Boosting the Performance of Video Compression Artifact Reduction with Reference Frame Proposals and Frequency Domain Information

Yi Xu[1]†, Minyi Zhao[1]†, Jing Liu[2]†, Xinjian Zhang[1], Longwen Gao[2], Shuigeng Zhou[1]*, Huyang Sun[2]
[1]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, China
[2]Bilibili, China
{yxu17,zhaomy20,zhangxj17,sgzhou}@fudan.edu.cn
{liujing04,gaolongwen,sunhuyang}@bilibili.com

Abstract

Many deep learning based video compression artifact removal algorithms have been proposed to recover high-quality videos from low-quality compressed videos. Recently, methods were proposed to mine spatiotemporal information via utilizing multiple neighboring frames as reference frames. However, these post-processing methods take advantage of adjacent frames directly, but neglect the information of the video itself, which can be exploited. In this paper, we propose an effective reference frame proposal strategy to boost the performance of the existing multi-frame approaches. Besides, we introduce a loss based on fast Fourier transformation (FFT) to further improve the effectiveness of restoration. Experimental results show that our method achieves better fidelity and perceptual performance on MFQE 2.0 dataset than the state-of-the-art methods. And ***our method won Track 1 and Track 2, and was ranked the 2nd in Track 3 of NTIRE 2021 Quality enhancement of heavily compressed videos Challenge****.*

Figure 1. Examples of high-frequency recovery. These visual cases are compressed frame (top left), prediction of STDF (top right), prediction of a model trained with $L1$ + FFT loss (**ours**) (bottom left), and the ground truth (bottom right).

1. Instruction

To handle the problems of huge storage cost and limited bandwidth while storing and transmitting multimedia data, lossy compression algorithms are commonly used to compress multimedia data (*e.g.* images, audios and videos). These irreversible compression algorithms often introduce compression artifacts that degrade the quality of experience (QoE), especially for videos. Accordingly, video compression artifact removal, which aims to reduce the introduced artifact and recover details for lossy compressed videos, becomes a hot topic in the multimedia field [11, 28, 7].

With the success of deep learning in text, image, and video processing, many deep neural network based compression artifact removal works have emerged and achieved significant performance improvement. The rapid progress in this low-level task can be attributed to deep neural networks [8, 37, 15, 4, 12], various video compression priors [6, 14, 33], and additional temporal information [34, 18, 29, 11, 28, 19, 7], respectively. Among them, [8, 15, 37] are designed for JPEG compression artifact removal and can be adopted for videos by restoring each frame individually. [6, 33, 14] are proposed based on the fact that I/P/B frames are compressed with different strategies and should be restored by individual models. These methods utilize a single frame as input but neglect temporal dependency with neighboring frames. To remedy this drawback, [34, 11] exploit two motion-compensated nearest peak-

*Corresponding author.
†Y. Xu, M. Zhao, J. Liu are co-first authors of the paper.

quality frames (PQFs) as reference frames, [18, 19] develop the deep Kalman filter network and capture spatiotemporal information from preceding frames, and [28, 7] employ the non-local ConvLSTM and deformable convolution respectively, to capture dependency among multiple neighboring frames. However, using only preceding frames omits the information from the followings; restoration with a pair of nearby PQFs leads to the missing of high-quality details from some other frames (as mentioned in [28]). Recent methods [28, 7] circumvent this problem but directly utilize the multiple adjacent frames as reference frames.

This paper is a summary of our method developed for NTIRE 2021 Quality enhancement of heavily compressed videos Challenge. We formulate an effective Reference Frame Proposal (abbreviated as RFP) strategy as an incremental technique equipped for methods incorporating multiple frames in this task. It is natural for RFP to be applied to [28, 7]. Considering that [28] suffers severe computation and memory costs and is hard to be extended to very deep models used for the Challenge, we applied our RFP to another state-of-the-art method STDF [7] during the competition. Besides, as shown in Fig. 1, over-smoothing harms the performance of enhanced frames a lot. Details and textures are almost removed after enhanced by STDF. The over-smoothing phenomenon indicates that high-frequency details are dropped [4, 5, 20, 17], thus we introduce an additional optimization objective based on fast Fourier transformation (FFT) to supervise the learning of frequency domain information. That is, we exploit both spatial and frequency supervision signals to train the model and complement missing details. Empirical experiments show that both the RFP strategy and the FFT loss lead to significant performance improvement, and combining these two techniques can further boost the performance. Moreover, we adopt a very deep Quality Enhancement (QE) module based on [36, 9] in the competition. In summary, the contributions of this work are as follows:

1. We propose an effective Reference Frame Proposal strategy by utilizing the neighboring compressed frames, which can be directly equipped for existing multi-frame approaches.

2. We introduce a loss based on FFT in this task to complement the missing high-frequency details.

3. We adopt an effective architecture for QE module, which can perform superior results with similar FLOPs and be extended to very deep models.

4. We conduct extensive experiments over MFQE 2.0 dataset and achieve state-of-the-art performance. Our solution is the winner of Track 1 and Track 2 - (Fixed QP, Fidelity / Perceptual), and won the 2nd place in Track 3 - (Fixed bit-rate Fidelity) of **NTIRE 2021 Quality enhancement of heavily compressed videos Challenge**.

2. Related Work

In this section, we review the related work of compression artifact reduction based on deep-learning techniques. Following the success of deep learning on ImageNet [21], many methods with neural networks have been proposed in this long-standing low-level task. According to the utilization of domain knowledge and the number of input frames, existing methods can be categorized into three groups: image-based approaches, single-frame, and multi-frame approaches, respectively.

Image-based Approaches. These approaches are proposed for image compression artifact removal [8, 37, 15, 10, 35, 4, 13, 16, 39]. When applied to the compressed videos, these methods are fed with a single frame and enhance it without knowledge of the video compression algorithms. For example, ARCNN [8] is the first work proposed for reducing JPEG compression artifacts. There are four convolutional layers without any pooling or fully connected layers. DnCNN [37] is another typical method that exploits deeper networks with batch normalization and residual learning. More recently, [35, 4] enhance visual quality via wavelet/frequency domain information. [16, 39] utilize the non-local mechanism for restoration in low-level tasks.

Single-frame Approaches. Some of such approaches [25, 6, 33, 32] employ knowledge of different coding modes in video compression algorithms, *e.g.* I/P/B frames. However, these methods omit the temporal information in frame sequence, and they are ineffective in handling some kinds of temporal noise, such as mosquito noise, edge floating, and flickering. Specifically, DS-CNN [33] and QE-CNN [32] were proposed with two independent models, and they are responsible for intra coding and inter coding modes, respectively.

Multi-frame Approaches. [18, 19] model this vision task as a Kalman filtering procedure, enhancing the frame sequence recursively and capturing temporal information from enhanced preceding frames. [18, 19] further incorporate quantized prediction residual in compressed code streams as strong prior knowledge. However, exploiting temporal information from only preceding frames is incomplete because B frames are compressed via preceding and following frames. Given that the quality of compressed frames in videos fluctuates dramatically, [34, 11] proposed MFQE to build temporal dependency with nearby higher-quality frames. In the MFQE series methods, a classifier is first employed for detecting PQFs, then PQFs are enhanced without reference frames, while non-PQFs take these PQFs as reference frames, compensate reference frames with optical flow and utilize a slow-fused strategy to capture spatial

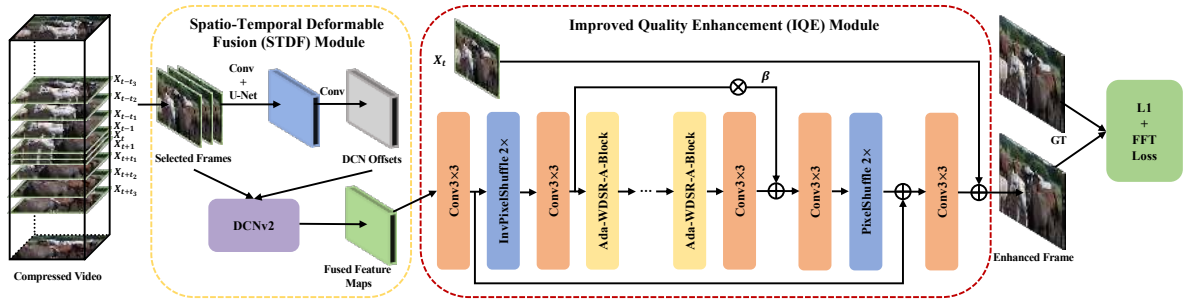


Figure 2. The framework of our method.

and temporal information from PQFs. Later, [29] was proposed with a modified convolutional LSTM. Due to the limitation of motion flow and the observation that high-quality patch also exists in nearby low-quality frames, [28, 7] utilize the non-local mechanism or deformable convolutional network for capturing spatiotemporal dependency in multiple adjacent frames.

Difference between Our Method and the Existing Multi-frame Ones. Mining spatiotemporal information from multiple frames becomes a trend for the quality enhancement of compressed videos. However, the state-of-the-art methods select reference frames in a naive form. In our method, a guidance technique is introduced for reference frame proposals in the preliminary step. Besides, to remedy the over-smoothing phenomenon in this task, an additional loss based on FFT is developed to help recover high-frequency details. Furthermore, we utilize a very deep model based on [36, 9] in the QE module.

3. Method

As for multi-frame approaches, most of them can be concluded as three essential components: *Reference Frame Proposal* (RFP) module, *Spatio-Temporal Feature Fusion* (STFF) module, and *Quality Enhancement* (QE) module. Recently, multi-frame approaches focus on improving the STFF module but still employ a naive reference frame proposal strategy in the RFP module. Thus, in this paper, we pay more attention to the other modules and loss function.

3.1. Reference Frame Proposal

The goal of video compression artifact reduction is to produce a high-quality frame $\hat{Y}_t$ from a compressed frame X_t of the original frame (the ground truth) Y_t, where $X_t \in \mathbb{R}^{C\times H\times W}$, C is the number of channels of a single frame, H and W are the width and height of input videos. In the RFP module, we need to select $2R$ frames from the compressed sequence $X = \{X_1, X_2, \cdots, X_t, \cdots, X_T\}$ as reference frames $\{X_{t+t_1}, \cdots, X_{t+t_{2R}}\}$ for the target frame X_t. Here, the first R frames $\{X_{t+t_1}, \cdots, X_{t+t_R}\}$ are the preceding

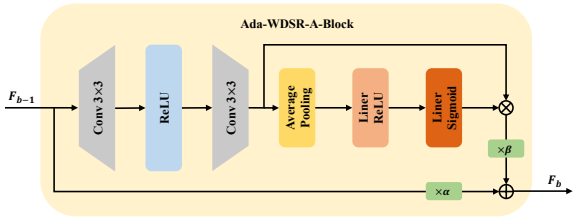


Figure 3. The architecture of Ada-WDSR-A-Block.

frames of the target frame X_t, $\{X_{t+t_{R+1}}, \cdots, X_{t+t_{2R}}\}$ are the following frames, and R is the number of reference frames in one direction. For the sake of simplicity, we take the preceding frames as example in the following.

Assume that $\{X_{t+t_1}, ..., X_{t+t_R}\}$ is an ordered sequence, and $t_1 < \cdots < t_R < 0$. Then, the rules of RFP can be described as follows:

1) As the preliminary step, we first extract the metadata from the HEVC bit-stream with HM-Decoder. The encoder configurations in Track 1/2 and Track 3 are different. Thus, we obtain the candidate frames for RFP with different metadata from bit-stream. In Track 1/2, we set the frame whose QP score is lower than that of the two adjacent frames as candidate frame. While in Track 3, all I/P frames are regarded as candidate frames.

2) We fixedly select adjacent frame X_{t-1} of X_t as the first reference frame by setting $t_R = -1$.

3) We recursively take the next preceding candidate frame of the last selected reference frame as a new reference frame until there are R reference frames or no candidate frames are left.

4) If there is no more candidate frame and the number of selected reference frames is smaller than R, then repeatedly pad it with the last selected frame until there are R frames.

3.2. Improved Quality Enhancement Module

The gist of the QE module is to take the fused feature from the STFF module (*i.e.*, the STDF module in Fig. 2) as input and produce a residual, which is used together with the compressed frame to reconstruct the enhanced frame. Apart from the STFF module, the QE module is another

critical factor for artifact reduction because it needs to explore spatiotemporal information and complement introduced artifacts. For a fair comparison, we employ the same QE module on the benchmark dataset MFQE 2.0. Furthermore, we adopt an Improved Quality Enhancement (IQE) module, which is extended to a very deep version to achieve better results in the challenge.

The framework of the IQE module is shown in Fig. 2. First, we utilize a global residual connection between the head and tail convolutional layers. Parallel to the global connection, the architecture consists of a down-sample module, a deep backbone, and an up-sample module. Among them, the down-sample module is composed of an inverse pixel shuffle layer [22] and a convolutional layer to decrease the spatial resolution, and the up-sample module utilizes an architecture contrast to that of the down-sample module. Between them, there is a skip connection with fixed residual scale $\beta = 0.2$ and a stack of Adaptive WDSR-A-Blocks [36] (Ada-WDSR-A-Block) followed by a convolutional layer.

The Ada-WDSR-A-Blocks are utilized to explore the complementary information for the compressed frame in our paper. The structure of Ada-WDSR-A-Block is illustrated in Fig. 3. For all Ada-WDSR-Blocks, scale r is set to 4. Comparing with WDSR-A-Blocks, there are two additional learnable parameters α and β in Ada-WDSR-Block, which are initialized with 1 and 0.2, respectively. Additionally, we deploy a channel-attention layer [38] before re-scaling the body stream in Ada-WDSR-A-Blocks.

3.3. Fast Fourier Transformation loss

To complement the missing high-frequency details caused by over-smoothing, we introduce a novel supervision signal based on fast Fourier transformation as a complementary loss. Concretely, we apply the fast Fourier transformation to both the ground truth Y_t and the prediction of the QE module, and then employ $L1$ loss on both amplitude and phase of them. The amplitude $A(\cdot)$ and the phase $P(\cdot)$ of a given frame X are calculated as follows:

$$\begin{aligned} X^f(u,v) &= \sum_{x}^{H}\sum_{y}^{W} X(x,y) e^{-j\,2\pi(\frac{u*x}{H}+\frac{v*y}{W})}, \\ A(X) &= \sqrt{Re(X^f)^2 + Im(X^f)^2}, \\ P(X) &= atan(Im(X^f)/Re(X^f)), \end{aligned} \tag{1}$$

where $X(x,y)$ denotes the value at spatial position (x,y), $Re(\cdot)$ and $Im(\cdot)$ are the real and imaginary parts of X^f. Accordingly, we have the following loss function:

$$\mathcal{L}_{FFT} = \|A(\hat{Y}) - A(Y)\|_2 + \lambda \|P(\hat{Y}) - P(Y)\|_2, \tag{2}$$

where λ is a trade-off hyper-parameter between amplitude and phase ($\lambda = 1$ in our implementation). With the FFT loss as a complementary supervision signal, our model is more powerful in high-frequency detail recovery.

4. Experiments

Actually, our techniques can be used in most multi-frame approaches. Here, we take the state-of-the-art STDF [7] as an example for evaluating our techniques. We conduct extensive experiments on the MFQE 2.0 dataset and the dataset provided by the competition. Our evaluation consists of three parts: 1) Ablation study on NTIRE 2021 Dataset [30]; 2) Comparison with state-of-the-art methods on MFQE 2.0 dataset [11] with five QPs; 3) Performance of our method on three tracks in NTIRE 2021 [31].

4.1. Datasets and Settings

MFQE 2.0 dataset [11]. It consists of 126 video sequences collected from Xiph.org [27], VQEG [24] and JCT-VC [3]. Resolutions of these video sequences vary from 352×240 to 2560×1680. For a fair comparison, we follow the settings in [11, 7]: 108 of them are taken for training and the remaining 18 for testing. All sequences are encoded in HEVC Low-Delay-P (LDP) configuration, using HM 16.20 with QP=22, 27, 32, 37 and 42.

NTIRE 2021 Dataset [30]. It is provided in NTIRE 2021 Quality enhancement of heavily compressed videos Challenge [31]. There are 200 videos for training in the competition, 20 for validation, and 20 for the final test. However, only videos in the training data are provided with uncompressed videos. Thus, in this paper, we split 200 videos into two parts: 190 for training and 10 for validation. Sequences are encoded in HEVC LDP configuration with QP=37 in Track 1 and 2, and encoded by FFmpeg supported with libx265 with fixed bit-rate 200kbps in Track 3. Our ablation study in Sec. 4.3 is evaluated with the settings in Track 1.

4.2. Implementation Details

In this paper, we take the state-of-the-art method STDF [7] as our baseline and conduct experiments by following the scheme of STDF. To achieve similar FLOPs of the IQE module to that of the QE module (R3L in STDF), we implement a shallow IQE module with 30 Ada-WDSR-A-Blocks, features in Ada-WDSR-Blocks are implemented with {32, 128, 32} channels in Sec. 4.3 and Sec. 4.4. For all datasets, models are trained by the Adam optimizer with an initial learning rate of 10^{-4}, which is decreased by half when 60% and 90% iterations are finished.

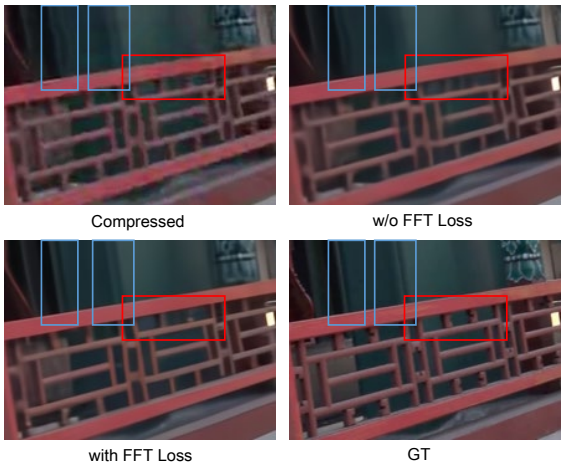
4.3. Ablation Study

As mentioned in Sec. 4.1, experiments of ablation study in this paper are conducted on the dataset from NTIRE 2021 with settings in Track 1. Experimental results included in

Table 1. Ablation study of our method at QP=37 over STDF. Experiments are presented with ΔPSNR (dB) and ΔSSIM ($\times 10^{-2}$) on validation sequences from NTIRE2021.

RFP	L1	L2	FFT	PSNR / SSIM
-	-	✓	-	0.72 / 1.572
✓	-	✓	-	0.74 / 1.607
-	✓	-	-	0.68 / 1.497
-	-	✓	✓	0.74 / 1.581
-	✓	-	✓	0.74 / 1.610
✓	✓	-	✓	**0.76 / 1.639**

(a) Visual case of RFP



(b) Visual case of FFT loss

Figure 4. Visual examples of ablation study. a) Results of RFP. Frames with the yellow box are the reference frames used by the original STDF. Frames with the red box are the references proposed by RFP. Reference frames from RFP provide additional details of the tree to enhance the compressed frame. b) Results of the FFT loss. Improvement by the FFT loss in visual perspective is bounded with rectangles.

Tab. 1 are STDF and STDF with our techniques. Among them, results with $L2$ loss in the second row of Tab. 1 is the performance of baseline STDF. As listed in Tab. 1, all experiments except that for the loss function and RFP strategy follow the same setting as STDF.

Effect of reference frame proposal. Here, we evaluate the effectiveness of utilizing our RFP strategy. First, we compare the performance between STDF (using $L2$ loss) and STDF with our RFP strategy (using RFP + $L2$ loss). The results in the 2nd and 3rd row in Tab. 1 show that utilization of RFP in STDF can improve the performance effectively. Visual examples in Fig. 4(a) also show that utilizing RFP brings benefit by learning missing details from adjacent $2R$ frames. Then, we further verify the effectiveness of RFP on the model trained with $L1$ and FFT loss. As shown in the last row in Tab. 1, PSNR/SSIM achieves improvement over that with $L1$ and FFT loss (in the 6th row). A similar conclusion can also be obtained on the MFQE 2.0 dataset. Thus, the results indicate that utilizing RFP makes the model achieve superior performance of restoration.

Effect of FFT loss. Considering that $L1$ loss achieves better performance than $L2$ loss in recently proposed low-level methods (*e.g.* [36, 38]) for the super-resolution task, we investigate the combination of $L1$/$L2$ loss and FFT loss to evaluate the effectiveness of FFT loss. Different from the conclusion in [36, 38], models trained with $L2$ loss achieve better performance than $L1$ loss in the 2nd and 4th rows in Tab. 1. However, the combination of $L1$ loss and FFT loss (in the 6th row) achieves a better result than the combination of $L2$ and FFT loss (in the 5th row). Besides the example illustrated in Fig. 1, we present additional visual examples in Fig. 4(b) for a further validation on the FFT loss.

4.4. Comparison with State-of-the-art Methods

To demonstrate the advantage of our method, we compare the performance of our method and state-of-the-art approaches, including image-based [8, 37, 15], singe-frame [25, 32] and multi-frame approaches [34, 11, 7]. For a fair comparison, our model is trained by following the training scheme of STDF. Results of video quality enhancement methods are cited from [11, 7].

Overall enhancement. Results of PSNR / SSIM improvement are presented in Tab. 2. Here, *same QE* in Tab. 2 indicates the model follows the same architecture of the QE module in STDF, which means that the differences between *same QE* and STDF are the REP strategy and the FFT loss. The improvement of *same QE* over STDF can be regarded as the benefit from the REP strategy and the FFT loss. Meanwhile, the variant *IQE* denotes the model with the improved QE module introduced in Sec. 3.2, from which a deeper version is designed as the final architecture used by us in the competition.

From Tab. 2, we can see that all multi-frame approaches outperform the methods for images or single frames due to the benefit of utilizing spatiotemporal information. Moreover, the fact that STDF with an effective RFP strategy and FFT loss achieves better results than all the existing methods further proves the importance of filtering input information and the limitation of the $L2$ loss function. Moreover, *IQE* further improves the performance on the benchmark

Table 2. Overall comparison for ΔPSNR (dB) and ΔSSIM ($\times 10^{-2}$) over test sequences at five QPs.

QP	Approach		AR-CNN [8]	DnCNN [37]	Li *et al.* [15]	DCAD [25]	DS-CNN [32]	MFQE 1.0 [34]	MFQE 2.0 [11]	STDF [7]	Ours *same QE*	Ours *IQE*
	Metrics		PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM
37	A	*Traffic*	0.24 / 0.47	0.24 / 0.57	0.29 / 0.60	0.31 / 0.67	0.29 / 0.60	0.50 / 0.90	0.59 / 1.02	0.73 / 1.15	0.71 / **1.18**	**0.96 / 1.50**
		PeopleOnStreet	0.35 / 0.75	0.41 / 0.82	0.48 / 0.92	0.50 / 0.95	0.42 / 0.85	0.80 / 1.37	0.92 / 1.57	1.25 / 1.96	**1.30 / 2.09**	**1.60 / 2.42**
	B	*Kimono*	0.22 / 0.65	0.24 / 0.75	0.28 / 0.78	0.28 / 0.78	0.25 / 0.75	0.50 / 1.13	0.55 / 1.18	0.85 / 1.61	**0.98 / 1.85**	**1.09 / 2.01**
		ParkScene	0.14 / 0.38	0.14 / 0.50	0.15 / 0.48	0.16 / 0.50	0.15 / 0.50	0.39 / 1.03	0.46 / 1.23	0.59 / 1.47	**0.62 / 1.58**	**0.79 / 2.00**
		Cactus	0.19 / 0.38	0.20 / 0.48	0.23 / 0.58	0.26 / 0.58	0.24 / 0.58	0.44 / 0.88	0.50 / 1.00	0.77 / 1.38	0.76 / **1.44**	**0.79 / 1.64**
		BQTerrace	0.20 / 0.28	0.20 / 0.38	0.25 / 0.48	0.28 / 0.50	0.26 / 0.48	0.27 / 0.48	0.40 / 0.67	0.63 / 1.06	**0.65 / 1.08**	**0.67 / 1.16**
		BasketballDrive	0.23 / 0.55	0.25 / 0.58	0.30 / 0.68	0.31 / 0.68	0.28 / 0.65	0.41 / 0.80	0.47 / 0.83	0.75 / 1.23	**0.86 / 1.43**	**0.91 / 1.90**
	C	*RaceHorses*	0.22 / 0.43	0.25 / 0.65	0.28 / 0.65	0.28 / 0.65	0.27 / 0.63	0.34 / 0.55	0.39 / 0.80	0.55 / 1.35	0.55 / 1.34	**0.58 / 1.61**
		BQMall	0.28 / 0.68	0.28 / 0.68	0.33 / 0.88	0.34 / 0.88	0.33 / 0.80	0.51 / 1.03	0.62 / 1.20	0.99 / 1.80	**1.08 / 2.00**	**1.25 / 2.26**
		PartyScene	0.11 / 0.38	0.13 / 0.48	0.13 / 0.45	0.16 / 0.48	0.17 / 0.58	0.22 / 0.73	0.36 / 1.18	0.68 / 1.94	0.67 / 1.91	**0.83 / 2.37**
		BasketballDrill	0.25 / 0.58	0.33 / 0.68	0.38 / 0.88	0.39 / 0.78	0.35 / 0.68	0.48 / 0.90	0.58 / 1.20	0.79 / 1.49	**0.82 / 1.51**	**0.91 / 1.90**
	D	*RaceHorses*	0.27 / 0.55	0.31 / 0.73	0.33 / 0.83	0.34 / 0.83	0.32 / 0.75	0.51 / 1.13	0.59 / 1.43	0.83 / 2.08	**0.86 / 2.15**	**0.95 / 2.42**
		BQSquare	0.08 / 0.08	0.13 / 0.18	0.09 / 0.25	0.20 / 0.38	0.20 / 0.38	-0.01 / 0.15	0.34 / 0.65	0.94 / 1.25	0.72 / 1.03	**1.28 / 1.86**
		BlowingBubbles	0.16 / 0.35	0.18 / 0.58	0.21 / 0.68	0.22 / 0.65	0.23 / 0.68	0.39 / 1.20	0.53 / 1.70	0.74 / 2.26	0.72 / 2.21	**0.91 / 2.88**
		BasketballPass	0.26 / 0.58	0.31 / 0.75	0.34 / 0.85	0.35 / 0.85	0.34 / 0.78	0.63 / 1.38	0.73 / 1.55	1.08 / 2.12	**1.12 / 2.23**	**1.29 / 2.65**
	E	*FourPeople*	0.37 / 0.50	0.39 / 0.60	0.45 / 0.70	0.51 / 0.78	0.46 / 0.70	0.66 / 0.85	0.73 / 0.95	0.94 / 1.17	**1.00 / 1.28**	**1.24 / 1.50**
		Johnny	0.25 / 0.10	0.32 / 0.40	0.40 / 0.60	0.41 / 0.50	0.38 / 0.40	0.55 / 0.55	0.60 / 0.68	0.81 / 0.88	**0.84 / 0.96**	**1.02 / 1.15**
		KristenAndSara	0.41 / 0.50	0.42 / 0.60	0.49 / 0.68	0.52 / 0.70	0.48 / 0.60	0.66 / 0.75	0.75 / 0.85	0.97 / 0.96	**1.03 / 1.09**	**1.23 / 1.23**
	Average		0.23 / 0.45	0.26 / 0.58	0.30 / 0.66	0.32 / 0.67	0.30 / 0.63	0.46 / 0.88	0.56 / 1.09	0.83 / 1.51	**0.85 / 1.58**	**1.03 / 1.90**
42	Average		0.29 / 0.96	0.22 / 0.77	0.32 / 1.05	0.32 / 1.09	0.31 / 1.01	0.44 / 1.30	0.59 / 1.65	– / –	**0.79 / 2.18**	**0.89 / 2.41**
32	Average		0.18 / 0.19	0.26 / 0.35	0.28 / 0.37	0.32 / 0.44	0.27 / 0.38	0.43 / 0.58	0.516 / 0.68	0.86 / 1.04	**0.93 / 1.16**	**1.08 / 1.36**
27	Average		0.18 / 0.14	0.27 / 0.24	0.30 / 0.28	0.32 / 0.30	0.27 / 0.23	0.40 / 0.34	0.49 / 0.42	0.72 / 0.57	**0.92 / 0.77**	**1.09 / 0.92**
22	Average		0.14 / 0.08	0.29 / 0.18	0.30 / 0.19	0.31 / 0.19	0.25 / 0.15	0.31 / 0.19	0.46 / 0.27	0.63 / 0.34	**0.83 / 0.46**	**0.96 / 0.53**

* Video resolution: Class A (2560 × 1600), Class B (1920 × 1080), Class C (832 × 480), Class D (480 × 240), Class E (1280 × 720).

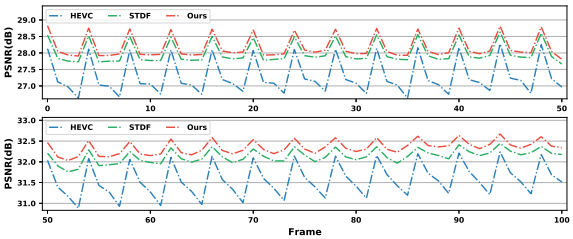


Figure 5. PSNR curves of HEVC baseline, STDF and ours on two test sequences at QP=37. Top: *ParkScene*. Bottom: *PartyScene*.

Table 3. Average PVD/SD of test sequences for PSNR at QP=27, 32, 37 and 42.

Method	27	32	37	42
HEVC	1.07 / 0.83	1.38 / 0.82	1.42 / 0.79	1.21 / 0.74
AR-CNN [8]	1.07 / 0.83	1.38 / 0.82	1.44 / 0.80	1.24 / 0.75
DnCNN [37]	1.06 / 0.83	1.40 / 0.83	1.44 / 0.80	1.24 / 0.75
Li *et al.* [15]	1.06 / 0.83	1.38 / 0.83	1.44 / 0.80	1.24 / 0.76
DCAD [25]	1.07 / 0.83	1.39 / 0.83	1.45 / 0.80	1.26 / 0.76
DS-CNN [32]	1.07 / 0.83	1.39 / 0.83	1.46 / 0.80	1.24 / 0.75
MFQE 1.0 [34]	0.84 / 0.81	1.07 / 0.77	1.05 / 0.73	0.82 / 0.69
MFQE 2.0 [11]	0.77 / 0.74	0.98 / 0.70	0.96 / 0.67	0.74 / 0.62
Ours *same QE*	**0.60 / 0.33**	**0.75 / 0.44**	**0.73 / 0.37**	**0.67 / 0.36**
ours *IQE*	**0.58 / 0.32**	**0.72 / 0.47**	**0.70 / 0.41**	**0.66 / 0.30**

dataset and achieves impressive results of $1.029dB/0.0190$ PSNR/SSIM improvement for QP=37, 23.9% and 25.8% higher than that of STDF. Similar improvements can also be observed for other QPs.

Quality fluctuation. Quality fluctuation is another observable measurement for the overall quality of enhanced videos. Drastic quality fluctuation of frames accounts for severe texture shaking and degradation of the quality of experience (QoE). Therefore, we present two PSNR curves of test sequences compressed by HEVC, the corresponding sequences enhanced by STDF and our method in Fig. 5. Comparing with STDF, our method achieves larger PSNR improvement over the compressed frames, especially for non-PQFs, which means that the quality of frames enhanced by our method fluctuates less than that by STDF. Besides, we also evaluate the fluctuation by Standard Deviation (SD), and Peak-Valley Difference (PVD) of each test sequence as in [28, 11, 34]. Results of PSNR are presented in Tab. 3, and our method still achieves impressive results of SD and PVD, which means that our method performs more stably than the other methods.

Rate-distortion performance. We then evaluate the rate-distortion of our method and compare it with state-of-the-art methods. For a simple illustration, we present only the results of compressed videos, the enhanced results of two state-of-the-art methods (MFQE 2.0 and STDF) and our method in Fig. 7. Here, we do not show the results of STDF at QP=42 due to the lack of data in [7]. From the curves in Fig. 7, we can see that our method performs better than the state-of-the-art approaches in rate-distortion performance. Following the experiments in [11], we also evaluate the BD-bitrate (BD-BR) reduction, which is calcu-

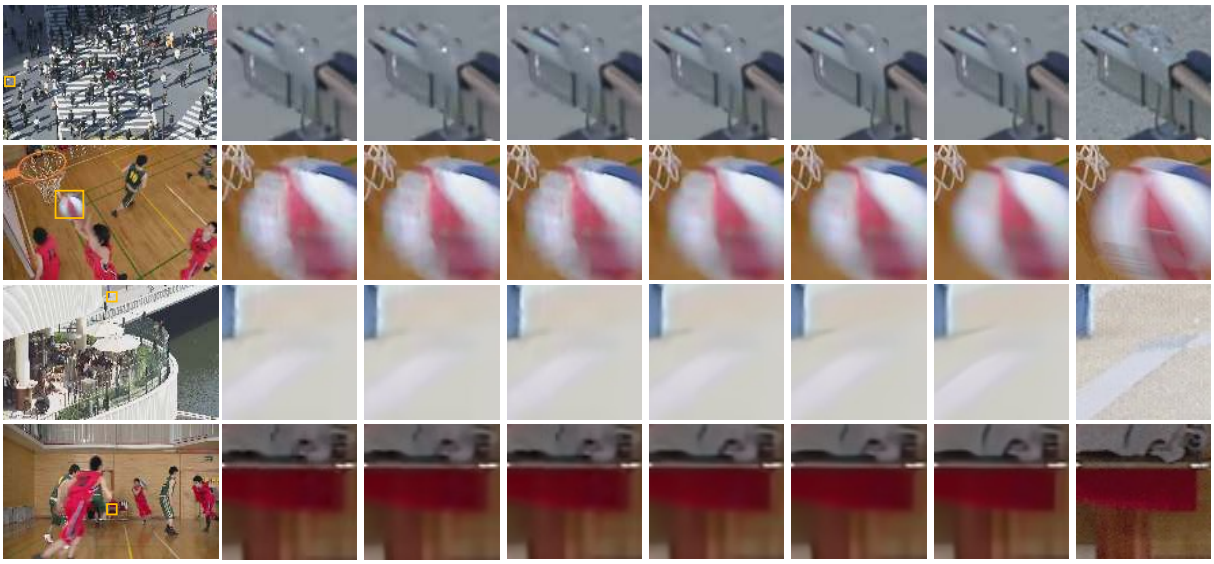


Figure 6. Qualitative examples at QP=37.

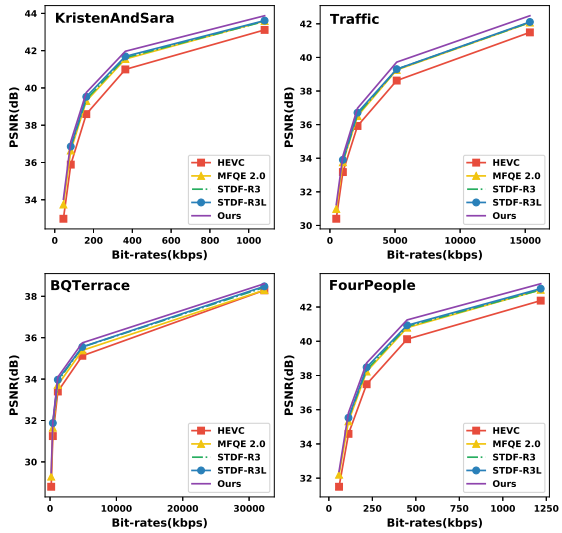


Figure 7. Rate-distortion performance of four test sequences.

lated over five PSNR results at QP=22, 27, 32, 37 and 42, while the result of STDF is obtained with four QPs. Average results of BD-BR reduction for MFQE 2.0, STDF, *same QE* and *IQE* are 14.06%, 20.79%, 22.49% and 25.86%, respectively. These results show the advantage of our techniques, and the methods with our techniques can achieve much better QoE under the same bit rate.

4.5. Qualitative Comparison

We also conduct qualitative comparison and present several visual examples at QP=37 in Fig. 6. We can see that the compressed frames suffer severe compression artifacts (*e.g.* missing vertical stripes, blocking artifact on the basketball). For the existing methods from the third to sixth columns, the enhanced patches are distorted by over-smoothing and temporal noise. However, our method restores much more detail or texture than the other methods. Compared with the baseline STDF, our method restores more details, especially for high-frequency information, such as sharpening edges. This means that by applying the technique introduced in our paper, multi-frame approaches can do restoration better than the original ones.

4.6. NTIRE 2021 Challenge

In the NTIRE 2021 Challenge on Quality enhancement of compressed videos [31], we won Track 1 and Track 2, and were the 2nd in Track 3. Detailed results are included in Tab. 4. Besides the techniques introduced above, the performance also relies on a much deeper IQE module and two ensemble strategies, *i.e.*, self-ensemble and gated fusion.

Deeper IQE module. In the competition implementation, we employ the IQE module with more Ada-WDSR-A-Blocks and wider features. Specifically, the number of channels for feature and blocks of Ada-Blocks in the deeper IQE module are 128 and 96, respectively. Thus, the number of feature channels in Ada-WDSR-A-Block is implemented as $\{64, 256, 64\}$.

Self-ensemble. In the competition, we utilize the self-ensemble strategy [2] that can boost the restoration through multiple trails of inputs with different augmentation operations. Unlike the conventional ensemble strategies that integrate results from multiple models, self-ensemble takes the frames transformed by different augmentation operations, and averages these different but related outputs with

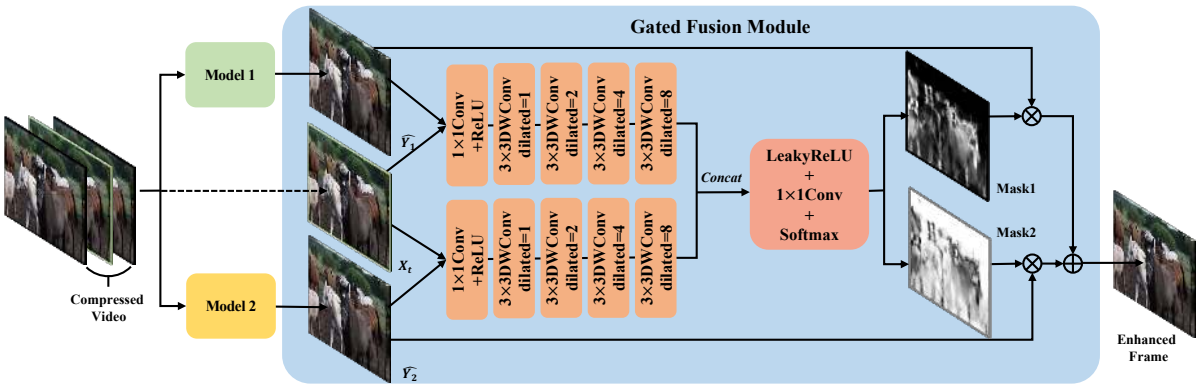


Figure 8. The architecture of gated fusion module.

Table 4. The final testing results of NTIRE Challenge on quality enhancement of heavily compressed videos.

Track 1			Track 2						Track 3		
Ranking	PSNR(dB)	MS-SSIM	Ranking	MOS↑	LPIPS↓	FID↓	VID↓	VMAF↑	Ranking	PSNR (dB)	MS-SSIM
1 (**Ours**)	**32.52**	**0.9562**	1 (**Ours**)	**71**	**0.0429**	**32.17**	**0.0137**	75.69	1	**30.37**	**0.9484**
2	32.49	0.9552	2	69	0.0483	34.64	0.0179	71.55	2 (**Ours**)	29.95	0.9468
3	32.04	0.9493	3	67	0.0561	46.39	0.0288	68.92	3	29.69	0.9423
4	31.90	0.9480	4	63	0.0561	50.61	0.0332	69.06	4	29.64	0.9405
5	31.86	0.9472	5	60	0.1018	72.27	0.0561	**78.64**	5	29.56	0.9403

the original to obtain the final predictions. In the competition, eight augmentation operations are exploited for evaluation. Empirically, experimental results on the validation dataset in Track 1 show that STDF with basic IQE module (shallow model) and deeper IQE module (deep model) can achieve 0.2 and 0.12 dB PSNR improvement by utilizing self-ensemble.

Gated Fusion module. Due to the limited official training data provided by the competition, models trained with only these data are easily dominated by the bias of training data. Meanwhile, limited clips mean rare scenes to be seen, but many unseen patterns may appear in inference, which restricts the performance of the model. However, directly using large-scale data collected by ourselves will destroy the original distribution of training data. To minimize the offset between the two datasets and gain benefit from extra data, we propose a novel module to improve the performance of enhancement at the bottom of the pipeline. As illustrated in Fig. 8, though each model has the same architecture (STDF with deeper IQE), one is trained on the official training sets, and the other is on the extra videos crawled from Bilibili and YouTube, named as BiliTube4k. Inspired by [23], we exploit a stack of layers to output the mask and aggregate the predictions of two models via produced mask. As shown in the middle of Fig. 8, the mask M in gated fusion module is of the same resolution of the target frame ranging from $[0, 1]$. Thus, the output of gated fusion module can be formulated as $\hat{Y} = M \bigotimes \hat{Y}_1 \bigoplus (1 - M) \bigotimes \hat{Y}_2$. The detail of network architecture can be referred to Fig. 8. Furthermore, such a structure of gated fusion module can be used in more models.

Other details. In Track 3, we utilize the model pre-trained in Track 1 as the pre-trained model, and then fine-tune it on training data of Track 3 with early stopping. As for Track 2, we reuse and freeze the models from Track 1, and attach ESRGAN [26] at the bottom of them. Specifically, we use the ESRGAN pre-trained on DIV2K dataset [1], remove the pixel shuffle layer, and employ the FFT loss. Then, two ESRGANs trained on different datasets are integrated with the gated fusion module to produce the final enhanced frames.

5. Conclusion

In this paper, we present a method for improving existing multi-frame approaches in video compression artifact reduction via integrating multiple frames and frequency domain information. Our method was developed for the NTIRE 2021 Challenge on Quality enhancement of heavily compressed videos Challenge, and won Track 1 and Track 2, and the 2nd place in Track 3. Through extensive experiments, we show that both our proposed reference frame proposal strategy and the FFT loss can achieve superior performance over state-of-the-art methods. In the future, more verification of our techniques is expected to be conducted on other multi-frame approaches.

6. Acknowledgement

Yi Xu, Minyi Zhao and Shuigeng Zhou were supported by 2019 Special Fund for Artificial Intelligence Innovation & Development, Shanghai Economy and Information Technology Commission (SHEITC), and partially by Science and Technology Commission of Shanghai Municipality Project (No. 19511120700).

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pages 126–135, 2017. 8

[2] Byeongyong Ahn, Gu Yong Park, Yoonsik Kim, and Nam Ik Cho. A self-ensemble approach for noise and compression artifacts removal using convolutional neural network. *IEIE Transactions on Smart Processing & Computing*, 7(4):296–304, 2018. 7

[3] Frank Bossen. Common test conditions and software reference configurations. In *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, 5th meeting, Jan. 2011*, 2011. 4

[4] Honggang Chen, Xiaohai He, Linbo Qing, Shuhua Xiong, and Truong Q Nguyen. DPW-SDNet: Dual pixel-wavelet domain deep cnns for soft decoding of JPEG-compressed images. In *CVPRW*, pages 711–720, 2018. 1, 2

[5] Kai Cui and Eckehard G Steinbach. Decoder side color image quality enhancement using a wavelet transform based 3-stage convolutional neural network. In *CVPRW*, page 0, 2019. 2

[6] Yuanying Dai, Dong Liu, and Feng Wu. A convolutional neural network approach for post-processing in HEVC intra coding. In *International Conference on Multimedia Modeling*, pages 28–39. Springer, 2017. 1, 2

[7] Jianing Deng, Li Wang, Shiliang Pu, and Cheng Zhuo. Spatio-temporal deformable convolution for compressed video quality enhancement. In *AAAI*, volume 34, pages 10696–10703, 2020. 1, 2, 3, 4, 5, 6

[8] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *ICCV*, pages 576–584, 2015. 1, 2, 5, 6

[9] Dario Fuoli, Zhiwu Huang, Martin Danelljan, and Radu Timofte. NTIRE 2020 challenge on video quality mapping: Methods and results. In *CVPRW*, June 2020. 2, 3

[10] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. Deep generative adversarial compression artifact removal. In *ICCV*, pages 4836–4845. IEEE, 2017. 2

[11] Zhenyu Guan, Qunliang Xing, Mai Xu, Ren Yang, Tie Liu, and Zulin Wang. MFQE 2.0: A new approach for multi-frame quality enhancement on compressed video. *TPAMI*, 2019. 1, 2, 4, 5, 6

[12] Jun Guo and Hongyang Chao. Building dual-domain representations for compression artifacts reduction. In *ECCV*, pages 628–644. Springer, 2016. 1

[13] Jun Guo and Hongyang Chao. One-to-many network for visually pleasing compression artifacts reduction. In *CVPR*, pages 4867–4876, 2017. 2

[14] Zhipeng Jin, Ping An, Chao Yang, and Liquan Shen. Quality enhancement for intra frame coding via cnns: An adversarial approach. In *ICASSP*, pages 1368–1372. IEEE, 2018. 1

[15] Ke Li, Bahetiyaer Bare, and Bo Yan. An efficient deep convolutional neural networks model for compressed image deblocking. In *ICME*, pages 1320–1325. IEEE, 2017. 1, 2, 5, 6

[16] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *NeurIPS*, pages 1680–1689, 2018. 2

[17] Jing Liu, Haiyan Wu, Yuan Xie, Yanyun Qu, and Lizhuang Ma. Trident dehazing network. In *CVPRW*, pages 430–431, 2020. 2

[18] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Zhiyong Gao, and Ming-Ting Sun. Deep kalman filtering network for video compression artifact reduction. In *ECCV*, pages 568–584, 2018. 1, 2

[19] Guo Lu, Xiaoyun Zhang, Wanli Ouyang, Dong Xu, Li Chen, and Zhiyong Gao. Deep non-local kalman network for video compression artifact reduction. *TIP*, 29:1725–1737, 2019. 1, 2

[20] Xiaotong Luo, Jiangtao Zhang, Ming Hong, Yanyun Qu, Yuan Xie, and Cuihua Li. Deep wavelet network with domain adaptation for single image demoireing. In *CVPRW*, pages 420–421, 2020. 2

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 2

[22] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, pages 1874–1883, 2016. 4

[23] Kai Tian, Yi Xu, Shuigeng Zhou, and Jihong Guan. Versatile multiple choice learning and its application to vision computing. In *CVPR*, pages 6349–6357, 2019. 8

[24] VQEG. VQEG video datasets and organizations. https://www.its.bldrdoc.gov/vqeg/video-datasets-and-organizations.aspx. 4

[25] Tingting Wang, Mingjin Chen, and Hongyang Chao. A novel deep learning-based method of improving coding efficiency from the decoder-end for HEVC. In *Data Compression Conference, 2017*, pages 410–419. IEEE, 2017. 2, 5, 6

[26] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCVW*, pages 0–0, 2018. 8

[27] Xiph.org. Xiph.org video test media. https://media.xiph.org/video/derf/. 4

[28] Yi Xu, Longwen Gao, Kai Tian, Shuigeng Zhou, and Huyang Sun. Non-local convLSTM for video compression artifact reduction. In *ICCV*, pages 7043–7052, 2019. 1, 2, 3, 6

[29] Ren Yang, Xiaoyan Sun, Mai Xu, and Wenjun Zeng. Quality-gated convolutional LSTM for enhancing compressed video. In *ICME*, pages 532–537. IEEE, 2019. 1, 3

[30] Ren Yang and Radu Timofte. NTIRE 2021 challenge on quality enhancement of compressed video: Dataset and study. In *CVPRW*, 2021. 4

[31] Ren Yang, Radu Timofte, et al. NTIRE 2021 challenge on quality enhancement of compressed video: Methods and results. In *CVPRW*, 2021. 4, 7

[32] Ren Yang, Mai Xu, Tie Liu, Zulin Wang, and Zhenyu Guan. Enhancing quality for HEVC compressed videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7):2039–2054, 2018. 2, 5, 6

[33] Ren Yang, Mai Xu, and Zulin Wang. Decoder-side HEVC quality enhancement with scalable convolutional neural network. In *ICME*, pages 817–822. IEEE, 2017. 1, 2

[34] Ren Yang, Mai Xu, Zulin Wang, and Tianyi Li. Multi-frame quality enhancement for compressed video. In *CVPR*, pages 6664–6673, 2018. 1, 2, 5, 6

[35] Jaeyoung Yoo, Sang-ho Lee, and Nojun Kwak. Image restoration by estimating frequency distribution of local patches. In *CVPR*, pages 6684–6692, 2018. 2

[36] Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, and Thomas Huang. Wide activation for efficient and accurate image super-resolution. In *BMVC*, 2019. 2, 3, 4, 5

[37] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP*, 26(7):3142–3155, 2017. 1, 2, 5, 6

[38] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 4, 5

[39] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *ICLR*, 2019. 2