Video Coding using Learned Latent GAN Compression

Mustafa Shukor
Bharath Bhushan Damodaran
InterDigital, Inc.
Rennes, France
mustafa.shukor97@gmail.com
bharath.damodaran@interdigital.com

Xu Yao
Pierre Hellier
InterDigital, Inc.
Rennes, France
pierre.hellier@interdigital.com

ABSTRACT
We propose in this paper a new paradigm for facial video compression. We leverage the generative capacity of GANs such as StyleGAN to represent and compress a video, including intra and inter compression. Each frame is inverted in the latent space of StyleGAN, from which the optimal compression is learned. To do so, a diffeomorphic latent representation is learned using a normalizing flows model, where an entropy model can be optimized for image coding. In addition, we propose a new perceptual loss that is more efficient than other counterparts. Finally, an entropy model for video inter coding with residual is also learned in the previously constructed latent representation. Our method (SGANC) is simple, faster to train, and achieves better results for image and video coding compared to state-of-the-art codecs such as VTM, AV1, and recent deep learning techniques. In particular, it drastically minimizes perceptual distortion at low bit rates.

1 INTRODUCTION
With the explosion of videoconferencing, the efficient transmission of facial video is a key industrial problem. In addition, building digital worlds raises the question of transmitting facial representations, where perceptual distortion matters more than exact fidelity. Image compression can be formulated as an optimization problem with the objective of finding a codec which reduces the bit-stream size for a given distortion level between the reconstructed image at the receiver side and the original one. The distortion mainly occurs due to the quantization in the codec, because the entropy coding method [26, 51] requires the discrete data to create the bit-stream. The compression quality of the codec depends on the modeling of the data distribution close to the real data distribution since the expected optimal code length is lower bounded by the entropy [54]. Existing compression methods suffer from various artefacts. Especially at very low bits-per-pixel (BPP), blocks and blur degrade the image quality, leading to a poorly photorealistic image.

Motivated by the appealing properties of the StyleGAN architecture for high quality image generation, we propose a new compression method for facial images and videos. The intuition is that the GAN latent representation associated to any face image is somehow disentangled, and a perceptual compression method should be easier to train using the latent, both for intra and inter coding. In addition, at extremely low BPP, a compressed latent code should always lead to a photorealistic image, hence leading to a compression technique that is perceptually more pleasant. To the best of our knowledge, no work has studied how the latent space of StyleGAN can be used for efficient perceptual compression.

However, we acknowledge that the method relies on the hypothesis that a GAN inversion technique can approximate accurately any input facial image. While this hypothesis is currently a limitation, we note that the performances of GAN inversion and generation methods have significantly improved recently and is still a hot topic [57, 66]. We take the leap of faith that the hypothesis will become valid in the near future. In addition, specialized GANs could be fine-tuned for a limited set of faces, leading to a valid GAN inversion in dedicated applications. For such applications of personalized videoconferencing, our system is valid if the fine-tuned GAN is known in advance at the receiver side. Finally, for the scenario of digital copies of our physical world (e.g., metaverse), transmitting facial videos requires perceptual distortions to be minimal, even at the cost of some loss of fidelity to the original face.

For real natural images, a StyleGAN encoder (GANs inversion) [49, 64] have been proposed to project any image onto the latent space of StyleGAN. Our objective is to compress this latent representations. Although, training GANs encoders and generators are computationally heavy, which is not desirable for compression.

In order to overcome the above challenges, we propose to learn a new proxy latent space representation using a diffeomorphic mapping function. This new latent representation is precisely learned to optimize the compression efficiency. The approach is illustrated in Figure 1. The advantages of such space are three-fold; (a) it makes...
the approach very efficient as we can use off the shelf pretrained StyleGAN encoder and generator without the need to retrain them for each quality level. (b) it allows to learn a representation optimal for compression, associated with an efficient entropy model. (c) it allows to obtain high quality photorealistic images thanks to StyleGAN. Finally, we extend this approach to videos and propose a new method to optimize for the inter-coding with residuals in the new latent representation, as illustrated in Figure 2.

While we focus in this paper on facial video compression, we think the method we propose here could be extended to different types of scenes. Recently, [53] has proposed to extend StyleGAN to ImageNet classes. As a result, learning scene-specific compression paves the way towards semantic perceptual compression methods.

Our contributions can be summarized as follows:

- We propose a new paradigm for facial video compression, leveraging the generative power of StyleGAN for artifact-free and high quality image compression.

- Without any GAN retraining, we propose to learn a proxy latent representation that effectively fits the entropy model, while using off-the-shelf pretrained GAN encoder/decoder models. We propose to learn a model for intra and inter video coding.

- We propose a new perceptual distortion loss that is more efficient to compute and leverages the multiscale and semantic representation in the latent space of StyleGAN.

- We show high quality and lower perceptually-distorted reconstructed images for low BPP compared to traditional and deep learning based methods for image compression. We show better qualitative and quantitative results compared to the most recent state-of-the-art methods for video compression.

The rest of the paper is organised as follows: section 2 presents related works on image and video compression. Our proposed method is detailed in section 3 and section 4 presents the experimental results.

2 RELATED WORK

Image Compression: Traditional image codecs such as JPEG [60], JPEG2000 [47] have been carefully human-engineered to extract and compress image features. Recently, deep learning based codecs [2, 4, 5, 42] have gained significant attraction in the community due to their superior performance. These approaches learn nonlinear transformations of the input data and an entropy model jointly with the objective of maximizing compression ratio while maintaining high reconstruction quality. To this end, several methods use variational autoencoder (VAE) type architecture [4, 14, 43] and achieve impressive performance at higher BPP, however they are sub-optimal at low BPP. Some papers have targeted the entropy model: [6] propose to use a hyper prior as a source of side information to capture the spatial dependencies. [43] follow a similar approach and augment the hierarchical model with an autoregressive one. [14] build the previous approach with a discretized Gaussian mixture entropy model with attention modules. Few papers have improved the model structure and the architecture [12, 13, 37] and some used RNNs [30, 58].

Others have leveraged the power of generative adversarial networks (GAN) [25] and used adversarial losses, especially for low bitrates [2, 50]. These approaches produce high quality and lower perceptually-distorted image reconstructions [42, 52], even for high bitrates [59]. These methods are computationally heavy as they require adversarial training for each quality level. We argue that this is not practical especially for data compression, where each quality level requires to train a new model from scratch.

The choice of distortion loss is crucial for better reconstruction, traditionally PSNR or MS-SSIM [63] are used. These metrics capture the pixel wise distortion and poorly capture the perceptual distortion. Several works have tried to remedy these limitations. Motivated by the success of perceptual losses such as LPIPS [68] or VGG16 [29] for various applications [17, 24, 36], some papers [11, 52] propose to include a perceptual distortion in addition to the...
pixel wise ones. These perceptual losses are based on computationally expensive networks such as VGG16 or learned perceptual metrics such as LPIPS. Moreover, the backbone networks are pretrained for an unrelated and discriminative task (e.g. Image classification pretraining on ImageNet).

**Video Compression**: Deep Video Compression systems also minimize the rate-distortion loss. In addition to the spatial redundancy (SR), temporal redundancy (TR) is reduced by incorporating motion estimation modules. Traditional methods [55, 65] rely on handcrafted and block based modules. Recently, deep learning based video compression systems proposed to replace the traditional modules by learned ones [41]. Motion estimation [18] and compensation are often used to address TR. Several improvements have been made to reduce TR: [38] leverage multiple frames to improve the motion compensation, and [27] use multi resolution flow maps to effectively compress locally and globally. However, training motion estimation modules make these compression systems less efficient during training. Some approaches perform TR reduction in latent space [16, 22, 28]. Others propose to perform frame interpolation [16], or an interpolation in the latent space of GANs [52]. Recently, [39] have proposed to train a variational autoencoder with an adversarial constraint. Temporal encoding is tackled using an LSTM network in latent space.

**Videoconferencing**: This is becoming ubiquitous nowadays due to the pandemic. Many papers have targeted facial video synthesis [40, 62, 67, 70], or compression artifacts removal [69]. However, there are few papers targeting specifically facial video compression. [62] was the first to propose a compression framework based on extracted keypoints and learned appearance model for video synthesis. Similarly, in [21, 45] a compression framework based on keypoints and adversarial training was proposed. Our approach is significantly simpler and more efficient to train as we do not incorporate pose, expression nor keypoints extraction, which is desirable to support a wide range of quality levels, as for deep learning based compression systems, each quality level a new model should be trained from scratch.

### 3 LEARNING PROXY REPRESENTATION FOR COMPRESSION

In this section, we describe our proposed method to learn the latent space dedicated for compression, illustrated in Figures 1 and 2. The method can be summarized as follows: an input facial video

\[ x = G(w) \quad w = M(z) \quad z \sim N(0, I) \quad (1) \]

It is shown that the latent space of GAN is semantically rich and enjoys several properties such as semantic interpolation [48]. In addition, the latent vector encoded in \( \mathbf{W}^* \) space of StyleGAN captures a hierarchical representation of the projected image. In our case, we use StyleGAN2 [34] which is an improved version.

**StyleGAN Encoder** The StyleGAN encoder [49] is a deterministic function denoted as \( E \). Its role is to project a real image into the latent space of StyleGAN (e.g. \( \mathbf{W} \) or \( \mathbf{W}^* \)), in such a way that the reconstructed image by the StyleGAN generator is minimally distorted (\( \hat{x} = G(E(x)) \approx x \)). In our case, the image is projected in \( \mathbf{W}^* \) with dimension of \( (18 \times 512) \) where each dimension controls a different convolution layer of the StyleGAN generator. Currently, the encoding based GAN inversion approaches are not ideal, which explains the difference between the projected and the original image.

#### 3.2 Image Compression (SGANC)

We assume that we have a pretrained and fixed StyleGAN2 generator \( G \) that considers a latent code \( w \in \mathbf{W}^* \) and generates a high resolution image of size \( 1024 \times 1024 \), and an encoder \( E \) (pretrained and fixed) that embeds any given image \( x \) to the latent code \( w \in \mathbf{W}^* \) such that \( G(E(x)) \approx x \). Our objective is to learn a new latent space \( \mathbf{W}^*_c \) optimal for image compression. The \( \mathbf{W}^*_c \) is obtained using a bijective transformation \( T : \mathbf{W}^* \rightarrow \mathbf{W}^*_c \), which is parametrized as a normalizing flows model (our work only requires the bijectivity, as such, no maximum-likelihood term is included in the training objective). \( T \) maps a latent code \( w \in \mathbf{W}^* \) into \( w^*_c \in \mathbf{W}^*_c \) such that the latent vectors \( w^*_c \) has the minimum entropy and the sufficient information to generate the original image with minimal distortion from the inverted latent code \( T^{-1}(w^*_c) \) using \( G \). The entropy model is trained on the latent codes \( w^*_c \in \mathbf{W}^*_c \) by minimizing the following rate loss after the quantization:

\[ R = -\mathbb{E} \sum_i^{D} \log_2 p_i(Q(T(w))) = -\mathbb{E} \sum_i^{D} \log_2 p_i(T(w) + \epsilon) \quad (2) \]

For (2) to be differentiable, following [5], we relax the hard quantization \( Q \) by adding uniform noise \( \epsilon \) to the latent vectors \( T(w) \).
where \(d\) is any distortion loss, in this paper we use the mean squared error (MSE) loss. The (3) can be seen as a perceptual distortion loss, and it is motivated by the fact that the latent space of GANs is semantically rich. We argue that this is true especially for StyleGAN, where the latent code is extracted from several layers of the StyleGAN encoder which allows to capture multiscale and semantic features/representations. Compared to existing perceptual losses, our loss does not require to generate the images during training or to compute heavy losses such as VGG16 or LPIPS in the image space. The total loss used to learn our proposed latent space \(W^*_c\) is a trade-off between the rate (2) and distortion (3) as shown below:

\[
L = R + \lambda D
\]

where \(\lambda\) is the trade-off parameter, and the transformation \(T\) and the entropy model \(p\) are jointly optimized. Once the optimization is completed, to create the bit-stream the latent codes in \(W^*_c\) are quantized with \(Q\) using rounding operator.

3.3 Video Compression (SGANC IC)

Here, we propose our approach for video compression, denoted as SGANC IC, using learned inter coding with residuals. Video compression methods and standards rely on motion estimation and compensation modules to leverage the temporal dependencies between frames. As our approach is formulated in the latent space, our inter coding schema is based on the successive latent differences, since these differences reflect the temporal changes.

Specifically, inter coding with residuals is performed in the latent space (Figure 2). Given a sequence of frames \(\{x_1, x_2, \ldots, x_t, \ldots\}\), a pretrained encoder \(E\) is first used to obtain the latent representation. Then, similarly to intra coding, a transformation \(T\) is learned to map these frames to \(W^*_n\), leading to a sequence of latent codes \(\{w^*_1, w^*_2, \ldots, w^*_t, \ldots\}\). The mapping \(T\) is learnt such that the sequence of latent codes in \(W^*_IC\) is optimal for inter-coding by taking the temporal dependencies into account. Our approach can be summarized as follows (the complete description of the algorithm can be retrieved in the supplementary material).

- The first latent code is coded using the method described in section 3.2: \(w^*_0\) (using the same entropy model or preferably another one trained for image compression). The following steps are repeated until the end of the video;
- The difference between two consecutive latent codes are computed and quantized: \(\hat{v}_t = Q(w^*_t - w^*_t-1)\).
- From the previous reconstructed code, an estimate of the latent code at frame \(t\) is obtained: \(\hat{w}_t = w^*_{t-1} + \hat{v}_t\).
- The residual between the estimated and the actual latent code is computed and quantized as \(\hat{r}_t = Q(w^*_t - \hat{w}_t)\) (for all the frames or each gap (i.e., \(g\) frames).
- The quantized difference \(\hat{v}_t\) and the residual \(\hat{r}_t\) are compressed using an entropy coding and sent to the receiver.
- On the receiver side, the current latent code is reconstructed from the residual and the estimated latent code (or only from the estimated latent code each \(g\) frames): \(\hat{w}^*_t = w^*_{t-1} + \hat{v}_t + \hat{r}_t\).
- The latent codes in \(W^*_IC\) are remapped to \(W^*\) to generate the images using the pretrained StyleGAN2 G.

To learn the new latent space \(W^*_IC\) for inter-coding, the transformation \(T\) and the entropy model \(p\) are learned to optimize the rate-distortion loss:

\[
\mathcal{L}_{IC} = \lambda d(T, T^{-1}(w^*_t)) - \mathbb{E} \left[ \sum_i D_q(\hat{p}_t(\hat{v}_t)) \right]
\]

As similar to section 3.2, we replace the quantization \(Q\) by adding uniform noise, and the entropy model is modeled as in [6]. While (5) has two entropy models, we show below, that it is sufficient to learn only one entropy model on the differences \(\hat{v}_t\), as \(\hat{r}_t\) admits the explicit probability distribution known as the Irwin-Hall distribution (the proof of the following Lemma 1 can be found in the supplementary material).

**Lemma 1.** Let \(Q(x) = x + \epsilon\) be the continuous relaxation of the quantization, where \(\epsilon\) follows the uniform distribution \(\mathcal{U}_{[-0.5, 0.5]}\). Let \(w^*_0, w^*_1, \ldots, w^*_t, \ldots \in \mathbb{R}^n\), \(w^*_0 = Q(w^*_0)\), and \(w^*_t = w^*_t - Q(w^*_t)\). The mapping \(T\) is learnt such that the sequence of latent codes in \(W^*_IC\) is optimal for inter-coding by taking the temporal dependencies into account. Our approach can be summarized as follows (the complete description of the algorithm can be retrieved in the supplementary material).

This leads to optimizing our latent space \(W^*_IC\) for the rate-distortion loss only with one entropy model for \(\hat{v}_t\) by discarding the last term in (5). During test, the explicit distribution of the residuals can be used for entropy coding.

**Stage-Specific entropy models:** In [33], it is shown that each stage/layer of the StyleGAN generator corresponds to a specific scale of details. To leverage this hierarchical structure, we propose stage specific entropy models. Specifically, the first layers, which correspond to coarse resolution (e.g., \(4^2-8^2\)) affect mainly high level aspects of the image, such the pose and face shape, while the last layers affect the low level aspects such as textures, colors and small micro structures. Here, we propose to leverage this hierarchical structure and weight the distortion \(\lambda\) differently for each layer of the generator (Note that, the latent code in \(W^*\) or \(W^*_IC\) consists of 18 latent codes of dimension 512 and each one corresponds to one layer in the generator, hence its dimension is (18, 512)). For practical reasons, we split the 18 layers in three stages (\(1=8, 8=13, 13=18\)), and learn one transformation and entropy model for each group. The distortion \(\lambda\) is chosen to be higher for the first layers and decreases subsequently.
Figure 2: An illustration of our proposed approach for Video compression using Inter Coding with residual during test (Section 3.2). \( Q \) corresponds to quantization, compression encoding and decoding. Here \( g = 1 \).

Figure 3: Rate distortion curves on MEAD intra dataset: for medium and large BPP, our method (SGANC in blue) is better in terms of LPIPS and MS-SSIM than VTM, AV1, MeanHP and HP. For high BPP, our method is better in terms of PSNR. The perceptual metrics LPIPS and PIM clearly show that our method outperforms existing methods perceptually.

4 EXPERIMENTS

4.1 Datasets and Implementation Details

Celeba-HQ [31], a dataset consists of 30000 high quality face images of 1024 \( \times \) 1024 resolution. FILMPAC, a dataset consists of 5 video clips with high resolution and the length varies between 60 and 260 frames. These videos can be found on the filmpac website [23] by searching their names (FP006734MD02, FP006940MD02, FP009971HD03, FP010363HD03, FP010263HD03). MEAD [61], a high resolution talking face video corpus for many actors with different emotions and poses. The training dataset for inter-coding consists of 2.5 k videos with frontal poses. For evaluation, we created MEAD-inter dataset consisting of 10 videos (selected from MEAD) of different actors with frontal pose. We also created MEAD-intra dataset consisting of 200 frames selected from these videos with frontal pose for evaluating image compression.

Dataset preprocessing: All the frames are cropped and aligned using the same preprocessing as that of FFHQ dataset [33], on which the StyleGAN is pretrained. As we compare the reconstructed image with the projected one for SGANC, we project all the frames (encode the original images and reconstruct them using StyleGAN2). All frames are of high resolution (1024 \( \times \) 1024). We also created MEAD-intra dataset consisting of 200 frames selected from these videos with frontal pose for evaluating image compression.

Implementation details: We used StyleGAN2 generator (\( G \)) [34] pretrained on FFHQ dataset [33]. The images are encoded in \( W^* \) using pSp, a pretrained StyleGAN2 encoder (\( E \)) [49]. The parameters of the generator and the encoder remain fixed in all the experiments. The latent vector dimension in \( W^* \) and \( W_e^* \) is 18 \( \times \) 512. The mapping function \( T \) is modeled using RealNVP architecture [15] without batch normalization. This NF model consists of 13 coupling layers, and each coupling layer consists of 3 fully connected (FC) layers for the translation function and 3 FC for the scale one with LeakyReLU as hidden activation and Tanh as output one (Total number of trainable parameters=20.5 M). For the entropy model, we have used the fully factorized entropy model [6] assuming the dimensions are independent, based on the implementation from the CompressAI library [7]. For the stage specific entropy model, \( \lambda \) is kept constant for the first stage and then decreased linearly to be \( 1e-2 \) smaller for the last layer. For all the experiments, we used Adam optimizer with \( \beta_1 = 0.9 \) and \( \beta_2 = 0.999 \), learning rate=\( 1e-4 \) and the batch size=8. Once the training is completed, we used RANS coder [19] to obtain the bit-stream.

Training: We encoded all the images once, and the training is performed solely using the latent codes. Thus, we do not need the generator nor the encoder during training, which makes the approach light and fast to train. For SGANC, We train on Celeba-HQ dataset. For SGANC IC, we trained on 2.5 k videos from the MEAD dataset [61], where each batch contains video slices of size = 9 frames. All the frames are preprocessed as in section 4.1.

4.2 Results

In this section we present the experimental results of our proposed methods. We used the following metrics to assess the quality of the compression methods: Peak Signal to Noise Ratio (PSNR), Multi Scale Structural Similarity (MS-SSIM) [63], Learned Perceptual Image Patch Similarity (LPIPS) [68], and the Perceptual Information Metric (PIM) with a mixture of 5 Gaussian distributions [10].
Figure 4: Qualitative results for image compression with comparable BPP for all methods (images are better seen zoomed). Other methods introduce blocking artifacts and blurring. Our method (SGANC) leads to high quality reconstruction and perceptually lower distortion.

Figure 5: Rate distortion curves on MEAD inter dataset. Our approach SGAN-IC leads to a significantly lower perceptual distortion as measured by the perceptual metrics LPIPS and PIM.

We report the size of the compressed images in bits per pixel (BPP), which is common in evaluating image and video compression. To exclude the distortion coming from the GAN inversion technique, all the frames are projected (i.e., inverted) using the encoder. The projected frames are used for all approaches in the main paper as a baseline to compute the distortion metrics. We would like to stress out that we evaluate in the core paper the reconstruction results with the projected images of the StyleGAN, and not the original frames. This is motivated by the fact that we aim here at measuring the performance of the compression method solely. However, for sake of completeness, results comparing to the original frames are
Figure 6: Qualitative results for video compression (images are better seen zoomed). Other methods introduce blocking artifacts and blurring. Our method (SGANC IC) leads to high quality reconstruction and a significantly lower perceptually distortion, as measured by the metrics LPIPS and PIM. The zoomed versions are displayed in supplementary material.

Figure 7: Original (left) and projected (right) frame, the reconstructed frame from StyleGAN is not perfect.

briefly presented in the 4.2.3 and detailed in the supplementary material.

4.2.1 Image Compression. We compare our method with the most recent state-of-the-art codecs such as VTM [1], AV1 [3] and deep compression models such as scale Hyper Prior (HP) [6], factorized entropy model with scale and mean hyperpriors (MeanHP) [43] and the anchor variant of [14]. MeanHP and [14] are trained with the MS-SSIM objective and HP with the MSE (we found that this leads to better results) on the Celeba-HQ dataset. All the methods are evaluated on the independent evaluation dataset: MEAD-Intra and FILMPAC.

We present the rate distortion curves for different BPP on the MEAD-intra dataset in Figure 3. As can be observed, for the perceptual metrics LPIPS and PIM, our method significantly outperforms the state-of-the-art methods by a large margin. Deep learning based methods are better than traditional codecs for high BPP but they are inferior for lower BPP. This observation highlights the potential of our method to obtain reconstruction with good perceptual quality. For classical metrics, our method is the best at medium and high BPP for MS-SSIM, and has the highest PSNR at high BPP. It is noted that the deep learning based competitors trained with MS-SSIM objective, are still not able to outperform the traditional codecs VTM and AV1. The above observations also holds on the evaluation of the FILMPAC dataset, please refer to the supplementary material.

Next we discuss the qualitative results, shown in figure 4. Perceptually, at a comparable BPP, AV1 introduces blocking artifacts while VTM generates blurry results. The reconstruction of MeanHP is not sharp enough and present distortion in color. Our method delivers perfect reconstruction compared to the projected image and preserves all the details. For example, the details of the eye brows are preserved where as the existing methods failed. Additional results are presented in the supplementary material.

4.2.2 Video Compression. We compare our method (SGANC IC) with state-of-the-art, deep learning based and traditional methods: Wang et al [62], DVC[41], Versatile Video Coding Test Model (VTM) [1], and H.265 standard [56]. We used the official implementation for both VTM (Random access with GOP=16) and H.265 (FFmpeg Library). The implementation details of all the methods are detailed in the supplementary material. Each method is evaluated on the MEAD-inter dataset. Comparison with learning based methods: First we compare our work with [62] using the implementation [44] at a comparable low bit-rate, the reconstructed frame from the compressed video and average metric over the frames are shown in Figure 6. It is clear, from the quantitative metrics and visual inspection, that our method significantly outperforms other ones. We encourage the reader to see the compressed videos in the supplementary file, where they clearly show that [62] suffers from blurred reconstruction, as well as artifacts probably due to unstable landmark detection.
Next we compare our method with the learning based video compression method DVC [41] using the pre-trained weights [20] at low bit-rate ($\lambda = 256$) using the I-compressed frames by H.265 ($q = 48$) and meanHP using our trained model for intra-compression. It has higher BPP compared to other methods, but still the reconstructions are very blurry and not sharp. The perceptual quality is drastically lower than our method. Despite being simple, our method better restored the edges and texture, and reconstructed frames with high perceptual quality than the learning based ones[41, 62].

**Comparison with traditional SOTA methods:** Here, we compare our method with traditional approaches. Figure 5 presents the quantitative evaluation curves on the MEAD-inter dataset and the metrics are first averaged over the frames in each video, and then averaged across all the videos. As similar to the observations in section 4.2.1, our proposed method for video compression achieved impressive performance and significantly outperformed the state-of-the-art methods especially with respect to perceptual metrics. In terms of PSNR, our method is lower than VTM and comparable with H.265 at low bit-rate, but achieved better PSNR at high BPP with both methods. Figure 6 compares the reconstruction quality with comparable BPP. Our method is almost artifacts free, photorealistic and perceptually more pleasant, while VTM leads to blurry results and H.265 exhibits blocking artifacts. For more visual results please refer to the supp. material. Note that, despite the high quality of our method, the reconstruction of StyleGAN (projected vs original) is still not perfect (Fig. 7). Once this factor is eliminated, the distortion coming from the quantization (projected vs ours) is negligible, which makes our method very promising taking into account of the exponential improvement of GAN generation and inversion.

**4.2.3 Comparison with original frames.** For the sake of completeness, we provide the visual comparison with original frames in Fig. 8. Despite the high quality and photo-realistic outputs, our method loses the fidelity in the reconstruction. We stress that this behavior is expected as off-the-shelf StyleGAN inversion is not perfect. Further we computed the FID score between the original frames and reconstructed frames with different methods. Our FID score (109.32) is similar to the learning based methods (DVC 116.09, Wang et.al 110.49). It is also noted that the loss of fidelity is the feature of the learning based compression using GAN loss functions[2] to generate photo-realistic outputs.

**4.2.4 Computational complexity.** Our method is computationally efficient and it only brings the minor computational burden to the entire workflow with off-the-shelf StyleGAN encoder and decoder. Our evaluation shows that the average computation time to map a point from $W$ to $\tilde{W}$ is 4.7ms. Regarding the complexity of the styleGAN encoder and generator, there are recent works that propose solutions for real-time StyleGAN encoding and decoding, even on mobile platforms[8, 9] and for real-time image editing using StyleGAN[35]. The training time of our method is also not longer, since we only retain the parametization (invertible property) of the flow transformation, thus no likelihood maximization and plus the dimension of the latent space is small compared to image space. Our training time is approximately 6 hours on a single V100 GPU.

5 CONCLUSION

We have proposed in this paper a new paradigm for facial image and video compression based on GANs. Our framework is efficient to train and leads to perceptually competitive results compared to the most efficient state-of-the-art compression systems. We believe that at low bitrates, our solution leads to a different and more acceptable type of distortion, since the reconstructed image is very sharp and photorealistic. Our approach is not restricted for faces since GANs have been proposed for various natural objects and have been recently extended to ImageNet scenes [53]. For a specific category of object, for which a GAN has not been trained already, our approach could be used after training the specific GAN. The main limitation of our method is the approximation of any input image using StyleGAN. Firstly, We believe that the continuous and impressive improvements of GANs inversion and generation will limit this limitation in the near future. For instance, the third generation of StyleGAN generator was very recently released [32]. Secondly, we believe that our method can have an impact in application scenario where perceptual distortion is more important than exact face fidelity, such as extremely low bandwidth videoconferencing, or interaction in digital worlds like the metaverse.
