Multi-level Wavelet-based Generative Adversarial Network for Perceptual Quality Enhancement of Compressed Video

Jianyi Wang[1 [0000−0001−7025−3626]], Xin Deng[2 [0000−0002−4708−6572]], Mai Xu[1,4⋆ [0000−0002−0277−3301]], Congyong Chen[3 [0000−0003−1682−7435]], and Yuhang Song[5 [0000−0002−7999−0291]]

[1] {School of Electronic and Information Engineering,
[2] School of Cyber Science and Technology,
[3] College of Software} Beihang University, Beijing, China
[4] Hangzhou Innovation Institute, Beihang University, Zhejiang, China
[5] Department of Computer Science, University of Oxford, United Kingdom
{iceclearwjy,cindydeng,maixu,ndsffx304ccy}@buaa.edu.cn,
yuhang.song@some.ox.ac.uk

Abstract. The past few years have witnessed fast development in video quality enhancement via deep learning. Existing methods mainly focus on enhancing the objective quality of compressed video while ignoring its perceptual quality. In this paper, we focus on enhancing the perceptual quality of compressed video. Our main observation is that enhancing the perceptual quality mostly relies on recovering high-frequency sub-bands in wavelet domain. Accordingly, we propose a novel generative adversarial network (GAN) based on multi-level wavelet packet transform (WPT) to enhance the perceptual quality of compressed video, which is called multi-level wavelet-based GAN (MW-GAN). In MW-GAN, we first apply motion compensation with a pyramid architecture to obtain temporal information. Then, we propose a wavelet reconstruction network with wavelet-dense residual blocks (WDRB) to recover the high-frequency details. In addition, the adversarial loss of MW-GAN is added via WPT to further encourage high-frequency details recovery for video frames. Experimental results demonstrate the superiority of our method.

Keywords: Video perceptual quality enhancement · Wavelet packet transform · GAN

1 Introduction

Nowadays, a large amount of videos are available on the Internet, such as YouTube and Facebook, which exerts huge pressure on the communication bandwidth. According to the Cisco Visual Networking Index [7], video causes 75% of Internet traffic in 2017, and this figure is predicted to reach 82% by 2022. As a result, video compression has to be applied to save the communication bandwidth

⋆ Corresponding author: Mai Xu

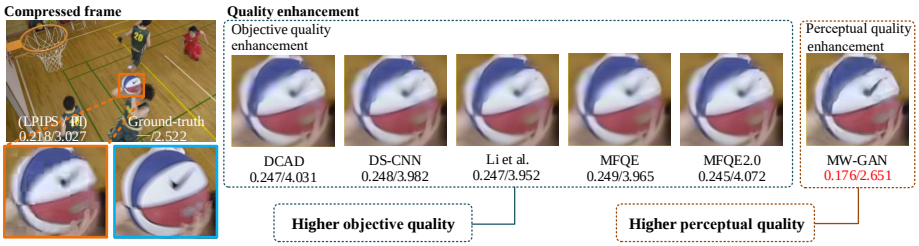

Fig. 1. Objective quality enhancement (traditional works) vs. perceptual quality enhancement(our work). State-of-the-art image [21] and video [41, 34, 42, 12] quality enhancement methods mainly focus on objective quality enhancement, but ignore perceptual quality, which leads to low perceptual quality with high LPIPS [44] and perceptual index (PI) [2]. As the first attempt to enhance the perceptual quality of compressed video, our method achieves better perceptual quality with lower LPIPS and PI. Zoom in for best view.

[32]. However, compressed video inevitably suffers from compression artifacts, which severely degrades the quality of experience (QoE) [22, 1]. Therefore, it is necessary to study on quality enhancement on compressed videos.

Recently, there have been extensive works for enhancing the quality of compressed images and videos [23, 11, 14, 9, 13, 37, 43, 21, 33, 41, 40, 42, 28]. Among them, a four-layer convolutional neural network (CNN) called AR-CNN was proposed in [9] to improve the quality of JPEG compressed images. Then, Zhang *et al.* proposed a denoising CNN (DnCNN) [43] for image denoising and JPEG image deblocking, through a residual learning strategy. Later, a multi-frame quality enhancement network (MFQE) was proposed in [42, 12] to improve the objective quality of compressed video, through leveraging the information from neighboring frames. Most recently, Yang *et al.* [39] have proposed a quality-gated convolutional long short-term memory (QG-ConvLSTM) network, which takes advantage of the bi-directional recurrent structure to fully exploit the useful information in a large range of frames. Unfortunately, the existing methods mainly focus on improving the objective quality of compressed images/videos while ignoring the perceptual quality. Actually, high objective quality, i.e., peak signal-to-noise ratio (PSNR), is not always consistent with the human visual system (HVS) [20]. Besides, according to the perception-distortion tradeoff [3], improving objective quality will inevitably lead to a decrease of perceptual quality. As illustrated in Figure 1, although the frames generated by state-of-the-art methods [21, 41, 34, 42, 12] have high PSNR values, they are not perceptually photorealistic with high LPIPS value [44] and perceptual index (PI) [2], due to the lack of high-frequency details and fine textures.

In this paper, we propose a multi-level wavelet-based generative adversarial network (MW-GAN) for perceptual quality enhancement of compressed video,

which recovers the high-frequency details via wavelet packet transform (WPT) [24] at multiple levels. The key insight to adopt WPT is that the high-frequency details are usually missing due to the compression and they can be regarded as the high-frequency sub-bands after WPT. Our MW-GAN has a generator and a discriminator. Specifically, the generator is composed of two main components: motion compensation and wavelet reconstruction. Motion compensation is first developed to compensate the motion between the target frame and its neighbors with a pyramid architecture following [12]. Then, the wavelet reconstruction network consisting of a bunch of wavelet-dense residual blocks (WDRB) is adopted to reconstruct the sub-bands of the target frame. Besides, WPT is also considered in the discriminator of MW-GAN, such that the generator is further encouraged to recover high-frequency details. As shown in Figure 1, our method is able to generate photorealistic frames with sufficient texture details.

To the best of our knowledge, our work is the first attempt to enhance the perceptual quality of compressed video, using a wavelet-based GAN. The main contributions of this paper are as follows: (1) We investigate that the high-frequency sub-bands in wavelet domain is highly related to the perceptual quality of compressed video. (2) We propose a novel network architecture called MW-GAN, which learns to recover the high-frequency information in wavelet domain for perceptual quality enhancement of compressed video. (3) Extensive experiments have been conducted to demonstrate the ability of the proposed method in enhancing the perceptual quality of compressed video.

2 Related Work

Quality enhancement of compressed images/videos. In the past few years, extensive works [23, 11, 14, 17, 5, 9, 13, 37, 43, 21, 4, 33, 41, 42, 12] have been proposed to enhance the objective quality of compressed images/videos. Specifically, for compressed images, Foi *et al.* [11] applied point-wise shape-adaptiveDCT (SADCT) to reduce the blocking and ringing effects caused by JPEG compression. Then, regression tree fields (RTF) were adopted in [14] to reduce JPEG image blocking effects. Moreover, [17] and [5] utilized sparse coding to remove JPEG artifacts. Recently, deep learning has also been successfully applied for quality enhancement. Particularly, Dong *et al.* [9] proposed a four-layer AR-CNN to reduce the JPEG artifacts of images. Afterwards, $\mathbf{D}^3$ [37] and deep dual-domain convolutional network (DDCN) [13] were proposed for enhancing the quality of JPEG compressed images, utilizing the prior knowledge of JPEG compression. Later, DnCNN was proposed in [43] for multiple tasks of image restoration, including quality enhancement. Li *et al.* [21] proposed a 20-layer CNN for enhancing image quality, which achieves state-of-the-art performance on objective quality enhancement of compressed images.

For the compressed videos, Wang *et al.* [34] proposed a deep CNN-based auto decoder (DCAD) which applies 10 CNN layers to reduce the distortion of compressed video. Later, DS-CNN [41] was proposed for video quality enhancement, in which two sub-networks are used to reduce the artifacts of intra- and

inter-coding frames, respectively. Besides, Yang *et al.* [42] proposed a multi-frame quality enhancement network (MFQE) to take advantage of neighbor high-quality frames. Most recently, a quality-gated convolutional long short-term memory (QG-ConvLSTM) [39] network was proposed to enhance video quality via learning the "forget" and "input" gates in the ConvLSTM [38] cell from quality-related features. Then, Guan *et al.* [12] proposed a new method for multi-frame quality enhancement called MFQE 2.0, which is an extended version of [42] with substantial improvement and achieved state-of-the-art performance. All the above methods aim at minimizing the pixel-wise loss, such as mean square error (MSE), to obtain high objective quality. However, according to [20], MSE cannot always reflect the perceptually relevant differences. Actually, minimizing MSE encourages finding pixel-wise averages of plausible solutions, leading to overly-smooth images/videos with poor perceptual quality [26, 16, 10].

Perception-driven super-resolution. To the best of our knowledge, there exists no work on perceptual quality enhancement of compressed video. The closest work to ours is the perception-driven image/video super-resolution that aims to restore perception-friendly high-resolution images/videos from low-resolution ones. Specifically, Johnson *et al.* [16] proposed to minimize perceptual loss defined in the feature space to enhance the perceptual image quality in single image super-resolution. Then, contextual loss [27] was developed to generate images with natural image statistics via using an objective that focuses on the feature distribution. SRGAN [20] was proposed to generate natural images using perceptual loss and adversarial loss. Sajjadi *et al.* [30] developed a similar method with the local texture matching loss. Later, Wang *et al.* [35] proposed spatial feature transform to effectively incorporate semantic prior in an image and improve the recovered textures. They also developed an enhanced version of SRGAN (ESRGAN) [36] with a deeper and more efficient network architecture. Most recently, TecoGAN [6] was proposed to achieve state-of-the-art performance in video super-resolution, in which a spatial-temporal discriminator and a ping-pong loss are applied. Similar to super-resolution, perceptual quality is also important for quality enhancement of compressed video. To the best of our knowledge, our MW-GAN in this paper is the first attempt in this direction.

3 Motivation for WPT

The wavelet transform allows the multi-resolution analysis of images [15], and it can decompose an image into multiple sub-bands, i.e., low- and high-frequency sub-bands. As verified in [8], the high-frequency sub-bands play an important role in enhancing the perceptual quality of a super-resolved image. Here, we further verify that the high-frequency sub-bands are also crucial for the perceptual quality enhancement of compressed video. Specifically, given a lossless video frame, a series of compressed frames are obtained via HM16.5 [32] at low-delay configuration, with the quantization parameter (QP) of 22, 27, 32 and 37. The corresponding wavelet sub-bands are achieved via WPT using *haar* filter at one level. For each frame, we obtain four sub-bands called LL, LH, H-

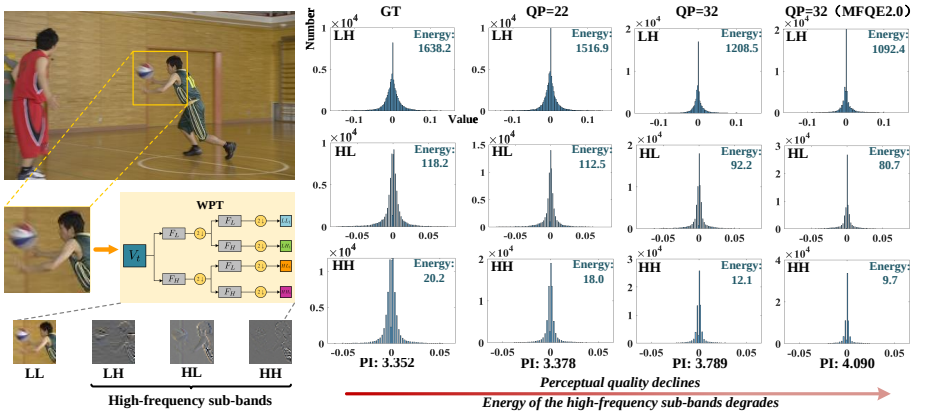

Fig. 2. Histograms of the wavelet sub-bands (zoom in for details). Compressed frames with higher QP have lower perceptual quality (i.e., higher PI) while high-frequency wavelets fade with degraded energy alongside the increase of the QP value. Besides, the state-of-the-art method [12] obtains low perceptual quality, due to the failure of enhancing high-frequency sub-bands. Similar results can be found for other state-of-the-art methods [42, 21, 41, 34].

L and HH. Here, LL is the low-frequency sub-band, LH, HL and HH are the sub-bands with high-frequency information at horizontal, vertical and diagonal directions, respectively. Figure 2 shows the histograms of wavelet coefficients of the LH, HL and HH sub-bands. As we can see, the compressed frames with higher QP values tend to have lower perceptual quality (i.e., higher PI) and insufficient high-frequency wavelet coefficients. Moreover, although the state-of-the-art method [12] can obtain high objective quality, it fails to recover the high-frequency sub-bands in wavelet domain, leading to low perceptual quality. In order to measure the richness of high-frequency wavelet coefficients, we further calculate the wavelet energy by summing up their squared coefficients. As shown in Table 1, the compressed frames with higher QP values tend to have lower wavelet energy in high-frequency sub-bands. All these demonstrate that the high-frequency wavelet sub-bands play an important role in enhancing the perceptual quality of compressed video.

Based on the above investigation, we propose our MW-GAN as follows. (1) Since high-frequency wavelet sub-bands are crucial for the perceptual quality, the generator of our MW-GAN directly outputs these wavelet sub-bands. (2) To restore the high-frequency wavelet sub-bands as much as possible, a wavelet-dense residual block (WDRB) is proposed in our MW-GAN to further recover the wavelet sub-bands. (3) The WPT is also applied in the discriminator of our MW-GAN for perceptual quality enhancement by encouraging the generated results to be indistinguishable from the ground-truth in wavelet domain.

Table 1. Wavelet energy of the high-frequency sub-bands (LH, HL, HH) per frame across the MFQE2.0 dataset [12]. Frames with higher QP values tend to have lower wavelet energy in high-frequency sub-bands.

Wavelet	Compressed frame				MFQE2.0 [12]				Ground-
sub-bands	QP=22	QP=27	QP=32	QP=37	QP=22	QP=27	QP=32	QP=37	truth
LH	1106.8	1043.3	956.1	835.4	872.0	844.2	809.9	715.9	1189.7
HL	1189.3	1130.2	1048.1	940.5	994.3	973.9	937.8	872.1	1273.0
HH	138.4	115.9	93.8	70.7	130.9	110.4	92.6	65.3	176.2

4 The Proposed MW-GAN

The architecture of the proposed MW-GAN is shown in Figure 3. To enhance the perceptual quality of a video frame $\mathbf{V}_t$, we simultaneously train a generator $G_{\boldsymbol{\theta}_G}$ parameterized by $\boldsymbol{\theta}_G$ and a discriminator $D_{\boldsymbol{\theta}_D}$ parameterized by $\boldsymbol{\theta}_D$ in an adversarial manner. The ultimate goal is to obtain the enhanced frame $\hat{\mathbf{O}}_t$ with high perceptual quality, similar to its ground-truth $\mathbf{O}_t$. To take advantage of the temporal information in adjacent frames, the generator $G_{\boldsymbol{\theta}_G}$ takes both $\mathbf{V}_t$ and its neighbor frames $\{\mathbf{V}_{t\pm n}\}_{n=1}^{N}$ as inputs. The details about the generator $G_{\boldsymbol{\theta}_G}$ are introduced in Section 4.1. To further encourage high perceptual quality, we propose a multi-level wavelet-based discriminator, which distinguishes the generated wavelet sub-bands from the ground-truths, as introduced in Section 4.2. Finally, we present the loss functions to train the MW-GAN in Section 4.3.

4.1 Multi-level wavelet-based generator

The generator $G_{\boldsymbol{\theta}_G}$ of our MW-GAN is mainly composed of two parts: motion compensation and wavelet reconstruction. Given the $2N+1$ frames as inputs, we first perform motion compensation across frames to align content across frames. After that, wavelet reconstruction is applied to reconstruct the target wavelet sub-bands using the information of sub-bands from both current and compensated frames. Finally, the output reconstructed sub-bands are used to obtain the enhanced frame through IWPT.

Motion compensation: To take advantage of the temporal information across frames, we adopt the motion compensation network with a pyramid architecture following [42]. To obtain better performance, we further make some modifications. First, for flow estimation, it is crucial to have large-size filters in the first few layers of the network to capture intense motion. Thus, the kernel sizes of the first two convolutional layers at each pyramid level are 7×7 and 5×5, respectively, instead of 3×3 in [42]. Second, we replace the pooling operation at each pyramid level by WPT. Since WPT is invertible, this pooling operation is able to make all information of the original frames be preserved, meanwhile efficiently enlarging the receptive field for flow estimation.

Wavelet reconstruction: Given the current and compensated frames, we further propose a wavelet reconstruction network to reconstruct the wavelet sub-

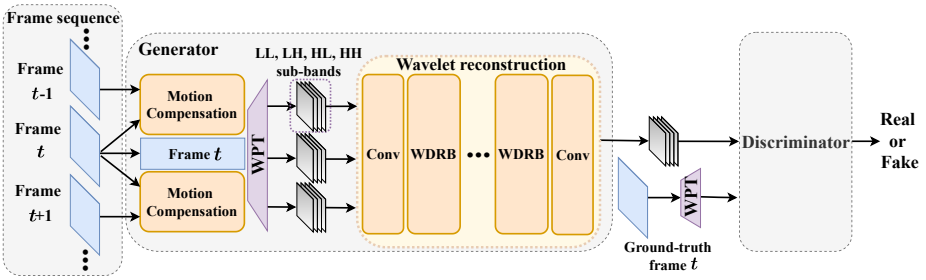

Fig. 3. Framework of our MW-GAN. Perceptual quality is enhanced via recovering wavelet sub-bands in wavelet domain. Specifically, we propose a generator with motion composition and WDRB to capture both temporal and high-frequency information. Then a multi-level wavelet-based discriminator is proposed to evaluate generated results in both pixel and wavelet domain.

bands $\hat{\mathbf{S}}_t$, which contains a cascade of wavelet-dense residual blocks (WDRB) . Specifically, WPT is first applied to generate sub-bands as inputs. Then, several convolutional layers are applied to extract corresponding feature maps. To further capture high-frequency details and reduce the computational cost, we develop the wavelet-based residual block (WDRB) with a residual-in-residual structure, as shown in Figure 4. Similar to [36], in our WDRB, several dense blocks are applied in the main path of the residual block to enhance the feature representation with high capacity. In addition, we adopt WPT and IWPT in the main path to learn the residual in wavelet domain. With this simple yet effective design, the receptive field is further enlarged without losing information, thus enabling the network to capture more high-frequency details. The efficiency of the WDRB is also investigated in the ablation study. The final output of $G_{\boldsymbol{\theta}_G}$ are the wavelet sub-bands $\hat{\mathbf{S}}_t$, and the enhanced frame $\hat{\mathbf{O}}_t$ can be obtained by combining $\hat{\mathbf{S}}_t$ via IWPT. Assuming that $\omega^{-1}(\cdot)$ denotes the IWPT operation, the output of the generator can be summarized as follows,

$$\begin{aligned} \hat{\mathbf{S}}_t &= G_{\boldsymbol{\theta}_G}(\mathbf{V}_t, \{\mathbf{V}_{t\pm n}\}_{n=1}^{N}), \\ \hat{\mathbf{O}}_t &= \omega^{-1}(\hat{\mathbf{S}}_t). \end{aligned} \tag{1}$$

4.2 Multi-level wavelet-based discriminator

In this section, a multi-level wavelet-based discriminator is proposed to encourage the generated results indistinguishable from the ground-truth. The structure of our discriminator is illustrated in Figure 4. Specifically, the discriminator takes the generated frames (obtained via IWPT) or the real ones (ground-truth) as input. Then, WPT is applied to both generated and real frames for obtaining their wavelet sub-bands at multiple levels. After that, wavelet sub-bands of different

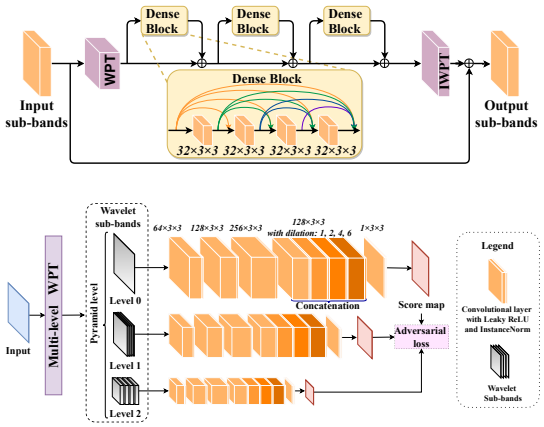

Fig. 4. Details about our framework. Top: Architecture of the proposed WDRB; Down: Architecture of the multi-level wavelet-based discriminator in MW-GAN.

levels are fed into the corresponding pyramid level. Each pyramid level employs a fully convolutional architecture, in which a group of convolutional layers with different dilations are applied to extract features with various receptive fields. Subsequently, the input wavelet sub-bands at each level are mapped into a single channel of the score map with the same size as the input, which measures the similarity between the generated and real frames. Let $\omega_l(\cdot)$ denote the l-th level WPT. The output of the discriminator is a set of score maps of different sizes, which can be formulated as follows,

$$D^l_{\boldsymbol{\theta}_D}(\mathbf{O}_t) = f_l(\omega_l(\mathbf{O}_t)), \tag{2}$$

where $f_l(\cdot)$ is the corresponding map function at the l-th level of the discriminator, L is the total number of the pyramid levels and $D^l_{\boldsymbol{\theta}_D}(\mathbf{O}_t)$ is the output score map at l-th level. Note that the $l = 0$ indicates that there is no WPT operation.

The advantage of the proposed multi-level wavelet-based discriminator is efficiency and efficacy. (1) The computational cost can be significantly reduced by adopting WPT, since the calculation of wavelet sub-bands can be regarded as forwarding through a single convolutional layer. In contrast, the traditional methods need to extract multi-scale features via a sequence of convolutional layers, thus consuming expensive computational complexity. (2) Our discriminator can distinguish the generated and real frames in both pixel ($l = 0$) and wavelet domains ($l > 0$), leading to better generative results.

4.3 Loss functions

In this section, we propose a loss function for training the MW-GAN. Specifically, we formulate the loss of the generator $L_G(\boldsymbol{\theta}_M; \boldsymbol{\theta}_G)$ as the weighted sum of a

motion loss $L_M(\boldsymbol{\theta}_M)$, a wavelet loss $L_W(\boldsymbol{\theta}_G)$ and an adversarial loss $L_{\text{Adv}}(\boldsymbol{\theta}_G)$. Then, $L_G(\boldsymbol{\theta}_M; \boldsymbol{\theta}_G)$ can be written as,

$$L_G(\boldsymbol{\theta}_M; \boldsymbol{\theta}_G) = L_W(\boldsymbol{\theta}_G) + \alpha L_M(\boldsymbol{\theta}_M) + \beta L_{\text{Adv}}(\boldsymbol{\theta}_G), \tag{3}$$

where α and β are the coefficients to balance different loss terms. In the following, we explain the different components of the loss functions.

Motion loss: Since it is hard to obtain the ground-truth of optical flow, the motion compensation network $M_{\boldsymbol{\theta}_M}$ parameterized by $\boldsymbol{\theta}_M$ is directly trained under the supervision of the ground-truth $\mathbf{O}_t$. That is, the neighboring frames $\{\mathbf{V}_{t\pm n}\}_{n=1}^N$ are first wrapped using the optical flow estimated by $M_{\boldsymbol{\theta}_M}$; then the loss between the compensated frames $\{M_{\boldsymbol{\theta}_M}(\mathbf{V}_{t\pm n})\}_{n=1}^N$ and the ground-truth $\mathbf{O}_t$ is minimized. Here, we adopt Charbonnier penalty function [19] as the motion loss, defined by:

$$L_M(\boldsymbol{\theta}_M) = \frac{1}{2N}\sum_{n=1}^{N}\sqrt{\|M_{\boldsymbol{\theta}_M}(\mathbf{V}_{t\pm n}) - \mathbf{O}_t\|_F^2 + \epsilon_m^2}, \tag{4}$$

where ϵ_m is a scaling parameter and is empirically set to 1×10^{-3}.

Wavelet loss: The wavelet loss function models how close the predicted sub-bands $\hat{\boldsymbol{S}}_t$ are to the ground truth $\boldsymbol{S}_t$. It is defined by a weighted Charbonnier penalty function in wavelet domain as follows:

$$L_W(\boldsymbol{\theta}_G) = \sqrt{\left\|\mathbf{W}^{1/2} \odot (\hat{\mathbf{S}}_t - \mathbf{S}_t)\right\|_F^2 + \epsilon_w^2}, \tag{5}$$

where $\odot$ represents dot product, $\hat{\mathbf{S}}_t = \{\hat{\mathbf{s}}_t^1, \hat{\mathbf{s}}_t^2, \ldots, \hat{\mathbf{s}}_t^{n_w}\}$ are the predicted sub-bands with the number of n_w in total and $\mathbf{S}_t$ are their corresponding ground-truths. In addition, $\mathbf{W} = \{w_1, w_2, \ldots, w_{n_w}\}$ is the weight matrix to balance the importance of each sub-band and ϵ_w is a scaling parameter set to 1×10^{-3}.

Adversarial loss: In addition to the above two loss functions, another important loss of the GAN is the adversarial loss. Note that we follow [25] to adopt ℓ_2 loss for training. Recall that $D^l_{\boldsymbol{\theta}_D}(\mathbf{O}_t)$ represents the outputs of the discriminator in (2) at the l-th level, taking $\mathbf{O}_t$ as input. Then, the discriminator loss can be defined as:

$$L_D(\boldsymbol{\theta}_D) = \frac{1}{2L}\mathbb{E}_{\mathbf{O}_t}[\sum_{l=0}^{L-1}\|D^l_{\boldsymbol{\theta}_D}(\mathbf{O}_t) - \mathbf{1}\|_F^2] + \frac{1}{2L}\mathbb{E}_{\hat{\mathbf{O}}_t}[\sum_{l=0}^{L-1}\|D^l_{\boldsymbol{\theta}_D}(\hat{\mathbf{O}}_t)\|_F^2], \tag{6}$$

where $\mathbb{E}_{\hat{\mathbf{O}}_t}[\cdot]$ represents the average for all $\hat{O}_t$ in the mini-batch and L is the number of the pyramid levels in the discriminator. Note that $\mathbf{1} \in \mathcal{R}^{W_l \times H_l}$, where W_l and H_l are the width and height of the score map at the l-th level, respectively. Symmetrically, the adversarial loss for generator is as follow:

$$L_{\text{Adv}}(\boldsymbol{\theta}_G) = \frac{1}{2L}\mathbb{E}_{\hat{\mathbf{O}}_t}[\sum_{l=0}^{L-1}\|D^l_{\boldsymbol{\theta}_D}(\hat{\mathbf{O}}_t(\boldsymbol{\theta}_G)) - \mathbf{1}\|_F^2]. \tag{7}$$

5 Experiments

In this section, the experimental results are presented to validate the effectiveness of our MW-GAN method. Section 5.1 introduces the experimental settings. Section 5.2 compares the results between our method and other state-of-the-art methods over the test sequences of JCT-VC [29]. In Section 5.3, the mean opinion score (MOS) test is performed to compare the subjective quality of videos by different methods. Finally, the ablation study is presented in Section 5.4.

5.1 Settings

Datasets: We train the MW-GAN model using the database introduced in [12]. Following [12], except the 18 common test sequences of Joint Collaborative Team on Video Coding (JCT-VC) [29], the other 142 sequences are randomly divided into non-overlapping training set (106 sequences) and validation set (36 sequences). All 160 sequences are compressed by HM16.5 [32] under Low-Delay configuration, setting QP to 32. Note that different from existing methods [21, 4, 33, 41, 42, 39, 12] that train models for each QP individually, we only train our MW-GAN under QP= 32, and we test on both QP= 32 and QP= 37. The results show the generalization ability of our method.

Parameter settings: Here, we mainly introduce the settings and hyperparameters of our experiments. Specifically, the total number of levels L of each component in MW-GAN is set to 3 and the generator takes 3 consecutive frames as inputs for a better tradeoff between performance and efficiency. The training process is divided into two stages. We first train our model without adversarial loss (i.e., $\beta = 0$). The motion loss weight α is initialed as 10 and decayed by a factor of 10 every 5×10^4 iterations to encourage the learning of motion compensation first. The iteration number and mini-batch size are 3×10^5 and 32, respectively, and the input frames are cropped to 256×256. The Adam algorithm [18] with the step size of 2.5×10^{-4} is adopted; the learning rate is initialized as 2×10^{-4} and decayed by a factor of 2 every 1×10^5 iterations. In the second stage, the pre-trained model is employed as an initialization for the generator. The generator is trained using the loss function in (3) with $\alpha = 1 \times 10^{-2}$ and $\beta = 5 \times 10^{-3}$. The iteration number and mini-batch size are 6×10^5 and 32, respectively, and the cropped size is reduced to 128×128. The learning rate is set to 1×10^{-4} and halved every 1×10^5 iterations. Note that the above hyperparameteres are tuned over the training set. We apply the Adam algorithm [18] and alternately update the generator and discriminator network until convergence. The wavelet sub-bands are obtained by wavelet packet decomposition with *haar* filter. More details can be found in the supplementary material.

5.2 Quantatative comparison

In this section, we evaluate the performance of our method in terms of learned perceptual image patch similarity (LPIPS) [44] and perceptual index (PI) [2], which are metrics widely used for perceptual quality evaluation. For results in

Table 2. Overall ΔLPIPS and ΔPI between enhanced and compressed frames on the test set of JCT-VC [29] at QP= 32 and QP= 37. Our MW-GAN achieves the best perceptual quality across all the test sequences.

QP		Video sequence	Li *et al.* [21]		DCAD [34]		DS-CNN [41]		MFQE [42]		MFQE 2.0 [12]		Ours	
			ΔLPIPS	ΔPI	ΔLPIPS	ΔPI	ΔLPIPS	ΔPI	ΔLPIPS	ΔPI	ΔLPIPS	ΔPI	ΔLPIPS	ΔPI
32	A	*Traffic*	0.021	0.501	0.019	0.419	0.017	0.373	0.016	0.441	0.014	0.430	**-0.032**	**-1.123**
		PeopleOnStreet	0.020	0.865	0.020	0.668	0.019	0.663	0.017	0.807	0.017	0.794	**-0.020**	**-0.558**
	B	*Kimono*	0.038	0.479	0.034	0.403	0.033	0.332	0.034	0.440	0.036	0.443	**-0.069**	**-1.561**
		ParkScene	0.010	0.299	0.010	0.377	0.009	0.291	0.012	0.239	0.010	0.346	**-0.032**	**-0.159**
		Cactus	0.032	0.264	0.028	0.264	0.027	0.236	0.028	0.248	0.028	0.274	**-0.109**	**-0.953**
		BQTerrace	0.028	0.384	0.025	0.413	0.025	0.362	0.026	0.418	0.026	0.447	**-0.099**	**-0.326**
		BasketballDrive	0.036	0.534	0.031	0.501	0.032	0.462	0.032	0.595	0.032	0.618	**-0.106**	**-0.921**
	C	*RaceHorses*	0.023	0.504	0.022	0.536	0.021	0.469	0.025	0.489	0.027	0.566	**-0.021**	**-0.579**
		BQMall	0.022	0.457	0.019	0.507	0.018	0.426	0.021	0.467	0.021	0.499	**-0.033**	**-0.908**
		PartyScene	0.032	0.436	0.027	0.381	0.024	0.347	0.025	0.408	0.025	0.412	**-0.075**	**-0.765**
		BasketballDrill	0.027	0.782	0.023	0.926	0.026	0.842	0.027	0.772	0.025	0.929	**-0.047**	**-0.532**
	D	*RaceHorses*	0.019	0.628	0.018	0.657	0.017	0.581	0.021	0.657	0.021	0.685	**-0.005**	**-0.199**
		BQSquare	0.012	0.285	0.012	0.599	0.011	0.382	0.013	0.344	0.011	0.570	**-0.037**	**-0.525**
		BlowingBubbles	0.013	0.492	0.014	0.585	0.011	0.458	0.012	0.394	0.009	0.472	**-0.039**	**-0.241**
		BasketballPass	0.020	0.537	0.016	0.564	0.016	0.517	0.020	0.574	0.019	0.594	**-0.021**	**-0.794**
	E	*FourPeople*	0.012	0.265	0.010	0.231	0.010	0.234	0.010	0.238	0.008	0.219	**-0.040**	**-1.129**
		Johnny	0.010	0.308	0.010	0.339	0.013	0.215	0.012	0.363	0.010	0.259	**-0.065**	**-1.171**
		KristenAndSara	0.016	0.278	0.015	0.316	0.016	0.360	0.015	0.338	0.014	0.352	**-0.026**	**-1.305**
		Average	0.022	0.461	0.020	0.483	0.019	0.419	0.020	0.457	0.020	0.495	**-0.049**	**-0.764**
37		Average	0.027	0.541	0.026	0.621	0.024	0.562	0.027	0.617	0.024	0.614	**-0.046**	**-0.993**

terms of other evaluation metrics, please see the supplementary material. We compare our method with Li *et al.* [21], DS-CNN [41], DCAD [34], MFQE [42] and MFQE 2.0 [12]. Among them, Li *et al.* is the latest quality enhancement methods for compressed images. DCAD and DS-CNN are single-frame enhancement methods for compressed videos, while MFQE and MFQE 2.0 are state-of-the-art multi-frame quality enhancement methods for compressed videos. All these methods are trained over the same training set as ours for fair comparison.

Table 2 reports the ΔLPIPS and ΔPI results which are calculated between enhanced and compressed frames averaged over each test sequence. Note that $\Delta\text{LPIPS} < 0$ and $\Delta\text{PI} < 0$ indicate improvement in perceptual quality. As shown in this table, our MW-GAN outperforms all 6 compared methods in terms of perceptual quality, for all test sequences, while all the compared methods fail to enhance the perceptual quality of compressed video. Specifically, at QP= 32, the average ΔLPIPS and ΔPI in our MW-GAN are -0.049 and -0.764, respectively, while other methods [43, 21, 41, 34, 42, 12] all have positive ΔLPIPS and ΔPI values. Furthermore, our MW-GAN also gains the largest decrease of ΔLPIPS with -0.046 and ΔPI with -0.993 at QP= 37, verifying the generalization ability of our MW-GAN for perceptual quality enhancement of compressed video.

5.3 Subjective Comparison

In this section, we mainly focus on the subjective evaluation of our method. Figure 5 visualizes the enhanced video frames of *Traffic* at QP= 32, *BasketballPass* at QP= 32, *BQTerrace* at QP = 37 and *RaceHorses* at QP= 32. We can observe

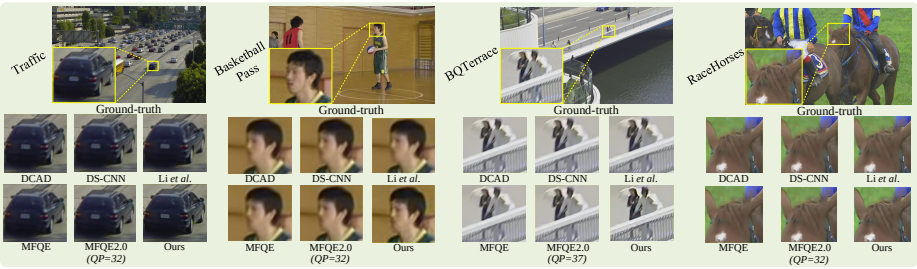

Fig. 5. Qualitative comparison on the test sequences of JCT-VC [29]. Our MW-GAN can generate much more realistic results (Zoom in for best view).

that our MW-GAN method outperforms other methods with sharper edges and more vivid details. For instance, the car in *Traffic*, the face in *BasketballPass*, the pedestrians in *BQTerrace* and the horse in *RaceHorses* can be finely restored in our MW-GAN, while the existing PSNR-oriented methods generate blurry results with low perceptual quality.

To further evaluate the subjective quality of our method, we also conduct a mean opinion score (MOS) test at QP= 37. Specifically, we asked 15 subjects to rate an integral score from 1 to 100 following [31] on the enhanced videos. The higher score indicates better perceptual quality. The subjects are required to rate the scores for compressed video sequences, raw sequences, sequences enhanced by Li *et al.* [21], MFQE [42], MFQE 2.0 [12], our MW-GAN method. Table 3 presents ΔMOS which is calculated between enhanced and compressed sequences. As can be seen, our MW-GAN method obtains the highest ΔMOS score for each class of sequences, and it obtains an average increase of 4.55 in terms of MOS, considerably better than other methods.

Table 3. ΔMOS calculated between enhanced and compressed sequences at QP = 37 over the test sequences of JCT-VC [29]. Our MW-GAN obtains the highest ΔMOS score for each class of sequences.

Sequence	Li *et al.* [21]	MFQE [42]	MFQE 2.0 [12]	Ours
A	-0.47	-0.83	-1.21	**4.74**
B	1.86	1.42	0.82	**3.95**
C	-4.76	-2.74	-5.07	**4.05**
D	-1.08	-0.30	-1.90	**5.15**
E	1.11	1.04	0.70	**4.86**
Average	-0.69	-0.28	-1.33	**4.55**

5.4 Ablation study

In this section, we mainly analyze the effectiveness of WPT applied in each component of our MW-GAN.

Effectiveness of WDRB in generator: In our generator, the multi-level WPT is achieved via adopting WDRB. Thus, it is essential to validate the effectiveness of the proposed WDRB. Here, we conduct the test with two strategies. (1) MW-GAN-RRDB: We directly replace the WDRB by the residual-in-residual dense block (RRDB) proposed in [36]. Note that the channels of the first convolutional layer before the first RRDB are increased to keep the structure of RRDB unchanged. (2) MW-GAN-CNN: We directly replace the WPT and IWPT in WDRB by average pooling and upsampling layers. Note that the above two strategies all lead to parameter increasing. Figure 6 shows the results of the above two ablation strategies over different classes of test sequences at QP= 32. We can see from this figure that WDRB has a positive impact on the perceptual quality, leading to 0.170 and 0.126 decrease in ΔPI compared with MW-GAN-CNN and MW-GAN-RRDB, respectively. Therefore, we can conclude that WDRB plays an effective role in MW-GAN.

Effectiveness of WPT in the discriminator: To evaluate the effectiveness of the multi-level wavelet-based discriminator, we add a baseline without WPT in the discriminator, i.e., MW-GAN-D w/o WPT. Specifically, we replace WPT in the discriminator with an average pooling layer that extracts features with the same size as wavelet sub-bands and keeps other components unchanged for a fair comparison. As Figure 6 shows, without WPT in the discriminator, the ΔPI score increases by average 0.192 over the whole test set, which indicates the effectiveness of the proposed multi-level wavelet-based discriminator.

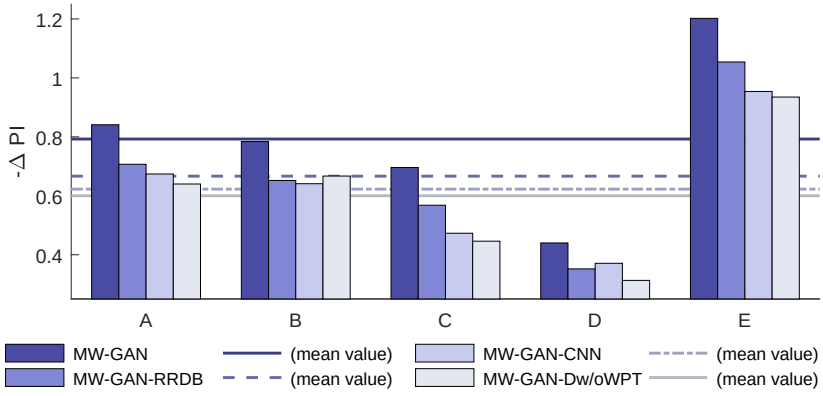

Fig. 6. Comparison between our MW-GAN and corresponding baselines on the test sequences of JCT-VC [29] according to $-\Delta$PI between enhanced and compressed sequences. Effectiveness of the proposed MW-GAN can be verified.

6 Generalization ability

In this section, we mainly focus on the generalization ability of the proposed MW-GAN. For more details, please see our supplementary material.

Transfer to H.264. To further verify the generalization capability of our MW-GAN, we also test on the video sequences compressed by other standards. Specifically, we compress the test sequences of JCT-VC with H.264 (the JM encoder with the low-delay P configuration) at QP= 32 and QP= 37. Then, we directly test the performance of our method over the above test sequences without fine-tuning. Consequently, the average PI decrease of test sequences is 0.548 at QP= 32 and 0.527 at QP=37, which implying the generalization ability of our MW-GAN approach across different compression standards.

Performance on other sequences. In addition to the common-used test sequences of JCT-VC, we also test the performance of our and other methods over the test set in [42], which is different from the test sequences of JCT-VC. Results show that our method has 0.039 LPIPS decrease and 1.058 PI decrease at QP= 32, while other methods all increase the PI value, which again indicates the superiority of our method according to the perceptual quality enhancement.

Perception-distortion tradeoff. Our work mainly focuses on enhancing perceptual quality of compressed video. However, according to the perception-distortion tradeoff [3], improving perceptual quality fully will inevitably lead to a decrease of PSNR. We further evaluate the objective quality of the frames generated by our method over the test sequences of JCT-VC. Results show that our method leads to 1.040 dB PSNR decrease at QP= 32 and 0.651 dB PSNR decrease at QP= 37. It is promising future work for perception-distortion tradeoff of compressed video quality enhancement, while at the current stage we mainly focus on perceptual quality enhancement (w/o considering PSNR) as the first attempt in this direction.

7 Conclusion

In this paper, we have proposed MW-GAN as the first attempt for perceptual quality enhancement of compressed video, which embeds WPT in both generator and discriminator for recovering high-frequency details. First, we find from the data analysis that perceptual quality is highly related to the high-frequency sub-bands in wavelet domain. Second, we design a wavelet-based GAN to recover high-frequency details in wavelet domain. Specifically, a WDRB is proposed in our generator for wavelet sub-band reconstruction and a multi-level wavelet-based discriminator is applied to further encourage high-frequency recovery. Finally, the ablation experiments showed the effectiveness of each component of MW-GAN. More importantly, both quantitative and qualitative results in extensive experiments verified that our MW-GAN outperforms other state-of-the-art methods according to perceptual quality enhancement of compressed video.

Acknowledgement This work was supported by the NSFC under Project 61876013, Project 61922009, and Project 61573037.

References

1. Bampis, C.G., Li, Z., Moorthy, A.K., Katsavounidis, I., Aaron, A., Bovik, A.C.: Study of temporal effects on subjective video quality of experience. IEEE Transactions on Image Processing (TIP) **26**(11), 5217–5231 (2017)
2. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: The 2018 pirm challenge on perceptual image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
3. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6228–6237 (2018)
4. Cavigelli, L., Hager, P., Benini, L.: Cas-cnn: A deep convolutional neural network for image compression artifact suppression. In: 2017 International Joint Conference on Neural Networks (IJCNN). pp. 752–759. IEEE (2017)
5. Chang, H., Ng, M.K., Zeng, T.: Reducing artifacts in jpeg decompression via a learned dictionary. IEEE Transactions on Signal Processing **62**(3), 718–728 (2013)
6. Chu, M., Xie, Y., Leal-Taixé, L., Thuerey, N.: Temporally coherent gans for video super-resolution (tecogan). arXiv preprint arXiv:1811.09393 (2018)
7. CVNI: Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021 white paper. In: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html (2017)
8. Deng, X., Yang, R., Xu, M., Dragotti, P.L.: Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision (CVPR). pp. 3076–3085 (2019)
9. Dong, C., Deng, Y., Change Loy, C., Tang, X.: Compression artifacts reduction by a deep convolutional network. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 576–584 (2015)
10. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 658–666 (2016)
11. Foi, A., Katkovnik, V., Egiazarian, K.: Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. IEEE Transactions on Image Processing (TIP) **16**(5), 1395–1411 (2007)
12. Guan, Z., Xing, Q., Xu, M., Yang, R., Liu, T., Wang, Z.: Mfqe 2.0: A new approach for multi-frame quality enhancement on compressed video. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) pp. 1–1 (2019)
13. Guo, J., Chao, H.: Building dual-domain representations for compression artifacts reduction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 628–644. Springer (2016)
14. Jancsary, J., Nowozin, S., Rother, C.: Loss-specific training of non-parametric image restoration models: A new state of the art. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 112–125. Springer (2012)
15. Jawerth, B., Sweldens, W.: An overview of wavelet based multiresolution analyses. SIAM review **36**(3), 377–412 (1994)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European conference on computer vision (ECCV). pp. 694–711. Springer (2016)

17. Jung, C., Jiao, L., Qi, H., Sun, T.: Image deblocking via sparse representation. Signal Processing: Image Communication **27**(6), 663–677 (2012)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR) (2015)
19. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 624–632 (2017)
20. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4681–4690 (2017)
21. Li, K., Bare, B., Yan, B.: An efficient deep convolutional neural networks model for compressed image deblocking. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 1320–1325. IEEE (2017)
22. Li, S., Xu, M., Deng, X., Wang, Z.: Weight-based r-λ rate control for perceptual hevc coding on conversational videos. Signal Processing: Image Communication **38**, 127–140 (2015)
23. Liew, A.C., Yan, H.: Blocking artifacts suppression in block-coded images using overcomplete wavelet representation. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) **14**(4), 450–461 (2004)
24. Mallat, S.: A wavelet tour of signal processing. Elsevier (1999)
25. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (CVPR). pp. 2794–2802 (2017)
26. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440 (2015)
27. Mechrez, R., Talmi, I., Shama, F., Zelnik-Manor, L.: Maintaining natural image statistics with the contextual loss. In: Asian Conference on Computer Vision (ACCV). pp. 427–443. Springer (2018)
28. Meng, X., Deng, X., Zhu, S., Liu, S., Wang, C., Chen, C., Zeng, B.: Mganet: A robust model for quality enhancement of compressed video. arXiv preprint arXiv:1811.09150 (2018)
29. Ohm, J.R., Sullivan, G.J., Schwarz, H., Tan, T.K., Wiegand, T.: Comparison of the coding efficiency of video coding standardsincluding high efficiency video coding (hevc). IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) **22**(12), 1669–1684 (2012)
30. Sajjadi, M.S., Scholkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4491–4500 (2017)
31. Seshadrinathan, K., Soundararajan, R., Bovik, A.C., Cormack, L.K.: Study of subjective and objective quality assessment of video. IEEE Transactions on Image Processing (TIP) **19**(6), 1427–1441 (2010)
32. Sullivan, G.J., Ohm, J.R., Han, W.J., Wiegand, T.: Overview of the high efficiency video coding (hevc) standard. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) **22**(12), 1649–1668 (2012)
33. Tai, Y., Yang, J., Liu, X., Xu, C.: Memnet: A persistent memory network for image restoration. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4539–4547 (2017)

34. Wang, T., Chen, M., Chao, H.: A novel deep learning-based method of improving coding efficiency from the decoder-end for hevc. In: 2017 Data Compression Conference (DCC). pp. 410–419. IEEE (2017)
35. Wang, X., Yu, K., Dong, C., Change Loy, C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 606–615 (2018)
36. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
37. Wang, Z., Liu, D., Chang, S., Ling, Q., Yang, Y., Huang, T.S.: D3: Deep dual-domain based fast restoration of jpeg-compressed images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2764–2772 (2016)
38. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems (NIPS). pp. 802–810 (2015)
39. Yang, R., Sun, X., Xu, M., Zeng, W.: Quality-gated convolutional lstm for enhancing compressed video. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). pp. 532–537. IEEE (2019)
40. Yang, R., Xu, M., Liu, T., Wang, Z., Guan, Z.: Enhancing quality for hevc compressed videos. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) (2018)
41. Yang, R., Xu, M., Wang, Z.: Decoder-side hevc quality enhancement with scalable convolutional neural network. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 817–822. IEEE (2017)
42. Yang, R., Xu, M., Wang, Z., Li, T.: Multi-frame quality enhancement for compressed video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6664–6673 (2018)
43. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Transactions on Image Processing (TIP) **26**(7), 3142–3155 (2017)
44. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018)