VFHQ: A High-Quality Dataset and Benchmark for Video Face Super-Resolution

Liangbin Xie* [1,2,3] Xintao Wang[3] Honglun Zhang[3] Chao Dong†[1] Ying Shan[3]

[1]Shenzhen Key Lab of Computer Vision and Pattern Recognition,
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3]ARC Lab, Tencent PCG

{lb.xie, chao.dong}@siat.ac.cn {xintaowang, honlanzhang, yingsshan}@tencent.com

Abstract

Most of the existing video face super-resolution (VFSR) methods are trained and evaluated on VoxCeleb1, which is designed specifically for speaker identification and the frames in this dataset are of low quality. As a consequence, the VFSR models trained on this dataset can not output visual-pleasing results. In this paper, we develop an automatic and scalable pipeline to collect a high-quality video face dataset (VFHQ), which contains over $16,000$ *high-fidelity clips of diverse interview scenarios. To verify the necessity of VFHQ, we further conduct experiments and demonstrate that VFSR models trained on our VFHQ dataset can generate results with sharper edges and finer textures than those trained on VoxCeleb1. In addition, we show that the temporal information plays a pivotal role in eliminating video consistency issues as well as further improving visual performance. Based on VFHQ, by analyzing the benchmarking study of several state-of-the-art algorithms under bicubic and blind settings.*

1. Introduction

As a special category of image super-resolution (SR) [12, 14, 30], face super-resolution (FSR) is an active research topic towards face-related applications, and has attracted increasing attention. Face super-resolution aims at restoring high resolution (HR) face images from low-resolution (LR) observations. Existing deep-learning-based methods mainly focus on exploiting the information of a single input image with the help of various priors, such as geometry facial priors [5,6,47], reference priors [11,27–29] or generative facial priors [41,46]. Thanks to the powerful capacity of convolutional neural networks (CNN) and the availability of high-quality face image datasets (*e.g.*, FFHQ [20]), some recent methods [5,41,46] can restore high-quality face images with the size up to 512×512 or even 1024×1024, from distorted face inputs.

Despite the rapid development of single-frame face SR, a few deep-learning-based methods [13, 19, 44] have tried to make progress for VFSR and their performance are even significantly inferior to the results of existing single face SR algorithms. We argue that it is the low quality of the training datasets that restrict the development of this field. The commonly-used dataset in VFSR is VoxCeleb1 [35] or VoxCeleb2 [7]. Though the image spatial size in those datasets can reach 800×800, the contents are blurry and have apparent video compression artifacts, as shown in the top of Fig. 2. Hence, algorithms trained with such datasets will inevitably retain those artifacts and are unable to generate high-quality details.

One may also want to directly apply single-frame face SR methods to videos. However, those approaches always lead to inconsistency among frames, which is a common issue in video applications [2, 23, 26]. Many works [23, 26] have shown that this inconsistency issue could be mitigated by training with multi-frame supervision. Moreover, exploiting multi-frame information will further improve the restoration performance [4, 40]. Therefore, it is highly desired to have a high-quality VFSR dataset. Constructing such a high-quality video face dataset is non-trivial work, as there are several complicated steps from the raw videos to the selected high-quality cropped face clips. In this work, we aim to establish an automatic and scalable pipeline to collect high-quality face clips from web videos. Based on this scalable pipeline, we have constructed the Video Face dataset with High Quality (VFHQ), which contains over 16,000 high-fidelity clips of diverse interview scenarios.

It is clear that the quality of VFHQ is superior to VoxCeleb1, as shown at the bottom of Fig. 2 and Fig. 4 (d). Besides, the clip resolution in VFHQ is between 700×700 and 1000×1000, which is close to the resolution of FFHQ images. To verify the necessity of VFHQ compared against VoxCeleb1 for video face SR, we train the BasicVSR [4], a state-of-the-art video SR method, on these two datasets respectively and compare their results accordingly. Equipped with VFHQ, BasicVSR can achieve more faithful results, and can restore more realistic textures with GAN training [15, 25], as shown in Fig. 1. We further experimentally show that directly applying single-frame face SR methods trained on FFHQ to restore distorted videos are sub-optimal. Instead, VFHQ not only contains high-fidelity details for

*Liangbin Xie is an intern in ARC Lab, Tencent PCG.
†Corresponding author.

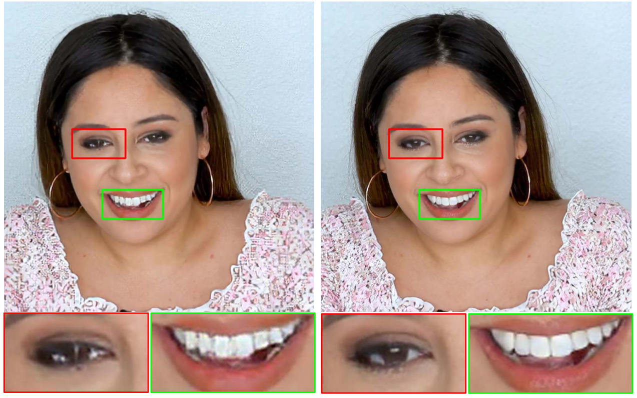


Figure 1. Visual comparison between BasicVSR-GAN models trained with Voxceleb1 and VFHQ dataset, respectively. The high-quality VFHQ dataset helps to recover more visual-pleasing results with finer details.

each individual frame, but also provides beneficial temporal information to promote better video consistency.

Based on the proposed VFHQ, we further conduct several benchmarking studies on the $\times 4$ bicubic and blind degradation settings. We summarize our contributions as follows. **1)** We establish an automatic and scalable pipeline to collect high-quality face clips from web videos, and construct a high-quality video face dataset VFHQ, which is superior to the commonly-used VoxCeleb1 in both quantitative and qualitative evaluation. **2)** We further verify the necessity of VFHQ compared against VoxCeleb1 and FFHQ. By using VFHQ, the recent works can achieve better performance than the same models trained on VoxCeleb1. Besides, compared against FFHQ, VFHQ can help to recover more details and mitigate the inconsistency issue in restored videos. **3)** Based on VFHQ, we evaluate several state-of-the-art methods in both bicubic and blind degradation settings to better understand the potential and limitations of those methods.

2. Related Work

Face super-resolution is an active problem in computer vision and can be divided into single-frame face super-resolution (SFSR) and video face super-resolution (VFSR).

This problem has been studied for a long time and please refer to [18] for a detailed survey. Different from single image SR methods [8, 12, 16, 21, 25, 31, 33, 38, 42, 48, 49] that directly learn a mapping from low-resolution image to their high-resolution counterpart, most SFSR methods attempt to integrate facial prior knowledge into the CNN architecture. There are three typical types of face-specific priors: geometry priors [5,6,43,47,50], reference-based priors [28,29] and generative facial priors [41, 46]. In contrast to the fast development of single-frame face SR, there are few attempts in VFSR [13, 34, 44] based on deep neural networks. All these methods focus on investigating the fusion of spatial and temporal information across frames or the fusion of aural and visual modalities. They do not consider the facial priors as SFSR does. Therefore, they are similar to general video super-resolution [4,40,45] except for the used dataset.

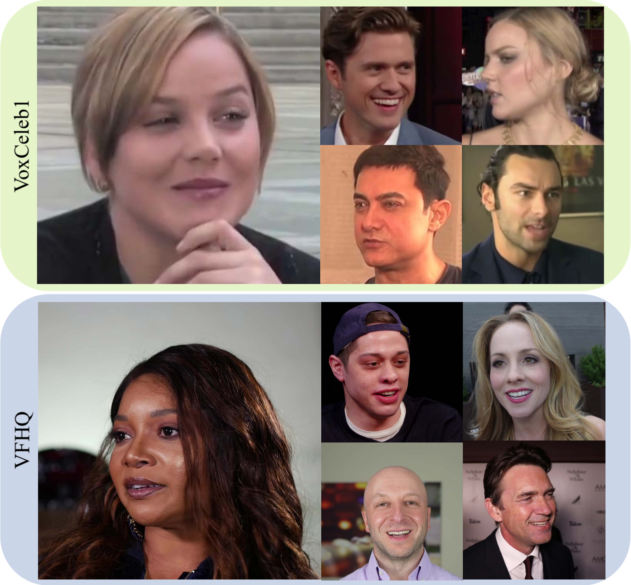


Figure 2. Visual comparisons between the two datasets: VoxCeleb1 (**top**) and VFHQ (**bottom**). Images are randomly selected from the dataset. VFHQ images have much higher quality. **Zoom in for best view**

The rapid development of the SFSR field can be partly attributed to the richness of image face datasets. There are several widely-used datasets for training and evaluating the SFSR methods, *e.g.*, Helen [24], CelebA [32], LFW [17], AFLW [22] and FFHQ [20]. Among them, FFHQ consists of $70,000$ high-quality images whose initial size exceeds 1024×1024. Based on the FFHQ dataset, some recent works [5, 41, 46] have achieved superior performance and can restore faces with faithful textures. Due to the low cost of taking high-definition face pictures and abundant online resources, it is easy to construct such high-quality image datasets without complicated pre-processing. In contrast to the abundant face image datasets, the most commonly used datasets in VFSR are VoxCeleb1 [35] and VoxCeleb2 [7]. Although these two datasets contain numerous utterances of celebrities, the resolution and quality of most videos are so poor that the models trained on these datasets do not have adequate ability to restore high-quality frames as SFSR methods. In order to fill the gap between the image face dataset and video face dataset, we propose a pipeline to extract high-quality face clips from web videos, and construct a high-quality video face dataset (VFHQ), which could promote the development of the VFSR field.

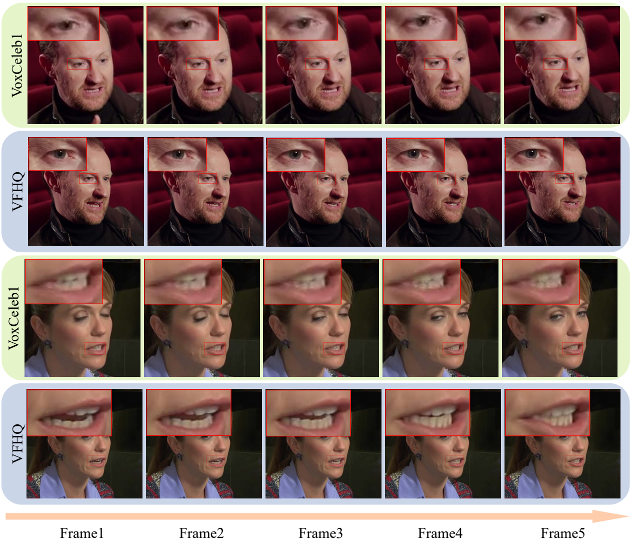


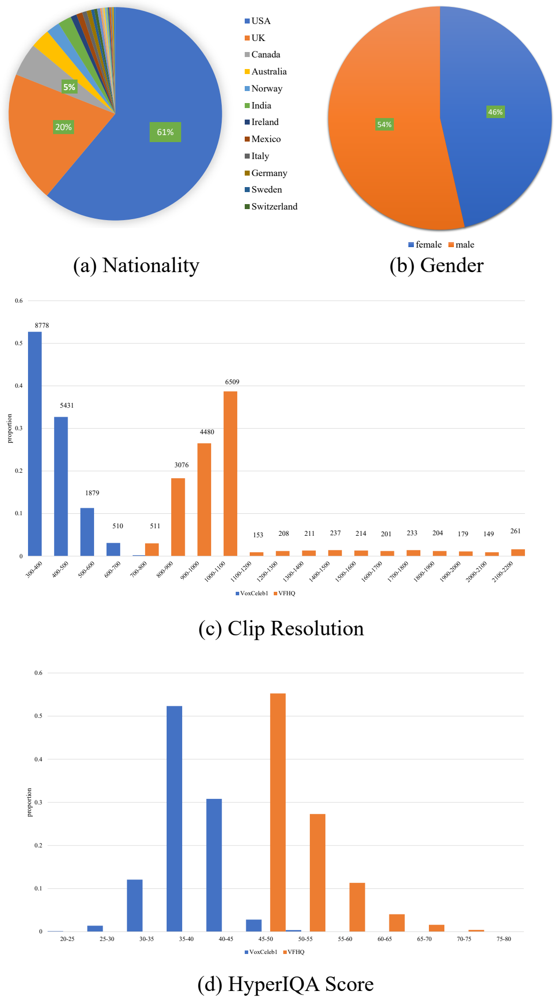
Figure 3. Visual comparison between VoxCeleb1 and VFHQ. For each dataset, we select five consecutive frames and the identity of selected videos is the same. The odd rows are the pictures of VoxCeleb1, while the even rows are the pictures of VFHQ. **Zoom in for best view**

3. Dataset Description

Following VoxCeleb1 [35], VFHQ is composed of clips for celebrities and is extracted from YouTube videos. The visual comparisons between VoxCeleb1 and VFHQ are shown in Fig. 2 and Fig. 3.

The pipeline adopted in collecting VoxCeleb1 and VoxCeleb2 is the same, which means the quality of these two datasets is nearly the same. Here, we only compare VFHQ with VoxCeleb1. In Fig. 2, we show several images randomly selected from VoxCeleb1. It can be observed that most of the frames in VoxCeleb1 are blurry and of low quality, while the face details in VFHQ are well preserved. We further select two sets of videos with the same identity from these two datasets, and show five consecutive frames within each video, as shown in Fig. 3. The frames in VFHQ retain relatively high quality across the whole video, while the frames in VoxCeleb1 are distorted with severe compression.

Moreover, we present the distribution of VFHQ celebrities in different aspects including nationality and gender. In our VFHQ celebrity list, we include persons that come from more than 20 distinct countries (Fig. 4 (a)). The proportion of men and women is roughly the same (Fig. 4 (b)). Compared with VoxCeleb1, the clip resolution of our VFHQ is much higher (Fig. 4 (c)). The hyperIQA (a blind image quality assessment (BIQA) method for authentically distorted images) [36] score of clips in VoxCeleb1 and VFHQ is shown in Fig. 4 (d), which quantitatively reflects the high-quality of VFHQ.



(a) Nationality

(b) Gender

(c) Clip Resolution

(d) HyperIQA Score

Figure 4. Distribution of the properties of the celebrities in our VFHQ in different aspects. As shown in (a), VFHQ includes persons that come from more than 20 distinct countries. In (b), we notice that the proportion of men and women is roughly the same. The figure (c) demonstrates that the distribution of clip resolution of our VFHQ is different from VoxCeleb1 and the resolution of VFHQ is much higher than VoxCeleb1. Above the bar is the number of clips. Note that we use the length of the shortest side as the clip resolution. The figure (d) shows that the quality of VFHQ is higher than VoxCeleb1 quantitatively.

4. Dataset Collection Pipeline

This section describes our multi-stage approach to collecting the VFHQ dataset, starting from YouTube videos. We adopt several CNN-based algorithms, including face detection (RetinaNet [10]), face recognition (ArcFace [9]), face alignment (AWing [39]), tracking (SORT [1]) and image quality assessment (HyperIQA [36]).

The pipeline involves 1) obtaining the raw videos from YouTube; 2) tracking faces by adopting RetinaNet and SORT algorithms; 3) confirming that the identity of each sub-video is the same by ArcFace; 4) selecting high quality sub-videos (top-three) within each video by calculating the assessment score and face landmark motions. Using this scalable pipeline, we have obtained $16,827$ video clips. We discuss the key stages in the following subsections.

4.1. Stage 1. Downloading videos from YouTube

Both VoxCeleb1 and VGGFace2 [3] provide a name list of celebrity, which contains $1,251$ and $9,131$ celebrities respectively. Based on these two lists, we crawl the corresponding videos from YouTube. Specifically, we append the word 'interview 4K' to the name of a celebrity in search query and download the top 20 videos for each celebrity.

4.2. Stage 2. Face tracking

For each frame within the video, we first use RetinaNet to detect face bounding boxes and filter out the detections with small sizes (less than 500×500). Then, all face detections are grouped together into face tracks by SORT. At this stage, we keep the tracks with the frame length between 100 and 2000.

4.3. Stage 3. Face verification

Based on the coarse tracks generated by the previous stage, we further refine the tracks to confirm that the detections in each track have the same identity. This is done by first using ArcFace to extract the feature of each detection and then calculating the L_2 similarity within every two features. The identities of two frames are considered to be different when the similarity is larger than a threshold 1.24. In this case, we will split a long clip into several short clips, in order to make sure that each frame within one short clip belongs to the same identity. At this stage, we also filter out clips that have less than 100 frames.

4.4. Stage 4. Selecting high-quality clips

For a clip that has not been filtered out, we are sure that it has a large spatial size (resolution), but we cannot guarantee its quality. Empirically, we find that HyperIQA [36] owns good generalization ability for face quality assessment in real scenes and we integrate it into our pipeline to help filter out low-quality clips.

Specifically, we first calculate the assessment score (the score of HyperIQA) AS_{frame} of each frame. The score ranges from 0 to 100 and the higher value represents better quality. Empirically, we compare it with the threshold 42. Once the assessment scores of more than four consecutive frames are less than this threshold, we discard these frames and divide the clip into two clips. After that, we calculate the average assessment score AS_{clip} of each clip and compare it with the overall threshold (we empirically set it to 45). The clips of which the average score is less than this threshold are discarded.

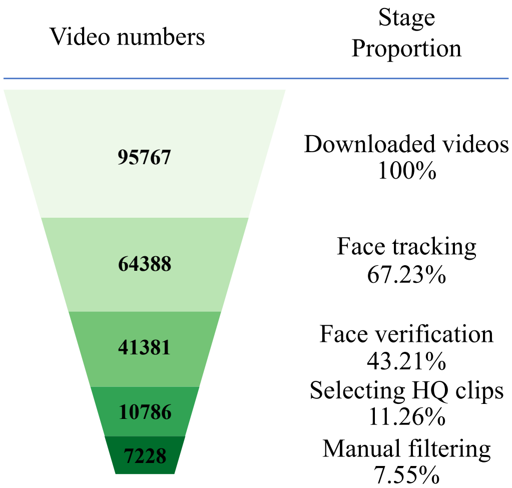


Figure 5. The number and proportion of remained videos after each stage. "Video number" means that the number of full videos crawled from YouTube. "Video clip number" means that the number of clips split from full videos. VFHQ includes $16,827$ clips from $7,228$ videos.

After the procedure of the above two steps, we find that some videos have a large number of clips that meet the requirements, while others have only one or two. To increase the dataset diversity and eliminate the imbalance, we finally select top-three high-quality clips for each video by considering both the assessment scores AS_{clip} and landmark motion M_{clip}:

$$Score_{clip} = \alpha AS_{clip} + \beta \hat{M}_{clip}, \tag{1}$$

The landmark motion M_{clip} is calculated on the 98 landmark points:

$$M_{clip} = \frac{1}{N \times 98} \sum_{i=1}^{N-1} \|\mathcal{L}_{i+1} - \mathcal{L}_i\|^2, \tag{2}$$

where $\mathcal{L}_i$ and $\mathcal{L}_{i+1}$ are the 98 landmark results of the i and $i+1$ frames. N is the total frame number in each clip. We further normalize M_{clip} by:

$$\hat{M}_{clip} = 0.25 M_{clip} + 42.5. \tag{3}$$

Empirically, we set α and β to 0.5 and 0.2, respectively, to balance their importance.

4.5. Stage 5. Manual filtering

Ideally, the proposed automatic pipeline can filter out all clips with distortion. However, there exists the generalization problem for HyperIQA and we need to manually verify the qualify of remained clips. Compared to directly manual filtering the remained clips (frame by frame) of stage 3, the amount of clips that need to be processed is greatly reduced and the process of verification takes less time. In practice, we uniformly select five frames for each clip and check their quality. The clip will be discarded when all five frames are obviously of low quality.

From stage 2 to stage 5, we discard the unsatisfied videos steadily. Fig. 5 shows the proportion of remaining videos after each stage. We have crawled a total of $95,767$ raw videos and finally obtain $16,827$ clips from $7,228$ videos. The percentage of final remaining high-quality videos is about 7.55%.

The diversity of motion is considered during collection and VFHQ clips can be categorized into 3 categories according to their motions. We calculate the average pixel displacement of each clip to perform such division. The ratio of large motion, middle motion and slow motion are 23.6%, 32.2% and 44.2%, respectively.

5. The necessity of VFHQ

The intuitive difference between VFHQ and the other two datasets (i.e, VoxCeleb1 and FFHQ) is that the quality of VFHQ is superior to VoxCeleb1 and FFHQ lacks temporal information. However, the effectiveness of VFHQ is still unclear. Hence, we further investigate two questions in the following section.

1. *The necessity of our proposed VFHQ compared against VoxCeleb1.* We verify this in the following experiments from two facets: 1) Is VFHQ a more suitable dataset for evaluating existing algorithms? 2) Will training on VFHQ improves the visual quality, for both the MSE-based and GAN-based methods?

2. *The necessity of our proposed VFHQ compared against FFHQ.* We verify this in the following experiments from two aspects: 1) How does the quality of VFHQ compared against FFHQ? 2) Will utilizing the temporal information help to relieve the video consistency issue and further enhance visual quality?

5.1. Experiment Settings

We compare different methods on three datasets: FFHQ, VoxCeleb1 and our proposed VFHQ. Specifically, we choose the representative image SR method – ESRGAN [42], the state-of-the-art video SR method – BasicVSR [4] and our implemented BasicVSR-GAN. The details of these methods and more experiments with other methods (RRDB, EDVR, EDVR-GAN) can be found in the supplementary materials.

Recent works [5,41,46] are focusing on restoring or generating high-quality faces whose sizes are up to 512×512. Following GFPGAN [41], we resize all the images to 512×512 as HR images. All experiments in this section are performed with a scaling factor of $\times 4$ between LR and HR images/frames. The corresponding LR images/frames are obtained by down-sampling the corresponding HR images/frames with the MATLAB bicubic kernel.

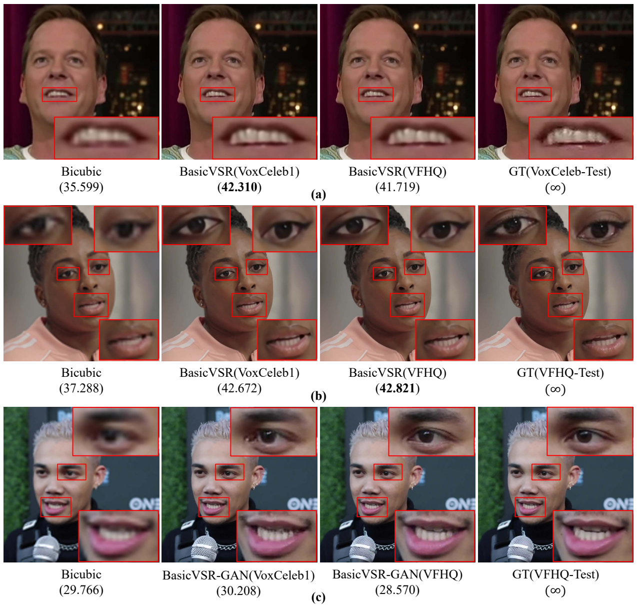


Figure 6. Qualitative comparisons of BasicVSR models trained with VoxCeleb1 and with VFHQ datasets. ((a) evaluated on VoxCeleb-Test, (b) evaluated on VFHQ-Test). (c) Qualitative comparisons of BasicVSR-GAN models trained with VoxCeleb1 and with VFHQ datasets. **Zoom in for best view**

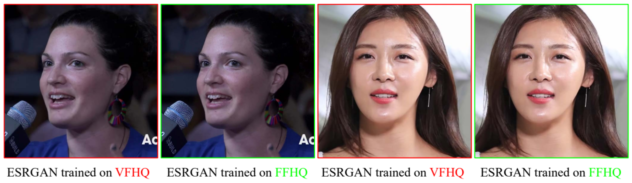


Figure 7. Visual comparisons between ESRGAN models trained with VFHQ and with FFHQ dataset, respectively. Both models restore similar details in face components. **Zoom in for best view**

For better evaluation, we construct two testing datasets, VoxCeleb1-Test and VFHQ-Test. VoxCeleb1-Test contains 20 sequences that are randomly selected from VoxCeleb2 [7]. VFHQ-Test is composed of 50 sequences that are randomly selected from VFHQ. Note that these two testing datasets have no overlap with their corresponding training datasets. All other clips that are not included in these two test datasets respectively construct two corresponding training datasets.

5.2. Comparisons with VoxCeleb1

VFHQ is proposed as a supplement to VoxCeleb1 and we hope VFHQ can be a new dataset for face SR. From Fig. 2 and Fig. 4, we can observe that VFHQ is superior to VoxCeleb1 in two facets: image quality and resolution distribution. To further explore the necessity of VFHQ for face SR, we train BasicVSR based on these two datasets and evaluate their results in VoxCeleb1-Test and VFHQ-Test.

Table 1. Quantitative results on VoxCeleb1-Test and VFHQ-Test. Trained on BasicVSR.

Methods	Training Datasets	VoxCeleb-Test		VFHQ-Test	
		PSNR	SSIM	PSNR	SSIM
BasicVSR	VoxCeleb1	43.367	0.9829	36.064	0.9410
	VFHQ	42.760	0.9817	36.399	0.9429

Table 2. Quantitative results with different training input frames for BasicVSR. Evaluated on VFHQ-Test."Length" indicates the input frame length of network during the training phase.

Method	Length	PSNR (dB)	SSIM
BasicVSR	L=1	35.213	0.9293
	L=7	36.258 (+1.045)	0.9412 (+0.0119)

The quantitative evaluation can be found in Tab. 1. Due to the difference in the distribution of VFHQ and VoxCeleb1, for a specific testing dataset, the model trained on the corresponding training dataset achieves the better performance on PSNR/SSIM metrics. However, from visual comparisons in Fig. 6 (a), we can find that: The quality of Ground-Truth (GT) in VoxCeleb1-Test is blurry with distortions. As PSNR is a pixel-wise metric, the restored facial components with lower quality may get a higher value. This phenomenon indicates that VoxCeleb1 dataset is not suitable for making paired test dataset to evaluate the performance of existing methods. Since many works [13, 44] evaluate their proposed methods based on the paired test dataset generated by VoxCeleb, we think that a high-quality test dataset to better evaluate existing algorithms is in urgent need. In Fig. 6 (b), it is clear that BasicVSR trained with VFHQ recovers more faithful details in the eyes than BasicVSR trained with VoxCeleb1.

For restoration task, GAN is a common technique for generating more realistic images. Therefore, based on VoxCeleb1 and VFHQ, we further fine-tune their corresponding BasicVSR, obtaining BasicVSR-GAN. As shown in Fig. 6 (c), BasicVSR-GAN trained with VoxCeleb1 fails to retain the fidelity of teeth and tends to generate artifacts, while BasicVSR-GAN trained with VFHQ obtains better teeth shape. Besides, when trained with VFHQ, BasicVSR-GAN is capable of recovering faithful details in the eyes.

In summary, compared against VoxCeleb1, the necessity of VFHQ reflects in two aspects. 1) It is a suitable dataset for evaluating existing algorithms, which can further promote researchers to propose better algorithms with better visual effects. 2) For an algorithm (e.g, BasicVSR), when trained with VFHQ rather than VoxCeleb1, the algorithm can restore more realistic textures. This phenomenon is more obvious when the algorithm is trained with GAN.

5.3. Comparisons with FFHQ

FFHQ [20] is a high-quality image dataset of human faces. Since both VFHQ and FFHQ are high-definition face datasets, we wonder how does the quality of VFHQ compare to FFHQ. To clarify this, we trained two ESRGAN models with VFHQ and with FFHQ datasets, respectively. The visual comparisons show in Fig. 7. We can observe that the ESRGAN model trained with VFHQ dataset can restore similar details in face components as the ESRGAN model trained on FFHQ dataset, reflecting the high-quality of our collected VFHQ.

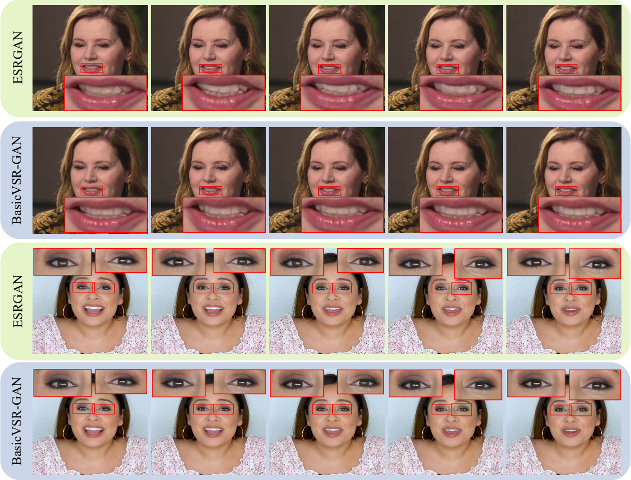


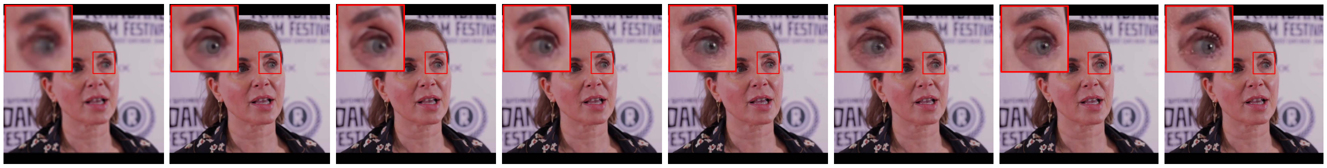
Figure 8. Qualitative comparisons between BasicVSR-GAN and ESRGAN. We present the result of five consecutive frames. **Top**: BasicVSR-GAN can restore the complete tooth shapes while ESRGAN mixes all the teeth together. **Bottom**: the bright spots in eyes keep changing for different frames in ESRGAN, while are consistent in BasicVSR-GAN. **Zoom in for best view**

In many scenarios that require face SR, directly applying single-frame face SR methods is an option. However, it is a sub-optimal option since it ignores the temporal information in the videos. By utilizing temporal information, there are two benefits. 1) It helps achieve better results by considering complementary information between adjacent frames. 2) It can mitigate inconsistency issues in restored videos.

To clarify this, we apply ESRGAN [42] trained with FFHQ [20] to VFHQ-Test and compare the results with BasicVSR-GAN, as shown in Fig. 8. We can draw the following observations. **1)** For facial components like teeth (1st and 2nd row), BasicVSR-GAN can restore the complete tooth shapes while ESRGAN mixes all the teeth together. **2)** Since ESRGAN is a single-frame method and do not consider the information among consecutive frames, each frame in the restored video is independent of each other. Although the motion contained in these frames are small, ESRGAN still leads to obviously pixel jittering. Specifically, the shape of teeth (top) and the bright spots in eyes (bottom) keep changing among the restored five frames. **3)** Exploiting temporal information in VFHQ is effective to eliminate the inconsistency and improve both the qualitative and quantitative performances.

Table 3. Benchmarking results with **bicubic** degradation model (evaluated on VFHQ-Test). Average PSNR/SSIM values for scaling factor ×4. **Red** and blue indicates the best and second best performance. The sampling interval in the testing phase is equal to 5.

Interval	Metrics	MSE-based				GAN-based		
		Bicubic	RRDB	EDVRM	BasicVSR	ESRGAN	EDVRM-GAN	BasicVSR-GAN
5	PSNR	31.964	35.332	36.090	**36.258**	32.803	33.592	32.327
	SSIM	0.8939	0.9302	0.9399	**0.9412**	0.8961	0.9089	0.8869



Bicubic RRDB EDVRM BasicVSR ESRGAN EDVRM-GAN BasicVSR-GAN GT

Figure 9. Qualitative comparisons by different models in ×4 bicubic degradation setting. **Zoom in for best view**

We also conduct comparisons under the same network architectures to validate the effectiveness of temporal information in improving performance. As shown in Tab 2, on the VFHQ-Test dataset, BasicVSR with seven input frame length (L=7) outperforms BasicVSR with only one frame (L=1, with equivalent computation FLOPs) by a large margin. It indicates that multi-frame temporal information is pivotal for improving the restoration performance of face videos.

In summary, the quality of VFHQ is comparable with FFHQ. For restoring distorted videos (especially with large motion), compared to FFHQ, the temporal information in VFHQ is pivotal for relieving the video consistency issue and improving the visual quality of restored videos.

6. Benchmark Experiments

6.1. Degradations

To comprehensively evaluate existing methods on VFHQ, we select two degradation models, the bicubic degradation model and the blind degradation model. The first one is classical in super-resolution and the second one is closer to real-world degradation. Details of these two degradations are described as follows.

Bicubic degradation model is implemented by adopting the Matlab function *imresize*. The downsample scale is ×4.

Blind degradation model [27, 41] is implemented by following the practice in [41]. Considering the compression type in image and video datasets is different, we use *FFMPEG* rather than JPEG to simulate the compression. To be specific, the degradation model is:

$$\boldsymbol{x} = [(\boldsymbol{y} \circledast \boldsymbol{k}_\sigma) \downarrow_r + \boldsymbol{n}_\delta]_{\mathrm{FFMPEG}_{\mathrm{crf}}} \tag{4}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are paired low-resolution and high-resolution clips. The $\boldsymbol{k}_\sigma$, r, $\boldsymbol{n}_\delta$ and crf are Gaussian blur kernel, down-sampler factor, additive white Gaussian noise, constant rate factor (decides how many bits will be used for each frame), respectively. The r equals 4 in experiments. The sampling range of σ, δ and crf are $\{0.1 : 10\}$, $\{0 : 10\}$, $\{18 : 25\}$, respectively. Note that for each individual training pair, we only sample one value for σ and crf, whereas the δ varies among frames in the clip by following [37].

6.2. Comparison in Bicubic Degradation

We conduct experiments with the MSE-based and GAN-based methods. Specifically, for MSE-based methods, we select RRDB [42], EDVRM [40], BasicVSR [4]. For GAN-based methods, we select ESRGAN [42] and EDVRM-GAN and BasicVSR-GAN, which are fine-tuned based on their corresponding PSNR-oriented models with generative adversarial loss. In the training phase, to increase the motion range, we interval sample a continuous video and input the newly composed video to the network. To be specifically, the sampling interval is $\{3 : 7\}$. In the testing phase, we also evaluate the performance of difference sampling intervals of test datasets. Here we only show the results whose sampling interval is equal to 5. Results of other intervals and more visual comparison among these methods can be found in the supplementary materials.

Tab. 3 shows a quantitative comparison between these methods. Consistent with the performance in the general video super-resolution field, BasisVSR achieves the best performance in PSNR and SSIM metrics. The visual comparison of these methods is shown in Fig. 9, for the current test image, we can find that ESRGAN, EDVRM-GAN and BasicVSR-GAN can restore faithful facial details. This indicates that for video face super-resolution task, specifically in ×4 bicubic degradation setting, current methods are capable of restoring high-quality face videos. In a larger scale ratio (e.g, ×8), the performance gap between these methods is larger and there needs more investigation for a larger scale ratio in the bicubic setting. Experiments for ×8 scale can be found in the supplementary materials.

Table 4. Benchmarking results with **blind** degradation model (evaluated on VFHQ-Test). Average PSNR/SSIM/LPIPS values for scaling factor ×4. **Red** and blue indicates the best and second best performance. The sampling interval in the testing phase is equal to 5.

Interval	Metrics	MSE-based			GAN-based			GAN-prior based	
		Bicubic	EDVRM	BasicVSR	EDVRM-GAN	BasicVSR-GAN	DFDNet	GFPGAN	GPEN
5	PSNR	26.842	29.457	**29.472**	26.682	25.813	25.178	25.978	26.672
	SSIM	0.7909	0.8428	**0.8430**	0.7638	0.741	0.7560	0.7723	0.7768
	LPIPS	0.4098	0.3288	0.3309	**0.3076**	0.3214	0.4008	0.3446	0.3607

Table 5. Quantitative results of combining MSE-based method and GAN-prior based method. Evaluated on VFHQ-Test. The sampling interval in the testing phase is equal to 5.

Interval	Metrics	EDVRM+GFPGAN	BasicVSR+GFPGAN
5	PSNR	27.879	27.868
	SSIM	0.8198	0.8195
	LPIPS	0.3265	0.3266

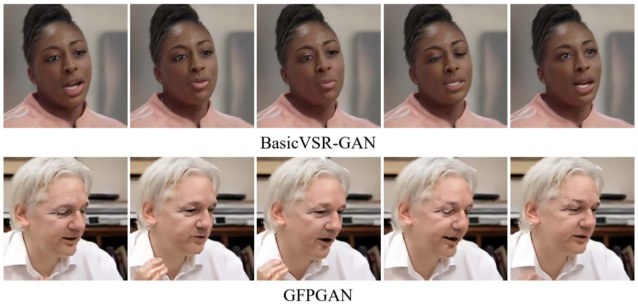
6.3. Comparison in Blind Degradation

Similar to the benchmarking study conducted in the bicubic degradation setting, we evaluate the MSE-based and GAN-based methods in the blind degradation setting. Considering that recent GAN-prior based methods [41, 46] and DFDNet [27] can restore realistic faces on both synthetic and real-world datasets, we also include those methods for comparison. Here, the testing datasets are synthesized based on the same degradation model used in the training pairs. For these three algorithms, we directly apply their released pre-trained models to distorted videos. We also show the restored results of which the sampling interval is equal to 5.

The quantitative results are listed in Tab. 4. We find that in the blind degradation setting, the gap between EDVR and BasicVSR on PSNR/SSIM metrics is smaller than the bicubic degradation. For LPIPS metric, we only evaluate the performance of five frames within each restored sequence and EDVRM-GAN achieves the best performance among these methods.

The strategy of applying both MSE-based method and GAN-prior method to restore distorted sequence is also adopted and the results are listed in Tab 5. Although the performance of EDVRM+GFPGAN and BasicVSR+GFPGAN is better than GFPGAN on LPIPS metric, their performance is inferior to their corresponding GAN-based methods. It indicates that end-to-end training is a better strategy. The design of combining MSE-based methods and GAN-prior based methods into a unified network is left as our future work.

Unlike the bicubic degradation setting, existing methods have limitations in the blind setting, as shown in Fig. 10. Specifically, for BasicVSR-GAN, although with end-to-end training, it can not restore realistic faces when the degradation of the input video is relatively severe (still in the range of training data distribution). For GFPGAN, it produces unnatural results for very large poses. Since it only takes the corresponding distorted face as the input, there exists obvious inconsistency in the restored videos. More visual results are shown in the supplementary materials.


BasicVSR-GAN

GFPGAN

Figure 10. Limitations of BasicVSR-GAN and GFPGAN.

7. Conclusion

Compared against high-quality face image datasets, the poor quality of training and testing video face datasets has restricted the development of multi-frame face SR research. To fill the gap between the image face dataset and video face dataset, we propose an automatic and scalable pipeline to collect high-quality face clips from web videos, and construct a Video Face dataset with High Quality (VFHQ). Based on VFHQ, we further reveal its importance for multi-frame face SR by exploring the necessity of VFHQ compared to VoxCeleb1 and FFHQ. In addition, we conduct benchmarking studies in bicubic and blind settings. Future work includes the investigation of generative facial priors in multi-frame face SR, with the help of VFHQ.

The proposed VFHQ may have some negative social impacts, like leaking privacy. To mitigate the influence of privacy, the selected identities are celebrities and the celebrity list comes from two public datasets [3, 35]. Users are required to read the license file provided by [3] carefully before downloading the data. We sincerely hope the collected VFHQ can promote the development of face-related applications.

Acknowledgement. This work is partially supported by the National Natural Science Foundation of China (61906184), the Joint Lab of CAS-HK, the Shenzhen Research Program (RCJC20200714114557087), the Shanghai Committee of Science and Technology, China (Grant No. 21DZ1100100).

References

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*. IEEE, 2016. 3

[2] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *TOG*, 2015. 1

[3] 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. 4, 8

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *CVPR*, 2021. 1, 2, 5, 7

[5] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *CVPR*, 2021. 1, 2, 5

[6] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*, 2018. 1, 2

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 1, 2, 5

[8] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019. 2

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 3

[10] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 3

[11] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *CVPRW*, 2019. 1

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*. Springer, 2014. 1, 2

[13] Chaowei Fang, Guanbin Li, Xiaoguang Han, and Yizhou Yu. Self-enhanced convolutional network for facial video hallucination. *TIP*, 2019. 1, 2, 6

[14] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*. IEEE, 2009. 1

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NIPS*, 2014. 1

[16] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018. 2

[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. 2

[18] Junjun Jiang, Chenyang Wang, Xianming Liu, and Jiayi Ma. Deep learning-based face super-resolution: A survey. *arXiv preprint arXiv:2101.03749*, 2021. 2

[19] Yonggang Jin and Christos-Savvas Bouganis. Robust multi-image based blind face hallucination. In *CVPR*, 2015. 1

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 6

[21] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 2

[22] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCVW*. IEEE, 2011. 2

[23] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 1

[24] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *ECCV*. Springer, 2012. 2

[25] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 1, 2

[26] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *NIPS*, 2020. 1

[27] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*. Springer, 2020. 1, 7, 8

[28] 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. In *CVPR*, 2020. 1, 2

[29] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *ECCV*, 2018. 1, 2

[30] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 1

[31] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 2

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 2

[33] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR*, 2020. 2

[34] Givi Meishvili, Simon Jenni, and Paolo Favaro. Learning to have an ear for face super-resolution. In *CVPR*, 2020. 2

[35] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 1, 2, 3, 8

[36] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *CVPR*, 2020. 3, 4

[37] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1354–1363, 2020. 7

[38] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 2

[39] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *ICCV*, 2019. 3

[40] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, 2019. 1, 2, 7

[41] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 1, 2, 5, 7, 8

[42] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 2, 5, 6, 7

[43] Jingwei Xin, Nannan Wang, Xinbo Gao, and Jie Li. Residual attribute attention network for face image super-resolution. In *AAAI*, 2019. 2

[44] Jingwei Xin, Nannan Wang, Jie Li, Xinbo Gao, and Zhifeng Li. Video face super-resolution with motion-adaptive feedback cell. In *AAAI*, 2020. 1, 2, 6

[45] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 2019. 2

[46] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, 2021. 1, 2, 5, 8

[47] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *ECCV*, 2018. 1, 2

[48] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *CVPR*, 2020. 2

[49] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2

[50] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *ECCV*. Springer, 2016. 2