

Audio-Assisted Face Video Restoration with Temporal and Identity Complementary Learning

Yuqin Cao¹, Yixuan Gao¹, Wei Sun², Xiaohong Liu¹, Yulun Zhang¹, Xiongkuo Min^{1*}

¹Shanghai Jiao Tong University

²East China Normal University

{caoyuqin, gaoyixuan, xiaohongliu, yulzhang, minxiongkuo}@sjtu.edu.cn, wsun@cee.ecnu.edu.cn

Abstract

Face videos accompanied by audio have become integral to our daily lives, while they often suffer from complex degradations. Most face video restoration methods neglect the intrinsic correlations between the visual and audio features, especially in mouth regions. A few audio-aided face video restoration methods have been proposed, but they only focus on compression artifact removal. In this paper, we propose a **General Audio-assisted face Video restoration Network (GAVN)** to address various types of streaming video distortions via identity and temporal complementary learning. Specifically, GAVN first captures inter-frame temporal features in the low-resolution space to restore frames coarsely and save computational cost. Then, GAVN extracts intra-frame identity features in the high-resolution space with the assistance of audio signals and face landmarks to restore more facial details. Finally, the reconstruction module integrates temporal features and identity features to generate high-quality face videos. Experimental results demonstrate that GAVN outperforms the existing state-of-the-art methods on face video compression artifact removal, deblurring, and super-resolution.

Introduction

Audio-visual (A/V) streaming services continue to grow in popularity and are rapidly becoming a crucial source of information in our daily lives. In A/V content, the speaker’s voice often garners the most attention, leading viewers to instinctively focus on the speaker’s face. However, in real-world scenarios, face videos often suffer from complex degradations. Face video restoration methods aim to restore degraded face videos to the authentic, high-quality and reliable ones. This not only improves users’ quality of experience (QoE) but also contributes to advancements in video compression technologies and supports visual tasks such as face recognition (Kong et al. 2021; Lau, Castillo, and Chellappa 2021; Zhang et al. 2011), privacy protection (Yu et al. 2016), and autonomous driving (Chen et al. 2015).

In the literature, most face restoration studies (Wang et al. 2019; Chan et al. 2022; Wang et al. 2023; Yang et al. 2021) only consider the image/video mode, while ignoring



Figure 1: Restoration results of our proposed GAVN with and without identity features and audio signals. The frames reconstructed without using identity features and audio signals appear blurry in the left eye and tongue, respectively.

the crucial impact of audio. The motivation of audio-aided face video restoration is the high correlation between visual and audio features, particularly the synchronization between audio and lip movements in face videos. Physiologically, sound production involves shaping it with facial muscles, particularly those in the lips that control airflow. Therefore, audio can significantly improve face video restoration quality whilst consuming minimal storage space. Moreover, the restored face video should strive to maintain the original appearance under various types of face actions. This is crucial since the face is highly structured and frequently used for personal information identification. It inspires us to leverage facial identity features to aid in video restoration.

Some researchers (Zhang et al. 2020; Guo, Zhang, and Wu 2020; Zhang and Wu 2022) have explored audio-assisted compressed face video restoration. However, there are still some issues that need to be addressed. Firstly, the existing audio-assisted face video restoration methods are exclusively designed to restore compressed face videos, neglecting the potential of utilizing audio for the restoration of other common distortion types. We propose a novel network, called **General Audio-assisted face Video Restoration Network (GAVN)** for the face video compression artifact removal, deblurring, and super-resolution. Secondly, in previous video restoration methods (Wang et al. 2019; Zhou et al. 2022; Li et al. 2021; Guo et al. 2022), input frames are often downsampled to align and capture features in the low-resolution space. Though this approach significantly reduces the computational cost, it results in the loss of facial detail features. In contrast, our GAVN extracts temporal features from multiple consecutive frames in the

*Corresponding Author

low-resolution space and captures identity features with the assistance of audio from individual concurrent frames in the high-resolution space. This allows us to preserve facial details whilst achieving computational efficiency. As illustrated in Fig. 1, both identity features and audio signals contribute to the restoration of facial details. Thirdly, incorporating face landmark features can further keep the identification information of the restored face (Chen et al. 2018; Li et al. 2018; Bulat and Tzimiropoulos 2018). Conventional face landmark detection methods are trained on high-quality face images and encounter difficulties in accurately detecting landmarks from low-quality face images. Therefore, we retrain the landmark detection model to precisely extract landmark features from low-quality face frames with the assistance of audio.

In this paper, we make three contributions to the face video restoration field.

- We propose GAVN for the face video compression artifact removal, deblurring, and super-resolution.
- Our GAVN integrates inter-frame temporal features and intra-frame identity features to restore face videos. With the assistance of audio and identity features, our GAVN outperforms the state-of-the-art (SOTA) video restoration methods.
- We conduct experimental analysis to systematically evaluate GAVN, including two scenarios: one involving various speakers with different identities and another involving a specific known speaker.

Related Work

Deep Video Restoration

Most video restoration methods are based on convolutional neural networks (CNNs), which can be divided into two categories: sliding window-based and recurrent-based methods. Sliding window-based methods (Wang et al. 2019; Isobe et al. 2020; Li et al. 2021, 2020) typically take a small segment of video frames as input and predict the center frame. Recurrent-based methods (Haris, Shakhnarovich, and Ukita 2019; Sajjadi, Vemulapalli, and Brown 2018; Chan et al. 2021, 2022) mainly use previously reconstructed high-quality frames or features for subsequent frame reconstruction. Some researcher utilize the recurrent structure to handle input of various lengths and capture long-term temporal information. For instance, Sajjadi *et al.* (Sajjadi, Vemulapalli, and Brown 2018) proposed a recurrent approach that integrates the previously estimated high-quality frame and the current low-quality frame to predict the current frame. Chan *et al.* (Chan et al. 2021) further utilized the bidirectional recurrent network and expanded it to grid propagation in (Chan et al. 2022).

Encouraged by the above research, our GAVN incorporates a recurrent structure to extract inter-frame temporal features, thereby mitigating quality fluctuations across frames. Different from the above methods, GAVN further leverages intra-frame identity features and audio features of the current frames to enhance facial details. Their combination outperforms the performance achieved by individual temporal features.

Temporal Alignment and Fusion

Due to the motions of the camera and object, the adjacent frames are often misaligned. Currently, many researchers (Wang et al. 2019; Tian et al. 2020; Luo et al. 2021) utilize deformable convolutions to align the neighboring frames to the reference frames. For example, Tian *et al.* (Tian et al. 2020) proposed TDAN that utilizes deformable convolution to align the reference frame and each supporting frame at the feature level without computing optical flow. Wang *et al.* (Wang et al. 2019) proposed the PCD module which extends TDAN to multi-scale alignment and performs alignment in a coarse-to-fine manner. In this paper, we take a step further by employing deformable convolution to align adjacent frames and skip frames, both forward and backward in time. This allows us to aggregate temporal features from different spatio-temporal locations and directions to obtain more comprehensive and abundant temporal features.

Methodology

Overview

The overall framework of the proposed GAVN is illustrated in Fig. 2. GAVN is a general architecture suitable for various face video restoration tasks, including face video compression artifact removal, deblurring, and super-resolution. GAVN takes $2N + 5$ consecutive low-quality frames $\mathbf{X}_{t\pm(N+2)} = \{X_{t-N-2}, \dots, X_{t+N+2}\}$ and the corresponding audio segments $\mathbf{A}_{t\pm m}$ as inputs to predict the high-quality frames $\hat{\mathbf{X}}_{t\pm N}$, which closely resemble the ground truth frames $\mathbf{Y}_{t\pm N}$. For the super-resolution task, low-quality frames are first upsampled to the same resolution as the ground truth frames through Bicubic interpolation.

GAVN consists of three modules: the inter-frame temporal module, the intra-frame identity module, and the reconstruction module. GAVN first utilizes the inter-frame temporal module to extract temporal features from low-quality frames $\mathbf{X}_{t\pm(N+2)}$. It utilizes deformable convolutions for frame alignment forward and backward in time. The inter-frame temporal module can be formulated as:

$$\mathbf{T}_{t\pm N} = \mathbf{G}_{\text{inter-frame}}(\mathbf{X}_{t\pm(N+2)}), \quad (1)$$

where $\mathbf{T}_{t\pm N}$ denotes as the temporal features of video frames $\mathbf{X}_{t\pm N}$ and $\mathbf{G}_{\text{inter-frame}}$ is the inter-frame temporal module. Since the audio is highly correlated with the movement of the mouth regions, the intra-frame identity module utilizes the current frame and the corresponding audio segments to extract identity features, that is:

$$\mathbf{I}_{t\pm N} = \mathbf{G}_{\text{intra-frame}}(\mathbf{X}_{t\pm N}, \mathbf{A}_{t\pm m}), \quad (2)$$

where $\mathbf{I}_{t\pm N}$ denotes as the identity features of video frames $\mathbf{X}_{t\pm N}$ and $\mathbf{G}_{\text{intra-frame}}$ is the intra-frame identity module. The audio also assists in detecting the face landmark, which helps the intra-frame identity module to extract identity features more accurately. Lastly, GAVN combines temporal features and identity features to predict the high-quality frames $\hat{\mathbf{X}}_{t\pm N}$:

$$\hat{\mathbf{X}}_{t\pm N} = \mathbf{R}(\mathbf{T}_{t\pm N}, \mathbf{I}_{t\pm N}) + \mathbf{X}_{t\pm N}, \quad (3)$$

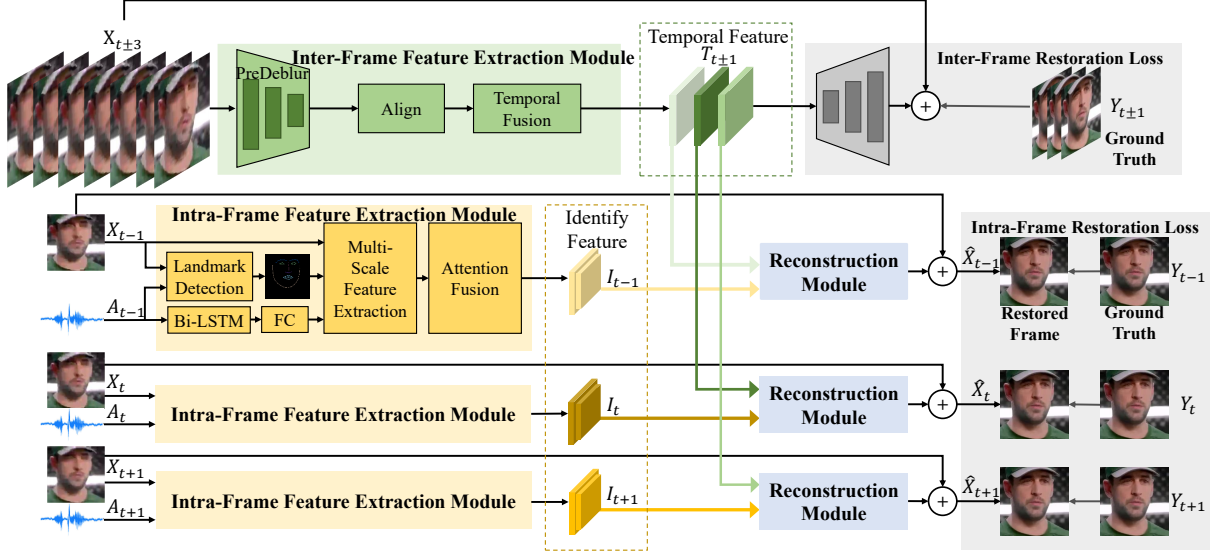


Figure 2: The framework of the proposed GAVN, which consists of three modules: (a) Inter-Frame Temporal Module: extract temporal features from the multiple consecutive frames. (b) Intra-Frame Identity Module: extract identity features with the assistance of audio and landmark features from the single frame. (c) Reconstruction Module: integrate temporal features and identity features to restore high-quality frames.

where R is the reconstruction module. GAVN can simultaneously restore multiple high-quality face video frames $\hat{X}_{t \pm N}$. During the model optimization process, GAVN first utilizes inter-frame reconstruction loss to optimize the inter-frame temporal module. It then employs intra-frame reconstruction loss to optimize the entire model, including the inter-frame temporal module, the intra-frame identity module, and the reconstruction module.

Inter-Frame Temporal Module

In order to fully exploit temporal correlations between adjacent frames in face videos, the inter-frame temporal module extracts temporal features from consecutive low-quality frames in the low-resolution space, thereby saving computational costs. Due to the inevitable movement of the camera and head, there is misalignment between the current frame and its neighboring frames. In order to capture temporal features more accurately, we employ deformable convolutions for both adjacent frame alignment and skip-frame alignment, in both forward and backward time directions. The architecture of the inter-frame temporal module is shown in the left half of Fig. 3a.

Specifically, the low-quality frames first feed into the pre-deblur module to downsample input frames with strided convolution layers and obtain 3-level pyramid features $F_{t \pm N}$ for each frame. Then, we employ deformable convolution to obtain four types of aligned features: forward adjacent, backward adjacent, forward skip-frame, and backward skip-frame aligned features. Here, we take the example of 7 consecutive frames as inputs:

$$FA_j = \begin{cases} \text{Align}(F_j, F_{j-1}), & j = t-2 \\ \text{Align}(F_j, FA_{j-1}), & j \in [t-1, t+2] \end{cases} \quad (4)$$

$$BA_j = \begin{cases} \text{Align}(F_j, F_{j+1}), & j = t+2 \\ \text{Align}(F_j, BA_{j+1}), & j \in [t-2, t+1] \end{cases} \quad (5)$$

$$FS_j = \begin{cases} \text{Align}(F_j, F_{j-2}), & j = t-1, t \\ \text{Align}(F_j, FS_{j-2}), & j = t+1, t+2 \end{cases} \quad (6)$$

$$BS_j = \begin{cases} \text{Align}(F_j, F_{j+2}), & j = t, t+1 \\ \text{Align}(F_j, BS_{j+2}), & j = t-2, t-1 \end{cases} \quad (7)$$

where FA_j , BA_j , FS_j , and BS_j denote forward adjacent frame aligned features, backward adjacent frame aligned features, forward skip-frame aligned features, and backward skip-frame aligned features of the j -th frame, respectively, and Align denotes the alignment operation. We utilize a deformable convolution operation to align neighboring frame features to reference frame features from coarse to fine.

Then, temporally fuse the four types of aligned features with the original features to obtain the temporal features. It can pay more attention to the aligned features which are more similar to the original features. We first perform the alignment process directly on the original features to ensure that the aligned features and the original features $F_{t \pm N}$ are of the same size. Then, compute the attention maps between the aligned features and the original features:

$$AF_j = \text{Align}(F_j, F_j), \quad (8)$$

$$\tilde{Y}_j = AF_j \odot \sigma(\text{Conv}(Y_j) \times \text{Conv}(AF_j)), \quad (9)$$

$$T_j = \text{Conv}([\tilde{F}S_j, \tilde{F}A_j, \tilde{A}F_j, \tilde{B}A_j, \tilde{B}S_j]), \quad (10)$$

where $Y \in \{FA, BA, AF, FS, BS\}$, σ is the sigmoid function, \odot is the element-wise multiplication, and T_j is the temporal features of the j -th frame.

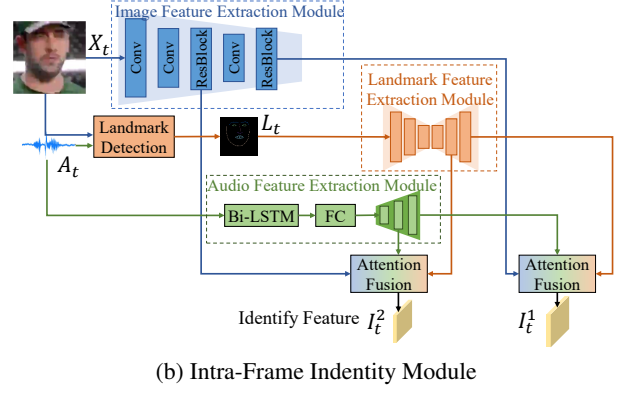
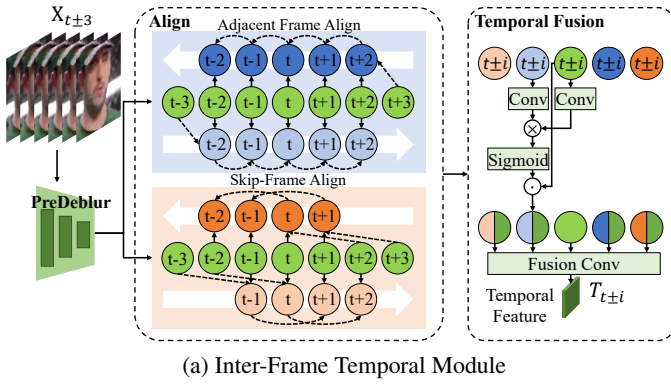


Figure 3: Details of the inter-frame temporal module and the intra-frame identity module. (a) The inter-frame temporal module utilizes deformable convolutions for aligning both adjacent frames and skip frames to extract temporal features. (b) The intra-frame identity module obtains identity features from the single frame aided by audio and landmark features

Intra-Frame Identity Module

Differing from the inter-frame temporal module, the intra-frame identity module does not include a predeblur module to downsample input frames. It integrates frame features, landmark features, and audio features to obtain identity features from the current single frame in the high-resolution space, as shown in Fig. 3b. This approach aims to extract facial details more precisely and accurately.

Since most face landmark detection methods are trained on high-resolution face images, it is difficult for them to detect face landmarks from distorted face images. To solve this problem, we pretrain the facial landmark detector network PFLD (Guo et al. 2019) on the training set. We utilize distorted face frames and corresponding audio segments as inputs and employ the landmark detection results from the original face frames as ground truth. After pretraining the PFLD network, we can predict relatively accurate face landmarks $L_{t±N}$ from distorted face images $X_{t±N}$ and the corresponding audio segments $A_{t±m}$. Then we utilize the frame feature extraction module, landmark feature extraction module, and audio feature extraction module to extract 2-level frame features, landmark features, and audio features:

$$\{F_{f,t}^1, F_{f,t}^2\} = H_{frame}(X_t), \quad (11)$$

$$\{F_{l,t}^1, F_{l,t}^2\} = H_{landmark}(L_t), \quad (12)$$

$$\{F_{a,t}^1, F_{a,t}^2\} = H_{audio}(A_t), \quad (13)$$

where $\{F_{f,t}^1, F_{f,t}^2\}$, $\{F_{l,t}^1, F_{l,t}^2\}$, and $\{F_{a,t}^1, F_{a,t}^2\}$ denote 2-level frame features, landmark features, and audio features of the t -th frame, respectively. H_{frame} , $H_{landmark}$, and H_{audio} are the frame feature extraction module, landmark feature extraction module, and audio feature extraction module, respectively.

After obtaining frame, landmark, and audio features, we fuse them to remove distortion and obtain identity features. Inspired by DAVID-Net (Zhang et al. 2020), we first compute the attention map from the audio and frame features. It indicates which regions of the audio features are more critical for video restoration. The same operation applies to the landmark features. Following this, the audio feature maps and landmark feature maps are element-wise multiplied with

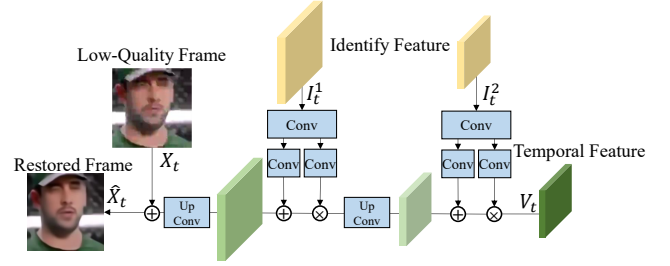


Figure 4: Details of the reconstruction module. It integrates temporal features and identity features to restore high-quality frames.

their corresponding spatial attention maps, and subsequently combined with the frame feature maps through several convolutional layers, as follows:

$$\hat{F}_{l,t}^k = F_{l,t}^k \odot \sigma(\text{Conv}([F_{l,t}^k, F_{f,t}^k])), \quad (14)$$

$$\hat{F}_{a,t}^k = F_{a,t}^k \odot \sigma(\text{Conv}([F_{a,t}^k, F_{f,t}^k])), \quad (15)$$

$$I_t^k = \text{Conv}([\hat{F}_{l,t}^k, F_{f,t}^k, \hat{F}_{a,t}^k]), k = 1, 2 \quad (16)$$

where k denotes the k -th level feature. Finally, we obtain 2-level identity features from the intra-frame identity module.

Reconstruction Module

Temporal features contain richer motion information, whereas identity features capture more spatial details. These features provide different characteristics and compensate for each other. Therefore, we utilize the reconstruction module to integrate both the temporal features and identity features to predict high-quality frames, as shown in Fig. 4. The intra-frame identity module captures 2 level identity features. The 2-th level identity feature I_t^2 has the same size as the temporal feature T_t . We first fuse them and upsample to the same size as the 1-th level identity feature:

$$\hat{I}_t^2 = \Phi_{UP}(T_t \times \text{Conv}(I_t^2) + \text{Conv}(I_t^2)), \quad (17)$$

where Φ_{UP} denotes the upsample residual convolution block, which upsamples the features by a factor of 2 using

the pixel shuffle. The same goes for 1-th level identity feature:

$$\hat{I}_t^1 = \Phi_{\text{UP}}(\hat{I}_t^2 \times \text{Conv}(I_t^1) + \text{Conv}(I_t^1)). \quad (18)$$

Finally, reduce the channel of the fused features \hat{I}_t^1 to 3 channels, and add to the input low-quality frames X_t to obtain the predicted high-quality frames \hat{X}_t .

Experiments

We conduct a comprehensive performance analysis for our GAVN method on two datasets: the VoxCeleb2 dataset (Chung, Nagrani, and Zisserman 2018; Nagrani, Chung, and Zisserman 2017) and the Obama dataset. The VoxCeleb2 and Obama datasets cover multi-speaker and single-speaker scenarios, respectively, ensuring that GAVN can generalize across diverse scenarios. Specifically, we validate GAVN across three types of face video restoration task, including compression artifact removal, deblurring, and super-resolution, to demonstrate its generalizability.

Dataset

VoxCeleb2 contains over 1 million speeches from 6,112 celebrities. Due to the limited computing resources, we randomly select a training set (200 celebrities with 4860 videos), a validation set (5 celebrities with 297 videos) and a testing set (20 celebrities with 894 videos) with no overlap of speakers. The Obama dataset consists of 180 high-quality weekly address videos from the White House website, ranging from one to six minutes. It is split into 150 videos for training, 5 videos for validation, and the rest 25 videos for testing. We crop and resize the face region of each frame to 224×224 resolution. To simulate real-world degradation, we apply three types of distortions: compression, blur, and low resolution. Compression videos are generated using FFmpeg with the x264 codec and a CRF of 45. For blur, we apply a Gaussian filter with a kernel size between 15 and 25. For low resolution, we use Bicubic interpolation with random downsampling factors from 2 to 8.

Evaluation Criteria

For image quality metrics, we utilize PSNR, SSIM, MSSSIM, and LPIPS. MS-SSIM is an extension of SSIM designed to measure structural similarity at multiple scales. LPIPS utilizes a deep neural network to capture perceptual similarity between images, which can provide a more accurate alignment with human judgments. For lip-sync quality metric, we use SyncNet’s confidence (Sync_c) and SyncNet’s distance (Sync_d) (Chung and Zisserman 2017). Sync_c can assess the synchronization quality between the restored face video and audio. Sync_d measures the distance between the lip and audio representations. Better face video restoration methods should have larger PSNR, SSIM, MSSSIM, Sync_c values and smaller LPIPS, Sync_d values.

Training Details

The training procedure of GAVN consists of two steps. In the first step, we only optimize the inter-frame temporal module. We only extract the temporal features from the

inter-frame temporal module to predict the restored frames. We utilize the Charbonnier penalty function (Lai et al. 2017) as the loss function. The training epoch is set to 20 epochs. The learning rate is set to 4×10^{-4} .

In the second step, we also utilize the Charbonnier penalty function as the loss function to optimize the entire model. We implement a warm-up training strategy: we utilize the inter-frame temporal module pretrained in the first step to extract temporal features and only train the intra-frame identity module and the reconstruction module in the first 5 epochs with the learning rate $lr = 4 \times 10^{-4}$. For the subsequent 15 epochs, we fine-tune the entire model with the learning rate $lr = 2 \times 10^{-4}$, including the inter-frame temporal module, the intra-frame identity module, and the reconstruction module.

We train the GAVN using the Adam optimizer (Kingma and Ba 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We implement the proposed GAVN model in PyTorch (Ketkar et al. 2021) and train it with one NVIDIA A100 GPU.

Comparison with SOTA Methods

We compare our GAVN with several SOTA restoration methods: DBPN (Haris, Shakhnarovich, and Ukita 2018), EDVR (Wang et al. 2019), BasicVSR++ (Chan et al. 2022), DAVID-Net (Zhang et al. 2020), and VRT (Liang et al. 2024). DBPN is proposed for image super-resolution, EDVR, BasicVSR++, and VRT are the unified framework extensible to various video restoration tasks, and DAVID-Net is designed for the task of audio-aided video compression artifact removal. We utilize the same datasets in our experiments to retrain all compared methods for three restoration tasks for a fair comparison.

The quantitative results on the VoxCeleb2 and Obama datasets are listed in Table 1, from which we have several noteworthy observations. Firstly, it is evident that GAVN achieves the best performance across all metrics, particularly in Sync_c and Sync_d . This indicates GAVN’s superior capability to reconstruct the mouth region based on audio information. It can significantly improve synchronization quality between the restored face videos and audio. Secondly, on the VoxCeleb2 and Obama datasets, GAVN is better than the audio-aided face video restoration method DAVID-Net. This indicates that GAVN can learn better correlations between facial dynamics and speakers’ voices and can leverage this knowledge to enhance face restoration. Thirdly, compared to the Obama dataset, GAVN has a more significant improvement on the VoxCeleb2 dataset. The larger and more diverse VoxCeleb2 dataset enables GAVN to more effectively improve face video restoration quality.

Qualitative results are presented in Figs. 5 and 6. GAVN exhibits richer detail recovery compared to other restoration methods, especially in the mouth and eye regions. As shown in the second row of Fig. 5, video restoration methods may mistakenly predict the open and closed states of the eyes. While GAVN can capture finer details and enhance quality of eye regions, providing clearer upper eyelid contour and highlights in the pupils. Moreover, GAVN can better restore the details of the mouth region, including the lips, upper and lower teeth, as well as the gaps between the teeth.

Task	Method	Training Frame	VoxCeleb2						Obama					
			PSNR \uparrow	SSIM \uparrow	MSSSIM \uparrow	LPIPS \downarrow	Sync \uparrow	Sync \downarrow	PSNR \uparrow	SSIM \uparrow	MSSSIM \uparrow	LPIPS \downarrow	Sync \uparrow	Sync \downarrow
Compression Artifact Removal	DBPN (Haris et al. 2018)	1	28.2960	0.8487	0.9294	0.1749	3.6849	9.9143	27.9026	0.8495	0.9469	0.1397	1.2573	9.4721
	EDVR (Wang et al. 2019)	5	28.5213	0.8518	0.9350	0.1686	4.3490	9.5611	28.3826	0.8531	<u>0.9567</u>	<u>0.1163</u>	1.4388	9.2704
	BasicVSR++ (Chan et al. 2022)	15	28.6857	0.8513	0.9348	<u>0.1665</u>	4.0944	9.7808	28.5223	0.8469	0.9504	0.1297	<u>1.5376</u>	<u>9.1289</u>
	DAVD-Net (Zhang et al. 2020)	5	28.7269	<u>0.8534</u>	0.9367	0.1694	<u>4.6788</u>	<u>9.3363</u>	<u>28.6645</u>	<u>0.8601</u>	0.9502	0.1330	1.3697	9.3129
	VRT (Liang et al. 2024)	16	28.4900	0.8504	0.9334	0.1703	3.9567	9.8124	28.0689	0.8351	0.9482	0.1349	1.2367	9.4621
	GAVN(Ours)	7	28.9780	0.8622	0.9391	0.1658	4.8463	9.2228	30.0221	0.8898	0.9671	0.1032	1.6664	9.0379
Deblur	DBPN (Haris et al. 2018)	1	38.0848	0.9617	0.9903	0.0645	7.0134	7.4523	34.5857	0.9357	0.9872	0.0794	1.9223	8.8230
	EDVR (Wang et al. 2019)	5	38.9772	0.9698	0.9950	0.0564	7.0534	7.4293	35.0969	0.9446	0.9921	0.0675	1.9918	8.7009
	BasicVSR++ (Chan et al. 2022)	15	38.5049	0.9675	0.9945	0.0620	7.0344	<u>7.4287</u>	34.7329	0.9406	0.9915	0.0740	1.9731	8.7216
	DAVD-Net (Zhang et al. 2020)	5	38.4667	0.9672	0.9943	0.0613	7.0441	7.4355	34.9129	0.9419	0.9913	<u>0.0651</u>	1.9551	8.7518
	VRT (Liang et al. 2024)	16	38.1804	0.9665	0.9940	0.0629	7.0251	7.4420	34.3923	0.9327	0.9814	0.0824	1.9012	8.9132
	GAVN(Ours)	7	39.3441	0.9716	0.9954	0.0523	7.0603	7.4251	35.2761	0.9461	0.9923	0.0610	2.0543	8.6771
Super Resolution	DBPN (Haris et al. 2018)	1	34.7139	0.9394	0.9778	0.0896	6.3998	7.8302	31.5422	0.9176	0.9793	0.0713	1.7924	8.9213
	EDVR (Wang et al. 2019)	5	35.1852	0.9455	0.9828	0.0725	6.6885	<u>7.6744</u>	32.0530	0.9243	0.9848	0.0575	1.8719	8.8151
	BasicVSR++ (Chan et al. 2022)	15	<u>35.5053</u>	<u>0.9496</u>	<u>0.9852</u>	<u>0.0663</u>	6.5442	7.7513	<u>32.9240</u>	<u>0.9339</u>	<u>0.9868</u>	<u>0.0540</u>	1.8090	8.8738
	DAVD-Net (Zhang et al. 2020)	5	34.5627	0.9391	0.9797	0.0806	6.5985	7.7462	31.9942	0.9215	0.9837	0.0670	<u>1.9343</u>	<u>8.7629</u>
	VRT (Liang et al. 2024)	16	34.9731	0.9378	0.9765	0.0871	6.4612	7.7419	30.2217	0.8929	0.9782	0.0781	1.8012	8.9552
	GAVN(Ours)	7	36.1462	0.9543	0.9859	0.0661	6.7240	7.6504	33.0620	0.9344	0.9873	0.0488	1.9428	8.7450

Table 1: Quantitative comparison of GAVN and SOTA restoration methods on the VoxCeleb2 dataset and Obama dataset for three restoration tasks (compression artifact removal, deblur, and super-resolution). The best and second-best performances for each metric are marked in boldface and underlined, respectively.



Figure 5: Qualitative results on VoxCeleb2 dataset. Distortion types from top to bottom: compression, blur, and low resolution.

We further conduct experiments at different distortion levels to validate the effectiveness of GAVN across various degree of low-quality videos. The different distortion lev-

els and the corresponding experimental results are listed in Table 2. It highlights the superiority of GAVN on restoring different quality videos.



Figure 6: Qualitative results on the Obama dataset. Distortion types from top to bottom: compression, blur, and low resolution.

Distortion	σ	EDVR	BasicVSR++	DAVD-Net	GAVN
Compression	35	31.5020	31.6120	31.7197	31.8327 (+0.1130)
	40	30.7648	30.8615	30.8785	31.0203 (+0.1418)
	45	28.5213	28.6857	28.7269	28.9780 (+0.2923)
Blur	17	39.2855	38.9902	38.7835	39.6316 (+0.3461)
	21	38.2728	37.8910	37.7774	38.7089 (+0.4361)
	25	37.1521	36.2650	36.5882	37.5581 (+0.4060)
Low Resolution	2	43.1496	42.4803	42.9954	44.1499 (+1.0003)
	4	36.2302	36.5015	35.3814	37.1964 (+0.6949)
	8	29.2927	29.7503	28.6765	30.1159 (+0.3656)

Table 2: PSNR at Different Distortion Levels on VoxCeleb2. σ represents the CRF scale, Gaussian blur, and downsampling scale respectively.

Task	Method	PSNR \uparrow	SSIM \uparrow	MSSSIM \uparrow	LPIPS \downarrow	Sync $_c$ \uparrow	Sync $_d$ \downarrow
Compression	w/o IF	28.7797	0.8547	0.9378	0.1759	4.7170	9.3214
	w/o AF	28.9480	0.8618	0.9382	0.1718	4.8340	9.2663
	GAVN(Ours)	28.9780	0.8622	0.9391	0.1658	4.8463	9.2228
Deblur	w/o IF	38.8352	0.9691	0.9949	0.0579	7.0517	7.4337
	w/o AF	39.0181	0.9701	0.9951	0.0540	7.0555	7.4311
	GAVN(Ours)	39.3441	0.9716	0.9954	0.0523	7.0603	7.4251
Super Resolution	w/o IF	35.7568	0.9512	0.9821	0.0680	6.7111	7.6538
	w/o AF	35.8427	0.9516	0.9842	0.0678	6.7125	7.6512
	GAVN(Ours)	36.1462	0.9543	0.9850	0.0661	6.7240	7.6504

Table 3: Quantitative results of ablation studies on the VoxCeleb2 dataset. IF: identity feature; AF: audio feature.

Ablation Studies

We conduct ablation studies to validate the importance of intra-frame identity features and audio features in GAVN. As shown in Table 3, removing intra-frame identity features eliminates the support of landmark and audio features, making GAVN a regular video restoration model. Consequently, its performance is comparable to BasicVSR++. As shown in the third row of Table 3, removing audio features significantly reduces both the quality of restored faces and the consistency between speech and lip movements.

Experiments on Real-World Degraded Face Videos

We collected 10 real-world degraded face videos from YouTube, featuring diverse genders, skin tones, and hair col-



Figure 7: Qualitative results on real-world videos.

Metrics	EDVR	BasicVSR++	DAVD-Net	VRT	GAVN(Ours)
NIQE \downarrow	6.5407	6.2761	6.1620	6.1762	6.1549
Sync $_c$ \uparrow	1.6590	1.6065	1.6581	1.6326	1.7895
Sync $_d$ \downarrow	7.2050	7.3655	7.1657	7.1825	7.1253

Table 4: Quantitative results on real-world videos.

ors. Since there are no corresponding high-quality undistorted face videos available, we employed NIQE, SyncNet’s confidence score (Sync $_c$), and average distance (Sync $_d$) to evaluate the naturalness and audio-visual synchronization of restored face videos. The quantitative and qualitative results on real-world restored face videos are presented in Table 4 and Fig. 7. It can be observed that our GAVN outperforms other methods, achieving the best performance in the restoration of real-world degraded face videos. This demonstrates that GAVN can better handle complex real-world distortions, making it suitable for a wider range of applications.

Conclusion

We propose a general audio-assisted face video restoration method GAVN for the face video compression artifact removal, deblurring, and super-resolution. GAVN leverages inter-frame temporal features and intra-frame identity features to restore face videos. Temporal features capture complex inter-frame motion information to restore frames coarsely and then identity features refine more facial details. We conducted experiments in two different scenarios. The results demonstrate that GAVN achieves SOTA performance.

Acknowledgments

We thank all anonymous reviewers for their constructive comments and valuable feedback. This work was supported in part by the National Natural Science Foundation of China under Grant 62522116, Grant 62271312, Grant 62132006, Grant 62301316 and Grant 62301310, and in part by STCSM under Grant 22DZ2229005.

References

- Bulat, A.; and Tzimiropoulos, G. 2018. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 109–117.
- Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4947–4956.
- Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5972–5981.
- Chen, C.; Seff, A.; Kornhauser, A.; and Xiao, J. 2015. Deep-driving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, 2722–2730.
- Chen, Y.; Tai, Y.; Liu, X.; Shen, C.; and Yang, J. 2018. Fs-rnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2492–2501.
- Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.
- Chung, J. S.; and Zisserman, A. 2017. Out of time: automated lip sync in the wild. In *ACCV*, 251–263.
- Guo, S.; Yang, X.; Ma, J.; Ren, G.; and Zhang, L. 2022. A differentiable two-stage alignment scheme for burst image reconstruction with large shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 17472–17481.
- Guo, X.; Li, S.; Yu, J.; Zhang, J.; Ma, J.; Ma, L.; Liu, W.; and Ling, H. 2019. PFLD: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859*.
- Guo, Y.; Zhang, X.; and Wu, X. 2020. Deep multi-modality soft-decoding of very low bit-rate face videos. In *Proceedings of the ACM International Conference on Multimedia*, 3947–3955.
- Haris, M.; Shakhnarovich, G.; and Ukita, N. 2018. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1664–1673.
- Haris, M.; Shakhnarovich, G.; and Ukita, N. 2019. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3897–3906.
- Isobe, T.; Li, S.; Jia, X.; Yuan, S.; Slabaugh, G.; Xu, C.; Li, Y.-L.; Wang, S.; and Tian, Q. 2020. Video super-resolution with temporal group attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8008–8017.
- Ketkar, N.; Moolayil, J.; Ketkar, N.; and Moolayil, J. 2021. Introduction to pytorch. *Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch*, 27–91.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kong, X.; Wang, K.; Wang, S.; Wang, X.; Jiang, X.; Guo, Y.; Shen, G.; Chen, X.; and Ni, Q. 2021. Real-time mask identification for COVID-19: An edge-computing-based deep learning framework. *IEEE Internet of Things Journal*, 8(21): 15929–15938.
- Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 624–632.
- Lau, C. P.; Castillo, C. D.; and Chellappa, R. 2021. At-facegan: Single face semantic aware image restoration and recognition from atmospheric turbulence. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(2): 240–251.
- Li, D.; Xu, C.; Zhang, K.; Yu, X.; Zhong, Y.; Ren, W.; Suominen, H.; and Li, H. 2021. Arvo: Learning all-range volumetric correspondence for video deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7721–7731.
- Li, W.; Tao, X.; Guo, T.; Qi, L.; Lu, J.; and Jia, J. 2020. Mucan: Multi-correspondence aggregation network for video super-resolution. In *Proceedings of the European Conference on Computer Vision*, 335–351.
- Li, X.; Liu, M.; Ye, Y.; Zuo, W.; Lin, L.; and Yang, R. 2018. Learning warped guidance for blind face restoration. In *Proceedings of the European Conference on Computer Vision*, 272–289.
- Liang, J.; Cao, J.; Fan, Y.; Zhang, K.; Ranjan, R.; Li, Y.; Timofte, R.; and Van Gool, L. 2024. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*.
- Luo, Z.; Yu, L.; Mo, X.; Li, Y.; Jia, L.; Fan, H.; Sun, J.; and Liu, S. 2021. Ebsr: Feature enhanced burst super-resolution with deformable alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 471–478.
- Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.
- Sajjadi, M. S.; Vemulapalli, R.; and Brown, M. 2018. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6626–6634.
- Tian, Y.; Zhang, Y.; Fu, Y.; and Xu, C. 2020. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3360–3369.

- Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.
- Wang, Z.; Zhang, Z.; Zhang, X.; Zheng, H.; Zhou, M.; Zhang, Y.; and Wang, Y. 2023. DR2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1704–1713.
- Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2021. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 672–681.
- Yu, J.; Zhang, B.; Kuang, Z.; Lin, D.; and Fan, J. 2016. iPrivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Transactions on Information Forensics and Security*, 12(5): 1005–1016.
- Zhang, H.; Yang, J.; Zhang, Y.; Nasrabadi, N. M.; and Huang, T. S. 2011. Close the loop: Joint blind image restoration and recognition with sparse representation prior. In *International Conference on Computer Vision*, 770–777.
- Zhang, X.; and Wu, X. 2022. Multi-modality deep restoration of extremely compressed face videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2024–2037.
- Zhang, X.; Wu, X.; Zhai, X.; Ben, X.; and Tu, C. 2020. Davd-net: Deep audio-aided video decompression of talking heads. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12335–12344.
- Zhou, K.; Li, W.; Lu, L.; Han, X.; and Lu, J. 2022. Revisiting temporal alignment for video restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6053–6062.