DAVD-Net: Deep Audio-aided Video Decompression of Talking Heads

Xi Zhang[1] Xiaolin Wu[2†] Xinliang Zhai[3] Xianye Ben[3] Chengjie Tu[4]
Shanghai Jiao Tong University[1] McMaster University[2] Shandong University[3] Tencent[4]
zhangxi_19930818@sjtu.edu.cn xwu@ece.mcmaster.ca
xinliangzhai@mail.sdu.edu.cn benxianye@sdu.edu.cn chengjietu@tencent.com

Abstract

Close-up talking heads are among the most common and salient object in video contents, such as face-to-face conversations in social media, teleconferences, news broadcasting, talk shows, etc. Due to the high sensitivity of human visual system to faces, compression distortions in talking head videos are highly visible and annoying. To address this problem, we present a novel deep convolutional neural network method for very low bit rate video reconstruction of talking heads. The key innovation is a new DCNN architecture that can exploit the audio-video correlations to repair compression defects in the face region. We further improve reconstruction quality by embedding into our DCNN the encoder information of the video compression standards and introducing a constraining projection module in the network. Extensive experiments demonstrate that the proposed DCNN method outperforms the existing state-of-the-art methods on videos of talking heads.

1. Introduction

Videos constitute roughly 80 percent of all IP traffic and still climbing. They are putting and will continue to put pressures on communication bandwidth and content storage. This makes video compression an indispensable enabling technology in today's digitally interconnected societies. For acceptable cost effectiveness, popular video compression methods (e.g., MPEG-4 [24], H.264 [31], HEVC [25]) have to compress the video data enough to achieve required savings in bandwidth and storage. For high compression ratio or low bit rates, lossy video compression inevitably produces objectionable artifacts, such as blocking, blurring, ringing and jaggies. Recently quite a few deep learning methods are proposed to remove video compression artifacts. Compared with the pure end-to-end DCNN approach for video compression [19, 14], the methods of compression artifacts removal [35, 20, 37] have the operational advantage of being compatible with existing video compression standards, as they are essentially a post-processing step of restoring already-decoded videos by the standards. We call this CNN-based video restoration strategy deep video decompression.

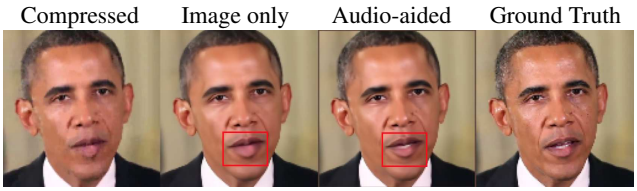

Figure 1. The reconstruction results without and with the assistance of audio signal. The lower teeth are completely missing in the frame reconstructed without using audio information.

In this work, we focus on CNN-based restoration of heavily compressed face videos with audio side information. Talking heads are arguably the most common and salient object in daily Internet video communications. For examples, conversational faces are the focal centerpiece in social media, teleconferences, Internet talk shows (TED and the alike), self media, etc. High reconstruction quality of faces in compressed videos is crucial to satisfactory user experiences, particularly when the network communication channels are congested and sporadic. Much to the advantage of algorithm or CNN designers, talking faces have very strong priors that can greatly reduce the solution space of the underlying inverse problem of video restoration. To start, the object is a face, a highly structured object; in many cases the face is known to be of a particular person. Furthermore, the audio of the speaking person is also available. Physiologically, facial muscles, particularly those in the lips, shape the sound and air stream into speech. This is why people can read lips, i.e., recognizing uttered words by watching the speaker's lips even without sound. It will be interesting to know how much one can improve the perceptual quality of compressed face videos by incorporating all the above priors into the design of deep decompression CNNs. To advance this line of enquiry, we design a

† Corresponding author.

novel neural network architecture, called Deep Audio-aided Video Decompression Network (DAVD-Net), for achieving the best possible quality of talking head videos even at very low bit rates (see Fig. 1). The success of the DAVD-Net hinges on how effectively the network can exploit the high correlation of a person's speech and her/his facial dynamics.

Besides accompanying audio, the structural information of the encoder in the video compression standards, such as coding block organization, quantization table, etc., also offers strong priors to reduce the uncertainty of the underlying video restoration problem. If used properly in the DAVD-Net, these priors can improve the video reconstruction quality further, but they are overlooked by the existing methods. To profit from the structural prior information of the encoder, we introduce a constraining projection module in the DAVD-Net. It refines the network output results by imposing both upper and lower bounds on the ground truth DCT coefficients of prediction residuals defined in video compression standards.

In summary, the major contributions of this research are the follows: (1) A baseline CNN method for deep decompression of talking head videos that outperforms existing CNN methods for the removal of compression artifacts, particularly at very low bit rates. (2) The DAVD-Net architecture design that exploits the joint audio-video statistics to extract additional performance gain over our baseline method out of the correlations of a talking head video and the associated speech. (3) A systematic performance evaluation and analysis of the DAVD-Net methodology with the availability of priors of different strengths: the DAVD-Net trained for compressed videos of a particular known speaker with and without his/her voices, and for generic talking heads with and without the accompanying voices.

This research will have a lasting practical significance even minus the pressure of communication bandwidth exerted by mass social media in extensively networked virtual communities. As the data volume of face-to-face conversations increases by order of magnitude from voice to video, backing up all conversational videos is unsustainable even for big social media service providers. The DAVD-Net technique allows such video contents to be archived in aggressively compressed form without the risk of fidelity loss, because it is capable of repairing compression defects if recalled in the future.

The rest of the paper is organized as follows. After a brief review of related works in Section 2, we present, in Section 3, the justifications and details of our network design. In Section 4, we describe the design of our experiments, explain the datasets used, and report our empirical findings. The experiments demonstrate that the proposed DAVD-Net outperforms the existing state-of-the-art methods on videos of talking heads for compression artifact reduction. Section 5 concludes the paper.

2. Related Work

Image compression artifact reduction. There is a large body of literature on removing compression artifacts in images [11, 40, 18, 3, 8]. The majority of the studies on the subject focus on post-processing JPEG images to alleviate compression noises, apparently because JPEG is the most widely used lossy compression standard.

Inspired by successes of deep learning in image restoration, a number of CNN-based compression artifacts removal methods were developed [9, 27, 13, 12]. Borrowing the CNN for super-resolution (SRCNN), Dong et al. [9] proposed an artifact reduction CNN (ARCNN). The ARCNN has a three-layer structure: a feature extraction layer, a feature enhancement layer, and a reconstruction layer. This CNN structure is designed in the principle of sparse coding. It was improved by Svoboda et al. [27] who combined residual learning and symmetric weight initialization. Guo et al. [13] and Galteri et al. [12] proposed to reduce compression artifacts by Generative Adversarial Network (GAN), as GAN is able to generate sharper image details. Zhang et al [38, 39] proposed to incorporate an ℓ_∞ fidelity criterion in the design of networks to protect small, distinctive structures in the framework of near-lossless image compression.

Deep restoration of compressed videos. All above methods for image compression artifact reduction can be viewed as of single-frame approach to video restoration without using any temporal correlations between neighboring frames. Yang et al. [37] introduced the first CNN-based multi-frame method for restoring compressed videos, which takes advantage of information in the neighboring frames. Xue et al. [35] proposed a multi-task learning approach to jointly carry out motion estimation and a video restoration task. He et al. [15] utilized the coding block information of the encoder structures to guide the video decompression process. Lu. et al. [20] modeled the video artifact reduction task as a Kalman filtering procedure and restored decoded frames through a deep Kalman filtering network. The main idea is to utilize the less noisy previously restored frames instead of directly decoded frames as temporal references. Recently, Xu et al. [34] introduced an non-local strategy in ConvLSTM to trace the spatiotemporal dependency in a video sequence, and achieved the state-of-the-art performance.

Joint audio-video generation and processing. Suwajanakorn et al. [26] proposed an interesting technique to automatically edit a video of a given speaker with accurate lip synchronization guided by his own audio in a different speech. Chung et al. [5] presented a method to animate a face image according to audio signals. Vougioukas et al. [29] proposed to do speech-driven animation with temporal GANS. Chen et al. [4] presented a method to do speech-driven facial animation with spatial attention. Afouras et al. [1] designed a deep audio-visual speech sep-

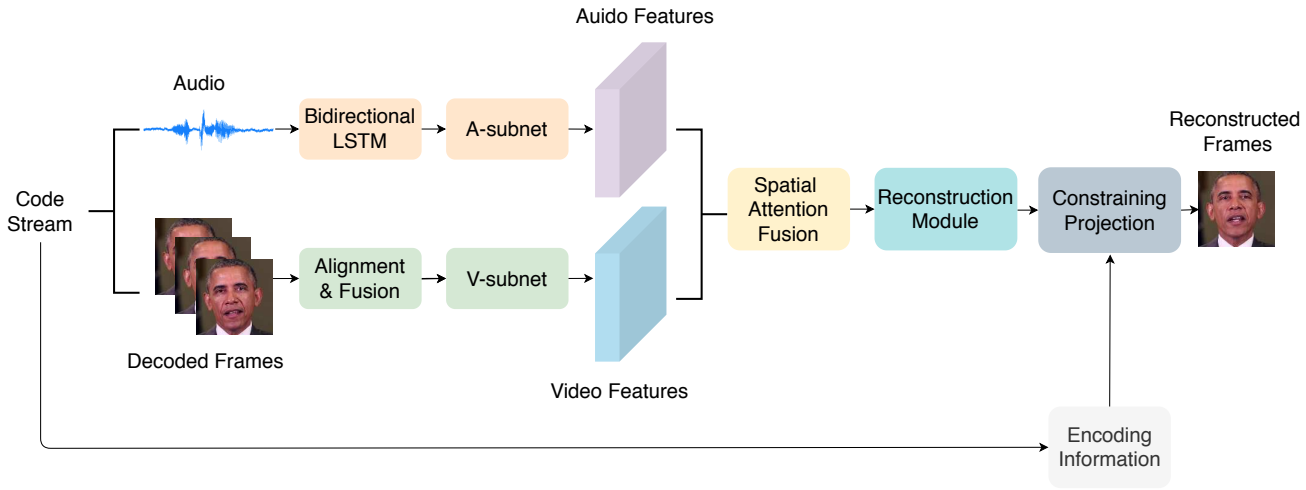

Figure 2. The framework of the proposed DAVD-Net.

aration network that is able to separate a speaker's voice from a noisy background given lip regions of the accompanying video of the speaker. Shlizerman et al. [23] presented a method that gets as input an audio of violin or piano playing, and generates a video of skeleton predictions to animate an instrument playing avatar. Wiles et al. [32] proposed a neural network that controls the pose and expression of a given face, using the talking head video of another person.

All the above CNN methods focused on video or audio generation with assistance of accompanying audio or video. To our best knowledge, no research has been reported on deep learning based methods for video artifacts reduction using joint audio-video statistics.

3. Methodology

3.1. Overview

Given an original video sequence $\{X_t|t = 0, 1, 2...\}$, $\{X_t\}$ must be compressed by a video compression standard (e.g. H.264/265) that removes spatial and temporal redundances to gain efficiency in transmission and storage. Then the compressed video will be decompressed by users to obtain a decoded video sequence, denoted by $\{Y_t|t = 0, 1, 2...\}$.

In the deep decompression task, the aim is to compute a refined reconstruction $\hat{X}_t$ from a decoded frame Y_t by maximally removing compression artifacts in Y_t. In order to utilize the temporal information, most existing methods take the current decoded frame and neighboring frames as input and output a restored current frame, that is:

$$\hat{X}_t = G(\mathcal{Y}_{t\pm n}) \tag{1}$$

where $\mathcal{Y}_{t\pm n} = \{Y_{t-n}, ..., Y_{t+n}\}$ denotes a consecutive $(2n + 1)$ compressed frames and G is the network to be optimized. In the context of talking heads videos, the audio signal is a useful piece of information for the video restoration due to the strong correlation between speech and facial movements. To factor in the audio, the video reconstruction problem should be reformulated as:

$$\hat{X}_t = G(\mathcal{Y}_{t\pm n},\ \mathcal{A}_{t\pm m}) \tag{2}$$

where $\mathcal{A}_{t\pm m} = \{A_{t-m}, ..., A_{t+m}\}$ is the consecutive audio signal temporally centered on A_t.

The overall architecture of the proposed DAVD-Net is shown in Fig. 2. It consists of two branches, for audio and video respectively. In the audio processing branch, we apply bidirectional LSTM to extract audio features, and feed them to a generation network that produces a cluster of 2-D feature maps in preparation for being combined with video feature maps. In the video processing branch, after alignment and fusion of neighboring decoded frames, we use several residual blocks to extract features of the aligned frames. Next, we design a spatial attention fusion module to dynamically fuse the audio and video features. A reconstruction module consisting of a cascade of 10 residual blocks operates on the fused video and audio features. Finally, before output, the reconstructed video is refined by a projection module that constrains the solution space by the quantization boundaries in the transform domain of the video compression standard.

Next, we detail the individual components of the proposed DAVD-Net.

3.2. Audio feature extraction

In our design the audio signal is represented by the standard Mel-frequency cepstral coefficients (MFCC) [33, 22]. When a person talks, at each time instance t, the facial image, particularly in parts around the mouth, depends not only on the current audio frame A_t but also on previous and future audio frames. For this reason, the network takes

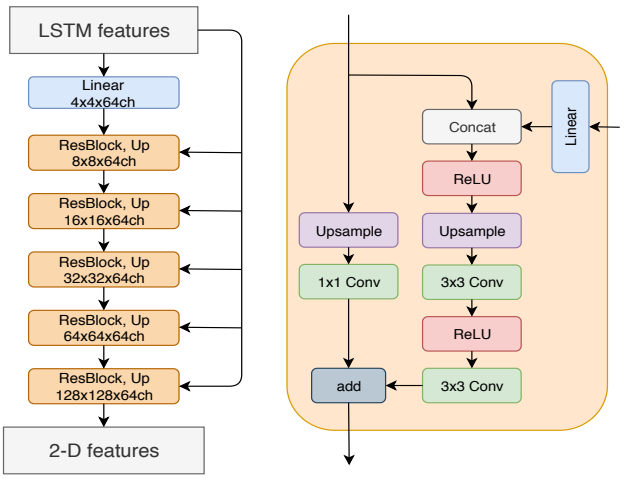

Figure 3. **Left**: The architecture layout of the A-subnet. **Right**: The detailed architecture of a upsampling ResBlock.

a consecutive audio feature sequence $\{A_{t-m}, ..., A_{t+m}\}$ as input in order to benefit from higher order statistical dependencies between the audio and video. To prepare the audio features for being combined with video features, we use a network block, called A-subnet, to extract and organize audio features in a 2D form. We do not directly use the MFCC coefficients to generate the 2-D feature maps. Instead, we adopt a three-layers bidirectional LSTM module to extract features from the MFCC coefficients. That is $\mathcal{L}_{t\pm m} = \text{LSTM}(\mathcal{A}_{t\pm m})$, where $\mathcal{L}_{t\pm m} = \{L_{t-m}, ..., L_{t+m}\}$ is the extracted LSTM feature sequence of equal length to $\mathcal{A}_{t\pm m}$.

The A-subnet takes the $\mathcal{L}_{t\pm m}$ as input and outputs a cluster of 2-D feature maps of the same size as the video frame. The sub-network consists of one linear layer and five upsampling residual blocks, are illustrated in Fig. 3.

3.3. Video feature extraction

In order to fully exploit spatiotemporal correlations in video signals, our network takes a group of consecutive $(2n + 1)$ compressed frames $\{Y_{t-n}, ..., Y_{t+n}\}$ as the second input in addition to associated audio. Due to motions of camera or/and object (head in our case), the current frame Y_t and its neighboring frames are misaligned. Aligning these video frames helps the CNN blocks for feature extraction to learn or predict spatial details more accurately. In recent studies on video super-resolution, Tian et al. [28] and Wang et al. [30] proposed to use deformable convolutions [7] to align each neighboring frame to the reference frame and have achieved the state-of-the-art performance in video super-resolution task. Inspired by their success, we also adopted the deformable convolutions to align the current frame and its neighboring frames in the DAVD-Net. After alignment, a network block called V-subnet is designated to extract features from the aligned video frames. The V-subnet is a cascade of 5 residual blocks.

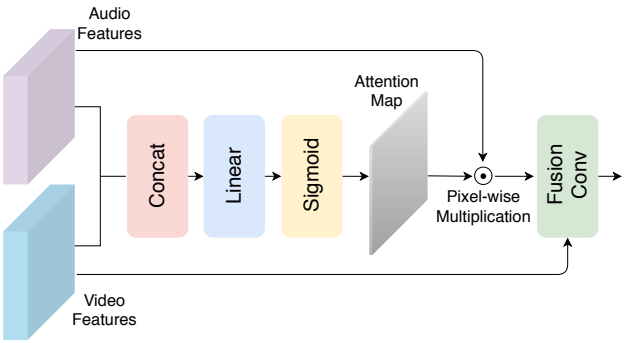

Figure 4. The detailed architecture of the proposed spatial attention fusion module.

3.4. Spatial attention fusion

After extracting temporally related audio and video features, the next task is to fuse them productively to facilitate the removal of compression artifacts. A simple approach is to concatenate the audio and video features directly. However, observing talking heads in videos reveals that the voice audio correlates most strongly to the image parts around the mouth (e.g., lips, cheeks, and chin), instead of the entire face. That is, the 2-D feature maps generated from the audio signal are not equally contributive in the spatial dimension, they should be used to guide the reconstruction of the mouth region.

However, natural head movements during talking change the position and even orientation of the speaker's mouth. To capture such dynamics the network needs to temporally adjust audio feature maps and video feature maps at pixel-level. To this end, we introduce a spatial attention fusion module to allow time varying associations of audio and video features, as illustrated in Fig. 4.

In our design, the network computes an attention map from the audio and video features with a range from 0 to 1, where 0 represents that the audio feature at this position is completely useless for recovery and 1 represents extremely helpful. The attention map M_t is formulated as:

$$M_t = Sigmod(\ Linear(\ [F_t^v, F_t^a]\ )\ ) \tag{3}$$

where F_t^v and F_t^a are the feature maps generated from video and audio, respectively. The $[\cdot, \cdot]$ denotes the concatenation operation. The sigmoid activation function is used to restrict the outputs M_t in [0, 1]. Next, the audio feature maps F_t^a are multiplied in a pixel-wise manner by the attention map and then aggregated with the video feature maps F_t^v using a few convolutional layers, that is:

$$F_{agg} = Conv(\ [F_t^v, M_t \odot F_t^a]\ ) \tag{4}$$

where F_{agg} is the aggregated feature map. Then as illustrated in 2, we feed the aggregated feature maps F_{agg} into a reconstruction module consisting of 10 residual blocks.

3.5. Constraining projection

Most existing methods for reducing video compression artifacts operate on the decoded frame sequences only, ignoring the encoder information contained in the code stream. Some researchers [15, 20, 10] did realize that the encoder prior information contained in the compressed video stream could help improve the performance of video restoration. However, they only fed the encoding prior like prediction residuals or unfiltered frames into neural networks along with the decoded frames, which is straightforward to do but has limited effect. A potentially highly profitable piece of information is left unused: the DCT coefficient quantization intervals, which can be extracted out of the compression standard code stream. To exploit the prior information we add a projection module; its role is to constrain the solution space by the quantization boundaries in the transform domain to further polish the reconstructed video. This projection module can be implemented using a piecewise linear activation function embedded in the neural network.

In modern video compression standards, prediction based coding is a core operation. In prediction based coding, given the original frame X_t to be coded, inter/intra frame prediction technique is used to obtain a prediction frame of X_t, denoted by P_t. Then the prediction residual $E_t = X_t - P_t$ will be transformed to DCT domain and quantized, followed by the entropy coding. In the encoding phase, the DCT coefficients of E_t (denoted by E_t^{dct}) are divided by a quantization table Q, and then are rounded to the nearest integers. When decoding, the decoder performs decompression by multiplying back the quantization table Q in the DCT domain. The overall quantization and dequantization process can be formulated as:

$$\hat{E}_t^{dct} = \left[(E_t^{dct})/Q\right] * Q \tag{5}$$

where $[\cdot]$ represents the round operation and $\hat{E}_t^{dct}$ denotes the decoded DCT coefficients of the prediction residual block. The decoded frame is obtained via inverse DCT transform of $\hat{E}_t^{dct}$ and adding the prediction frame, that is $Y_t = P_t + \hat{E}_t$, where $\hat{E}_t = \text{IDCT}(\hat{E}_t^{dct})$.

Eq.5 implies the following DCT coefficient range constraint:

$$\hat{E}_t^{dct} - Q/2 \leq E_t^{dct} \leq \hat{E}_t^{dct} + Q/2 \tag{6}$$

That is, from the decoded DCT coefficients of the prediction error, we can derive the lower and upper bounds of the original DCT coefficients of the prediction residuals.

We can enforce the DCT coefficient bounds in the decision of the DAVD-Net by inserting a projection module before the final output (see Fig 5). Letting $\tilde{Y}_t$ be the output of the reconstruction module, the projection module imposes

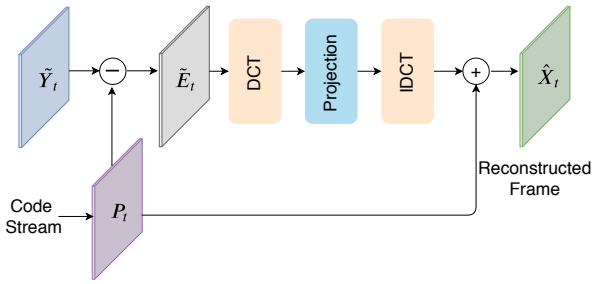

Figure 5. The architecture of the constraining projection module.

constrains on the DCT coefficients of $\tilde{E}_t = \tilde{Y}_t - P_t$:

$$F(\tilde{E}_t^{dct}) = \begin{cases} L(i,j), & \tilde{E}_t^{dct}(i,j) < L(i,j) \\ \tilde{E}_t^{dct}(i,j), & \tilde{E}_t^{dct} \in [L(i,j), U(i,j)] \\ U(i,j), & \tilde{E}_t^{dct}(i,j) > U(i,j) \end{cases} \tag{7}$$

where $L = \hat{E}_t^{dct} - Q/2$ and $U = \hat{E}_t^{dct} + Q/2$, i and j are the indexes in the DCT domain. The projection function $F(\cdot)$ can be implemented as a piecewise linear activation in the neural network. Finally, the reconstructed frame $\hat{X}_t$ is given by

$$\hat{X}_t = \text{IDCT}(F(\tilde{E}_t^{dct})) + P_t \tag{8}$$

In order to implement the proposed projection module, encoder structural information like the DCT transform block partition, prediction frame and prediction residual image are required. In video compression standards, the pixels are organized by a hierarchical block structure, and the transform block is the basic coding unit. Unlike the fixed 8×8 transform block size used in JPEG, video coding standards like (H.264/265) adopt different transform block sizes in different regions, adaptively determined according to the image content. However, the common video decoding tools like FFmpeg is not able to extract the DCT transform block partition and other encoding information from the code stream. To overcome this difficulty and get all encoder priors required by the projection module, we developed a tool to extract the encoding information from the compressed code stream, including transform block partition, prediction frame and prediction residual, etc. Some pieces of encoding information of H.264 (prediction frame, prediction residual and transform unit (TU) partition are shown in Fig.6.

4. Experiments

To systematically evaluate and analyze the performance of the DAVD-Net methodology conditioned on priors of different strengths, we conduct extensive experiments on two datasets: Obama dataset (single person) and VoxCeleb2 dataset (multiple persons) [21, 6]. In these two sets of experiments, the DAVD-Net is trained for compressed videos

Figure 6. The illustration of encoder information of the H.264 video compression standard. **Left**: prediction frame in Y channel. **Middle**: prediction residual image in Y channel. **Right**: transform unit (TU) partition.

of a particular known speaker with and without his voices, and for generic talking heads with and without the accompanying voices, respectively.

4.1. Data Preparation

Obama Dataset. We collect 198 high-quality Barack Obama's Weekly Address videos from YouTube. Each video is approximately three to six minutes long and 790 minutes in total. This dataset is divided into two parts: 160 videos for training/validation, and the rest 38 videos for testing. We detect and crop the face region from each frame and then resize it to 128×128 resolution.

VoxCeleb2 Dataset. VoxCeleb2 is an audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube. It contains speech from speakers spanning a wide range of different ethnicities, accents, professions and ages. All speaking face-tracks are captured "in the wild", with background chatter, laughter, overlapping speech, pose variation and different lighting conditions. Specifically, VoxCeleb2 contains over 1 million utterances for 6,112 celebrities. Due to the limited computing resources, we train the model using a subset (200 celebrities) randomly selected from the VoxCeleb2 development set and test the trained model using the VoxCeleb2 test set.

The compressed videos are all generated by FFmpeg with x264 video codec and rate control parameter CRF=42 and CRF=45.

4.2. Training Details

We apply an end-to-end training of all modules presented int the DAVD-Net, except of the constraining projection module, which has no parameters to be learned. The channel size in each residual block is set to 64. The upsample module in the A-subnet is implemented by deconvolution. Mini-batch size is set to 32. We use the cropped RGB face images of size 128×128 and 224×224 as input for Obama dataset and VoxCeleb dataset, respectively. The window size for video signal is 5 and for audio signal is 21. That is, $\mathcal{Y}_{t\pm2} = \{Y_{t-2}, Y_{t-1}, Y_t, Y_{t+1}, Y_{t+2}\}$ and $\mathcal{A}_{t\pm10} = \{A_{t-10}, A_{t-9}, ..., A_{t+9}, A_{t+10}\}$. The training loss is set to L_1 loss only, defined by $L_1(X_t, \hat{X}_t) = ||X_t - \hat{X}_t||_1$.

We train the DAVD-Net with Adam optimizer [17] by setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$, with initializing learning rate as 1×10^{-4}. We implement the proposed DAVD model in PyTorch [16] and train it with NVIDIA 2080 Ti GPUs. The training takes about 2 days (50 epochs) to converge on the VoxCeleb2 dataset.

4.3. Comparison with State-of-the-art Methods

We compare our DAVD-Net with several state-of-the-art methods on video restoration: DSCNN [36], DKFN [20], MFQE [37], EDVR [30]. where DSCNN, DKFN and MFQE are particularly designed for the task of video compression artifacts removal and EDVR is claimed as a unified framework that is extensible to various video restoration tasks.

To demonstrate the advantages of tailoring the restoration network to talking heads over generic objects, we compare our network with the DKFN and MFQE models pretrained by the authors using generic videos (the only two pretrained models available to us). On both Obama and VoxCeleb2 datasets, the DAVD-Net outperforms DKFN and MFQE by as much as 2.5dB. This is not surprising, because the other two networks are all trained using generic video datasets like Vimeo-90K [35] or JCT-VC [2]. In order to factor out the effects of different training sets and conduct a fair comparison, we retrain all CNN networks in the comparison group from scratch using the same datasets (Obama and VoxCeleb2) in our experiments.

The quantitative results on Obama dataset and VoxCeleb2 dataset are shown in Tables 1 and 2, respectively. On the obama dataset, our DAVD-Net outperforms the existing methods by a large margin, which means that if the network is trained for a particular known person, the assistance of accompanying voices can achieve significant performance gains on face video restoration task. On the VoxCeleb2 dataset, the proposed DAVD-Net achieves the reasonable performance gain over the existing state-of-the-art methods. This implies that by trained with a large dataset

Figure 7. Qualitative results of different methods in the comparison group on Obama dataset.

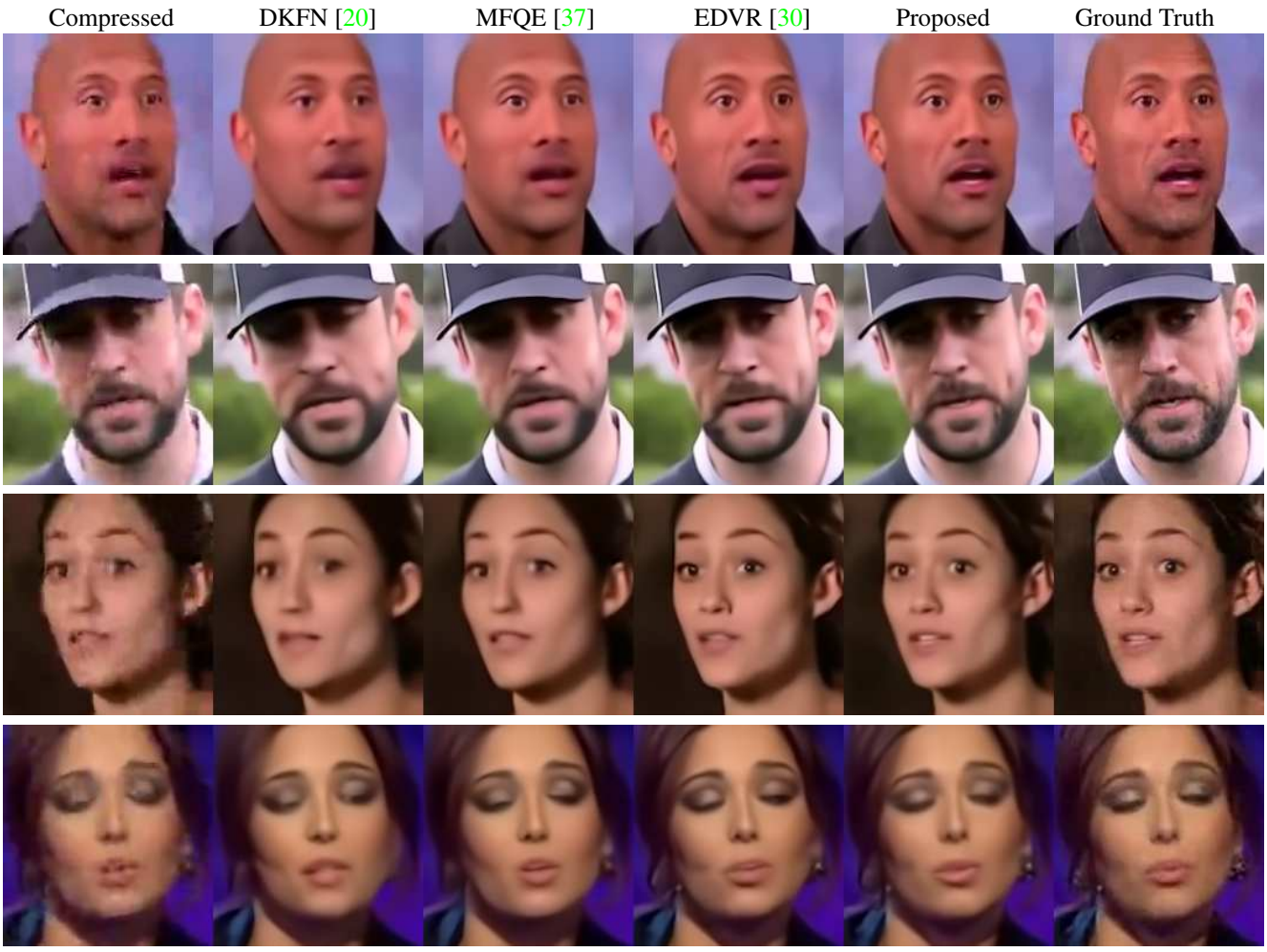

Figure 8. Qualitative results of different methods in the comparison group on VoxCeleb2 dataset.

Table 1. Average quantitative results (PSNR/SSIM) on Obama dataset for compression quality parameter CRF=42 and CRF=45.

Methods	CRF=42	CRF=45
DSCNN [36]	32.91/0.9311	31.19/0.9142
DKFN [20]	33.15/0.9348	31.40/0.9180
MFQE [37]	33.21/0.9365	31.48/0.9183
EDVR [30]	33.45/0.9402	31.64/0.9217
DAVD(Ours)	**33.94/0.9468**	**32.08/0.9272**

Table 2. Average quantitative results (PSNR/SSIM) on VoxCeleb2 dataset for compression quality parameter CRF=42 and CRF=45.

Methods	CRF=42	CRF=45
DSCNN [36]	29.18/0.8602	27.61/0.8231
DKFN [20]	29.49/0.8661	27.85/0.8259
MFQE [37]	29.54/0.8683	27.91/0.8264
EDVR [30]	29.73/0.8710	28.03/0.8289
DAVD(Ours)	**30.12/0.8741**	**28.39/0.8335**

containing various persons with accompanying voices, neural network can figure out the common correlations between face dynamics and speaking voices, and utilize it to guide the restoration of faces.

Qualitative results are presented in Figs. 7 and 8. If trained for a particular person, with the assistance of accompanying voices, DAVD can restore facial features better (note clearer teeth, sharper lips and muscle contours) than other methods, as shown in Fig. 7. If trained for generic talking heads, the improvement in perceptual image quality is lesser than the case of training the network for a given person. Nevertheless, the results of DAVD still appear to be better than the other methods, see mouth regions in Fig. 8. Please refer to the supplementary material for more qualitative results including images and videos.

4.4. Ablation Studies

In this subsection, we test various ablations of our full architecture to evaluate the effects of each component of the proposed network.

Ablation study of voices. Here we evaluate the effectiveness of using voices in face video restoration. We build a baseline that only contains the video feature extraction branch and the reconstruction module, then train it using Obama and VoxCeleb2 datasets. The performance of the baseline is shown in the first row of Table 3 and 4. As expected, the performance of our baseline is comparable to EDVR because they have similar structure and complexity. Then we combine the audio feature extraction branch into network and fuse the audio and video features with a simple convolutional layer. As shown in the second row of Table 3 and 4, the accompanying voices can significantly improve the quality of restored faces by 0.25dB in Obama dataset and 0.17dB in VoxCeleb2 dataset.

Table 3. Quantitative results (PSNR/SSIM) of ablation studies on the Obama dataset. VPB: video processing branch; APB: audio processing branch; SAF: spatial attention fusion; CPM: constraining projection module.

VPB	APB	SAF	CPM	PSNR/SSIM
✓				33.42/0.9401
✓	✓			33.67/0.9445
✓	✓	✓		33.78/0.9452
✓	✓	✓	✓	33.94/0.9468

Table 4. Quantitative results (PSNR/SSIM) of ablation studies on the VoxCeleb2 dataset. VPB: video processing branch; APB: audio processing branch; SAF: spatial attention fusion; CPM: constraining projection module.

VPB	APB	SAF	CPM	PSNR/SSIM
✓				29.71/0.8702
✓	✓			29.88/0.8731
✓	✓	✓		29.97/0.8735
✓	✓	✓	✓	30.12/0.8741

Ablation study of spatial attention fusion. We further evaluate the effectiveness of the proposed spatial attention fusion module. By aggregating the audio and video features using a spatial attention module instead of a simple convolutional layer, recovery performance increases by about 0.11dB in Obama dataset and 0.09dB in VoxCeleb2 dataset, as shown in the third row of Table 3 and 4.

Ablation study of constraining projection. As shown in the last row of Table 3 and Table 4, by introducing the constraining projection module into the network, restoration performance increases by 0.16dB in Obama dataset and 0.15dB in VoxCeleb2 dataset, respectively.

5. Conclusion

We propose and justify a novel DCNN design for restoring highly compressed videos of talking heads. The key innovation is a new DCNN architecture that can exploit the audio-video correlations to repair compression defects in the face region. We also embed into our network the structural information of the encoder in the video compression standards and introduce a constraining projection module in the network to further improve the restoration. Experiments show that the proposed DAVD outperforms existing methods appreciably.

Acknowledgements. This research is supported by Natural Sciences and Engineering Research Council of Canada (NSERC), Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project, NO.2019JZZY010119). This research is also supported in part by 111 plan B07022.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018. 2

[2] Frank Bossen et al. Common test conditions and software reference configurations. *JCTVC-L1100*, 12, 2013. 6

[3] Huibin Chang, Michael K Ng, and Tieyong Zeng. Reducing artifacts in jpeg decompression via a learned dictionary. *IEEE transactions on signal processing*, 62(3):718–728, 2014. 2

[4] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019. 2

[5] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. 2

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

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 4

[8] Yehuda Dar, Alfred M Bruckstein, Michael Elad, and Raja Giryes. Postprocessing of compressed images via sequential denoising. *IEEE Transactions on Image Processing*, 25(7):3044–3058, 2016. 2

[9] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 576–584, 2015. 2

[10] Longtao Feng, Xinfeng Zhang, Shanshe Wang, Yue Wang, and Siwei Ma. Coding prior based high efficiency restoration for compressed video. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 769–773. IEEE, 2019. 5

[11] Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. *IEEE Transactions on Image Processing*, 16(5):1395–1411, 2007. 2

[12] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. Deep generative adversarial compression artifact removal. *arXiv preprint arXiv:1704.02518*, 2017. 2

[13] Jun Guo and Hongyang Chao. One-to-many network for visually pleasing compression artifacts reduction. *arXiv preprint arXiv:1611.04994*, 2016. 2

[14] Amirhossein Habibian, Ties van Rozendaal, Jakub M Tomczak, and Taco S Cohen. Video compression with rate-distortion autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7033–7042, 2019. 1

[15] Xiaoyi He, Qiang Hu, Xiaoyun Zhang, Chongyang Zhang, Weiyao Lin, and Xintong Han. Enhancing hevc compressed videos with a partition-masked convolutional neural network. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 216–220. IEEE, 2018. 2, 5

[16] Nikhil Ketkar. Introduction to pytorch. In *Deep learning with python*, pages 195–208. Springer, 2017. 6

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Yu Li, Fangfang Guo, Robby T Tan, and Michael S Brown. A contrast enhancement framework with jpeg artifacts suppression. In *European Conference on Computer Vision*, pages 174–188. Springer, 2014. 2

[19] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11006–11015, 2019. 1

[20] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Zhiyong Gao, and Ming-Ting Sun. Deep kalman filtering network for video compression artifact reduction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 568–584, 2018. 1, 2, 5, 6, 7, 8

[21] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 5

[22] Md Sahidullah and Goutam Saha. Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. *Speech Communication*, 54(4):543–565, 2012. 3

[23] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7574–7583, 2018. 3

[24] Thomas Sikora. The mpeg-4 video standard verification model. *IEEE Transactions on circuits and systems for video technology*, 7(1):19–31, 1997. 1

[25] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012. 1

[26] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95, 2017. 2

[27] Pavel Svoboda, Michal Hradis, David Barina, and Pavel Zemcik. Compression artifacts removal using convolutional neural networks. *arXiv preprint arXiv:1605.00366*, 2016. 2

[28] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally deformable alignment network for video super-resolution. *arXiv preprint arXiv:1812.02898*, 2018. 4

[29] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. *arXiv preprint arXiv:1805.09313*, 2018. 2

[30] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 4, 6, 7, 8

[31] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding stan-

dard. *IEEE Transactions on circuits and systems for video technology*, 13(7):560–576, 2003. 1

[32] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–686, 2018. 3

[33] Min Xu, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, and Qi Tian. Hmm-based audio keyword generation. In *Pacific-Rim Conference on Multimedia*, pages 566–574. Springer, 2004. 3

[34] Yi Xu, Longwen Gao, Kai Tian, Shuigeng Zhou, and Huyang Sun. Non-local convlstm for video compression artifact reduction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7043–7052, 2019. 2

[35] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 1, 2, 6

[36] Ren Yang, Mai Xu, and Zulin Wang. Decoder-side hevc quality enhancement with scalable convolutional neural network. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 817–822. IEEE, 2017. 6, 8

[37] Ren Yang, Mai Xu, Zulin Wang, and Tianyi Li. Multi-frame quality enhancement for compressed video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6664–6673, 2018. 1, 2, 6, 7, 8

[38] Xi Zhang and Xiaolin Wu. Near-lossless ℓ_∞-constrained image decompression via deep neural network. In *2019 Data Compression Conference (DCC)*, pages 33–42. IEEE, 2019. 2

[39] Xi Zhang and Xiaolin Wu. Ultra high fidelity image compression with ℓ_∞-constrained encoding and deep decoding. *arXiv preprint arXiv:2002.03482*, 2020. 2

[40] Xinfeng Zhang, Ruiqin Xiong, Xiaopeng Fan, Siwei Ma, and Wen Gao. Compression artifact reduction by overlapped-block transform coefficient estimation with block similarity. *IEEE transactions on image processing*, 22(12):4613–4626, 2013. 2