FaceLift: Semi-supervised 3D Facial Landmark Localization - Supplementary Material

David Ferman Pablo Garrido Gaurav Bharaj
Flawless AI

A. Implementation Details

3D Landmark Transformer Architecture Fig. 1 presents a more detailed figure of our 3D landmark transformer architecture. 3D head pose and facial landmarks are estimated via cross-attention and self-attention heads and MLP layers.

Loss Implementation Details In order to compute occlusion-aware masks, $m \in \{0,1\}^N$, used in Eq. 4, we apply the predicted rotation matrix to a template of normal vectors for each landmark, and threshold the dot product with the forward vector to obtain the mask. We obtain our normal template by selecting the landmarks on a face mesh, and computing the normals at those locations. We set the threshold so that products above 0.5 were considered visible, while lowering this threshold to -0.1 for the nose bridge. We found this conservative masking strategy reasonable in our experiments.

Multi-view Camera Optimization In obtaining 3D pseudo-labels for 3D-aware GAN-generated samples, we perform a multi-view 3D landmark optimization over detections from renders of camera views, $\tilde{c}_i \in \tilde{C}$, represented by (α, β) azimuth and elevation pairs. Fig. 2 illustrates all these $|\tilde{C}| = 41$ sample views.

B. Evaluation Set Preparation

When comparing our model on the DAD3D-Heads [3] dataset, we upsample the meshes to ensure that the mesh is dense enough that the distance between vertices is much smaller than the model's inconsistencies.

Due to the enormous size of the Multiface [6] dataset, we sample a subset for our evaluations. We selected 6 sequences: *Neutral Eyes Open, Relaxed Mouth Open, Open Lips Mouth Stretch Nose Wrinkled, Mouth Nose Left, Mouth Open Jaw Right Show Teeth, Suck Cheeks In*, which include closed eyes, wide mouth openings, and asymmetric facial deformations. The data covers a wide range of cameras, and we discard several in which the face is not visible, including cameras numbered 400055, 400010, 400067, 400025, 400008, and 400070. To eliminate redundancy in the evaluation set, we sample every 15 frames from the downloaded sequences.

C. Additional Experiments

Pseudo-labels Visualized In Fig. 3, we visualize 3D-aware GAN samples, obtained via 3D pseudo-labeled IDE-3D [4] latent renders, which are sampled from our augmented camera space, $\mathbb{C}$.

Additional Qualitative Results on CelebV-HQ Additional qualitative results on the CelebV-HQ [8] dataset are shown in Fig. 4.

Evaluations on Additional DAD3D-Heads Categories In Tab. 1, we report the DAD3D-Heads [3] evaluation results for additional categories, including image quality, lighting, gender, and age.

Loss Function Ablations We compare the loss function used by our method, Laplacian Log Likelihood, with other common loss functions, L1 and MSE, in Tab. 2. Our choice yields the best results on our benchmark datasets.

Cross-Dataset Evaluations Our investigations into cross-dataset evaluations reveal a notable limitation in model generalizability between datasets with differing labeling conventions. We report cross-dataset evaluations in Tab. 3 on both AFLW2000-3D [9] and the DAD3D-Heads [3] validation set, comparing our method with methods trained on DAD3D-Heads and 300WLP [9], noting that 300WLP's compatible evaluation set is the AFLW2000-3D dataset. We observe that despite a global alignment in how landmarks are defined, cross-dataset scores of every SoTA model are all worse than the SoTA models of the compatible dataset. This is expected due to the local definition bias w.r.t. a different dataset's landmark definition, which yields a consistent error. For each dataset, our model achieves the best cross-dataset score. The cross-dataset metrics do not disentangle the local definition bias from some notion of actual error with respect to the model's landmark definition. Intuitively, if our model's landmark definition were the midway interpolation between the two dataset definitions, our model would incur half of the error from local definition bias than that of the other models. Hence, for fair comparisons, we compare against other methods using our proposed NMLC metric, which removes the local definition bias from the evaluated error. Nevertheless, cross-dataset evaluation remains a useful proxy for assessing

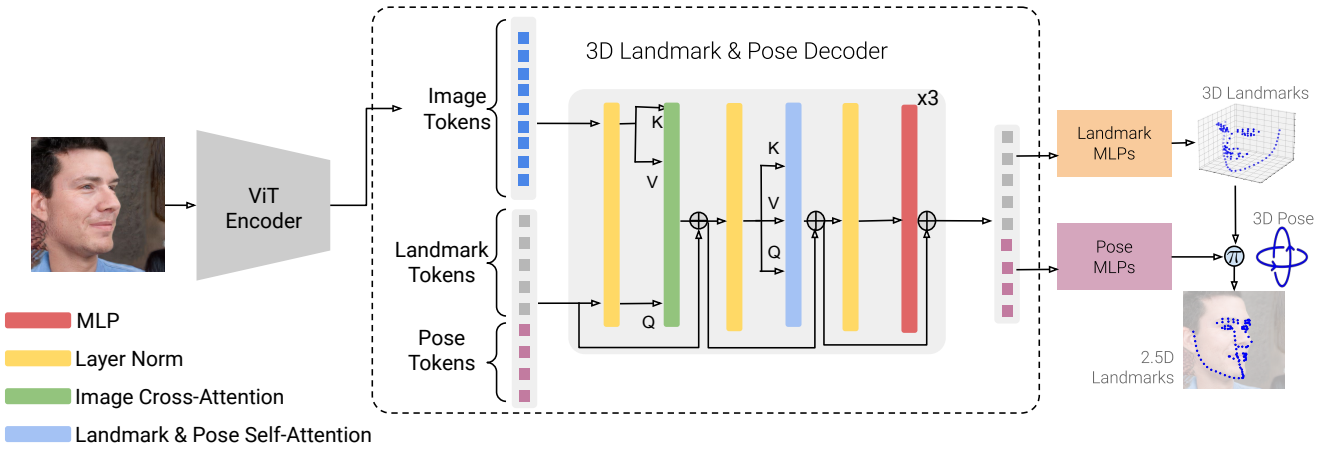

Figure 1. **3D Landmark Transformer Architecture.**

	Quality		Standard Light		Gender			Age			
Model	High	Low	True	False	female	male	undefined	child	young	middle aged	senior
SynergyNet [5]	2.27	3.70	2.55	3.29	2.35	2.78	6.15	2.44	2.74	2.95	2.67
3DDFA [9]	2.80	4.50	2.95	4.34	2.99	3.41	6.71	2.94	3.45	3.53	3.30
3DDFA+ [7]	2.70	4.04	2.88	3.79	2.81	3.21	5.70	2.82	3.22	3.26	3.14
3DDFAv2 [2]	2.13	2.97	2.33	2.67	2.17	2.51	3.77	2.23	2.45	2.49	2.43
DAD-3DNet★ [3]	1.84	2.48	1.99	2.26	1.87	2.15	2.88	1.83	2.05	2.16	1.96
DAD-3DNet+★ [7]	1.84	2.43	1.98	2.21	1.87	2.13	2.75	1.83	2.03	2.13	1.97
FAN3D [1]	1.99	3.22	2.21	2.91	2.08	2.51	4.46	2.02	2.32	2.67	2.36
Ours	1.68	2.28	1.81	2.07	1.72	1.95	2.72	1.70	1.90	1.95	1.85
Ours (Resnet50)	1.92	2.75	2.11	2.45	2.03	2.23	3.55	1.91	2.24	2.29	2.07
Ours (Resnet152)	1.81	2.47	1.96	2.23	1.87	2.08	3.05	1.80	2.07	2.09	1.98
Ours (MF only)	1.80	2.48	1.96	2.24	1.87	2.12	2.78	1.83	2.10	2.08	1.88
Ours (MV only)	2.01	2.85	2.17	2.60	2.10	2.33	3.75	2.03	2.36	2.35	2.21
Ours (100)	1.89	2.50	2.02	2.30	1.93	2.14	3.13	1.92	2.14	2.14	2.03
Ours (1k)	1.78	2.37	1.91	2.17	1.81	2.04	2.89	1.82	2.01	2.03	1.94

Table 1. SoTA evaluation (top) and ablations (bottom) on DAD-3DHeads [3], for additional categories. We report the NMLC for each model when averaging across various facial regions and categories. {Model}★ denotes the model was trained on the data samples used for our evaluation.

Model	Multiface	DAD3D-Heads
Ours (LLL Loss)	2.52	1.68
Ours (L1 Loss)	2.82	2.08
Ours (MSE Loss)	3.01	2.49

Table 2. Ablation studies concerning the loss function used, where Ours uses Laplacian Log Likelihood (LLL), evaluated on the full set of landmarks from Multiface [6] and DAD3D-Heads [3]. We report the NMLC for each model, when averaging across various facial regions.

the global consistency of landmark definitions across models, a presupposition integral to the NMLC metric.

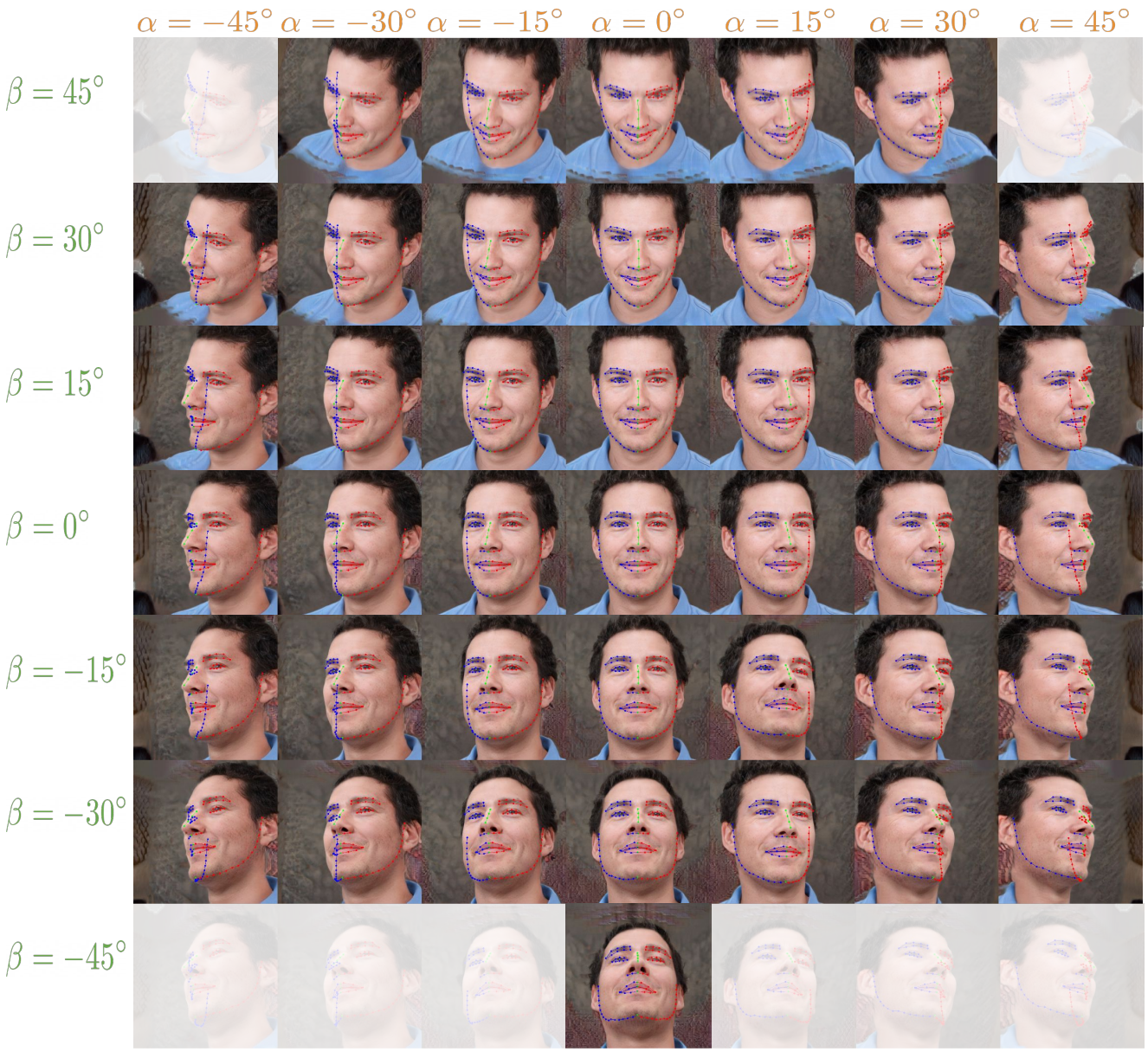

Figure 2. **Masked Multiview 3D Landmark Optimization's Camera "Rig":** Sample camera views used to perform the masked multiview 3D landmark optimization.

Figure 3. **3D-aware GAN Pseudo-labeled Samples.** Our approach can faithfully reconstruct 3D landmarks under extreme 3D head poses, and face outline landmarks are not affected by inherent GAN noise around face boundaries.

Model	Training Set	AFLW2000-3D-reannotated NME	DAD3D-Heads NME
FAN3D [1]	300WLP	2.85	3.83✓
SynergyNet [5]	300WLP	2.65	3.46✓
3DDFAv2 [2]	300WLP	3.33	3.10✓
DAD-3DNet [3]	DAD3D-Heads	5.10✓	2.71
DAD-3DNet+ [7]	DAD3D-Heads+	5.00✓	2.71
Ours	FaceLift	3.51✓	2.78✓

Table 3. Cross-dataset evaluation of NME on AFLW2000-3D-reannotated [9] and the validation set of DAD3D-Heads [3]. ✓denotes that the score is cross-dataset, meaning the training set definition is not compatible with the evaluation dataset and definition. We see that while our model is the best on the cross-dataset comparisons for each dataset, compatible SoTA models yield better scores since they do not incur the local definition bias of cross-dataset evaluation.

Figure 4. **Additional Qualitative Results on CelebV-HQ [8] dataset.** Here, the blue, green, and red axes represent Cartesian coordinates and denote the forward, up, and right vectors, respectively. Our approach can faithfully reconstruct 3D landmarks under challenging 3D head poses and harsh lighting.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, pages 1021–1030. IEEE CS, 2017. 2, 4

[2] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, pages 152–168. Springer, 2020. 2, 4

[3] Tetiana Martyniuk, Orest Kupyn, Yana Kurlyak, Igor Krashenyi, Jiri Matas, and Viktoriia Sharmanska. Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image. In *CVPR*, pages 20910–20920. IEEE, 2022. 1, 2, 4

[4] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. IDE-3D: interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM TOG*, 41 (6):270:1–270:10, 2022. 1

[5] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *3DV*, pages 453–463. IEEE, 2021. 2, 4

[6] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason M. Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. *CoRR*, abs/2207.11243, 2022. 1, 2

[7] Libing Zeng, Lele Chen, Wentao Bao, Zhong Li, Yi Xu, Junsong Yuan, and NimaKalantari. 3d-aware facial landmark detection via multiview consistent training on synthetic data. In *CVPR*. IEEE, 2023. 2, 4

[8] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *ECCV*, pages 650–667. Springer, 2022. 1, 5

[9] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155. IEEE CS, 2016. 1, 2, 4