Supplementary Material for Learning Motion Refinement for Unsupervised Face Animation

Anonymous Author(s)
Affiliation
Address
email

In this supplementary, we present detailed information on the architectural design, evaluation metrics, cross-identity evaluation, component ablations, and a more comprehensive parameter analysis. Additionally, we include video results demonstrating same-identity reconstruction and cross-identity animation, which can be found in the provided files under the name "scorr_animation.mp4". To facilitate understanding and further research, we are also sharing the code files, with our core module implementation available in "modules/scorr.py".

Architecture details: Our system mainly consists of a keypoint detector, a dense motion module, the source and driving structure encoders, a flow updater, and an image generator. The keypoint detector and dense motion modules are implemented using blocks similar to those used in previous works [5, 8]. In contrast, we provide architecture details of our source and driving structure encoders and image generator in Figure A1. All the modules are included in our provided codes for easy reference and implementation.

Evaluation metrics: We mainly introduce four metrics that utilized third-party models for evaluation.

- Average Keypoint Distance (AKD [6]). This metric computes the average keypoint distance between generated and ground-truth images. It is designed to evaluate the pose quality of the generated images. We use existing detectors [2] to extract the facial landmarks.
- Average Euclidean Distance (AED [6]). This metric is designed to assess the identity quality of generated images based on specific feature representations, that extracted from a pre-trained facial identification network [1]. The average Euclidean distance between generated and ground-truth video frames is computed.
- Average Rotation Distance (ARD [3]). We use the toolbox py-feat [4] to extract the Euler angles of the head poses, and then compute the average Euler angles distance between the generated and driving images. This metric evaluates the head pose quality.
- Action Units Hamming distance (AUH [3]). This metric measure the quality of facial expression, it computes the average Hamming distance between action units of generated and driving images. We use the toolbox py-feat [4] to extract facial action units.

Table A1: Cross-identity evaluation on the Voxceleb1 dataset.

	FOMM	MRAA	LIA	DAM	DaGAN	TPSM	FNeVR	Ours
ARD	3.122	2.678	3.883	2.669	3.090	2.724	2.755	**2.399**
AUH	0.850	0.729	0.772	0.717	0.751	0.668	0.751	**0.625**

Cross-identity animation evaluation: In the cross-identity scenario, we do not have ground-truth label videos, so we use the ARD and AUH metrics to evaluate the facial expression and head pose quality. Table A1 illustrates that our method surpasses others by a significant margin in both metrics, indicating its efficacy in capturing finer motions with good precision.

Component ablations: We conducted a thorough examination of the inputs and outputs of our non-prior based motion refinement module through detailed ablations. Specifically, we explored the impacts of the model variants of the source structure encoder without the input of a source image, the

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Table A2: Component ablations of the inputs and outputs of the proposed non-prior based motion refinement module. We present results on the Voxceleb1 dataset.

	L1	PSNR	LPIPS	AKD	AED
w/o occlusion	0.0354	**25.53**	**0.151**	1.177	0.108
w/o flow	0.0361	25.40	0.155	1.190	0.117
w/o warped source feature	0.0355	25.48	0.153	0.183	0.113
w/o source image	0.0355	25.51	**0.151**	1.180	0.109
Ours full	**0.0353**	25.51	0.152	**1.176**	**0.107**

Table A3: Analysis on the different setting of r. We present results on the Voxceleb1 dataset. Our method is generally robust to the pyramid levels and the patch radius.

	L1	PSNR	LPIPS	AKD	AED
$P = 0, r = 2$	0.0355	25.48	**0.151**	1.185	0.109
$P = 0, r = 3$	0.0354	25.52	**0.151**	1.174	0.107
$P = 0, r = 4$	**0.0351**	**25.54**	**0.151**	1.185	0.111
$P = 1, r = 2$	0.0354	25.51	**0.151**	1.186	0.110
$P = 1, r = 3$	0.0353	25.51	**0.151**	1.176	0.107
$P = 1, r = 4$	0.0353	25.51	**0.151**	1.183	0.109
$P = 2, r = 2$	0.0355	25.52	**0.151**	**1.173**	**0.106**
$P = 2, r = 3$	0.0355	25.49	0.152	1.184	0.109
$P = 2, r = 4$	0.0355	25.43	0.152	1.192	0.112

flow updater without the input of warped source features, and the flow updater without the output of occlusion or flow (i.e. only updating either occlusion or flow and keeping the other the same as in the initialization process). The results are presented in Table A2. It is noteworthy that both the warped source feature and the source image play an important role in the motion refinement process as demonstrated by the decreased AKD and AED metric values on the w/o warped source feature and w/o source image variants. Importantly, if we solely update the occlusion map without updating the motion flow, a significant decrease in AKD and AED metrics occurs, validating the significance of our motivation to refine the coarse motion flow estimated by prior-based motion models. We do observe a slight performance decrease when not updating the occlusion map, further demonstrating the importance of refining motion flow.

Parameter analysis: In the main paper, we examined the sampled patch radius r in the structure correlation volume. Here we continue to explore this parameter along with the pyramid levels we utilized for the correlation volume, as both of these parameters can expand the the search space of the structure correlation volume, resulting in the expanded correlation feature dimensions. Specifically, we pool the structure correlation volume $C \in \mathcal{R}^{h\times w\times h\times w}$ in the last two dimensions to obtain the pyramidal structure correlation volume $\{C^i \in \mathcal{R}^{h\times w\times h/2^i\times w/2^i}\}_{i=0}^{P}$, In the main paper's experiments, we set P to 1. Note that in each iteration, we sample patch correlation features on all C^i's in the pyramid, as the pyramid design aims to capture more rich motion features of different scales, which is inspired by the optical flow method RAFT [7]. As presented in Table A3, when we increased the pyramid levels, we observed no significant performance improvements. This indicates that the single level pyramid provided sufficient motion information for our motion flow refinement. In summary, we empirically set $P = 1, r = 3$ in our primary experiments to demonstrate our approach's generality, even though we could potentially achieve even better performance with other parameter choices.

References

[1] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2), 2016. 1

[2] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 1

[3] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14398–14407, 2021. 1

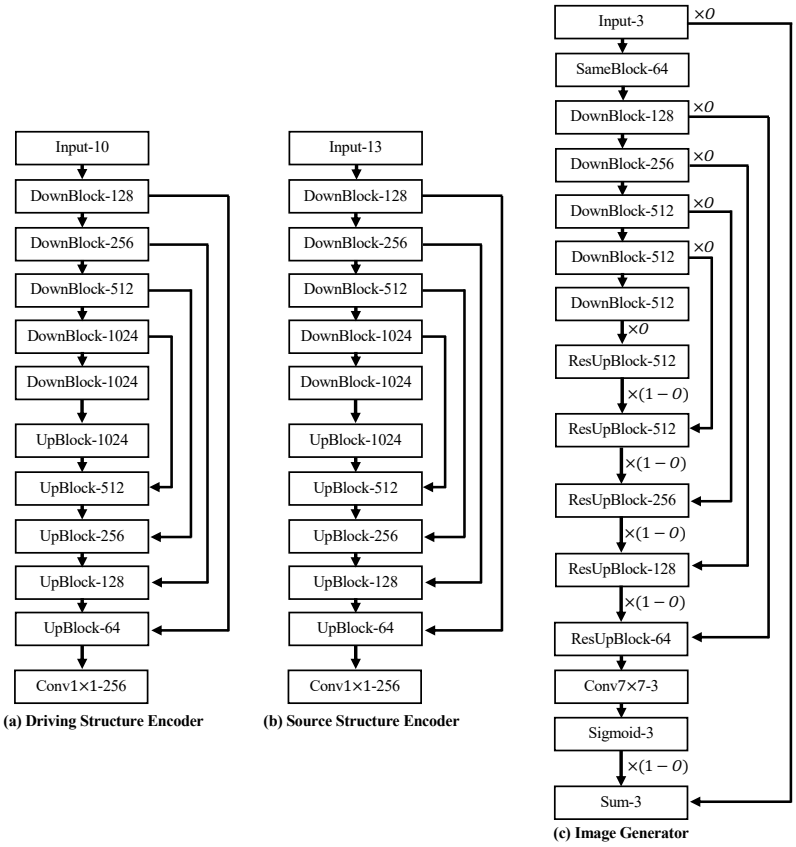

Figure A1: Detailed architectures of the driving structure encoder, the source structure encoder, and the image generator.

[4] Eshin Jolly, Jin Hyun Cheong, Tiankang Xie, Sophie Byrne, Matthew Kenny, and Luke J Chang. Py-feat: Python facial expression analysis toolbox. *arXiv preprint arXiv:2104.03509*, 2021. 1

[5] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 1

[6] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, 2019. 1

[7] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[8] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 1