

LEARNING PERCEPTUAL COMPRESSION OF FACIAL VIDEO

Anonymous authors

Paper under double-blind review

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 each video frame (intra compression), as well as the successive differences between frames (inter compression). Each frame is inverted in the latent space of StyleGAN, where 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 (LPIPS, VGG16). Finally, an entropy model for video inter coding with residual is also learned in the previously constructed latent space. 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.

1 INTRODUCTION

With the explosion of videoconferencing, the efficient transmission of facial video is a key industrial problem. 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 (Rissanen & Langdon, 1979; Gray, 2011) 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 (Shannon, 1948). 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, especially for 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. However, we acknowledge that the method relies on a strong hypothesis that a GAN inversion technique can approximate accurately any given image. While this hypothesis is currently strong, we note that the performances of GAN inversion and generation methods have significantly improved recently. We take the leap of faith that the hypothesis will be valid in the near future.

For real natural images, a StyleGAN encoder (GANs inversion) (Richardson et al., 2021; Wei et al., 2021) have been proposed to project any image onto the latent space of StyleGAN. Our objective is to compress the latent representations. Although, retraining the encoder and the generator for different BPP is computationally heavy and costly, especially, using a heavy perceptual distortion losses in the image space (*i.e.*, VGG16, LPIPS). 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 a space are two-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 space dedicated for

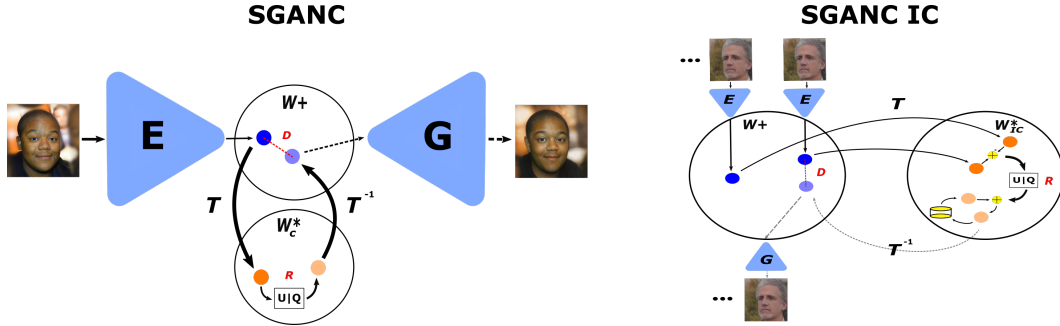


Figure 1: An illustration of our proposed approach. E and G are the pretrained and fixed StyleGAN2 encoder (Richardson et al., 2021) and generator (Karras et al., 2020) respectively. Only the mapping T and the entropy model are trained using the joint rate (R) and distortion (D) loss. **On the left** we show SGANC for image compression; On the encoding side, the images are projected in the latent space (\mathcal{W}^+) then mapped to \mathcal{W}_c^* where the quantization/compression is done. On the decoding side, the obtained bit-stream is decompressed, then mapped to \mathcal{W}^+ before generating the reconstructed image. **On the right** we show SGANC for video compression using inter coding (SGANC IC); we quantize the consecutive latent code difference and reconstruct the current latent code from this difference and the last reconstructed latent code.

compression in which an optimal entropy model can be learned. Finally, we extend this approach to videos and propose a method to optimize for the inter-coding with residuals in the new latent representation, as illustrated in Figure 2.

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 StyleGAN encoder/decoder models. We propose to learn a model for the intra image coding and the 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 H.265 and VTM 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 (Wallace, 1992), JPEG2000 (Rabani, 2002) are carefully human-engineered to achieve very effective performance. However, they are not data-dependent and limited to linear transformations. Recently, deep learning based codecs (Ballé et al., 2017; 2016; Agustsson et al., 2019; Mentzer et al., 2020) have gained significant attraction in the community due to their superior performance and ability to learn the data dependent non-linear transformations. These approaches learn nonlinear transformations of the input data and an entropy model jointly with the objective of optimal trade-off between efficient compression and reconstruction quality. Specifically, they minimize the following rate-distortion loss:

$$L = -\mathbb{E}[\log_2 \mathbf{P}_z] + \lambda \mathbb{E}[d(\mathbf{x}, \hat{\mathbf{x}})], \quad (1)$$

where x, \hat{x} are the original and the reconstructed images, z is the corresponding latent representation, $P_z(\cdot)$ is the entropy model of the latent distribution and $d(\cdot, \cdot)$ is a distortion loss.

To achieve optimal rate-distortion trade-off, several methods use variational autoencoder (VAE) type architecture (Ballé et al., 2016; Minnen et al., 2018; Cheng et al., 2020) and achieve impressive performance at higher BPP, however they are sub-optimal at low BPP. Some papers have targeted the entropy model: Ballé et al. (2018) propose to use a hyper prior as a source of side information to capture the spatial dependencies. Minnen et al. (2018) follow a similar approach and augment the hierarchical model with an autoregressive one. Cheng et al. (2020) build the previous approach with a discretized Gaussian mixture entropy model with attention modules.

Other papers have improved the model structure and the architecture (Li et al., 2018; Cheng et al., 2018; 2019) and some used recurrent neural networks (Toderici et al., 2017; Johnston et al., 2018). Others, have leveraged the power of generative adversarial networks (GAN) (Goodfellow et al., 2014) and used adversarial losses, especially for low bitrates (Rippel & Bourdev, 2017; Agustsson et al., 2019). These approaches produce high quality and lower perceptually-distorted image reconstructions (Santurkar et al., 2018; Mentzer et al., 2020), even for high bitrates (Tschannen et al., 2018). 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 (Wang et al., 2003) are used. These metrics capture the pixel wise distortion and poorly capture the perceptual distortion. Moreover, Blau & Michaeli (2018) shows that there is a trade-off between pixel wise distortion and perceptual quality. This observation is seen clearly for low BPP, where traditional codecs favor blocking artifacts and deep compression systems show blurred and other types of artifacts.

Several works have been tried to remedy these limitations; motivated by the success of using perceptual losses such as LPIPS (Zhang et al., 2018) or VGG16 (Johnson et al., 2016) in other applications (Ledig et al., 2017; Dosovitskiy & Brox, 2016; Gatys et al., 2016), some papers (Santurkar et al., 2018; Chen et al., 2020) propose to include a perceptual distortion loss in addition to the pixel wise ones. These perceptual distortion 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, and on different datasets, such as Image classification on ImageNet. To remedy this, we propose an alternative and more efficient perceptual distortion loss in the latent space.

Video Compression: Deep Video Compression systems also minimize the rate-distortion loss equation 1. In addition to the spatial redundancy (SR), temporal redundancy (TR) is reduced by incorporating motion estimation modules. Traditional methods (Wiegand et al., 2003; Sullivan et al., 2012b) rely on handcrafted and block based modules. Recently, deep learning based video compression systems proposed to replace the traditional modules by learned ones (Lu et al., 2019). Motion estimation (Dosovitskiy et al., 2015) and compensation are often used to address TR. Several improvements have been made to reduce TR; Lin et al. (2020) leverage multiple frames to improve the motion compensation, and Hu et al. (2020) use multi resolution flow maps to effectively compress locally and globally. However, training motion estimation modules is difficult since they require large annotated data.

Some approaches perform TR reduction in latent space (Feng et al., 2020; Djelouah et al., 2019; Hu et al., 2021). Others propose to do frames interpolation (Djelouah et al., 2019), or an interpolation in the latent space of GANs (Santurkar et al., 2018).

3 LEARNING PROXY REPRESENTATION FOR COMPRESSION

In this section, we describe our proposed method to learn the latent space dedicated for compression. The schematic diagram is illustrated in Figures 1 and 2. The method can be summarized as follows: an input facial video frame is first projected in the latent space of StyleGAN \mathcal{W}^+ . Our method consists in learning a transformation T from the original latent space \mathcal{W}^+ to a new proxy representation denoted as \mathcal{W}_c^* . The transformation T is defined as a diffeomorphic mapping (*i.e.*, normalizing flow) and is learnt so that the intra image compression and the inter video compression

are optimal in \mathcal{W}_c^* . It is noted that learning a proxy representation in \mathcal{W}_c^* for compression is also motivated by the fact that the rate distortion loss can not be optimized in the original space \mathcal{W}^+ without training the encoder/generator. Below, we first briefly describe some background material (section 3.1), and then present our methods for images or intra compression (section 3.2) and for videos or inter compression (section 3.3).

3.1 BACKGROUND

StyleGAN Generator StyleGAN (Karras et al., 2019) is the state of the art unconditional GAN for high quality realistic image generation. It consists of a mapping function that takes a vector sampled from a normal distribution ($z \sim \mathcal{N}(0, \mathbf{I})$) and maps it to an intermediate latent space \mathcal{W} using a fully connected network (M) before feeding it to multiple stages (*i.e.*, \mathcal{W}^+ space) of the generator (G) to generate an image (x) from the distribution of the real images ($x \sim p_x$):

$$\mathbf{x} = G(\mathbf{w}) \quad \mathbf{w} = M(\mathbf{z}) \quad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}) \quad (2)$$

It is shown that the latent space of GAN is semantically rich and enjoys several properties such as semantic interpolation (Radford et al., 2016). In addition, the latent vector encoded in \mathcal{W}^+ space of StyleGAN captures a hierarchical representation of the projected image. In our case, we use StyleGAN2 (Karras et al., 2020) which is an improved version.

StyleGAN Encoder The StyleGAN encoder (Richardson et al., 2021) is a deterministic function denoted as E . Its role is to project a real image into the latent space of StyleGAN (*e.g.*, \mathcal{W} or \mathcal{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 \mathcal{W}^+ with dimension of (18×512) 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 $\mathbf{w} \in \mathcal{W}^+$ and generates a high resolution image of size 1024×1024 , and an encoder E (pretrained and fixed) that embeds any given image \mathbf{x} to the latent code \mathbf{w} in \mathcal{W}^+ such that $G(E(\mathbf{x})) \simeq \mathbf{x}$. Our objective is to learn a new latent space \mathcal{W}_c^* optimal for image compression. The \mathcal{W}_c^* is obtained using a bijective transformation $T : \mathcal{W}^+ \rightarrow \mathcal{W}_c^*$, which is parametrized as a normalizing flows model (note that our work only requires the bijectivity, as such, no maximum-likelihood term is included in the training objective). T maps a latent code $\mathbf{w} \in \mathcal{W}^+$ into $\mathbf{w}_c^* \in \mathcal{W}_c^*$ such that the latent vectors \mathbf{w}_c^* has the minimum entropy and the sufficient information to generate the original image with minimal distortion from the inverted latent code $T^{-1}(\mathbf{w}_c^*)$ using G . The entropy model is trained on the latent codes $\mathbf{w}_c^* \in \mathcal{W}_c^*$ by minimizing the following rate loss after the quantization:

$$\mathcal{R} = -\mathbb{E} \left[\sum_i^D \log_2 p_i(Q(T(\mathbf{w}))) \right] = -\mathbb{E} \left[\sum_i^D \log_2 p_i(T(\mathbf{w}) + \epsilon) \right] \quad (3)$$

For equation 3 to be differentiable, following (Ballé et al., 2017), we relax the hard quantization Q by adding uniform noise ϵ to the latent vectors $T(\mathbf{w})$. p_i is the i^{th} dimension of the probability density function in \mathcal{W}_c^* , D is the latent vector dimension, $\mathbf{w} = E(\mathbf{x})$ where \mathbf{x} is the input image and ϵ is sampled from a uniform distribution $\mathcal{U}_{[-0.5, 0.5]}$. The entropy model \mathbf{p} is modeled as the fully factorized entropy model and it is also parametrized by a neural network as in (Ballé et al., 2018). The transformed/quantized latent codes should also have sufficient information to reconstruct the image, and this is achieved by minimizing a distortion loss. In general the distortion loss is computed in the image space, although, we propose a new efficient distortion loss directly in the latent space:

$$\mathcal{D} = d(\mathbf{w}, T^{-1}(T(\mathbf{w}) + \epsilon)), \quad (4)$$

where d is any distortion loss, in this paper we use the mean squared error (MSE) loss. The equation 4 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

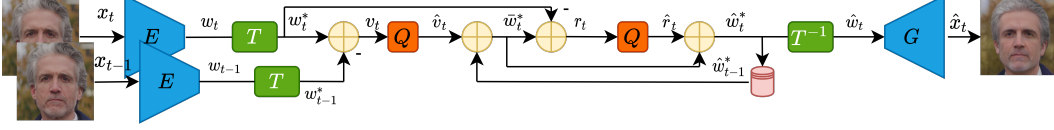


Figure 2: An illustration of our proposed approach for Video compression using Inter Coding with residual during test (Section 3.3). Q corresponds to quantization, compression encoding and decoding. Here $g = 1$.

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 \mathcal{W}_c^* is a trade-off between the rate (equation 3) and distortion (equation 4) as shown below

$$\mathcal{L} = \mathcal{R} + \lambda \mathcal{D} \quad (5)$$

Where λ 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 \mathcal{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. We argue that this leads to efficient compression due to the good properties of the latent space of GANs (*e.g.*, linear interpolation).

Specifically, inter coding with the residuals is performed in the latent space (Figure 2). Given a sequence of frames $\{x_1, x_2, \dots, x_{t-1}, x_t, \dots\}$, 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 \mathcal{W}_{IC}^* , leading to a sequence of latent codes $\{w_1^*, w_2^*, \dots, w_{t-1}^*, w_t^*, \dots\}$. The mapping T is learnt such that the sequence of latent codes in \mathcal{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 appendix A.2):

- The first latent code is coded using the method described in section 3.2: \hat{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 to obtain $\hat{v}_t = Q(w_t^* - w_{t-1}^*)$.
- From the previous reconstructed code, an estimate of the latent code at frame t is obtained as $\bar{w}_t = \hat{w}_{t-1}^* + \hat{v}_t$.
- The residual between the estimated and the actual latent code is computed and quantized as $\hat{r}_t = Q(w_t^* - \bar{w}_t^*)$ (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): $\hat{w}_t^* = \bar{w}_t^* + \hat{r}_t$.
- The latent codes in \mathcal{W}_{IC}^* are remapped to \mathcal{W}^+ to generate the images using the pretrained StyleGAN2 G .

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

$$\mathcal{L}_{IC} = d(w_t, T^{-1}(\hat{w}_t^*)) - \lambda \left(\mathbb{E} \left[\sum_i^D \log_2 p_i(\hat{v}_t) \right] + \mathbb{E} \left[\sum_i^D \log_2 q_i(\hat{r}_t) \right] \right) \quad (6)$$

As similar to section 3.2, we replace the quantization Q by adding uniform noise, and the entropy model is modeled as in (Ballé et al., 2018). While equation 6 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 Irwin-Hall distribution (the proof of the following Lemma 1 can be found in the appendix A.1).

Lemma 1 *Let $Q(x) = x + \epsilon$ be the continuous relaxation of the quantization, where ϵ follows the uniform distribution $\mathcal{U}_{[-0.5, 0.5]}$. Let $\mathbf{w}_0^* \in \mathcal{R}^n, \hat{\mathbf{w}}_0^* = Q(\mathbf{w}_0^*)$, and $\bar{\mathbf{w}}_t^* = \hat{\mathbf{w}}_{t-1}^* + Q(\mathbf{w}_t^* - \mathbf{w}_{t-1}^*)$, $t = 1, \dots, n$. If $\hat{r}_t = Q(\mathbf{w}_t^* - \bar{\mathbf{w}}_t^*)$ is the quantized residual defined for every t such that $t \equiv 0 \pmod{g}$, then \hat{r}_t follows the Irwin-Hall distribution with the parameter $3 + (g - 1)$ and the support being shifted by $-0.5 * (3 + (g - 1))$.*

This leads to optimizing our latent space \mathcal{W}_{IC}^* for the rate-distortion loss only with one entropy model for \hat{v}_t by discarding the last term in (equation 6). During test, the explicit distribution of the residuals can be used for entropy coding.

Stage-Specific entropy models: In Karras et al. (2019), 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 λ differently for each layer of the generator (Note that, the latent code in \mathcal{W}^+ or \mathcal{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 λ is chosen to be higher for the first layers and decreases subsequently.

4 EXPERIMENTS

4.1 DATASETS

Celeba-HQ (Karras et al., 2018), a dataset consists of 30000 high quality face images of 1024×1024 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 website¹ by searching their names (FP006734MD02, FP006940MD02, FP009971HD03, FP010363HD03, FP010263HD03).

MEAD (Wang et al., 2020), 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 methods.

Dataset preprocessing: All the frames are cropped and aligned using the same preprocessing method as that of FFHQ dataset (Karras et al., 2019), 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 (1024×1024).

4.2 IMPLEMENTATION DETAILS

We used StyleGAN2 generator (G) (Karras et al., 2020) pretrained on FFHQ dataset (Karras et al., 2019). The images are encoded in \mathcal{W}^+ using a pretrained StyleGAN2 encoder (E) (Richardson et al., 2021). The parameters of the generator and the encoder remain fixed in all the experiments. The latent vector dimension in \mathcal{W}^+ and \mathcal{W}_c^* is 18×512 . The mapping function T is modeled using RealNVP architecture (Dinh et al., 2017) without batch normalization. It consists of 13 coupling layers, and each coupling layer consists of 3 fully connected (FC) layers for the translation function and

¹<https://filmpac.com/>

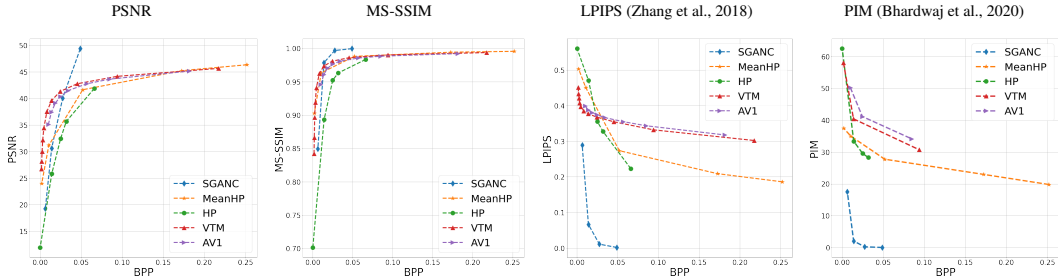


Figure 3: Rate distortion curves on MEAD intra dataset: for medium and large BPP, our method (SGANC in blue) is better in terms of LPIPS and MS-SSIM than VTM (Red), AV1 (Purple), MeanHP (Orange) and HP (Green). For high BPP, our method is better in terms of PSNR. The perceptual metrics LPIPS and PIM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

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 (Ballé et al., 2018) based on the implementation from the CompressAI library (Bégaint et al., 2020). For the stage specific entropy model, λ 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 Range Asymmetric Numeral System coder (Duda, 2013) 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 (Wang et al., 2020), where each batch contains video slices of size = 9 frames. All the frames are preprocessed as in section 4.1.

4.3 RESULTS

In this section we present the experimental results of our proposed method for image and video compression. 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) (Wang et al., 2003), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), and the Perceptual Information Metric (PIM) with a mixture of 5 Gaussians (Bhardwaj et al., 2020).

LPIPS and PIM are two perceptual metrics; PIM is recently proposed and follow an information-theoretic approach, inspired from the human perceptual system (Bhardwaj et al., 2020). The authors of LPIPS, on the other hand, show that it is consistent with the human perception. 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. For the sake of completion, we also present evaluation results with the original images in the appendix A.4.

4.3.1 IMAGE COMPRESSION

We compare our method with the most recent state-of-the-art codecs such as VTM (VTM, 2020), AV1 (AV1, 2018) and deep compression models such as scale Hyper Prior (HP) (Ballé et al., 2018), factorized entropy model with scale and mean hyperpriors (MeanHP) (Minnen et al., 2018) and the anchor variant of Cheng et al. (2020). MeanHP and (Cheng et al., 2020) 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.

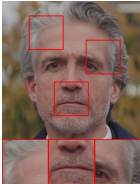
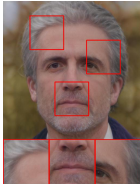
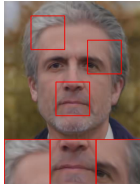
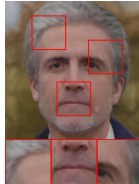
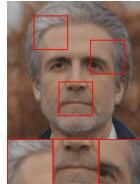
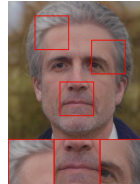
Model	Original	Projected	VTM	AV1	MeanHP	SGANC
						
BPP			0.0155	0.0156	0.0119	0.0146
PSNR			37.443	32.128	35.946	29.31
MS-SSIM			0.951	0.934	0.953	0.970
LPIPS			0.296	0.332	0.361	0.062
PIM			21.075	24.168	39.671	1.767

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. In particular, the perceptual metrics LPIPS and PIM show a significant impact of our method.

Quantitative results: 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 figure 17 in the appendix.

Qualitative results: Next we discuss the qualitative results, shown in figure 4. The visual inspection reveals that 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 images are displayed in appendix A.4.

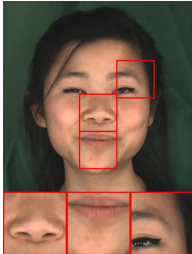
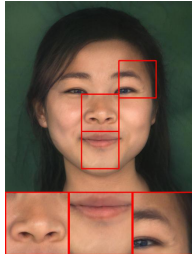
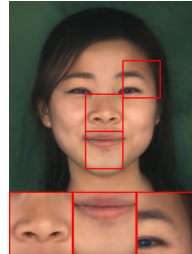
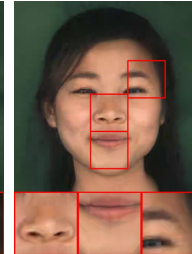
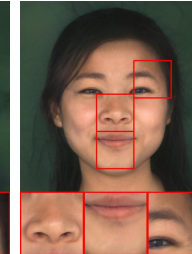
Model	Original	Projected	VTM	H.265	SGANC IC
					
BPP			0.0018	0.0021	0.0017
PSNR			38.94	34.87	32.23
MSSSIM			0.978	0.953	0.985
LPIPS			0.2800	0.3267	0.0465
PIM			24.97	29.76	1.221

Figure 5: 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.

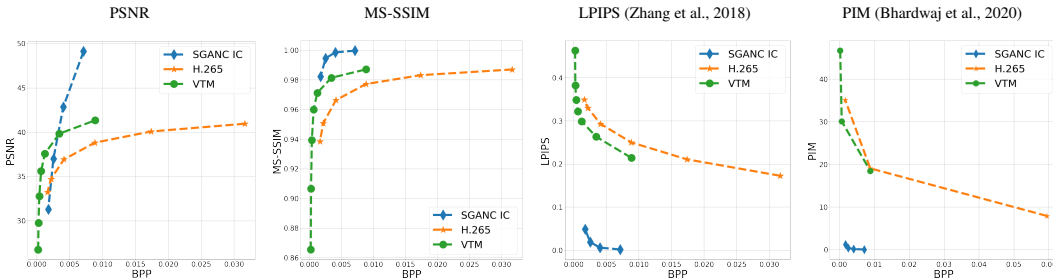


Figure 6: Rate distortion curves on MEAD inter dataset. Our approach SGANC-IC leads to a significantly lower perceptual distortion as measured by the perceptual metrics LPIPS and PIM.

4.3.2 VIDEO COMPRESSION

We compare our method (SGANC IC) with the most recent and performing two state-of-the-art methods: Versatile Video Coding Test Model (VTM) (VTM, 2020) and H.265 standard (Sullivan et al., 2012a). We used the official implementation for both VTM (Random access with GOP=16) and H.265 (FFmpeg Library). Each method is evaluated on the **MEAD-inter** dataset.

Quantitative results: Figure 6 presents the quantitative evaluation curves on the **MEAD-inter** dataset. The computed metrics are first averaged across the frames in each video, and then averaged across all the videos. As similar to the observations in section 4.3.1, our proposed method for video compression achieved impressive performance and significantly outperformed the state-of-the-art methods with respect to perceptual metrics. Especially, with the perceptual distortion metric PIM, our method is significantly better. In terms of PSNR, our method is comparable with H.265 at low bit-rate and achieved better performance at high BPP with both methods.

Qualitative results: Figure 5 compares the reconstruction quality of all the methods with a comparable BPP on a single frame extracted from the compressed video. 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 appendix A.4. Note that, despite the high quality of our method, the reconstruction of StyleGAN (projected vs original) is still not perfect. Once this factor is eliminated, the distortion coming from the quantization (projected vs our method) is negligible which makes our method very promising taking into account of the exponential improvement of GAN generation and inversion.

The ablation studies related to the different distortion losses, the alternative choices of transformation T and other design choices are discussed in the appendix A.3 due to the space limitation.

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. 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 as well as the compression bottleneck. The main limitation of our method is the approximation of any input image using StyleGAN. We believe that the continuous and impressive improvements of GANs inversion and generation will limit this limitation in the near future.

REFERENCES

Versatile Video Coding. <https://jvet.hhi.fraunhofer.de/>, 2020.

Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *2019 IEEE/CVF Inter-*

- national Conference on Computer Vision (ICCV)*, pp. 221–231, 2019. doi: 10.1109/ICCV.2019.00031.
- AOM AV1. Versatile Video Coding. <https://aomedia.googlesource.com/aom/>, 2018.
- Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimization of nonlinear transform codes for perceptual quality. In *2016 Picture Coding Symposium (PCS)*, pp. 1–5. IEEE, 2016.
- Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.
- Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.
- Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.
- Sangnie Bhardwaj, Ian Fischer, Johannes Ballé, and Troy Chinen. An unsupervised information-theoretic perceptual quality metric. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13–24. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/00482b9bed15a272730fcb590ffebddd-Paper.pdf>.
- Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018. doi: 10.1109/CVPR.2018.00652.
- Li-Heng Chen, Christos G Bampis, Zhi Li, Andrey Norkin, and Alan C Bovik. Perceptually optimizing deep image compression. *arXiv preprint arXiv:2007.02711*, 2020.
- Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Deep convolutional autoencoder-based lossy image compression. In *2018 Picture Coding Symposium (PCS)*, pp. 253–257. IEEE, 2018.
- Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Deep residual learning for image compression. In *CVPR Workshops*, pp. 0, 2019.
- Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7939–7948, 2020.
- Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *ICLR 2017: International Conference on Learning Representations*, 2017.
- Abdelaziz Djelouah, Joaquim Campos, Simone Schaub-Meyer, and Christopher Schroers. Neural inter-frame compression for video coding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6420–6428, 2019. doi: 10.1109/ICCV.2019.00652.
- Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29:658–666, 2016.
- Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2758–2766, 2015.
- Jarek Duda. Asymmetric numeral systems: entropy coding combining speed of huffman coding with compression rate of arithmetic coding. *arXiv preprint arXiv:1311.2540*, 2013.
- Runsen Feng, Yaojun Wu, Zongyu Guo, Zhizheng Zhang, and Zhibo Chen. Learned video compression with feature-level residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 120–121, 2020.

- Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016. doi: 10.1109/CVPR.2016.265.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- Robert M. Gray. *Entropy and Information Theory*. Springer Publishing Company, Incorporated, 2nd edition, 2011. ISBN 9781441979698.
- Zhihao Hu, Zhenghao Chen, Dong Xu, Guo Lu, Wanli Ouyang, and Shuhang Gu. Improving deep video compression by resolution-adaptive flow coding. In *European Conference on Computer Vision*, pp. 193–209. Springer, 2020.
- Zhihao Hu, Guo Lu, and Dong Xu. Fvc: A new framework towards deep video compression in feature space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1502–1511, 2021.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.
- N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, Volume 2*. Wiley Series in Probability and Statistics. Wiley, 1995. ISBN 9780471584940. URL <https://books.google.fr/books?id=BTANEAAAQBAJ>.
- Nick Johnston, Damien Vincent, David Minnen, Michele Covell, Saurabh Singh, Troy Chinen, Sung Jin Hwang, Joel Shor, and George Toderici. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4385–4393, 2018.
- Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.
- Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.
- Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.
- Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017. doi: 10.1109/CVPR.2017.19.
- Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang. Learning convolutional networks for content-weighted image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3214–3223, 2018.
- Jianping Lin, Dong Liu, Houqiang Li, and Feng Wu. M-lvc: multiple frames prediction for learned video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3546–3554, 2020.
- Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11006–11015, 2019.
- Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33, 2020.

- David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *NeurIPS*, 2018.
- Majid Rabbani. JPEG2000: Image Compression Fundamentals, Standards and Practice. *Journal of Electronic Imaging*, 11(2), 2002. doi: 10.1117/1.1469618. URL <https://doi.org/10.1117/1.1469618>.
- Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR 2016 : International Conference on Learning Representations 2016*, 2016.
- Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2287–2296, 2021.
- Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *International Conference on Machine Learning*, pp. 2922–2930. PMLR, 2017.
- J. Rissanen and G. G. Langdon. Arithmetic coding. *IBM Journal of Research and Development*, 23(2):149–162, 1979. doi: 10.1147/rd.232.0149.
- Shibani Santurkar, David Budden, and Nir Shavit. Generative compression. In *2018 Picture Coding Symposium (PCS)*, pp. 258–262, 2018. doi: 10.1109/PCS.2018.8456298.
- C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.
- Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012a. doi: 10.1109/TCSVT.2012.2221191.
- Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012b. doi: 10.1109/TCSVT.2012.2221191.
- George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell. Full resolution image compression with recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5306–5314, 2017.
- Michael Tschannen, Eirikur Agustsson, and Mario Lucic. Deep generative models for distribution-preserving lossy compression. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 5933–5944, 2018.
- G. K. Wallace. The jpeg still picture compression standard. *IEEE Trans. on Consum. Electron.*, 38(1):xviii–xxxiv, February 1992. ISSN 0098-3063. doi: 10.1109/30.125072. URL <https://doi.org/10.1109/30.125072>.
- Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pp. 700–717. Springer, 2020.
- Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pp. 1398–1402 Vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216.
- Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Lu Yuan, Gang Hua, and Nenghai Yu. A simple baseline for stylegan inversion. *arXiv preprint arXiv:2104.07661*, 2021.
- T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the h.264/avc video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, 2003. doi: 10.1109/TCSVT.2003.815165.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

A APPENDIX

The Appendix is organized as follows: in section A.1, we prove that the distribution of residuals follows a known Irwing-Hall distribution, in section A.2 we detail the algorithms for training and testing of the inter coding with residual approach. In section A.3 we detail the ablation study for image and video compression. Finally, in section A.4 we provide more results.

A.1 EXPLICIT DISTRIBUTION OF THE RESIDUALS: PROOF OF LEMMA 1

Let $Q(x) = x + \epsilon$ be the continuous relaxation of the quantization, and ϵ follows the uniform distribution $\mathcal{U}_{[-0.5, 0.5]}$. Let the t be the frame index ranging from $\{0, 1, \dots, K-1\}$.

Let $\mathbf{w}_0^* \in \mathcal{R}^n$ and $\hat{\mathbf{w}}_0^* = Q(\mathbf{w}_0^*)$. Let us define the following

$$\begin{aligned}\hat{\mathbf{v}}_t &= Q(\mathbf{w}_t^* - \mathbf{w}_{t-1}^*) = \mathbf{w}_t^* - \mathbf{w}_{t-1}^* + \epsilon, \\ \bar{\mathbf{w}}_t &= \hat{\mathbf{w}}_{t-1}^* + \hat{\mathbf{v}}_t,\end{aligned}\tag{7}$$

Similarly, we define

$$\hat{\mathbf{r}}_t = \begin{cases} Q(\mathbf{w}_t^* - \bar{\mathbf{w}}_t^*), & \forall t \equiv 0 \pmod{g} \\ 0 & \text{else} \end{cases}\tag{8}$$

Now we prove that $\hat{\mathbf{r}}_t$ follows the Irwin-Hall or the uniform sum distribution (Johnson et al., 1995). we have:

$$\begin{aligned}\hat{\mathbf{r}}_t &= Q(\mathbf{w}_t^* - \bar{\mathbf{w}}_t^*) = \mathbf{w}_t^* - \bar{\mathbf{w}}_t^* + \epsilon_1 \\ &= \mathbf{w}_t^* - (\hat{\mathbf{w}}_{t-1}^* + \hat{\mathbf{v}}_t) + \epsilon_1 \\ &= \mathbf{w}_t^* - (\hat{\mathbf{w}}_{t-1}^* + \mathbf{w}_t^* - \mathbf{w}_{t-1}^* + \epsilon_1) + \epsilon_2 \\ &= \mathbf{w}_t^* - \hat{\mathbf{w}}_{t-1}^* - \mathbf{w}_t^* + \mathbf{w}_{t-1}^* - \epsilon_2 + \epsilon_1 \\ &= \mathbf{w}_{t-1}^* - \hat{\mathbf{w}}_{t-1}^* - \epsilon_2 + \epsilon_1\end{aligned}\tag{9}$$

For $t \equiv 0 \pmod{g}$, the reconstructed latent code is computed from the residual, and can be written as:

$$\begin{aligned}\hat{\mathbf{w}}_t^* &= \hat{\mathbf{r}}_t + \bar{\mathbf{w}}_t^* \\ &= Q(\mathbf{w}_t^* - \bar{\mathbf{w}}_t^*) + \bar{\mathbf{w}}_t^* \\ &= \mathbf{w}_t^* - \bar{\mathbf{w}}_t^* + \epsilon + \bar{\mathbf{w}}_t^* \\ &= \mathbf{w}_t^* + \epsilon\end{aligned}\tag{10}$$

Otherwise, it is the estimated latent code $\bar{\mathbf{w}}_t^*$, thus $\hat{\mathbf{w}}_t^*$ can be written as:

$$\hat{\mathbf{w}}_t^* = \begin{cases} \mathbf{w}_t^* + \epsilon, & \forall t \equiv 0 \pmod{g} \\ \bar{\mathbf{w}}_t^* & \text{else} \end{cases}\tag{11}$$

For $g = 1$, equation 9 becomes:

$$\begin{aligned}\hat{\mathbf{r}}_t &= \mathbf{w}_{t-1}^* - \hat{\mathbf{w}}_{t-1}^* - \epsilon_2 + \epsilon_1 \\ &= \mathbf{w}_{t-1}^* - \mathbf{w}_{t-1}^* + \epsilon_3 - \epsilon_2 + \epsilon_1 \\ &= \epsilon_3 - \epsilon_2 + \epsilon_1\end{aligned}\tag{12}$$

For $g > 1$, Let us define the following quantity:

$$\hat{\mathbf{m}}_t = \begin{cases} \hat{\mathbf{r}}_t, & \forall t \equiv 0 \pmod{g} \\ Q(\mathbf{w}_t^* - \bar{\mathbf{w}}_t^*) & \text{else} \end{cases}\tag{13}$$

Using equation 11 and equation 13:

$$\begin{aligned}\hat{\mathbf{r}}_t &= \mathbf{w}_{t-1}^* - \hat{\mathbf{w}}_{t-1}^* - \epsilon_2 + \epsilon_1 \\ &= \mathbf{w}_{t-1}^* - \bar{\mathbf{w}}_{t-1}^* - \epsilon_2 + \epsilon_1 \\ &= \hat{\mathbf{m}}_{t-1} - \epsilon_2 \\ &= \dots \\ &= \hat{\mathbf{m}}_{t-(g-1)} - \sum_{i=1}^{g-1} \epsilon_i\end{aligned}\tag{14}$$

Similarly to equation 9, we can replace $\hat{\mathbf{m}}_{t-(g-1)}$:

$$\hat{\mathbf{r}}_t = \mathbf{w}_{t-g}^* - \hat{\mathbf{w}}_{t-g}^* + \epsilon_1 - \sum_{i=1}^{g-1} \epsilon_i \quad (15)$$

As $t-g \equiv 0 \pmod{g}$ (as the residual is used only each $t \equiv 0 \pmod{g}$), from equation 11, we can replace $\hat{\mathbf{w}}_{t-g}^*$ by its relaxed approximation:

$$\begin{aligned} \hat{\mathbf{r}}_t &= \mathbf{w}_{t-g}^* - (\mathbf{w}_{t-g}^* + \epsilon_2) + \epsilon_1 - \sum_{i=1}^{g-1} \epsilon_i \\ &= \epsilon_1 - \epsilon_2 + \epsilon_3 - \sum_{i=1}^{g-1} \epsilon_i \end{aligned} \quad (16)$$

Which is the sum of $3 + (g-1)$ independent random variables following the uniform distribution.

We showed $\hat{\mathbf{r}}_t$ follows the uniform sum distribution, which is the Irwin-Hall distribution $IH(x; n)$ with parameter $n = 3 + (g-1)$ and $x \in [-n \times 0.5, n \times 0.5]$. Since the ϵ is the uniform distribution $\mathcal{U}_{[-0.5, 0.5]}$, in our case the support of the Irwin-Hall distribution is shifted by $-n \times 0.5$.

A.2 INTER CODING (SGANC IC)

In this section we present the algorithms for video compression using inter coding with residual during train and test:

Algorithm .1: SGANC IC: during test

Result: Compressed frames sequence: $\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_{t-1}, \hat{\mathbf{x}}_t, \dots, \hat{\mathbf{x}}_N\}$

- 1 Initialization: frames sequence: $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{t-1}, \mathbf{x}_t, \dots, \mathbf{x}_N\}$, N =number of frames, Encoder (E), Generator (G), Transformation T , entropy coder (EC) and decoder (ED), quantizer (Q), g for residual coding;
- 2 $\hat{\mathbf{w}}_0^* = ED(EC(Q(T(E(\mathbf{x}_0)))));$ // Intra Coding of the first frame
- 3 $t = 1;$
- 4 **while** $t < N$ **do**
- 5 $\mathbf{w}_t^* = T(E(\mathbf{x}_t));$ // Encode the frames and map them to \mathcal{W}_c^*
- 6 $\mathbf{w}_{t-1}^* = T(E(\mathbf{x}_{t-1}));$
- 7 $\hat{\mathbf{v}}_t = ED(EC(Q(\mathbf{w}_t^* - \mathbf{w}_{t-1}^*)));$ //Quantize, Compress and Decompress (Receiver) the differences
- 8 $\bar{\mathbf{w}}_t^* = \hat{\mathbf{w}}_{t-1}^* + \hat{\mathbf{v}}_t;$ // Compute an estimate of the latent code
- 9 **if** $t \% g == 0$ **then**
- 10 $\hat{\mathbf{r}}_t = ED(EC(Q(\mathbf{w}_t^* - \bar{\mathbf{w}}_t^*)));$ // Quantize, Compress and Decompress (Receiver) the residual
- 11 $\hat{\mathbf{w}}_t^* = \bar{\mathbf{w}}_t^* + \hat{\mathbf{r}}_t;$ // Reconstruct the latent code
- 12 **else**
- 13 $\hat{\mathbf{w}}_t^* = \bar{\mathbf{w}}_t^*;$
- 14 **end**
- 15 $\hat{\mathbf{x}}_t = G(T^{-1}(\hat{\mathbf{w}}_t^*));$ // Reconstruct the image
- 16 $t = t + 1;$
- 17 **end**

Algorithm .2: SGANC IC: during training

Result: Transformation (T), Entropy model (p)

```

1 Initialization: video dataset encoded as latent codes, where each video=
  { $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{t-1}, \mathbf{w}_t, \dots$ }, number of frames in each batch ( $N$ ), dataset size ( $S$ ), Encoder
  (E), Generator (G);
2 while  $i < S$  do
3    $t, L = 1, 0$ ;
4   while  $t < N$  do
5      $\mathbf{w}_t^*, \mathbf{w}_{t-1}^* = T(\mathbf{w}_t), T(\mathbf{w}_{t-1})$ ; // Map the latent codes to  $\mathcal{W}_c^*$ 
6      $\hat{\mathbf{v}}_t = Q(\mathbf{w}_t^* - \mathbf{w}_{t-1}^*)$ ; // Quantize (adding noise) the differences
7      $\bar{\mathbf{w}}_t^* = \hat{\mathbf{w}}_{t-1}^* + \hat{\mathbf{v}}_t$ ; // Compute an estimate of the latent code
8      $\hat{\mathbf{w}}_t^* = \bar{\mathbf{w}}_t^*$ ;
9      $L = L + \mathcal{L}$ ; // Compute the loss  $\mathcal{L}$  (equation 17)
10     $t = t + 1$ ;
11  end
12  Update the parameters of  $T$  and  $p$  to minimize  $L$ ;
13   $i = i + 1$ ;
14 end

```

A.3 ABLATION STUDY

In this section we detail the ablation study for image and video compression.

A.3.1 IMAGE COMPRESSION (DISTORTION LOSS)

In this section we compare different choices of the distortion loss. Specifically, we compare the MSE loss in the image space (Img. D), the MSE loss in the latent space \mathcal{W}^+ (Lat. D), the combination of both (Lat.-Img. D), the LPIPS loss in the image space (LPIPS D) and the combination of the MSE and the LPIPS losses in the image space (LPIPS-Img. D). The implementation details are the same as in section 4.2 except for the choice of the distortion loss and we use a batch size of 4 when training with LPIPS. We compare on both MEAD intra and FILMPAC dataset.

Results: From Figures 7, 8, we can notice that our loss (Lat. D) outperforms the the MSE and the LPIPS distortion in the image space. Moreover, the loss in the image space produces some artifacts and images are blurred while the loss in the latent space is almost artifacts free. There is no benefit of using the loss in the image space in addition to ours. Note that, the training using only the latent loss is faster than others (e.g., the training took; 5 days (LPIPS.-Img. D), 4 days (Lat.-Img. D) and 6 hours (Lat. D)) and occupies smaller space in GPU (e.g., 24 GB with batch size=4 for LPIPS.-Img. D, 10 GB with batch size =4 for Img. D and Lat.-Img. D and less than 2 GB with batch size =8 for Lat. D).

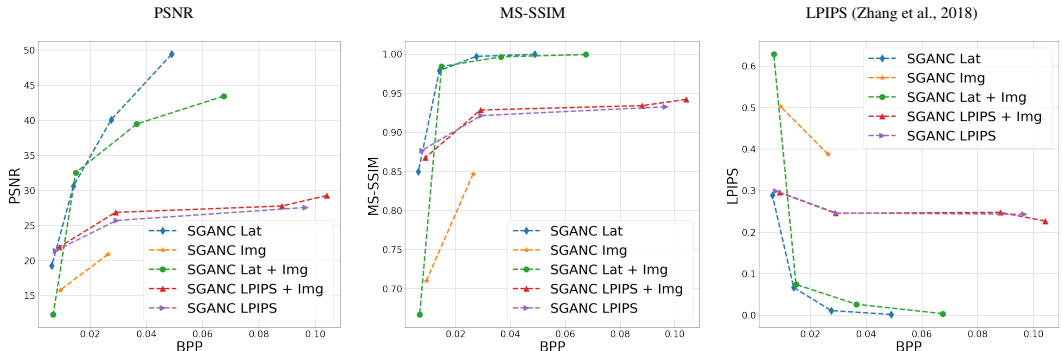


Figure 7: Ablation study for the choice of the distortion loss on MEAD intra dataset. we compare the MSE loss in the image space (SGANC Img, Orange), the MSE loss in the latent space \mathcal{W}^+ (SGANC Lat, Blue), the combination of both (SGANC Lat + Img, Green), the LPIPS loss in the image space (SGANC LPIPS, Purple) and the combination of the MSE and the LPIPS losses as done in other works (SGANC LPIPS + Img, Red). The loss in the latent space (SGANC Lat, ours) outperforms the MSE and LPIPS losses in the image space. Using the MSE loss with our loss does not seem to improve the results (SGANC Lat + Img).



Figure 8: Ablation study for the choice of the distortion loss. We compare the effect of taking the MSE loss in; the latent space (Lat. D, ours), in the image space (Img. D), both in the latent and image space (Lat.-Img. D) and in the image space with the LPIPS perceptual distortion (LPIPS.-Img. D). The loss in the image space favors some artifacts such as white patches and the images are blurred. In addition, the LPIPS loss produce blurred images ((LPIPS.-Img. D and LPIPS. D). Our loss produce artifacts free images. There is no benefit of using additionally the image space loss. The bpp is the taken as the average over the 5 images in each row

A.3.2 IMAGE COMPRESSION (TRANSFORMATION T)

In this section, we investigate the importance of using a bijective transformation (*i.e.*, Normalizing Flows). To this end, we replace the Real NVP model with an autoencoder (AE) and retrain it with the entropy model on Celeba-HQ with the same implementation details as in section 4.2. The AE consists of 7 layers with the same dimension (*i.e.*, 512) for the encoder and 7 layers with the same dimension for the decoder with ReLU activation functions.

Results: From Figure 9, we can notice that parametrizing the transformation T as Normalizing Flows (*i.e.*, Real NVP) leads to better results. For instance, some facial attributes are changed when using an AE (such as the age and the skin color), as well as the person’s identity.



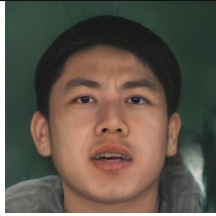
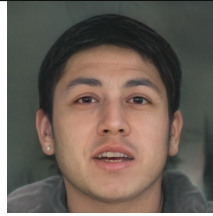
Model	Original	Projected	SGANC	AEC
				
BPP			0.014	0.017

Figure 9: Qualitative results for image compression: Using an AE (AEC) instead of NF leads to high distortion and a change in the facial attributes and the identity of the person. Our method (SGANC) leads to high quality reconstruction and perceptually lower distortion. Images are better seen zoomed.

A.3.3 VIDEO COMPRESSION (INTER CODING: EXPLICIT SPARSIFICATION)

Having only few dimensions that change between two consecutive latent codes is efficient for entropy coding. Thus, we have investigated how to explicitly sparsify these differences and add an $L1$ regularization on the latent codes differences during training. The loss becomes:

$$\mathcal{L}_{IC-L1} = \mathcal{L}_{IC} + \lambda_{L1} |\mathbf{w}_{t-1}^* - \mathbf{w}_t^*| \quad (17)$$

Note that, as the transformation T and the entropy model are trained jointly, we expect that the latent codes in \mathcal{W}_c^* to be transformed to fit efficiently the entropy model. In the following, we assess whether explicit sparsification brings additional improvement (Section A.3.4).

A.3.4 VIDEO COMPRESSION (INTER CODING)

In this section we present the ablation study for SGANC IC, in which we investigate the effect of the following parameters; parameter g for the residual coding (g), $L1$ regularization equation 17, residual coding (res) vs intra coding (intra) each g , stage specific entropy model; using different entropy models and distortion λ for the different stages of StyleGAN2 (SS). We thus compare the following variants:

- SGANC IC intra g10: replacing the residual coding at each $g = 10$ by intra coding using one of the image processing models that we train in section 3.2.
- SGANC IC res g10: doing residual coding each $g = 10$.
- SGANC IC res g80: doing residual coding each $g = 80$.
- SGANC IC res g10 L1: Adding the $L1$ regularization equation 17 during training and doing residual coding each $g = 10$.
- SGANC res g10 SS: Using 3 entropy models and 3 T models for each stage of the StyleGAN2 (1-7, 7-13, 13-18). In addition, training with layer specific distortion lambda; $\lambda = w\lambda$ (where $w = 1$ for the first stage and decrease from 1 to 0.01 for the second and third stage).
- SGANC res g10 SS L1: residual coding each $g = 10$, stage specific entropy models and $L1$ regularization.
- SGANC res g2 SS: residual coding each $g = 2$, stage specific entropy.
- SGANC res g1 SS L1: Same as before but doing a residual coding each frame ($g = 1$).

From Figure 10, we can notice the following:

- Decreasing g leads to better results, at the expense of increased BPP especially for $g = 1$.
- Significant improvements is obtained by performing residual coding (SGANC IC intra g10 vs SGANC IC res g10) and using stage specific entropy models (SGANC IC res g10 vs SGANC IC res g10 SS).
- Only a marginal improvement was obtained by adding the $L1$ regularization during training (SGANC res g10 vs SGANC res g10 L1). In addition, the improvement becomes negligible when using SS entropy models (SGANC res g10 SS vs SGANC res g10 SS L1).

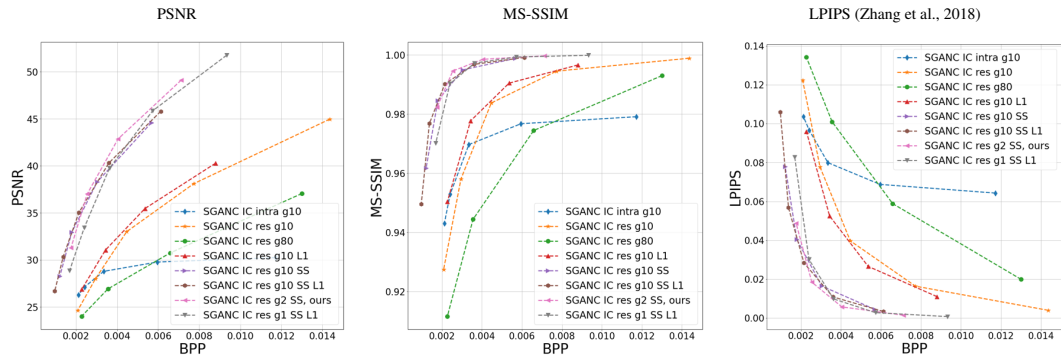


Figure 10: Ablation study for video compression using inter coding with residuals on MEAD inter dataset. res: do residual coding each g frames, intra: do intra coding each g frames, SS: stage specific entropy model and transformation T , L1: add L1 regularization during training (section A.3.3), EM: entropy model

A.4 MORE RESULTS

In this section we show more quantitative and qualitative results for image and video compression.

A.4.1 IMAGE COMPRESSION

In this section we show the results for image compression (Figures 11, 12, 15, 16, 13, 14). Contrary to the results presented in the main paper, we here present for sake of completeness the results by comparing to the original image (except for our method).

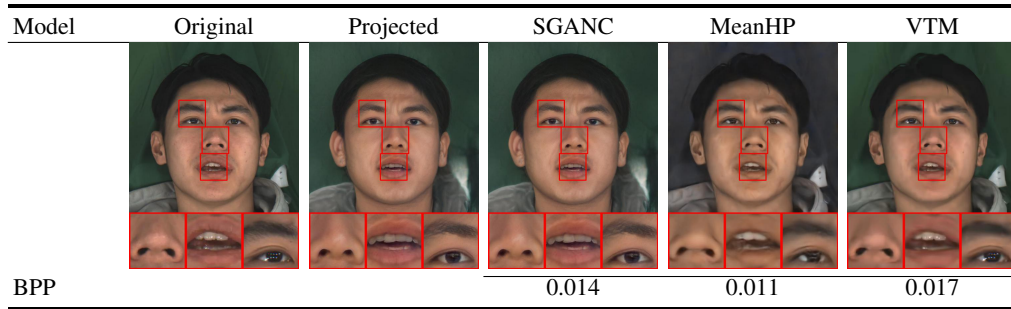


Figure 11: Qualitative results for image compression: Other methods introduce blocking artifacts and blurring. Our method (SGANC) leads to high quality reconstruction and perceptually lower distortion. Images are better seen zoomed.

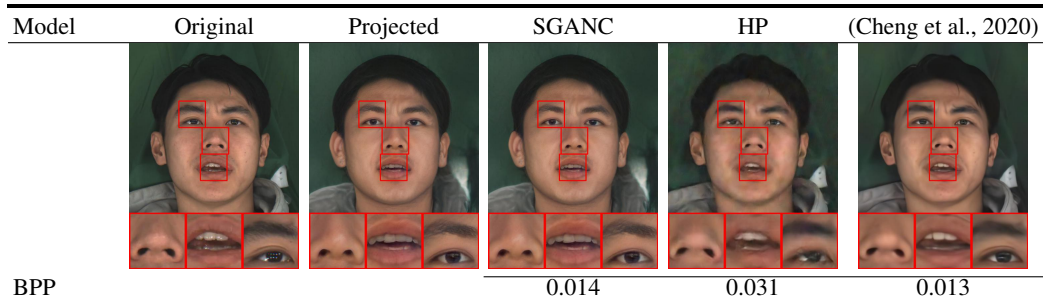


Figure 12: Qualitative results for image compression: Other methods introduce blocking artifacts and blurring. Our method (SGANC) leads to high quality reconstruction and perceptually lower distortion. Cheng et al. (2020) is better than HP, but still the results are blurred. Images are better seen zoomed.

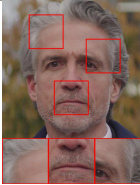
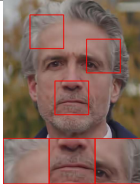
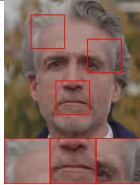
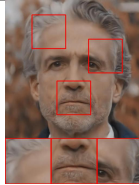
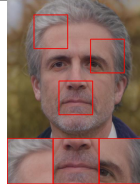
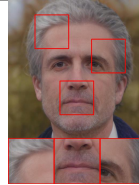
Model	Original	VTM	AV1	MeanHP	Projected	SGANC $l=1e-5$
						
PSNR		35.46	32.79	31.64		29.31
MS-SSIM		0.933	0.894	0.945		0.970
BPP		0.0157	0.0127	0.0129		0.0146
LPIPS		0.274	0.360	0.344		0.062

Figure 13: Qualitative results for image compression: Other methods introduce blocking artifacts and blurring. Our method (SGANC) leads to high quality reconstruction and perceptually lower distortion. Images are better seen zoomed.

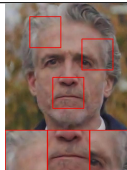
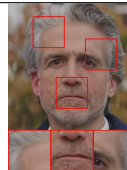
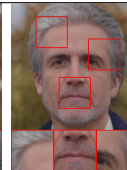
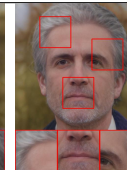
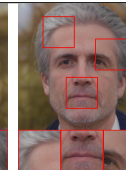
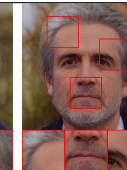
Model	Original	HP	(Chen et al., 2020)	Projected	SGANC $l=1e-4$	SGANC $l=1e-5$	SGANC $l=1e-6$
							
PSNR		33.50	34.79		38.85	29.31	18.71
MS-SSIM		0.915	0.959		0.995	0.970	0.834
BPP		0.0217	0.0152		0.0280	0.0146	0.0066
LPIPS		0.320	0.281		0.010	0.062	0.250

Figure 14: Qualitative results for image compression: Other methods introduce blocking artifacts and blurring. Our method (SGANC) leads to high quality reconstruction and perceptually lower distortion. Images are better seen zoomed.

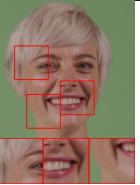

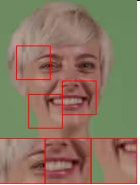
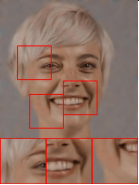
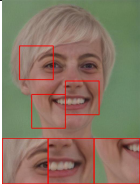
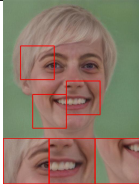
Model	Original	VTM	AV1	MeanHP	Projected	SGANC $l=1e-5$
						
PSNR		38.73	35.51	27.15		30.56
MS-SSIM		0.966	0.940	0.953		0.967
BPP		0.0102	0.0095	0.0098		0.0149
LPIPS		0.1758	0.2201	0.4470		0.061

Figure 15: Qualitative results: Other methods introduce blocking artifacts and blurring. Our method (SGANC) leads to high quality reconstruction and perceptually lower distortion. Images are better seen zoomed





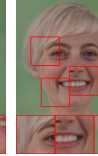
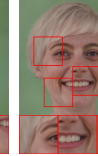
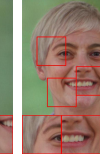
Model	Original	HP	(Chen et al., 2020)	Projected	SGANC $l=1e-4$	SGANC $l=1e-5$	SGANC $l=1e-6$
							
PSNR		34.98	36.89		41.43	30.56	17.97
MS-SSIM		0.938	0.978		0.995	0.967	0.716
BPP		0.0216	0.0089		0.0289	0.0149	0.0068
LPIPS		0.2498	0.1718		0.008	0.061	0.317

Figure 16: Qualitative results: Other methods introduce blocking artifacts and blurring. Our method (SGANC) leads to high quality reconstruction and perceptually lower distortion. Images are better seen zoomed

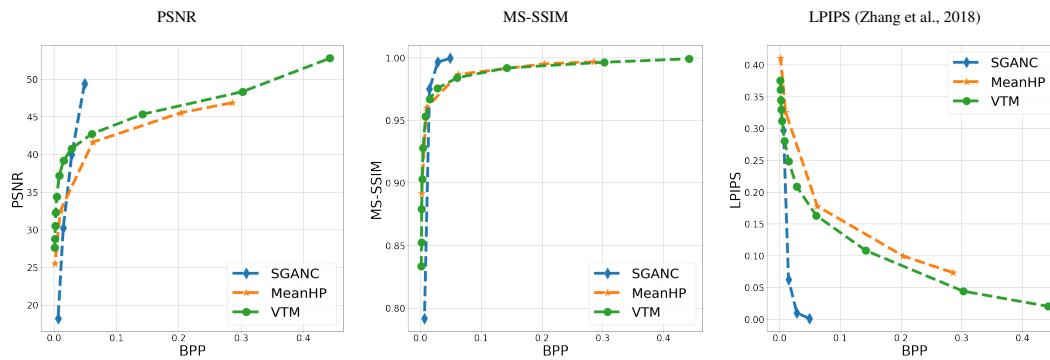


Figure 17: Rate distortion curves on FILMPAC dataset: for medium and large BPP, Our method (Blue) is better in terms of LPIPS and MS-SSIM than VTM (Green), MeanHP (Orange). For high BPP, our method is better in terms of PSNR.

A.4.2 VIDEO COMPRESSION

In this section, we show quantitative and qualitative results for video compression. Contrary to our main results in the paper, we here compress the original video and compute the metrics with original frames (except for our approach still compared to the projected frames).

From Figures 18, 20, 21, 22 and 23, we can notice that VTM and H.265 introduces blocking and blurring artifacts which is not case for our approach SGANC IC. We can also notice similar observations in Figures 24, 25, 26 and 27.

From Figure 19, we can notice that our methods are better perceptually (LPIPS and MS-SSIM) than VTM and H.265.

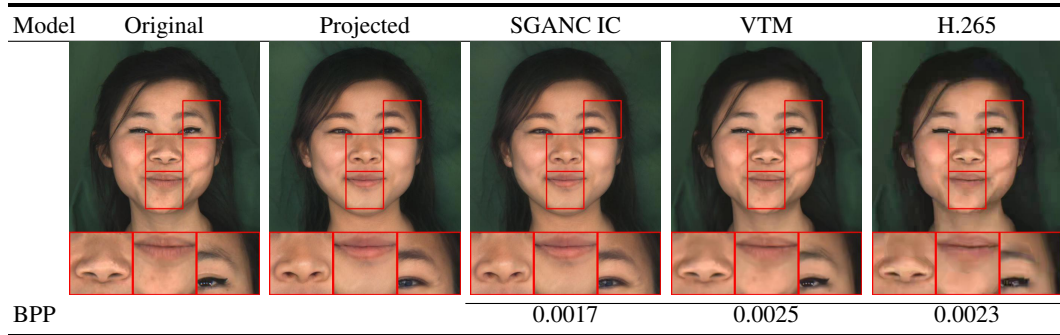


Figure 18: Qualitative results for video compression: Other methods introduce blocking artifacts and blurring. Our method (SGANC IC) leads to high quality reconstruction and perceptually lower distortion. Images are better seen zoomed

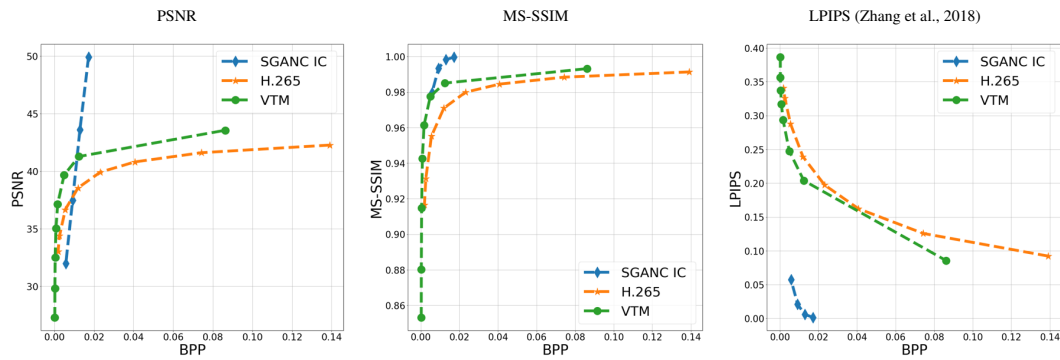


Figure 19: Rate distortion curves for video compression on Filmpac dataset. The perceptual metric LPIPS clearly shows a perceptual improvement of our method.

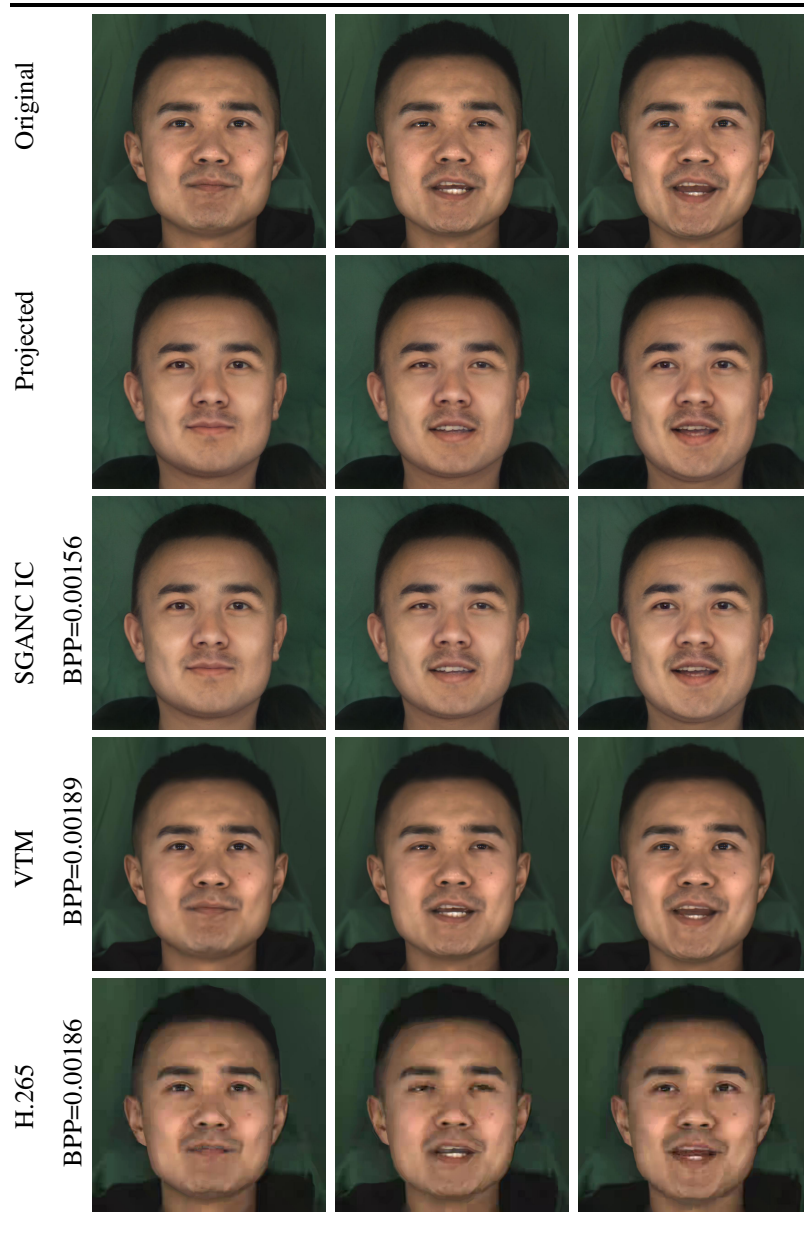


Figure 20: Qualitative results on MEAD inter dataset for extreme BPP, computed for three video frames. H.265 shows blocking artifacts and blurring, VTM shows blurring especially at the edges of the face and the hair while our method (SGANC) is almost artifacts free and with high quality images.

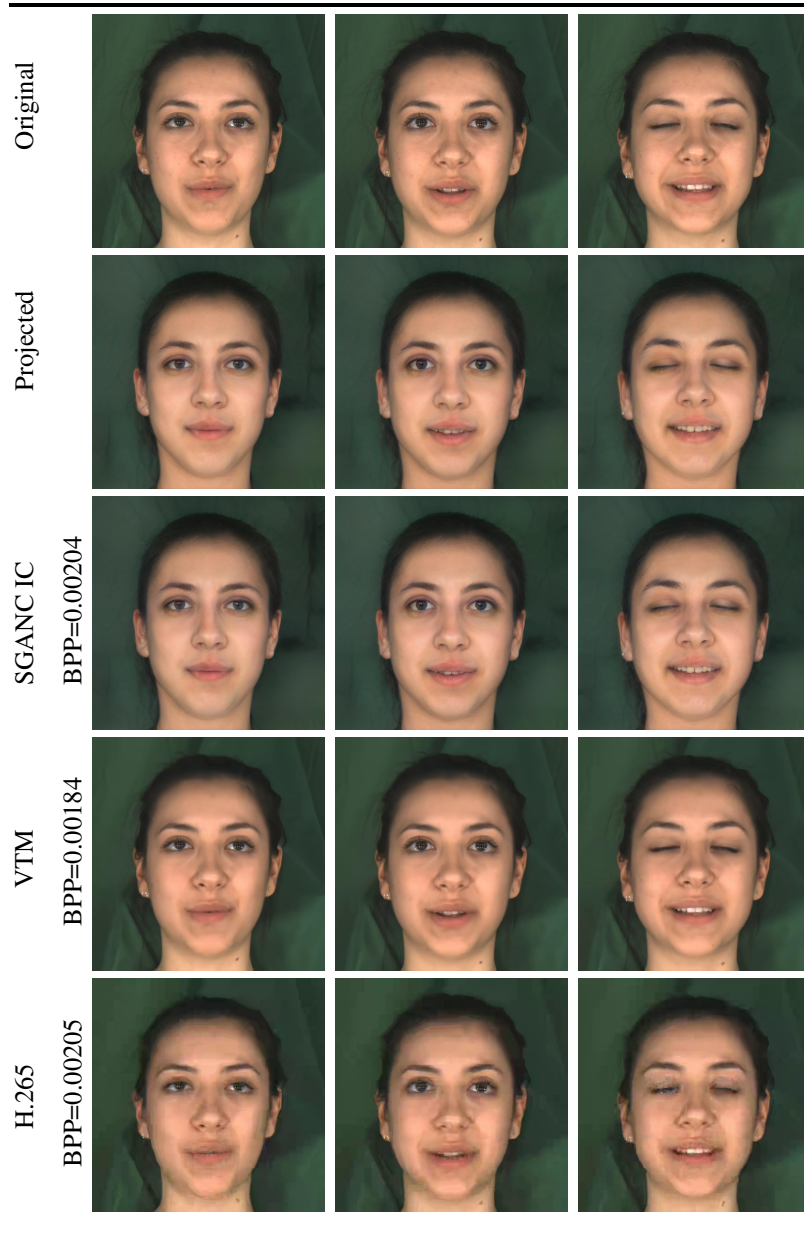


Figure 21: Qualitative results on MEAD inter dataset for extreme BPP, computed for three video frames (columns). H.265 shows blocking artifacts and blurring, VTM shows blurring especially at the edges of the face and the hair while our method (SGANC) is almost artifacts free and with high quality images.

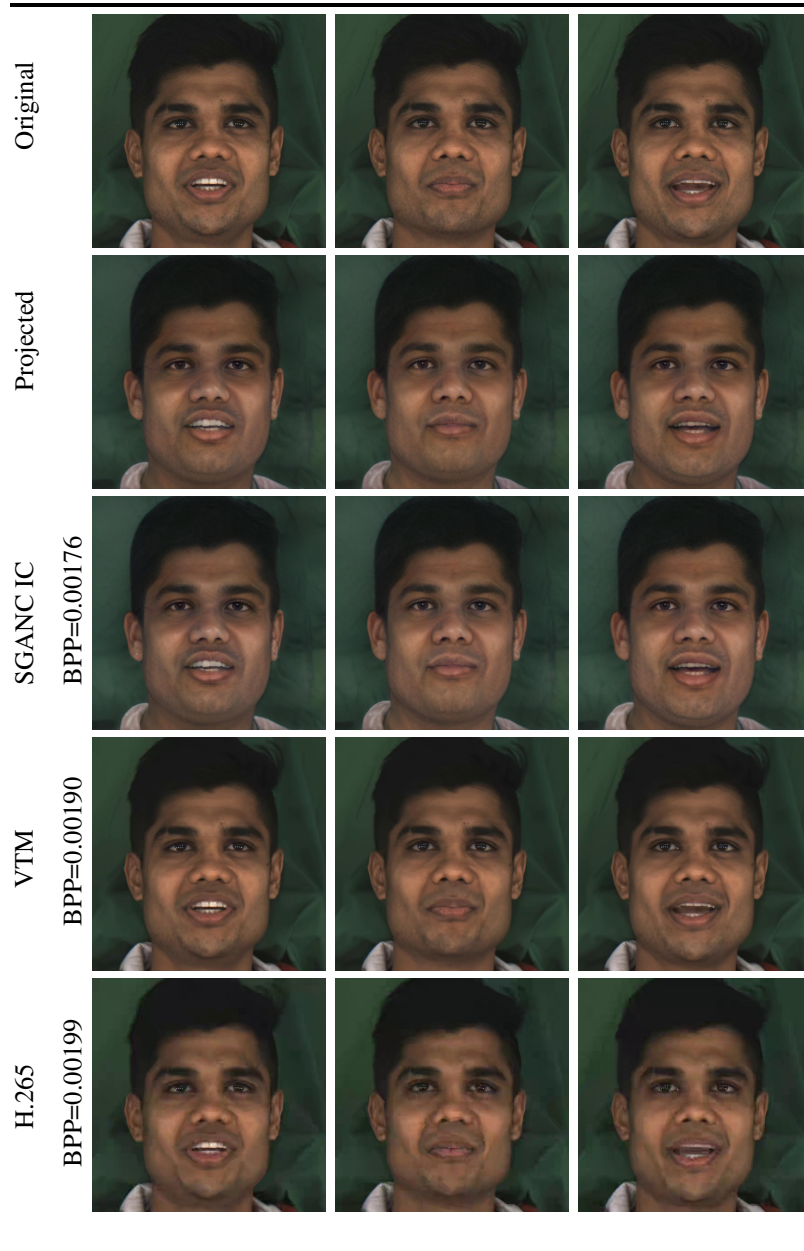


Figure 22: Qualitative results on MEAD inter dataset for extreme BPP, computed for three video frames. H.265 shows blocking artifacts and blurring, VTM shows blurring especially at the edges of the face and the hair while our method (SGANC) is almost artifacts free and with high quality images.

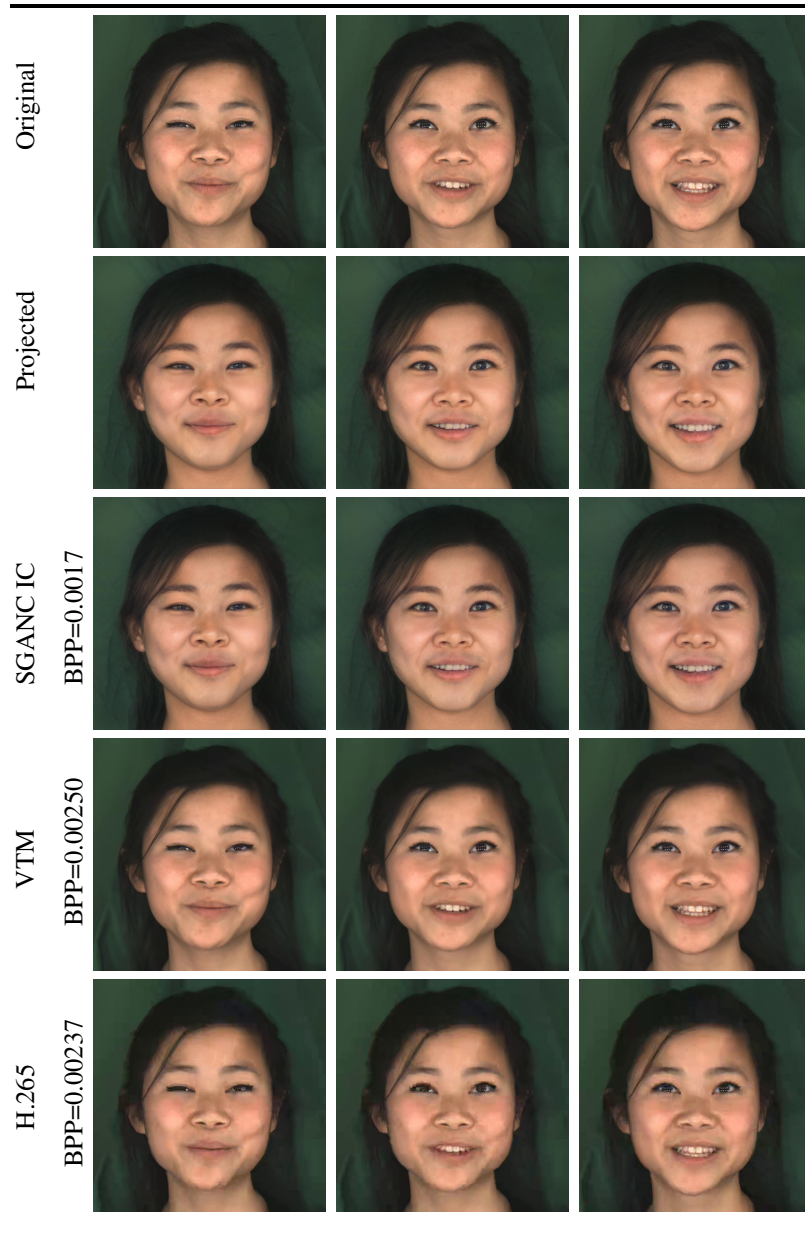


Figure 23: Qualitative results on MEAD inter dataset for extreme BPP: H.265 shows blocking artifacts and blurring, VTM shows blurring especially at the edges of the face and the hair while our method (SGANC) is almost artifacts free and with high quality images.

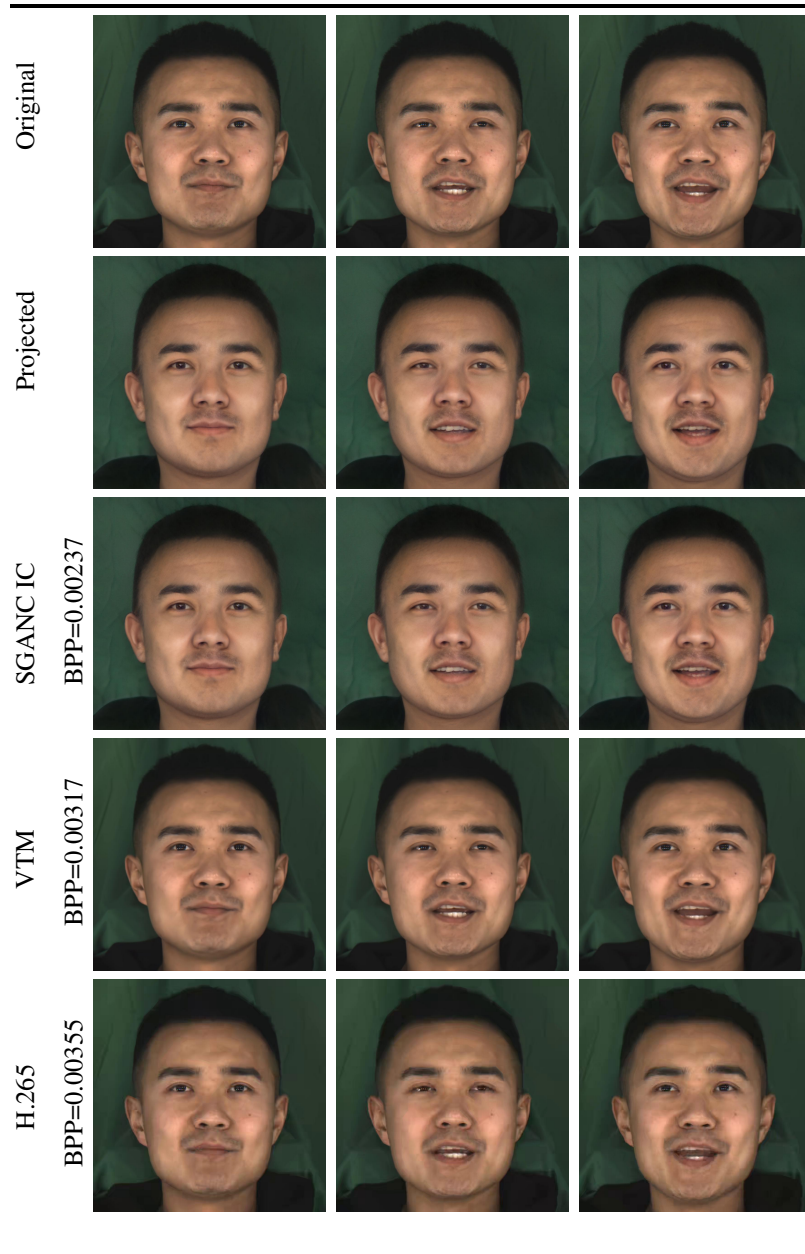


Figure 24: Qualitative results on MEAD inter dataset: H.265 shows blocking artifacts and blurring, VTM shows blurring especially at the edges of the face and the hair while our method (SGANC) is almost artifacts free and with high quality images.

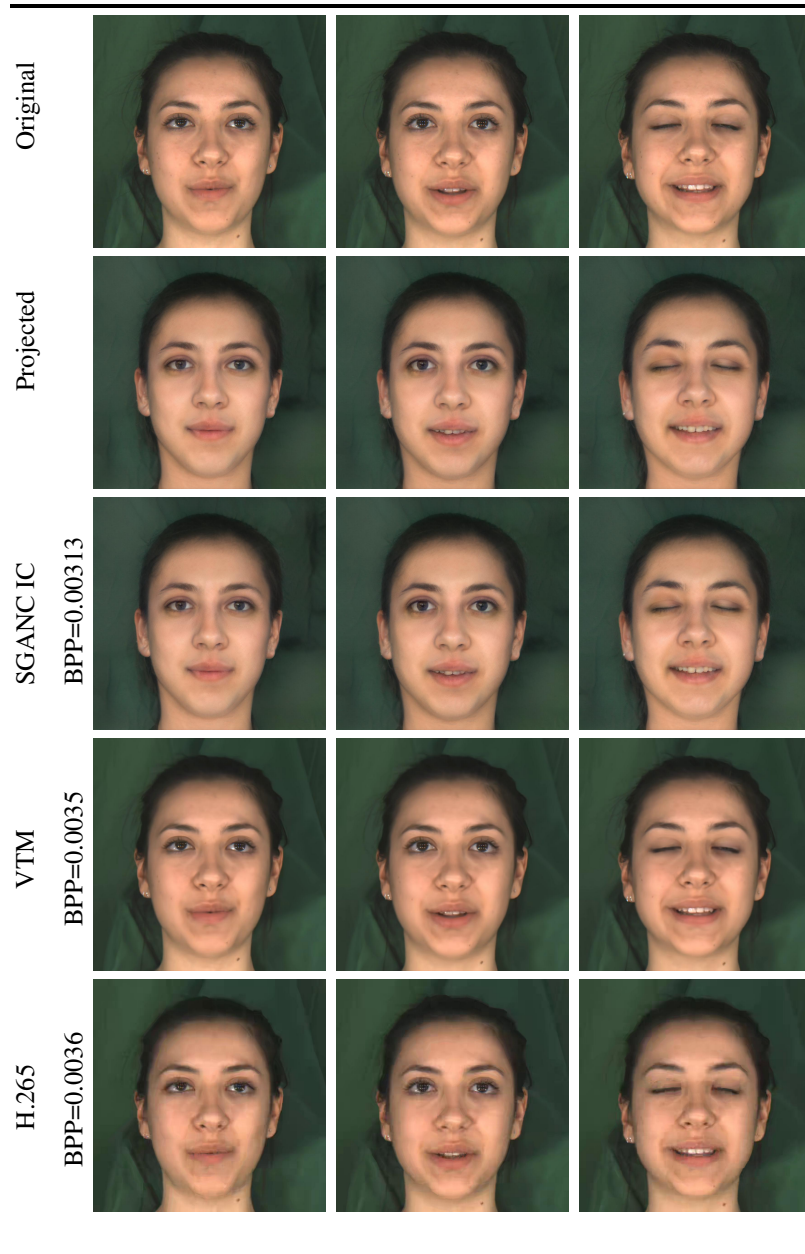


Figure 25: Qualitative results on MEAD inter dataset: H.265 shows blocking artifacts and blurring, VTM shows blurring especially at the edges of the face and the hair while our method (SGANC) is almost artifacts free and with high quality images.

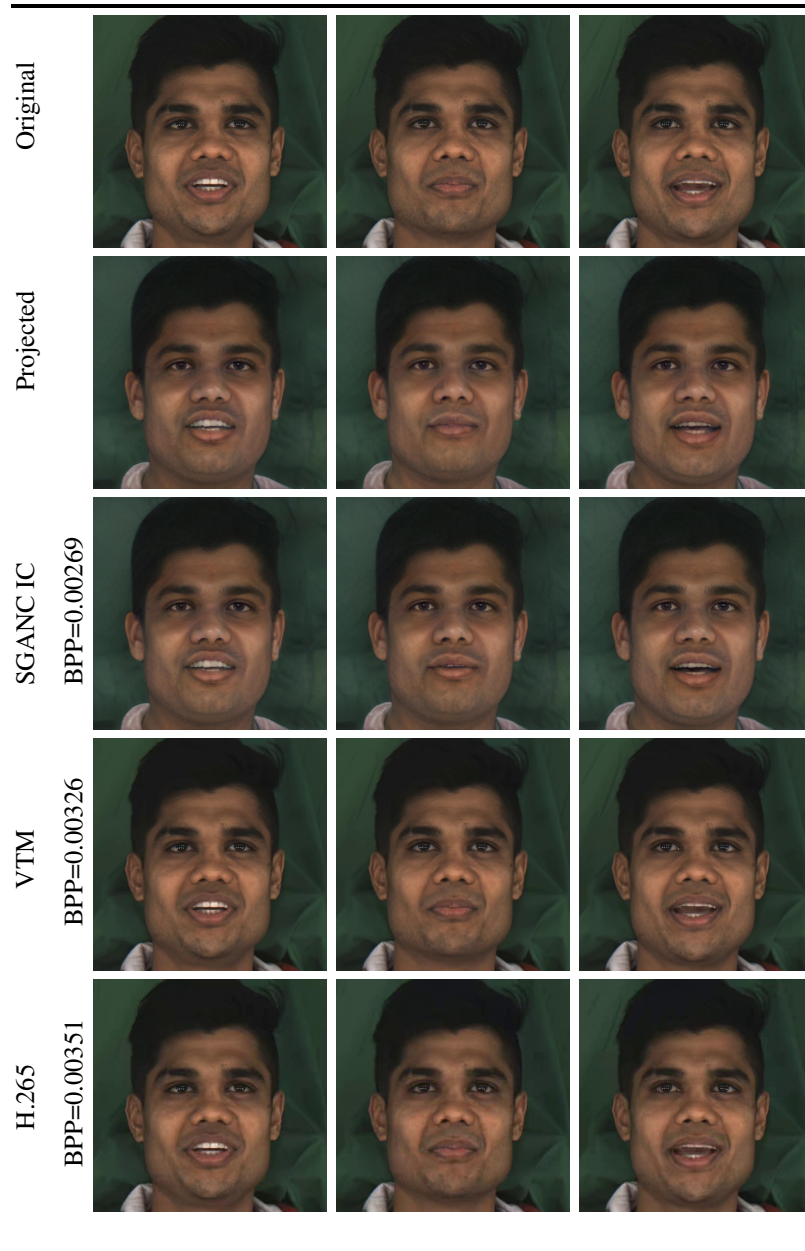


Figure 26: Qualitative results on MEAD inter dataset: H.265 shows blocking artifacts and blurring, VTM shows blurring especially at the edges of the face and the hair while our method (SGANC) is almost artifacts free and with high quality images.

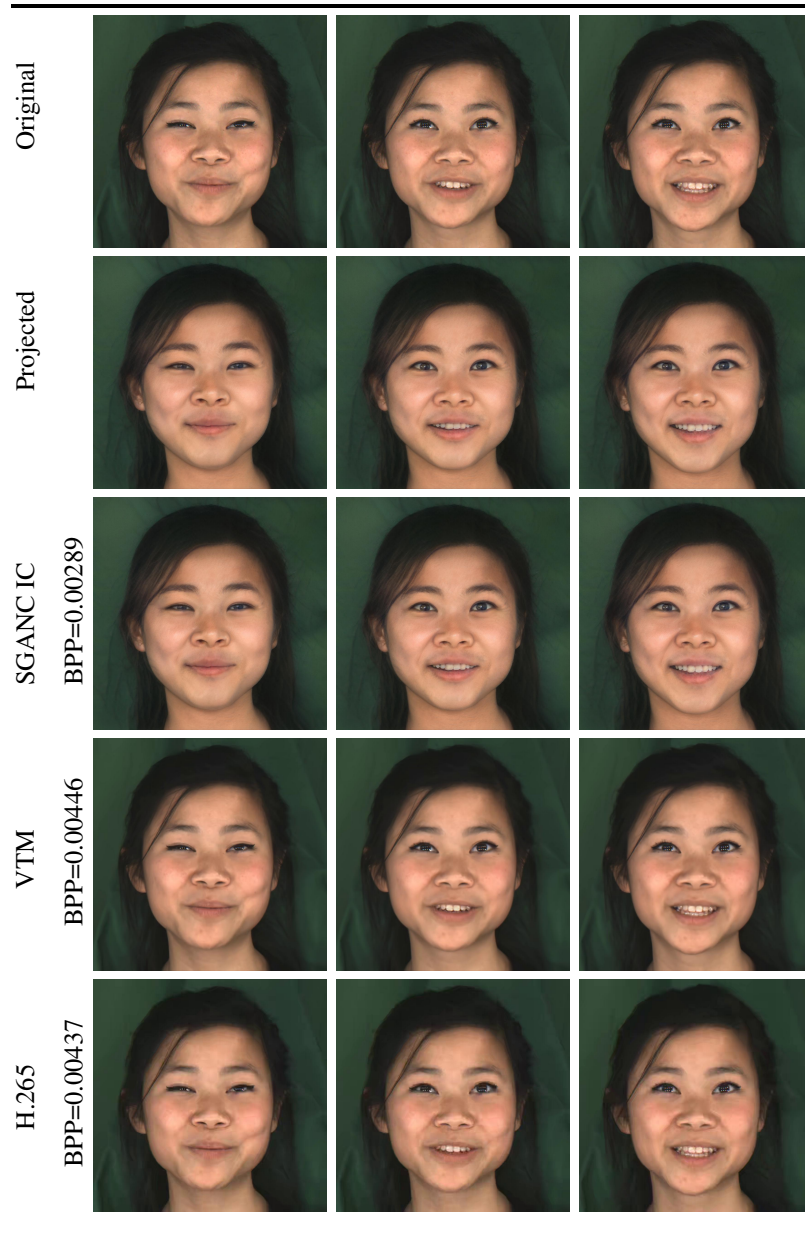


Figure 27: Qualitative results on MEAD inter dataset: H.265 shows blocking artifacts and blurring, VTM shows blurring especially at the edges of the face and the hair while our method (SGANC) is almost artifacts free and with high quality images.