Perceptual Quality Improvement in Videoconferencing Using Keyframes-Based GAN

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo, *Senior Member, IEEE*

***Abstract*—In the latest years, videoconferencing has taken a fundamental role in interpersonal relations, both for personal and business purposes. Lossy video compression algorithms are the enabling technology for videoconferencing, as they reduce the bandwidth required for real-time video streaming. However, lossy video compression decreases the perceived visual quality. Thus, many techniques for reducing compression artifacts and improving video visual quality have been proposed in recent years. In this work, we propose a novel GAN-based method for compression artifacts reduction in videoconferencing. Given that, in this context, the speaker is typically in front of the camera and remains the same for the entire duration of the transmission, we can maintain a set of reference keyframes of the person from the higher-quality I-frames that are transmitted within the video stream and exploit them to guide the visual quality improvement; a novel aspect of this approach is the update policy that maintains and updates a compact and effective set of reference keyframes. First, we extract multi-scale features from the compressed and reference frames. Then, our architecture combines these features in a progressive manner according to facial landmarks. This allows the restoration of the high-frequency details lost after the video compression. Experiments show that the proposed approach improves visual quality and generates photo-realistic results even with high compression rates.**

***Index Terms*—Video restoration, generative adversarial networks, videoconferencing.**

I. Introduction

In THE latest years videoconferencing has become a primary means of personal and business communication all over the world, also because of the emergence of the COVID-19 pandemic.

Lossy video compression algorithms such as H.264 and H.265 allow to decrease the bandwidth required for video transmission but introduce compression artifacts that reduce the perceived quality of the video stream. The degradation of the visual quality worsens the user experience, even making it unacceptable in certain cases.

For these reasons, the development of methods for video quality enhancement constitutes a very active area of research. In the latest years, Generative Adversarial Networks (GANs) have emerged as one of the most promising and powerful tools for several image and video processing tasks, thanks to their ability to generate photorealistic and perceptually satisfying results [1], [2], [3].

Applying deep learning-based enhancement methods to videos has several advantages. Firstly, these methods can be applied as post-processing steps to existing video compression and transmission systems without requiring to change any component and being independent of the specific video codec employed. Secondly, enhancing the visual quality of videos reduces compression artifacts and other types of degradation, thus improving the user experience. Finally, the improvement in the perceived quality makes it possible to transmit videos with higher compression rates, consequently reducing the needed bandwidth. For example, [4] uses semantic video coding and a GAN to obtain a quality comparable to the one obtained by standard H.264 with three times the bandwidth. [5] proposes a talking-head synthesis approach that reconstructs a video using one-tenth of the original bandwidth.

Contributions: In this work we propose a novel GAN-based approach for improving visual quality in videoconferencing. In videoconferencing the background has so little relevance [6] that some commercial solutions provide features to blur or replace the background with a virtual one. For this reason, we focus on the enhancement of the framed person, and in particular on the head area, because it is the most expressive and important part of interpersonal communications. Our approach is based on the assumption that the subject speaking in front of the camera stays the same for a relatively long consecutive time frame, so that we can exploit for enhancement the previous high-quality reference keyframes of the Group of Pictures (GOP) coding (i.e. the so-called *I-frames*), used in video compression algorithms as the base for motion-based compression. In particular, we propose a novel policy to create and update a set of reference keyframes in order to keep this set small, and thus memory efficient, and also to make it effective for the improvement of the visual quality. Our model extracts multi-scale features of the compressed frame and a reference keyframe and then combines them according to the facial landmarks (see Fig. 1). The feature fusion is performed with Adaptive Spatial Feature Fusion (ASFF) [7] and Spatial Feature Transform (SFT) [8] blocks in a progressive manner that

Manuscript received 14 July 2022; revised 24 December 2022; accepted 14 March 2023. Date of publication 5 April 2023; date of current version 8 January 2024. This work was supported by the European Commission through European Horizon 2020 Programme under Grant 951911 - AI4Media. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wengang Zhou. *(Corresponding author: Lorenzo Agnolucci.)*

Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo are with the Department of Information Engineering - MICC, University of Florence, 50139 Firenze, Italy (e-mail: lorenzo.agnolucci@unifi.it; marco.bertini@unifi.it; alberto.delbimbo@unifi.it).

Leonardo Galteri is with the MICC, Universita degli Studi di Firenze, 50134 Firenze, Italy (e-mail: leonardo.galteri@unifi.it).

Code and pre-trained networks are publicly available at https://github.com/LorenzoAgnolucci/Keyframes-GAN.

Digital Object Identifier 10.1109/TMM.2023.3264882

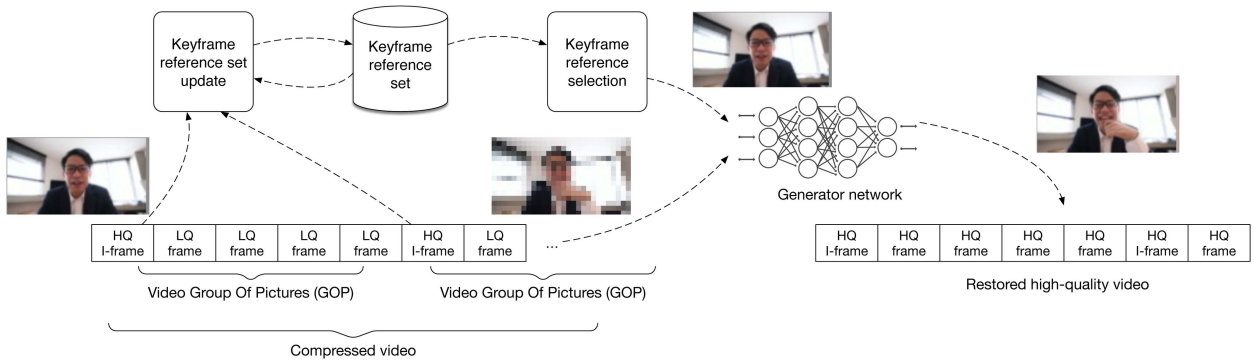

Fig. 1. Overview of the proposed system at runtime. High quality reference keyframes (video I-frames) are used in our GAN-based approach to improve the visual quality of the video conference stream. The algorithm used to update the keyframe reference set is a key element to improve the visual quality of the restored frames.

helps in restoring coarse-to-fine details. We designed a pipeline for video enhancement that involves preserving a limited number of keyframes extracted from the video stream and using the most useful ones as a reference for restoring the compressed frame. The experiments and the comparison with competing state-of-the-art approaches show that our proposed method is very effective in generating photo-realistic results even with high compression rates.

II. Related Works

a) Video Coding: Some interesting initial works have addressed the quality improvement of videos and images using coding based on neural networks [9], [10]. These approaches are currently not deployable with satisfying visual results due to an unbearable computational cost. Moreover, fully learned compression requires the standardization and diffusion of a novel technology, which is a very high market barrier to practical use.

b) Video Quality Improvement: Recently, many learning-based image enhancement techniques have been proposed [1], [2], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. Such approaches learn deep convolutional architectures, often based on GANs, to restore low-quality images corrupted by compression artifacts into high-quality ones, and deal with generic video content. [21] presents a Multi-Frame Quality Enhancement approach for compressed videos. After observing that the quality of compressed videos fluctuates across frames, the authors developed a BiLSTM-based detector to locate Peak Quality Frames (PQFs), that is frames that have a higher quality than their neighbors, whose information can be exploited to reduce the distortion of low-quality frames. A non-PQF and its nearest two PQFs are the input of a multi-frame CNN, composed of a motion compensation and a quality enhancement subnet. [22] presents EDVR, a video restoration framework with enhanced deformable convolutions. A pyramid, cascading and deformable module uses deformable convolutions in a coarse-to-fine manner to align the features of the reference frame to that of its neighboring frames and then a temporal and spatial attention fusion module combines them.

c) Face Quality Improvement: Face super-resolution has been addressed in [23], where the authors proposed GWAInet, a GAN-based approach that performs $8\times$ face super-resolution using a HR reference image of the same person depicted in the LR image. A warper subnetwork aligns the contents of the reference image to the input image. Then, after extracting the features of the LR and HR images, a feature fusion chain combines them to exploit the reference image. A peculiarity of this method is that it does not require facial landmarks for the training. In [24] super-resolution of extremely degraded faces is dealt with a GAN that produces a coarse SR image. Then, the result is refined by exploiting facial components extracted from multiple high-quality warped images of the same person or a similar one. In [25] the problem of face quality improvement is formulated as a dual-blind restoration problem, lifting the requirements of both the degradation and structural prior for training. The authors present HiFaceGAN, a collaborative suppression and replenishment framework with a nested architecture for multi-stage face renovation with hierarchical semantic guidance. [26] proposes a GAN prior embedded network for blind face restoration, using a U-shaped DNN for face restoration as a decoder. PSFR-GAN, a GAN-based Progressive Semantic-aware Style Transformation framework presented in [27], uses a face parsing network to obtain a segmentation map given an LQ face image. The input image and the segmentation map are exploited to produce a multi-scale pyramid of the inputs modulating different scale features with a semantic-aware style transfer approach. A semantic aware style loss accounts for each semantic region individually. In [28] blind face restoration task is tackled with a Guided Face Restoration Network (GFRNet) that takes advantage of a high-quality reference image of the same identity. A warper subnetwork reduces the difference in pose and expression between the two images to better recover fine and identity-aware facial details with a reconstruction subnetwork. The Deep Face Dictionary Network (DFDNet) proposed in [29] attempts to overcome the main limitation of reference-based methods by observing that facial components are similar between different people. Multi-scale dictionaries of facial parts are built offline with K-means from high-quality images. The features in the dictionaries most similar to the facial components of the degraded input are leveraged for restoration by means of Dictionary Feature Transfer and Spatial Feature Transform blocks. In [7] blind face restoration is tackled by exploiting a high-quality image

selected from multiple available images of the same person as a reference to restore a degraded one. The features of the guidance image are warped to the low-quality ones according to the facial landmarks to reduce the difference in pose and expression. Multiple Adaptive Spatial Feature Fusion blocks combine the degraded and guidance features by generating an attention mask with facial landmarks to guide the restoration of the facial components. In [4] a method that combines semantic video coding and GAN-based video quality restoration is proposed for video conference systems, using a perceptual loss that accounts separately for the background and the foreground face. [30] presents HeadGAN, a method for head reenactment that conditions head synthesis on 3D face representations from a driving video. Audio features are exploited to better synthesize mouth movements. When driving and reference identities coincide, HeadGAN can be used for face reconstruction. In [31] facial priors encapsulated in a pretrained GAN (GFP-GAN) are incorporated for blind face restoration by means of channel-split Spatial Feature Transform layers. Unlike GAN inversion methods, GFP-GAN can restore faces with a single forward pass. [32] tackles blind face restoration with a GAN that uses multi-scale facial features. A feature prior loss aims to reduce the difference in the feature space between the input and restored images, thus preserving the overall image content and spatial structure information. [33] proposes a restoration with memorized modulation framework for blind face restoration. Low-level spatial feature embedding, wavelet memory embedding and disentangled high-level noise embedding are combined with adaptive attention maps. [34] presents DAVD-Net, a DCNN architecture that exploits the audio-video correlations to remove compression artifacts in close-up talking head videos. The audio features are extracted with a BiLSTM and organized in a 2D form. The video and audio features are aggregated with a spatial attention module. To further improve the restoration the structural information of the encoder in the video compression standards is embedded into the network by adding a constraining projection module. In [35] face quality of compressed videos is enhanced with MRS-Net+, a multi-level architecture comprised of one base and two refined enhancement levels which restore small, medium and large-scale faces, respectively. A landmark-assisted pyramid alignment subnet is developed to align faces across consecutive frames. [36] and [37] exploit a multi-modality neural network to restore strongly compressed face videos. They both use video and audio signals, combined with codec information in [36] and with an emotion state in [37]. [38] presents a multi-task face restoration network that relies on network architecture search to restore images affected by various degradations. Additionally, during training clean images of the same subject as the degraded image are exploited by means of an identity loss. [39] proposes a method based on fully-spatial attention to tackle blind face restoration. A multi-head cross-attention layer takes the features of a degraded face as queries while the key-value pairs are from high-quality facial priors. The key-value pairs are sampled from a reconstruction-oriented high-quality dictionary.

Even if our aim is to improve the perceptual quality of videos we did not follow the standard multi-frame restoration approach that is commonly used in video restoration tasks, such as in MFQE 2.0 [21], MRS-NET+ [35] or DAVD-Net [34], because it usually involves looking also at future frames and this is not possible in a real-time stream. Surely taking into account only past neighboring frames is a possibility, but we preferred to consider possibly very distant I-frames and not necessarily the closest one. This preference is possible in videoconferencing because the subject usually is the same for the entire transmission so old I-frames can still be very useful in restoring the current compressed frame. This is similar to exemplar-guided face image restoration techniques but given that our method is applied to videos we can exploit multiple I-frames from the same video stream as possible references, dynamically updating the set of keyframes with the policy we designed to obtain the best performance. Precisely the LFU-inspired update strategy for the dynamic set of keyframes is what mainly differentiates our work from exemplar-based face restoration methods that constitute the state-of-the-art. For instance, ASFFNet [7] relies on a given set of reference images representing the same person and it can not handle a dynamic set of references, nor a policy for updating it. Similarly, DFDNet [29] needs an offline-generated dictionary of features of different subjects, therefore it can not exploit high-quality I-frames of the same subject that arrive in real-time.

III. Proposed Approach

Since its introduction in [40], the Generative Adversarial Network (GAN) framework has emerged as a powerful tool for various image and video synthesis tasks, such as image-to-image translation [41], face reenactment [42] and pose transfer [43]. Compared to other deep generative models, like Deep Boltzmann Machines [44] or Variational AutoEncoders [45], GANs proved to be able to generate more photorealistic results [3], [46], and have been successfully used to improve the visual quality of images [2] and videos [20]. Our method is based on such a framework.

A. Proposed Architecture

We propose a novel GAN architecture shown in Fig. 2 and inspired by [7] and [29]. Similarly to [7], we adopt the ASFF block and Moving Least Squares for warping. Differently from [7], we warp directly the reference image and not its features and we extract and fuse features at multiple scales in a progressive manner to help the network in restoring coarse-to-fine details. We took inspiration from [29] in the use of multi-scale features and of the SFT block, but we leverage a high-quality image of the same person to better restore subject-specific details. Differently from both [7] and [29], we select our reference image from the best-performing set of high-quality keyframes coming from the same video, which is built and updated with our proposed policy.

Our architecture is based on U-Net [47] and it is composed of an encoder, that processes the input so that it is smaller in terms of spatial dimensions but deeper in terms of the number of channels, and by a decoder, that inverts the process. Multi-scale reference features are combined with the features of the degraded image in a progressive manner. This approach can make the network

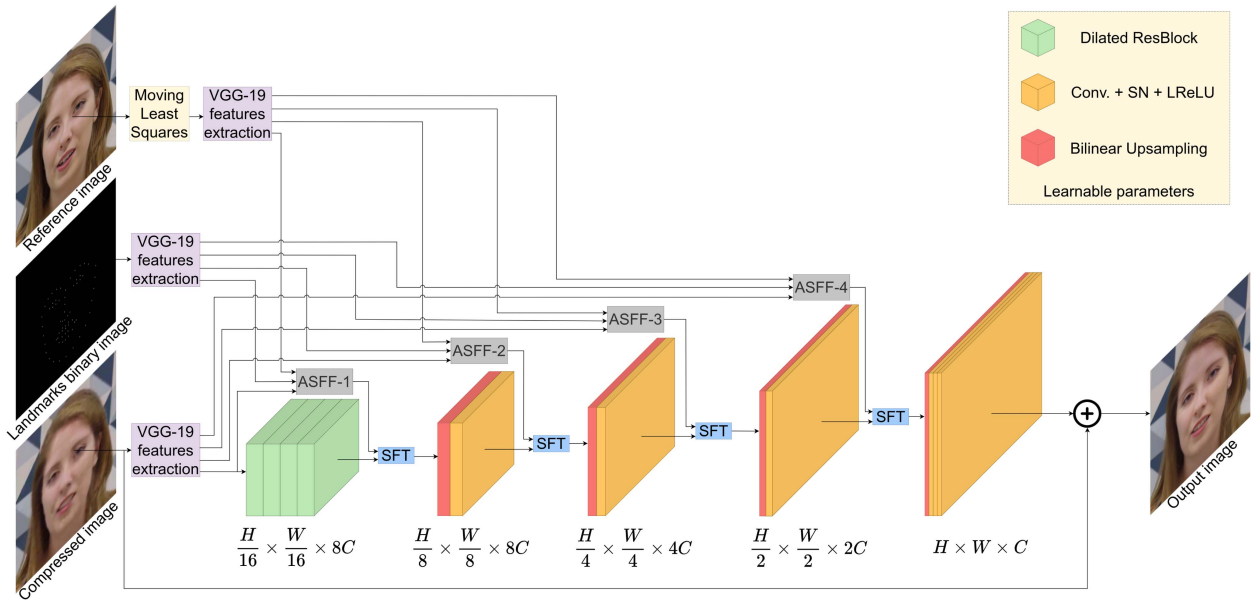

Fig. 2. Overview of the proposed architecture. Best viewed in color on PDF.

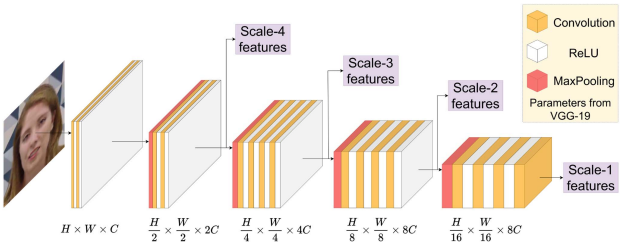

Fig. 3. Diagram of the multi-scale feature extraction with VGG-19.

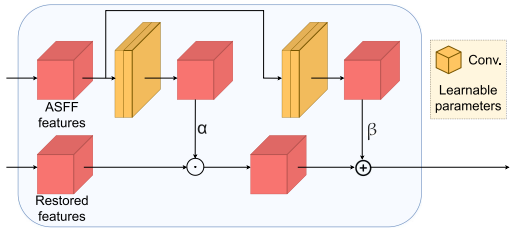

Fig. 4. Structure of the SFT block.

learn coarse-to-fine details and is beneficial to the restoration process.

Our model takes 3 inputs:

- a degraded (i.e. highly compressed) image;
- a high-quality reference image (i.e. a video *I-frame*);
- a binary image that is white only in correspondence with the facial landmarks of the compressed image.

The model produces a restored image from the compressed one.

We use a pre-trained VGG-19 [48] to extract multi-scale features from the degraded, reference and landmarks binary images. The reference (guidance) image is previously warped to the degraded one based on the facial landmarks using Moving Least Squares (Section III-C). We extract features at 4 different scales from the layers `relu_2_2`, `relu_3_4`, `relu_4_4` and `conv_5_4` of the VGG-19. The feature extraction is depicted in Fig. 3.

To align the warped reference and degraded features we adopt AdaIN [49]. This helps reduce the difference in style and illumination between the two images and thus improves the restoration. We denote by F^d and F^g the degraded and guidance features. The AdaIN can be written as

$$F^{g,a} = \sigma(F^d)\left(\frac{F^g - \mu(F^g)}{\sigma(F^g)}\right) + \mu\left(F^d\right) \tag{1}$$

where $\sigma(\cdot)$ and $\mu(\cdot)$ represent the mean and the standard deviation.

After going through multiple dilated residual blocks, the degraded features are progressively upsampled by enlarging the spatial resolution and reducing the number of channels. At the same time, they are combined with the reference features by means of Adaptive Spatial Feature Fusion (Section III-B) and Spatial Feature Transform (SFT) [8] blocks.

The SFT block generates affine transformation parameters for spatial-wise feature modulation incorporating some prior condition. The scale α and the shift β parameters are learned from the features outputted by the corresponding ASFF block. The output of the SFT block is formulated as

$$SFT = \alpha \odot F^r + \beta \tag{2}$$

where $\odot$ is the element-wise product and F^r are the restored features, that is the features originated from the degraded ones and restored in the decoding part of the architecture. Fig. 4 shows the structure of the SFT block.

Following [4], we train the network to learn the residual image, so there is a skip connection between the degraded image and the restored output. This choice reduces the overall training time and improves its stability.

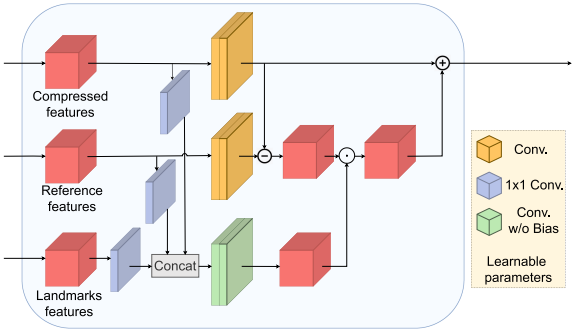

Fig. 5. Structure of the ASFF block.

B. ASFF Block

The fusion of the features of the reference and degraded images is a fundamental part of exemplar-based approaches, as it allows to fully exploit the information supplied by the guidance image. Adopting a concatenation-based approach, as in [23], [28], does not take full advantage of the reference features.

Thus, in our multi-scale architecture, we rely on multiple Adaptive Spatial Feature Fusion (ASFF) blocks [7]. While the reference image generally contains more high-quality details, the degraded image should have more weight in the reconstruction of the overall face components. For example, if the mouth of the reference image is closed while that of the compressed image is open, the reconstruction of the teeth should be mainly based on the restored features from the degraded image. For this reason, ASFF blocks generate an attention mask based on the degraded image facial landmarks to guide the fusion of the guidance and restored features. Fig. 5 shows the structure of the ASFF block.

C. Warping Reference With Moving Least Squares

For most guided face restoration methods, the performance is diminished by the pose and expression difference between reference and degraded images because it introduces artifacts in the reconstruction result. Thus, we spatially aligned the reference and compressed images with an image deformation method based on Moving Least Squares (MLS) [50].

Let p and q be respectively the sets of facial landmarks of the reference and degraded image, with $|p| = |q| = N$. In our case, $N = 68$. We aim to find a deformation function f to apply to all the points of the reference image. Given a point v in the image, we solve for the best affine transformation $l_v(x)$ that minimizes

$$\sum_{i=1}^{N} w_i \left| l_v(p_i) - q_i \right|^2 \quad \text{where} \quad w_i = \frac{1}{|p_i - v|^2} \tag{3}$$

Because the weights w_i are dependent on the point of evaluation v we obtain a different transformation $l_v(x)$ for each v. We define the deformation function f to be $f(v) = l_v(v)$.

Since $l_v(x)$ is an affine transformation we can rewrite it in terms of a linear transformation matrix M

$$l_v(x) = (x - p_*)M + q_* \tag{4}$$

where p_* and q_* are weighted centroids

$$p_* = \frac{\sum_{i=1}^{N} w_i p_i}{\sum_{i=1}^{N} w_i} \qquad q_* = \frac{\sum_{i=1}^{N} w_i q_i}{\sum_{i=1}^{N} w_i}$$

Based on this insight, the least squares problem of (3) can be rewritten as

$$\sum_{i=1}^{N} w_i \left| \hat{p}_i M - \hat{q}_i \right|^2 \tag{5}$$

where $\hat{p}_i = p_i - p_*$ and $\hat{q}_i = q_i - q_*$. The affine deformation that minimizes (5) is

$$M = \left(\sum_{i=1}^{N} \hat{p}_i^T w_i \hat{p}_i \right)^{-1} \sum_{j=1}^{N} w_j \hat{p}_j^T \hat{q}_j$$

With this closed-form solution for M, we can write a simple expression for the deformation function f

$$f(v) = (v - p_*) \left(\sum_{i=1}^{N} \hat{p}_i^T w_i \hat{p}_i \right)^{-1} \sum_{j=1}^{N} w_j \hat{p}_j^T \hat{q}_j + q_* \tag{6}$$

Applying this deformation function to each point of the reference image lets to warp it according to the facial landmarks of the degraded image.

D. Keyframes Selection and Set Maintenance

Although warping with MLS helps to reduce the distance between the compressed and reference images, if they are too different the results will still be sub-optimal. Thus it is natural to select the optimal reference keyframe as the one that has a similar pose and expression to the degraded image, instead of simply using the previous keyframe. We measure the similarity between a keyframe and the degraded frame with the Euclidean distance between the sets of facial landmarks. Considering videoconferencing, assuming that the talking subject stays the same, even very old keyframes can be useful. So, as the video progresses, one can save a limited set of keyframes, to reduce memory requirements, and then use the most similar one as a reference to restore the current compressed frame. This novel method is the key to improving the overall restoration quality of the video and limits the cases in which the compressed and reference frames are very different.

We took inspiration from the Least-Frequently Used (LFU) cache replacement strategy: for each keyframe of the set, we keep count of how many times it was selected for reconstruction and when a new keyframe is received from the video stream the least used is evicted. However, in this way, the first keyframes of the video would be excessively rewarded. Indeed, since for the first seconds of the video they are the only ones available as a reference they can be used not because of similarity with the compressed frame but for lack of alternatives. To overcome this problem we apply an exponential decay to the number of uses, i.e. when a new keyframe arrives the counter of the number of uses of all the keyframes of the set is halved.

E. Training Losses

As in [7], to train our model we employed a weighted sum of reconstruction and photo-realistic losses. We denote by I_D, I_R and I_{GT} the degraded, reconstructed and ground-truth (i.e. high-quality uncompressed) images, respectively.

The reconstruction loss constrains the reconstructed image to faithfully approximate the ground-truth one and is composed of two terms. First, we relied on the Mean Square Error (MSE), defined as

$$\ell_{MSE} = \frac{1}{CHW} \left\| I_R - I_{GT} \right\|^2 \tag{7}$$

where C, H and W denote the channel, height and width of the image. Second, we adopted the perceptual loss [51], [52], [53], defined on the VGG-19 feature space. The perceptual loss is formulated as

$$\ell_{perc} = \sum_{l \in L} \frac{1}{C_l H_l W_l} \left\| \Psi_l(I_R) - \Psi_l(I_{GT}) \right\|^2 \tag{8}$$

where Ψ_l represents the features from the l-th layer of a pre-trained VGG-19 model and $L = \{\texttt{relu\_2\_2}, \texttt{relu\_3\_4}, \texttt{relu\_4\_4}, \texttt{conv\_5\_4}\}$. We also experimented using VGG-Face [54] for the perceptual loss, in particular by extracting the output taken from the third convolutional layer of the fifth block before the ReLU activation, but the results were worse than with VGG-19.

The photo-realistic loss also contains two terms. First, we used the style loss [55] that is defined on the Gram matrix of the feature map for each layer in L

$$\ell_{style} = \sum_{l \in L} \frac{1}{C_l H_l W_l} \left\| \Psi_l(I_R)^T \Psi_l(I_R) - \Psi_l(I_{GT})^T \Psi_l(I_{GT}) \right\|^2 \tag{9}$$

Second, we employed the hinge version of the adversarial loss [56], [57]. We adopted multi-scale discriminators [58], that is 4 discriminators that have the same network structure but operate at different image scales. The adversarial loss can be formulated as

$$\begin{aligned} \ell_{adv,D} = & -\sum_{r \in R} \left[\mathbb{E}_{I_{GT}^{\downarrow r} \sim P(I_{GT}^{\downarrow r})} \left[\min\left(0, -1 + D(I_{GT}^{\downarrow r})\right) \right] \right. \\ & \left. + \mathbb{E}_{I_R^{\downarrow r} \sim P(I_R^{\downarrow r})} \left[\min\left(0, -1 - D\left(I_R^{\downarrow r}\right)\right) \right] \right] \\ \ell_{adv,G} = & -\sum_{r \in R} \lambda_{adv,r} \, \mathbb{E}_{I_D^{\downarrow r} \sim P(I_D^{\downarrow r})} \left[D\left(G\left(I_D^{\downarrow r}\right)\right) \right] \end{aligned} \tag{10}$$

where $_{\downarrow r}$ denotes the downsampling operation with scale factor $r \in R = \{1, 2, 4, 8\}$ and $\lambda_{adv,r}$ are the trade-off parameters for each scale discriminator. $\ell_{adv,D}$ and $\ell_{adv,G}$ are used to update respectively the discriminators and the generator. To stabilize the learning of the discriminators we adopted SNGAN [59], incorporating the spectral normalization after each convolutional layer of the discriminator. Spectral normalization is based on regularizing the spectral norm of each layer of the discriminator by simply dividing the weight matrix by its largest eigenvalue.

The overall training loss is defined as

$$\ell_{total} = \lambda_{MSE}\ell_{MSE} + \lambda_{perc}\ell_{perc} + \lambda_{style}\ell_{style} + \lambda_{adv}\ell_{adv,G} \tag{11}$$

where λ_{MSE}, λ_{perc}, λ_{style}, and λ_{adv} are the tradeoff parameters.

IV. Results

A. Datasets

Similarly to [4], we used the Deep Fake Detection (DFD) dataset [60], which is composed of 363 high-resolution and high-quality videos depicting different activities performed by 28 actors. Then, we selected 55 videos of actions in which the actor is talking while facing the camera as in a setup of a video conference (i.e. "podium speech" and "talking against wall" scenes) for an overall size of $\sim$ 40 GB and a duration of $\sim$ 40 minutes. The first 22 identities were utilized for training and the last 6 for testing.

We also employed the High-Definition Talking Face (HDTF) dataset [61], which contains 362 videos collected from YouTube with a resolution of 720P or 1080P. We used the "WDA" subset since it is composed of the videos that have the highest quality among those in the whole dataset, for a total of 193 videos. Since the videos have a much larger duration than those of the DFD dataset, we used only the initial 30 seconds to reduce the computational cost; this does not hamper the evaluation since the visual content remains extremely similar. We relied on this dataset only for testing purposes, to compare the proposed approach with competing state-of-the-art methods and to evaluate the generalization capabilities of the models trained on the DFD dataset.

Starting from the raw (Constant Rate Factor 0) version of the original sequences, each video was compressed with the H.264 codec and CRF 32 and 42 using `FFmpeg` [62]. Then, only during training, the frames of each sequence were extracted by sampling one frame every five, both for the raw and compressed versions. In addition, for the compressed versions, the frames were extracted starting from a given offset measured in the number of frames to skip. This was because for the training the reference frames (i.e. the raw ones) need to precede the compressed ones. The offset used in the experiments was equal to 5.

Both for training and testing we relied on `dlib` [63] to detect the face rectangle and the 68 facial landmarks of each frame. Then, we leveraged an affine transformation to perform the crop and alignment of the detected faces based on the set of facial landmarks. Each reference image was warped to the corresponding degraded one with Moving Least Squares to reduce the difference in pose and expression. To this end, we extracted the facial landmarks of both images and then applied the MLS algorithm presented in Section III-C. Finally, we used the facial landmarks of the compressed frame to generate the landmarks binary images. After the preprocessing, we ended up with 9,007 images for the training set and 12,568 images for the test set, considering the DFD dataset. Instead, all the 175,832 frames of the HDTF dataset were used for testing.

B. Training Setup

To train both the generator and the discriminator we employed the ADAM optimizer [64] with batch size 4, learning rate 10^{-4} and momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. We trained all the models for 15 epochs because after that the outputs did not change significantly. We adopted several data augmentation techniques, such as shifting, 90° rotations and cutout [65]. We performed a grid search to find the optimal trade-off parameters for the training losses. After that, they were set as follows: $\lambda_{MSE} = 300$, $\lambda_{perc} = 10$, $\lambda_{style} = 1$, $\lambda_{adv} = 2$, $\lambda_{adv,1} = 4$, $\lambda_{adv,2} = 2$, $\lambda_{adv,4} = 1$ and $\lambda_{adv,8} = 1$. The 4 layers used to compute the perceptual loss were given the same weight, equal to 1. During testing, we set the maximum cardinality of the set of keyframes to 10.

C. Evaluation Metrics

The performance is evaluated using six full-reference and two no-reference visual quality metrics. Regarding the full-reference metrics, we employed: *1)* Peak Signal-to-Noise Ratio (PSNR), which is often used to evaluate reconstruction and compression artifacts reduction, despite its issues in estimating the perceived quality [66], [67]; *2)* Structural Similarity Index Measure (SSIM) [68], another commonly used metric, although it is known that it doesn't perform well on the output of generative models [69]; *3)* Learned Perceptual Image Patch Similarity (LPIPS) [70], using, in particular, the version with AlexNet [71] backbone. Typically LPIPS measures are in contrast with SSIM, i.e. distortions that are low for LPIPS are high in SSIM and vice-versa. LPIPS has been shown to have a very strong correlation with perceived visual quality; *4)* CONTRastive Image QUality Evaluator-Full Reference (CONTRIQUE-FR) [72], using, in particular, the LIVE _ FR model downloaded from the official repository; *5)* Video Multimethod Assessment Fusion (VMAF) [73], a full reference perceptual video quality assessment model that combines multiple elementary quality metrics; *5)* Video Multimethod Assessment Fusion - No Enhancement Gain (VMAF-NEG) [74], which subtracts the effect of image enhancement from the VMAF score. Indeed, VMAF tends to overpredict the perceptual quality when image enhancement techniques, such as sharpening or histogram equalization, are performed [74]. Both VMAF and VMAF-NEG include an elementary metric that accounts for the temporal difference between adjacent frames of the videos, thus evaluating the presence of motion jitter and flicker. Regarding the no-reference metrics, we relied on: *1)* Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [75], which evaluates the naturalness of an image; *2)* CONTRastive Image QUality Evaluator (CONTRIQUE) [72], using, in particular, the LIVE model downloaded from the official repository.

D. Baselines

We compare the proposed approach with several state-of-the-art methods: six methods for blind face restoration, HiFaceGAN [25], PSFR-GAN [27], GFP-GAN [31], GPEN [26], DFDNet [29] and ASFFNet [7], and one for face super-resolution, GWAINet [23]. DFDNet, PSFR-GAN, GPEN and GFP-GAN do not use a reference image but utilize extra face prior, respectively some offline-generated dictionaries of facial components, a segmentation mask and pretrained GANs. Instead, GWAINet exploits a reference image that is warped to the compressed one by means of a warper network. HiFaceGAN does not require any additional information w.r.t. the compressed input image. The most similar to our work is ASFFNet, which leverages a reference image and a binary landmark image. As ASFFNet needs a given static set of reference images, we make all the keyframes in the video available to it as possible guidance. Therefore, ASFFNet actually has an advantage over our approach, as, in our case, we limit the maximum cardinality of the set of keyframes to 10.

E. Quantitative Results

The quantitative results for the DFD dataset are reported in Table I. The proposed method achieves the best performance for the LPIPS metric, which is the most indicative full-reference perceptual metric, as well as in terms of CONTRIQUE, CONTRIQUE-FR and VMAF-NEG. PSFR-GAN performs better with regard to the signal metrics PSNR and SSIM, while GWAINet achieves the best result for BRISQUE. However, manual inspection shows that the images produced by GWAINet include excessive high-frequency artifacts and thus we did not consider this approach in the other experiments. GFP-GAN obtains the best VMAF value, probably because of its tendency to saturate colors and increase contrast at the cost of loss of photorealism, as is visible from the qualitative results. This tendency is similar to the application of image enhancement methods, which are known to boost the VMAF score [74]. In support of this theory, we can notice the large difference from the VMAF-NEG score, which in contrast is not affected by image enhancement techniques. Our method achieves both the second-best VMAF value and the best VMAG-NEG value, proving its ability to obtain great overall video quality while preserving photorealism. Moreover, the VMAF and VMAF-NEG scores show that our video results are temporal consistent and do not present too much motion jitter and flicker or mosquito noise.

In the second experiment, reported in Table II, we compare the proposed method with the baselines on the HDTF dataset. It is important to note that our model has not been trained on this dataset so that we can evaluate its generalization capabilities. Again, the proposed approach outperforms the other methods in terms of LPIPS, CONTRIQUE, CONTRIQUE-FR and VMAF-NEG. Manual examination of the results shows that this may be motivated by the fact that several competing approaches tend to add (or, on the opposite, hide) skin imperfections or boost excessively the color of lips and eyebrows.

Overall, our method is the one that performs best with the highest consistency, as none of the baselines achieves better performance on multiple metrics simultaneously. The results obtained for the HDTF dataset also prove that the proposed model is capable of generalization. In addition, we argue that the metrics for which our method performs best, namely LPIPS, CONTRIQUE, CONTRIQUE-FR, and VMAF-NEG are those

TABLE I
QUANTITATIVE COMPARISON BETWEEN THE PROPOSED APPROACH AND OTHER STATE-OF-THE-ART METHODS FOR CRF 42 ON DFD DATASET [60]

Method	PSNR ↑	SSIM ↑	LPIPS ↓	BRISQUE ↓	CONTRIQUE ↓	CONTRIQUE-FR ↓	VMAF ↑	VMAF-NEG ↑
GWAINet [23]	22.25	0.608	0.129	**24.18**	50.16	20.79	44.65	36.60
HiFaceGAN [25]	29.38	0.828	0.075	28.41	48.75	18.67	47.77	45.11
PSFR-GAN [27]	**29.68**	0.833	0.057	29.07	46.87	16.46	48.55	46.22
GFP-GAN [31]	27.51	0.822	0.081	34.17	50.84	23.01	**57.55**	48.51
GPEN [26]	27.61	0.813	0.075	28.67	49.42	21.36	55.86	49.26
DFDNet [29]	27.03	0.827	0.065	32.38	46.84	16.04	55.15	48.95
ASFFNet [7]	28.29	**0.834**	0.062	29.67	46.27	17.48	51.74	46.84
Ours	26.19	0.779	**0.037**	27.41	**44.95**	**13.16**	56.87	**54.20**

Best and second best results are in bold and underlined, respectively. ↑= higher values are better, ↓= lower values are better.

TABLE II
QUANTITATIVE COMPARISON BETWEEN THE PROPOSED APPROACH AND OTHER STATE-OF-THE-ART METHODS FOR CRF 42 ON HDTF DATASET [61]

Method	PSNR ↑	SSIM ↑	LPIPS ↓	BRISQUE ↓	CONTRIQUE ↓	CONTRIQUE-FR ↓	VMAF ↑	VMAF-NEG ↑
HiFaceGAN [25]	**30.70**	**0.864**	0.047	31.71	41.50	10.69	44.50	42.40
PSFR-GAN [27]	30.31	0.853	0.046	**30.01**	44.99	13.57	40.18	38.40
GFP-GAN [31]	28.19	0.846	0.064	35.02	46.80	13.95	**51.06**	41.27
GPEN [26]	27.72	0.817	0.061	30.62	46.15	13.31	47.52	40.57
DFDNet [29]	28.16	0.847	0.050	32.10	47.50	12.04	49.48	43.90
ASFFNet [7]	27.88	0.835	0.058	32.18	45.39	12.90	40.15	35.72
Ours	30.39	0.862	**0.028**	33.34	**37.35**	**7.76**	47.82	**45.07**

Best and second best results are in bold and underlined, respectively. ↑= higher values are better, ↓= lower values are better.

that correlate best with the actual quality of the restored frames. In Appendix A we provide some examples that support this argument.

F. Qualitative Results

Qualitative results for the DFD dataset are shown in Fig. 6. Our approach outperforms all the baselines in generating photorealistic and detailed results. GWAINet, HiFaceGAN and PSFR-GAN produce unsatisfactory images that still present visible artifacts, see for example the mouth in the second row. GFP-GAN and GPEN generate detailed but artsy and not photorealistic results, as the eyes in the first and fifth rows. DFDNet and ASFFNet achieve a better tradeoff between details and photorealism but, as can be seen in the last row, still produce visible artifacts. Our model exploits the reference keyframe and reproduces the high-frequency details lost after such strong compression without loss of photorealism. It is interesting to note that often the reference image (i.e. the bottom-left image in the input column) is not too similar to the degraded image, but the proposed method is still able to exploit it. For example, in the last row, the reference image has open eyes while the compressed one has them closed, and despite this, our model correctly depicts the restored frame with closed eyes.

Fig. 7 shows the qualitative results for the HDTF dataset. Again, our method produces the most detailed and photorealistic images. All the baselines generate blurry hair in both the first and second rows, as well as a not detailed beard in the third row. In the first row, PSFR-GAN, GFP-GAN, GPEN and ASFFNet mistake the shadow of the glasses for their border and thus produce unrealistic results. In the fourth row, GPEN and DFDNet hallucinate moles that are not present in the ground truth. In the fifth row, our method is the only one capable of depicting the eyes as closed without adding artifacts. In the last row, GFP-GAN and GPEN add traces of glasses, while ASFFNet exploits the reference incorrectly and portrays the eyes as open. In general, our method is the one that most consistently generates satisfactory results that are similar to the ground truth.

G. Subjective Experiments

In this experiment we conducted a subjective test based on the three-alternative forced choice (3-AFC) methodology, using the *AVrate Voyager* tool [76], [77]. The test included the inspection of 15 sets of videos, 8 from the DFD dataset and 7 from the HDTF one, so as to maintain the completion time of around 15-20 minutes and avoid excessive fatigue as recommended by ITU-R BT.500-13 [78]. Each original video was compressed with CRF 42 and restored using our proposed method, GPEN [26] and GFP-GAN [31]; using 3-AFC allowed to reduce the number of required comparisons [79]. Participants (18, i.e. almost double the minimum required [80]) were requested to choose the reconstruction that matched more closely the original high-quality video, without considering aesthetic preferences. The position of the results of all the methods was changed randomly for each evaluation. Fig. 8 reports the percentages of the forced choices for the 15 sets. The much larger preference given to our proposed method can be attributed to the fact the proposed GAN introduces fewer high-frequency details and color shifts than the GPEN [26] and GFP-GAN [31]; these additions tend to be more visible in a video sequence than when evaluating separately the frames using the quality metrics.

H. Inference Time

We compared the Frames Per Second (FPS) processed by our model with the baselines. The experiments were performed on an NVIDIA RTX 2080 Ti GPU. As shown in Table III, our

Fig. 6. Qualitative comparison between the proposed approach and the baselines for the DFD dataset and CRF 42. The bottom-left image in the input column represents the reference frame exploited by our approach. Best viewed in full screen.

TABLE III
FPS COMPARISON BETWEEN THE PROPOSED APPROACH AND OTHER STATE-OF-THE-ART METHODS

Method	# parameters	FPS
HiFaceGAN [25]	72.22M	<u>44</u>
PSFR-GAN [27]	<u>67.26M</u>	28
GFP-GAN [31]	86.44M	**49**
GPEN [26]	71.00M	39
DFDNet [29]	113.31M	4
ASFFNet [7]	**23.62M**	24
Ours	96.35M	<u>44</u>

Best and second best results are in bold and underlined, respectively.

method achieves a number of FPS similar to or better than the baselines but outperforms them in terms of quality. Given that our model runs at almost 45 FPS, it proves to be capable of real-time inference and therefore suitable for videoconferencing.

I. Ablation Studies

a) Architecture: We performed ablation studies to evaluate the importance of each component of our architecture. In particular, we measure the effect of using: *i)* Multi-scale features; *ii)* ASFF blocks; *iii)* SFT blocks. We start from a single-scale features model that considers only the features with the smallest size but with the most channels and that relies on concatenation instead of the ASFF and SFT blocks. Then, we gradually add each component: first individually and then in combination with each other. The results are reported in Table IV. Our experiments show how the use of multi-scale features is the most important component of the architecture, followed by the ASFF and SFT blocks. Additionally, we substitute the ASFF and SFT blocks one at a time with SPADE [81], a spatially-adaptive denormalization block. The proposed architecture outperforms both versions that make use of SPADE, proving that ASFF and SFT blocks are more effective in our architecture.

b) Keyframes Selection Policy: Tables V and VI compare the proposed LFU policy update method with a different approach that maximizes the diversity of the keyframes, called "Max

Fig. 7. Qualitative comparison between the proposed approach and the baselines for the HDTF dataset and CRF 42. The bottom-left image in the input column represents the reference frame exploited by our approach. Best viewed in full screen.

TABLE IV
ABLATION STUDIES WITH CRF 42 ON THE DFD DATASET

Ablation	PSNR ↑	SSIM ↑	LPIPS ↓	BRISQUE ↓	CONTRIQUE ↓	CONTRIQUE-FR ↓	VMAF ↑	VMAF-NEG ↑
SSF w/o ASFF w/o SFT	25.41	0.736	0.078	33.30	49.95	16.50	48.19	46.19
SSF w/ ASFF w/ SFT	24.99	0.736	0.079	30.96	49.97	15.83	47.90	45.83
MSF w/o ASFF w/o SFT	**26.19**	0.777	0.039	**25.84**	**44.34**	13.84	54.49	52.34
MSF w/o ASFF w/ SFT	26.17	0.776	0.038	26.21	45.25	13.87	55.29	52.99
MSF w/ ASFF w/o SFT	26.12	0.776	0.038	29.17	45.86	13.39	56.33	53.84
MSF w/ ASFF w/ SPADE	26.17	0.777	0.038	28.14	45.06	14.18	55.18	52.97
MSF w/ SPADE w/ SFT	26.18	0.776	0.039	27.57	45.12	13.78	55.87	53.39
MSF w/ ASFF w/ SFT	**26.19**	**0.779**	**0.037**	27.41	44.95	**13.16**	**56.87**	**54.20**

SSF stands for single-scale features, MSF for multi-scale features. Best results are in bold. ↑= higher values are better, ↓= lower values are better.

distance". The "Max distance" policy consists in maximizing the Euclidean distance between the facial landmarks of the frames of the set, in order to have a wide range of poses and expressions. The idea is that in this way, every future frame of the video should always have a reference in the set that is not too different. For each new keyframe, its distance to all the keyframes in the set is computed. Then, between all the possible combinations of frames, we choose the group of keyframes that maximizes the total distance, so the new keyframe is not necessarily added to the set.

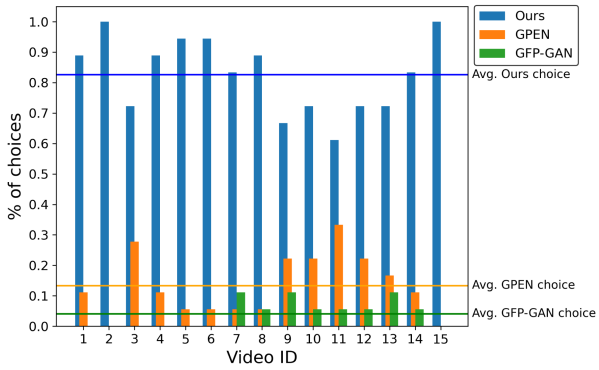

Fig. 8. Subjective results using 3-AFC. Videos from 1 to 8 belong to the DFD dataset, the others to the HDTF dataset.

TABLE V
ABLATION STUDIES ON THE KEYFRAMES SELECTION POLICY FOR THE DFD DATASET AND CRF 32

Strategy	PSNR ↑	SSIM ↑	LPIPS ↓	BRISQUE ↓	CONTRIQUE ↓	CONTRIQUE-FR ↓	VMAF ↑	VMAF-NEG ↑
Max distance	29.29	0.844	0.022	**27.89**	43.83	13.79	66.99	63.53
LFU	**29.32**	**0.845**	**0.021**	27.92	**43.77**	**13.75**	**67.18**	**63.69**

Best results are in bold. ↑= higher values are better, ↓= lower values are better.

TABLE VI
ABLATION STUDIES ON THE KEYFRAMES SELECTION POLICY FOR THE DFD DATASET AND CRF 42

Strategy	PSNR ↑	SSIM ↑	LPIPS ↓	BRISQUE ↓	CONTRIQUE ↓	CONTRIQUE-FR ↓	VMAF ↑	VMAF-NEG ↑
Max distance	**26.22**	0.776	0.039	27.50	45.09	13.29	55.85	53.24
LFU	26.19	**0.779**	**0.037**	**27.41**	**44.95**	**13.16**	**56.87**	**54.20**

Best results are in bold. ↑= higher values are better, ↓= lower values are better.

TABLE VII
ABLATION STUDIES ON THE MAXIMUM CARDINALITY OF THE SET OF REFERENCES FOR THE DFD DATASET AND CRF 42

Max cardinality	PSNR ↑	SSIM ↑	LPIPS ↓	BRISQUE ↓	CONTRIQUE ↓	CONTRIQUE-FR ↓	VMAF ↑	VMAF-NEG ↑
1	26.23	**0.779**	0.038	27.68	45.07	13.18	56.63	53.95
3	**26.25**	**0.779**	**0.037**	27.50	45.01	13.19	56.82	54.14
5	**26.25**	**0.779**	**0.037**	27.46	45.00	**13.16**	**56.87**	**54.20**
10	26.19	**0.779**	**0.037**	**27.41**	**44.95**	**13.16**	**56.87**	**54.20**

Best results are in bold. ↑= higher values are better, ↓= lower values are better

Table V reports the results obtained with CRF 32 and Table VI those for CRF 42. The maximum number of keyframes in the group was set to 10 in both cases. The results show that the proposed LFU strategy outperforms the "Max distance" one for almost all the metrics.

c) Keyframes Set Cardinality: Regarding the dimension of the set of keyframes, we expect that as the maximum cardinality increases, the results will improve. In fact, having more possible references available, it is less likely that a compressed frame has no similar reference. The results reported in Table VII confirm our assumption, but the increase in performance is not too significant. However, we set the maximum cardinality to 10 because the time needed to choose the best keyframe is still about 0.1 milliseconds so a higher number of keyframes to choose from does not impact the computational complexity significantly.

d) Feature Extractor: In this experiment We replace the VGG-19 backbone with different feature extractors. In particular, we exploit the small and large versions of MobileNetV3 [82], a popular and light CNN designed for mobile platforms which, from our experiments, reduces the inference time of our model by about two times. Table VIII reports the quantitative results.

TABLE VIII
ABLATION STUDIES ON FEATURE EXTRACTOR FOR THE DFD DATASET AND CRF 42

Features extractor	# parameters	PSNR ↑	SSIM ↑	LPIPS ↓	BRISQUE ↓	CONTRIQUE ↓	CONTRIQUE-FR ↓	VMAF ↑	VMAF-NEG ↑
MobileNetV3 Small	**2.72M**	25.87	0.784	0.047	33.18	44.96	13.00	50.30	48.19
MobileNetV3 Large	8.45M	26.11	**0.786**	0.045	31.37	45.94	**11.91**	50.26	48.01
VGG-19	96.35M	**26.19**	0.779	**0.037**	**27.41**	**44.95**	13.16	**56.87**	**54.20**

Best results are in bold. ↑= higher values are better, ↓= lower values are better.

TABLE IX
ABLATION STUDIES ON THE DISCRIMINATOR FOR THE DFD DATASET AND CRF 42

Discriminator	PSNR ↑	SSIM ↑	LPIPS ↓	BRISQUE ↓	CONTRIQUE ↓	CONTRIQUE-FR ↓	VMAF ↑	VMAF-NEG ↑
Single-scale	**26.49**	**0.804**	0.039	32.97	**44.87**	**11.88**	50.44	48.85
Multi-scale	26.19	0.779	**0.037**	**27.41**	44.95	13.16	**56.87**	**54.20**

Best results are in bold. ↑= higher values are better, ↓= lower values are better.

TABLE X
ABLATION STUDIES ON THE DISCRIMINATOR FOR THE HDTF DATASET AND CRF 42

Discriminator	PSNR ↑	SSIM ↑	LPIPS ↓	BRISQUE ↓	CONTRIQUE ↓	CONTRIQUE-FR ↓	VMAF ↑	VMAF-NEG ↑
Single-scale	30.34	**0.873**	0.030	34.25	37.52	8.22	40.21	38.94
Multi-scale	**30.39**	0.862	**0.028**	**33.34**	**37.35**	**7.76**	**47.82**	**45.07**

Best results are in bold. ↑= higher values are better, ↓= lower values are better.

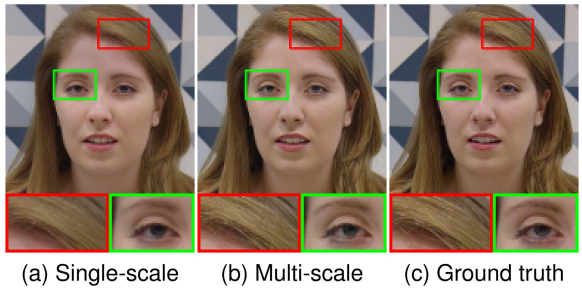

Fig. 9. Qualitative results for different discriminators for max cardinality 10 and CRF 42 on the DFD dataset.

As expected, the version with the VGG-19 outperforms the MobileNetV3 ones, but the number of parameters is an order of magnitude greater. However, looking at the qualitative results obtained with the MobileNetV3 as the feature extractor we noticed how they were still more than acceptable, proving how effective our approach is, and suggesting that these backbones could be used for deployment on mobile devices.

e) Discriminator: We substitute the multi-scale discriminators with a standard single-scale discriminator. Consequently, we also replace the adversarial loss described in (10) with the following one:

$$\begin{aligned} \ell_{adv,D} = & -\mathbb{E}_{I_{GT}\sim P(I_{GT})}\left[\min(0,-1+D(I_{GT}))\right] \\ & -\mathbb{E}_{I_R\sim P(I_R)}\left[\min(0,-1-D(I_R))\right] \\ \ell_{adv,G} = & -\mathbb{E}_{I_D\sim P(I_D)}\left[D(G(I_D))\right] \end{aligned} \tag{12}$$

$\ell_{adv,D}$ and $\ell_{adv,G}$ were used to update respectively the discriminator and the generator.

Tables IX and X report the quantitative results for the DFD and HDTF datasets, respectively. Even if the version with the single-scale discriminator outperforms the multi-scale one for some metrics, the qualitative results show clearly that the use of the multi-scale discriminators allows to obtain less blurry and more sharp and detailed outputs. This is proven also by the lower values of the LPIPS metric for both datasets. For instance, Fig. 9 shows how the multi-scale version has less blurred and

more detailed hair and eyes than the single-scale one, as well as an overall color more faithful to the ground truth. Additionally, the multi-scale discriminators let to achieve higher VMAF and VMAF-NEG values, which correspond to a better temporal consistency.

V. Conclusion

In this paper we have proposed a novel GAN-based method and a keyframes selection system that improves the visual quality of videoconference videos enhancing the appearance of faces. A key element of the system is the policy that updates a set of previous I-frames and exploits them to improve the visual quality improvement process. The proposed approach improves over competing state-of-the-art methods in terms of perceptual metrics and is rated much better in terms of fidelity by human evaluators.

Appendix
Quantitative Results Analysis

In Section IV-E we reported the quantitative results for the DFD and HDTF datasets. Our method obtains the best performance in terms of LPIPS, CONTRIQUE, CONTRIQUE-FR and VMAF-NEG. We argue that these metrics best correlate with the perceived visual quality. In Figs. 10 and 11 we show two examples supporting our argument. In Fig. 10 we compare a frame restored by our method and by GWAINet and present the corresponding values of no-reference metrics BRISQUE and CONTRIQUE. The proposed approach clearly generates a more satisfying image than GWAINet, which adds high-frequency artifacts. We argue that these artifacts deceive BRISQUE, which mistakes them for high-frequency details that are distinctive of high-quality images [83]. In Fig. 11 we report the values of the full-reference metrics PSNR, SSIM, LPIPS and CONTRIQUE-FR obtained by our approach and HiFaceGAN for a restored frame. Again, the proposed method produces a more detailed and photorealistic image, while HiFaceGAN generates a frame with visible artifacts. However, HiFaceGAN obtains better values for PSNR and SSIM. PSNR and SSIM are signal-based metrics that do not correlate well with the perceived visual quality for the output of generative models [66], [67], [69]. On the contrary, LPIPS and CONTRIQUE-FR are perceptual-based metrics and are good indicators of the actual perceived quality of an image.

Regarding VMAF, it is known that image enhancement techniques tend to boost its values [74]. As Fig. 12 shows, some baselines, such as GFP-GAN, saturate colors and increase the contrast of the restored frames, making them more visually pleasing but less similar to the ground truth. This tendency is similar to the application of image enhancement methods. We argue that this is the reason why such baselines perform so well in terms of VMAF. Our argument is supported by the large difference between the values of the baselines for VMAF and VMAF-NEG, which is not affected by image enhancement techniques, in Tables I and II. On the contrary, our method obtains high values for both metrics without a substantial difference between them, meaning that they are due to the actual quality and temporal consistency of the results, and not due to color enhancement.

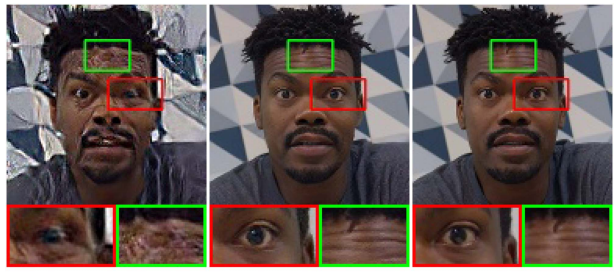

Fig. 10. Comparison between our method and GWAINet [23]. The reported values represent BRISQUE ↓/CONTRIQUE ↓, respectively, where ↓ means that lower values are better. Best results for each image are highlighted in bold.

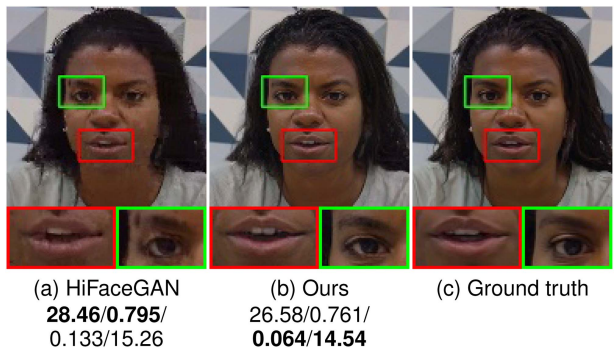

Fig. 11. Comparison between our approach and HiFaceGAN [25]. The reported values represent PSNR ↑/SSIM ↑/LPIPS ↓/CONTRIQUE-FR ↓, respectively. ↑= higher values are better, ↓= lower values are better. Best results for each image are highlighted in bold.

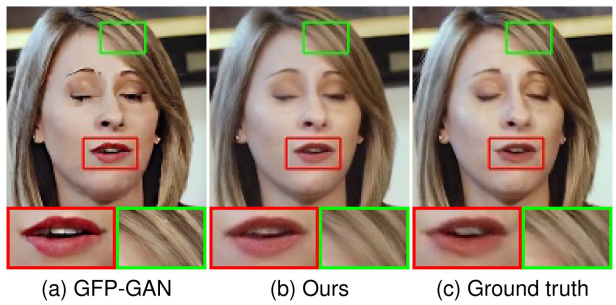

Fig. 12. Comparison between our approach and GFP-GAN.

References

[1] L. Galteri, L. Seidenari, M. Bertini, and A. D. Bimbo, "Deep generative adversarial compression artifact removal," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4836–4845.
[2] L. Galteri, L. Seidenari, M. Bertini, and A. D. Bimbo, "Deep universal generative adversarial compression artifact removal," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2131–2145, Aug. 2019.
[3] M.-Y. Liu, X. Huang, J. Yu, T.-C. Wang, and A. Mallya, "Generative adversarial networks for image and video synthesis: Algorithms and applications," *Proc. IEEE*, vol. 109, no. 5, pp. 839–862, May 2021.
[4] L. Galteri, M. Bertini, L. Seidenari, T. Uricchio, and A. Del Bimbo, "Increasing video perceptual quality with GANs and semantic coding," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 862–870, . [Online]. Available: https://doi.org/10.1145/3394171.3413508
[5] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10034–10044.

[6] M. Wijnants, S. Coppers, G. Rovelo Ruiz, P. Quax, and W. Lamotte, "Talking video heads: Saving streaming bitrate by adaptively applying object-based video principles to interview-like footage," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 2449–2458. [Online]. Available: https://doi.org/10.1145/3343031.3351045

[7] X. Li et al., "Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2703–2712.

[8] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 606–615.

[9] O. Rippel and L. Bourdev, "Real-time adaptive image compression," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2922–2930.

[10] O. Rippel et al., "Learned video compression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3454–3463.

[11] L. W. Kang, C. C. Hsu, B. Zhuang, C. W. Lin, and C. H. Yeh, "Learning-based joint super-resolution and deblocking for a highly compressed image," *IEEE Trans. Multimedia*, vol. 17, no. 7, pp. 921–934, Jul. 2015.

[12] C. Dong, Y. Deng, C. Change Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 576–584.

[13] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2810–2818.

[14] P. Svoboda, M. Hradiš, D. Barina, and P. Zemčík, "Compression artifacts removal using convolutional neural networks," *J. WSCG*, vol. 24, no. 2, pp. 63–72, 2016.

[15] Z. Wang et al., "D3: Deep dual-domain based fast restoration of JPEG-compressed images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2764–2772.

[16] L. Yu, T. Tillo, J. Xiao, and M. Grangetto, "Convolutional neural network for intermediate view enhancement in multiview streaming," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 15–28, Jan. 2018.

[17] L. Cavigelli, P. Hager, and L. Benini, "CAS-CNN: A deep convolutional neural network for image compression artifact suppression," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2017, pp. 752–759.

[18] J. Yoo, S.-h. Lee, and N. Kwak, "Image restoration by estimating frequency distribution of local patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6684–6692.

[19] D. Maleki, S. Nadalian, M. M. Derakhshani, and M. A. Sadeghi, "Block-CNN: A deep network for artifact removal and image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 2555–2558.

[20] F. Vaccaro, T. Uricchio, M. Bertini, and A. D. Bimbo, "Fast video visual quality and resolution improvement using SR-UNet," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 1221–1229.

[21] Z. Guan et al., "MFQE 2.0: A new approach for multi-frame quality enhancement on compressed video," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019, pp. 949–963.

[22] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1954–1963.

[23] B. Dogan, S. Gu, and R. Timofte, "Exemplar guided face image super-resolution without facial landmarks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1814–1823.

[24] X. Li et al., "Recovering extremely degraded faces by joint super-resolution and facial composite," in *Proc. IEEE Int. Conf. Tools Artif. Intell.*, 2019, pp. 524–530.

[25] L. Yang et al., "HiFaceGAN: Face renovation via collaborative suppression and replenishment," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 1551–1560.

[26] T. Yang, P. Ren, X. Xie, and L. Zhang, "GAN prior embedded network for blind face restoration in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 672–681.

[27] C. Chen et al., "Progressive semantic-aware style transformation for blind face restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11891–11900.

[28] X. Li et al., "Learning warped guidance for blind face restoration," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 278–296.

[29] X. Li et al., "Blind face restoration via deep multi-scale component dictionaries," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 399–415.

[30] M. C. Doukas, S. Zafeiriou, and V. Sharmanska, "HeadGAN: One-shot neural head synthesis and editing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 14378–14387.

[31] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9164–9174.

[32] Y. Liu, "Face restoration network with feature prior," in *Proc. IEEE Int. Conf. Comput. Sci., Artif. Intell. Electron. Eng.*, 2021, pp. 222–226.

[33] J. Li, H. Huang, X. Jia, and R. He, "Universal face restoration with memorized modulation," 2021, *arXiv:2110.01033*.

[34] X. Zhang, X. Wu, X. Zhai, X. Ben, and C. Tu, "DAVD-Net: Deep audio-aided video decompression of talking heads," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020.

[35] T. Liu, M. Xu, S. Li, R. Ding, and H. Liu, "MRS-Net for enhancing face quality of compressed videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2881–2894, May 2021.

[36] X. Zhang and X. Wu, "Multi-modality deep restoration of extremely compressed face videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2024–2037, Feb. 2023.

[37] Y. Guo, X. Zhang, and X. Wu, "Deep multi-modality soft-decoding of very low bit-rate face videos," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 3947–3955.

[38] R. Yasarla, H. R. V. Joze, and V. M. Patel, "Network architecture search for face enhancement," 2021, *arXiv:2105.06528*.

[39] Z. Wang, J. Zhang, R. Chen, W. Wang, and P. Luo, "RestoreFormer: High-quality blind face restoration from undegraded key-value pairs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17512–17521.

[40] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[41] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5967–5976.

[42] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy, "ReenactGAN: Learning to reenact faces via boundary transfer," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 622–638.

[43] T.-C. Wang et al., "Few-shot video-to-video synthesis," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[44] A. Fischer and C. Igel, "An introduction to restricted Boltzmann machines," in *Proc. Iberoamerican Congr. Pattern Recognit.*, 2012, pp. 14–36.

[45] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[46] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," 2016, *arXiv:1701.00160*.

[47] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[49] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2017, pp. 1510–1519.

[50] S. Schaefer, T. McPhail, and J. Warren, "Image deformation using moving least squares," in *Proc. ACM SIGGRAPH*, 2006, pp. 533–540.

[51] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[52] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 105–114.

[53] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 658–666.

[54] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 41.1–41.12.

[55] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2414–2423.

[56] J. H. Lim and J. C. Ye, "Geometric GAN," 2017, *arXiv:1705.02894*.

[57] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[58] T.-C. Wang et al., "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8798–8807.

[59] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2018.

[60] A. Rössler et al., "FaceForensics : Learning to detect manipulated facial images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1–11.
[61] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3661–3670.
[62] F. Bellard, "FFmpeg." Accessed: Aug. 23, 2021. [Online]. Available: https://www.ffmpeg.org/
[63] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.
[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[65] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.
[66] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, pp. 800–801, 2008.
[67] Q. Huynh-Thu and M. Ghanbari, "The accuracy of PSNR in predicting video quality for different video scenes and frame rates," *Telecommun. Syst.*, vol. 49, no. 1, pp. 35–48, 2012.
[68] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[69] H. Ko, D. Y. Lee, S. Cho, and A. C. Bovik, "Quality prediction on deep generative images," *IEEE Trans. Image Process.*, vol. 29, pp. 5964–5979, 2020.
[70] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.
[72] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Image quality assessment using contrastive learning," *IEEE Trans. Image Process.*, vol. 31, pp. 4149–4161, 2022.
[73] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," 2016. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652
[74] Z. Li, K. Swanson, C. Bampis, L. Krasula, and A. Aaron, "Toward a better quality metric for the video community," 2020. [Online]. Available: https://netflixtechblog.com/toward-a-better-quality-metric-for-the-video-community-7ed94e752a30
[75] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput.*, 2011, pp. 723–727.
[76] R. Rao Ramachandra Rao, S. Göring, and A. Raake, "Towards high resolution video quality assessment in the crowd," in *Proc. Int. Conf. Qual. Multimedia Experience*, 2021, pp. 1–6.
[77] S. Göring, R. Rao Ramachandra Rao, S. Fremerey, and A. Raake, "AVRate voyager: An open source online testing platform," in *Proc. IEEE 23rd Int. Workshop Multimedia Signal Process.*, 2021, pp. 1–6.
[78] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, Rec. ITU-R BT.500-13, ITU, Geneva, Switzerland, 2012.
[79] J. B. Phillips and H. Eliasson, "Chapter 5: Subjective image quality assessment," in *Camera Image Quality Benchmarking*. Hoboken, NJ, USA: Wiley, 2017.
[80] S. Winkler, "On the properties of subjective ratings in video quality experiments," in *Proc. IEEE Int. Workshop Qual. Multimedia Experience*, 2009, pp. 139–144.
[81] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019.
[82] A. Howard et al., "Searching for MobileNetv3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.
[83] L. Seidenari, L. Galteri, P. Bongini, M. Bertini, and A. Del Bimbo, "Language based image quality assessment," in *Proc. ACM Multimedia Asia*, 2022, pp. 1–7. [Online]. Available: https://doi.org/10.1145/3469877.3490605

Lorenzo Agnolucci received the M.S. degree (*cum laude*) in computer engineering in 2021 from the University of Florence, Florence, Italy, where he is currently working toward the Ph.D. degree with Media Integration and Communication Center. His research interests include machine learning and computer vision, focusing on low-level vision, video restoration, and multimodal learning.

Leonardo Galteri received the master's degree in computer engineering from the University of Florence, Florence, Italy, in 2014, and the Ph.D. degree in 2018, with a thesis that focused on semantic video compression and object detection. He is currently a Research Fellow with Telematic University Pegaso, Naples, Italy. He has authored eight journal papers and more than 15 peer-reviewed conference papers, attaining an H-index with more than 700 citations. His research interests include the application of deep learning techniques for video restoration and deepfake detection. He is a co-founder of Small Pixels, an academic spin-off working on visual quality improvement based on AI.

Marco Bertini is currently an Associate Professor of computer science with the School of Engineering, University of Florence, Florence, Italy, and the Director of the Media Integration and Communication Center (MICC) at the same university. His research interests include computer vision, multimedia, pattern recognition, and their application to different domains, such as cultural heritage. He has been the General Co-Chair, Program Co-Chair, and Area Chair of several international conferences and workshops on multimedia and computer vision, such as ACM MM, ICMR, and CBMI. He was an Associate Editor for IEEE TRANSACTIONS ON MULTIMEDIA. He has been involved in different roles in more than ten EU research projects. He is a co-founder of Small Pixels, an academic spin-off working on visual quality improvement based on AI.

Alberto Del Bimbo (Senior Member, IEEE) received the master's degree (*cum laude*) in electrical engineering in 1977. He is currently a Full Professor of computer engineering, and the Director of the Media Integration and Communication Center, University of Florence, Florence, Italy. From 1996 to 2000, he was the President of the IAPR Italian Chapter and from 1998 to 2000, the Member-at- Large with the IEEE Publication Board. His research interests include multimedia information retrieval, pattern recognition, and computer vision. Prof. Bimbo was the recipient of the SIGMM Technical Achievement Award for Outstanding Technical Contributions to Multimedia Computing, Communications, and Applications. He was nominated the ACM Distinguished Scientist in 2016. He is a co-founder of Small Pixels, an academic spin-off working on visual quality improvement based on AI.