# HEVC Compression Artifact Reduction with Generative Adversarial Networks

Shiqi Yu, Bolin Chen, Yiwen Xu, Weiling Chen, Zhonghui Chen, Tiesong Zhao
*Fujian Key Lab for Intelligent Processing and Wireless Transmission of Media Information*
*Fuzhou University, Fuzhou, China*
t.zhao@fzu.edu.cn

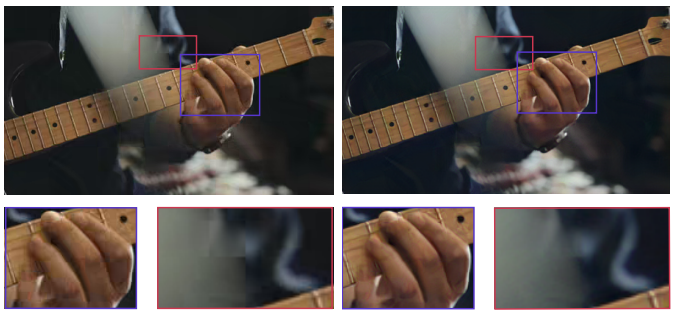
***Abstract*—Video compression technology is significant to video transmission and storage. However, compression artifacts arise in videos. Specifically, coarse quantization eliminates video details and degrades visual quality. Most of artifact reduction methods use filter processing or Mean Square Error (MSE) loss that leads to over-smoothing results. Moreover, most of the methods target to reduce single image compression artifact instead of video artifact. In this paper, we present an adversarial learning method with recurrent framework called Video Artifact Reduction Generative Adversarial Network (VRGAN). Our network contains a generator with recurrent framework that improves video consistency, a dense block that enhances receptive field for large transform unit, and a relativistic discriminator that evaluates the relationship between the generated frames and the original high-quality frames. Our VRGAN is able to generate more realistic videos. The effectiveness in reducing video compression artifacts of the method has been demonstrated qualitatively and quantitatively. The performance comparison with previous works shows the superiority of the proposed method.**

***Index Terms*—Visual Quality, Video Coding, Compression Artifact Reduction, Video Restoration, Adversarial Learning.**

## I. Introduction

According to the prediction of Cisco Visual Networking Index [1], video traffic will occupy 80% of all Internet traffic by 2022. With the increase of video capture devices combined with the popularity of panoramic video, High Definition (HD)/Ultra HD resolution video and 3D video, there exists great demand to high-performance video compression. Uncompressed video is difficult to be transmitted and stored for its huge data size. New standards of lossy video compression ($e.g.$, High Efficiency Video Coding (HEVC), VP9) give great convenience to video transmission and storage. For example, compared with Advanced Video Coding (AVC), HEVC standard saves 50% of storage, reduces the burden on server's bandwidth and prevents stalling in online high-resolution video playback. While watching HD online videos at low bit rates, it is easy to find obvious artifacts on videos. Compression artifacts arise as the loss of detail accompanied by the appearance of blocking and noise. Furthermore, video compression artifacts comprise blocking, blurring, ringing, floating, color bleeding and flickering, $etc.$ Fig. 1 shows that the decoded video consist of various of compression artifacts such as blocking, ringing and blurring.

Videos compressed by larger quantization parameters are easier to be transferred with smaller bit occupation. However, in compressed videos, artifact arises and obviously degrades visual quality as a result of adding visible boundaries and making video details blurry. A large number of methods have been proposed to reduce compression artifacts. Spatial filters [2] have been integrated to compression standard. Shape adaptive DCT domain filters [3] have been employed to reduce image noise successfully. Recently, convolutional neural networks have been demonstrated the effectiveness of artifact reduction. Mean Square Error (MSE) loss function which measures distortion is widely used to solve distortion related issues. Various convolutional neural networks combined with MSE loss function are proposed to reduce artifacts. However, as a result of being weakly correlated to perceptual quality [4], MSE loss leads to over-smoothed result with loss of high frequency details. Recently, Generative Adversarial Networks (GANs) are developed to generate complex images [5]. In the previous work of compression artifact removal, generative adversarial learning has been demonstrated great capability in the artifact reduction [6]. Owing to the success in avoiding blurring while not relying on manual feature extractor, the images recovered by GANs are extremely clear and realistic. Even though high-frequency information is retained in generated images of GANs, the networks are not capable of video discontinuity repairing. Moreover, most of the artifact reduction methods aim to process single image. Video information between adjacent frames remain an underused resource.



(a) (b)

Fig. 1. (a): A frame from HEVC compressed video. Regions of compression artifact are magnified. (b): The frame processed by the proposed method reaches better visual quality.

In this paper, we employ GAN to reduce intra-frame and inter-frame artifacts. The method generates visual pleasing video while keeping the coherence of frames. Our contributions can be summarized as follows:

- Firstly, the discontinuity of compressed videos degrades visual quality. We apply the recurrent framework to our model to keep the coherence among the adjacent frames.
- Secondly, HEVC videos are compressed by large transform units, which cause large scale artifact arising. We define a large receptive field generator network with dilated convolution for HEVC. Mixed convolutions are applied to reduce large scale artifacts and recover contents from larger area.
- Thirdly, the generator of standard GAN is unable to receive feedback from the original high-quality images. We adopt the relativistic discriminator to assess the probability of the generated frame being more realistic than the high-quality frame.

## II. Related Work

In the artifact reduction task, methods can be classified as processing-based and learning-based. Furthermore, some methods target for denoising, while the other methods target for restoration.

Spatial domain processing-based algorithms for artifact reduction are commonly suitable for being embedded into encoding process. For example, in-loop deblocking filter [2] of H.265/HEVC is an adaptive filter designed for different kinds of coding unit boundary. Frequency domain processing-based algorithms work as filters to process the frequency domain information of image. SA-DCT [3] presented by Foi $et\ al.$ is a DCT domain method, which uses pointwise arbitrary adaptive shape discrete cosine transform with lower complexity to enhance the quality of image while avoiding Gibbs phenomenon. Wu $et\ al.$ utilized Meyer algorithm to perform wavelet transform for JPEG2000 image artifact reduction [7].

Learning-based algorithms have been applied to artifact reduction with the development of deep convolutional neural networks. Most of learning-based algorithms are used as post-processing techniques. For image artifact, recent work AR-CNN [8] proposed by Dong $et\ al.$ applied convolutional neural network and transfer learning to artifact reduction. Dai $et\ al.$ proposed VRCNN [9], which embedded CNN processing into the intra coding of HEVC to reduce artifact. Galteri $et\ al.$ proposed deep generative adversarial method for JPEG image artifact reduction, which greatly improved visual quality [6]. For multi-frame recurrent artifact reduction, Lu $et\ al.$ proposed deep Kalman filter [10] to enhance video quality by optimal estimation. In [11], Meng $et\ al.$ designed bidirectional convolutional LSTM to enhance video quality.

Processing-based methods have a tendency to make the image blur as a result of averaging. Furthermore, in some cases, ringing may arise as a result of Gibbs phenomenon. As for learning-based methods, although MSE loss or other full-reference measurements reduce pixel-wise distortion, they are still having little correlation to visual quality. GAN has been applied for artifact reduction of JPEG, however, the GAN-based method for JPEG artifact [6] has not considered the inter-frame processing. Besides, HEVC compression applies much larger transform unit than JPEG compression. Therefore, the GAN-based method for JPEG artifact is not suitable for HEVC video artifact. The proposed VRGAN has frame recurrent framework and larger receptive field to address these problems. Besides, we adopt the Relativistic Standard GAN (RSGAN) [12] to enhance the capability of our model.

## III. Proposed Method

The task of compression artifact reduction is to reconstruct an artifact reduced video from a compressed video. In this section, we formulate the video compression artifact reduction problem and describe our optimization objective for HEVC compression artifact reduction.

### A. Problem Formulation

A formulation of degradation model in video compression can be described as:

$$V_n^{LQ} = f_c(V_n^{HQ}, Qp), \tag{1}$$

where $n$ denotes video index, $V_n^{HQ}$ is the $n^{th}$ high-quality video input to video encoder, $f_c$ is the quantization loss of the input video, $V_n^{LQ}$ is the decoded video, and $Qp$ is the quantization parameter.

Video artifact reduction is to recover compressed videos $V_n^{LQ}$ from observed degraded video information. We target to solve the function that acquires the parameters $\hat{\theta}$ to minimize the loss function between the generated videos and the high-quality videos:

$$\hat{\theta} = \arg\min_{\theta}(\sum_{n\in N} l_g(A(S_n^{LQ}), S_n^{HQ})). \tag{2}$$

In the formula above, we show the optimization target of the proposed method. Function $A$ is the convolution neural networks. The inputs of $A$ are the low-quality frame sets. $\hat{\theta}$ denotes the parameters of the proposed generator. Loss function $l_g$ denotes the generator loss which measures the distance between the original high-quality video and each frame generated from low-quality video. The low-quality video frame sets $S_n^{LQ}$ and the high-quality sets $S_n^{HQ}$ of $N$ videos are input to the loss function, respectively. The network training process is to minimize the loss function $l_g$ by optimizing the networks A.

### B. Recurrent Framework

As is described in Fig. 2, to maintain the coherence of the generated frames, we use flow estimation. The motion compensation is the key process of our recurrent method. We train the flow estimation model $f_{est}$ based on the methods proposed in [13] and [14] for motion compensation. The flow estimation is to predict the variety of location coordinate from two adjacent frames. The relative relationship of the previous frame $I_{n-1}^{LQ}$ and the current frame $I_n^{LQ}$ is predicted from pixel to coordinate residual. Therefore, the motion between two frames is estimated. The input of the flow estimation nets is the current frame and the previous frame. The flow estimation nets output the predicted location coordinate residual $F$ of $I_n^{LQ}$ and $I_{n-1}^{LQ}$ as Eq. (3) shows:

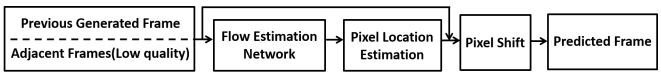



Fig. 2. The figure shows the recurrent frame prediction procedure of our method. Pixel location maps are estimated from the flow estimation network. Flow guided pixel shift is applied to the previous generated frame.

$$F = f_{est}(I_n^{LQ}, I_{n-1}^{LQ}). \tag{3}$$

The pixel location coordinate residual is added to the location map of the previous frame to estimate the pixel location of current frame. The previous frame of the first low-quality frame and the generated frame are initialized to constant matrices. To estimate pixel location $\hat{L}_n$ of the current frame, location residual $F$ which estimated by flow nets is added to the location $L_{n-1}$ of the previous frame as motion compensation:

$$\hat{L}_n{=}L_{n-1} + F. \tag{4}$$

The previous generated frame $\hat{I}_{n-1}$ implements pixel location shift guided by the predicted location coordinate $\hat{L}_n$. Then, the prediction result and the current low-quality frame are input to the generator. The generator receives combined channels of location estimated frame and compressed current frame. The generator reconstructs the current frame as:

$$\hat{I}_n = G(f_l(\hat{L}_n, \hat{I}_{n-1}), I_n^{LQ}). \tag{5}$$

In the equation above, $f_l$ denotes the transform from motion compensated location coordinate to image pixel shift, $\hat{I}_n$ denotes the current frame generated. The input of $f_l$ is the predicted location coordinate $\hat{L}_n$ and the previous generated frame $\hat{I}_{n-1}$. The generator receives the input, which includes the frame after pixel shift and the current frame $I_n^{LQ}$. Then, $G$ generates the quality enhanced current frame $\hat{I}_n$.

### C. Network Architecture

GANs [5] are a class of network architectures that a generative model $G$ and a discriminative model $D$ contest with each other. The discriminator $D$ targets to conduct the generator $G$ to generate images following more realistic probability distributions. Conditional GAN [15] is adopted as training steps in VRGAN. Specifically, when the generative network generates a image with high quality, the discriminator is updated to output a number close to 1, otherwise the discriminator is updated to output a number close to 0.

In Fig. 3, we show the generator architecture which uses motion compensated image and current decoded frame as input. The generator consists of two convolution layers in the front, a dense block in the middle and two convolution layer in the end. SELU [16] is used as activation function. Scaled residual connection [17] is applied between the low quality frame and the output of the generator. The scaling factor is set to 0.1 for training stability. The generator consists of four convolution layers and a densely connected block in total.

Enlarging receptive field needs larger convolution kernels or deeper networks which have more training complexity are more difficult to be trained. Dilated convolution [18] enlarges the coverage area of the convolution cores while

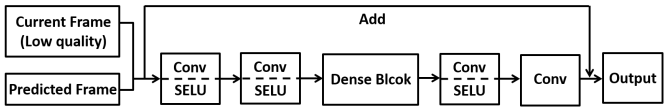



Fig. 3. The figure shows the generator of our method. The input of generator is the combination of the flow estimation network output and the low-quality frame. The output of generator is the quality enhanced current frame with the same size and the channel number of the low-quality frame.

avoiding computational complexity increase caused by larger kernel. Moreover, compared with stride length increase, dilated convolution keeps the resolution of the input, while stride length adjustment significantly reduces the resolution. We apply dense block [19] with mixed convolutional [20] layers to our generator. In these mixed layers, dilated convolutions combine standard convolutions for blind spots removal. The mixed convolutions we adopted enlarge the receptive field. The mixed convolutions are able to reduce larger-scale artifacts without blind spots.

The largest transform unit of HEVC is $32 \times 32$, which is larger than JPEG $8 \times 8$ transform unit. Thus, we use dilated convolution to enlarge the receptive field. Specifically, we adopt mixed convolutional layers. We set several dilation rates. For example, the dilation rates are set to 2, 4, 8 and 16 in Fig. 4. The areas which dilated convolution kernels cover are $5 \times 5$, $9 \times 9$, $17 \times 17$ and $33 \times 33$, respectively. The receptive field can be computed as follow:

$$m_i = (m_{i-1} - 1)s_i + k_i. \tag{6}$$

In this formula, $i$ denotes the ordinal of the layer. $m_i$ is the receptive field of the $i$th layer. $s_i$ is the stride of the layer while $k_i$ is the kernel size of the layer. Computed by this function, our generator with mixed convolutions in Fig. 4 reaches $71 \times 71$ receptive field, while AR-CNN [8] reaches $19 \times 19$ and the method from Galteri et. al [6] reaches $22 \times 22$. In HEVC compression, the largest transform unit is $32 \times 32$. Our receptive field is larger than $32 \times 32$.

The dense block is shown in Fig. 4. Each mixed convolutional layer is densely connected by channel-wise concatenation. Densely connection is employed to enhance the capability of the generator.

The network architecture of discriminator is similar to JPEG artifact reduction network based on GAN from Galteri *et al* [6]. We substitute all the Leaky ReLU activation functions in the discriminator network with SELU [16]. Sigmoid function is removed. Each frame is divided into $16 \times 16$ channel-wise combined sub-patches [6] for ringing artifact reduction.

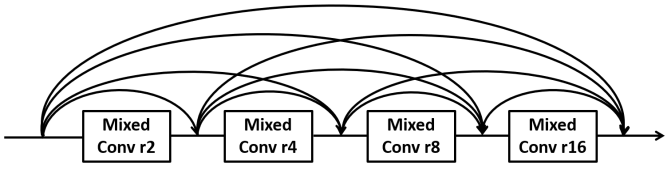



Fig. 4. The figure shows the dense block of our method. We employ mixed convolutional architecture.

### D. Relativistic Discriminator

We use RSGAN [12] as adversarial loss function, which is shown to reach better fitting ability. VRGAN measures relativistic distance. The relativistic loss enhances the discriminator and improves the stability of the training process. The generator gets gradients from generated frame $x_f$ and original high-quality frame $x_r$ in VRGAN rather than only gets gradients from generated result in Conditional GAN (CGAN). The adversarial loss functions are as follows:

$$L_G^R = -\mathbb{E}_{(x_r,x_f)}[\log(\text{sigmoid}(D_R(x_f, x_r)))], \quad (7)$$

where $L_G^R$ is the adversarial loss for generator and $L_D^R$ is the adversarial loss for discriminator. The adversarial loss functions of discriminator $D_R$ and generator are symmetrical:

$$L_D^R = -\mathbb{E}_{(x_r,x_f)}[\log(\text{sigmoid}(D_R(x_r, x_f)))]. \quad (8)$$

### E. Loss Function

For video compression artifact reduction, the discriminator loss in detail is designed as follows:

$$l_d = -\sum_{n\in N}[\log(\text{sigmoid}(D(I_n^{HQ}) - D(\hat{I}_n)))]. \quad (9)$$

The discriminator is trained using function $l_d$. In this loss function, $N$ is the number of total frames in the video while $n$ is the $n^{th}$ frame of the video. For video compression artifact reduction, $D(\hat{I}_n)$ is the discriminator output of the generated frame while $D(I_n^{HQ})$ is the discriminator output of the original high-quality frame.

Our generator loss consists of reconstruction loss, flow loss, perceptual loss and adversarial loss. Eq. (10) shows the generator loss:

$$l_g = \alpha l_{flow} + \beta l_p + \lambda l_{advg} + \mu l_{rec}. \quad (10)$$

The adversarial loss for generator $l_{advg}$ is $L_G^R$ above. The relativistic discriminator evaluates the relative quality of the generated video and the original high-quality video. The generator optimized by relativistic discriminator generates better visual quality frames. The adversarial loss of generator for video artifact is expressed in Eq. (11):

$$l_{advg} = -\sum_{n\in N}[\log(\text{sigmoid}(D(\hat{I}_n) - D(I_n^{HQ})))], \quad (11)$$

where the adversarial loss for generator of the proposed method is symmetrical to the discriminator loss.

The reconstruction loss $l_{rec}$ is L1 loss of the high-quality video and the generated video. We adopt L1 loss for low frequency content reconstruction and noise reducing of the generated video. The generative adversarial artifact removal has the drawback of adding extra noise and distortion. Therefore, the reconstruction loss is adopted to recover the low frequency details while reducing extra noise and distortion.

The flow loss $l_{flow}$ is adopted to keep the correction of the pixel location in the estimated frame:

$$l_{flow} = \sum_{n\in N}\frac{1}{WH}(f_l(\hat{L}_n, I_{n-1}^{LQ}) - I_n^{HQ})^2. \quad (12)$$

On the basis of the recurrent framework, current frame is influenced by several previous frames. In the formula, $W$ and $H$ are width and height. Distance is computed between each pixel shifted image and the corresponding original high-quality video frame. Function $f_l$ denotes the pixel shift.

Perceptual loss described in [21] and [22] is applied to reconstruct feature presentation of the generated video. The formula of the perceptual loss $l_p$ is given by:

$$l_p = \sum_{n\in N}\sum_{x\in W_j}\sum_{y\in H_j}\frac{1}{W_j H_j}(\phi_j(\hat{I}_n)_{x,y} - \phi_j(I_n^{HQ})_{x,y})^2, \quad (13)$$

where function $\phi_j$ is the $15^{th}$ convolution layer of a pre-trained VGG19 network before activation. Besides, $W_j$ and $H_j$ denote the width and height of the feature maps of the layer.

## IV. Experiment

### A. Implementation Details

To evaluate the effectiveness of our method, we employ Vimeo90k [23] septuplets dataset for training and testing. The dataset consists of 91701 video sequences. We choose 88950 sequences for training and 2751 sequences for testing. Every sequence consists of 7 frames in the same scene. Video sequences from the dataset are compressed by HEVC. The quantization parameter (Qp) of compression is set to 32 and 37. The resolution of the sequences is $448 \times 256$. The sequences are directly input to the networks with $448 \times 256$ complete size. The output resolution is $448 \times 256$ as a result of padding. Considering the sequences are encoded by HEVC in YUV format, we convert sequence color RGB to YUV according to ITU-R BT.601 standard and extract luminance channel Y to evaluate the quality of the sequences. Our models are trained with 2080Ti GPUs.

We use Adam optimizer where $\beta_1$ is 0.9, $\beta_2$ is 0.99 and the learning rate is 0.0001. The dilation rates are set to 2, 4 and 8. The model training of VRGAN has been carried on for 500,000 iterations. The batch size is set to 3 sequences (21 frames per iteration in total) in VRGAN training process.

### B. Results Comparison

For comparison, we denote the HEVC compression without loop-filter as HEVC and denote the HEVC compression with loop-filter as HEVC-LF. The input video of models are compressed by HEVC without loop-filter. We compress the videos with different quantization parameters. The models are trained and evaluated with quantization parameter Qp=32 and Qp=37 separately. We train our model and compare performance with the HEVC compression, HEVC-LF and two learning-based approaches. For comparison, we trained AR-CNN which designed for JPEG artifact reduction from Dong $et\ al.$ and the method proposed by Galteri $et\ al.$ which designed based on GAN for JPEG artifact reduction. AR-CNN is a deep convolutional network for image artifact reduction. The deep convolutional network of AR-CNN removes blocking and ringing artifacts while avoiding blurring and retaining details.

The method from Galteri *et al.* is a conditional generative adversarial framework for image compression artifact reduction, which realistically recovers high frequency details.

Compression artifact reduction is one of the low-level vision problems in the field of computer vision. It has been proved in [24] that perceptual quality and distortion are contradictory with each other in low-level vision tasks. In these cases, no-reference measurement for perceptual quality is anti-correlated to full-reference measurement for distortion. Hence, we adopt no-reference quality assessment natural image quality evaluator (NIQE) [25] and video intrinsic integrity distortion evaluation oracle (VIIDEO) [26] to evaluate perceptual quality of single frame and multi-frame results respectively.

The visual quality comparison of reconstructed images is depicted in Fig. 5. It can be observed that our method generates clearer image compared to other methods. Moreover, blocking and ringing artifacts have been reduced. Our method generates more realistic details and sharper edges compared to AR-CNN as a result of adversarial learning. We use NIQE to evaluate single image visual quality. NIQE is based on fitting multivariate Gaussian density model, which is obtained from perceptually relevant spatial domain natural scene statistic features. Therefore, NIQE has high consistency with subjective image evaluation. The result of our method has the lowest NIQE value. This indicates that the result of our method reaching the best quality in single frame no-reference measurement NIQE compared to other methods while outperforming 1.11dB video PSNR compared to method from Galteri *et al.*

To evaluate the distortion, we perform experiments to compare the average PNSR of different methods on the evaluation dataset which contains 2751 sequences. In Table I, the input sequences of the artifact reduction algorithms are all compressed by no loop-filter setting. To avoid denominator MSE value being equal to zero, MSE value of each frame in the video is accumulated for PSNR computing. In Table I, Ours-M is VRGAN that only trained with MSE loss to optimize the networks. Our model only optimized by MSE loss outperforms other methods in PSNR measurement. In addition, GAN-base methods are compared separately in the following experiment results displayed in table II. We trained Ours-M for distortion comparison and VRGAN for perceptual quality comparison. In Table I, the average, min and max PSNR of Ours-M is the highest (38.86dB when Qp=32, 36.10dB when Qp=37). Therefore, the architecture of our generator is more suitable for HEVC artifact reduction than the benchmarks.

Differing from MSE optimized methods, GAN-based methods are suitable for image perceptual quality enhancement. Therefore, we adopt video sequence quality assessment to compare our VRGAN with GAN-based method from Galteri *et al.* in both full-reference and no-reference evaluation. In Table II and Table III, our VRGAN outperforms Galteri's in the both full-reference average PNSR, max PNSR and no-reference VIIDEO [26] video quality assessment.

To evaluate the multi-frame perceptual quality of the generated sequences, we perform multi-frame no-reference assessment experiment. Single frame and recurrent artifact reduction methods are compared in the experiment. Learning-based single image artifact reduction methods are compared with VRGAN to verify the effectiveness of recurrent framework of VRGAN. We use VIIDEO [26] as quality assessment method. VIIDEO score is computed by perceptually relevant models which is based on the distribution of the multi-scale wavelet coefficient of video frame difference. Therefore, the discontinuity of adjacent frames is detected in the experiment. In Table III, our method outperforms other methods and is closer to the score 0.699 that from the VIIDEO measurement of original high-quality videos. The result of the proposed method reaches the best quality in video no-reference measurement compared to other methods. Furthermore, the result shows the effectiveness of recurrent framework for video consistency improvement. Our VRGAN generates more realistic videos than the previous single image artifact reduction methods.

TABLE I
AVERAGE PSNR COMPARISON ON VIMEO90K DATASET.

| Dataset | Setting | HEVC | HEVC-LF [2] | AR-CNN [8] | Ours-M |
|---|---|---|---|---|---|
| Average | Qp=32 | 38.02 | 38.44 | 38.55 | **38.86** |
| | Qp=37 | 35.16 | 35.57 | 35.86 | **36.10** |
| Max | Qp=32 | 52.18 | 52.22 | 51.74 | **52.61** |
| | Qp=37 | 50.94 | 51.10 | 51.23 | **51.89** |
| Min | Qp=32 | 30.62 | 30.87 | 30.79 | **30.96** |
| | Qp=37 | 27.63 | 27.80 | 27.85 | **27.93** |

TABLE II
AVERAGE PSNR COMPARISON ON VIMEO90K DATASET. (GAN-BASED)

| Dataset | Setting | Galteri's [6] | VRGAN |
|---|---|---|---|
| Average | Qp=32 | 35.60 | **37.02** |
| | Qp=37 | 33.98 | **34.17** |
| Max | Qp=32 | 46.72 | **50.08** |
| | Qp=37 | 45.63 | **49.95** |
| Min | Qp=32 | 29.66 | 29.17 |
| | Qp=37 | 27.01 | 26.95 |

TABLE III
AVERAGE VIIDEO SCORE COMPARISON ON VIMEO90K DATASET. (SMALLER BEING BETTER)

| Dataset | Setting | Original | AR-CNN | Galteri's [6] | VRGAN |
|---|---|---|---|---|---|
| Vimeo | Qp=32 | 0.699 | 0.796 | 0.743 | **0.695** |
| | Qp=37 | 0.699 | 0.822 | 0.753 | **0.720** |

## V. CONCLUSION

We presented VRGAN, a generative adversarial method with frame recurrent framework for HEVC video compression artifact reduction. The visual quality of the compressed video has been improved effectively by VRGAN. Full-reference quality assessment suggests that VRGAN produces results that are comparable to the benchmarks. Furthermore, the no-reference quality assessments suggest that VRGAN generates realistic videos.

(a) Original (b) HEVC 37.04/24.60 (c) HEVC-LF [2] 37.51/25.18

(d) AR-CNN [8] 37.78/25.03 (e) Galteri's [6] 35.41/24.48 (f) VRGAN (ours) 36.52/**24.46**

Fig. 5. The figure shows qualitative results comparison at the 7th frames of sequences. The Qp is set to 32. The left side number under the image is PSNR of 7 frames, while the right number is NIQE score of 7th frame (smaller being better). Details are magnified by bicubic interpolation for comparison.

## REFERENCES


[1] Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022 white paper. (2019) [online] Available: https://www.cisco.com/c/en/us/solutions/collateral/serviceprovider/visual-networking-index-vni/white-paper-c11-738429.pdf.

[2] A. Norkin, G. Bjontegaard, A. Fuldseth, M. Narroschke, M. Ikeda, K. Andersson, M. Zhou, and G. Van der Auwera, "HEVC Deblocking Filter," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1746–1754, Dec. 2012.

[3] A. Foi, K. Dabov, V. Katkovnik, and K. Egiazarian, "Shape-adaptive DCT for denoising and image reconstruction," in *Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning*, vol. 6064. International Society for Optics and Photonics, 2006, p. 60640N.

[4] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[6] L. Galteri, L. Seidenari, M. Bertini, and A. D. Bimbo, "Deep Generative Adversarial Compression Artifact Removal," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, Oct. 2017, pp. 4836–4845.

[7] M.-T. Wu, "Wavelet transform based on Meyer algorithm for image edge and blocking artifact reduction," *Information Sciences*, vol. 474, pp. 125–135, Feb. 2019.

[8] C. Dong, Y. Deng, C. C. Loy, and X. Tang, "Compression Artifacts Reduction by a Deep Convolutional Network," in *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, Dec. 2015, pp. 576–584.

[9] Y. Dai, D. Liu, and F. Wu, "A convolutional neural network approach for post-processing in HEVC intra coding," in *International Conference on Multimedia Modeling*. Springer, 2017, pp. 28–39.

[10] G. Lu, W. Ouyang, D. Xu, X. Zhang, Z. Gao, and M.-T. Sun, "Deep Kalman filtering network for video compression artifact reduction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 568–584.

[11] X. Meng, X. Deng, S. Zhu, S. Liu, C. Wang, C. Chen, and B. Zeng, "Mganet: A robust model for quality enhancement of compressed video," *arXiv preprint arXiv:1811.09150*, 2018.

[12] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," *arXiv preprint arXiv:1807.00734*, 2018.

[13] M. S. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6626–6634.

[14] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[16] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 971–980.

[17] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[18] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations (ICLR)*, 2016, pp. 1–13.

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[20] Z. Zhang, X. Wang, and C. Jung, "DCSR: Dilated Convolutions for Single Image Super-Resolution," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1625–1635, 2018.

[21] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[22] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 63–79.

[23] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, pp. 1–20, 2017.

[24] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6228–6237.

[25] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[26] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2015.