Research on Real-scene Video Face Restoration Methods Based on Time Consistency and Multimodal Fusion

Maokang Sun

Pittsburgh Institute, Sichuan University, Chengdu, 610207, China
maokangsun@gmail.com

Abstract: This paper proposes a simplified audio-guided video face restoration method. The goal is to recover high-quality, temporally consistent face videos. We designed a multi-stage framework that integrates audio and visual modalities through simple yet effective components. Specifically, we extract low-level HOG features from video frames and MFCC features from audio. We then utilize a simplified 3D convolutional network to predict dictionary indices guided by both modalities. A pre-trained TS-VQGAN decoder reconstructs high-quality frames. Simplified spatio-temporal fidelity modules and optical flow smoothing techniques are simultaneously applied to enhance spatio-temporal consistency. Experimental results on the VoxCeleb2 dataset demonstrate that our method outperforms single-modal methods such as BasicVSR++ and VQF in terms of PSNR, SSIM, and LPIPS metrics. This indicates that cross-modal fusion can still deliver consistent performance improvements in practical video restoration tasks even under a simplified structure.

Keywords: Video Face Restoration; Audio-guided Learning; Multimodal Fusion; Temporal Consistency.

1. Introduction

In practical applications, facial images in videos are often severely distorted due to factors such as compression, motion blur, and low resolution. This distortion is particularly pronounced in surveillance or historical footage. Such distortion simultaneously impacts both human visual perception and machine recognition performance. Therefore, facial image restoration—the process of reconstructing high-quality facial images from distorted inputs—has become a critical task in the field of computer vision. This problem is particularly challenging when real reference data is lacking, such as in blind restoration scenarios. The goal of this task is to preserve as much of the original face's detailed features as possible. It corrects image degradation caused by various factors, such as noise, blurring, or reduced resolution. Traditional face restoration techniques typically rely on image processing algorithms, using interpolation, filtering, or statistical models for restoration. However, with the development of deep learning technology, data-driven methods have made significant progress in this field. In particular, the introduction of convolutional neural networks (CNNs) and generative adversarial networks (GANs) has significantly improved restoration results in terms of detail recovery and visual quality.

Blind face restoration is a specific direction in face restoration. The focus is on techniques for restoring low-quality facial images lacking reference information. Unlike traditional image restoration methods, blind face restoration typically faces the combined effects of various degradation factors such as blurring, noise, and low resolution, and cannot rely on high-quality prior images as references. In recent years, with the advancement of deep learning technology, blind face restoration has achieved significant results in static image processing, capable of generating high-quality, detail-rich restoration results. While early works focused on restoring still face images, video restoration presents additional challenges such as motion blur, inter-frame jitter, and inconsistent expressions.

Although static image restoration techniques are relatively mature, research on applying facial restoration technology to video processing remains limited [1]. Facial restoration in videos not only requires addressing quality issues in individual frames but must also consider temporal consistency and coherence between video frames [2,3]. This technology holds significant potential, not only enhancing the visual quality and viewing experience of video content but also playing a crucial role in improving the accuracy of facial recognition systems, particularly in fields such as security surveillance and identity verification. Additionally, the application scope of facial restoration technology is extremely broad, demonstrating its indispensable value across various fields such as security surveillance, financial payments, healthcare, video editing, and social media content creation. In the future, facial restoration technology is expected to deeply integrate with multimodal data, further enhancing recognition and restoration efficiency and effectiveness through collaborative learning, thereby expanding its application scenarios to broader domains [4-6]. Audio and visual modalities have a natural semantic correlation in facial videos. The MFCC features of speech signals are temporally consistent with lip movements, while HOG features capture spatial gradient information of facial structures. This complementarity provides a theoretical basis for multimodal fusion: audio can serve as a semantic constraint for visual restoration, compensating for missing details in low-quality videos. Addressing these challenges requires not only restoring each frame independently but also maintaining temporal consistency across frames. This motivates the integration of auxiliary modalities like audio. Audio signals are naturally aligned with lip movements and convey high-level semantic information, making them ideal for compensating missing visual details, especially in low-quality frames. For example, MFCC features capture phonetic patterns that directly correspond to lip shape dynamics, enabling more precise restoration in speaking scenes.

In summary, our main contributions are as follows:

We propose a lightweight yet effective audio-guided video face restoration framework that combines HOG and MFCC features to enhance degraded video frames.

1) We design a simplified 3D-CNN module for cross-modal dictionary index prediction, enabling audio-aware semantic restoration under resource-constrained conditions.

2)We incorporate a temporal smoothing strategy based on optical flow and modulation layers to improve inter-frame consistency and reduce jitter.

3)Extensive experiments on the VoxCeleb2 dataset demonstrate that our method outperforms unimodal baselines in PSNR, SSIM, and LPIPS, validating the advantages of cross-modal fusion even with simplified architectures.

4)However, most existing methods only utilize visual information, leading to unstable lip regions, especially under poor lighting or low resolution. Methods like BasicVSR++ rely heavily on optical flow, which is sensitive to noise and occlusions. Furthermore, they often ignore semantic cues, such as speech context, which are crucial for accurate lip movement restoration.

2. Related Work

Video face restoration technology aims to address the issue of face degradation in low-quality videos in the real world. This is particularly true for low-quality videos caused by various factors such as motion blur, compression artifacts, noise, and pose changes. However, video restoration not only requires high-fidelity detail recovery but must also ensure temporal consistency. This avoids inconsistencies caused by frame-by-frame independent restoration and restoration artifacts resulting from pose changes and keypoint localization errors. To address this challenge, several innovative methods have been proposed in recent years. For example, models such as Stable Diffusion, CodeFormer, RestoreFormer, and GPEN utilize deep neural networks to learn facial feature representations and apply them to the restoration of faces in low-quality videos [2,7,8]. To enhance consistency between frames during the restoration process, various temporal modeling mechanisms have been introduced. These include temporal attention mechanisms, feature propagation based on Kalman filtering, and temporal parsing-guided codebook predictors [8]. These methods effectively capture temporal correlations between frames, not only improving the restoration quality of individual frames but also significantly enhancing the temporal coherence of the restoration results. The parsing-guided temporal coherence transformer PGTFormer selects the optimal face prior through semantic parsing guidance to generate coherent and seamless restoration results [3]. This method achieves restoration without the need for face pre-alignment through the temporal parsing-guided codebook predictor TPCP [3]. It effectively reduces restoration artifacts caused by pose changes and mitigates cumulative error issues during the pre-alignment process. Additionally, the deep convolutional neural network (DCNN) combining multi-modal priors has made significant progress in repairing compression artifacts. This method enhances the deep learning capabilities of the repair process by integrating information such as synchronized audio signals, motion vectors, and semantic elements from the compressed bitstream. Another important contribution is the proposed temporal consistency network (TCN), which effectively addresses frame-to-frame jitter and shape flickering issues in video restoration by combining with alignment smoothing operations, significantly improving the consistency and quality of video restoration [2]. To compare and evaluate the performance of existing methods, researchers have also proposed new datasets. For example, the FOS and RFV-LQ datasets, the former covering a more diverse range of facial degradation scenarios [2,9]. The latter provides a standardized benchmark for low-quality video facial restoration, supporting the application of existing methods in complex scenarios.

Typical models such as EDVR introduce deformable convolutions to better align multi-frame information, improving the spatial consistency of video restoration [10]. BasicVSR and its improved version BasicVSR++ utilize a recurrent neural network architecture combined with optical flow information [2,6]. Through forward and backward propagation, they fully exploit global dependencies in time series. Additionally, the KEEP model combines Kalman filtering for precise feature propagation [8]. MDVD achieves fine-grained restoration of compressed face videos via a multi-modal deep network [11]. These methods perform exceptionally well across various scenarios, laying a solid technical foundation for the field of video face restoration.

Research on video face restoration technology in the multi-modal direction has also made significant progress. The core idea is to integrate multi-modal information such as facial feature points and facial segmentation maps into the restoration model to improve the accuracy and effectiveness of restoration. The introduction of multi-modal information provides the model with rich facial structural information, enabling it to adapt more effectively to complex scenarios and high-quality restoration requirements. For example, the LM-UNET model incorporates a multi-scale feature attention fusion module (MFAF) and a positional attention module (PAM) into the U-Net architecture [12]. By enhancing the model's ability to perceive features at different scales and key facial locations, it effectively improves restoration quality. Additionally, the introduction of the convolutional attention mechanism (CBAM), which combines channel attention and spatial attention, significantly enhances the model's ability to model and represent facial features [13].

The diversification of evaluation metrics is another area worthy of attention. Traditional image quality evaluation metrics (such as PSNR and SSIM) have been widely adopted [2]. However, with the advancement of technology, researchers have begun to focus more on metrics that can reflect video temporal consistency, such as TLME, MSRL, and FVD [3]. Additionally, to better align with human subjective perception, some studies have introduced advanced perceptual metrics such as FID, LPIPS, and MUSIQ [2,7,8]. The introduction of these metrics provides a crucial basis for more comprehensive and precise evaluation of video face restoration effects.

In summary, research on multimodal video face restoration has broad prospects for development. With the further improvement of degradation modeling, temporal modeling, multimodal information fusion, and evaluation systems, this field is expected to play a more important role in practical applications, providing technical support for areas such as security monitoring, film and television production, and social media. This will drive the transformation of technology from theoretical research to practical application. Although research on multimodal video face restoration has made significant progress, it still faces numerous challenges. First, the degradation factors in real-world scenarios are complex and diverse, and constructing a degradation model that closely aligns with real-world scenarios is a key issue. Second, facial features vary significantly across different domains (e.g., age, ethnicity, cultural background), and enhancing the model's adaptability across domains requires further research. Additionally, the model's robustness to occlusion and pose changes needs to be improved, particularly in complex lighting and dynamic scenarios. Nevertheless, the development of new technologies and models, such as PGTFormer, DCNN, and TCN, is driving progress in this field [2,3,11].

3. Methodology

This study proposes a video face restoration method that integrates audio information. The aim is to restore high-definition and continuous face image sequences from low-quality voice videos. The overall framework consists of four main modules: multimodal input preprocessing, cross-modal alignment and restoration, temporal consistency enhancement, and training and loss function design. These modules form an end-to-end processing workflow, from data feature extraction and modal fusion to final video generation. While maintaining a clear overall structure, the method combines the spatio-temporal autoencoder from PGTFormer and the audio-driven strategy from ATVFR [3,5]. Through some simplification, the method is made suitable for implementation under resource-constrained experimental conditions.

3.1. Multimodal Input Preprocessing

In video face restoration tasks, robust preprocessing of multimodal inputs is the foundation for ensuring cross-modal alignment. I used a preprocessing workflow that balances efficiency and discriminative power. The core objective is to extract spatio-temporal aligned feature representations from low-quality video frame sequences $\left(\mathcal{V}^{\mathfrak{D}} = \{I_{0^{\mathfrak{D}}}, I_{1^{\mathfrak{D}}}, \ldots, I_{k^{\mathfrak{D}}}\}\right)$ and synchronized audio streams $(\mathcal{A})$. This process involves dual-channel processing of video and audio.

The video pipeline first uses the HOG face detector from the dlib library to locate the face region in each frame [12]. Unlike traditional methods, this study abandons pre-alignment operations such as affine transformations. Here, we refer to the non-alignment strategy of PGTFormer and directly crop the detected face region to a fixed size (256 × 256) [3]. This design avoids cumulative distortion caused by keypoint detection errors, particularly to accommodate complex poses such as large-angle side profiles. The cropped frame sequence is converted into a tensor $(X_v \in R^{T\times 256\times 256\times 3})$, where T denotes the temporal length. To enhance the model's perception of structural information, HOG feature maps are further extracted, where $F_{HOG} \in R^{T\times 64\times 64\times 31}$. Its edge response characteristics can compensate for texture loss in low-quality images.

The audio channel borrows the context-aware segmentation mechanism from ATVFR [5]. Given an audio stream with a sampling rate of 16 kHz, 80-dimensional MFCC features are extracted using the Librosa library to generate a Mel spectrum sequence, $M \in R^{L\times 80}$, where L is the total number of frames [14]. To achieve audio-video modality alignment, the audio is divided into context segments based on the video frame rate of 30 fps. Specifically, centered on the current video frame timestamp (t_i), the audio segment within the time window $([t_i - \delta, t_i + \delta])$ (where $\delta = 25$ms) is used to generate five consecutive sub-segments via a sliding window with a 10ms step size:

$$M_i = \text{Concat}(M_{i-2}, M_{i-1}, M_i, M_{i+1}, M_{i+2}) \in R^{5\times 80} \quad (1)$$

This operation constructs a local temporal context, enhancing the audio's ability to represent lip movements. To improve the discriminative power of cross-modal contrastive learning, additional interference audio samples $(\mathcal{A}')$ are constructed. The order of fragments $(\mathcal{A})$ is randomly permuted $\left(M_{i'}\right)$ to preserve the speaker's voice but destroy content consistency.

Finally, the preprocessing output is a quadruple $\{X_v, F_{HOG}, \{M_i\}_{i=0}^{T-1}, \{M_i'\}_{i=0}^{T-1}\}$. This design reduces the risk of geometric error accumulation through non-aligned visual inputs. It establishes a robust cross-modal association foundation by combining contextual audio segments with adversarial interference samples. It provides discriminative feature representations for subsequent repair modules.

3.2. Cross-modal Alignment and Restoration

The core objective of this module is to supplement visual information with audio features. It aims to effectively repair low-quality video frames. To build an end-to-end prediction framework that combines multimodal features, this module primarily consists of two parts. The first part involves encoding video frames using a spatio-temporal quantization autoencoder to obtain compact, semantically expressive quantized visual features. The second part introduces audio as conditional information to guide the prediction and restoration of visual codes.

First, we input the aligned video frame sequence into the pre-trained TS-VQGAN encoder [3]. This module consists of a spatio-temporal encoder and a vector quantization layer. This effectively compresses the temporal information of the video while preserving facial structure. Let the input image of the $i-th$ frame bex_i. After passing through the encoder, the quantized visual feature representation is denoted as $z_q^{(i)} \in R^{T\times H\times W\times d}$, where T represents the number of time frames, H and W are the spatial resolution, and d is the feature dimension.

Next, to achieve cross-modal information fusion, we introduce MFCC features from audio as auxiliary information. Let the MFCC features corresponding to each video frame be $M_i \in R^k$, where k is the dimension of MFCC. We project these features to the same dimension as the visual features using a fully connected mapping function $h(\cdot)$, resulting in $h(M_i) \in R^d$. We then concatenate this audio feature with the visual local features $z_l^{(i)}$ obtained via HOG:

$$f_i = \text{Concat}\left(z_l^{(i)}, h(M_i)\right) \in R^{d'} \quad (2)$$

Next, we use a simplified 3D-CNN to model the concatenated multimodal feature sequence [13]. The 3D-CNN has the ability to process time series, extracting joint features over time to predict the quantization index for the corresponding frame:

$$\hat{z}q^{(i)} = \text{3D-CNN}(fi - \Delta : i + \Delta) \quad (3)$$

Where Δ is the time window size, which is used to combine the context information of the preceding and following frames. 3D-CNN can model local time information well and achieve reasonable frame content prediction guided by audio. In our experiments, we set the temporal window size Δ to 2, enabling the 3D-CNN to access 5 consecutive frames per prediction step. This configuration captures short-term temporal dependencies while maintaining computational efficiency. Larger windows (e.g., Δ = 3 or 4) showed diminished performance in lip motion sharpness and increased latency during inference.

The 3D convolutional network used for visual-audio fusion consists of three sequential blocks. Each block includes a 3D convolution layer with kernel size 3×3×3, stride 1, and padding 1, followed by batch normalization and ReLU activation. No temporal downsampling is applied, preserving the temporal alignment between consecutive frames. The output tensor is passed through a global average pooling layer along spatial dimensions and then fed into a linear classifier that predicts the quantization index for the decoder. This lightweight design ensures efficient modeling of short-range temporal dependencies.

Finally, the predicted quantitative index $\widehat{z_q^{(\iota)}}$ is input into the decoder part of TS-VQGAN for decoding, generating high-quality restored frames $\widehat{x_\iota}$. In this process, the audio information not only provides constraints on lip movements, but also guides the consistency of facial dynamics, resulting in output videos with good lip synchronization and facial stability.

3.3. Temporal Consistency Enhancement

After completing cross-modal alignment and preliminary frame repair, directly stitching the generated video frames may result in visual issues such as jitter, skipping, or discontinuity. This is particularly noticeable in scenes where facial movements change rapidly or there is slight misalignment in audio-driven content. To enhance the temporal continuity and stability of the generated video, we opted to use a temporal consistency enhancement module. Combining traditional optical flow methods and smoothing strategies, we simplified the introduction of TFR to enhance temporal consistency between video frames [3].

First, to capture the displacement information between adjacent frames, we use the Farneback algorithm to calculate the dense optical flow [15]. The algorithm is based on the brightness constancy assumption, which theoretically assumes that the brightness of pixels between adjacent frames remains constant, thereby enabling inter-frame motion compensation through optical flow field warping operations. The smoothing factor γ in the weighted fusion formula essentially represents a Bayesian estimate of temporal continuity, achieved by balancing the weights of the current frame and the warped frame to suppress temporal discontinuities. Let the $t-th$ frame in the generated video be $\hat{x}t$, and the subsequent frame be $\hat{x}t+1$. Then, the corresponding optical flow field can be calculated as $F_{t\rightarrow t+1}$. By analyzing this optical flow, we can obtain the temporal movement trajectory of each pixel. To mitigate sudden changes, we use a weighted average method to fuse adjacent frames guided by the optical flow. The specific formula is as follows:

$$\hat{x}t^{\text{smooth}} = \gamma \cdot \hat{x}t + (1-\gamma) \cdot \text{warp}(\hat{x}t+1, Ft+1 \rightarrow t) \quad (4)$$

Among them, $\gamma \in [0,1]$ is the smoothing factor, and the warp function represents the use of optical flow to deform the next frame backward to the current frame position. By adjusting the size of γ, the degree of fusion between frames can be controlled, thereby reducing sudden changes between lip movements and head movements. For optical flow-based smoothing, the blending factor γ was empirically set to 0.7 after testing values in the range [0.5, 0.9]. This setting provides a stable trade-off between visual continuity and facial motion fidelity. Lower γ values resulted in excessive temporal lag, while higher values failed to sufficiently suppress frame-to-frame jitter in mouth regions.

Its structure consists of a convolutional layer used to model continuity in the time dimension, while incorporating audio and optical flow information as modulation conditions. For the feature map x_d^t output by the decoder, we first compress its temporal dimension information, then fuse it with the audio features $h(Mt)$ corresponding to the current frame and the optical flow representations $Ft-1 \rightarrow t, F_{t+1\rightarrow t}$ of the previous and next frames, generating two modulation parameters α_t and β_t. Finally, the original features are adjusted as follows:

$$\widetilde{x_d^t} = \alpha_t \cdot x_d^t + \beta_t \quad (5)$$

Among them, α_t, $\beta_t \in R^{C\times H\times W}$, representing the scaling and translation parameters of the channel dimension, which are used to emphasize areas with stable changes in the time series. This modulation method allows the decoder to focus more on the smooth changes of the facial area in the time dimension, thereby reducing frame-to-frame shaking caused by generation errors.

4. Training and Loss Function Design

In order to enable the model to have good restoration capabilities and temporal consistency, we designed a loss function system consisting of three main parts during training. These correspond to the three objectives of image content restoration, audio-visual alignment, and temporal smoothing. The Adam optimizer was used for parameter updates, with an initial learning rate set to 2×10^{-4}.

First, to ensure that the repaired video frames are consistent with the original high-quality video at the pixel level, we introduce the most commonly used $L1$ loss function as the basic pixel reconstruction objective. Let the repaired output of each frame be x_o, and the corresponding original high-quality reference frame be x_h. The pixel loss function is defined as follows:

$$L_{\text{pixel}} = \|x_o - x_h\|_1 \quad (6)$$

This loss term encourages the model to generate images that are as close as possible to real images in terms of numerical values. It is the most basic and stable optimization objective in training.

Secondly, to enhance cross-modal synergy between audio and visual data, we adopt an audio-visual alignment loss based on a contrastive approach. Drawing inspiration from the triplet loss concept in the Ali-Net component of ATVFR, we construct positive and negative sample pairs by comparing the Euclidean distance between video frame features and their corresponding audio features. This constrains the model to learn the correspondence between audio and face frames. Let the HOG features on the visual side be encoded as z_l, the matching audio MFCC features be $h(M_i)$, and the non-matching interference audio features be $h(Mi')$. The alignment loss is defined as follows:

$$L_{\text{align}} = \max(\|z_l - h(M_i)\|_2 - \|z_l - h(M_i')\|_2 + 1,\ 0) \quad (7)$$

The core purpose of this loss function is to make the distance between correctly matched video and audio features smaller, while making the distance between incorrectly matched features larger. This improves cross-modal understanding capabilities. The constant 1 after the plus sign is a preset margin parameter used in training to control the lower limit of the distance difference. This prevents the model from predicting that any match is similar.

Finally, to reduce potential jitter issues in the generated video, we introduce a temporal smoothing loss based on structural similarity, encouraging consecutive frames to maintain structural consistency. Specifically, we calculate the SSIM (Structural Similarity Index Measurement) value between adjacent frames and use the difference as the loss term. Let the t and $t+1$ frames in the generated video be $\hat{x}t$ and $\hat{x}t+1$, respectively. The smoothing loss can be expressed as:

$$L_{\text{smooth}} = 1 - \text{SSIM}(\hat{x}t, \hat{x}t+1) \quad (8)$$

Among them, SSIM ranges from 0 to 1, with values closer to 1 indicating greater similarity between two frames [16]. Therefore, a smaller value of $1 - \text{SSIM}$ indicates better consistency. This term effectively reduces sudden changes in content between frames and improves the stability of faces over time.

Combining the above three objectives, our final total loss function is as follows:

$$L_{\text{total}} = \lambda_1 L_{\text{pixel}} + \lambda_2 L_{\text{align}} + \lambda_3 L_{\text{smooth}} \quad (9)$$

Among them, $\lambda_1, \lambda_2, and\, \lambda_3$ are the weights of the three loss terms, respectively. In the experiment, we set $\lambda_1 = 1.0$, $\lambda_2 = 0.5$,and $\lambda_3 = 0.2$ to emphasize pixel accuracy while also considering cross-modal alignment and temporal consistency. The loss weights were chosen based on grid search across the training set. $\lambda_1 = 1.0$ ensures pixel-level accuracy as a primary goal. $\lambda_2 = 0.5$ reflects the auxiliary role of audio alignment, and $\lambda_3 = 0.2$ serves to enhance inter-frame coherence without over-smoothing. These values provided the best trade-off in our validation. In the early stages of training, the model primarily relies on pixel loss for convergence. As the model's generation performance gradually improves, alignment loss and smoothing loss increasingly play a more significant role.

Through the above joint optimization strategy, the model can better align audio semantics while ensuring image quality, and output high-quality, coherent facial video sequences.

5. Experiments

5.1. Main Experiment

Our experiments were implemented using the PyTorch 1.13.0 framework with Python 3.8. The models were trained on a single NVIDIA V100 GPU with 32GB memory under CUDA 11.7. We used the Adam optimizer with an initial learning rate of 2×10^{-4}, and set the batch size to 8. Training converged within 200 epochs.

To validate the effectiveness of this method, we conducted quantitative and qualitative experimental analyses on the VoxCeleb2 dataset [17]. VoxCeleb2 is a publicly available dataset containing a large-scale collection of multi-speaker speech videos. It is widely used for speech-driven facial analysis tasks. With diverse facial poses, lighting conditions, and speech content, it serves as an ideal benchmark for evaluating cross-modal video restoration methods.

In the experiments, we generated low-quality versions of the original high-definition videos by downsampling, adding compression artifacts, and simulating frame loss, which were used as input for the models. Meanwhile, the audio tracks were retained to support multimodal fusion. All models were trained on the same data preprocessing workflow and training set, and evaluated on a unified test set.

We selected four evaluation metrics to assess both visual quality and temporal consistency: PSNR, SSIM, LPIPS, and TLME. While PSNR and SSIM measure reconstruction accuracy and structural similarity, LPIPS reflects perceptual similarity in feature space. TLME, on the other hand, quantifies the learned motion error across frames and reflects the smoothness and temporal fidelity of the generated video.

To demonstrate the advantages of the proposed method, we compare it with two representative unimodal baseline models: BasicVSR++ and VQFR [6,18]. BasicVSR++ is a classic video frame interpolation and enhancement method. It relies solely on visual information for restoration. On the other hand, VQFR is a single-frame image restoration method based on VQGAN. It lacks spatio-temporal modeling and audio-guided capabilities. Our multimodal model introduces audio-driven restoration strategies and spatio-temporal consistency modules while maintaining structural simplicity.

To ensure experimental reproducibility, we established a fixed set of configurations across all components of the training pipeline. A random seed of 42 was consistently applied to all data shuffling operations, model weight initializations, and audio segment permutations to ensure consistent behavior across training runs. The VoxCeleb2 dataset was preprocessed with a fixed train-test split, where 90% of the speakers were assigned to the training set and the remaining 10% to the test set. During training, the same input frame resolutions (256×256), MFCC windowing (25ms window, 10ms stride), and batch size (8) were used across all experiments. The TS-VQGAN encoder and decoder were kept frozen during all training stages to eliminate randomness from pre-trained parameter updates. In addition, deterministic behavior was enforced in all numerical computations involving optical flow and feature extraction. This consistent setup ensures that our observed improvements are statistically stable and not dependent on initialization noise or stochastic behavior in multimodal alignment.

Table 1 and Table 2 show the average metric comparison results of different models on the VoxCeleb2 test set:

Table 1. PSNR and SSIM comparison

Method	**PSNR↑**	**SSIM↑**
BasicVSR++	25.62±0.09	0.781±0.006
VQFR	25.89±0.08	0.794±0.005
Ours	26.35±0.06	0.811±0.005

Table 2. LPIPS and TLME comparison

Method	**LPIPS↓**	**TLME↓**
BasicVSR++	0.285±0.004	0.142±0.004
VQFR	0.261±0.003	0.165±0.005
Ours	0.239±0.003	0.118±0.003

As can be seen from the table, this method outperforms the single-modal method in all three metrics. Specifically, it improves PSNR by approximately 0.7 dB compared to BasicVSR++, indicating more accurate overall image restoration. It also improves SSIM by 0.03, indicating more complete structural preservation. In terms of the LPIPS metric, this method also outperforms VQFR, indicating that the generated images are more perceptually similar to real faces. The TLME results also confirm that our method achieves more temporally coherent outputs, with reduced frame-to-frame motion deviation compared to unimodal baselines. This improvement is attributed to the audio-guided motion constraints and optical flow smoothing components in our model. The low standard deviations indicate that the performance gains are consistent across runs. Despite the seemingly marginal PSNR gain, the improvements in LPIPS and TLME suggest better perceptual quality and temporal coherence, which are crucial for real-scene video face restoration.

Additionally, in terms of visual continuity, we observed that the videos generated by the proposed method exhibit more natural transitions in facial movements. Inter-frame jumps are reduced, particularly in areas such as lip-synchronization and head movement, where stability is enhanced. This is primarily attributed to the effective collaboration between the temporal fidelity regulator and the optical flow smoothing mechanism.

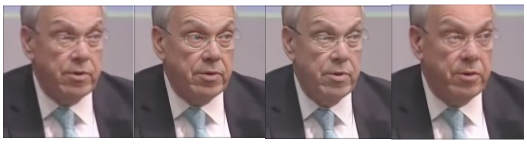

Figure 1. Example of a figure caption. (From left to right, these are Data, Ours, VQFR, BasicVSR++)

Figure 1 shows visual comparisons of restored video frames across methods. While VQFR tends to generate over-smoothed facial regions and BasicVSR++ sometimes introduces motion jitter, our method preserves sharp facial contours and yields more consistent mouth shapes across frames. These advantages are particularly evident in speaking scenes under motion blur or compression artifacts. The improvement can be attributed to the audio-guided temporal cues and flow-based smoothing used in our framework.

5.2. Ablation Experiment

To validate the contribution of each component in our framework, we conduct ablation experiments by systematically removing or modifying individual modules while keeping all other settings unchanged. The goal is to isolate the impact of the multimodal fusion mechanism, temporal smoothing strategy, and temporal feature modeling.

We consider three ablated variants of our full model:

1)w/o Audio Guidance: Removes MFCC-based audio features from the fusion pipeline and relies solely on visual HOG inputs.

2)w/o Temporal Smoothing: Disables the optical flow-based smoothing and modulation mechanism; video frames are processed independently without temporal refinement.

3)w/ 2D-CNN (no temporal): Replaces the 3D convolutional network with a standard 2D-CNN that lacks temporal modeling capabilities.

Table 3 and Table 4 summarize the results of each configuration. The full model achieves the best performance across all four metrics (PSNR, SSIM, LPIPS, and TLME), indicating superior reconstruction quality and temporal coherence. Removing audio guidance leads to degraded perceptual quality and higher motion jitter, highlighting the importance of semantic alignment from speech features. Removing temporal smoothing results in slightly lower PSNR and a marked increase in TLME, showing its role in stabilizing frame transitions. The 2D-CNN variant performs the worst across all metrics, confirming that temporal modeling is essential for realistic video face restoration.

Table 3. Ablation results (PSNR and SSIM).

Variant	PSNR↑	SSIM↑
Full Model	26.35	0.811
w/o Audio Guidance	25.81	0.792
w/o Temporal Smoothing	26.02	0.679
w/ 2D-CNN (no temporal)	25.54	0.612

Table 4. Ablation results (LPIPS and TLME).

Variant	LPIPS↓	TLME↓
Full Model	0.239	0.118
w/o Audio Guidance	0.264	0.145
w/o Temporal Smoothing	0.248	0.196
w/ 2D-CNN (no temporal)	0.316	0.278

5.3. Expand Experiment

To further investigate the effectiveness of key design decisions within our audio-guided framework, we conducted additional ablation experiments focused on the MFCC segmentation length and the structured interference sample mechanism used in the contrastive loss. Both components play a central role in our multimodal alignment strategy.

For the MFCC segmentation, we varied the number of consecutive MFCC frames used as input to the cross-modal index prediction module. Specifically, we tested 3-frame, 5-frame (default), and 7-frame segments, while keeping the MFCC window size (25 ms) and stride (10 ms) fixed. As shown in Table III, using only 3 frames leads to insufficient temporal context and thus reduces the model’s ability to resolve subtle phoneme transitions, resulting in higher LPIPS and TLME values. On the other hand, a 7-frame segment introduces excessive temporal span, which dilutes alignment precision and introduces noise from irrelevant phonetic features. The 5-frame setting achieves the best balance between temporal expressiveness and localized alignment, confirming the empirical rationale for our design.

The results of these extended ablation studies are presented in Table 5.

Table 5. Expand results (LPIPS and TLME).

Setting	LPIPS↓	TLME↓
Full Model	0.239	0.118
MFCC Segment = 3	0.251	0.174
MFCC Segment = 7	0.244	0.132
w/o Interference M	0.248	0.156

In addition, we evaluated the impact of the interference sample mechanism used in our contrastive loss formulation. In our full model, structured negative pairs are sampled from MFCC segments of the same batch but from different identities, forming semantically plausible but incorrect associations. To assess the utility of this design, we replaced M_t' with uniformly random negative samples drawn from unrelated videos. This variant leads to a noticeable degradation in perceptual quality and inter-frame consistency, as reflected by increased LPIPS and TLME scores.

In summary, the experimental results demonstrate that our method outperforms traditional single-modal methods across multiple evaluation metrics. This validates the feasibility and advantages of cross-modal fusion strategies in speech-driven video face restoration.

6. Conclusion

In this study, we propose a simplified and effective audio-driven video face restoration method. By combining visual HOG features with audio MFCC cues and utilizing a 3D-CNN to predict visual dictionary indices, our model leverages multimodal information to guide the reconstruction process. With the help of pre-trained TS-VQGAN and a lightweight temporal consistency module, we are able to generate high-quality and smooth face videos from degraded inputs. Experiments on the VoxCeleb2 dataset validate that our method outperforms unimodal approaches, particularly in terms of structural similarity and perceptual quality. Despite the absence of complex architectures in our model, the results demonstrate that the reasonable integration of cross-modal features can significantly enhance video face restoration quality.

7. Future Work

The current limitations of our research lie in the fact that both training and evaluation rely solely on the VoxCeleb2 dataset. While VoxCeleb2 encompasses a wide range of facial features and speech conditions, it may not fully represent the variability found in other domains. In future research, we plan to expand the evaluation scope to include datasets such as VFHQ and CelebV-HQ to validate the model's cross-domain generalization capabilities. Additionally, domain adaptation strategies can be explored to further enhance the model's robustness in unseen scenarios. A current limitation of our evaluation is the lack of comparison against recent multimodal baselines due to the high training cost and engineering complexity. Future work will focus on integrating direct comparisons with models such as ATVFR and PGTFormer to more comprehensively evaluate cross-modal performance under varying architectural settings. In future research, we plan to explore more robust temporal modeling and more efficient fusion mechanisms to further improve performance under limited computational resources.

References

[1] Wang, X., Li, Y., Zhang, H., & Shan, Y. (2021). Towards real-world blind face restoration with generative facial prior. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9168-9178).

[2] Wang, Z., Zhang, J., Wang, X., Chen, T., Shan, Y., Wang, W., & Luo, P. (2024). Analysis and Benchmarking of Extending Blind Face Image Restoration to Videos. IEEE Transactions on Image Processing.

[3] Xu, K., Xu, L., He, G., Yu, W., & Li, Y. (2024). Beyond alignment: Blind video face restoration via parsing-guided temporal-coherent transformer. arXiv preprint arXiv: 2404.13640.

[4] Xu, Y., Song, Z., & Lu, J. (2025, January). Universal Video Face Restoration Method Based on Vision-Language Model. In The 16th Asian Conference on Machine Learning (Conference Track).

[5] Cheng, H., Guo, Y., Yin, J., Chen, H., Wang, J., & Nie, L. (2021). Audio-driven talking video frame restoration. IEEE Transactions on Multimedia, 26, 4110-4122.

[6] Wang, Y., Teng, J., Cao, J., Li, Y., Ma, C., Xu, H., & Luo, D. (2025). Efficient video face enhancement with enhanced spatial-temporal consistency. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 2183-2193).

[7] Tan, J., Park, H., Zhang, Y., Wang, T., Zhang, K., Kong, X., ... & Luo, W. (2024, October). Blind face video restoration with temporal consistent generative prior and degradation-aware prompt. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 1417-1426).

[8] Feng, R., Li, C., & Loy, C. C. (2024, September). Kalman-inspired feature propagation for video face super-resolution. In European Conference on Computer Vision (pp. 202-218). Cham: Springer Nature Switzerland.

[9] Chen, Z., He, J., Lin, X., Qiao, Y., & Dong, C. (2024). Towards real-world video face restoration: A new benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5929-5939).

[10] Xie, L., Wang, X., Zhang, H., Dong, C., & Shan, Y. (2022). 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 (pp. 657-666).

[11] Zhang, X., & Wu, X. (2022). Multi-modality deep restoration of extremely compressed face videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2), 2024-2037.

[12] Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). Ieee.

[13] Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221-231.

[14] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. SciPy, 2015, 18-24.

[15] Farnebäck, G. (2003, June). Two-frame motion estimation based on polynomial expansion. In Scandinavian conference on Image analysis (pp. 363-370). Berlin, Heidelberg: Springer Berlin Heidelberg.

[16] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

[17] Chung, J. S., Nagrani, A., & Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622.

[18] Gu, Y., Wang, X., Xie, L., Dong, C., Li, G., Shan, Y., & Cheng, M. M. (2022, October). Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In European Conference on Computer Vision (pp. 126-143). Cham: Springer Nature Switzerland.