A PARALLEL NETWORK FOR COMPRESSED VIDEO ENHANCEMENT

A PARALLEL NETWORK FOR COMPRESSED VIDEO ENHANCEMENT

BY

WEI HAO, B.Eng.

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

Master of Applied Science (2021) McMaster University

(Electrical and Computer Engineering) Hamilton, Ontario, Canada

TITLE:	A Parallel Network for Compressed Video Enhancement
AUTHOR:	Wei Hao B.Eng. (Electrical and Information Engineering), Beihang University, Beijing, China
SUPERVISOR:	Dr. Jun Chen
NUMBER OF PAGES:	xv, 46

Lay Abstract

The quality of video is improving as cameras improve, but the size of the video is also increasing. As a result, we will need to compress the video. Video compression, on the other hand, is always accompanied with a loss of video quality. Deep learning approaches have made tremendous progress in improving the quality of compressed video in recent years. In this paper, we propose an effective method PEN for Video Quality Enhancement(VQE) task by parallel processing of multiple frames.

Abstract

Recent years, we have witnessed significant progress in the quality enhancement of compressed video by deep learning methods. In this paper, we propose an effective method for Video Quality Enhancement(VQE) task. Our method is realized via **A Parallel Network for Compressed Video Enhancement(PEN)**. To tackle optical flow estimates and complicated motion, PEN has two branches which are **Offset Deformable Fusion Network(ODFN)** and **Complex Motion Solution Network(CMSN)**. During the alignment stage, existing methods typically estimate optical flow for temporal motion compensation. However, because the compressed video may be severely distorted as a result of various compression artifacts, the estimated optical flow is typically inaccurate and unreliable. Therefore in ODFN we use deformable convolution to align frames in a fast and efficient way. At the same time, we adopt pyramidal processing and cascading refinement in CMSN which can address complex motions and large parallax problems in alignment. Furthermore, we use the target frame's neighbor Peak Quality frames(PQFs) as reference frames, which adjusts for video quality variations. Extensive experiments show that our method has improved the average video quality by 0.7 decibel.

To the past and future

Acknowledgements

This thesis would not be achievable without a very dedicated group of people.

First and foremost, I'd like to convey my gratitude to my supervisor, Dr. Jun Chen. Thank you for giving me the opportunity to study and for guiding me through two years of education. I learned how to think about and solve difficulties under his influence. He instilled in me a rigorous and logical attitude to problem-solving.

Next, I would like to sincerely thank my committee members, Dr. Sorina Dumitrescuand and Dr. Xiaolin Wu for their valuable suggestions and kind comments.

Then I'd like to express my appreciation to Yankun Yu and Minghan Fu, two of my closest pals. We shared two joyful and fulfilling years together, studying and having fun together, and they were there for me in both happy and sad times.

I owe Kangdi Shi a huge debt of gratitude. In Canada, he is my first and best partner. Kangdi provided me with a great deal of advice and assistance over our two years of collaboration. He is constantly eager to respond to my many inquiries. He will switch on the computer, even if it is in the middle of the night, to assist me in solving certain problems.

Finally, I'd like to express my gratitude to my family. Thank you to my parents and grandparents for always being pleased with my progress and ensuring that I am well fed. Thank you to my sister Yimeng, brother-in-law Jaden, and aunt Carrie for their support and presents. Last but not least, I'd like to show my thankfulness to you, my uncle. He is the

individual who has aided me the most throughout my academic career. Despite the fact that he is usually strict with me, it is because of his strict criteria that I was able to study abroad and graduate with honours. His video coding knowledge and advice will help me become a better engineer in the future.

Contents

List of Figures

List of Tables

Notation and Abbreviations

Notation

sigmoid	A sigmoid function is a mathematical function having a characteristic "S"-shaped curve or sigmoid curve.
$A \bigodot B$	element-wise multiplication

Abbreviations

H.264	Advanced Video Coding
HEVC	High Efficiency Video Coding
YUV	Color encoding system typically used as part of a color image pipeline
PNG	Portable Network Graphics
QP	Quantization parameter
QoE	Quality of Experience
SROCC	Spearman Rank-Order Correlation Coefficient

CNN	Convolutional Neural Network
JEPG	Joint Photographic Experts Group
VQE	Video Quality Enhancement
PQF	Peak Quality Frames
VQF	Calley Quality Frames
DCN	Deformable Convolutional Network
PEN	Parallel Enhancement Network
RTF	Regression Tree Fields
IQA	Image Quality Assessment
BRISQUE	Blind/Referenceless Image Spatial QUality Evaluator
PSNR	Peak signal-to-noise ratio
SSIM	structural similarity index measure
MSE	Mean square error
SSE	Sum of Squared Error
ODFN	Offset Deformable Fusion Network
TPAN	Temporal and Spatial Attention Network
PCD	Pyramid,Cascading and Deformable Convolutions
TSA	Temporal and Spatial Attention

Chapter 1

Introduction

The popularity of video on the Internet has increased dramatically in recent years. With limited bandwidth, video compression is necessary to reduce the bit rate. However, compression algorithms, such as H.264/AVC(Wiegand *et al.* (2003)) and H.265/HEVC (Sullivan *et al.* (2012)), frequently create numerous video distortions, especially at low bit rates. As shown in Figure 1.2, such artifacts may considerably jeopardize video quality, resulting in the deteriorated Quality of Experience (QoE). The distorted contents of low-quality compressed video may also impair subsequent vision tasks (e.g., recognition, detection, and tracking) in low-bandwidth applications. As a result, it's critical to increase the compressed video's quality.

In the past few decades, extensive work has been done on artifact removal or quality enhancement of a single compressed image. Earlier research mainly focused on the enhancement of single image quality by traditional methods(Foi *et al.* (2007),Zhang *et al.* (2013),Liew and Hong Yan (2004),Shen *et al.* (2011),Chang *et al.* (2014)). Traditional methods reduced artifacts by optimizing the transform coefficients for specific compression standard, thus they are hard to extend to other compression schemes.

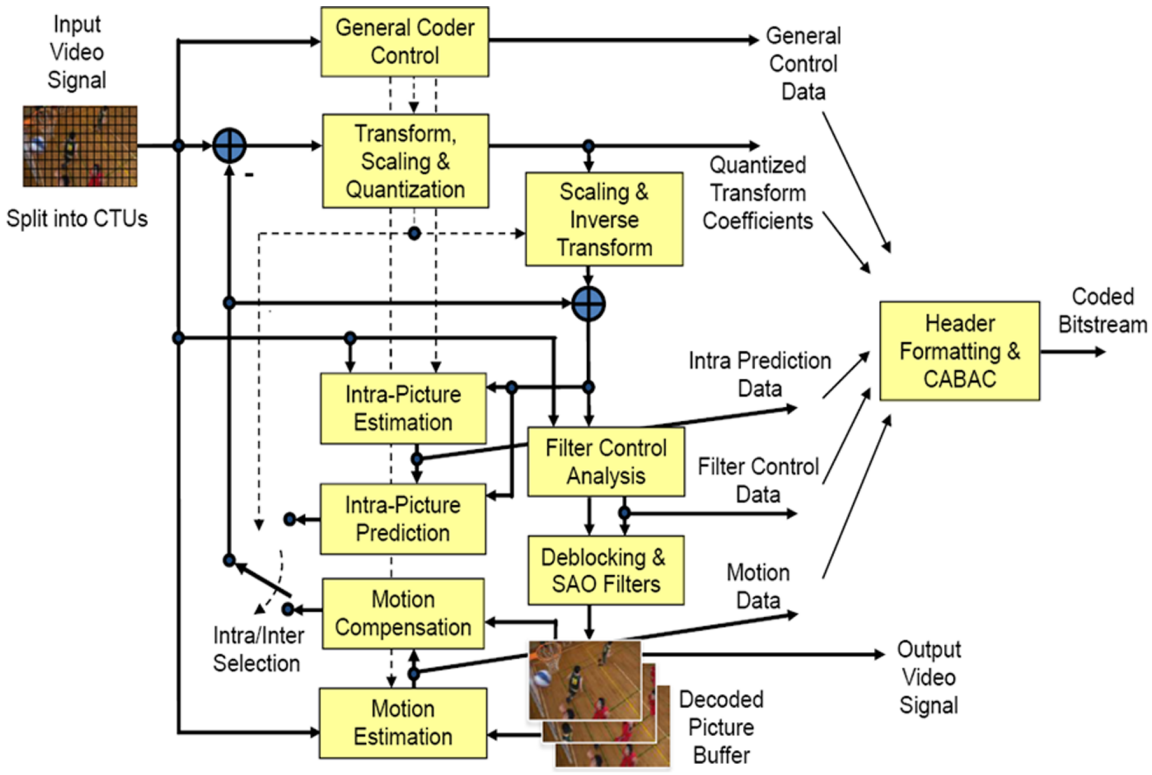

Figure 1.1: Illustration of traditional compression process

With the recent advances in Convolutional Neural Networks (CNNs), CNN-based methods have also emerged for image quality enhancement(Dong *et al.* (2015),Tai *et al.* (2017)). For example, Dong *et al.* (2015) designed a four-layer Convolutional Neural Network (CNN), named AR-CNN, which considerably improves the quality of JPEG images. Deep learning methods usually learn Non-linear mappings that can directly regress images without artifacts from a large amount of training data to obtain impressive results with high efficiency. However, these methods cannot be directly extended to compressed video since they treat frames independently and thus fail to exploit temporal information.

On the other hand, there is only limited study on quality enhancement for compressed

Figure 1.2: Illustration of compression artifacts.

video. Yang et al. first proposed the Multi-Frame Quality Enhancement (MFQE 1.0) approach to leverage temporal information for Video Quality Enhancement(VQE) (Yang *et al.* (2018)). High-quality frames in the compressed video are utilized as reference frames to help enhancing the quality of neighboring low-quality target frames. Recently, an upgraded version MFQE 2.0 (Guan *et al.* (2021)) was introduced to improve the efficiency of MF-CNN further and achieved the state-of-the-art performance. In order to aggregate information from the target frame and reference frames, both MFQE methods adopt a widely used temporal fusion scheme that incorporates dense optical flow for motion compensation(Kappeler *et al.* (2016)).

Deng et al. proposed a fast yet effective method for compressed video quality enhancement by incorporating a novel Spatio-Temporal Deformable Fusion (STDF)(Deng *et al.* (2020)) scheme to aggregate temporal information. Specifically, the proposed STDF takes

a target frame along with its neighboring reference frames as input to jointly predict an offset field to deform the Spatio-Temporal sampling positions of convolution.

In this paper, inspired by STDF(Deng *et al.* (2020)), we introduce a Parallel Enhancement Network of Compressed Video scheme(PEN) for VQE task. The PEN has two branches which are Offset Deformable Fusion Network(ODFN) and Complex Motion Solution Network(CMSN). The first ODFN branch is primarily concerned with flow prediction in the multiframes alignment task. We first attempted optical flow estimation(Dosovitskiy *et al.* (2015)) in our network during the alignment stage. We discovered that optical flow estimation may be suboptimal in the context of the VQE task. Because compression artifacts can heavily distort video content and disrupt pixel-wise distances between frames, the estimated optical flow tends to be inaccurate and unreliable, thereby resulting in ineffective quality enhancement. Optical flow estimation needs to be repeatedly performed for different reference target frame pairs in a pairwise manner, which involves substantially increased computational cost to explore more reference frames. Therefore we followed Deng *et al.* (2020) to use deformable convolutional networks(DCN)(Dai *et al.* (2017)) to aggregate temporal information while avoiding explicit optical flow estimation. Then to address complex motions and large parallax problems in alignment. Inspired by Wang *et al.* (2019), we add a second branch CMSN to improve the network's performance and robustness. The CMSN mainly contains two parts: Pyramid, Cascading and Deformable convolutions (PCD) alignment module at the feature level and the Temporal and Spatial Attention(TSA) fusion module at the image level.

At last, inspired by (Yang *et al.* (2018); Guan *et al.* (2021)), we found that the frame's PSNR varies significantly fluctuated. This indicates that there exists considerable quality fluctuation in compressed video sequences for (MPEG-1, MPEG-2, MPEG-4, H.264/AVC

and HEVC). Due to that, we decided to take the Peak Quality frames(PQFs) as the reference frame instead of neighboring frames. The main contributions of this paper are summarized as follows:

1.An end-to-end CNN-based method was proposed for the VQE task.

2.Compare the proposed PEN to prior fusion schemes analytically and experimentally, and demonstrate its enhanced flexibility and robustness.

Chapter 2

Related Work

2.1 Image and Video Quality Enhancement

Over the past decade, an increasing number of works have focused on quality enhancement for compressed image.(Foi *et al.* (2007),Zhang *et al.* (2013),Liew and Hong Yan (2004),Shen *et al.* (2011),Chang *et al.* (2014),Dong *et al.* (2015),Tai *et al.* (2017),Guo and Chao (2016),Wang *et al.* (2016),Zhang *et al.* (2017),Li *et al.* (2015)). Specifically, Foi *et al.* (2007) applied point-wise Shape-Adaptive DCT (SA-DCT) to reduce the blocking and ringing effects caused by JPEG compression. Later, (Jancsary *et al.* (2012)) proposed the approach of reducing JPEG image blocking effects by adopting Regression Tree Fields (RTF). Recently, deep learning has also been successfully applied to improve the visual quality of compressed images. Particularly, Dong *et al.* (2015) proposed a four-layer AR-CNN to reduce the JPEG artifacts of images. Afterward, D3(Wang *et al.* (2016)) and Deep Dual-domain Convolutional Network (DDCN)(Guo and Chao (2016)) were proposed as advanced deep networks for the quality enhancement of JPEG image, utilizing the prior knowledge of JPEG compression. Later, DnCNN was proposed in (Zhang *et al.* (2017)) for

several tasks of image restoration, including quality enhancement. Li *et al.* (2015) proposed a 20-layer CNN for enhancing image quality. Most recently, the memory network (MemNet) (Tai *et al.* (2017)) has been proposed for image restoration tasks, including quality enhancement. In the MemNet, the memory block was introduced to generate the long-term memory across CNN layers, which successfully compensate the middle and high-frequency signals distorted during compression. It achieves the state-of-the-art quality enhancement performance for compressed images.

On the other hand, MFQE 1.0 (Yang *et al.* (2018)) pioneered the application of multi-frame CNN to take advantage of temporal information for compressed video quality enhancement, where high-quality frames are utilized to enhancing the quality of the adjacent low-quality frames. To exploit long-range temporal information, Yang et al. later introduced a modified convolutional long short-term memory network(Yang *et al.* (2019)) for video quality enhancement. Most recently, Guan *et al.* (2021) proposed MFQE 2.0 to upgrade several key components of MFQE 1.0. Deng *et al.* (2020) proposed STDF in 2020 to further improve the performance of the deep learning method on video quality enhancement. The STDF(Deng *et al.* (2020)) takes a target frame along with its neighboring reference frames as input to jointly predict an offset field to deform the Spatio-Temporal sampling positions of convolution and achieved state-of-the-art performance in terms of accuracy and speed.

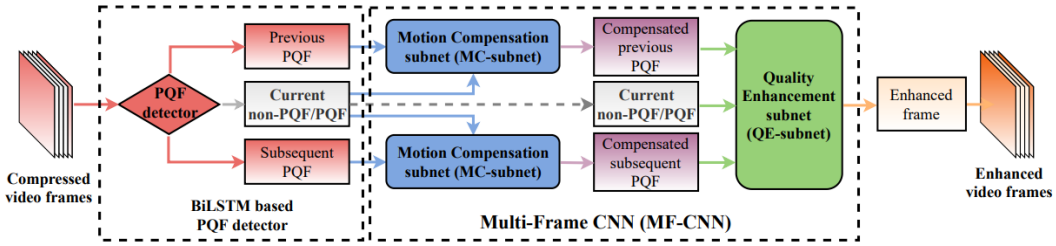

Figure 2.1: The framework of MFQE(source: Guan *et al.* (2021))

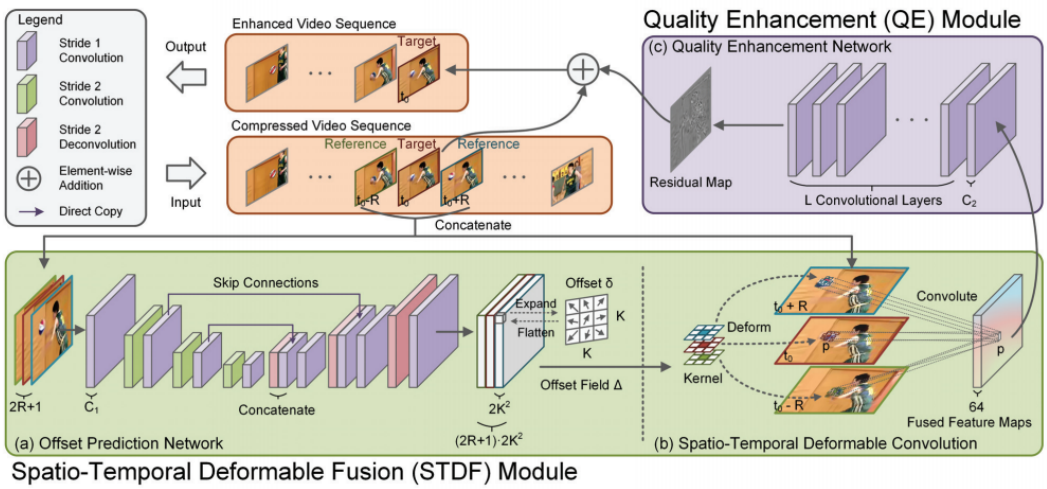

Figure 2.2: The framework of STDF(source:Deng *et al.* (2020))

2.2 Multi-frame Super-Resolution

In the early years, Brandi *et al.* (2008) and Song *et al.* (2011) proposed to enhance video resolution by taking advantage of high-resolution vital frames. Recently, many multi-frame super-resolution approaches have employed deep neural networks. For example, Huang *et al.* (2018) developed a Bidirectional Recurrent Convolutional Network (BRCN), which improves the super-resolution performance over traditional single-frame approaches. Kappeler et al. proposed a Video Super-Resolution network (VSRnet)(Kappeler *et al.* (2016)), in which the neighboring frames are warped according to the estimated motion, and then both the current and warped neighboring frames are fed into a super-resolution CNN to enhance the resolution of the current frame. Later, Li and Wang (2017) proposed replacing VSRnet with a deeper network with residual learning strategy. All these multi-frame methods exceed the limits of single-frame approaches (e.g., SRCNN(Dong *et al.* (2016))) for super-resolution, which only utilize the spatial information within one single frame. Then, Caballero *et al.* (2017) designed a spatial transformer motion compensation network to detect the optical flow for warping neighboring frames. The current and warped neighboring frames were then fed into the Efficient Sub-Pixel Convolution Network (ESPCN)(Shi *et al.* (2016)) for super-resolution. Wang *et al.* (2019) proposed a unified framework, called EDVR, which is extensible to various video restoration tasks, including super-resolution and deblurring.

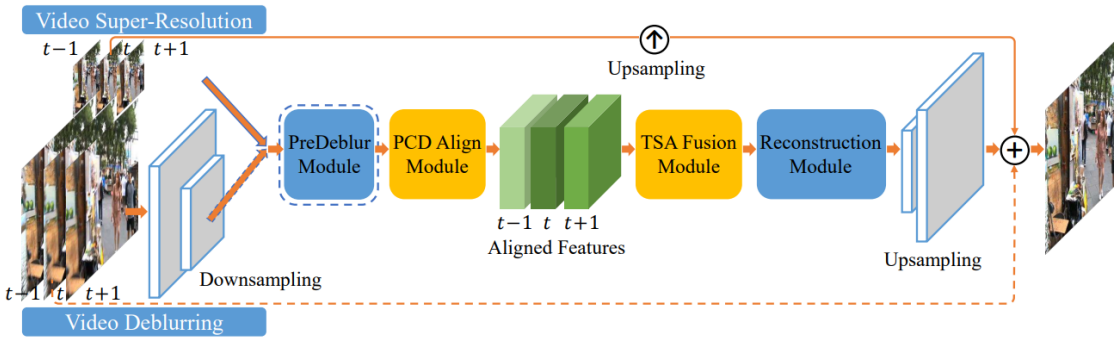

Figure 2.3: The framework of EDVR(source:Wang *et al.* (2019))

2.3 Image Quality Assessment

Mittal et al. proposed a natural scene statistic-based distortion-generic blind/no-reference (NR) image quality assessment (IQA) model that operates in the spatial domain. To illustrate a new practical application of Blind/Referenceless Image Spatial QUality Evaluator(BRISQUE), Mittal et al. describe a non-blind image denoising algorithm that can be augmented with BRISQUE in order to perform blind image denoising. Results show that BRISQUE augmentation leads to performance improvements over the state-of-the-art methods.

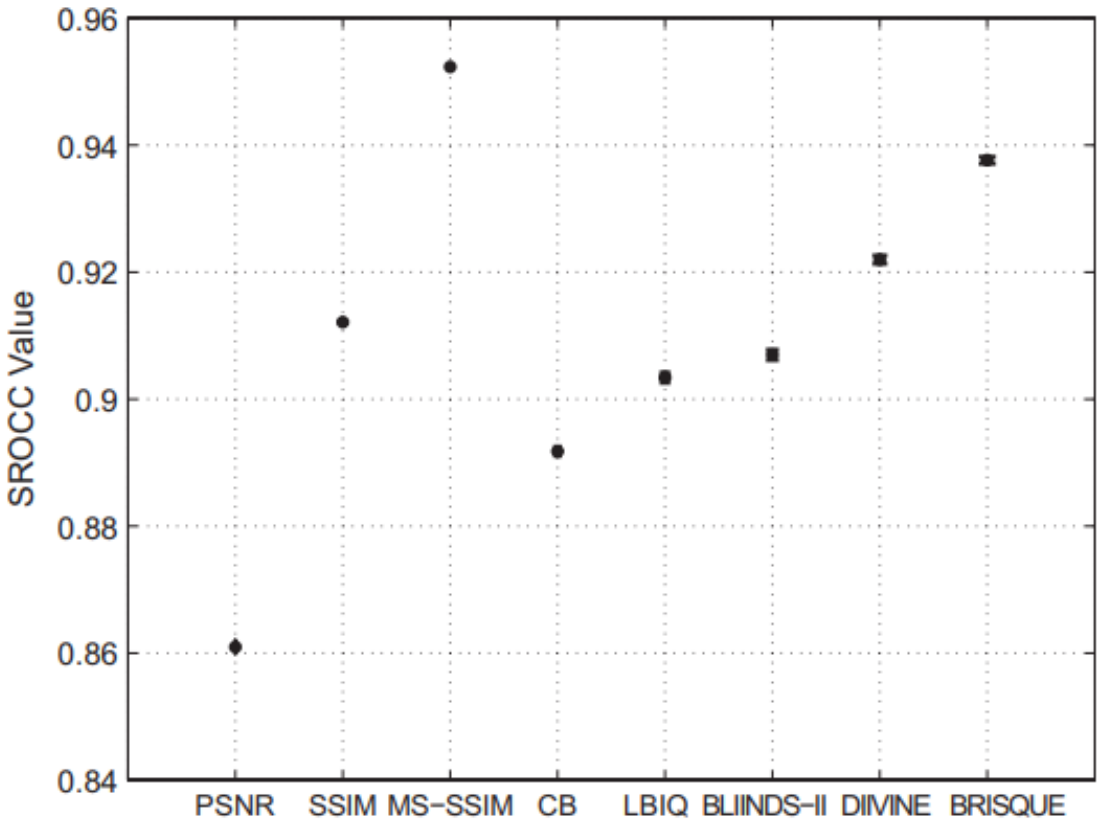

Figure 2.4: The performance of Brisque(source:Mittal *et al.* (2012))

2.4 Frame-level quality fluctuation

Yang et al. first observed that the PSNR varies significantly fluctuates across the compressed frames. Fig 2.5 shows the PSNR curves of 6 video sequences, which are compressed by different compression standards. It can be seen that PSNR fluctuates significantly across the compressed frames. This indicates that there exists considerable quality fluctuation in compressed video sequences for MPEG-1, MPEG-2, MPEG-4, H.264/AVC and HEVC. In addition, Fig. 2.6 visualizes the subjective results of some frames in one video sequence, which is compressed by the latest HEVC standard. We can see that visual quality varies across compressed frames, also implying the frame-level quality fluctuation.

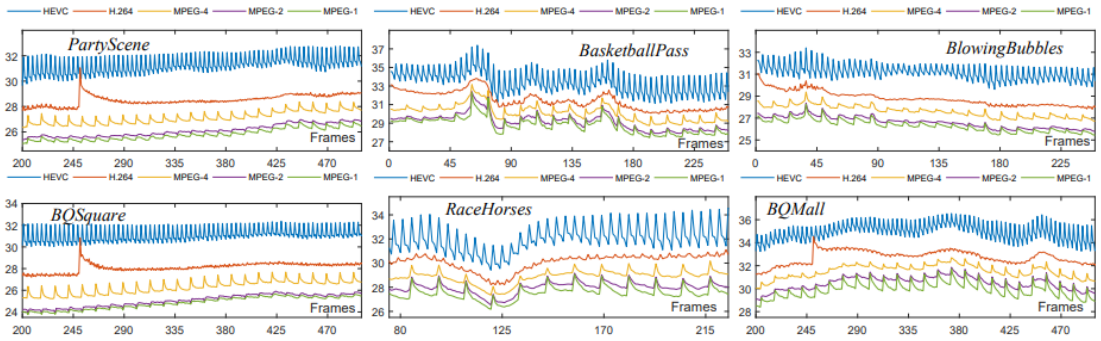

Figure 2.5: PSNR (dB) curves of compressed video by various compression standards.(source:Guan *et al.* (2021))

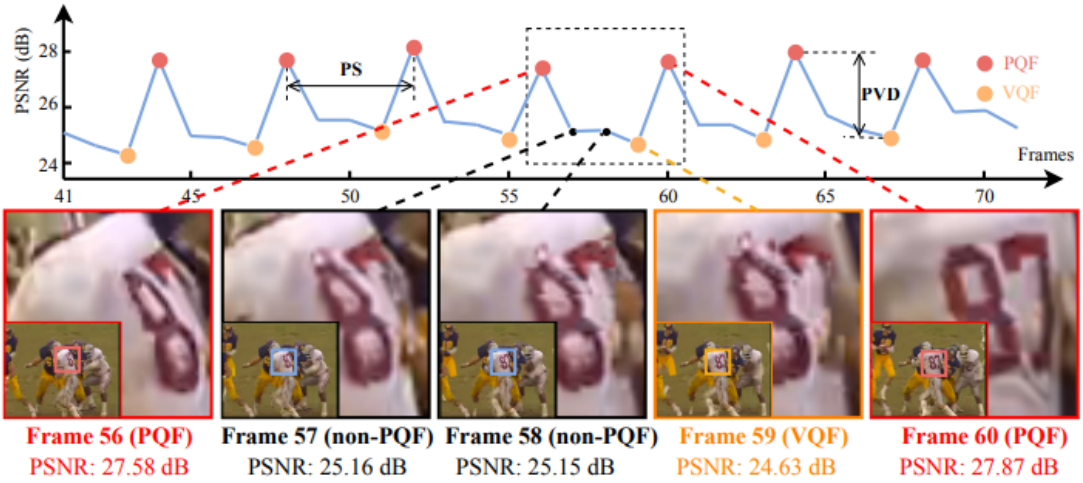

Figure 2.6: An example of frame-level quality fluctuation in video Football compressed by HEVC.(source:Guan *et al.* (2021))

2.5 Optical Flow Prediction

2.5.1 Learning Optical Flow with Convolutional Networks

In FlowNet1.0(Dosovitskiy *et al.* (2015)), the authors proposed and compared two architectures: FlowNetSimple and FlowNetCorr. Both of the two architectures are end-to-end learning approaches. In FlowNetSimple, as shown in Fig.2.7, the authors simply stack two sequentially adjacent input images together and feed them through the network. Compared with FlowNetSimple(top), FlowNetCorr(bottom) first produces representations of the two images separately, and then combines them together in the 'correlation layer', and learns the higher representation together. Both of the two architectures have refinements which are used for upsampling resolution.

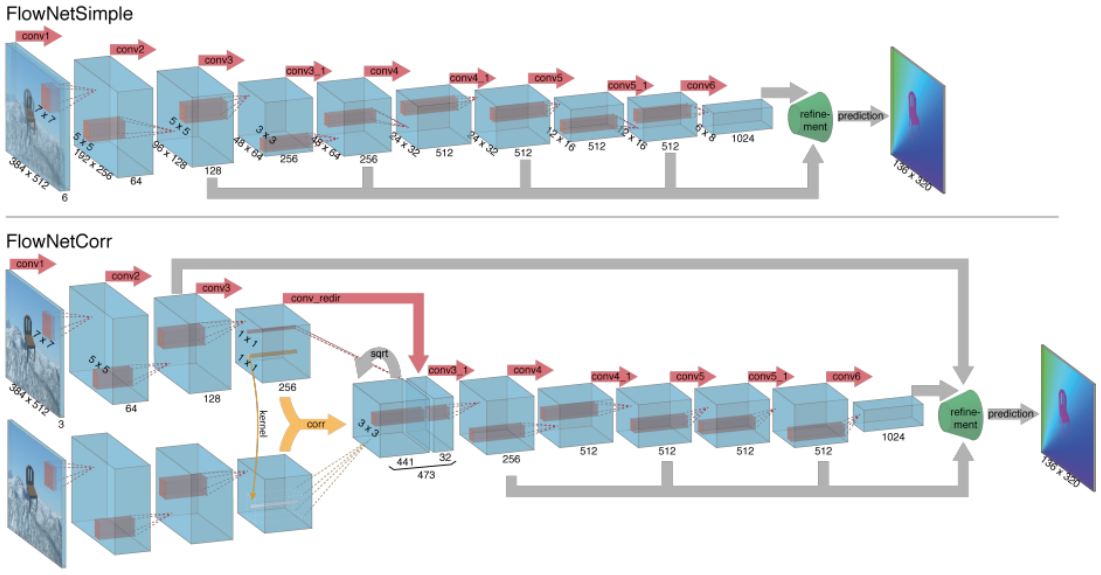

Figure 2.7: The two network architectures: FlowNetSimple (top) and FlowNetCorr (bottom).(source:Dosovitskiy *et al.* (2015))

FlowNet2.0(Ilg *et al.* (2017)) is much better than FlowNet1.0(Dosovitskiy *et al.* (2015)). Compared with FlowNet1.0, FlowNet2.0 has a large improvement in quality as well as speed. The main architecture is shown in Fig 2.8. It has four main contributions:

1. The schedule of presenting data is significant in training progress.
2. Proposed a stacked architecture.
3. Introduced a sub network specializing on small motions.
4. Proposed a fusion architecture.

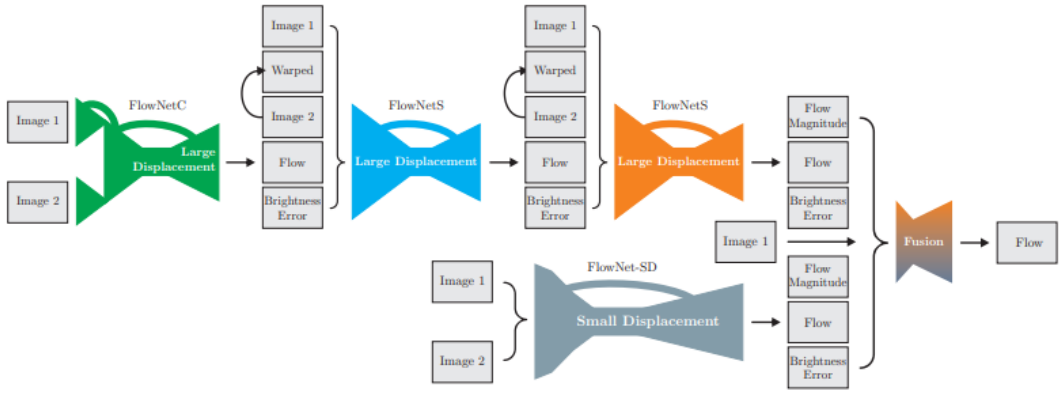

Figure 2.8: Schematic view of complete flownet2 architecture.(source:Ilg *et al.* (2017))

2.5.2 Learning Optical Flow with Deformable Convolutional Networks

Dai *et al.* (2017) first proposed deformable convolutions, in which additional offsets are learned to allow the network to obtain information beyond its regular local neighborhood, improving the capability of regular convolutions. Later, several works (Tian *et al.* (2020),Wang *et al.* (2019)) extended it along temporal direction to implicitly capture motion cues for video-related applications and achieved better performance than traditional methods.

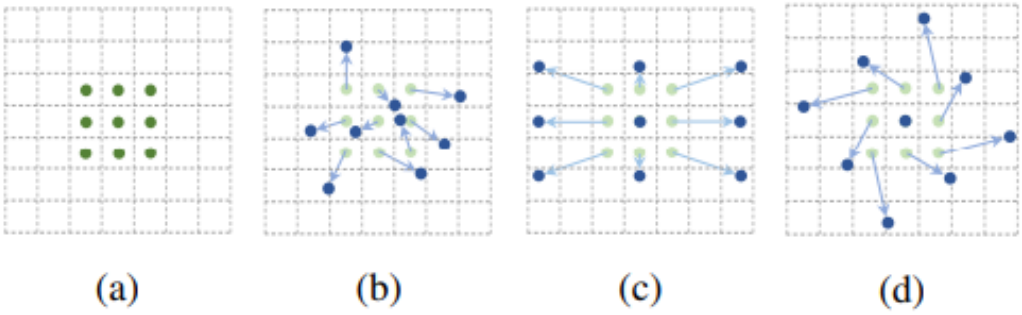

Figure 2.9: Illustration of the sampling locations in 3 × 3 standard and deformable convolutions. (a) regular sampling grid (green points) of standard convolution. (b) deformed sampling locations (dark blue points) with augmented offsets (light blue arrows) in deformable convolution. (c)(d) are special cases of (b), showing that the deformable convolution generalizes various transformations for scale, (anisotropic) aspect ratio and rotation.(source:Dai *et al.* (2017))

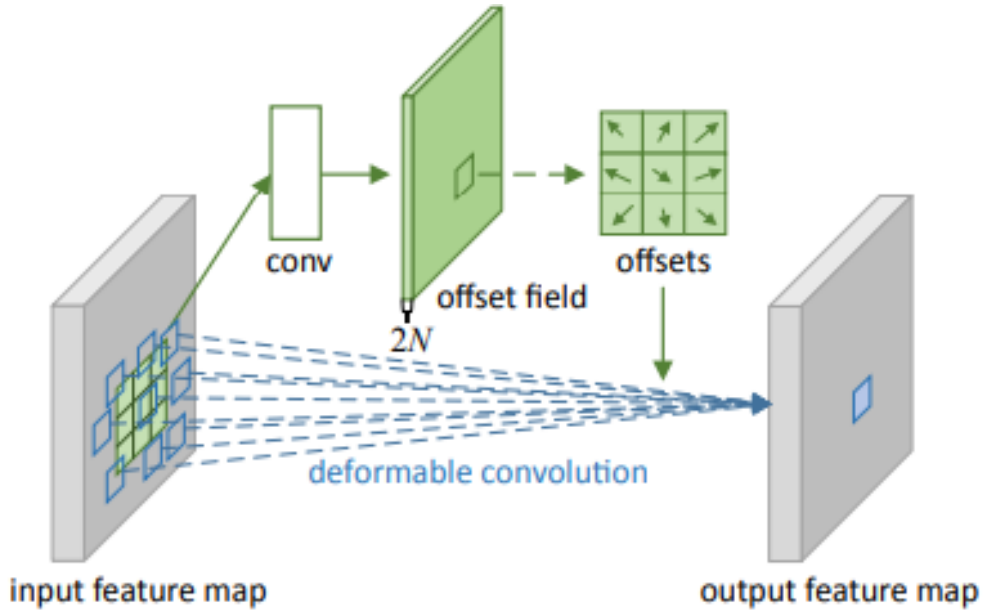

Figure 2.10: Illustration of 3 × 3 deformable convolution.(source:Dai *et al.* (2017))

However, these methods perform deformable convolution in a pairwise manner, thus fail to fully explore temporal correlations across multiple frames. Deng *et al.* (2020) proposed STDC to jointly consider a video clip rather than splitting it into several reference-target frame pairs, leading to more effective use of contextual information.

Chapter 3

Proposed Method

3.1 overview

Given a compressed video, the aim is to obtain a network design /solution capable of producing high quality results with the best perceptual quality and fidelity to the reference ground truth. To be specific, we conduct the enhancement separately for each compressed frame $I_{t_0}^{LQ} \in \mathbb{R}^{H \times W}$ at time t_0. We take the preceding and succeeding peak PSNR frames as reference to help enhancing quality of each target $I_{t_0}^{LQ}$. The enhanced solution $\hat{I}_{t_0}^{LQ} \in \mathbb{R}^{H \times W}$ can then be expressed as

$$\hat{I}_{t_0}^{HQ} = \mathcal{F}_\theta(I_{t_0-1}^{LQ}, I_{t_0}^{LQ}, I_{t_0+1}^{LQ}) \tag{3.1.1}$$

where $\mathcal{F}_\theta$ represents the proposed quality enhancement model and θ are the learnable parameters.

Figure 3.1 demonstrates the framework of our method.

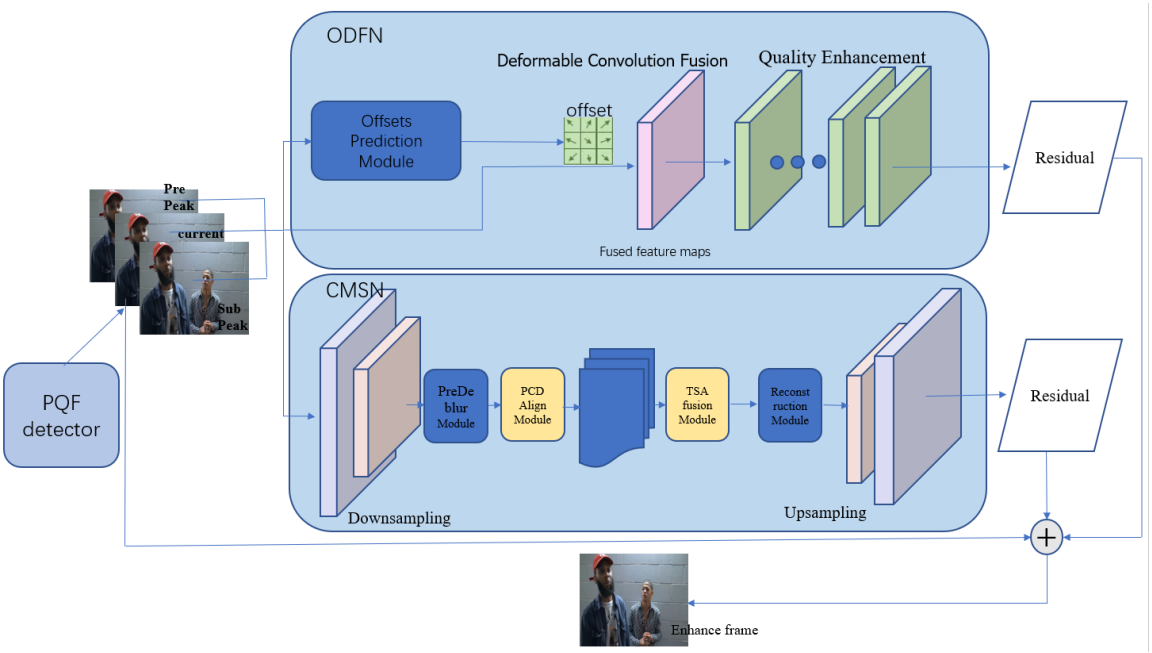

Figure 3.1: Overview of the proposed framework for compressed video quality enhancement. Firstly, we choose the current frame and its neighbor peak Quality Frames as the input data. Then send them to the two branches ODFN and CMSN. As a result, complementary information from both target and reference frames can be fused within the operation. Finally, we add the two residual frames on the raw target frame.

3.2 PQF Detection

In the training step, we calculated the PSNR of each frame in video series. Only frames whose quality is better than its neighbors' will be labeled as a PQF. Then we send the target frame and its neighboring PQFs to the input of the PEN. In the validation and test steps, due to the absence of raw videos, we used Brisque (Mittal et al.) to perform blind/Referenceless image quality. The blue line in Figure 3.2 represents the result of PSNR, while the red line is the result of Brisque. We can observe that PQFs are in very similar positions, which demonstrates the correctness of using Brisque to detect PQFs.

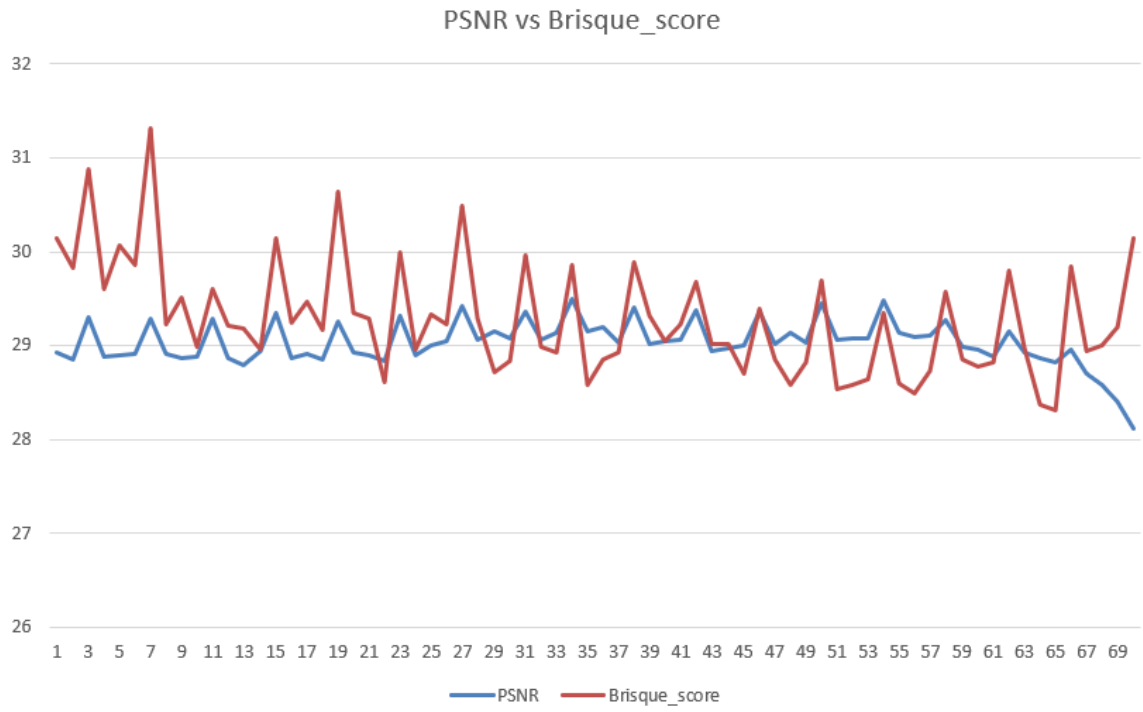

Figure 3.2: Comparison of PSNR and Brisque score

3.3 Offset Deformable Fusion Network(ODFN)

The ODFN can be divided into two separate parts: Deformable Convolution Fusion Module and Quality Enhancement Module.

3.3.1 Deformable Convolution Fusion Module

For a three frames input group $I_{t_0-1}^{LQ}$,$I_{t_0}^{LQ}$,$I_{t_0+1}^{LQ}$,the most straightforward fusion method is to apply convolution directly on the compressed frames. For example, Karpathy et al. have proposed a large-scale method EF for video classification in 2014 (Karpathy *et al.* (2014)) as:

$$F(p) = \sum_{t_0-1}^{t_0+1} \sum_{K=1}^{K^2} W_{t,k} \cdot I_t^{LQ}(\mathbf{p} + \mathbf{p}_k) \tag{3.3.1}$$

where *F* is the resulting feature map, *K* represents the size of convolution kernel, $W_t \in \mathbb{R}^{K^2}$is the kernel for t-th channel, $\mathbf{p}$ indicates arbitrary spatial position and $\mathbf{p}_k$ represents the regular sampling offsets. For example, $\mathbf{p}_k \in \{(-1,-1),(-1,0)...(1,1)\}$for *K*=3. Despite the high efficiency, EF(Karpathy *et al.* (2014)) may easily introduce noisy content and reduce the performance of subsequent enhancement due to temporal motion. Inspired by Dai *et al.* (2017). and Deng *et al.* (2020), we address this issue by introducing a Spatio-Temporal deformable Convolution to augment the regular sampling offset with extra learnable offset $\delta_{t,\mathbf{p}} \in \mathbb{R}^{2K^2}$ as

$$\mathbf{p}_k \leftarrow \mathbf{p}_k + \delta_{(t,\mathbf{p}),k}. \tag{3.3.2}$$

It is worth noting that the deformable offset $\delta_{t,\mathbf{p}}$ are position-specific, i.e., individual $\delta_{t,\mathbf{p}}$ will be assigned for each convolution window centered at spatio-temporal position $(t,\mathbf{p})$.

Thus, spatial deformations as well as tempora dynamics within the video clip can be simultaneously modeled. Since the learnable offsets can be fractional, we follow Dai *et al.* (2017) and Deng *et al.* (2020) to apply the differentiable bilinear interpolation to sample sub-pixel $I_t^{LQ}(\mathbf{p} + \mathbf{p}_k)$.

Unlike previous VQE methods(Yang *et al.* (2018); Guan *et al.* (2021)) which perform explicit motion compensation before fusion to alleviate the effect of temporal motion. We implicitly combines motion cues with position-specific sampling while conducting fusion. This leads to higher flexibility and robustness because adjacent convolution windows can sample contents independently. Unlike STDF(Deng *et al.* (2020)) which takes the neighboring 2R+1 frames as input, our method takes the proceeding and succeeding Peak Quality Frames that contain richer information compared with neighboring Low Quality Frames.

3.3.2 Quality Enhancement Module

The main idea of QE module is to fully explore complementary information from fused feature maps F and accordingly generate the enhanced target frame $I_{t_0}^{HQ}$. In order to take advantage of residual learning(Kim *et al.* (2016)), we first learn a non-linear mapping $\mathcal{F}_{\theta_{qe}}(\cdot)$ to predict the enhancement residual as

$$\hat{\mathcal{R}}_{t_0}^{HQ} = \mathcal{F}_{\theta_{qe}}(F). \tag{3.3.3}$$

The enhanced target frame can then be generated as

$$\hat{I}_{t_0}^{HQ} = \hat{\mathcal{R}}_{t_0}^{HQ} + I_{t_0}^{LQ}. \tag{3.3.4}$$

The last convolutional layer outputs the enhancement residual. Without bells and whistles, such plain QE network is able to achieve satisfactory enhancement results.

3.4 Complex Motion Solution Network(CMSN)

The input of CMSN is the same as the input of ODFN, we have added two modules to this branch:Pyramid, Cascading and Deformable convolutions (PCD) alignment module at the feature level and the Temporal and Spatial Attention(TSA) fusion module at the image level(Wang *et al.* (2019)). These can improve the network's ability for complex motion and large parallax issues in alignment and fusion.

3.4.1 PCD Module

Compared with previous deforamble convolution fusion methods, we followed Wang *et al.* (2019) to use a PCD module which mainly introduces two well established principles in optical flow: pyramidal processing(Ranjan and Black (2017)) and cascading refinement(Hui *et al.* (2018)) which can address complex motions and large parallax problems. The offset and alignment feature prediction of the pyramid structure is shown in Figure 3.3. As shown with purple lines, to generate feature at the l-th level, we use strided convolution filters to downsample the features at the (l-1)-th pyramid level by a factor of 2, obtaining L-level pyramids of feature representation. At the l-th level, offsets and aligned features are predicted also with the $\times 2$ upsampled offsets and aligned features from the upper (l+1)-th level, respectively (Dark and light green lines). Following the pyramid structure, a subsequent deformable alignment is cascaded to further refine the coarsely aligned features (the part with light yellow background in Figure 3.3). PCD module in such a coarse-to-fine manner improves the alignment to the sub-pixel accuracy. It is noteworthy that the PCD alignment module is jointly learned with the whole framework, without additional supervision(Tian *et al.* (2020)) or pretraining on other tasks.

3.4.2 TSA Module

Inter-frame temporal relation and intra-frame spatial relation are critical in fusion because
1. Different neighboring frames are not equally informative due to occlusion, blurry regions and parallax problems;
2. Misalignment and unalignment arising from the preceding alignment stage adversely affect the subsequent reconstruction performance.

Therefore, dynamically aggregating neighboring frames in pixel-level is indispensable for effective and efficient fusion. In order to address the above problems, we added a TSA fusion module(Wang *et al.* (2019)) to assign pixel-level aggregation weights on each frame. Specifically, we adopt temporal and spatial attentions during the fusion process, as shown in Figure 3.4.

Intuitively, a neighboring frame that is more similar to the reference one should be paid more attention. For each frame i $\in [-N : N]$, the similarity distance h can be calculated as:

$$h(F_{t+i}, F_t) = sigmoid(\theta(F_{t+i})^T \phi(F_t)) \tag{3.4.1}$$

where $\theta(F_{t+i})^T$ and $\phi(F_t)$ are two feature maps, which can be achieved with simple convolution filters. The sigmoid activation function is used to restrict the outputs in [0, 1], stabilizing gradient back-propagation. Note that, the temporal attention is spatial-specific for each spatial location. The temporal attention maps are then multiplied in a pixel-wise manner to the original aligned features F_{t+i}. An extra fusion convolution layer is adopted to aggregate these attention-modulated features $\hat{F}_{t+i}$:

$$\hat{F}_{t+i} = F_{t+i} \bigodot h(F_{t+i}, F_t) \tag{3.4.2}$$

$$F_{fusion} = Conv([\hat{F}_{t-1}, \hat{F}_t, \hat{F}_{t+1}]) \tag{3.4.3}$$

where $\bigodot$ denote the element-wise multiplication. Spatial attention masks are then computed from the fused features. A pyramid design is employed to increase the attention receptive field. After that, the fused features are modulated by the masks through element-wise multiplication and addition, similar to Wang *et al.* (2019).

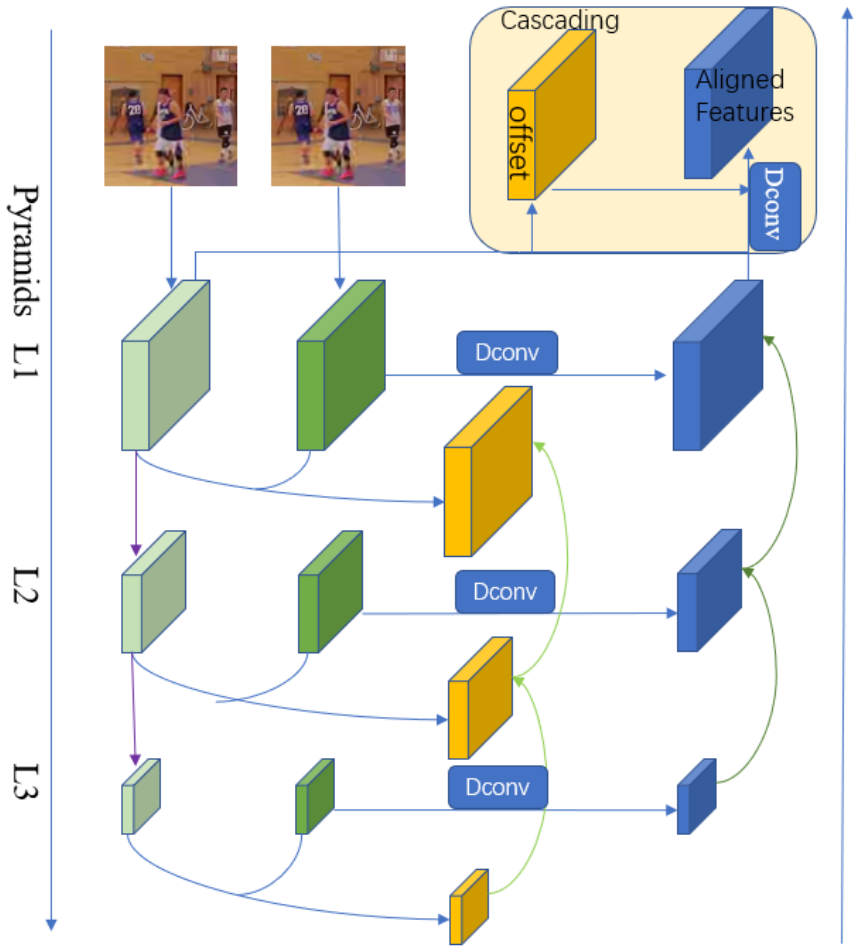

Figure 3.3: PCD alignment module with Pyramid, Cascading and Deformable convolution.

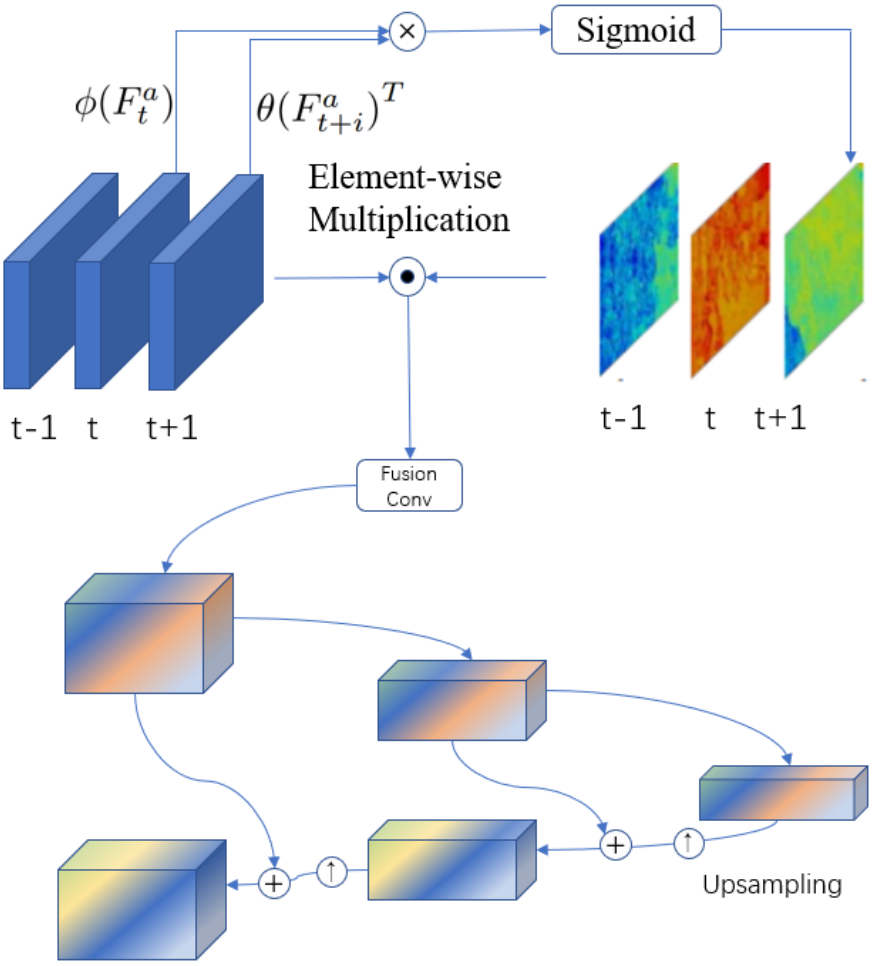

Figure 3.4: TSA fusion module with Temporal and Spatial Attention.

3.5 Training Scheme

Since ODFN and CMSN are fully-convolutional and thus differentiable, we jointly optimize them in an end-to-end fashion. The overall loss function $\mathcal{L}$ is set to the Sum of Squared Error (SSE) between the enhanced target frame$\hat{I}_{t_0}^{HQ}$ and the raw one $I_{t_0}^{HQ}$ as:

$$\mathcal{L} = ||\hat{I}_{t_0}^{HQ} - I_{t_0}^{HQ}||_2^2 \tag{3.5.1}$$

Note that, as there is no ground-truth for deformable offsets, learning for offset prediction network is totally unsupervised and fully driven by the final loss $\mathcal{L}$.

Chapter 4

Experiments and Results

4.1 Datasets

The organiser of NTIRE 2021 provided us with a total of 230 uncompressed videos. There are currently 230 videos in the collection, with a variety of content types, motion types, and frame rates. The training set consists of 200 videos. The video is compressed at a given bitrate and a fixed QP. For fixed-rate, the videos are compressed in the YUV domain by x265 of ffmpeg 4.3.1 at 200kbps. The raw YUV videos are losslessly compressed to mkv via ffmpeg x265 to reduce the file sizes. For fixed-QP, videos are compressed in the YUV domain by the Low-delay P mode HM 16.20 at QP 37. We calculated the PSNR of each frame and labeled out the PQFs for training.

Because we don't have the original lossless videos during the real video improvement process, we followed Mittal et al.Mittal *et al.* (2012) to detect each frame's quality and score them which is used for labeling PQFs. The compressed videos and their corresponding label files are then used as input for validation and test.

4.2 Implementation Details

The proposed method is implemented based on PyTorch framework. For training, we randomly crop 64 × 64 clips from raw and the corresponding compressed videos as training samples. Data augmentation (i.e., rotation or flip) is further used to better exploit those training samples. We train all models using Adam optimizer(Kingma and Ba (2015)) with β_1 = 0.9, β_2 = 0.999 and $\epsilon = 10^{-8}$. Learning rate is initially set to 10^{-4} and retained throughout training. We adopt Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM)(Zhou Wang *et al.* (2004)) to evaluate quality enhancement performance, which measure the improvement of the enhanced video from the compressed one.

4.3 Comparison to State-of-the-arts

We compared the proposed method with the state-of-the-art video quality enhancement methods: STDFDeng *et al.* (2020),MFQEYang *et al.* (2018),EDVRWang *et al.* (2019). For fair comparison, all video quality enhancement methods are retrained on our training set.

The quantitative accuracy values are presented in Table 1. On the 20 test videos, our method consistently outperformed all other methods in terms of average PSNR and SSIM, as can be seen. For fixed-rate, our method can achieve 27.7 db, which can improve video quality by an average of 0.7db. For fixed-QP, our method can achieve 29.76 db, which can improve video quality by an average of 0.5db. Comparing these two sets of experiments, the reason for better performance on fixed-rate is that the quality of videos compressed by fixed-rate is worse. The network can work more efficient.

Module	PSNR	SSIM	MSE
COMPRESSED	27.09421633	0.999453141	167.3619876
MFQE	27.4443455	0.999634534	155.3454354
STDF	27.4695685	0.999671607	154.7390707
EDVR	27.56421974	0.999746318	152.4928956
PEN	27.6984277	0.999784724	149.0298929

Table 4.1: Qualitative results of our module compared to others on three measure methods/fixed-rate

Module	PSNR	SSIM	MSE
COMPRESSED	29.30123438	0.999905059	85.59711258
MFQE	29.52641928	0.999963452	79.74057344
STDF	29.51714474	0.999950214	83.1527412
EDVR	29.64081145	0.999943558	81.21837553
PEN	29.76772013	0.99996444	78.89882611

Table 4.2: Qualitative results of our module compared to others on three measure methods/fixed-QP

Fig 4.1-4.6 are the histograms of the results, which give a more visual indication of the better results of our model. For both PSNR and SSIM, our network achieves the best results.

Fig 4.7 and 4.8 shows that our results not only achieve higher PSNR/SSIM/MSE but also better perceptually quality than reference methods. For example, in the horse picture, we can see smoother lines and less blurring on the horse's body.

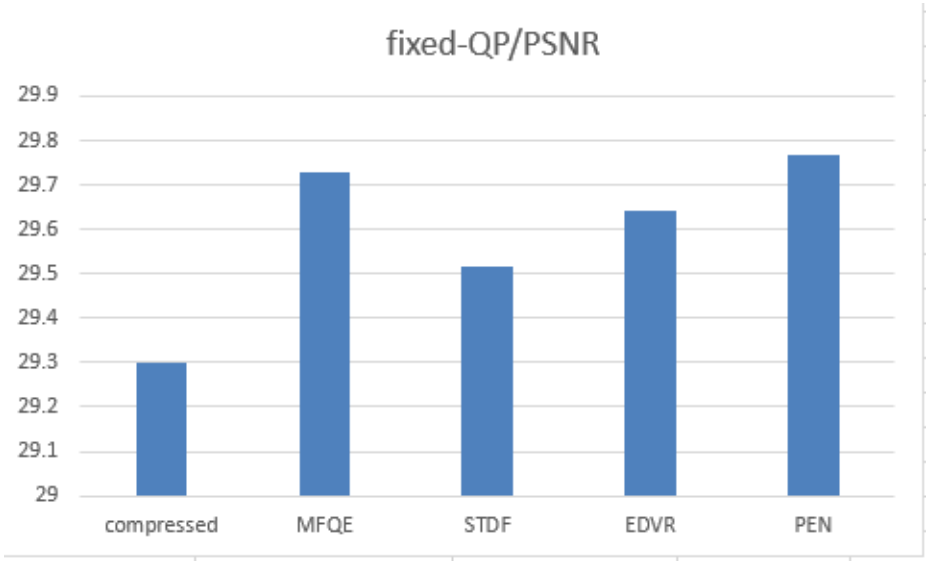

Figure 4.1: Histogram of experimental comparison of PSNR results/fixed-QP

Figure 4.2: Histogram of experimental comparison of SSIM results/fixed-QP

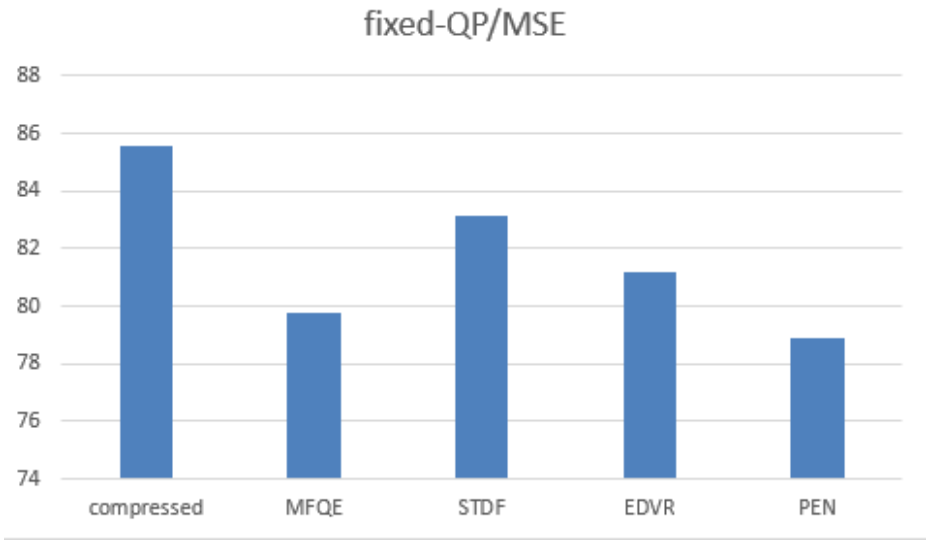

Figure 4.3: Histogram of experimental comparison of MSE results/fixed-QP

Figure 4.4: Histogram of experimental comparison of PSNR results/fixed-rate

Figure 4.5: Histogram of experimental comparison of SSIM results/fixed-rate

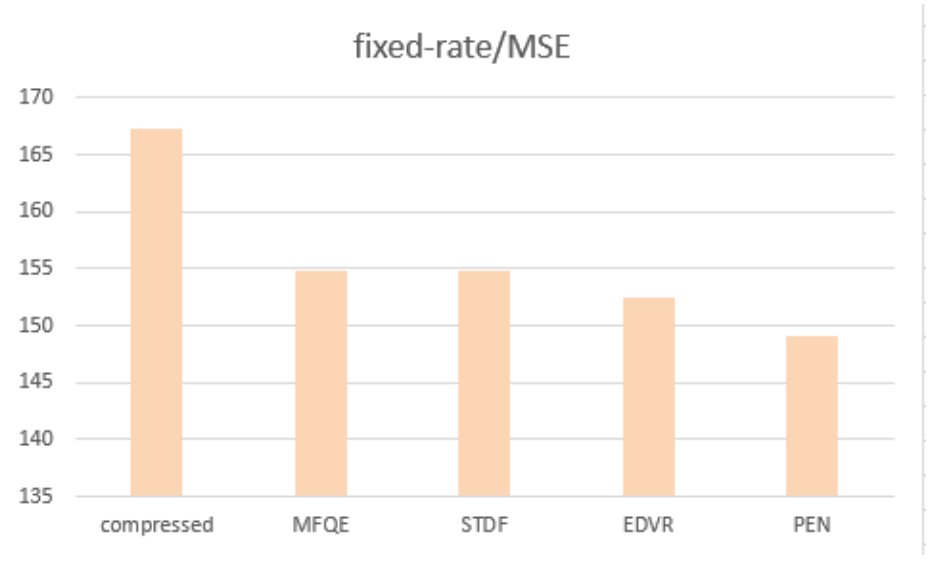

Figure 4.6: Histogram of experimental comparison of MSE results/fixed-rate

Figure 4.7: Qualitative results of videos compressed by fixed rate.

Figure 4.8: Qualitative results of videos compressed by fixed QP.

4.4 Ablation Study

4.4.1 PQF detector

In this section, we validate the necessity and effectiveness of utilizing PQFs to enhance the quality of non-PQFs. To this end, we retrain the PEN approach to enhance non-PQFs with the help of adjacent frames, instead of PQFs. The retrained model is represented by PEN_NF (i.e.,PEN with Neighboring Frames), and the experimental results are shown in Table 4.3, which are obtained by averaging over all 20 test sequences. We can see that our approach without considering PQFs can only result in 27.44 dB for PSNR gain. By contrast, as aforementioned, our approach with PQFs can achieve 27.7 dB enhancement in PSNR.

4.4.2 Deformable Offset Prediction

To demonstrate the effectiveness of deformable fusion, we compare it with a previous fusion scheme-flownet(Dosovitskiy *et al.* (2015),Ilg *et al.* (2017)). The experimental results are shown in Table 4.3, which are obtained by averaging over all 20 test sequences. We can see that our approach with flownet can only result in 27.45dB for PSNR gain.

4.4.3 Ensemble models

As shown in Table 4.3, the single ODFN can result in 27.47db for PSNR while the single CMSN can result in 27.56db, and the whole PEN approach can achieve 27.7db in PSNR. This indicates that combining two branches is a good idea.

Module	PSNR	SSIM	MSE
COMPRESSED	27.09421633	0.999453141	167.3619876
PEN_NF	27.443040304	0.999454351	169.34234425
flownet	27.450897878	0.999458685	155.4656307
ODFN	27.4695685	0.999671607	154.7390707
CMSN	27.56421974	0.999746318	152.4928956
PEN	27.6984277	0.999784724	149.0298929

Table 4.3: ablation study

Chapter 5

Conclusion

In this paper, we have proposed a CNN-based PEN approach for improving the quality of compressed video by reducing compression artifacts. Different from the current multi-frame quality enhancement approaches, we use the neighboring Peak Quality frames as reference to improve the quality of target frame. We also implement Deformable Convolution on fusion stage to improve the accuracy and efficiency. In the mean time, we have integrated a second branch which applies temporal and spatial attention to improve the quality and stability of the output. In the future, we will mainly focus on three points:

1. Modify the loss function further. Compare the benefits and drawbacks of L1 and L2, and decide whether to regularize.

2. Not only using Peak PSNR frames as reference frames, but also use adjacent low-PSNR frames with useful information.

3. In order to evaluate performance under different compression levels, the compression Quantization Parameters (QPs) can set to different values like 27,29,31,33.

Bibliography

Brandi, F., de Queiroz, R., and Mukherjee, D. (2008). Super resolution of video using key frames. In *2008 IEEE International Symposium on Circuits and Systems*, pages 1608–1611.

Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., and Shi, W. (2017). Real-time video super-resolution with spatio-temporal networks and motion compensation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2848–2857.

Chang, H., Ng, M. K., and Zeng, T. (2014). Reducing artifacts in jpeg decompression via a learned dictionary. *IEEE Transactions on Signal Processing*, **62**(3), 718–728.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. (2017). Deformable convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 764–773.

Deng, J., Wang, L., Pu, S., and Zhuo, C. (2020). Spatio-temporal deformable convolution for compressed video quality enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**(07), 10696–10703.

Dong, C., Deng, Y., Loy, C. C., and Tang, X. (2015). Compression artifacts reduction

by a deep convolutional network. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 576–584.

Dong, C., Loy, C. C., He, K., and Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**(2), 295–307.

Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., van der Smagt, P., Cremers, D., and Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. pages 2758–2766.

Foi, A., Katkovnik, V., and Egiazarian, K. (2007). Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. *Image Processing, IEEE Transactions on*, **16**, 1395 – 1411.

Guan, Z., Xing, Q., Xu, M., Yang, R., Liu, T., and Wang, Z. (2021). Mfqe 2.0: A new approach for multi-frame quality enhancement on compressed video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**(3), 949–963.

Guo, J. and Chao, H. (2016). Building dual-domain representations for compression artifacts reduction. In *ECCV*.

Huang, Y., Wang, W., and Wang, L. (2018). Video super-resolution via bidirectional recurrent convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**(4), 1015–1028.

Hui, T., Tang, X., and Loy, C. C. (2018). Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8981–8989.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. pages 1647–1655.

Jancsary, J., Nowozin, S., and Rother, C. (2012). Loss-specific training of non-parametric image restoration models: A new state of the art. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, pages 112–125, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kappeler, A., Yoo, S., Dai, Q., and Katsaggelos, A. K. (2016). Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, **2**(2), 109–122.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Kim, J., Lee, J. K., and Lee, K. M. (2016). Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *CoRR*, **abs/1412.6980**.

Li, D. and Wang, Z. (2017). Video superresolution via motion compensation and deep residual learning. *IEEE Transactions on Computational Imaging*, **3**(4), 749–762.

Li, S., Xu, M., Deng, X., and Wang, Z. (2015). Weight-based r rate control for perceptual hevc coding on conversational videos. *Signal Processing: Image Communication*, **38**, 127–140. Recent Advances in Saliency Models, Applications and Evaluations.

Liew, A. W. . and Hong Yan (2004). Blocking artifacts suppression in block-coded images using overcomplete wavelet representation. *IEEE Transactions on Circuits and Systems for Video Technology*, **14**(4), 450–461.

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, **21**(12), 4695–4708.

Ranjan, A. and Black, M. J. (2017). Optical flow estimation using a spatial pyramid network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2720–2729.

Shen, S., Fang, X., and Wang, C. (2011). Adaptive non-local means filtering for image deblocking. In *2011 4th International Congress on Image and Signal Processing*, volume 2, pages 656–659.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883.

Song, B. C., Jeong, S., and Choi, Y. (2011). Video super-resolution algorithm using bi-directional overlapped block motion compensation and on-the-fly dictionary training. *IEEE Transactions on Circuits and Systems for Video Technology*, **21**(3), 274–285.

Sullivan, G. J., Ohm, J., Han, W., and Wiegand, T. (2012). Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, **22**(12), 1649–1668.

Tai, Y., Yang, J., Liu, X., and Xu, C. (2017). Memnet: A persistent memory network for image restoration. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4549–4557.

Tian, Y., Zhang, Y., Fu, Y., and Xu, C. (2020). Tdan: Temporally-deformable alignment network for video super-resolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3357–3366.

Wang, X., Chan, K. C., Yu, K., Dong, C., and Loy, C. C. (2019). Edvr: Video restoration with enhanced deformable convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Wang, X., Chan, K. C. K., Yu, K., Dong, C., and Loy, C. C. (2019). Edvr: Video restoration with enhanced deformable convolutional networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1954–1963.

Wang, Z., Liu, D., Chang, S., Ling, Q., Yang, Y., and Huang, T. S. (2016). D3: Deep dual-domain based fast restoration of jpeg-compressed images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2764–2772.

Wiegand, T., Sullivan, G. J., Bjontegaard, G., and Luthra, A. (2003). Overview of the h.264/avc video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, **13**(7), 560–576.

Yang, R., Xu, M., Wang, Z., and Li, T. (2018). Multi-frame quality enhancement for compressed video. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6664–6673.

Yang, R., Sun, X., Xu, M., and Zeng, W. (2019). Quality-gated convolutional lstm for enhancing compressed video. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 532–537.

Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, **26**(7), 3142–3155.

Zhang, X., Xiong, R., Fan, X., Ma, S., and Gao, W. (2013). Compression artifact reduction by overlapped-block transform coefficient estimation with block similarity. *IEEE Transactions on Image Processing*, **22**(12), 4613–4626.

Zhou Wang, Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, **13**(4), 600–612.