Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Neural texture transfer assisted video coding with adaptive up-sampling

Li Yu [a,b], Wenshuai Chang [a,b], Weize Quan [c,d], Jimin Xiao [e], Dong-Ming Yan [c,d], Moncef Gabbouj [a,f,*]

[a] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China
[b] Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China
[c] National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing 100049, China
[d] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[e] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215028, PR China
[f] Department of Computing Sciences, Tampere University, Tampere, Finland

ARTICLE INFO

Keywords:
High-efficiency video coding (HEVC)
Reference-based super-resolution
Low bitrate
Video compression
Deep learning
Machine learning

ABSTRACT

Deep learning techniques have been extensively investigated for the purpose of further increasing the efficiency of traditional video compression. Some deep learning techniques for down/up-sampling-based video coding were found to be especially effective when the bandwidth or storage is limited. Existing works mainly differ in the super-resolution models used. Some works simply use a single image super-resolution model, ignoring the rich information in the correlation between video frames, while others explore the correlation between frames by simply concatenating the features across adjacent frames. This, however, may fail when the textures are not well aligned. In this paper, we propose to utilize neural texture transfer which exploits the semantic correlation between frames and is able to explore the correlated information even when the textures are not aligned. Meanwhile, an adaptive group of pictures (GOP) method is proposed to automatically decide whether a frame should be down-sampled or not. Experimental results show that the proposed method outperforms the standard HEVC and state-of-the-art methods under different compression configurations. When compared to standard HEVC, the BD-rate (PSNR) and BD-rate (SSIM) of the proposed method are up to -19.1% and -26.5%, respectively.

1. Introduction

According to recent statistics, video traffic will account for about 82% of all Internet traffic by 2022 [1]. Video has become one of the major ways for information transmission and communication. At the same time, new video types, including Ultra-High Definition (UHD), Virtual Reality (VR), Wide Color Gamut (WCG), and High Dynamic Range (HDR), are emerging. These new video types provide a better user experience but at the cost of dramatically increased data volumes. Meanwhile, the number of video cameras in use keeps boosting in recent years, such as surveillance cameras, laptops, and smartphones. Consequently, the total amount of global video data doubles every two years, which has been the bottleneck for data processing, storage, and transmission [2]. Therefore, more advanced video compression techniques are of vital importance, which will support more efficient storage and transmission of videos.

During the last three decades, the development of traditional statistical video compression methods [3–7] has somewhat saturated and most recent endeavors turned to deep learning models [8–10], which have proved their capacity to discover knowledge from unstructured massive data and provide data-driven predictions. Deep learning has the potential to provide new opportunities for further upgrading video coding technologies. Inspired by the recent advances in deep learning, many works have been proposed to leverage deep learning in video compression and achieve significant improvements [11,12]. Among these, Convolution Neural Network (CNN) based video enhancement was proposed as a post-processing procedure at the decoder to improve the perceptual quality of the reconstructed video [13–15]. To further increase the compression ratio, some works propose to down-sample the video prior to encoding and up-sample the decoded video using CNN-based video super-resolution model, which is known as down/up-sampling-based coding [16–18]. As reported in [19,20], the coding of low-resolution video can perform both subjectively and objectively better than the direct coding of full-resolution version at low bit rates.

Prior work in down/up-sampling-based coding is mainly inspired by CNN-based single image super-resolution (SR), which outperforms traditional methods with a huge margin [21]. For example, a CNN-based up-sampling scheme is proposed in [16] for intra-frame coding. A video coding scheme is proposed in [22], where a group of pictures

[*] Corresponding author at: School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China.
E-mail addresses: li.yu@nuist.edu.cn (L. Yu), moncef.gabbouj@tuni.fi (M. Gabbouj).

https://doi.org/10.1016/j.image.2022.116754
Received 12 October 2021; Received in revised form 23 May 2022; Accepted 27 May 2022
Available online 3 June 2022

(GOP) is entirely down-sampled and compressed, and each frame is individually up-sampled using trained CNN models. Although these methods have demonstrated the potential of down/up-sampling-based coding with CNN for improving compression performance, they have not exploited the correlation between neighboring frames.

To this end, several attempts have been made to use the correlation between neighboring frames to further enhance the performance [17,23,24]. Lin et al. [25] proposed to adaptively divide frames into keyframes (KFs) and nonkey frames (NKFs), which are encoded at the original resolution and at a reduced resolution, respectively. At the decoder, NKFs are reconstructed with the corresponding motion estimation block in KFs using CNN. In addition to frame-level down/up-sampling-based coding, block-level down/up-sampling-based coding are also proposed to improve the performance [17,26]. In [17], each block in the P/B frame can either be compressed at the original resolution or down-sampled and compressed at a lower resolution. At the decoder, low-resolution blocks are up-sampled by the CNN models. The block-based scheme provides the flexibility to deal with the spatially variant texture and motion characteristics in natural videos. However, it is quite computational intensive for block-level processing.

In this paper, frame-level down/up-sampling-based coding is employed to reduce the computational complexity. Meanwhile, a neural texture transfer-based frame up-sampling is proposed to cope with the spatially variant texture and motion characteristics in natural videos. As in the existing frame-level schemes [23,25], features of the low-resolution frame and the full-resolution reference frame are simply concatenated, where only co-located features can be fused. However, the best matching feature is not necessarily co-located, thus leading to a sub-optimal result. While with neural texture transfer [27], a multi-level matching is conducted in the neural space, instead of the raw pixel space, to adaptively transfer texture from the reference images to the target image. This matching scheme facilitates the semantic texture transfer, which provides robust results even when irrelevant reference images are provided. To achieve optimized performance, the neural texture transfer model was fine-tuned on HEVC compressed video sequences. The proposed approach has been compared to both the HEVC anchor [28] and the block-level scheme [17], with results demonstrating consistent improvements on the HEVC common test sequences [29] for different QP ranges. Specifically, the main contributions of our work are as follows:

- We propose to use the semantic texture transfer for down/up-sampling-based video coding, which exploits the semantic correlation between the reference frames and the target frame, leading to significant enhancement of the frame-level SR performance.
- We propose a non-uniform compression scheme, where frames are adaptively compressed at the original or reduced resolution. Thus, frames encoded at the original resolution can be used as reference frames for the restoration of other frames.
- Our model outperforms the state-of-the-art block-level down/up-sampling-based coding scheme while requiring a lower computational complexity.

The rest of the paper is organized as follows. In Section 2, the deep learning-based up-sampling methods and down/up-sampling-based video coding methods are reviewed. Section 3 introduces the proposed neural texture transfer assisted frame-level down/up-sampling-based coding scheme, the CNN architecture of the neural texture transfer, and the training strategy. Section 4 describes the experiments and results. Finally in Section 6, conclusion and future work are presented.

2. Related work

In this section, we briefly overview the most related works, including deep-learning-based image/video super-resolution and down/up-sampling-based video coding methods.

2.1. Deep learning-based image/video super resolution

CNN-based image/video SR methods can be classified into two categories, namely, Single Image Super-Resolution (SISR) and Reference-based Super Resolution (RefSR).

2.1.1. Single Image Super Resolution (SISR)

SISR is an ill-posed problem, which is defined as recovering a high-resolution (HR) image from its low-resolution (LR) observation. As a pioneering work based on CNN, [30] proposed the super-resolution CNN (SRCNN), which used a three-layer full convolution network to learn the complex non-linear mapping between LR and HR, and achieved great improvement over previous works. Based on SRCNN, VDSR [31] further improved the SR performance by increasing the network depth and introducing a residual network in the reconstruction model. DRCN [32] was the first to introduce Recurrent Neural Network (RNN) in the SR task. It enhances the SR performance by using RNN, residual network, and a broader receptive field. Inspired by densely connected convolutional networks [33], SRDenseNet [34] used densely connected structures to fuse features at different levels and achieved better performance. As the network depth continues to deepen, the performance of super-resolution is also improved, but the huge network parameters bring complex calculations. Therefore, some lightweight super-resolution methods have also been proposed. [35] uses grouped convolution to reduce the amount of calculation, [36] proposes a recursive block method to reduce the amount of parameters, and each recursive block parameter is fixed to be used recursively many times. Although the above works achieved high PSNR scores, the high-frequency details are lost in the reconstructed images, which leads to poor subjective quality. Thus in SRGAN [37], the LR image was fed into a Generative Adversarial Network (GAN) to achieve a visually plausible result, instead of pursuing a high PSNR score.

2.1.2. Reference-based Super Resolution (RefSR)

RefSR compensates for the lost details in the LR images by utilizing rich textures in the HR references (Ref) to relax the ill-posedness issue and produce more realistic and finer textures with the aid of reference images. Traditional RefSR methods assume the reference images share similar content as that of the LR image with a good alignment. In order to get well-aligned reference image and LR image pairs, the landmark method was proposed in [38], which searched for well-matched reference images from the Internet for the recovery of LR image. In CrossNet [39], the reference image and the LR image are aligned with the assistance of optical flow. However, the performance of these methods degrade significantly and even become worse than SISR methods if the reference image and LR image are not well aligned in terms of content. Yet, an ideal RefSR algorithm should outperform SISR when good reference images are provided and could achieve a comparable performance as SISR when reference images are not provided or do not possess relevant texture at all. Thus, in Super-Resolution by Neural Texture Transfer (SRNTT) [27], the textures are adaptively transferred from the reference images to the LR image in the feature space, instead of the pixel domain. Specifically, SRNTT conducts local texture matching and learns the complicated dependency between LR and Ref textures in the feature space. Then, matched textures are transferred to the final output through a deep model, while suppressing dissimilar textures. Thus, even if a totally irrelevant reference image is given, SRNTT can still achieve at least a similar result as SISR methods. Inspired by the SRNTT model, we build a novel SR model which can adaptively search for and transfer related textures between the reference frame and the LR frame in this work.

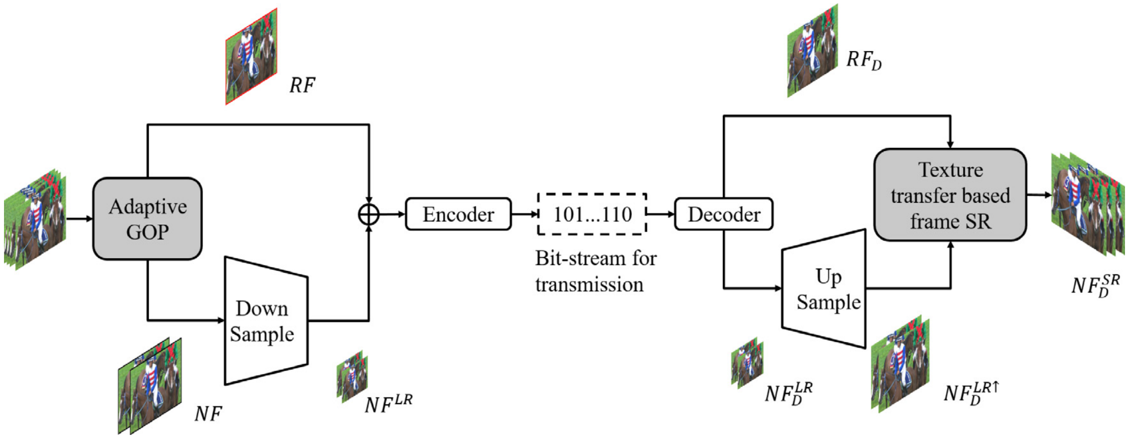


Fig. 1. Diagram of the proposed methods (in gray boxes), integrated into a typical video encoding/decoding flow. The video frames are adaptively divided into reference frames (RF) and non-reference frames (NF) by the proposed adaptive GOP method. The RFs are encoded at full resolution, while NFs are encoded at reduced resolution using the standard HEVC encoder. At the decoder, all frames are decoded by the standard HEVC decoder first. Then, the decoded full-resolution frame RF_D is used to facilitate the super-resolution of low-resolution NF_D^{LR} using the proposed neural texture transfer network.

2.2. Down/up-sampling-based video coding

Down/up-sampling-based video coding reduces the total bitrate at the encoder by down-sampling the video spatially and recovers the video with the assistance of SR at the decoder. The rapid development of SR triggers the improvement of a down/up-sampling-based video coding scheme. This down/up-sampling-based video coding scheme achieves better rate–distortion performance when the storage or transmission bandwidth is limited. Currently, the majority of such schemes down-samples the video with a ratio of 1/2 or 1/4 at the encoder and then recovers them to the original size via SR at the decoder. Experimental results showed that, compared with the standard codec (such as JPEG, H.264/AVC, HEVC, etc.), the above scheme achieves a better rate–distortion performance. The down/up-sampling-based video coding scheme can be implemented at a block-level or frame-level, leading to two groups of methods.

2.2.1. Block-level down/up-sampling-based video coding

As for block-level methods, a CNN-based model for intra coding was proposed in [16]. This was later extended to inter coding, where inter-frame correlations are exploited to further improve the performance [17]. In [26], the block resolution was adaptively tuned in the enhancement layer encoder, which achieves better performance than HEVC with low computational complexity.

2.2.2. Frame-level down/up-sampling-based video coding

Traditional methods for frame-level approach achieved better RD performance at low bitrates [40–42]. Further improvement was achieved in [43], where an image quality enhancement CNN was used prior to the SR process, which alleviates the compression artifacts effectively. All of these approaches are based on SISR technology suffering from a common drawback, that is the generated high-resolution image has various types of degradation, including those caused by video down-sampling and artifacts generated during the SR process. To address this problem, [20] employed an end-to-end deep convolutional neural network to directly train the correlation degradation model. In [44], the image enhancement network was applied before super-resolution to reduce the artifacts brought by video compression. Lin et al. [25] adaptively divided frames into keyframes (KFs) and non-key frames (NKFs), which were encoded at the original resolution and at a reduced resolution, respectively. At the decoder, NKFs were reconstructed with the corresponding motion estimation block in the KFs using CNN. In order to improve the subjective quality, the GAN-based method [45] was proposed to compensate for various degradation caused by the compression and down/up-sampling process. Nonetheless, the texture of the output image is not real.

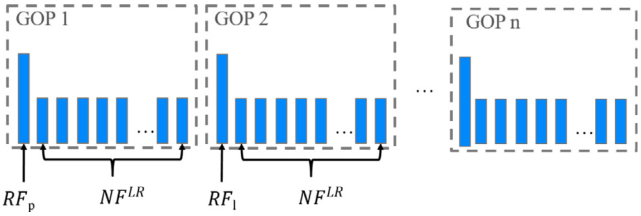


Fig. 2. Proposed adaptive GOP, where frames in the video sequences are adaptively divided into GOPs, with the first frame of each GOP encoded in full resolution (denoted as RF) and others at reduced resolution (denoted as NF^{LR}).

3. Methodology

As shown in Fig. 1, our proposed method first divides the video sequence adaptively into groups of pictures, called as adaptive GOP. The first frame of GOP (denoted as RF) remains full resolution, while the remaining frames in GOP (denoted as NF) are down-sampled to a reduced resolution (denoted as NF^{LR}) using bicubic down-sampling (×1/2). As for encoding, RF is encoded at full resolution. Then, the encoder reconstructed version RF_D is bicubic down-sampled (×1/2) to assist the encoding of NF^{LR}. Only the full-resolution version of the compressed RF is transmitted to the receiver.

While at the decoder, a similar process is performed. That is, the decoded RF_D is bicubic down-sampled to assist the decoding of the following frames in the GOP. Next, all the decoded frames at reduced resolution NF_D^{LR} are up-sampled (×2) using bicubic interpolation (the output is denoted as $NF_D^{LR\uparrow}$). Then, $NF_D^{LR\uparrow}$ and RF_D are fed into the neural texture transfer model, where texture details in RF_D are used to facilitate the reconstruction of $NF_D^{LR\uparrow}$.

3.1. Adaptive GOP

The proposed adaptive GOP method divides the video sequence into groups of pictures as shown in Fig. 2, where the first frame of each GOP (i.e. RF) is used to assist the reconstruction of the rest frames in the GOP (i.e. NF^{LR}). Obviously, the reconstruction achieves better performance when frames in one GOP share similar contents and textures. Thus, we divide the GOP based on the similarity between frames. Consecutive frames sharing similar contents will be placed in one GOP. Usually, the similarity between frames decreases as the further apart are the frames, this will lead to small GOP sizes. The extreme case is one frame per GOP, which turns out to be all-intra encoding. This

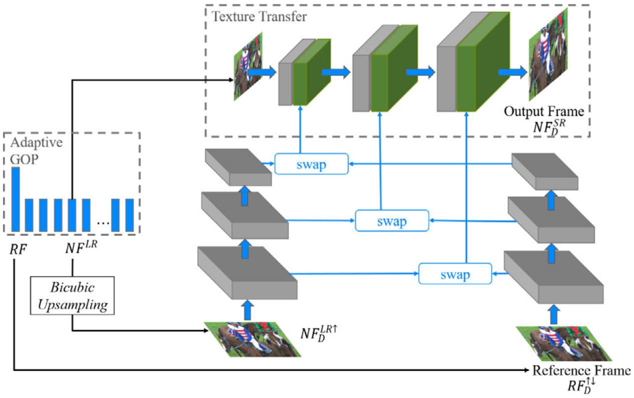


Fig. 3. Neural texture transfer based frame super-resolution model.

contradicts the aim of bit saving. Thus, the proposed adaptive GOP method aims to find a tradeoff between the reconstruction performance (i.e. smaller GOP) and the bit saving (i.e. larger GOP).

We start from an initial GOP size of N. Assume the first frame in the current and next GOP are RF_p and RF_l respectively, the distance between these two frames is $N-1$. The mean absolute error (MAE) between RF_p and RF_l is

$$MAE_{p,l} = \frac{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left|f_l(i,j) - f_p(i,j)\right|}{m \times n}, \tag{1}$$

where $f_p(i,j)$ and $f_l(i,j)$ represent the pixel at position (i,j) in RF_p and RF_l, and the resolution of each frame is (m,n). Then, the MAEs between RF_p and its subsequent frames in current GOP are calculated:

$$MAE_x = \frac{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left|f_p(i,j) - f_x(i,j)\right|}{m \times n}, \tag{2}$$

where x is the frame index in the range of (p,l). Based on Eqs. (1) and (2), the relative changing ratio (RCR) is calculated as below:

$$RCR = \frac{MAE_x}{MAE_{p,l}}, \tag{3}$$

which is used as the similarity measurement in our proposed adaptive GOP method. When RCR is larger than a predefined threshold T, the xth frame will be set as another RF, and the frames from p to $(x-1)$ forms one GOP. Then, another round of evaluation will be applied to determine the next GOP, until all GOPs are formed.

3.2. Texture transfer based frame SR model

Fig. 3 shows the overview of the neural texture transfer model, which is inspired by the reference-based image super-resolution work [27]. The input of the model is a couple of frames, i.e. RF_D of full resolution and NF_D^{LR} of down-sized (×1/2) resolution. While the output of the model is the ×2 up-sampled version of NF_D^{LR}, i.e. NF_D^{SR}. The ground truth is the corresponding original version of NF. The goal of the network is to search for matching textures from RF_D in the feature space, as the features are more robust to variations in color and illumination. The search process is conducted by the swap unit in a multi-scale way, where the NF_D^{LR} is first bicubic up-sampled to the same size as RF_D. Both semantic and textural similarities are evaluated in a swap unit to only transfer related textures while suppressing unrelated textures. The swapped texture feature maps are then merged into a base deep generative network at different layers, as shown in the texture transfer unit. The output frame is generated through layers in the texture transfer unit to reach the target resolution.

The reconstruction loss L_{rec} is the L_1 loss between NF_D^{SR} and NF, which is calculated as follows:

$$L_{rec} = \left\|NF - NF_D^{SR}\right\|_1. \tag{4}$$

The texture difference between NF_D^{SR} and RF_D (denoted as texture loss, L_{tex}) is also considered, enforcing the adaptive texture transfer from RF_D to NF_D^{SR}. Specifically, the L_{tex} is computed as follows:

$$L_{tex} = \sum_l \lambda_l \left\|Gr\left(\phi_l\left(NF_D^{SR}\right)\cdot S_l^*\right) - Gr\left(M_l \cdot S_l^*\right)\right\|_F, \tag{5}$$

where $Gr(\cdot)$ denotes the calculation of the Gram matrix. l represents each neural layer, and λ_l is the normalization factor corresponding to that layer. M_l is the exchange feature map obtained from $RF_D^{\uparrow\downarrow}$, and S_l^* is the weight map calculated by each patch in $NF_D^{LR\uparrow}$ and the most similar patch in $RF_D^{\uparrow\downarrow}$. F denotes the Frobenius norm.

Besides the texture and reconstruction losses, L_{tex} and L_{rec} mentioned above, perceptual loss L_{per} and adversarial loss L_{adv} are also used in the loss function. The perceptual loss [46] has been widely used in recent SR tasks for better visual quality. The L_{per} is computed as follows:

$$L_{\text{per}} = \left\|\phi_i\left(NF\right) - \phi_i\left(\boldsymbol{NF}_D^{SR}\right)\right\|_2, \tag{6}$$

where ϕ_i denotes the feature map of the ith layer in the VGG19. As GANs [47] can significantly enhance the sharpness and visual quality of the synthesized images, we adopt WGAN-GP [48], which proposes a gradient norm penalty to make the training stable and achieve better results. The L_{adv} is computed as follows:

$$\begin{aligned} L_{adv} = & \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_g}[D(\tilde{\boldsymbol{x}})] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_r}[D(\boldsymbol{x})] \\ & +\lambda \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}}\sim\mathbb{P}_{\hat{\boldsymbol{x}}}}\left[\left(\left\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\right\|_2 - 1\right)^2\right]. \end{aligned} \tag{7}$$

Ultimately, the overall loss is defined as:

$$\begin{aligned} L = & \lambda_{tex} \times L_{tex} + \lambda_{rec} \times L_{rec} \\ & +\lambda_{per} \times L_{per} + \lambda_{adv} \times L_{adv}. \end{aligned} \tag{8}$$

The texture matching procedure plays a vital role, which enables the texture transfer between misaligned patches. Through extensive experiments, we found that optimal matching performance was obtained when the QPs for encoding RF and NF follows the following rule:

$$QP(RF) = QP(NF) + 10. \tag{9}$$

where constant 10 is selected by thorough experiments.

3.3. Training configuration

For referenced SR problems, the similarity between a low-resolution image and the Ref images has a vital influence on the final result. In general, frame pairs with different levels of similarity should be provided for training. Here, we train our network on the CUFED dataset [49], which includes image couples of four similarity levels. There are a total of 11,871 paired patches of size 160 × 160 in the CUFED training set.

The model is first trained on the original CUFED dataset with 30 epochs to learn the texture transfer for various similarity levels. Then, the model is further fine-tuned with a compressed CUFED dataset with another 5 epochs, to adapt to various compression artifacts. The compression is fulfilled with HEVC HM 12.1,[1] using Low-Delay P configuration. The QP for the Ref image is set to 46 and the QP for the low-resolution image is 36. The intra period is set to 4.

The network is implemented with Tensorflow 1.13.1. Adam optimizer is used for training the model with a learning rate of 10^{-4}. When the trained model is used in the pipeline as shown in Fig. 1, RF and NF can be directly inputted into the model without any cropping.

[1] https://hevc.hhi.fraunhofer.de/trac/hevc/milestone/HM-12.1

Table 1
BD-rate results of our scheme compared to HEVC (Y stands for Y-PSNR, S stands for Y-SSIM, QP 42–51).

Class	Sequence	RA		LDB		LDP	
		Y (%)	S (%)	Y (%)	S (%)	Y (%)	S (%)
Class A (2560 × 1600)	Traffic	−12.4	−13.2	~	~	~	~
	PeopleOnStreet	−5.4	−6.4	~	~	~	~
Class B (1920 × 1072)	Kimono	−15.4	−19.6	−15.1	−19.3	−19.7	−29.4
	ParkScene	−17.6	−17.9	−16.0	−16.0	−18.4	−17.1
	Cactus	−13.1	−9.7	−11.3	−7.7	−12.7	−9.0
	BasketballDrive	−2.5	−3.5	−2.1	−2.6	−3.9	−7.2
	BQTerrace	−2.5	5.2	0.4	8.7	1.9	9.7
Class C (832 × 480)	BasketballDrill	−6.8	−5.7	−4.9	−5.7	−6.3	−8.4
	BQMall	5.2	5.0	6.8	6.4	5.8	5.4
	RaceHorses	−10.0	−9.2	−9.4	−8.8	−10.9	−10.8
Class D (416 × 240)	BasketballPass	−8.1	−7.2	−8.2	−6.9	−9.6	−8.4
	BQSquare	−6.5	−6.6	−5.9	−5.9	−5.5	−5.8
	BlowingBubbles	−10.9	−7.3	−8.9	−5.0	−10.2	−6.0
	RaceHorses	−8.7	−8.5	−8.5	−8.6	−9.0	−9.0
Class E (720p)	FourPeople	~	~	−3.0	−7.5	−3.3	−8.4
	Johnny	~	~	−5.5	−6.0	−6.5	−8.3
	KristenAndSara	~	~	−4.2	−8.6	−5.0	−10.7
Average	Class A	−8.9	−9.8	~	~	~	~
	Class B	−10.2	−9.1	−8.8	−7.4	−10.6	−10.6
	Class C	−3.9	−3.3	−2.5	−2.7	−3.8	−4.6
	Class D	−8.6	−7.4	−7.9	−6.6	−8.6	−7.3
	Class E	~	~	−4.2	−7.4	−4.9	−9.1
Average of Classes A–E		**−7.9**	**−7.4**	**−5.9**	**−6.0**	**−7.0**	**−7.9**

4. Experiment

4.1. Experimental settings

The proposed scheme is implemented based on the reference software of HEVC (HM 12.1). The HEVC common test sequences [29] are used to evaluate the performance of the proposed method, with various resolutions, known as Class A, B, C, D, E.[2] None of these sequences was used in training the SR model in Fig. 3. The Low-Delay P (LDP), Low-Delay B (LDB), and Random Access (RA) configurations are used in the following experiments, with QP values of 32, 37, 42, and 47. The initial GOP size N is set as 4.

The RCR threshold T is set as 0.55. We empirically set the weights in Eq. (8) as $\lambda_{tex} = 1e-4$, $\lambda_{rec} = 1.0$, $\lambda_{per} = 1e-4$, $\lambda_{adv} = 1e-6$.

Our method is compared to the HEVC anchor and the block-level down/up-sampling-based coding method [17]. To evaluate the performance, we adopt PSNR and SSIM [50] for the Y-component. The BD-rate [51] is also used to compare different coding schemes.

The experiments are conducted on a PC equipped with Intel(R) Core(TM) i7-9700K CPU, 32 GB of RAM, NVIDIA GeForce RTX 2080Ti GPU.

4.2. Performance comparison with the standard HEVC

Table 1 summarizes the BD-rate between our method and the standard HEVC. The QPs of NF are between {32, 35, 38, 41}, while QPs of RF between {42, 45, 48, 51}. Generally, our proposed method outperforms the standard HEVC over all encoding configurations for sequences of different resolutions. The average BD-rate (PSNR/SSIM) of RA is (−8.2%/−7.2%), LDB is (−6.2%/−5.8%) and LDP is (−7.3%/−7.6%). Since our method adopts the reference-based super-resolution network, the texture of the reference frame plays a very important role in the reconstruction effect. For video sequences with simple motion and small changes in video frames, the adjacent frames have more similar textures, and naturally the reconstructed video frames have better visual quality. Therefore, our method presents higher coding gain for sequences with uncomplicated motion than those with complex motion. For example, the Kimono sequence in Class B achieves the highest BD-rate (PSNR/SSIM) reduction when encoded using the LDP (−19.7%/−29.4%).

When comparing the results of sequences with different resolutions, our method has a higher gain for high-resolution sequences. For example, RaceHorses has two sequences with different resolutions (Class C and Class D), and their sequence contents are consistent. For low resolution Class D RaceHorses sequences, our method achieves BD-rate (PSNR/SSIM) for RA of (−8.7%/−8.5%), LDB of (−8.5%/−8.6%) and LDP of (−9.0%/−9.0%). However, for high-resolution Class C RaceHorses sequences, our method achieves the reduction of BD-rate (PSNR/SSIM) for RA of (−10.0%/−9.2%), LDB of (−9.4% / −8.8%) and LDP of (−10.9%/−10.8%). This is because high-resolution images contain more information than low-resolution images for texture transfer.

The R–D curves of our proposed method and standard HEVC under different encoding configurations for different video sequences are shown in Fig. 4. For all sequences, our scheme achieves a better performance than the standard HEVC at low bit rates. At the same time, competing performance is achieved at high bit rates.

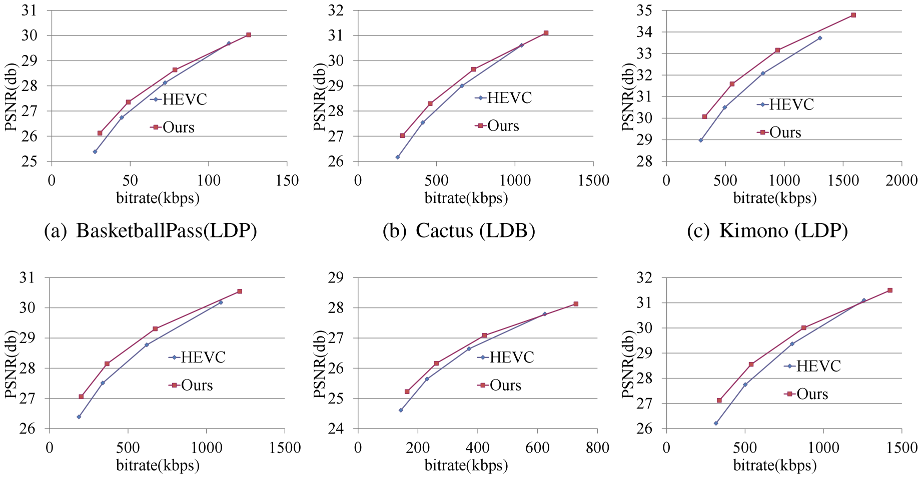
4.3. Performance comparison with other methods

In order to compare with existing work [17], we choose a similar QP setting, i.e. the QPs of NF are {32, 35, 38, 41} and the QPs of RF are {42, 45, 48, 51}. The method in [17] uses the relevant information of the surrounding reference frames to reconstruct the video frames at the block level. Different from the block-based reference that has less useful information to be used, our method migrates similar textures at the frame level, and exploits the texture information of three different scales for the reconstruction, hence exploiting the global information of the reference frame relative to the block level. Table 2 presents the BD-rates of our method compared with [17] under three different configurations of RA, LDB, and LDP. Overall, we can observe that our method outperforms [17] in sequences of Class A, B, D, and E, but not for sequences of Class C. This is because sequences in Class C, such as BQMall, contain complex textures, which are severely lost during the

[2] In order to meet the resolution requirements of HEVC CU division, the resolutions are slightly cropped, such as Class B sequences are cropped into a resolution of 1920 × 1072.

Table 2
BD-rate results of our scheme and [17] compared to HEVC (Y stands for Y-PSNR, S stands for Y-SSIM, QP 32–47). Bold values show the best BD-rate (%) on average.

Class	Sequence	RA				LDB				LDP			
		Y (%)		S (%)		Y (%)		S (%)		Y (%)		S (%)	
		[17]	Ours	[17]	Ours	[17]	Ours	[17]	Ours	[17]	Ours	[17]	Ours
Class A (2560 × 1600)	Traffic	−6.6	−8.5	−9.6	−10.4	~	~	~	~	~	~	~	~
	PeopleOnStreet	−3.6	−2.4	−3.9	−3.4	~	~	~	~	~	~	~	~
Class B (1920 × 1072)	Kimono	−5.8	−11.7	−7.4	−16.5	−2.7	−19.1	−4.5	−19.6	−4.2	−16.2	−9.6	−26.5
	ParkScene	−2.3	−13.5	−2.6	−13.4	−1.7	−14.5	−2.1	−14.0	−2.3	−14.4	−3.7	−13.3
	Cactus	−3.8	−9.4	−4.0	−6.1	−3.3	−9.8	−3.8	−5.3	−3.8	−8.6	−4.7	−5.0
	BasketballDrive	−7.5	0.3	−8.6	0.1	−7.0	−4.8	−9.2	−1.3	−9.5	−0.1	−13.5	−3.2
	BQTerrace	−1.5	2.5	−1.0	11.1	−2.9	−2.3	−2.4	5.8	−3.7	−2.4	−3.2	6.8
Class C (832 × 480)	BasketballDrill	−6.4	−3.1	−7.7	−3.8	−5.7	−4.3	−7.9	−4.6	−7.9	−2.7	−10.3	−2.9
	BQMall	−2.3	9.8	−3.1	7.9	−1.9	8.4	−3.1	6.8	−3.3	11.1	−4.8	8.4
	RaceHorses	−5.9	−8.3	−8.2	−5.8	−2.9	−8.5	−4.2	−5.6	−4.1	−8.6	−6.5	−6.7
Class D (416 × 240)	BasketballPass	−1.0	−5.1	−0.7	−4.3	−1.0	−6.8	−1.1	−5.3	−0.6	−6.7	−0.6	−5.6
	BQSquare	0.1	−4.1	−0.4	−3.7	0.2	−3.3	−0.0	−1.6	−0.4	−2.5	−1.5	−0.7
	BlowingBubbles	−2.6	−7.8	−2.7	−5.2	−1.8	−6.4	−2.3	−4.0	−2.8	−8.2	−3.9	−5.2
	RaceHorses	−4.1	−5.1	−5.0	−5.1	−1.3	−4.8	−1.7	−4.4	−2.5	−5.7	−3.4	−5.7
Class E (720p)	FourPeople	~	~	~	~	−2.1	−2.3	−3.0	−9.4	−2.2	−1.2	−3.5	−9.8
	Johnny	~	~	~	~	−2.9	−3.6	−1.8	−2.8	−4.0	−3.3	−3.6	−2.7
	KristenAndSara	~	~	~	~	−2.7	−2.8	−3.5	−9.1	−2.2	−2.0	−3.8	−11.4
Average	Class A	−5.1	**−5.5**	−6.8	**−6.9**	~	~	~	~	~	~	~	~
	Class B	−4.2	**−6.4**	−4.8	**−5.0**	−3.5	**−10.1**	−4.4	**−6.9**	−4.7	**−8.3**	−6.9	**−8.2**
	Class C	**−4.9**	−0.5	**−6.3**	−0.6	**−3.5**	−1.5	**−5.1**	−1.1	**−5.1**	−0.1	**−7.2**	−0.4
	Class D	−1.9	**−5.5**	−2.2	**−4.6**	−1.0	**−5.3**	−1.3	**−3.8**	−1.6	**−5.8**	−2.4	**−4.3**
	Class E	~	~	~	~	−2.6	**−2.9**	−2.8	**−7.1**	**−2.8**	−2.2	−3.6	**−8.0**



(a) BasketballPass(LDP) (b) Cactus (LDB) (c) Kimono (LDP)

(d) ParkScene (RA) (e) RaceHorses (LDB) (f) Traffic (RA)

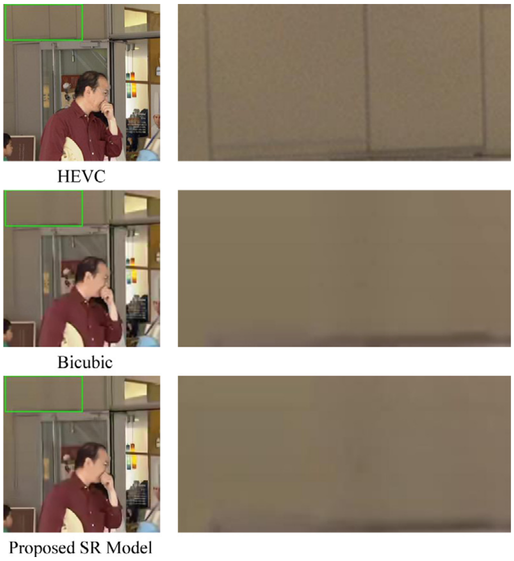
Fig. 4. Rate–distortion (R–D) curves of several sequences: (a) BasketballPass (LDP); (b) Cactus (LDB); (c) Kimono (LDP); (d) ParkScene (RA); (e) RaceHorses (LDB); (f) Traffic (RA). For each chart, the *x*-axis (horizontal) represents bitrate (kbps); *y*-axis (vertical) represents PSNR (dB).

down/up-sampling process. As a result, the details for textual matching are not available in the proposed SR model, leading to the poor performance. For example, in sequence BQMall, as shown in Fig. 5, the textures of the wall and the billboard are lost during the down/up-sampling process. Thus, when using the reference frame for texture matching and migration, the texture information in the reference frame cannot be correctly matched and inaccurate details are migrated.

The average performance gain at B and D resolution is better than the other two resolutions. Under B resolution, the highest BD-rate (PSNR) reduction of LDB (−10.1%) and BD-rate (SSIM) reduction of LDP (−8.2%) are achieved, while the method proposed in [17] provides BD-rate (PSNR) reduction of LDB (−3.5%) and BD-rate (SSIM) reduction of LDP (−6.9%). At D resolution, the highest BD-rate (PSNR) reduction of LDP (−5.8%) and BD-rate (SSIM) reduction of RA (−4.6%) are obtained, while [17] provides BD-rate (PSNR) reduction of LDP (−1.6%) and BD-rate (SSIM) reduction of RA (−2.2%). From the results of a single test sequence, for the sequence with little motion change, our method can make full use of the information of the reference frame. For example, the Kimono test sequence at B resolution achieved the highest BD-rate (PSNR) reduction of LDB (−19.1%) and BD-rate (SSIM) reduction of LDP (−26.5%).

4.4. Subjective comparison

For video compression, the quality of a video frame reconstructed at the decoder should be evaluated both objectively (PSNR/SSIM), as well as subjectively. In the latter, an evaluation that conforms to the human visual perception is an effective video quality evaluation metric. Therefore, we make a subjective comparison of the video frames in Traffic, PeopleOnStreet, Cactus, and ParkScene under the Class A test sequence. For class B video sequences, the Bicubic method is used to downsample the video sequence. It can be seen from Fig. 6 that our proposed method can restore a clearer texture in medium and low quality frames. For better comparison, some selected areas are enlarged. In these figures, the frames encoded using the proposed



HEVC

Bicubic

Proposed SR Model

Fig. 5. Analysis of the poor performance for sequence BQMall. From top to bottom are the results of HEVC (LDP, QP42), bicubic down/up-sampling, and the proposed SR model. It can be seen that the detailed textures are lost during the down/up-sampling process, such as the vertical lines on the wall. Thus in the subsequent texture search in the proposed SR model, the vertical lines cannot be matched and recovered.

Table 3
BD-rate results with AKS/Adaptive GOP and Texture Transfer.

Class	Sequence	AKS [25]	Adaptive GOP	Texture Transfer	BD-rate
Class A	Traffic			✓	−7.2
		✓		✓	−3.7
			✓	✓	−12.4
Class B	Cactus			✓	−8.3
		✓		✓	−5.2
			✓	✓	−13.1
Class C	RaceHorses			✓	−4.9
		✓		✓	−2.5
			✓	✓	−8.7
Class D	BlowingBubbles			✓	−7.1
		✓		✓	−2.9
			✓	✓	−10.9
Class E	Johnny			✓	−3.3
		✓		✓	−1.4
			✓	✓	−5.5

Table 4
Ablation studies of the loss functions.

L_{rec}	L_{tex}	L_{per}	L_{adv}	PSNR	SSIM
✓				27.0615	0.7203
✓	✓			27.5613	0.7332
✓	✓	✓		27.5409	0.7338
✓	✓		✓	27.5462	0.734
✓	✓	✓	✓	27.5459	0.7345

model have fewer artifacts, especially in the selected regions. At the same time, for sequences of different scenes, our solution offers good texture restoration results. For medium quality frames, the characters in the Traffic and Cactus sequences are clearer and the patterns are more complete. Our results also produce fewer artifacts. For outdoor sports sequences such as ParkScene and PeopleOnStreet, the texture restoration effect of the pillars in the park and the texture of the shirt are closer to those in the original images, and the restoration results also present a finer surface and clearer edge contours. For low quality frames, our method also achieves good results. In the RaceHorses sequence, the video frame recovered by our method is more detailed, such as the reflective part on the boot. In BlowingBubbles, our method makes the contour between fingers more obvious and closer to the original image. For BQTerror sequence, our method can better restore the texture boundary region of the chair. Finally, for the BasketballDrill sequence, our method is more precise for the texture detail restoration of the basketball net, which is close to the original image. As can be seen, our method has good subjective quality when there are sharp edges and more details in medium and low quality frames. Recall that the signal-to-noise ratio of the proposed algorithm is higher than that of the standard HEVC coding. As a result, our method has both better objective video quality and subjective quality than the standard HEVC.

5. Ablation studies

5.1. The effect of texture transfer based frame SR model

As shown in Fig. 7, the PSNR value of each frame before and after the texture transfer-based SR are depicted. The blue curves are the PSNR values of each frame before the texture transfer-based SR, i.e. only bicubic up-sampled. While the red curves are the PSNR values of each frame after the proposed texture transfer-based SR model. Overall, the PSNR gain of the proposed texture transfer-based SR model is from 0.8 to 1.3 dB, and the average gain is around 1 dB.

At the same time, we also conducted experiments on the HEVC standard test sequences. We randomly selected one sequence from Class A–E for experiments. When only texture transfer was used, the experimental results were shown in Table 3, and achieved certain coding performance improvement. BD-rate was reduced by −7.2%, −8.3%, −4.9%, −7.1%, −3.3% respectively.

5.2. The effect of adaptive GOP

The proposed method uses RF to reconstruct NF, so the quality of RF has an important impact on the result of NF reconstruction. For a fixed GOP, the intra-frame information of the RF is not necessarily applicable to all NFs in the current GOP. As we get farther from the reference frame, the similarity between RF and subsequent NF gradually decreases, and even scene jumps occur. At this time, NF cannot make full use of RF information. Therefore, the adaptive GOP method is adopted to find two RFs adaptively, and the key information of RF is effectively used in the most similar GOP. As a result, the bit rate is slightly increased, but the video quality is greatly improved. In order to validate the effectiveness of the proposed adaptive GOP, we compare the RD curve of adaptive GOP with the curve with a fixed GOP of 4 in Fig. 8. It can be seen that the RD curve of adaptive GOP achieves a better performance than the RD curve with a fixed GOP of 4. For all sequences, adaptive GOP achieves better results than fixed GOP of 4. For example, for sequence Traffic in Class A, when fixed GOP is used, BD-rate was reduced by −7.2%. When adaptive GOP is used, BD-rate was reduced by −12.4%, which is −5.2% higher than that of fixed GOP. These results illustrate that the proposed adaptive GOP method in this model is effective. Specifically, the adaptive GOP algorithm enables the video frames in each GOP to maintain a high temporal correlation. Therefore, the down-sampled frames can make full use of the relevant information of the reference frame for better reconstruction.

5.3. Contribution of each loss function

In this section, we discuss the contribution of each loss function. As shown in Table 4, when only L_{rec} loss is used, the PSNR and SSIM are the lowest, because L_{rec} loss only carries out point-to-point learning without considering the texture difference between RF and NF. When L_{tex} is added, PSNR and SSIM are greatly improved, which enables the

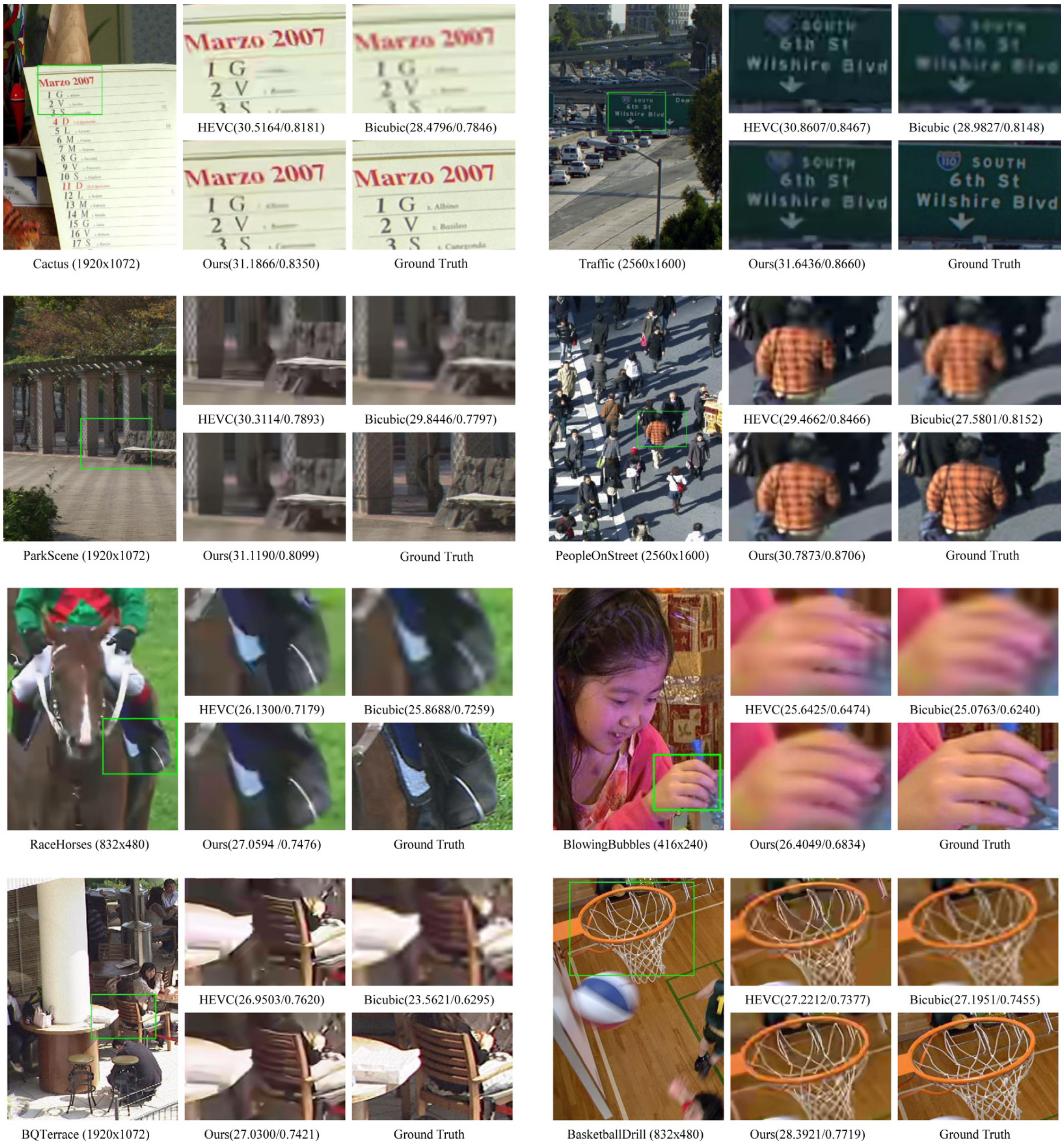


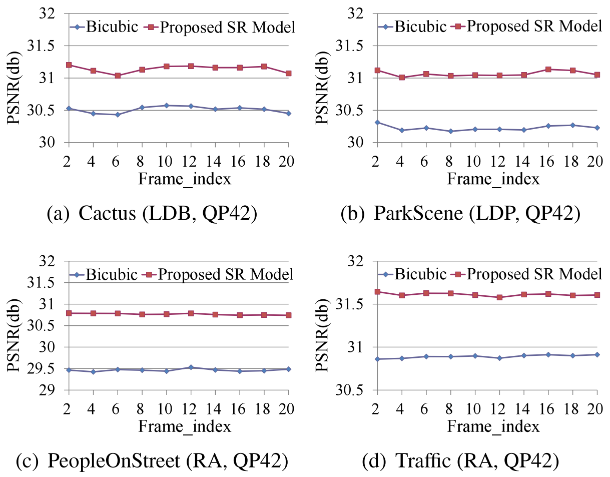
Fig. 6. Subjective comparisons between HEVC, bicubic and our method (PSNR/SSIM) over standard HEVC test sequences: Cactus (LDP, QP42), Traffic (RA, QP42), ParkScene (LDB, QP42), PeopleOnStreet (RA, QP42), RaceHorses (LDP, QP45), BlowingBubbles (LDB, QP45), BQTerrace (RA, QP45), BasketballDrill (RA, QP48).

network to more effectively learn the texture difference between RF and NF, so as to force the adaptive texture to transfer from RF to NF. L_{per} and L_{adv} are proposed to improve the visual effect. It can be seen from the table that, after adding L_{per} and L_{adv}, the objective evaluation index PSNR decreases slightly, while the SSIM index, which is more in line with human vision, is further improved. From the penultimate two and three rows in the table, it can be seen that L_{adv} is more effective in our network than L_{per}. When all losses are used, PSNR can maintain a high level, and SSIM reaches the highest level. Therefore, we finally selected the fusion of four losses in our work.

6. Conclusion

A neural texture transfer-assisted video coding with an adaptive up-sampling scheme is proposed in this paper. This scheme adaptively decides whether a frame should be down-sampled or not. In the decoder, the down-sampled frames are restored by exploring their correlations with the frames that are not down-sampled using neural texture transfer in a multi-scale manner. Experimental results show that, compared with HEVC and state-of-the-art method [17], our model provides better performance in terms of PSNR, SSIM, and visual perception, with up to (−19.1%) BD-rate (PSNR) and (−26.5%) BD-rate (SSIM) reduction.

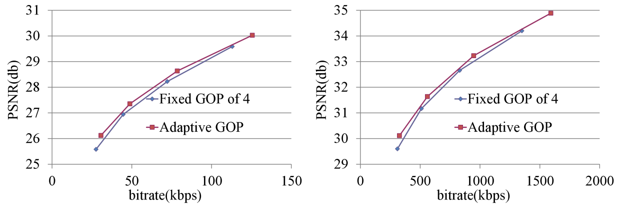
In the future, we plan to expand this work in several directions. First, the quality of the reference frame directly affects the quality of the final reconstruction. We would like to improve the visual quality of the reference frame before the neural texture transfer, so as to alleviate the degradation caused by video compression. Second, we plan to use a wider range of reference video frames to jointly restore the down-sampled video frame, so as to obtain better performance.



(a) Cactus (LDB, QP42) (b) ParkScene (LDP, QP42)

(c) PeopleOnStreet (RA, QP42) (d) Traffic (RA, QP42)

Fig. 7. Comparison of the effect of image reconstruction with proposed SR model. Results for sequences: (a) BasketballPass (LDP, QP42); (b) Cactus (LDB, QP42); (c) PeopleOnStreet (RA, QP42); (d) Traffic (RA, QP42). The red line represents proposed SR model, the blue line represents bicubic. We use the odd-numbered frames of the first 20 frames as a reference to restore the even-numbered frames. For each chart, the *x*-axis (horizontal) represents frame_index; meanwhile, *y*-axis (vertical) represents PSNR (dB).



(a) BasketballPass (LDB) (b) Kimono (RA)

Fig. 8. Rate–distortion (R–D) curve comparison between adaptive GOP and fixed GOP of 4 for sequences BasketballPass (LDB) and Kimono (RA). For each chart, the *x*-axis (horizontal) represents bitrate (kbps), and *y*-axis (vertical) represents PSNR (dB).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62002172, and Grant 61972323; and in part by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant 19KJB510040; and in part by the Nanjing Scientific Innovation Foundation for the Returned Overseas Chinese Scholars under Grant R2019LZ04; and in part by the Jiangsu Provincial Double-Innovation Doctor Program; and in part by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), China under Grant 202100002; and in part by the Startup Foundation for Introducing Talent of NUIST, China under Grant 2018r080.We acknowledge the High Performane Computing Center of Nanjing University of Information Science & Technology for their support of this work.

References

[1] V. Cisco, Cisco visual networking index: Forecast and trends, 2017–2022, White Paper 1, 2018.

[2] Y. Zhang, S. Kwong, S. Wang, Machine learning based video coding optimizations: A survey, Inform. Sci. 506 (2020) 395–423, http://dx.doi.org/10.1016/j.ins.2019.07.096, URL: http://www.sciencedirect.com/science/article/pii/S0020025519307145.

[3] Z. Peng, C. Huang, F. Chen, G. Jiang, X. Cui, M. Yu, Multiple classifier-based fast coding unit partition for intra coding in future video coding, Signal Process., Image Commun. 78 (2019) 171–179.

[4] M. Zhang, W. Zhou, H. Wei, X. Zhou, Z. Duan, Frame level rate control algorithm based on GOP level quality dependency for low-delay hierarchical video coding, Signal Process., Image Commun. 88 (2020) 115964.

[5] H. Yin, H. Cai, E. Yang, Y. Zhou, J. Wu, An efficient all-zero block detection algorithm for high efficiency video coding with RDOQ, Signal Process., Image Commun. 60 (2018) 79–90.

[6] X.H. Van, J. Ascenso, F. Pereira, HEVC backward compatible scalability: A low encoding complexity distributed video coding based approach, Signal Process., Image Commun. 33 (2015) 51–70.

[7] X. Li, X. Lan, M. Yang, J. Xue, N. Zheng, A new compressive sensing video coding framework based on Gaussian mixture model, Signal Process., Image Commun. 55 (2017) 66–79.

[8] F. Raufmehr, M.R. Salehi, E. Abiri, A neural network-based video bit-rate control algorithm for variable bit-rate applications of versatile video coding standard, Signal Process., Image Commun. 96 (2021) 116317.

[9] Q.-T. Vien, T.T. Nguyen, H.X. Nguyen, Deep-NC: A secure image transmission using deep learning and network coding, Signal Process., Image Commun. (2021) 116490.

[10] S. Kuanar, K.R. Rao, C. Conly, N. Gorey, Deep learning based HEVC in-loop filter and noise reduction, Signal Process., Image Commun. 99 (2021) 116409.

[11] D. Liu, Y. Li, J. Lin, H. Li, F. Wu, Deep learning-based video coding: A review and a case study, ACM Comput. Surv. 53 (2020) 1–35.

[12] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, S. Wanga, Image and video compression with neural networks: A review, IEEE Trans. Circuits Syst. Video Technol. (2019).

[13] F. Zhang, C. Feng, D.R. Bull, Enhancing VVC through CNN-based post-processing, in: 2020 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2020, pp. 1–6.

[14] Z. Guan, Q. Xing, M. Xu, R. Yang, T. Liu, Z. Wang, MFQE 2.0: A new approach for multi-frame quality enhancement on compressed video, IEEE Trans. Pattern Anal. Mach. Intell. (2019).

[15] L. Yu, T. Tillo, J. Xiao, M. Grangetto, Convolutional neural network for intermediate view enhancement in multiview streaming, IEEE Trans. Multimed. 20 (2017) 15–28.

[16] Y. Li, D. Liu, H. Li, L. Li, F. Wu, H. Zhang, H. Yang, Convolutional neural network-based block up-sampling for intra frame coding, IEEE Trans. Circuits Syst. Video Technol. 28 (2017) 2316–2330.

[17] J. Lin, D. Liu, H. Yang, H. Li, F. Wu, Convolutional neural network-based block up-sampling for HEVC, IEEE Trans. Circuits Syst. Video Technol. 29 (2018) 3701–3715.

[18] Y. Li, D. Liu, H. Li, L. Li, Z. Li, F. Wu, Learning a convolutional neural network for image compact-resolution, IEEE Trans. Image Process. 28 (2018) 1092–1107.

[19] A.M. Bruckstein, M. Elad, R. Kimmel, Down-scaling for better transform compression, IEEE Trans. Image Process. 12 (2003) 1132–1144.

[20] K. Takahashi, T. Naemura, M. Tanaka, Rate-distortion analysis of super-resolution image/video decoding, in: 2011 18th IEEE International Conference on Image Processing, IEEE, 2011, pp. 1629–1632.

[21] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2015) 295–307.

[22] D. Bull, F. Zhang, M. Afonso, Description of SDR video coding technology proposal by University of Bristol (JVETJ0031), in: The JVET Meeting, ITU-T and ISO/IEC, 2018.

[23] A. Kappeler, S. Yoo, Q. Dai, A.K. Katsaggelos, Video super-resolution with convolutional neural networks, IEEE Trans. Comput. Imaging 2 (2016) 109–122.

[24] J. Schneider, J. Sauer, C. Rohlfing, Adaptive resolution change Using Uncoded Areas and dictionary learning-based super-resolution in versatile video coding, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2020, pp. 2203–2207.

[25] H. Lin, X. He, L. Qing, Q. Teng, S. Yang, Improved low-bitrate HEVC video coding using deep learning based super-resolution and adaptive block patching, IEEE Trans. Multimed. 21 (2019) 3010–3023.

[26] G. Herrou, W. Hamidouche, L. Morin, Low-complexity scalable encoder based on local adaptation of the spatial resolution, in: 2019 IEEE International Conference on Image Processing, ICIP, IEEE, 2019, pp. 3552–3556.

[27] Z. Zhang, Z. Wang, Z. Lin, H. Qi, Image super-resolution by neural texture transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7982–7991.

[28] G.J. Sullivan, J.-R. Ohm, W.-J. Han, T. Wiegand, Overview of the high efficiency video coding (HEVC) standard, IEEE Trans. Circuits Syst. Video Technol. 22 (2012) 1649–1668.

[29] F. Bossen, et al., Common test conditions and software reference configurations, 2013, JCTVC-L1100 12, 7.

[30] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 295–307, http://dx.doi.org/10.1109/TPAMI.2015.2439281.

[31] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 1646–1654, http://dx.doi.org/10.1109/CVPR.2016.182.

[32] J. Kim, J.K. Lee, K.M. Lee, Deeply-recursive convolutional network for image super-resolution, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 1637–1645, http://dx.doi.org/10.1109/CVPR.2016.181.

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

[34] T. Tong, G. Li, X. Liu, Q. Gao, Image super-resolution using dense skip connections, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4799–4807.

[35] A. Esmaeilzehi, M.O. Ahmad, M. Swamy, SRNHARB: A deep light-weight image super resolution network using hybrid activation residual blocks, Signal Process., Image Commun. (2021) 116509.

[36] A. Esmaeilzehi, M.O. Ahmad, M. Swamy, MuRNet: A deep recursive network for super resolution of bicubically interpolated images, Signal Process., Image Commun. 94 (2021) 116228.

[37] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, 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, 2017, pp. 4681–4690.

[38] H. Yue, X. Sun, J. Yang, F. Wu, Landmark image super-resolution by retrieving web images, IEEE Trans. Image Process. 22 (2013) 4865–4878.

[39] H. Zheng, M. Ji, H. Wang, Y. Liu, L. Fang, Crossnet: An end-to-end reference-based super resolution network using cross-scale warping, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 88–104.

[40] G. Georgis, G. Lentaris, D. Reisis, Reduced complexity superresolution for low-bitrate video compression, IEEE Trans. Circuits Syst. Video Technol. 26 (2016) 332–345, http://dx.doi.org/10.1109/TCSVT.2015.2389431.

[41] V.-A. Nguyen, Y.-P. Tan, W. Lin, Adaptive downsampling/upsampling for better video compression at low bit rate, in: 2008 IEEE International Symposium on Circuits and Systems, ISCAS, 2008, pp. 1624–1627, http://dx.doi.org/10.1109/ISCAS.2008.4541745.

[42] J. Wu, Y. Xing, G. Shi, L. Jiao, Image compression with downsampling and overlapped transform at low bit rates, in: 2009 16th IEEE International Conference on Image Processing, ICIP, 2009, pp. 29–32, http://dx.doi.org/10.1109/ICIP.2009.5414006.

[43] L. Feng, X. Zhang, X. Zhang, S. Wang, R. Wang, S. Ma, A dual-network based super-resolution for compressed high definition video, in: Pacific Rim Conference on Multimedia, Springer, 2018, pp. 600–610.

[44] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, W. Shi, Real-time video super-resolution with spatio-temporal networks and motion compensation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4778–4787.

[45] M. Shen, P. Xue, C. Wang, Down-sampling based video coding with super-resolution technique, in: Proceedings of 2010 IEEE International Symposium on Circuits and Systems, 2010, pp. 673–676, http://dx.doi.org/10.1109/ISCAS.2010.5537494.

[46] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[47] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[48] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, Adv. Neural Inf. Process. Syst. 30 (2017) 5767–5777.

[49] Y. Wang, Z. Lin, X. Shen, R. Mech, G. Miller, G.W. Cottrell, Event-specific image importance, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4810–4819.

[50] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004) 600–612.

[51] G. Bjontegaard, Calculation of average PSNR differences between RD-curves, 2001, VCEG-M33.