Towards Real-world Video Face Restoration: A New Benchmark

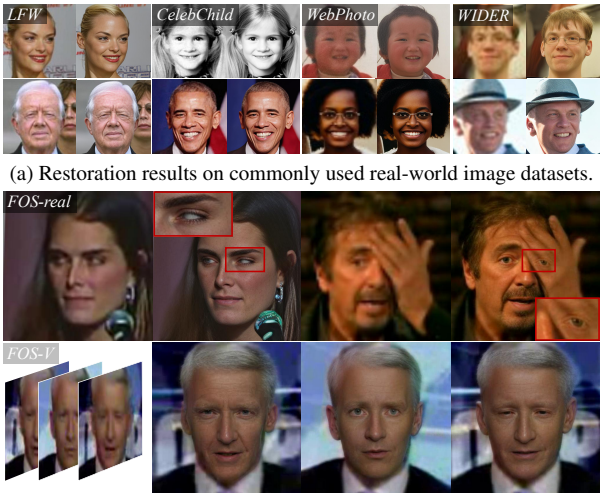
Ziyan Chen[1,3*], Jingwen He[2,3*], Xinqi Lin[1], Yu Qiao[3], Chao Dong[1,3†]
[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]The Chinese University of Hong Kong
[3]Shanghai AI Laboratory
{chen.ziyan, hejingwenhejingwen, linxinqi007}@outlook.com, {chao.dong, yu.qiao}@siat.ac.cn

Abstract

*Blind face restoration (BFR) on images has significantly progressed over the last several years, while real-world video face restoration (VFR), which is more challenging for more complex face motions such as moving gaze directions and facial orientations involved, remains unsolved. Typical BFR methods are evaluated on privately synthesized datasets or self-collected real-world low-quality face images, which are limited in their coverage of real-world video frames. In this work, we introduced new real-world datasets named **FOS** with a taxonomy of "**F**ull, **O**ccluded, and **S**ide" faces from mainly video frames to study the applicability of current methods on videos. Compared with existing test datasets, FOS datasets cover more diverse degradations and involve face samples from more complex scenarios, which helps to revisit current face restoration approaches more comprehensively. Given the established datasets, we benchmarked both the state-of-the-art BFR methods and the video super resolution (VSR) methods to comprehensively study current approaches, identifying their potential and limitations in VFR tasks. In addition, we studied the effectiveness of the commonly used image quality assessment (IQA) metrics and face IQA (FIQA) metrics by leveraging a subjective user study. With extensive experimental results and detailed analysis provided, we gained insights from the successes and failures of both current BFR and VSR methods. These results also pose challenges to current face restoration approaches, which we hope stimulate future advances in VFR research.*

1. Introduction

Blind face restoration (BFR) aims at recovering a realistic and faithful face image from a corrupted one where the specific degradation is unknown. It is a challenging problem as the real-world scenarios contain complicated degradations and diverse facial poses as well as expressions. Existing blind face restoration works [13, 14, 46, 48, 54, 57] have achieved unprecedented progress by incorporating powerful generative models as facial priors, such as generative adversarial network (GAN) [12, 20–22] and vector-quantized autoencoder [10, 39, 44]. These blind face restoration algorithms [46, 54, 57] have shown a huge impact in the open source community, as well as its significant commercial potential in image enhancement applications.



(a) Restoration results on commonly used real-world image datasets.

(b) Restoration failures on our proposed FOS datasets.

Figure 1. The face restoration results achieved by CodeFormer [57] on widely used real-world datasets (a) [46, 57] and our proposed FOS datasets (b). (**Zoom in for details**)

Figure 1a shows that state-of-the-art BFR methods can yield incredible results in both identity preservation and facial features/texture details regeneration on widely used real image testing datasets (*e.g.*, WebPhoto-Test [46], LFW-Test [17], Wider-Test [57], CelebChild [46]). However, these commonly used test datasets cover limited scenarios. Specifically, WebPhoto-Test and Wider-Test are mainly about human photography, where (1) most faces are assured to have no occlusion, (2) diversity of the gaze and head di-

*Equal Contribution. This work is done in Shanghai AI Laboratory.
†Corresponding author.

(a) **FOS-syn**. The FOS-syn dataset contains a total number of 3, 150 face images derived from widely-used CelebA-HQ [19].

(b) **FOS-real**. The FOS-real dataset involves a total number of 4, 253 face images extracted from face videos.

(c) **FOS-V**. The FOS-V dataset consists of 3,316 face clips originating from YouTube videos.

Figure 2. Overview of the proposed FOS datasets.

rections are lacking, and (3) the facial expressions are not diverse enough. Although these missing features may be marginal in the image cases, they are commonly found in video cases. Therefore, when applying video test samples to existing BFR methods, we find their clear limitations in handling complex image faces or video motions. Figure 1b indicates obvious failures of the most recent state-of-the-art BFR method CodeFormer [57] in restoring side/occluded faces or maintaining stability in face video frames.

These observations revealed limitations of the existing image test datasets. Furthermore, real video face test sets remain lacking. Therefore, we introduce a new benchmark to advance future VFR research in this work. Specifically, we construct three test datasets, **FOS-V**, **FOS-real** and **FOS-syn**, with three representative face poses/conditions highlighted as ***Full***, ***Occluded***, and ***Side***. An overview of our FOS datasets is presented in Table 1, while visualized samples are in Figure 2. Compared to the existing test sets, FOS datasets show larger diversity that involves heterogeneous scenes (*e.g.* interviews, sports, concerts, and old movies) and a wide range of faces with different facial expressions, poses, and ethnicities.

Given the collected FOS testing datasets, we benchmark 11 representative state-of-the-art BFR methods and 4 VSR implementations [50] to reveal their potential and limitations for the FVR task. In addition, we empirically discovered that the commonly used metrics are inconsistent with the human eye. We thus explore the effectiveness of current image quality assessment (IQA) metrics and the face IQA (FIQA) by conducting a subjective user study based on the benchmarking results. This work aims to stimulate future advances in video face restoration. The contributions of this work can be summarized below.

1. We introduce a real-world video face dataset, FOS-V, filling a gap in the VFR field where real test sets are missing.
2. By benchmarking the state-of-the-art BFR and VSR on **FOS**, we identify future challenges of FVR task. None of the existing methods can achieve satisfying intra-frame reconstruction performance while maintaining inter-frame stability.
3. In addition, we leverage a user study to investigate the effectiveness of IQA and FIQA metrics. The results show that NIQE [35] and BRISQUE [34] are significantly inconsistent with subjective scores, while FIQA [15, 38, 43] metrics yield a tremendously strong correlation with the human eye.
4. A new metric VIDD is introduced to assess inter-frame stability of videos.

2. Related Work

Blind face restoration. Recent works exploit multiple facial priors to help the blind restoration problem. These prior-based methods can be three mainstreams: 1) Geometric priors. Spatial information such as facial parsing maps [5], facial landmarks [6, 24], and facial component heatmaps [55] are used to provide additional information on face shape and details. 2) Reference priors. Typically, a reference face or facial component dictionaries obtained from high-quality faces [9, 26, 27, 29] are leveraged to guide the face recovery. 3) Generative priors. A pretrained generative network, *e.g*. StyleGAN [20, 21] or VQ-GAN [10], is employed to provide realistic facial information [13, 46, 48, 57]. GFP-GAN [46] and GPEN [54] made the first endeavor to extract fidelity information from low-quality input face images to balance the realness and fidelity. In addition, CodeFormer [57], RestoreFormer [48], and VQFR [13] exploit to fuse high-quality priors from the learned vector-quantized dictionary and information from the low-quality inputs to enable the discovery of natural and realistic faces that well approximate the target faces.

Video restoration Although BFR methods have achieved great success in single image restoration, video face restoration draws little attention from researchers. A few attempts have been made in video face super resolution [11, 32, 51], most of which focus on a fusion of inter-frame spatial and temporal information, or aural and visual modalities. Recently, [58] made a new attempt to apply the diffusion model to the VFR task. It is also worth noting that some modest progress has been made in general video super resolution. For example, EDVR [45] and BasicVSR [3] demonstrated their appealing performance in both temporal consistency and single-frame restoration by leveraging temporal alignment, and aggregation.

Datasets CelebA-Test [19] is a widely used synthetic test set for single-image blind face restoration. The commonly-used testing datasets in the real-world scenarios, LFW-Test [17], WIDER-Test [57], WebPhoto-Test [46] and CelebChild-Test [46] are either collected from the Internet or originated from other face-related tasks such as face verification or face detection. The face images of these testing datasets tend to be photos with a frontal view. However, the absence of more hard testing samples prevents evaluations from fully reflecting the models' performance in the real world. Furthermore, a standard and real-world video face testing set is still missing. VFHQ [50] dataset is proposed as an alternative to VoxCeleb [7, 37] with higher-quality face images and enables an improvement in VSR methods.

3. Benchmark Settings

In this section, we introduce the established benchmark dataset and our evaluation settings. To establish the benchmark dataset, two real-world datasets (FOS-real and FOS-V) and a synthesized dataset (FOS-syn) are introduced. We first elaborate on data collection and categorization of two collected real-world datasets. Then we present how the synthetic test set is constructed with the widely-used CelebA-HQ [19]. Table 1 presents an overview of the FOS datasets. For evaluation, a user study is first conducted to study the effectiveness of the commonly used metrics. Based on these conclusions, the following assessment metrics are employed. 1) Traditional IQA metrics, including PSNR, SSIM [47], LPIPS [56] and FID [16], 2) one new general IQA metric (MANIQA [53]) and one FIQA metric (SER-FIQ [43]).

3.1. Dataset

Data collection. Our datasets derive from videos on the Internet. Video data instead of image data is collected since face cases from video frames are more diverse and general. This is also a major difference between our datasets and the existing real-world datasets [46, 57]. First, we include YouTube-Faces (YTF) [49] and YouTube-Celebrities (YTCeleb) [25] datasets as one of our resources. Specifically, the YTF dataset contains 3,425 videos of 1,595 different identities, and the average frame number of one video clip is 181.3. The YTCeleb dataset contains 1,910 videos of 47 people with an average of 163.0 frames. The frame sizes of these video clips from them are around $[300, 500]$, and the proportion of face area is less than 1/8 in most cases. Each frame is associated with unknown degradations, such as compression artifacts, blur, and various noises. The contents of YTF and YTCeleb datasets are mainly about celebrities talking in show or interview scenes. To increase the diversity of our test data, we further download 137 videos from YouTube using queries like *old movie*, *aging*, *symphony performance*, and *baseball game etc*. The average duration of these videos are around 10 minutes. Following VFHQ, we pre-process all the collected raw videos to obtain cropped face video clips. The whole process can finally generate 3,316 clips which have the following properties: a) the face area is fully covered and roughly centered; b) the resolution is fixed to 128×128; c) the frame length of a single clip is within [50, 1500]; d) a clip only contains one identity. (See more details in supplementary file)

Data categorization. After data collection, we build out test datasets by specifying the collected face images into three categories: 1) *full*: a full face is a front face, and its major facial features (eyes, cheek, nose, mouth, and jaw) are not occluded by other objects; 2) *occluded*: one or more facial features are occluded or truncated; 3) *side*: a side face refers to a face with incomplete facial features (*e.g*., one eye is hidden) due to a change in head pose. Figure 2 presents some examples of three categories. We first use Hope-Net [40] to estimate the head pose and automatically

Table 1. Overview of the existing public real-world testing datasets and our FOS datasets.

Dataset	# Full	# Occluded	# Side	# Total	Real/synthetic	Descriptions
CelebChild [46]	325	27	8	360	real	childhood and recent photos of celebrities
LFW-Test [46]	1566	113	32	1711	real	snapshots of celebrities
WebPhoto-Test [46]	383	19	5	407	real	old photos
WIDER-Test [57]	894	58	18	970	real	group photos
VFHQ-Test [50]	-	-	-	50	synthetic	videos of interviews
FOS-syn (Ours)	1020	1097	1032	3150	synthetic	photos of celebrities
FOS-real (Ours)	1633	1136	1484	4253	real	images of sports, old movies, interviews, concerts, *etc.*
FOS-V (Ours)	-	-	-	3316	real	videos of sports, old movies, interviews, concerts, *etc.*

determine a *full* or *side* face. The head pose estimation network Hope-Net outputs the orientation degrees of one human head regarding three egocentric rotation angles: *yaw*, *pitch* and *roll* [36]. Then, we calculate a head pose score by assigning weights to each angle to determine a *side* face. In this way, a face with a larger head pose movement will be assigned to a greater score. (See details in the supplementary file) We manually select the occluded subset according to occluded eyes, nose, or mouth from a face image.

FOS-syn consists of pairs of a ground truth face image and the synthesized counterpart. The ground-truth images originate from the resized 512×512 CelebA-HQ dataset. We first categorize the CelebA-HQ dataset based on the proposed data categorization approach and obtain 1,021, 1,097, and 1,032 images as *full*, *occluded*, and *side* subsets, respectively. To generate the synthesized low-quality images, we adopt the commonly used degradation model [13, 14, 46, 48]: $I^d = \{[(I^h \otimes \mathbf{k}_\sigma) \downarrow_r + \mathbf{n}_\delta]_{JPEG_q}\} \uparrow_r$, where I^d, I^h denote the low-quality image and the high-quality counterpart, respectively. First, the high-quality image I^h is convolved with Gaussian blur kernel $\mathbf{k}_\sigma$, Then, resampling with scale factor r is performed. Next, additive Gaussian noise $\mathbf{n}_\delta$ to the resampled image, and the JPEG compression with quality factor q is applied. Finally, the LQ image is resized back to 512×512. We follow the widely-used setting in [13, 46, 48] and randomly sample σ, r, δ, and q from $[1, 10]$, $[0.8, 8]$, $[0, 20]$, and $[60, 100]$, respectively. The resampling is based on bilinear interpolation, and the blur kernel size is fixed to 41.

FOS-real is derived from extracted video frames with a stride of 5 from our collected and pre-processed clips. Creating image datasets from video resources can capture more diverse head poses, facial expressions, and gaze directions (see Figure 2). After data categorization, we obtain 1,633 face images assigned to *full*, 1,136 images assigned to *occluded*, and 1,484 images assigned to *side*. We exclude video frames with severely truncated/occluded faces or extremely low quality.

FOS-V is a video test dataset with 3,316 processed face clips from the real world. This dataset involves heterogeneous scenes such as interviews, sports, nature, and old movies, where the faces are of various ethnicities, across a wide range of ages, and have diverse facial expressions as well as head motions (see Figure 1 and supplementary file).

3.2. Evaluation

In this section, we unify the evaluation protocol and illustrate the subjective evaluation criteria. We also specify the used IQA metrics and the proposed new metrics.

Evaluation protocol. For image test sets (**FOS-syn**, **FOS-real**), the evaluation is conducted on aligned face images. Since **FOS-syn** is already aligned, it can be directly used for quantitative comparison. For the unaligned **FOS-real**, we follow [50] to use RetinaNet [41] for facial landmark detection and OpenCV's `warpAffine` function for face alignment. This way, we obtain an aligned version of **FOS-real** for both quantitative assessment and subjective comparison. The video test sets (**FOS-V**, VFHQ-Test [50]) are based on a different evaluation protocol. In particular, the full-reference metrics are calculated upon the original unaligned video frames, while the no-reference IQA/FIQA metrics are calculated on the aligned video frames for a more specific evaluation on face quality. Therefore, an additional operation – pasting back to the original LQ video frame is required for full-reference metrics calculation. Following the impactive works [46, 54, 57], we use ParseNet [31] to predict the parsing map of each aligned and restored face image for only pasting the restored face region to the original LQ video frame. This strategy guarantees that only the face area is restored, while the background remains unchanged.

Subjective evaluation criteria. For subjective comparison, we conduct a user study to evaluate results on **FOS-real** and **FOS-V**. For image comparison on **FOS-real**, we specify two evaluation dimensions: *fidelity* and *realness*. One is for evaluating the identity preservation of the low-quality face image while the other is for evaluating face generation quality without referring to the LQ input. The video comparison on **FOS-V** also has two evaluation dimensions: *stability* (temporal consistency) and *reconstruction performance*. For each evaluation dimension, we employ a five-point grading system: 5–Outstanding; 4–Good; 3–Acceptable; 2–Insufficient; 1–Fail. Note that one-point (Fail) means the restoration performance is far from acceptable, such as exaggerated face distortion regarding *fidelity* and extremely poor image quality regarding *realness*; three-point (Acceptable) indicates the restored result is overall satisfactory but still needs improvement in terms of fine-grained facial attribute reconstruction/generation (*e.g.*, pupil, teeth, *etc.*); five-point (Outstanding) means that the restoration result

Figure 3. Visual examples of different scores on our five-point grading system. The number of stars lit corresponds to the score rated. (**Zoom in for details**)

maintains the identity perfectly or is of extraordinary face quality regarding both facial features (*e.g.*, eyes, mouth, *etc.*) and texture details (*e.g.*, hair, skin, *etc.*). Please see Figure 3 and the supplementary file for visualized illustrations of our five-point grading system.

Evaluation metrics. We first adopt several widely acknowledged full-/no-reference metrics (PSNR, SSIM, LPIPS [56], FID) for quantitative measurement. Then, we investigate the effectiveness of five no-reference image quality assessment algorithms (NIQE, BRISQUE [34], hyperIQA [42], MUSIQ [23], MANIQA [53]) and four face image quality assessment algorithms (IFQA [18], FaceQnet [15], SER-FIQ [43], SDD-FIQA [38]) by computing the correlation coefficients (PLCC and SROCC) with human opinion scores obtained by user study. It is observed that MANIQA and SER-FIQ are the most relevant with subjective scores (see Figure 6a) among these selected IQA and FIQA algorithms, respectively. Therefore, we use these two metrics on all test datasets.

Moreover, we propose a new Video IDentity Distance (VIDD) metric to evaluate the temporal consistency, *i.e.* the *stability*, between consecutive video frames. For an input clip consisting of N frames $\{x_1, ..., x_N\}$, a facial feature descriptor $\mathcal{F}$ is used to extract the semantic feature of each input x_i. The VIDD score indicates a mean distance of inter-frame features as

$$\text{VIDD} = \frac{1}{N}\sum_{i=1}^{N-1} ||\mathcal{F}(x_i), \mathcal{F}(x_{i+1})||_2. \quad (1)$$

In particular, we use a pretrained ArcFace [8] model as the descriptor $\mathcal{F}$ to extract discriminative facial features. Referring to Figure 4, the VIDD score demonstrates its significant consistency with the subjective scores, which will be elaborated in detail in section 4.1.

4. Experiments

We evaluate 11 representative state-of-the-art BFR methods on **FOS-syn** and **FOS-real**: GAN prior-based methods (GFP-GAN [46], GLEAN [2], GPEN [54], PULSE [33]), VQ-based methods (VQFR [13], CodeFormer [57], RestoreFormer [48]), reference-based method (DMDNet[1] [30]), and other methods (GCFSR [14], PSFRGAN [4], Hi-FaceGAN [52]). Furthermore, we evaluate 4 VSR models trained on VFHQ [50](EDVR [45], EDVR-GAN, BasicVSR [3], BasicVSR-GAN) on our real-world **FOS-V** dataset and the synthesized VFHQ-Test [50]. As a comparison, 6 most recent BFR methods (GFP-GAN [46], GPEN [54], VQFR [13], CodeFormer [57], GCFSR [14], RestoreFormer [48]) are also evaluated on those video datasets in a single-frame processing manner. [2]

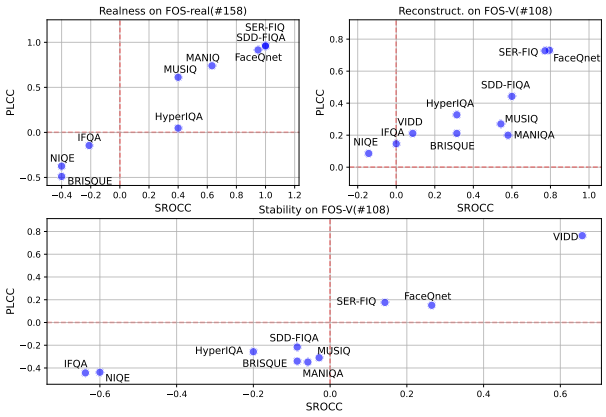


Figure 4. SROCC v.s PLCC results of 10 IQA/FIQA algorithms and proposed stability evaluation metric VIDD on **FOS-real**(#158) regarding *realness* and **FOS-V**(#108) regarding *reconstruction performance* and *stability*.

4.1. Investigation into The Evaluation Metrics

The motivation of this section is that we empirically found that the commonly used metrics are inconsistent with the human eye. To better evaluate face restoration results, we first investigate several recently proposed state-of-the-art NR-IQA algorithms (HyperIQA[42], MUSIQ [23], MANIQA [53]), and two traditional NR-IQA algorithms (NIQE [35], BRISQUE [34]). In addition, we select four recently proposed state-of-the-art FIQA algorithms (IFQA [18], FaceQnet [15], SER-FIQ [43], SDD-FIQA [38]) for more extensive exploration. Notably, we exclude the commonly used FID since this metric needs to be calculated between data distributions, which may be inaccurate for small-scale test sets.

We select 158 image samples from **FOS-real** as **FOS-real** (#158), and 108 video samples as **FOS-V** (#108). Methods on these two selected datasets are used for user rating. Based on the obtained subjective scores regarding the *realness* evaluation dimension, we calculate their correlation coefficients with the results of these selected

[1] An extension of DFDNet [28].

[2] DMDNet is excluded from the comparison since detecting 68 landmarks often fails under complex conditions.

IQA/FIQA algorithms. The SROCC v.s PLCC results are illustrated in Figure 4. Three FIQA algorithms, SER-FIQ, SDD-FIQA, and FaceQnet exhibit tremendously strong correlations with human ratings. For IQA algorithms, MANIQA and MUSIQ show good consistency. However, the commonly used NIQE presents a negative correlation. Furthermore, we provide the SROCC v.s. PLCC results on **FOS-V**(#108) in terms of *reconstruction performance*. SER-FIQ is still fairly aligned with subjective scores, while SDD-FIQA, MANIQA, and MUSIQ show an obvious performance drop. In addition, although we exclude FID from the investigation, it yields noticeable relevance with SER-FIQ scores (see Table 3).

Based on these observations, we adopt the best-performing FIQA and IQA algorithms, SER-FIQ, and MANIQA to evaluate all test sets. More importantly, we highly recommend that readers focus more on SER-FIQ scores than MANIQA and FID in the paper. To demonstrate the effectiveness of our proposed VIDD, we also compute the correlation coefficients concerning the obtained subjective scores on **FOS-V**(#108) according to *stability*. In the bottom Figure 4, the proposed VIDD yields 0.763 on SROCC and 0.657 on PLCC, which stands out among all explored metrics.

4.2. Benchmark Results on Image Datasets.

FOS-syn. We first compare the results of 10 blind face restoration methods[3] on the synthesized FOS-syn dataset. As seen from Table 2, GCFSR, CodeFormer, and GFP-GAN could achieve more satisfactory results in terms of all full-reference metrics (PSNR, SSIM, and LPIPS), indicating their great identity preservation and perceptual quality. Regarding the no-reference metrics, we observe that PULSE obtains the best results in SER-FIQ, but performs the worst with regard to PSNR, SSIM, and LPIPS. This manifests that PULSE generates high-quality face images without considering maintaining the identity information. In comparison, CodeFormer, RestoreFormer, VQFR, GFP-GAN gain good results on both SER-FIQ and LPIPS. In summary, on the synthesized dataset, GCFSR and DMDNet [30] show superior restoration performance, while RestoreFormer and CodeFormer exhibit better generation ability. It is clear that no method can surpass the others on all metrics and datasets.

FOS-real. We evaluate 11 BFR methods on our real-world image dataset. The qualitative comparison can be found in Figure 5 (see more results in supplementary file). Three no-reference metrics FID, MANIQA [53], and SER-FIQ [43] are adopted for quantitative comparison. Table 3 reports the performance comparison. By excluding the results of PULSE, we find that CodeFormer, VQFR, GFP-GAN, and RestoreFormer yield superior performance regarding the FIQA metric SER-FIQ. Regarding the IQA metric MANIQA, the top 5 methods are GPEN, GFP-GAN, CodeFormer, VQFR, and GCFSR. By carefully examining the FID results, we observe that CodeFormer, RestoreFormer, VQFR, and GFP-GAN attain more satisfactory results than other methods. Notably, all methods perform worse on both *occluded* and *side* subsets, especially regarding FID and SER-FIQ. However, MANIQA shows less significant differences among the three subsets since it is a general IQA metric.

[3]We exclude GLEAN for comparison as its blind version involves training on CelebA-HQ dataset.

4.3. Benchmark Results on Video Datasets.

FOS-V. We also conduct a performance comparison among BFR and VSR methods on our FOS-V dataset (See Table 5). The VSR methods achieve the best performance in VIDD, showing high stability on the video dataset. For IQA metrics, the BFR methods achieve superior overall performance compared to VSR methods. However, they fall behind in maintaining video restoration stability referring to VIDD. Among BFR methods, GPEN obtains excellent performance in both FID and the general metric MANIQA. CodeFormer demonstrates its superiority in real-world scenes with a significantly higher SER-FIQ score, but its stability falls out of the top five. GFP-GAN ranks second in the SER-FIQ score and achieves good performance regarding VIDD, indicating its excellent balance on *reconstruction performance* and *stability*.

We additionally leverage a synthesized video test set, **VFHQ-Test**, to our evaluation. We compare 6 state-of-the-art BFR and 4 VSR ($\times 4$) methods on the synthesized VFHQ-Test in Table 4. BasicVSR obtains the best result in PSNR and SSIM, while EDVR ranks second. As for LPIPS, EDVR-GAN and BasicVSR-GAN achieve first and second place, respectively. This implies that GAN training improves perceptual quality. For inter-frame stability (VIDD), BasicVSR and EDVR yield significantly better results than others due to the multi-frame and MSE-oriented training strategies. Regarding the no-reference IQA/FIQA metrics (SER-FIQ, MANIQA, and FID), CodeFormer, GPEN, and RestoreFormer achieve superior results over other methods.

4.4. Subjective Comparison

We next conduct a user study to evaluate the most recent 6 BFR and 4 VSR methods. We select a total of 158 images from **FOS-real**: 68 images from *full*, 45 images from *occluded*, and 45 images from *side*. Our hand-picking ensures diversity by only keeping one face image for one identity. Moreover, we carefully adopt images with diverse facial expressions, poses, and conditions. For video results comparison, we select a total of 108 video clips from **FOS-V**. The restored results together with the original LQ images, are sent to 28 users for grading in each evaluation dimension.

Table 2. Comparison of 10 state-of-the-art BFR methods on **FOS-syn**(#3150). Results are reported in PSNR, SSIM, LPIPS, SER-FIQ, MANIQA, FID. **Red** and blue indicate the best and second best results. The top five results are marked as gray.

	PSNR↑				SSIM↑				LPIPS↓				SER-FIQ↑				MANIQA↑				FID↓			
	F.	O.	S.	Total	F.	O.	S.	Total	F.	O.	S.	Total	F.	O.	S.	Total	F.	O.	S.	Total	F.	O.	S.	Total
PULSE [33]	22.01	20.56	21.24	21.25	0.623	0.584	0.643	0.616	0.466	0.518	0.513	0.500	0.734	0.700	0.661	0.698	0.590	0.573	0.554	0.572	88.29	67.86	70.66	75.40
PSFR-GAN [4]	24.30	23.70	25.15	24.37	0.620	0.596	0.669	0.628	0.409	0.444	0.440	0.431	0.515	0.413	0.365	0.430	0.626	0.612	0.582	0.606	68.77	58.86	63.38	63.56
HiFaceGAN [52]	24.75	23.64	24.71	24.35	0.622	0.592	0.644	0.618	0.418	0.456	0.460	0.445	0.766	0.616	0.531	0.637	0.595	0.553	0.530	0.559	71.90	62.59	72.12	68.73
GPEN [54]	24.60	23.41	24.75	24.24	0.664	0.635	0.707	0.668	0.394	0.402	0.388	0.395	0.768	0.629	0.556	0.650	0.695	0.695	0.674	0.688	69.39	63.67	66.95	66.60
GCFSR [14]	26.31	25.11	26.43	25.93	0.699	0.677	0.734	0.703	0.311	0.353	0.354	0.340	0.772	0.654	0.579	0.668	0.656	0.640	0.622	0.639	79.71	64.36	65.31	69.65
DMDNet [30]	25.67	24.48	25.79	25.30	0.681	0.656	0.719	0.685	0.355	0.398	0.389	0.381	0.769	0.637	0.555	0.653	0.622	0.586	0.558	0.588	67.61	60.91	61.91	63.41
GFP-GAN [46]	25.27	23.95	25.13	24.76	0.673	0.642	0.702	0.672	0.331	0.381	0.385	0.366	0.771	0.662	0.589	0.673	0.664	0.660	0.641	0.655	78.06	62.75	63.39	67.92
VQFR [13]	23.95	22.49	23.69	23.36	0.650	0.618	0.685	0.651	0.328	0.373	0.375	0.359	0.765	0.664	0.587	0.672	0.654	0.644	0.624	0.641	75.90	61.68	66.26	67.79
RestoreFormer [48]	24.74	23.41	24.68	24.26	0.636	0.609	0.671	0.638	0.328	0.373	0.381	0.361	0.781	0.656	0.576	0.670	0.648	0.634	0.613	0.631	72.09	56.92	56.81	61.80
CodeFormer [57]	25.29	25.17	25.17	25.21	0.661	0.692	0.692	0.682	0.317	0.358	0.358	0.345	0.783	0.680	0.603	0.688	0.666	0.666	0.645	0.659	80.90	65.02	68.78	71.40

Table 3. Comparison of 11 state-of-the-art BFR methods on **FOS-real**(#4253). Results are reported in SER-FIQ, MANIQA, and FID. **Red** and blue indicate the best and second best performance. The top five results are marked as gray.

	SER-FIQ↑				MANIQA↑				FID↓			
	F.	O.	S.	Total	F.	O.	S.	Total	F.	O.	S.	Total
PULSE [33]	0.710	0.674	0.638	0.675	0.565	0.549	0.542	0.553	58.90	61.28	61.16	60.32
PSFR-GAN [4]	0.664	0.511	0.379	0.523	0.579	0.550	0.532	0.555	46.47	64.93	75.74	61.61
HiFaceGAN [52]	0.631	0.469	0.313	0.477	0.502	0.463	0.427	0.465	63.46	94.00	117.88	90.61
GLEAN [2]	0.685	0.511	0.387	0.534	0.591	0.567	0.529	0.563	52.07	69.79	79.95	66.53
GPEN [54]	0.679	0.519	0.406	0.541	0.672	0.661	0.641	0.658	57.40	76.37	82.45	71.21
GCFSR [14]	0.675	0.517	0.391	0.534	0.609	0.582	0.560	0.585	46.77	69.79	68.65	60.55
DMDNet [30]	0.673	0.507	0.373	0.524	0.578	0.543	0.509	0.545	41.25	65.76	70.68	58.07
GFP-GAN [46]	0.687	0.532	0.419	0.552	0.638	0.622	0.599	0.620	44.27	65.50	64.44	56.98
VQFR [13]	0.695	0.533	0.425	0.558	0.614	0.593	0.567	0.592	42.99	65.25	65.33	56.73
RestoreFormer [48]	0.683	0.522	0.402	0.542	0.596	0.576	0.541	0.572	39.85	64.98	66.53	55.87
CodeFormer [57]	0.709	0.557	0.444	0.576	0.634	0.621	0.599	0.618	43.98	63.09	63.96	56.05

Table 4. Comparison of 6 BFR and 4 VSR methods on **VFHQ-Test**(#50). **Red** and blue indicate the best and second best results. Top five results are marked as gray.

	PSNR↑	SSIM↑	LPIPS↓	SER-FIQ↑	MANIQA↑	FID↓	VIDD↓
GPEN [54]	26.232	0.766	0.368	0.674	0.580	89.85	0.41
GCFSR [14]	26.747	0.783	0.355	0.662	0.533	98.63	0.42
GFP-GAN [46]	26.547	0.776	0.360	0.670	0.551	95.08	0.42
VQFR [13]	25.645	0.760	0.365	0.677	0.547	91.98	0.58
RestoreFormer [48]	26.094	0.755	0.381	0.649	0.552	87.26	0.46
CodeFormer [57]	26.420	0.771	0.360	0.699	0.552	97.16	0.46
BasicVSR [3]	29.353	0.848	0.319	0.671	0.326	131.45	0.31
EDVR [45]	29.269	0.846	0.322	0.677	0.326	131.02	0.33
EDVR-GAN [45]	26.384	0.766	0.304	0.674	0.555	89.14	0.39
BasicVSR-GAN [3]	25.820	0.759	0.319	0.651	0.550	87.45	0.44

Table 5. Comparison of 6 BFR and 4 VSR methods on **FOS-V**(#3316). **Red** and blue indicate the best and second best performance. Top five results are marked as gray.

	SER-FIQ↑	MANIQA↑	FID↓	VIDD↓
GPEN [54]	0.596	0.639	79.21	0.51
GCFSR [14]	0.572	0.473	98.53	0.48
GFP-GAN [46]	0.601	0.515	96.50	0.48
VQFR [13]	0.596	0.514	85.60	0.62
RestoreFormer [48]	0.556	0.524	86.82	0.50
CodeFormer [57]	0.616	0.520	98.85	0.50
BasicVSR [3]	0.567	0.300	123.26	0.36
EDVR [45]	0.567	0.302	124.20	0.38
EDVR-GAN [45]	0.562	0.499	84.84	0.44
BasicVSR-GAN [3]	0.520	0.505	83.19	0.52

Table 6. The subjective scores achieved by 4 BFR methods on **FOS-real**(#158) for each subset. Point ≥ 3.5 is marked as red; point ≥ 3 is marked as blue.

	F.		O.		S.	
	Real.↑	Fidel.↑	Real.↑	Fidel.↑	Real.↑	Fidel.↑
CodeFormer [57]	3.64	3.64	3.43	3.47	3.29	3.31
RestoreFormer [48]	2.72	2.90	2.57	2.83	2.36	2.63
VQFR [13]	3.42	3.32	2.92	2.89	2.91	2.85
GFP-GAN [46]	3.05	3.22	2.83	3.04	3.00	3.23

The results of the user study are organized in Figure 6a and Table 6. The results of GPEN, GCFSR, EDVR, and EDVR-GAN are in the supplementary files due to the space limit.

From Table 6, CodeFormer achieves overall great subjective scores for both *realness* and *fidelity* on **FOS-real**(#158): $>$ 3.5 points on *full* subeset; $>$ 3.0 points on *occluded* and *side* subsets. These results indicate that CodeFormer yields superior reconstruction performance for *full* subset while showing robustness on other occasions.

While most methods obtain better results in *realness* than *fidelity*, VQFR shows an obvious advantage in *realness* over *fidelity*. Additionally, VQFR mainly performs well ($>$ 3.0) on *full* subset but less effective on *occluded* and *side* ($<$ 3.0). GFP-GAN exhibits robust performance by obtaining $>$ 3.0 point in most cases. As shown in Figure 6a, we find that the subjective score distribution of CodeFormer is more compact while that of RestoreFormer is rather flat, which implies their robust and less-than-robust performance on the test set, respectively. In general, the average score of all

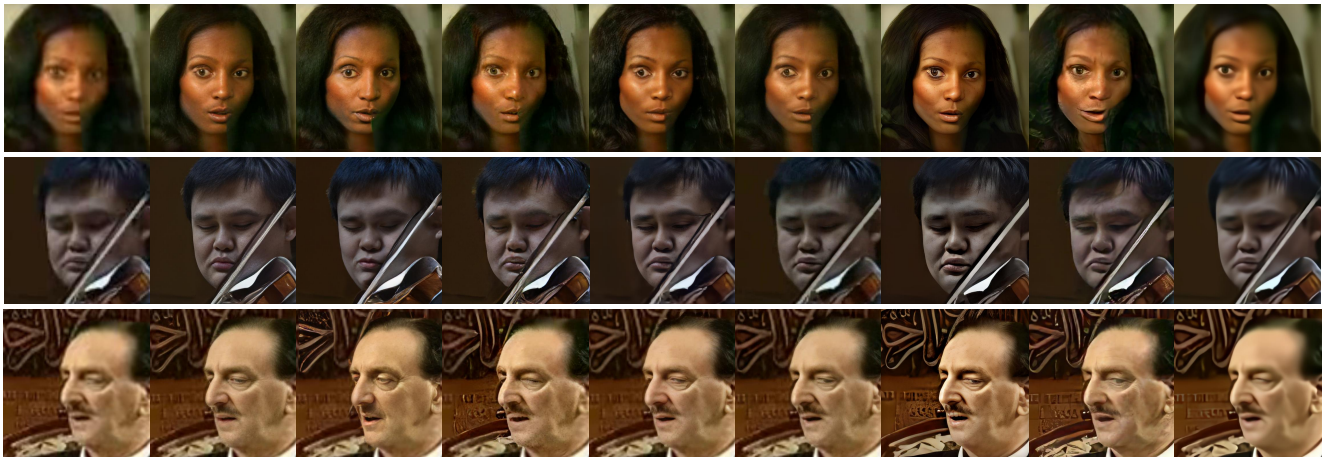
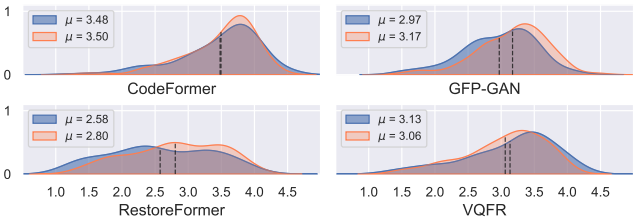
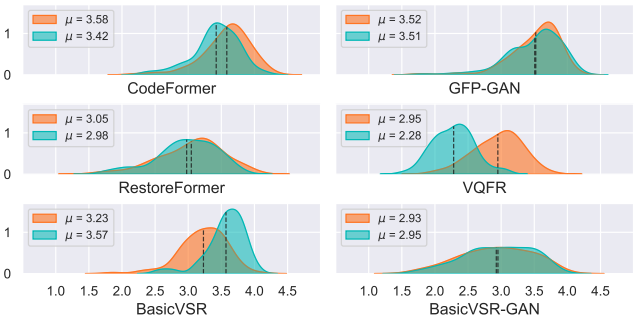


Figure 5. Qualitative comparison of both state-of-the-art BFR methods and VSR methods on **FOS-real**. (**Zoom in for details**)



(a) Subjective score distributions of 4 BFR methods on **FOS-real** (#158) regarding *realness* (blue) and *fidelity* (coral).



(b) Subjective score distributions on 4 BFR methods plus 2 VSR methods on **FOS-V** (#108) regarding *reconstruction performance* (orange) and *stability* (green).

Figure 6. 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.

methods is less than 3.5 points in *occluded* and *side* cases, indicating that the robustness of existing BFR methods still needs to be improved.

From Figure 6b, BasicVSR exhibits the highest *stability* performance (point 3.57) among all methods and shows a small variance in subjective scores. However, its reconstruction only achieves less than 3.3 points, indicating a gap between the restoration performance and stability maintenance. CodeFormer achieves the best reconstruction performance (3.58 points) but performs less well than BasicVSR in stability. Specifically, there exists a huge gap between the reconstruction and stability performance of VQFR. BFR methods can generate visually pleasant results but underperform in maintaining temporal consistency. This inspires us to combine the advantages of BFR and VSR methods to achieve more stable and realistic video face restoration.

5. Conclusion

In this work, we propose new benchmark datasets and present a comprehensive benchmark study of the state-of-the-art BFR and VSR methods. The established benchmark dataset **FOS** consists of face samples mainly from videos, which involve more complex scenarios than existing test datasets. The benchmarking results posed new challenges and identified the future direction of restoration on face videos. Based on our evaluation and analysis, we present the overall conclusions below, which we hope will shed some light on future FVR advances. 1) The current BFR methods have difficulty generalizing to cases in complex scenarios, such as faces with large pose movements or object occlusion. This generalization issue makes extending BFR methods to VFR solutions more challenging. 2) A more balanced trade-off is expected in future VFR solutions. None existing BFR or VSR methods produce high-quality faces while maintaining iter-frame stability when restoring video faces. 3) We narrow the gap between the current qualitative evaluation and the subjective scores. The effectiveness of the commonly used evaluation metrics is revisited and studied by leveraging a user study. The results show that NIQE and BRISQUE, are inconsistent with the human eye. Meanwhile, the recently proposed state-of-the-art IQA/FIQA metrics are potential candidates for face quality evaluation.

The collected FOS datasets may have some negative social impacts such as privacy leaking. To mitigate the influence of privacy, the data are derived from two public datasets and self-collected internet data which involves mainly celebrities. Any user acquires the datasets are required to follow the licence provided by [1].

References

[1] 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. 8

[2] Kelvin C.K. Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. *arXiv: Computer Vision and Pattern Recognition*, 2020. 5, 7

[3] Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. *Computer Vision and Pattern Recognition*, 2021. 3, 5, 7

[4] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K. Wong. Progressive semantic-aware style transformation for blind face restoration. *Cornell University - arXiv*, 2020. 5, 7

[5] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee Wong. Progressive semantic-aware style transformation for blind face restoration. 2023. 3

[6] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. *Cornell University - arXiv*, 2017. 3

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *Cornell University - arXiv*, 2018. 3

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 5

[9] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. *Computer Vision and Pattern Recognition*, 2019. 3

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1, 3

[11] Chaowei Fang, Guanbin Li, Xiaoguang Han, and Yizhou Yu. Self-enhanced convolutional network for facial video hallucination. *IEEE transactions on image processing*, 2020. 3

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[13] 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. 1, 3, 4, 5, 7

[14] 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. 1, 4, 5, 7

[15] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019. 2, 5

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3

[17] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 1, 3

[18] Byungho Jo, Donghyeon Cho, In Kyu Park, and Sungeun Hong. Ifqa: Interpretable face quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3444–3453, 2023. 5

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv: Neural and Evolutionary Computing*, 2017. 2, 3

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv: Neural and Evolutionary Computing*, 2018. 1, 3

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[22] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021. 1

[23] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 5

[24] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. *arXiv preprint arXiv:1908.08239*, 2019. 3

[25] 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. 3

[26] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. *Cornell University - arXiv*, 2018. 3

[27] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. *Springer International Publishing eBooks*, 2020. 3

[28] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *Computer*

Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, pages 399–415. Springer, 2020. 5

[29] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. 2020. 3

[30] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. 2022. 5, 6, 7

[31] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. 4

[32] Givi Meishvili, Simon Jenni, and Paolo Favaro. Learning to have an ear for face super-resolution. *arXiv: Computer Vision and Pattern Recognition*, 2019. 3

[33] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. 2020. 5, 7

[34] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708, 2012. 2, 5

[35] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 2, 5

[36] 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. 4

[37] Arsha Nagrani, Joon Son, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. 2023. 3

[38] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. Sdd-fiqa: unsupervised face image quality assessment with similarity distribution distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7670–7679, 2021. 2, 5

[39] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 1

[40] 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. 3

[41] 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. 4

[42] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Ge Xin, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. *Computer Vision and Pattern Recognition*, 2020. 5

[43] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5651–5660, 2020. 2, 3, 5, 6

[44] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1

[45] 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, 5, 7

[46] 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. 1, 3, 4, 5, 7

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[48] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. 2022. 1, 3, 4, 5, 7

[49] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011. 3

[50] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 2, 3, 4, 5

[51] Jingwei Xin, Nannan Wang, Jie Li, Xinbo Gao, and Zhifeng Li. Video face super-resolution with motion-adaptive feedback cell. *Proceedings of the ... AAAI Conference on Artificial Intelligence*, 2020. 3

[52] Lingbo Yang, Chang Liu, Pan Wang, Shanshe Wang, Peiran Ren, Siwei Ma, and Wen Gao. Hifacegan: Face renovation via collaborative suppression and replenishment. *arXiv: Computer Vision and Pattern Recognition*, 2020. 5, 7

[53] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 3, 5, 6

[54] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. *Cornell University - arXiv*, 2021. 1, 3, 4, 5, 7

[55] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European conference on computer vision (ECCV)*, pages 217–233, 2018. 3

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3, 5

[57] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. 2022. 1, 2, 3, 4, 5, 7

[58] Zihao Zou, Jiaming Liu, Shirin Shoushtari, Yubo Wang, Weijie Gan, and Ulugbek S Kamilov. Flair: A conditional diffusion framework with applications to face video restoration. *arXiv preprint arXiv:2311.15445*, 2023. 3