Towards Real-world Video Face Restoration: A New Benchmark (Supplementary Material)

In section 1, we elaborate on raw data pre-processing and data categorization of the FOS dataset and present more data samples of face images and video clips. Furthermore, more details about the subjective evaluation system, results, and qualitative comparison samples on both face images and videos are included in section 2.

1. Dataset

Pre-processing. The raw videos are preprocessed to extract face-centered clips for easy and fast adoption. Specifically, RetinaNet [10] is applied to detect bounding boxes of faces in all video frames. Note that small ones (less than 20×20) are filtered out and all the remaining face bounding boxes (size of $20 \sim 300$) are padded with 30% length of their heights/widths to achieve full coverage for the face area. Next, we implement a face tracking algorithm, SORT [1], to group those processed face detections into face tracks. For each track, a fixed rectangle that can cover all the bounding boxes will be adopted for cropping video clips. Then, we resize all the video clips to a fixed size of 128×128 and cut the frame lengths to less than 1500. Finally, face verification is utilized to remove those clips that have several identities. The whole process can finally generate 3,316 clips, of which 1,484 clips originated from YTF dataset [14], 410 clips come from YTCeleb dataset [6] and 1,422 clips are from self-collected videos.

Data Categorization. To reduce the cost of data categorization, an approach is first designed to classify the side faces automatically. Based on the head pose estimation model Hope-Net [9], three meta directions (*i.e*. the yaw, roll, and pitch angles) of faces in the input images can be obtained. Then we calculate a head pose score by assigning different weights to each angle for determining a *side* face. Formally, given a human head from a face image, the three degrees of freedom of a human head, *i.e*. the egocentric rotation angles *yaw*, *roll*, and *pitch* are denoted as $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, respectively. We determine the head pose score τ of the target face in the given image by assigning different weights to each angle and calculating the L_∞ norm as:

$$\tau = ||\alpha\mathbf{x}, \beta\mathbf{y}, \gamma\mathbf{z}, \alpha\mathbf{x} + \beta\mathbf{y} + \gamma\mathbf{z}||_\infty.$$

In this work, the parameter α, β, and γ are set to 1.0, 0, and 1.2, respectively (since rotation in the roll direction can be offset by face alignment). Figure 1 illustrates the above-mentioned three degrees of freedom with the defined direction symbol marked. Next, we manually select the *occluded* subset since the occlusion situation is rather complicated and unpredictable.

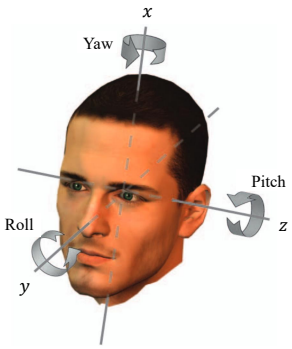


Figure 1. A schematic diagram originated from [8] shows the egocentric rotation angles *pitch*, *roll* and *yaw*, *i.e*. the three degrees of freedom of a human head.

More data statistics. We present the clip length distribution of FOS-V in Figure 2.

More data samples. We present more sample face images from FOS-real dataset as shown in Figure 3, and more sample frames of face clips from FOS-V dataset in Figure 5. The sample face images/clips originated from various real-world scenarios and thus involve faces with rich character-

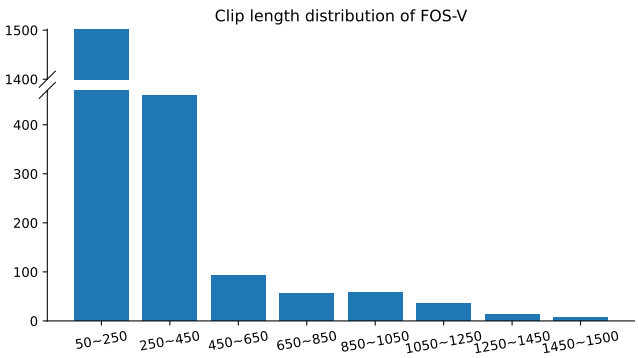


Figure 2. Clip length distribution of FOS-V.

Figure 3. Sample face images from the **FOS-real** dataset. Faces with different races, ages, expressions, and gaze directions are included.

istic information (*e.g*. races, ages, expressions, and gaze directions).

2. Evaluation

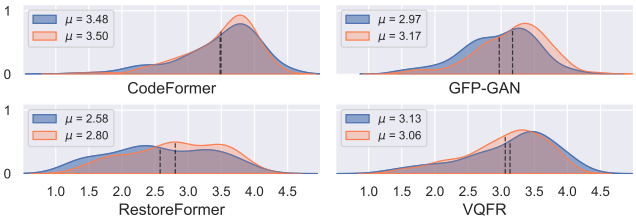
2.1. Subjective Evaluation

Subjective evaluation system. For subjective comparison, we conduct user studies to evaluate results achieved on **FOS-real** and **FOS-V**. A total of 30 volunteers are invited and briefly trained to perform the evaluation based on the proposed subjective system. Meanwhile, we design and deploy a user interface on a publicly accessible website for quick and convenient access by the invited volunteers. Figure 6 shows demo pages of our user study interface. We invite volunteers to score the given restored images and videos from different evaluation dimensions according to a five-point rating system. For 128 groups of images, the volunteers need to assign subjective scores from $1 \sim 5$ to the restored images in two dimensions: *realness* and *fidelity*, as depicted in Figure 6a. The same volunteers will continue scoring 108 restored video groups in *reconstruction performance* and *stability*, as shown in Figure 6b. Note that volunteers are guided in the user training to 1) focus more on the facial area instead of the background when rating, which is consistent with real-world face restoration applications; 2) decouple the two evaluation dimensions when rating both images and videos.

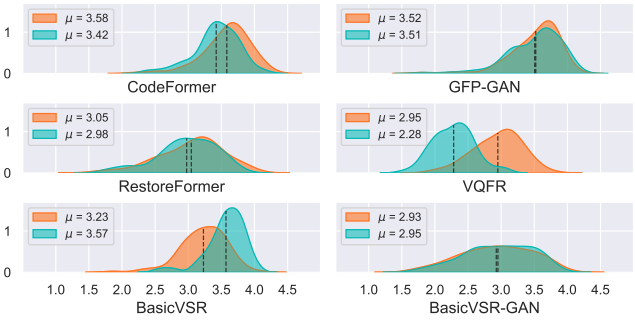
More subjective evaluation results. We present the complete results of the subjective comparison in Table 1 and Figure 4. A total of 6 BFR methods (CodeFormer [16], VQFR [4], RestoreFormer [13], GFP-GAN [12], GCFSR [5], GPEN [15]) and 4 VSR meth-

Table 1. User study statistics of different baseline methods. Point ≥ 3.5 is marked as red and point ≥ 3 is marked as blue.

	FOS-real								FOS-V	
	F.		O.		S.		Total		Total	
	Real.↑	Fidel.↑	Real.↑	Fidel.↑	Real.↑	Fidel.↑	Real.↑	Fidel.↑	Reconst.↑	Stability↑
GPEN [15]	3.30	3.39	3.12	3.28	3.21	3.31	3.22	3.33	3.60	3.47
GCFSR [5]	2.91	3.26	2.94	3.23	2.84	3.15	2.90	3.22	3.40	3.33
GFP-GAN [12]	3.05	3.22	2.83	3.04	3.00	3.23	2.97	3.17	3.52	3.51
VQFR [4]	3.42	3.32	2.92	2.89	2.91	2.85	3.13	3.06	2.95	2.28
RestoreFormer [13]	2.72	2.90	2.57	2.83	2.36	2.63	2.58	2.80	3.05	2.98
CodeFormer [16]	3.64	3.64	3.43	3.47	3.29	3.31	3.48	3.50	3.19	3.46
EDVR [11]	-	-	-	-	-	-	-	-	3.19	3.46
EDVR-GAN	-	-	-	-	-	-	-	-	3.07	3.08
BasicVSR [3]	-	-	-	-	-	-	-	-	3.23	3.57
BasicVSR-GAN	-	-	-	-	-	-	-	-	2.93	2.95



(a) Subjective score distributions of 6 BFR methods on **FOS-real** (#158) regarding *realness* (blue) and *fidelity* (coral).



(b) Subjective score distributions on 6 BFR methods plus 4 VSR methods on **FOS-V** (#108) regarding *reconstruction performance* (orange) and *stability* (green).

Figure 4. The distribution of subjective scores obtained on **FOS-real** (#158) and **FOS-V** (#108). The mean scores are denoted on the left of each subfigure.

ods (BasicVSR [3], BasicVSR-GAN, EDVR [11], EDVR-GAN) are evaluated based on the designed subjective evaluation system.

2.2. Qualitative Comparison

More results on single face images. More qualitative comparison results on full, occluded, and side faces of FOS-real image dataset based on 7 BFR methods (CodeFormer [16], VQFR [4], RestoreFormer [13], GFP-GAN [12], DMDNet [7], GCFSR [5], GPEN [15]) and 4 VSR methods (BasicVSR [3], BasicVSR-GAN, EDVR [11], EDVR-GAN) are shown in Figure 7. Note that the results of VSR methods are included since we select the images of FOS-real which also exists in FOS-V as frames.

More results on video face clips. Figure 8 shows the qualitative comparison of 6 BFR methods (with DMDNet excluded) plus 4 VSR methods on the FOS-V video dataset.

3. Social ethical concerns

Copyright claims. All images and videos from the FOS datasets are obtained from the Internet which are not properties of our institutions. Their copyright remains with the original owners of the video. FOS datasets are only available for non-commercial use. Any users can only acquire the datasets following the license provided by [2].

Potential data biases. The distribution of identities in the FOS datasets may be unrepresentative of the global human population. Please be careful of unintended societal, gender, racial, and other biases when evaluating models on FOS datasets.

Potential malicious use. This benchmark is proposed to encourage the development of BFR and VFR fields, whose typical application scenes include old photo/video restoration and low-quality face image/video restoration. However, images and videos may contain faces whose identities are intentionally blurred or obscured for privacy protection. Potential malicious use BFR methods to restore these privacy-protected cases may raise social ethical concerns.

References

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. 1

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 3

[3] Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, and

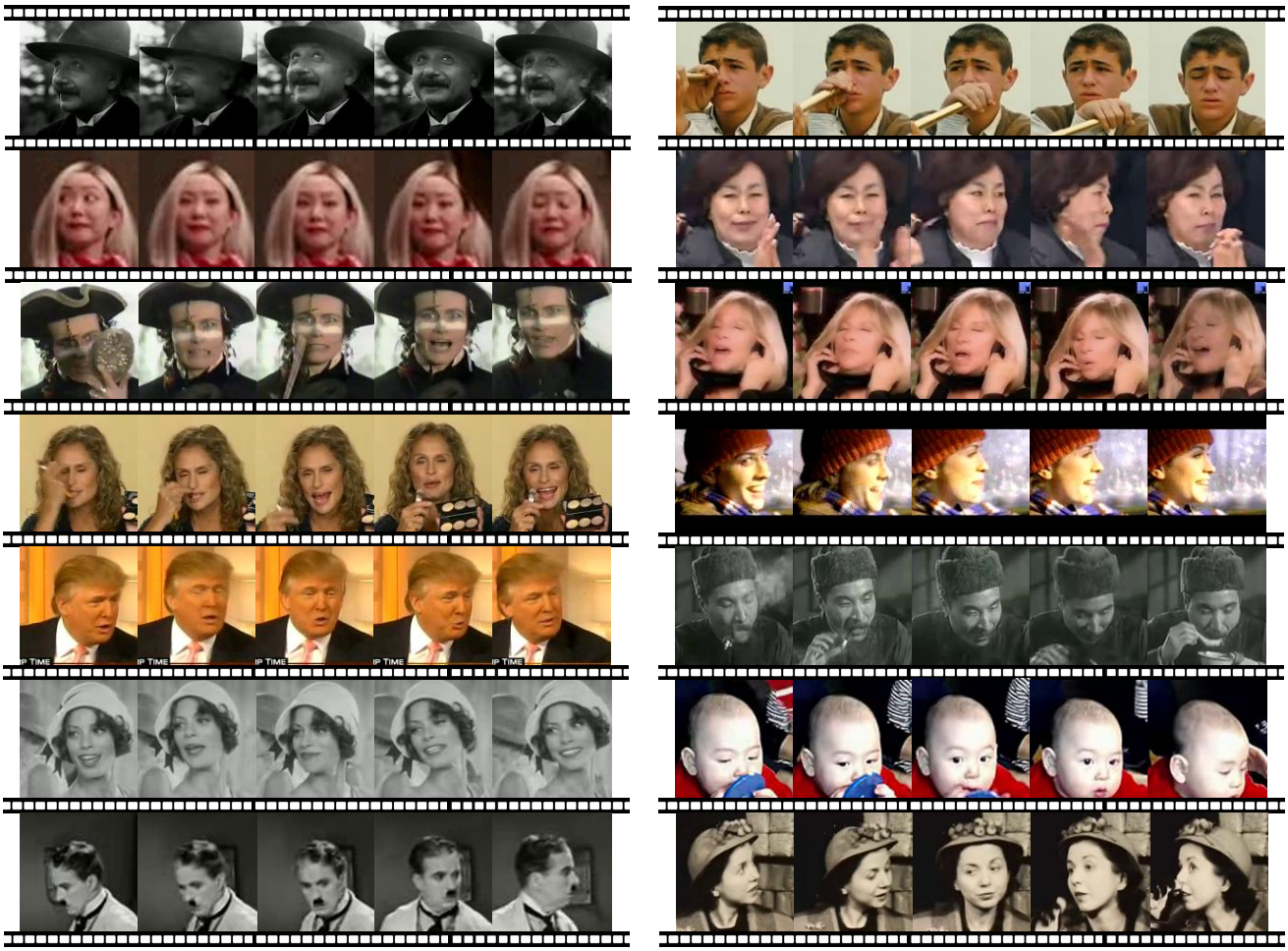


Figure 5. Sample frames of face clips from the **FOS-V** dataset. The clips involve both large-motion and stable human face frames.

Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. *Computer Vision and Pattern Recognition*, 2021. 3

[4] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. 2, 3

[5] Jingwen He, Wu Shi, Kai Chen, Lean Fu, and Chao Dong. Gcfsr: a generative and controllable face super resolution method without facial and gan priors. 2, 3

[6] Minyoung Kim, Sanjiv Kumar, Vladimir Pavlovic, and Henry Rowley. Face tracking and recognition with visual constraints in real-world videos. In *2008 IEEE Conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 1

[7] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. 2022. 3

[8] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4): 607–626, 2008. 1

[9] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018. 1

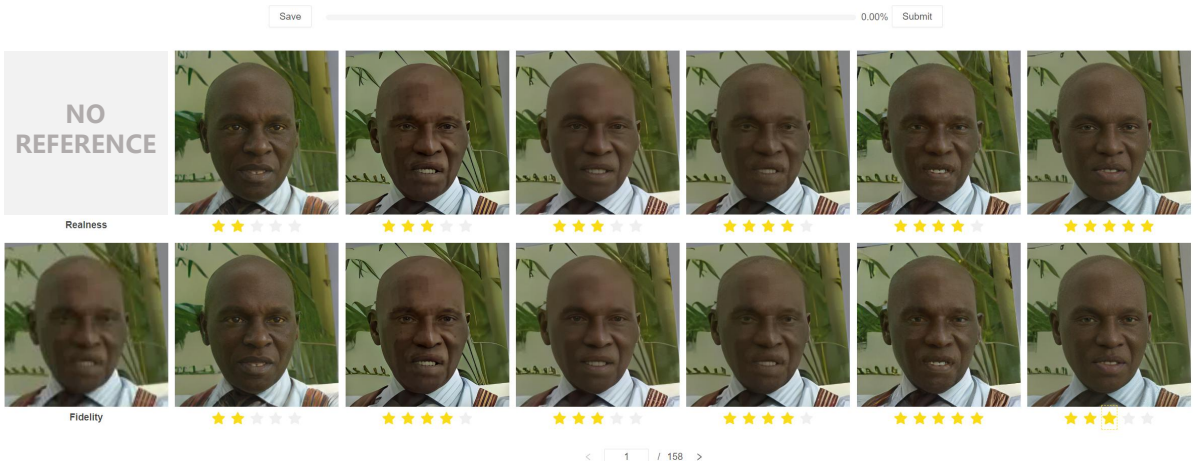
[10] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. 1

[11] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. *arXiv: Computer Vision and Pattern Recognition*, 2019. 3

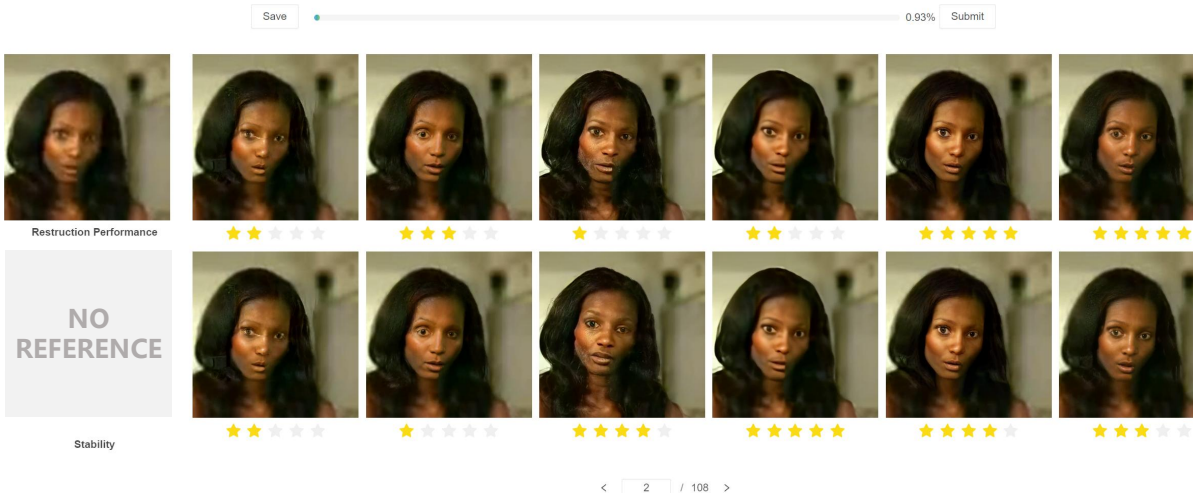
[12] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. *Computer Vision and Pattern Recognition*, 2021. 2, 3

[13] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. 2022. 2, 3

[14] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition



(a) The interface of subjective evaluation on **images**.



(b) The interface of subjective evaluation on **videos**.

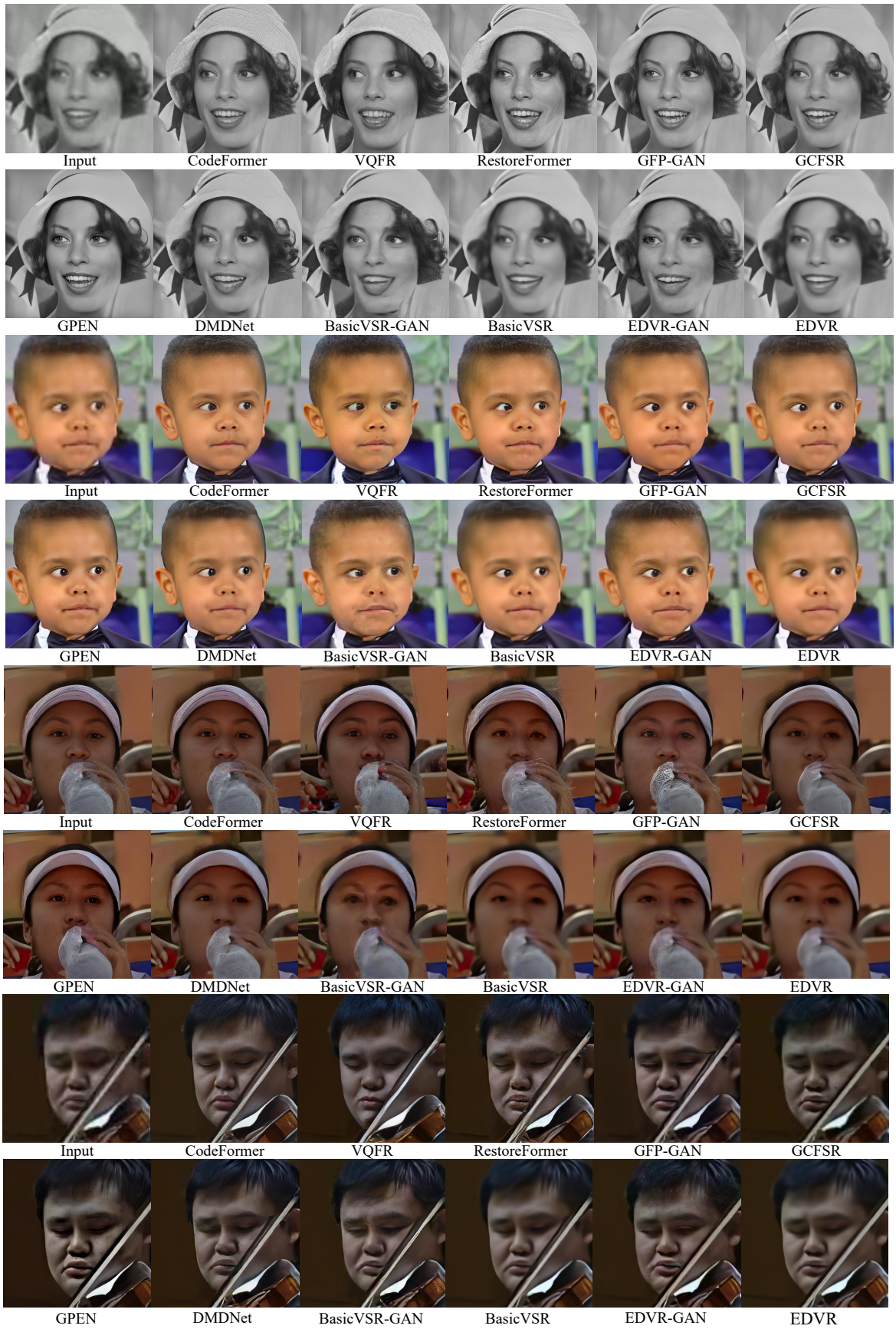
Figure 6. The demo page screenshots of our designed user study interface. Volunteers are invited to score the given restored results from different evaluation dimensions according to a five-point rating system.

in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011. 1

[15] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. *Cornell University - arXiv*, 2021. 2, 3

[16] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. 2022. 2, 3


Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR
Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR
Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR
Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR


Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR
Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR
Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR
Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR


Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR
Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR
Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR
Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR


Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR
Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR
Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR
Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
DMDNet
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR

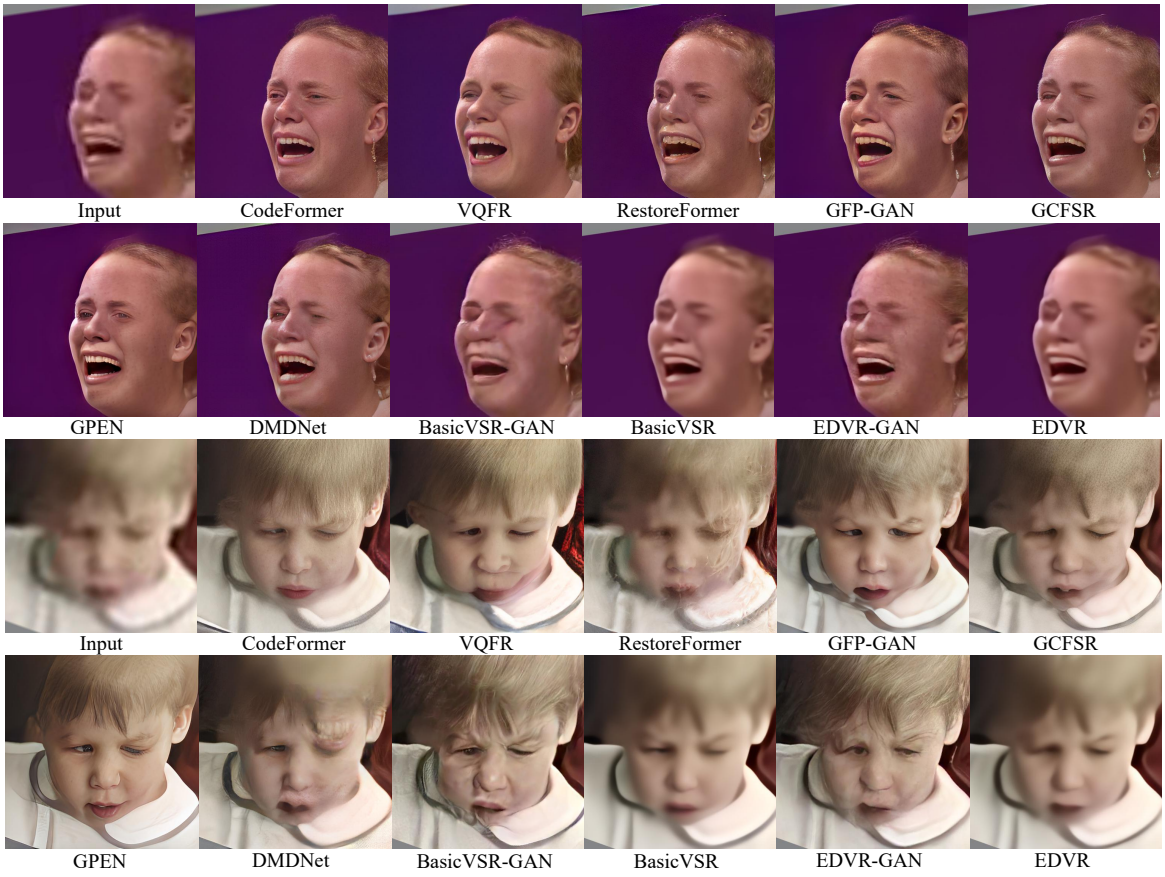
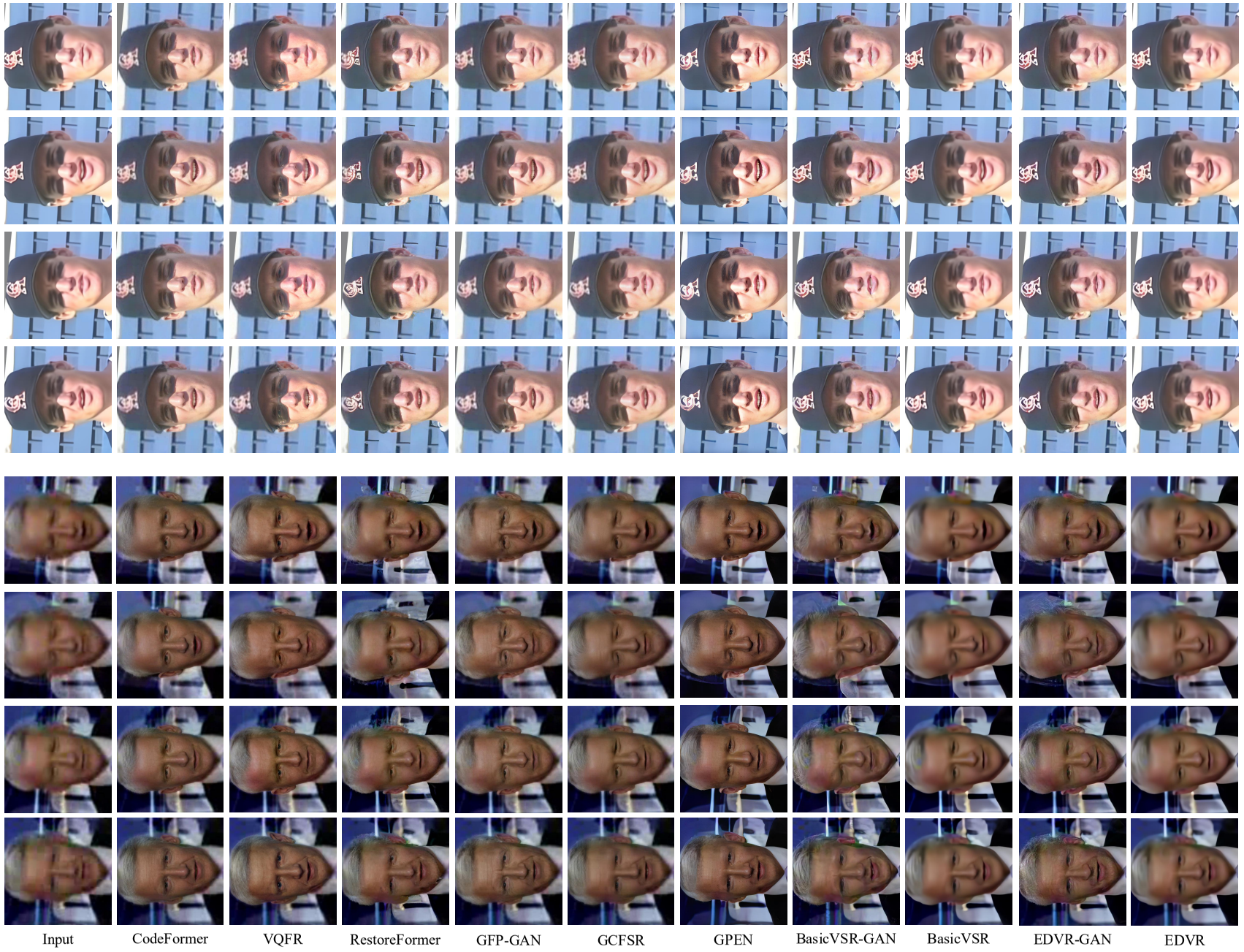


Figure 7. Qualitative comparison of 11 baselines in **FOS-real** image dataset. (**Zoom in for details**)


Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR


Input
CodeFormer
VQFR
RestoreFormer
GFP-GAN
GCFSR
GPEN
BasicVSR-GAN
BasicVSR
EDVR-GAN
EDVR

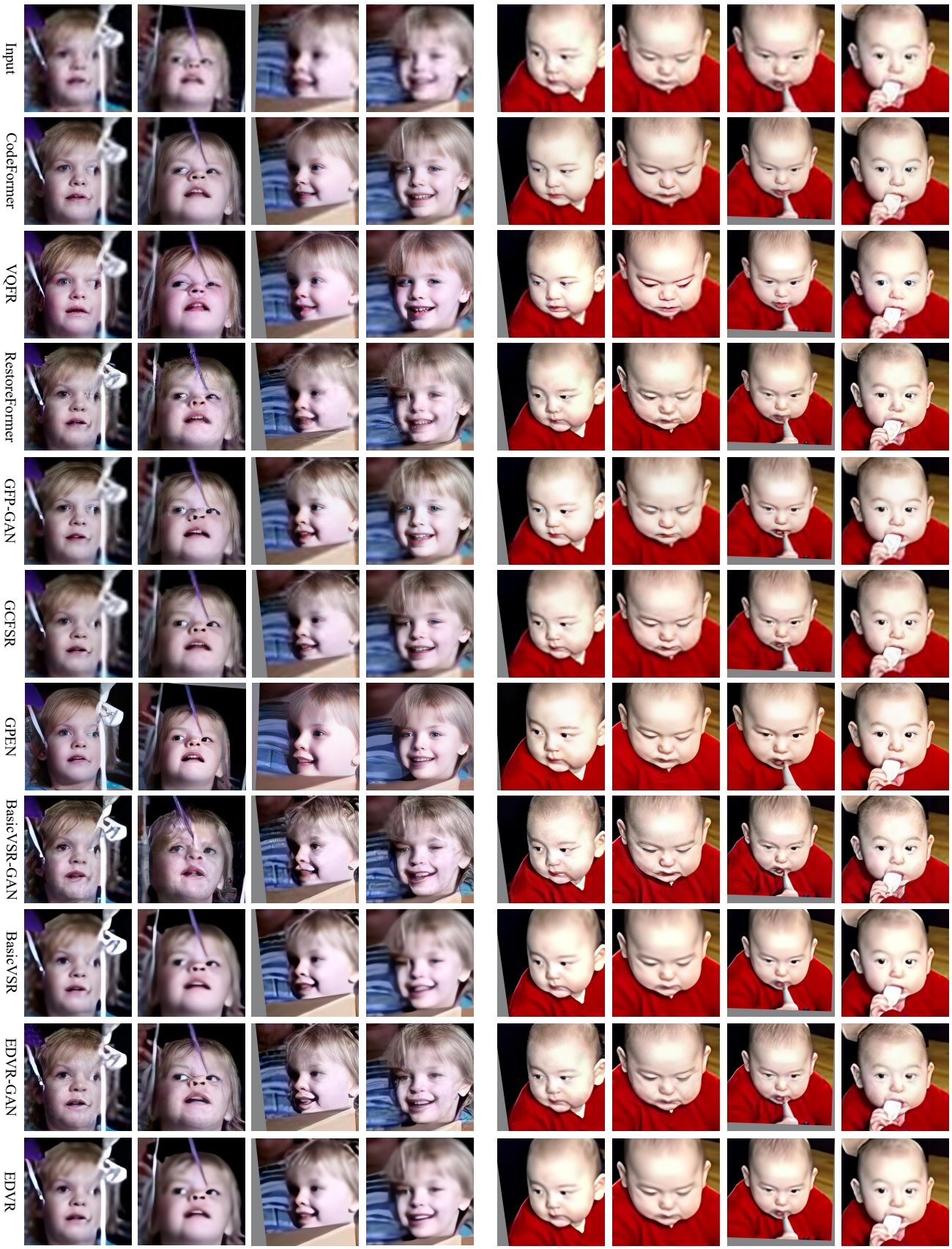


Figure 8. Qualitative comparison of 10 baselines in **FOS-V** video dataset. (**Zoom in for details**)