

Supplemental Material for FacialFlowNet: Advancing Facial Optical Flow Estimation with a Diverse Dataset and a Decomposed Model

Anonymous Authors

In this supplementary material, we first introduce the details of our UV-Texture extraction module (Sec. 1). Following that, we conduct additional experiments to further validate our approach and dataset (Sec. 2). Finally, we present more visualization results in Sec. 3.

1 UV-TEXTURE EXTRACTION PIPELINE

As shown in Fig. 1, we adapted FFHQ-UV’s [3] pipeline to create FLAME-based [10] UV-textures. This pipeline consists of two steps: UV-texture extraction (1.1) and UV-texture completion (1.2).

1.1 UV-Texture Extraction

The process of extracting UV-texture from a face image, also termed "unwrapping," requires a single-image 3D face shape estimator. We use DECA [5] to reconstruct 3D meshes for input faces. Subsequently, facial UV-textures are unwrapped by projecting the face images onto the 3D meshes. This yields three texture maps, T_l , T_r , and T_f for left, right, and frontal views, respectively. We then perform color matching between them to prevent color jumps and linearly blend them together with pre-defined visibility masks, resulting in a complete texture map T .

1.2 UV-Texture Completion

To create a complete texture, we begin by manually crafting an image that includes areas such as the neck, ears, hair, and eyes—referred to as the temple texture \bar{T} . Following this, we blend the extracted facial texture T with the temple texture \bar{T} using color matching, and then apply Laplacian pyramid blending [4]. To address the eyes’ textures, we utilize a facial parsing model to extract the eye regions from the frontal view image. Subsequently, we use a circular detection method to extract images of the two pupils and paste them onto the corresponding positions in the texture. Through these steps, we obtain the final UV-texture T_{final} .

1.3 UV-Texture Samples

We have implemented the aforementioned pipeline on images from FFHQ-Norm [3], producing high-quality UV maps at a resolution of 1024×1024 . Samples of these UV-textures, along with the corresponding rendered images, are illustrated in Fig. 5.

2 MORE EXPERIMENTS

2.1 Comparison with Facial Alignment Methods

We also compared our method with 3DDFA_V2 [6, 7], a state-of-the-art facial alignment method. Facial alignment is typically achieved through facial landmark alignment methods, so we calculated the endpoint error (EPE) between the landmarks predicted by 3DDFA_V2 and the ground truth labels provided by CK+ [11]. The quantitative results in Tab. 1 indicate that, compared to 3DDFA_V2,

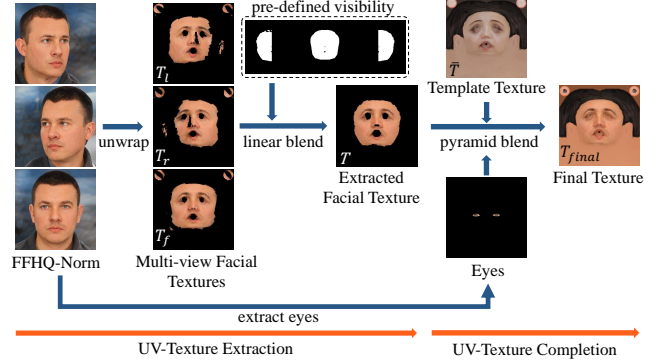


Figure 1: The proposed UV-Texture extraction pipeline. It takes normalized multi-view faces from FFHQ-Norm [3] as input, and outputs a normalized texture UV-map.

Table 1: Comparison with facial alignment method. The best result is indicated in bold.

Methods	3DDFA_V2 [6, 7]	DecFlow(Ours)
EPE	7.49	3.47

our facial flow provides more accurate tracking of facial landmarks, which might be helpful for facial alignment tasks. Visual samples in Fig. 2 also support the same conclusion, as the result of our method aligns more closely with the ground truth.

2.2 Ablation Study

One Decoder: We also explored an alternative network structure, utilizing a single decoder to simultaneously estimate both facial flow and head flow. In this architecture, we employ a decoder with two independent GRU blocks. During the iterative optimization process of optical flow, for instance, iterating 20 times, the first GRU block is used for optimization in the initial 10 iterations. It is simultaneously constrained by the flow labels from FFN-H to generate head flow. In the subsequent 10 iterations, the second GRU block is employed for optimization, supervised by the optical flow labels from FFN-F, building upon the existing head flow. With this structure, we could generate head flow and facial flow simultaneously using a single decoder, denoted as DecFlow (One Dec.). The results in Tab. 2 indicate that, while this network structure is capable of decomposing facial flow, it tends to reduce its accuracy. Which validates the necessity of using two independent decoders.

Without Background Images: The addition of static real images to the dataset is hypothesized to improve the model’s robustness. To validate this hypothesis, we utilize masks to eliminate background

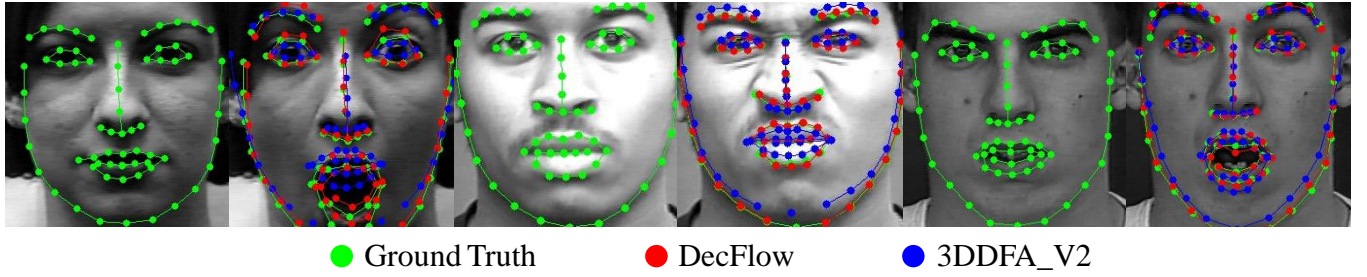


Figure 2: Comparison with facial alignment method.

Table 2: Ablation study of decomposed flow decoder and the background images. (V.) and (L.) denote the vertex and landmark coordinates, respectively. (One Dec.) refers to the structure with only one decoder, and (w/o Back.) indicates training the model on the dataset without background images.

Model	FFN	CK+(V.)	CK+(L.)
GMA [9]	0.142	4.73	3.59
DecFlow (One Dec.)	0.140	4.68	3.68
DecFlow (w/o Back.)	0.333	4.68	3.57
DecFlow	0.132	4.67	3.47

images from the dataset, as illustrated in Fig. 3 (d). Subsequently, we train a new model on this dataset, denoted as DecFlow (w/o Back.). The evaluation results in Tab. 2 show that the model trained on images without background images challenges in handling regions beyond the face. Additionally, it also leads to a reduction in the performance of facial optical flow.

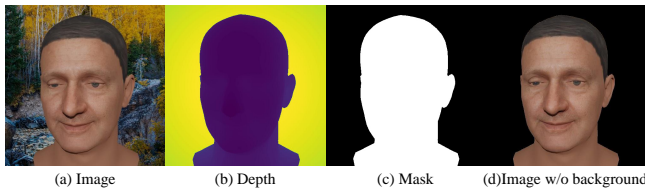


Figure 3: The ground truth labels for depth information.

2.3 Dynamic Facial Expression Recognition

We further assess the functionality of our proposed framework in dynamic facial expression recognition using a CNN network by Allaert et al. [2]. This network is composed of three successive layers of convolution, ReLU activation, and max pooling and the architecture ends with connected layers for the classification. Although we acknowledge that this architecture and its meta-parameters may not represent the state-of-the-art, our goal in this evaluation is to compare different optical flow approaches in low-complexity contexts to minimize learning biases.

We divide CK+ [11] into training and testing sets with a ratio of 7:3, and employed AUC (Area Under the Curve) and ACC (accuracy) as the evaluation metrics. The dynamic facial expression

recognition experiment involved facial optical flow estimated by various recent methods, as well as the facial and expression flow estimated by our network.

Tab. 3 presents the classification results for the six facial expressions. After finetuning on FFN-F, various methods showed different degrees of performance improvement, with the maximum enhancement being 9.6% (SKFlow 81.5 to 89.3). In the classification of expressions like Fear, facial flow alone achieves only 55.8% AUC, while utilizing the decomposed expression flow elevates it to 80.3%. Compared to GMA with general datasets (C+T+S), our approach and dataset demonstrated an improvement of up to 7.6% of AUC (88.3 to 95.0) and 28.0% of ACC (57.0 to 73.0). It highlights the superiority of our approach in estimating facial optical flow and the significance of decomposed expression flow in analyzing facial movements.

3 MORE VISUALIZATION

3.1 Extra Data

In addition to optical flow labels, we also generate ground truth depth information for each frame in FacialFlowNet, as depicted in Fig. 3. Given that in our dataset, the background image is positioned on a stationary plane behind the facial model, we can derive a mask for the moving region in the image based on the depth information. In Section 5.3 of the manuscript, we utilize this mask to assess only the end-point error (EPE) of the moving regions. We hope that this multi-modality data can assist researchers in facial analysis.

3.2 Visual Results

Samples of both FFN-F and FFN-F are presented in 5. Additional estimated flows for real-world images are presented in 6 and 7.

REFERENCES

- [1] Muhannad Alkaddour, Usman Tariq, and Abhinav Dhall. 2021. Self-Supervised Approach for Facial Movement Based Optical Flow. *IEEE Transactions on Affective Computing, IEEE Transactions on Affective Computing* (May 2021).
- [2] Benjamin Allaert, Isaac Ronald Ward, Ioan Marius Bilasco, Chaabane Djeraba, and Mohammed Bennamoun. 2019. Optical Flow Techniques for Facial Expression Analysis: Performance Evaluation and Improvements. *Le Centre pour la Communication Scientifique Directe - HAL - Diderot, Le Centre pour la Communication Scientifique Directe - HAL - Diderot* (Apr 2019).
- [3] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. 2022. FFHQ-UV: Normalized Facial UV-Texture Dataset for 3D Face Reconstruction. (Nov 2022).
- [4] Peter J. Burt and Edward H. Adelson. 1983. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics* (Oct 1983), 217–236. <https://doi.org/10.1145/245.247>

Table 3: The results of dynamic facial expression recognition. The evaluation metric of this table is AUC and ACC. DecFlow(F) and DecFlow(E) represent facial flow and expression flow. The best and second-best results are indicated in bold and underlined, respectively.

Methods	AUC of Each Emotion(%) \uparrow						Metrics(%) \uparrow	
	Disgust	Fear	Happy	Sad	Surprise	Angry	Mean AUC	ACC
RAFT [13]	97.6	51.2	94.2	90.6	<u>98.3</u>	94.7	87.8	55.0
GMA [9]	<u>98.4</u>	54.6	93.5	88.7	97.5	97.3	88.3	57.0
SKFlow [12]	88.8	46.4	85.2	81.1	96.4	91.3	81.5	49.0
FlowFormer [8]	94.6	48.2	91.4	85.8	94.5	96.5	85.2	63.0
RAFT+FFN	97.5 \downarrow 0.1	55.8 \uparrow 4.6	<u>99.4</u> \uparrow 5.2	<u>91.6</u> \uparrow 1.0	97.4 \downarrow 0.9	97.2 \uparrow 2.5	89.8 \uparrow 2.0	58.0 \uparrow 3.0
GMA+FFN	96.9 \downarrow 1.5	53.1 \downarrow 1.5	98.1 \uparrow 4.6	90.7 \uparrow 2.0	97.7 \downarrow 0.2	<u>97.5</u> \uparrow 0.2	89.0 \uparrow 0.7	<u>64.0</u> \uparrow 7.0
SKFlow+FFN	97.7 \uparrow 8.9	54.0 \uparrow 7.6	98.2 \uparrow 13.0	<u>91.6</u> \uparrow 10.5	97.1 \uparrow 0.7	97.2 \uparrow 5.9	89.3 \uparrow 7.8	60.0 \uparrow 11.0
FlowFormer+FFN	98.0 \uparrow 3.4	53.4 \uparrow 5.2	99.1 \uparrow 7.7	<u>91.6</u> \uparrow 5.8	97.5 \uparrow 3.0	<u>97.7</u> \uparrow 1.2	89.5 \uparrow 4.3	61.0 \downarrow 2.0
FMB [1]	95.2	45.1	83.0	79.3	87.4	86.7	79.5	50.0
DecFlow(F)	98.0	<u>57.5</u>	<u>99.4</u>	91.3	97.8	97.2	<u>90.2</u>	<u>64.0</u>
DecFlow(E)	99.3	80.3	99.9	95.1	98.4	96.6	95.0	73.0



Figure 4: The UV-Textures extracted with the proposed pipeline and the rendered images.

[5] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. *ACM Transactions on Graphics* (Aug 2021), 1–13. <https://doi.org/10.1145/3450626.3459936>

[6] Jianzhu Guo, Xiangyu Zhu, and Zhen Lei. 2018. 3DDFA. <https://github.com/cleardusk/3DDFA>.

[7] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *Proceedings of*

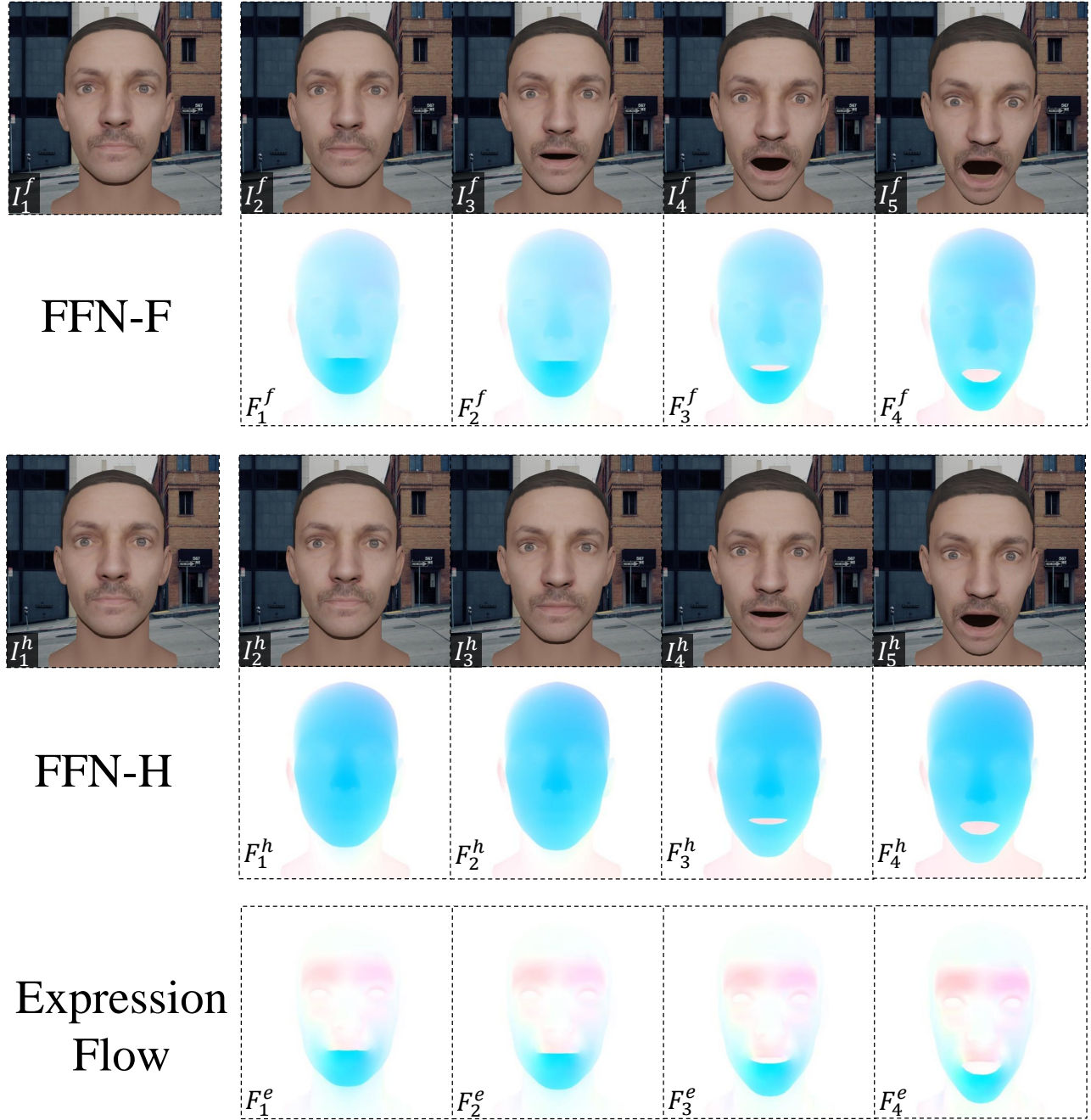


Figure 5: Samples of the FacialFlowNet dataset. I_t^f and I_t^h represent the t^{th} frame in FFN-F and FFN-H respectively. F_t^f , F_t^h , and F_t^e are facial flow, head flow, and expression flow, respectively.

the European Conference on Computer Vision (ECCV).

- [8] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. 2022. FlowFormer: A Transformer Architecture for Optical Flow. *ECCV* (2022).
- [9] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. 2021. Learning To Estimate Hidden Motions With Global Motion Aggregation. *Cornell University - arXiv, Cornell University - arXiv* (Apr 2021).

- [10] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics* (Dec 2017), 1–17. <https://doi.org/10.1145/3130800.3130813>
- [11] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. <https://doi.org/10.1109/cvprw.2010.5543262>

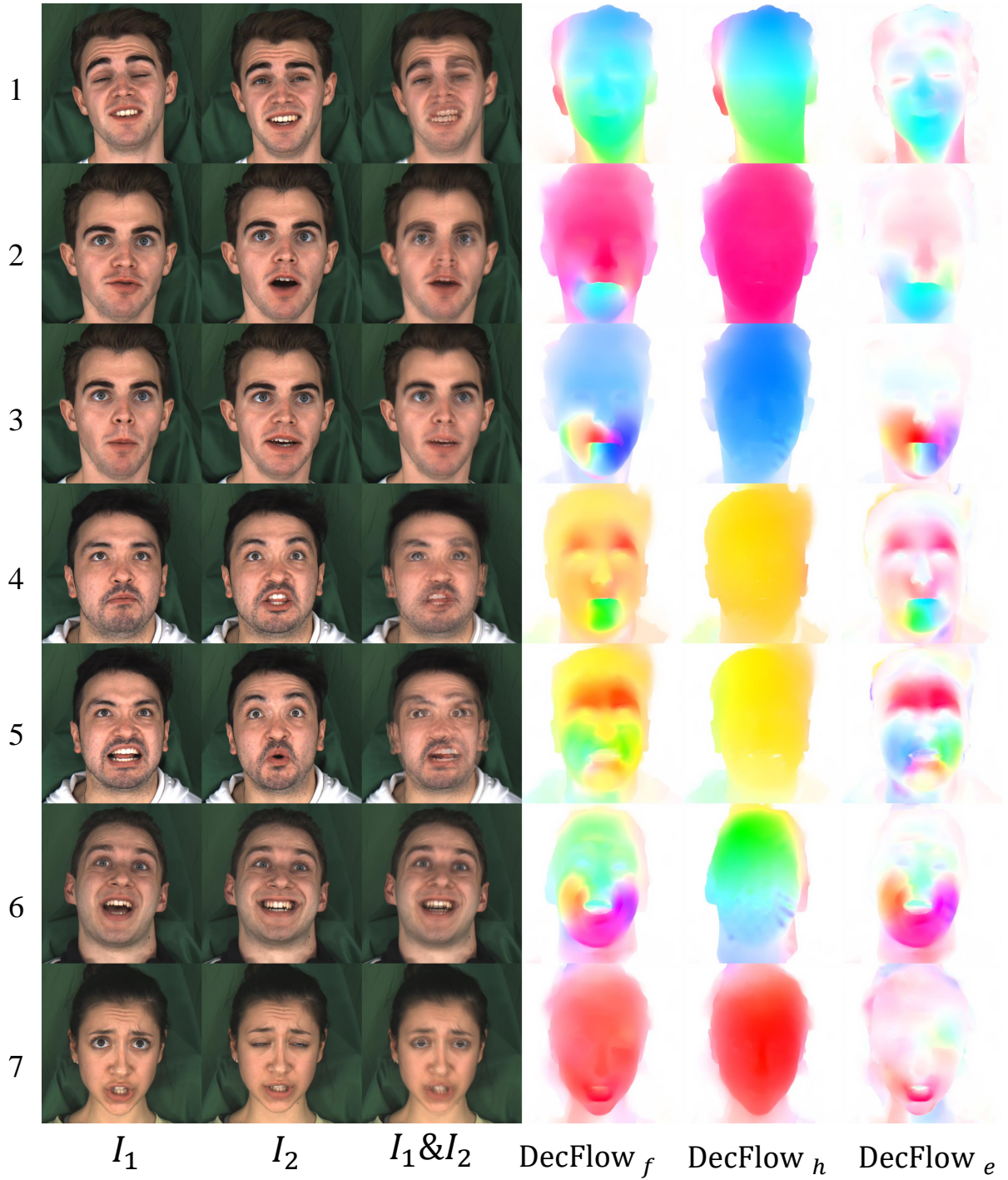


Figure 6: Qualitative results on real-world images from MEAD [14]. DecFlow_f , DecFlow_h , and DecFlow_e denote facial, head, and expression flow obtained by our method, respectively.

[12] Shangkun Sun, Yuanqi Chen, Yu Zhu, Guodong Guo, and Ge Li. 2022. SKFlow: Learning Optical Flow with Super Kernels. (May 2022).

[13] Zachary Teed and Jia Deng. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. 402–419. https://doi.org/10.1007/978-3-030-58536-5_24



Figure 7: Qualitative results on real-world images from the Internet. DecFlow_f , DecFlow_h , and DecFlow_e denote facial, head, and expression flow obtained by our method, respectively.

[14] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. *MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation*. 700–717. https://doi.org/10.1007/978-3-030-58589-1_42

[//doi.org/10.1007/978-3-030-58589-1_42](https://doi.org/10.1007/978-3-030-58589-1_42)