tab. A3, respectively.

C.1. Error Maps of REALY Benchmark

In Fig. A10, we present additional comparisons of the error map on the REALY benchmark [1]. The RGB color in ground-truth regions is mapped from the vertex-to-vertex error between the ground-truth and predicted shape according to the evaluation protocol in [1]. Compared to previous methods, HiFace reaches the smallest error and the best reconstruction quality.

C.2. Reconstruction Comparisons

Qualitative Comparisons. In Fig. A11 and Fig. A12, we present additional comparisons of HiFace to previous coarse shape reconstruction methods [5, 8, 19, 23, 18] and detail reconstruction methods [20, 2, 6, 3, 16]. The input images show diversity w.r.t. ethnicity, gender, age, BMI, pose, environment, occlusion, and expression. Compared to previous methods, HiFace is robust to occlusions, extreme poses, and diversity expressions. HiFace reconstructs realistic coarse shapes, better expressions, and realistic details.

User Study. To demonstrate that our reconstructed 3D faces present visually better results and are faithfully aligned

Figure A11. **Comparison on coarse shape reconstruction.** From left to right: Input image, Deep3D [5], 3DDFA-v2 [8], SynergyNet [19], MICA [23], Dense [18], and HiFace (Ours). Note that MICA focuses on identity reconstruction, lacking the consideration of expression.

with human perception, we present a user study by inviting 77 volunteers with a computer science background to vote for the best-reconstructed shapes, from sampled faces in CelebA [11], FFHQ [10], and AFLW2000 [21]. Specif-

Figure A12. **Comparison on detail shape reconstruction.** From left to right: Input image, FaceScape [20], Unsup [2], DECA [6], EMOCA [3], FaceVerse [16], and HiFace (Ours). "⊠" indicates this method fails to return any reconstruction.

ically, we separately compare methods for coarse shape reconstruction [5, 8, 19, 23, 18], and detailed reconstruction [20, 2, 6, 3, 16]. The results are summarized in Tab. A2.

Tab. A2 shows that more than half of users perceive

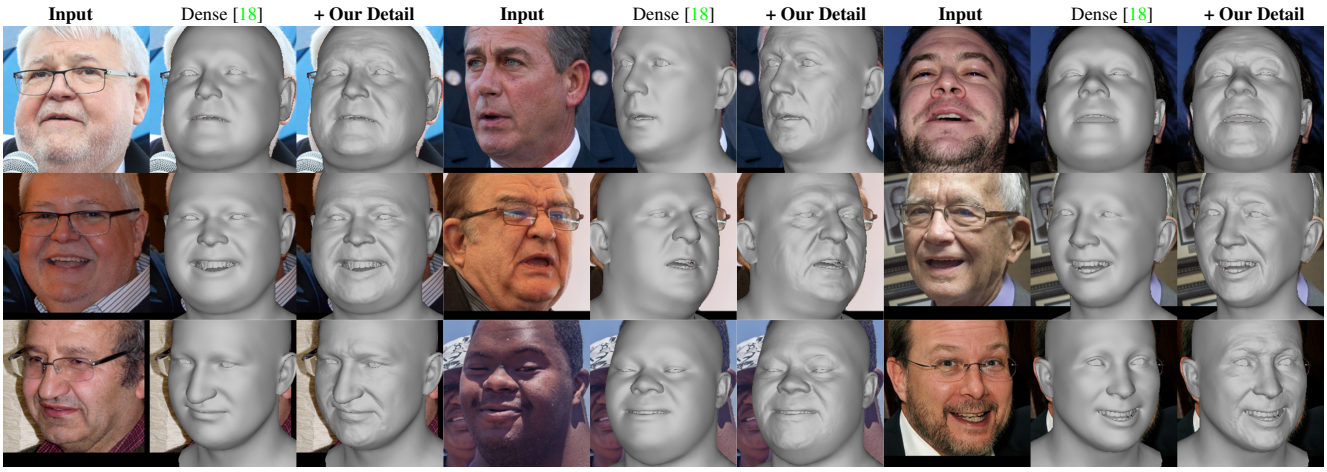

Figure A13. **Illustration on the flexibility of SD-DeTail.** Given the identity and expression coefficients (β, ξ) from the optimization-based method [18], SD-DeTail can generate realistic details based on the coarse shape and further improve the visual quality.

Figure A14. **Illustration on the flexibility of SD-DeTail on video reconstruction (part 1).** We visualized the reconstruction quality of Dense [18] with/without our SD-DeTail and compare them with prior art [6, 3]. Videos are taken from YouTube.

that our reconstructed shapes are more similar to the given images, and 80.52% users vote that HiFace reconstructs more realistic details compared to others. As a comparison, the second-best detail reconstruction method [6] has

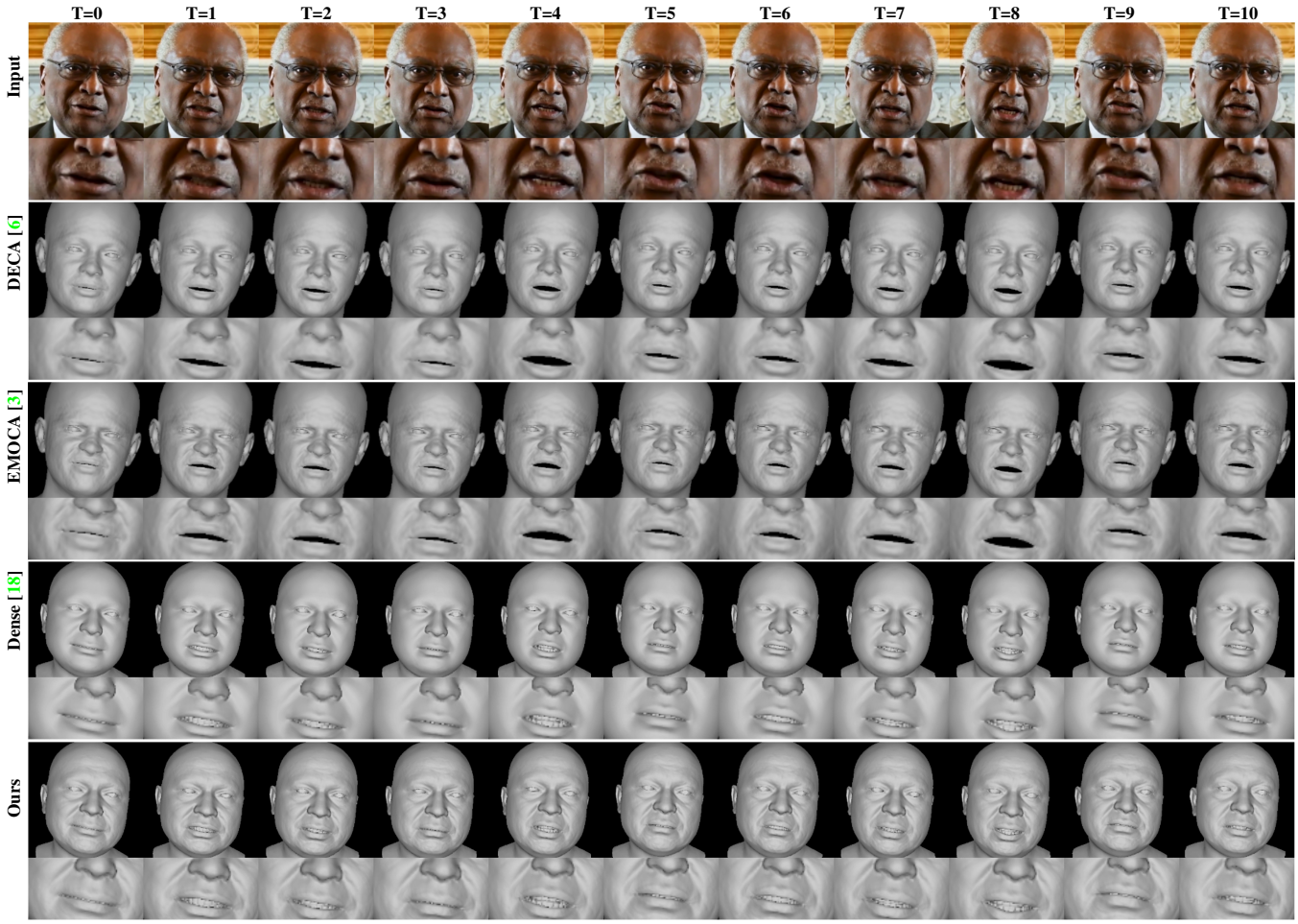

Figure A15. **Illustration on the flexibility of SD-DeTail on video reconstruction (part 2).** We visualized the reconstruction quality of Dense [18] with/without our SD-DeTail and compare them with prior art [6, 3]. Videos are taken from YouTube.

Table A2. **User study results on the reconstructed shape and details.** HiFace achieves the best results in coarse shape and details according to human perception compared to prior art [6, 18].

Group	Best method	2nd best method	Other methods
Coarse	**54.55%** (Ours)	35.06% (Dense [18])	10.39% ([23, 5, 19, 8])
Detail	**80.52%** (Ours)	11.69% (DECA [6])	7.79% ([3, 16, 2, 20])

only 11.69% votes. It demonstrates the superiority of our methods in reconstructing coarse shapes and details.

C.3. Flexibility of SD-DeTail

We introduce the settings of plugging SD-DeTail into optimization-based methods such as Dense [18]. Specifically, for a given image, we leverage optimization-based methods to regress the identity coefficient β and expression coefficient ξ, and use our feature extractor in HiFace to regress the static coefficient φ. Then these three coefficients serve as the input of SD-DeTail to synthesize the realistic displacement maps. Finally, we integrate the coarse shape from the optimization-based methods with the synthesized displacement map to obtain the final detailed shapes. In Fig. A13, we present additional results to justify the flexibility of SD-DeTail. SD-DeTail can be easily plugged into previous optimization-based methods and introduces realistic details on the coarse shapes to improve the visualized quality.

In addition, we also compare the reconstruction quality in the video sequences and further demonstrate that, given the coarse shape obtained by existing optimization-based methods, the proposed SD-DeTail has the advantage of achieving realistic detailed results based on the existing coarse shape. As Fig. A14-A16 show, we perform reconstruction on the video sequences by comparing Dense [18] with/without our SD-DeTail to prior art [6, 3]. Fig. A14-A16 demonstrate that our SD-DeTail significantly improve the visualized quality compared to the coarse shape from [18]. Compared to prior art [6, 3], our SD-DeTail captures subtle and realistic details and outperforms previous

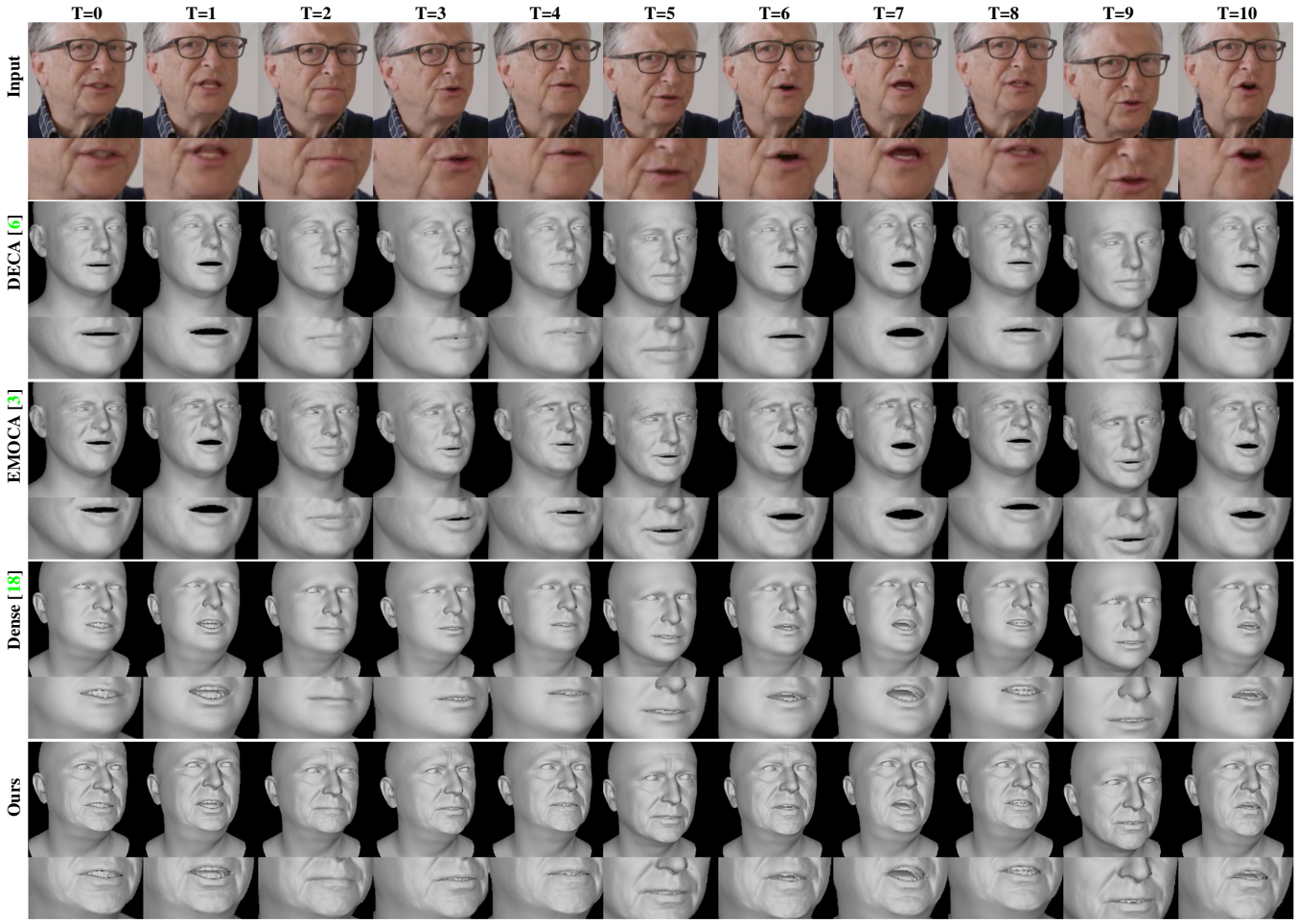

Figure A16. **Illustration on the flexibility of SD-DeTail on video reconstruction (part 3).** We visualized the reconstruction quality of Dense [18] with/without our SD-DeTail and compare them with prior art [6, 3]. Videos are taken from YouTube.

Table A3. **User study results on detail and expression transfer.** We group the driving images into young people (A) and elder people (B), and ask artists to score the animation quality of static details and dynamic expressions of the transferred images, corresponding to the 1st and 2nd row of each source image in Fig. A17 and Fig. A18. We report the average scores, median scores, and standard deviation. "5" indicates the best score while "1" indicates the worst score.

Driving Group		A (1-5)						B (6-10)						All					
Source	Method	Static			Dynamic			Static			Dynamic			Static			Dynamic		
		avg.	med.	std.	avg.	med.	std.	avg.	med.	std.	avg.	med.	std.	avg.	med.	std.	avg.	med.	std.
I	DECA [6]	1.63	1.50	0.72	2.50	2.50	1.05	3.27	3.00	0.65	2.73	2.50	0.80	2.45	2.50	1.07	2.62	2.50	0.93
	EMOCA [3]	1.93	2.00	0.82	1.97	2.00	0.92	3.20	3.00	0.62	2.40	2.50	0.74	2.57	3.00	0.96	2.18	2.00	0.85
	Ours	**4.40**	**4.50**	0.60	**4.63**	**5.00**	0.44	**4.13**	**4.00**	0.35	**4.53**	**4.50**	0.44	**4.27**	**4.00**	0.50	**4.58**	**4.50**	0.44
II	DECA [6]	1.70	2.00	0.62	2.47	2.00	1.08	3.37	3.00	0.52	3.10	3.00	0.78	2.53	2.75	1.02	2.78	3.00	0.98
	EMOCA [3]	2.07	2.00	0.73	2.37	2.00	1.03	3.73	4.00	0.68	2.07	2.00	0.84	2.90	3.00	1.09	2.22	2.00	0.90
	Ours	**4.17**	**4.00**	0.49	**4.80**	**5.00**	0.41	**4.30**	**4.00**	0.59	**4.47**	**4.00**	0.40	**4.23**	**4.00**	0.54	**4.63**	**5.00**	0.43

methods by a large margin.

C.4. Detail Animation Comparisons

While previous state-of-the-art methods [6, 3] directly concatenate the person-specific identity features with expression-aware features and decode them into a displacement map. In our ablation studies, we have demonstrated that simply predicting the dynamic details is rather challenging to achieve satisfactory results (see Fig. 9). Here we present additional comparisons on detail animation to justify our claims.

Qualitative Comparisons. In Fig. A17 and Fig. A18, we

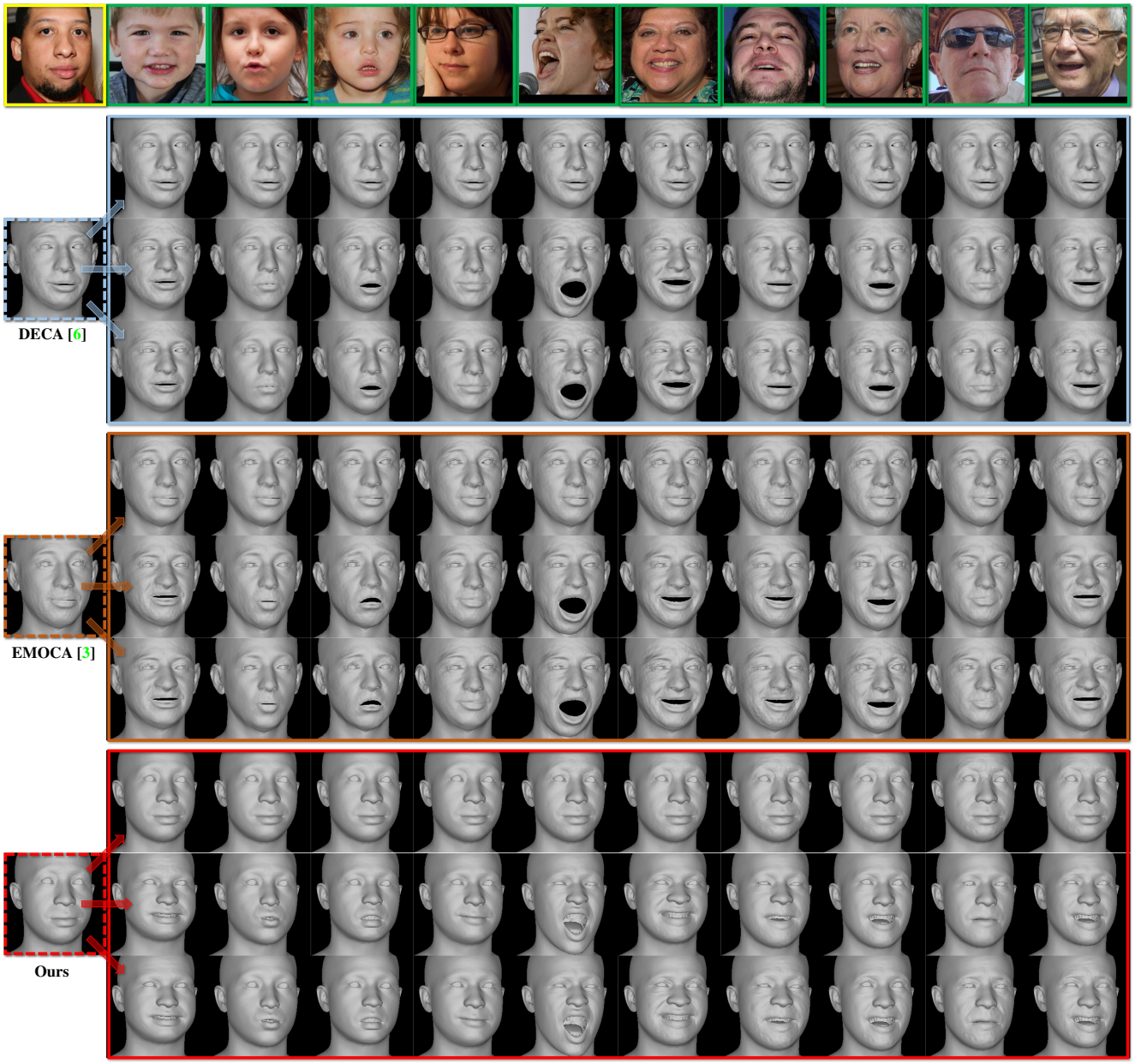

Figure A17. **Comparison on face animation (part 1).** Given a source image (yellow box), we use the driving images (green box) to drive its person-specific details and expressions. For each method, we manipulate the static (1st-row), dynamic (2nd-row), or both (3rd-row) factors. However, DECA [6] (blue box) and EMOCA [3] (orange box) can animate the expression-driven details but lack realistic, and cannot transfer the static details from the driving images well. As a comparison, HiFace (red box) is flexible to animate details from static, dynamic, or both factors, and presents vivid animation quality with realistic shapes.

manipulate the static and/or dynamic details of the source image by assigning the static and/or dynamic codes from the driving image. Specifically, the static coefficient in HiFace (or called "detail code" in DECA [6] and EMOCA [3]) is encoded from the driving image. The dynamic factor is based on the expression coefficient in our HiFace, while in DECA [6] and EMOCA [3], it is based on the "expression parameter" and "jaw pose". We present the comparisons of the animation results in Fig. A17 and Fig. A18.

We can see that the details are not well decoupled in previous methods. For example, when we manipulate the static factor to the source image and the driving image is a young child (*e.g.*, second column), the output shape should have presented "young" details. However, results in previous methods still exhibit "noisy" details, which correspond to "old" wrinkles. It demonstrates that such implicit learning

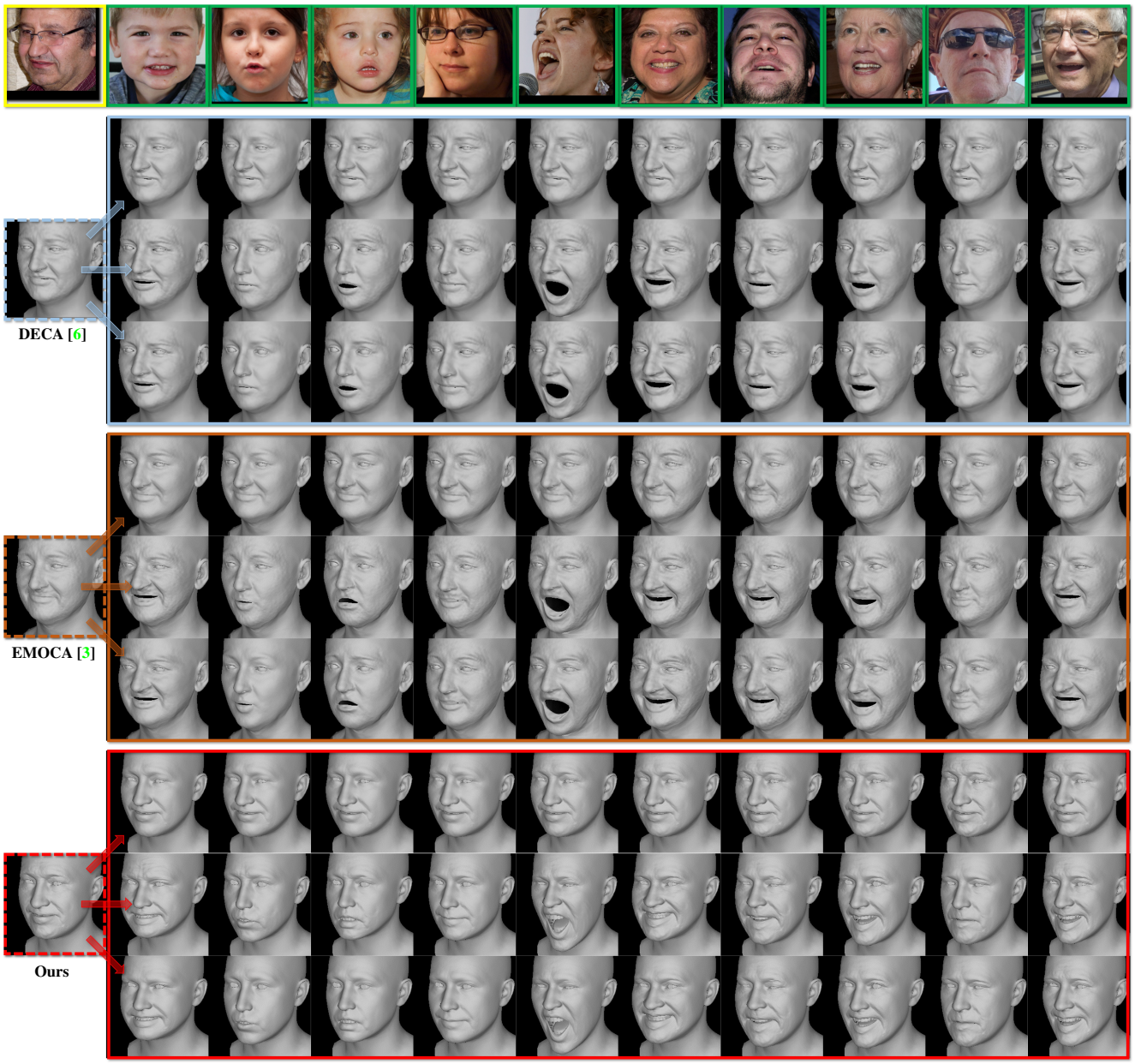

Figure A18. **Comparison on face animation (part 2).** Given a source image (yellow box), we use the driving images (green box) to drive its person-specific details and expressions. For each method, we manipulate the static (1st-row), dynamic (2nd-row), or both (3rd-row) factors. However, DECA [6] (blue box) and EMOCA [3] (orange box) can animate the expression-driven details but lack realistic, and cannot transfer the static details from the driving images well. As a comparison, HiFace (red box) is flexible to animate details from static, dynamic, or both factors, and presents vivid animation quality with realistic shapes.

is hard to decouple the static and dynamic factors well.

As a comparison, our novelty and insights lie in the essence of 3DMMs that simplify the 2D-to-3D difficulty by statistical models. We successfully reconstruct plausible static and dynamic details by simplifying such difficulty into feasible regression and interpolation tasks. The details generated by HiFace are naturally decoupled into static and dynamic factors for animation. For example, if we animate the static factor, the facial details present variation among different age groups, and when we animate the dynamic factor, the expression-driven details are well transferred (see "Ours" in Fig. A17 and Fig. A18 from left to right).

User Study. We also present another user study to investigate the objective evaluation from 5 experienced artists in estimating the expression and detail transfer quality.

More specifically, given the source images in Fig. A17 and Fig. A18 (noted as subject I and subject II in Tab. A3), we ask the artists to mark scores ranging from 1 to 5 (5 indicates the best score) for each driving sample. The scores are evaluated based on the animation quality w.r.t. the static details and the dynamic expressions from the driving images. The driving images are classified into: A. the young group (1-5 images) and B. the elder group (6-10 images). The quantitative results are summarized in Tab. A3.

According to Tab. A3, we demonstrate that our reconstruction and animation results are better aligned with human perception. For the static factor, we can see the results of previous methods [6, 3] present an imbalanced distribution, *i.e.*, when the driving images are young people's, the score is, in general, worse than that of the elder group. As for the expression transfer, previous methods reach worse scores when the driving images contain extreme expressions (corresponding to a larger standard deviation in Tab. A3). As a comparison, our method presents higher scores with smaller variances among different age groups, which demonstrates the power of our model in transferring novel expressions and details.

D. Limitations & Future Work

This paper proposes a novel approach to reconstructing animatable details from monocular images. While we manage to synthesize realistic details and demonstrate higher accuracy compared to previous state-of-the-art, our work still has limitations. We pinpoint these challenges in the 3D face community and leave them for future work.

Facial Appearance Model. We use the vanilla albedo 3DMM to linearly represent the facial appearance. While we focus more on the geometry shape, such albedo inherently lacks details and indirectly influences the training of HiFace. In the future, we plan to integrate the diffuse model and spectral model to present high-fidelity facial appearance and extend our HiFace with photo-realistic textures.

Reconstruction Quality. While we achieve state-of-the-art reconstruction quality in the REALY [12] benchmark in terms of the overall quality in Tab. 1. We notice HiFace does not perform the best in the mouth and cheek, which are highly emotional and structural regions. To address this problem, we leave it for future work to incorporate the emotion-aware perceptual loss and structure-aware constraints in these regions to further improve the reconstruction quality of HiFace.

Displacement Prior. We demonstrate the necessity to leverage the statistical model to constrain the displacement distribution. However, due to the high expense of capturing large-scale and high-quality displacements for training a non-linear model. We choose the common practice of linear PCA model to build displacement bases for it is easy to implement and data amount friendly. We trade off the learning difficulty and personalized details (*e.g.*, nevus) through the statistical model. Therefore, it is still challenging to recover pore-level details. In addition, we notice that the imbalanced data (*i.e.*, a majority of young scans and images with fewer children and elders) also influence the representation of our model. In the future, we plan to capture and synthesize more class-balanced data to train a non-linear model and leverage the versatile generative models to reconstruct high-resolution displacement maps for more realistic 3D faces.

Evaluation on Facial Animation. We present the visualized comparisons in transferring facial expressions and their details. However, we can only refer to quantitative comparisons for missing such a benchmark to estimate the transfer accuracy. We leave it for future work to construct a benchmark to evaluate the quality of expression transfer.

E. Potential Social Impact

While this paper successfully reconstructs 3D shapes with animatable details from the monocular images, it is not intended to create content that is used to mislead or deceive. Therefore, this paper does not raise disinformation or immediate security concerns. However, like other related 3D face reconstruction and animation techniques, it could still potentially be misused for impersonating humans. We condemn any behavior to create misleading or harmful content of real persons. We encourage researchers in the 3D face community to consider the questions about preventing privacy disclosure before applying the model to the real world.

References

[1] Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 14

[2] Yajing Chen, Fanzi Wu, Zeyu Wang, Yibing Song, Yonggen Ling, and Linchao Bao. Self-supervised learning of detailed 3d face reconstruction. *IEEE Transactions on Image Processing*, 29:8696–8705, 2020. 14, 16, 18

[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. 14, 16, 17, 18, 19, 20, 21, 22

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 14

[5] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with

weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 13, 14, 15, 16, 18

[6] 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):1–13, 2021. 13, 14, 16, 17, 18, 19, 20, 21, 22

[7] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018. 13

[8] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020. 14, 15, 16, 18

[9] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 14

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 15

[11] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 13, 15

[12] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 13

[13] Chirag Raman, Charlie Hewitt, Erroll Wood, and Tadas Baltrusaitis. Mesh-tension driven expression-based wrinkles for synthetic faces. In *WACV*, 2023. 13

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 13

[15] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *European Conference on Computer Vision*, pages 53–70. Springer, 2020. 14

[16] 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. 14, 16, 18

[17] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 13

[18] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, pages 160–177. Springer, 2022. 13, 14, 15, 16, 17, 18, 19

[19] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *2021 International Conference on 3D Vision (3DV)*, 2021. 14, 15, 16, 18

[20] 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 601–610, 2020. 14, 16, 18

[21] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3990–3999, 2017. 15

[22] Qi Zheng, Jiankang Deng, Zheng Zhu, Ying Li, and Stefanos Zafeiriou. Decoupled multi-task learning with cyclical self-regulation for face parsing. In *Computer Vision and Pattern Recognition*, 2022. 14

[23] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*. Springer International Publishing, Oct. 2022. 14, 15, 16, 18