Video Compression Artifact Reduction by Fusing Motion Compensation and Global Context in a Swin-CNN Based Parallel Architecture

Xinjian Zhang[1,2], Su Yang[1,2*], Wuyang Luo[1,2], Longwen Gao[3], Weishan Zhang [4]

[1]School of Computer Science, Fudan University
[2]Shanghai Key Laboratory of Intelligent Information Processing
[3]Bilibili
[4]Department of Software Engineering, China University of Petroleum (East China)
{zhangxj17, suyang, wyluo18}@fudan.edu.cn, gaolongwen@bilibili.com, zhangws@upc.edu.cn

Abstract

Video Compression Artifact Reduction aims to reduce the artifacts caused by video compression algorithms and improve the quality of compressed video frames. The critical challenge in this task is to make use of the redundant high-quality information in compressed frames for compensation as much as possible. Two important possible compensations: Motion compensation and global context, are not comprehensively considered in previous works, leading to inferior results. The key idea of this paper is to fuse the motion compensation and global context together to gain more compensation information to improve the quality of compressed videos. Here, we propose a novel Spatio-Temporal Compensation Fusion (STCF) framework with the Parallel Swin-CNN Fusion (PSCF) block, which can simultaneously learn and merge the motion compensation and global context to reduce the video compression artifacts. Specifically, a temporal self-attention strategy based on shifted windows is developed to capture the global context in an efficient way, for which we use the Swin transformer layer in the PSCF block. Moreover, an additional Ada-CNN layer is applied in the PSCF block to extract the motion compensation. Experimental results demonstrate that our proposed STCF framework outperforms the state-of-the-art methods up to 0.23dB (27% improvement) on the MFQEv2 dataset.

1 Introduction

With the development of video-based applications, video data have gradually become dominating digital network traffic (Hoang and Zhou 2021). To tackle the huge space cost and limited bandwidth in video data storage and transmission, lossy video compression algorithms, such as H.264/AVC (Wiegand et al. 2003) and H.265/HEVC (Sullivan et al. 2012), are widely used to compress video data. However, lossy video compression can also lead to various compression artifacts such as blocking, blurring, ringing, edge/texture floating, and jerkiness (Zeng et al. 2014; Deng et al. 2020). Such undesirable compression artifacts severely affect the quality of experience (QoE). Furthermore, these distorted contents in the compressed video also reduce the performance of downstream video-based tasks, such as video grounding (Zhang et al. 2020a; Liu et al. 2021a) and video summarization (Zhu et al. 2020; Apostolidis et al. 2021). Consequently, video compression artifact reduction has emerged as an important research topic in computer vision.

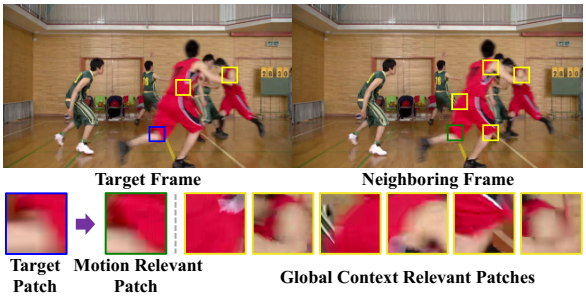

Figure 1: A patch with artifacts (the blue one) can be compensated with motion relevant or time line related patches on account of the redundancy existing in the two dimensions, which are marked as green and yellow, respectively.

In recent years, deep neural networks have achieved significant performance improvement in video compression artifact reduction. These works roughly fall into two types: Motion compensation based (Yang et al. 2018b; Xue et al. 2019; Guan et al. 2019; Deng et al. 2020; Zhu et al. 2019; Zhao, Xu, and Zhou 2021) and global context based (Lu et al. 2019; Xu et al. 2019; Hou, Zhao, and Wang 2021). In detail, motion compensation of a target patch uses the information extracted from motion relevant patch, as shown in Figure 1. The motion relevant patch is always explored by optical flow or other motion alignment methods. For instance, MFQE1.0 (Yang et al. 2018b) and MFQE2.0 (Guan et al. 2019) adopt a widely used temporal fusion scheme that incorporates dense optical flow for motion compensation. To improve the poor result of the optical flow estimation algorithm on distorted video frames, STDF (Deng et al. 2020) employs the deformable convolution (Zhu et al. 2019) to capture the motion compensation among multiple neighboring frames. RFDA (Zhao, Xu, and Zhou 2021) extends the STDF with recursive fusion and deformable spatiotemporal attention modules to model the motion compensation within a long temporal range. Unlike motion compensation,

*Corresponding author.

the global context captured from the highly correlated but possibly motion-misaligned patches (global context relevant patches in Figure 1) is also helpful for video compression artifact reduction. Several non-local based methods (Lu et al. 2019; Xu et al. 2019; Hou, Zhao, and Wang 2021) are proposed to extract the global context. However, non-local suffers from the huge computation cost. A work (Xu et al. 2019) analyses the difference between global context and motion compensation: The motion compensation is restricted by a fixed flow magnitude depending on the motion field (locally aggregated information); The global context is warped from multiple positions depending on the similarity determined by the feature. Hence, motion compensation and global context can be considered as complementary information. Recent works only exploit one of them to enhance the quality of compressed videos. Thus, although these methods have made great progress in this task, their performance is still limited due to the incomplete compensation information.

To address the issues mentioned above, we fuse motion compensation and global context from multiple preceding and following frames of the compressed frame to boost the performance of video compression artifact reduction. To this end, we propose a novel end-to-end Spatio-Temporal Compensation Fusion (STCF) framework. Our framework consists of the Spatiotemporal Alignment (SA) module and the Quality Enhancement (QE) module. The SA module is used to align the motion of input frames with the predicted offsets and provide the shallow temporal feature. The QE module infers the high-quality frame from the fused compensation produced by the efficient Parallel Swin-CNN Fusion (PSCF) block. In detail, the motion compensation is learned from the shallow temporal feature via the Ada-CNN layer in PSCF. To reduce the computational complexity for global context extraction, we further design a shifted windows (Swin) based temporal self-attention to compute the intra-frame and inter-frame patch-wise similarity. We select the optimal fusion scheme to aggregate motion compensation and global context through experiments. Our framework can fully use the neighboring highly-related compensation information to enhance the compressed frame, which outperforms all existing methods on the MFQEv2 dataset.

Major contributions of this work include: 1) We propose and verify a new solution for video compression artifact reduction by exploiting both motion compensation and global context. 2) We develop an STCF framework with the novel PSCF block. The PSCF block is efficient and effective that can be easily extended to various video low-level tasks. 3) We design a Swin-based temporal self-attention strategy that dramatically reduces the calculation cost and enables high resolution training and inference. 4) We conduct extensive experiments over the MFQEv2 dataset to evaluate the proposed method. Our method achieves state-of-the-art performance for video compression artifact reduction.

2 Related Work

2.1 Image Compression Artifact Reduction

In early studies, prior knowledge plays a vital role in image compression artifact reduction since it is an ill-posed inverse problem. The prior knowledge includes: The quantization step (Liu et al. 2016), sparse representation (Song et al. 2020), non-local self-similarity (Zhang et al. 2013), and graph (Mu et al. 2020). Recently, deep learning (DL) has made breakthrough progress in JPEG compression artifacts removal. ARCNN (Dong et al. 2015) first introduces the DL-based method by designing a four-layer CNNs architecture. Several deep networks (Zhang et al. 2017, 2020b) are well designed in residual architectures. Many non-local based methods (Liu et al. 2018; Zhang et al. 2019) utilize the long-range dependencies of the whole image to reduce artifacts. They compute the self-similarity between each pixel and its neighbors to capture the intra-frame global context.

2.2 Video Compression Artifact Reduction

Most previous methods (Dai, Liu, and Wu 2017; Yang, Xu, and Wang 2017; Jin et al. 2018; Yang et al. 2018a; Wang, Chen, and Chao 2017) take only a single frame to reduce video compression artifacts, neglecting the spatiotemporal correlation between neighboring frames. Thus, recent works are proposed to exploit the spatiotemporal information such as motion compensation and global context from neighboring frames. TOFlow (Xue et al. 2019) designs DL-based motion estimation and video processing components with a joint training strategy to handle various low-level vision tasks. MFQE1.0 (Yang et al. 2018b) and MFQE2.0 (Guan et al. 2019) propose multi-frame quality enhancement (MFQE) and utilize motion compensation of the two nearest peak quality frames extracted by optical flow estimation algorithm to enhance low-quality frames. Since compression artifacts could seriously distort video contents and break pixel-wise correspondences between frames, the estimated optical flow tends to be inaccurate and unreliable. Consequently, the STDF (Deng et al. 2020) leverages the deformable convolution to align the motion and capture the motion compensation. RFDA (Zhao, Xu, and Zhou 2021) further improves the STDF with recursive fusion and deformable spatiotemporal attention modules to model the motion compensation within a long temporal range. C3D layers are also utilized to learn the motion compensation (Ding et al. 2021). Several works (Lu et al. 2019; Hou, Zhao, and Wang 2021) apply the non-local method to acquire the global context among neighboring frames. A two-stage non-local similarity approximation (Xu et al. 2019) is developed to reduce the calculation and memory cost of the non-local operation. Differently, here, we adopt a Swin-based temporal self-attention to model both intra-frame and inter-frame global contexts in a low computation-cost way. Besides, previous methods fail to comprehensively consider both motion compensation and global context, and the video enhanced by them still suffers from artifacts. Our framework can leverage much more compensation, significantly improving the restoration performance.

2.3 Shifted Windows based Self-Attention

Recently, self-attention (Vaswani et al. 2017) has gained much popularity in the computer vision field. The self-attention learns to attend to important image regions by exploring the global context between different regions. Al-

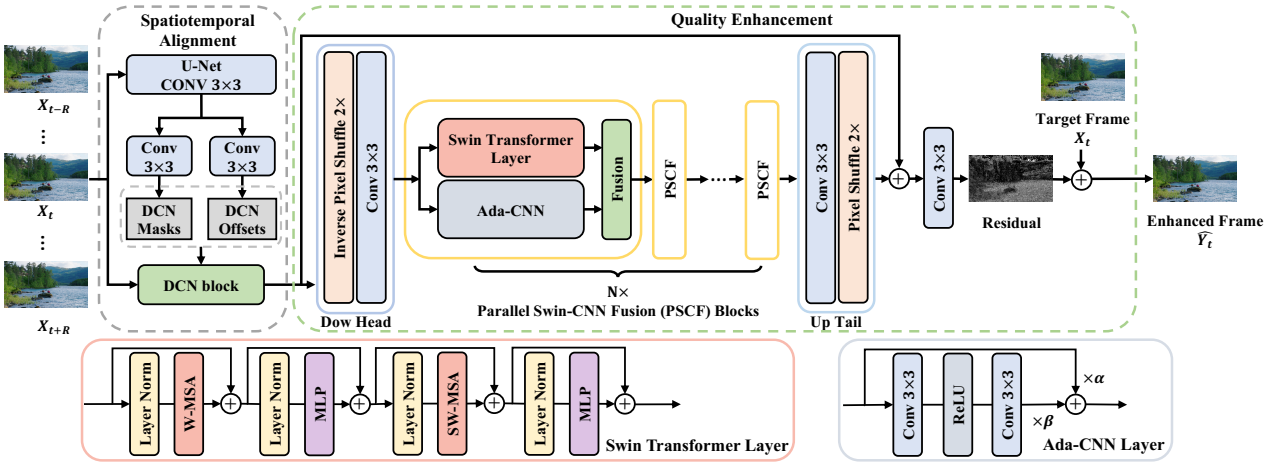

Figure 2: Architecture of our Spatio-Temporal Compensation Fusion (STCF) framework.

though the self-attention shows impressive performance on various tasks, the high computational cost makes it difficult to be applied to low-level vision tasks. The shifted windows (Swin) self-attention (Liu et al. 2021b,a) brings greater efficiency by limiting self-attention computation to non-overlapping local windows while allowing cross-window interaction. After that, Swin self-attention has been extensively applied to image low-level vision tasks (Liang et al. 2021; Wang et al. 2022; Zhang et al. 2022). As our proposed temporal self-attention gives rise to huge computational complexity, making it difficult to train and infer on high-resolution videos, we introduce the Swin self-attention mechanism to solve this problem, which makes our temporal self-attention workable for high-resolution videos.

3 Proposed Method

3.1 Overview

Given a T-frames compressed video $V^{LQ} = \{X_t \in \mathbb{R}^{C\times H\times W} | t = 1, ..., T\}$, where the C is the number of channels of the compressed frame, and H, W represent the frame height and width, respectively, video compression artifact reduction aims to reduce the artifacts in compressed frame X_t and enhance it to a high-quality frame $\hat{Y}_t$. In our method, we take the preceding and succeeding R frames as input to enhance the quality of the target frame X_t. Thus, the corresponding input reference frames are $I_t = \{X_{t-R}, \ldots, X_t, \ldots, X_{t+R}\}$.

In this paper, we consider employing motion compensation and global context to enhance the quality of the compressed frame. For motion compensation, it is necessary to predict the motion offsets to align the motion in consecutive frames (Deng et al. 2020; Zhao, Xu, and Zhou 2021), so we can get the aligned shallow feature to extract the deep motion compensation. Besides, our proposed temporal self-attention method requires the aligned shallow feature to extract the global context. As shown in Figure 2, we design a Spatiotemporal Alignment (SA) module, which is responsible for aligning the motion in reference frames and providing the shallow temporal feature. The structure and settings of the SA module are the same as those of the STDF module in (Deng et al. 2020), which is commonly used in recent works (Zhao, Xu, and Zhou 2021; Xu et al. 2021).

After that, we can easily apply our proposed Parallel Swin-CNN Fusion (PSCF) blocks to capture and fuse motion compensation and global context. In the PSCF block, the motion compensation is captured by the Ada-CNN layer, while the Swin transformer layer models the global context following the Swin-based temporal self-attention mechanism. The fusion layer in PSCF will aggregate the motion compensation and global context together.

To reduce the computational complexity, we utilize Down Head to reduce the spatial resolution of features. The UP Tail utilizes an architecture contrary to that of the Down Head to restore the original resolution behind the final PSCF block. Finally, a convolutional layer combines the fused compensation with the shallow feature from a skip connection to generate the residual for the enhanced frame. All these components make up our Quality Enhancement (QE) module. Overall, the enhanced frame $\hat{Y}_t$ of the compressed frame X_t can be generated as follow:

$$\begin{aligned} f_t^S &= \mathcal{F}_\theta(I_t), \\ \hat{Y}_t &= \mathcal{F}_\phi(f_t^S, X_t), \end{aligned} \tag{1}$$

where $\mathcal{F}_\theta(\cdot)$ represents the *Spatiotemporal Alignment* module, and $\mathcal{F}_\phi(\cdot)$ is the *Quality Enhancement* module. θ and ϕ are the learnable parameters of the corresponding modules. $f_t^S \in \mathbb{R}^{C_f\times H\times W}$ is the shallow temporal feature, where C_f indicates the number of channels for it.

3.2 Motion Compensation via Ada-CNN

Through the SA module, we extract the aligned shallow features of the reference frames I_t. To obtain motion compensation, we need to further extract the deep spatiotemporal information from the shallow feature. Thus, we propose the Ada-CNN, and the structure of Ada-CNN is illustrated in Figure 2.

The design of Ada-CNN is based on the residual block with two improvements: 1) **Wide activation.** Inspired by

WDSR-A (Yu et al. 2018), we use the 3×3 convolution to enlarge the channel number of the input feature by a factor of 4 before ReLU activation; 2) **Adaptive parameters.** We add two additional learnable parameters α and β, which are initialized with 1 and 0.2, respectively. After these improvements, the Ada-CNN can be extended to a very deep version to achieve better results.

3.3 Global Context via Swin-based Temporal Self-Attention

In image-oriented vision tasks, self-attention (Vaswani et al. 2017) is widely used to extract the global context within a single image. However, video-oriented tasks require computing both intra-frame and inter-frame global context information, which is beyond the capability of self-attention. At the same time, due to the high computational complexity, self-attention is difficult to train or infer with high-resolution data. To address these issues, we first extend self-attention to a temporal self-attention by replacing the key and value vectors. Then, we introduce the Swin self-attention (Liu et al. 2021b) in our temporal self-attention by an approximation method to reduce the computational complexity. In the following, we present how to learn the global context via Swin-based temporal self-attention.

Temporal self-attention. Given a specific compressed frame X_{t_0} (t_0^{th} frame of V^{LQ}) as target frame, and the corresponding reference frames $I_{t_0} = \{X_t | t = t_0 - R, \dots, t_0 + R\}$, different from self-attention, our temporal self-attention computes the query vector Q^L from the target compressed frame X_{t_0} while the key vectors K_t^R and value vectors V_t^R are calculated from the reference frames:

$$\begin{aligned} Q^L &= W_Q \mathcal{F}_\tau(X_{t_0}), \\ K_t^R &= W_K \mathcal{F}_\tau(X_t), t \in \{t_0 - R, \dots, t_0 + R\}, \\ V_t^R &= W_V \mathcal{F}_\tau(X_t), t \in \{t_0 - R, \dots, t_0 + R\}, \end{aligned} \tag{2}$$

where $\mathcal{F}_\tau(\cdot)$ represents the feature extractor, τ is learnable parameters, and $W_Q, W_K, W_V \in \mathbb{R}^{D \times C_\tau}$ are the projection matrices. Provided C_τ is the channel number of the feature from $\mathcal{F}_\tau(\cdot)$, and D the channel number of projected features, the global context G_{t_0} of X_{t_0} is calculated as follows:

$$\begin{aligned} G_{t_0} = \{ &\textit{SoftMax}(\frac{Q^L (K_t^R)^\top}{\sqrt{D}}) V_t^R | \\ &t = t_0 - R, \dots, t_0 + R\}. \end{aligned} \tag{3}$$

Since Q^L is computed from the target frame while K_t^R and V_t^R are from the reference frames, the G_{t_0} aggregates the correlation between elements (pixels or patches) in the reference frames. However, temporal self-attention suffers from high computational complexity. The computational complexity of temporal self-attention (omit the $\mathcal{F}_\tau(\cdot)$) is:

$$\Omega(TMA) = (2R + 1) \times (4hwD^2 + 2(hw)^2 D), \tag{4}$$

where h and w are the height and width of Q^L. The computation of temporal self-attention is generally unaffordable for a large hw. The Swin scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window interaction. So, we attempt to introduce the Swin self-attention in temporal self-attention. However, it requires Q, K, and V computed from the same frame feature.

Approximation of the query vector. As shown in Figure 1, the motion relevant patch in the neighboring frame has the same content as the target patch. Therefore, the global context queried by the motion relevant patch should be consistent with that queried by the target patch. Motivated by this, we align the motion between reference frames and the target frame to get the aligned frames $I_{t_0}^{align} = \{X_t^{align} | t = t_0 - R, \dots, t_0 + R\}$, and make an approximation to Q^L:

$$\begin{aligned} Q^L &= W_Q \mathcal{F}_\tau(X_{t_0}) \approx W_Q \mathcal{F}_\tau(X_t^{align}) \\ &= Q_t^R, t \in \{t_0 - R, \dots, t_0 + R\}. \end{aligned} \tag{5}$$

Combining Equations (2), (3), and (5), we can recalculate the global context on the $I_{t_0}^{align}$ as follows:

$$G_{t_0} = \{\textit{SoftMax}(\frac{W_Q H_\tau (W_K H_\tau)^\top}{\sqrt{D}}) W_V H_\tau\}, \tag{6}$$

where $H_\tau = \mathcal{F}_\tau(I_{t_0}^{align})$ is the feature map of the aligned frames. After approximation, the Q, K, and V are all computed from H_τ, so we can directly apply the Swin self-attention to reduce the computational complexity. Supposing each window in Swin self-attention contains $M \times M$ pixels, the computational complexity of Swin-based temporal self-attention will be reduced to:

$$\Omega(SWTMA) = (2R + 1) \times (4hwD^2 + 2M^2 hwD). \tag{7}$$

In summary, we decompose the temporal self-attention mechanism into two processes: Motion alignment and Swin self-attention, which can reduce the computational complexity without affecting the performance. Especially, the H_τ is the feature of motion-aligned reference frames, which is exactly the shallow feature f_t^S calculated by the SA module. Consequently, we only need to employ the Swin self-attention after the SA module. In practice, instead of Swin self-attention, we apply the whole Swin transformer layer (detailed in Figure 2) proposed in (Liu et al. 2021b) to capture the global context for better performance.

3.4 Parallel Swin-CNN Fusion Blocks

We parallelize the Ada-CNN layer with the Swin transformer layer and further utilize a fusion layer to learn and fuse the different compensations. This parallel structure referred to as Parallel Swin-CNN Fusion (PSCF) block. The QE module takes N PSCF blocks to continuously refine the extracted compensation information. We have conducted extensive experiments on the number of blocks N and the fusion scheme used in the fusion layer, which will be presented in Sec 4.4.

Before the PSCF, a Down Head is used to reduce the spatial feature resolution. Correspondingly, an Up Tail is placed after the last PSCF block to restore the resolution. For the Swin transformer layer, they also play the role of patch partition layer and patch merging layer, respectively. The Down Head comprises an inverse pixel shuffle layer (Shi et al. 2016) and a 3×3 convolutional layer. The Up Tail utilizes an architecture contrast to that of the Down Head.

QP	Approach		AR-CNN ICCV2015	DnCNN TIP2017	MFQE 2.0 TPAMI2019	STDF-R3L AAAI2020	RFDA MM2021	BasicVSR++ CVPR2022	Ours N=16
37	A	*Traffic*	0.24 / 0.47	0.24 / 0.57	0.59 / 1.02	0.73 / 1.15	0.80 / 1.28	**0.94 / 1.52**	0.91 / 1.44
		PeopleOnStreet	0.35 / 0.75	0.41 / 0.82	0.92 / 1.57	1.25 / 1.96	1.44 / 2.22	1.37 / 2.23	**1.62 / 2.43**
	B	*Kimono*	0.22 / 0.65	0.24 / 0.75	0.55 / 1.18	0.85 / 1.61	1.02 / 1.86	**1.41 / 2.18**	1.21 / 1.94
		ParkScene	0.14 / 0.38	0.14 / 0.50	0.46 / 1.23	0.59 / 1.47	0.64 / 1.58	**0.86 / 2.25**	0.74 / 1.79
		Cactus	0.19 / 0.38	0.20 / 0.48	0.50 / 1.00	0.77 / 1.38	0.83 / 1.49	0.62 / 1.51	**0.93 / 1.61**
		BQTerrace	0.20 / 0.28	0.20 / 0.38	0.40 / 0.67	0.63 / 1.06	0.65 / 1.06	0.71 / **1.25**	**0.75 / 1.25**
		BasketballDrive	0.23 / 0.55	0.25 / 0.58	0.47 / 0.83	0.75 / 1.23	0.87 / 1.40	1.02 / 1.53	**1.09 /1.59**
	C	*RaceHorses*	0.22 / 0.43	0.25 / 0.65	0.39 / 0.80	0.55 / 1.35	0.48 / 1.23	**0.76 / 1.84**	0.69 / 1.59
		BQMall	0.28 / 0.68	0.28 / 0.68	0.62 / 1.20	0.99 / 1.80	1.09 / 1.97	1.17 / **2.24**	**1.25** / 2.21
		PartyScene	0.11 / 0.38	0.13 / 0.48	0.36 / 1.18	0.68 / 1.94	0.66 / 1.88	0.44 / 1.71	**0.73 / 2.28**
		BasketballDrill	0.25 / 0.58	0.33 / 0.68	0.58 / 1.20	0.79 / 1.49	0.88 / 1.67	0.87 / 1.67	**0.96 / 1.76**
	D	*RaceHorses*	0.27 / 0.55	0.31 / 0.73	0.59 / 1.43	0.83 / 2.08	0.85 / 2.11	**1.02 / 2.74**	**1.02** / 2.47
		BQSquare	0.08 / 0.08	0.13 / 0.18	0.34 / 0.65	0.94 / 1.25	1.05 / 1.39	0.61 / 0.93	**1.06 / 1.48**
		BlowingBubbles	0.16 / 0.35	0.18 / 0.58	0.53 / 1.70	0.74 / 2.26	0.78 / 2.40	0.69 / **2.65**	**0.80** / 2.53
		BasketballPass	0.26 / 0.58	0.31 / 0.75	0.73 / 1.55	1.08 / 2.12	1.13 / 2.24	1.22 / **2.66**	**1.32** / 2.63
	E	*FourPeople*	0.37 / 0.50	0.39 / 0.60	0.73 / 0.95	0.94 / 1.17	**1.13** / 1.36	**1.13 / 1.38**	1.11 / 1.33
		Johnny	0.25 / 0.10	0.32 / 0.40	0.60 / 0.68	0.81 / 0.88	0.90 / 0.94	0.99 / 0.97	**1.00 / 1.13**
		KristenAndSara	0.41 / 0.50	0.42 / 0.60	0.75 / 0.85	0.97 / 0.96	1.19 / **1.15**	**1.20** / 1.13	1.12 / 1.11
	Average		0.23 / 0.45	0.26 / 0.58	0.56 / 1.09	0.83 / 1.51	0.91 / 1.62	0.95 / 1.80	**1.02 / 1.81**
42	Average		0.29 / 0.96	0.22 / 0.77	0.59 / 1.65	0.76 / 2.04	0.82 / 2.20	– / –	**0.88 / 2.34**
32	Average		0.18 / 0.19	0.26 / 0.35	0.52 / 0.68	0.86 / 1.04	0.87 / 1.07	0.89 / 1.25	**1.07 / 1.32**
27	Average		0.18 / 0.14	0.27 / 0.24	0.49 / 0.42	0.72 / 0.57	0.82 / 0.68	– / –	**1.05 / 0.88**
22	Average		0.14 / 0.08	0.29 / 0.18	0.46 / 0.27	0.63 / 0.34	0.76 / 0.42	– / –	**0.93 / 0.54**

Table 1: Overall performance comparison in terms of ΔPSNR (dB) / ΔSSIM ($\times 10^{-2}$) over the test sequences at five QPs. Video resolutions: Class A (2560×1600), Class B (1920×1080), Class C (832×480), Class D (480×240), Class E (1280×720).

4 Experiments

4.1 Dataset

Following recent works (Deng et al. 2020; Zhao, Xu, and Zhou 2021; Xu et al. 2021), we conduct our experiments on the MFQEv2 dataset. It contains 126 video sequences collected from Xiph[1], VQEG[2], and JCT-VC (Bossen 2011) with various resolutions and contents. According to the common setting in the above works: 108 of them are taken for training and the remaining 18 for testing. All video sequences are compressed by HM 16.20 with HEVC Low-Delay-P (LDP) configuration. In order to evaluate performance under different compression levels, the compression is conducted with five different Quantization Parameters (QPs): 22, 27, 32, 37, and 42.

4.2 Implementation Details

In our experiments, we set STDF-R3L (Deng et al. 2020) and RFDA (Zhao, Xu, and Zhou 2021) as our baseline. Following their settings, we set $R = 3$. The number of channels C_f for the feature fed into the QE module is 64. The window size M is 8, and the number of layers N for the PSCF block is 16. In the training phase, we randomly crop 128×128 clips from raw, and the corresponding compressed videos as training samples. The data augmentation (flip and rotation) is adopted to further expand training samples. We only employ the Charbonnier Loss (Charbonnier et al. 1994) to optimize the model. Our model is trained by the Adam optimizer (Kingma and Ba 2014) with an initial learning rate of 10^{-4}. The learning rate will be set to 10^{-5} when 80% iterations are reached. Our model is trained on 4 NVIDIA 16G V100 GPUs with Pytorch1.10. For evaluation, the same as the baseline works, we only apply quality enhancement on Y-channel in YUV/YCbCr space. We adopt improvement over Peak Signal-to-Noise Ratio (ΔPSNR) and Structural Similarity (ΔSSIM) (Wang et al. 2004) to evaluate quality enhancement performance. Code is available[3].

4.3 Comparison to State-of-the-Art Methods

We compare our method with recently proposed state-of-the-art methods on MFQEv2 dataset, including: **AR-CNN** (Dong et al. 2015), **DnCNN** (Zhang et al. 2017), **MFQE2.0** (Guan et al. 2019), **STDF-R3L** (Deng et al. 2020), and **RFDA** (Zhao, Xu, and Zhou 2021). The results of these methods are partially cited from (Zhao, Xu, and Zhou 2021). The latest **BasicVSR++** (Chan et al. 2022) shows excellent performance in video low-level tasks (Yang 2021, 2022). The official release of BasicVSR++ is pre-trained with other datasets and fine-tuned with complicated boosting schemes, so we retrain it on the MFQEv2 dataset (QP37 and QP32) with the same settings as the other baselines for a fair comparison.

Overall performance. Table 1 presents the results of PSNR/SSIM improvement. The results show that our method performs among the best across five QPs in terms

[1]https://media.xiph.org/video/derf

[2]https://vqeg.org/video-datasets-and-organizations.aspx

[3]https://github.com/WilliammmZ/STCF

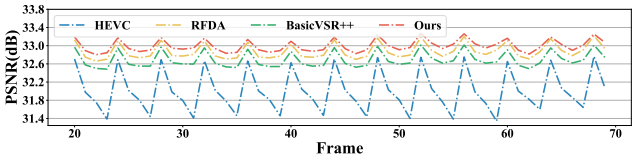

Figure 3: PSNR curves of HEVC, RFDA, BasicVSR++, and ours on test sequence *Cactus* at *QP*=37.

Method	QP27	QP32	QP37
HEVC	1.07 / 0.83	1.38 / 0.82	1.42 / 0.79
AR-CNN	1.07 / 0.83	1.38 / 0.82	1.44 / 0.80
DnCNN	1.06 / 0.83	1.40 / 0.83	1.44 / 0.80
DCAD	1.07 / 0.83	1.39 / 0.83	1.45 / 0.80
DS-CNN	1.07 / 0.83	1.39 / 0.83	1.46 / 0.80
MFQE 2.0	0.77 / 0.74	0.98 / 0.70	0.96 / 0.67
RFDA	0.63 / 0.61	0.70 / 0.63	0.69 / **0.61**
BasicVSR++	-	0.73 / 0.67	0.71 / 0.66
Ours	**0.57 / 0.58**	**0.62 / 0.59**	**0.61 / 0.61**

Table 2: Averaged PVD/SD of test sequences for PSNR at *QP*=27, 32, and 37.

of average ΔPSNR and ΔSSIM. Our method achieves **0.06-0.23 dB (up to 27%)** improvement in PSNR compared with RFDA. Besides, Our method achieves **0.07-0.18 dB (up to 20%)** improvement in PSNR compared with BasicVSR++. Similar improvements can be observed for SSIM. However, the RFDA employs the motion compensation among all preceding frames. BasicVSR++ adopts bidirectional propagation to exploit motion compensation from the entire input video. Differently, our proposed method enhances the compressed frame by global context and motion compensation information explored from only 6 neighboring frames. It evidences the importance of the fused compensation. These all verify the superiority of our method in video compression artifact reduction.

Quality Fluctuation. Quality fluctuation is another index to evaluate the quality of a whole enhanced video (Guan et al. 2019). Drastic quality fluctuation often causes severe temporal inconsistency and degradation of QoE. We evaluate the fluctuation by Standard Deviation (SD) and Peak-Valley Difference (PVD) (Xu et al. 2019) on each test sequence. The averaged PVD and SD over all test sequences are presented in Table 2. We can see that our method has the smallest averaged PVD and SD, which means our approach is more stable than the other baselines. Figure 3 shows a group of PSNR curves. Each group consists of four PSNR curves of a test sequence compressed by HEVC and the three corresponding sequences enhanced by STDF-R3L, BasicVSR++, and our method. Compared with the other methods, our method achieves larger PSNR improvement over the compressed frames but minor fluctuation.

Rate-distortion. Here, we evaluate the rate-distortion of our method and compare it with state-of-the-art approaches. For simplicity of illustration, we present the results of the compressed videos (HEVC), two state-of-the-art methods (RFDA and BasicVSR++), and our method in Figure 4.

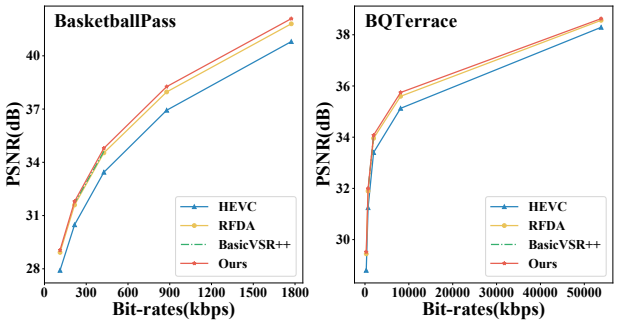

Figure 4: Rate-distortion performance on two test videos in MFQEv2. Our method outperforms the other approaches at various bitrates.

Method	Run-time	#Params	ΔPSNR
RFDA	582.4	1.27	0.91
BasicVSR++	172.0	7.40	0.95
Ours	1486.9	1.96	1.02
Ours-Light	842.6	1.13	0.99

Table 3: Inference run-time (ms) and parameter scale (M) comparison. For run-time, all methods are retested on a 720p video with the NVIDIA Tesla P40 GPU.

Due to the lack of data, it only displays the results of the BasicVSR++ in QP37 and QP32. From Figure 4, we can see that for a similar bitrate, our method gets a larger PSNR than the other approaches, which indicates that our method performs better than the other approaches in rate-distortion.

Efficiency of the methods. To explore the inference efficiency, we calculate the inference run-time and parameter size metrics between our method and other state-of-the-art methods. As our method has to explore both motion compensation and global context, it consumes more time than the other methods. However, the run-time of our method is acceptable, as our method has a smaller number of parameters. Specifically, our method gets an averaged ΔPSNR improvement of 7.4% (from 0.95 dB to 1.02 dB for QP37) and has about 73.5% fewer parameters than BasicVSR++. It demonstrates that our method achieves a good trade-off between performance and efficiency. We also provide a lightweight version framework with only 8 PSCF blocks. A smaller number of parameters allows us to test with a larger batch size, which can further improve the run-time.

Qualitative Results. As shown in Figure 5, we conduct a qualitative comparison to compare the reduction effect of different methods. The compressed patches in the $2nd$ column suffer from various compression artifacts (the ringing in BasketballDrive, and the blurring and edge floating in BQTerrace). Although image artifacts reduction methods can somehow reduce those artifacts, the enhanced frames usually become over-blurred and lack details, while STDF-R3, RFDA, and BasicVSR++ suffer from over-smoothing. However, our method restores more details or textures in the enhanced frame. All these indicate the strong capability of our method for handling distortions.

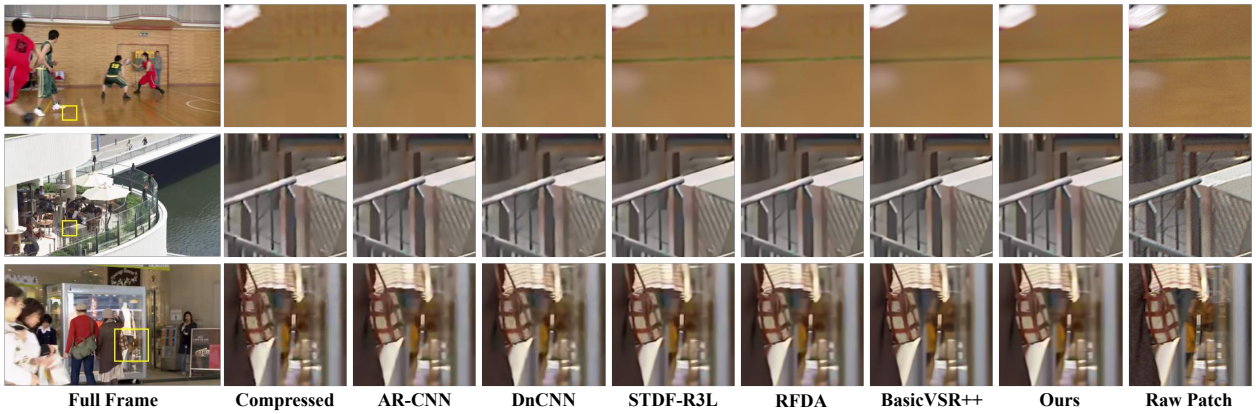

Figure 5: Qualitative results on the state-of-the-art methods and our method. The test video name (from top to bottom): BasketballDrive, BQTerrace, and BQMall.

Method	Fusion Scheme	ΔPSNR	ΔSSIM
Only Ada-CNN	-	0.78	1.48
Only Swin	-	0.89	1.60
Full PSCF	Add	0.97	1.73
	Concat & Conv	**0.99**	**1.77**

Table 4: Ablation study of the PSCF block with $QP = 37$. We train our methods with 8-layer PSCF and fixed window size 8 for quick experiment. (Conv: 1×1 convolution.)

#Blocks (N)	Window size (M)	ΔPSNR	ΔSSIM
4		0.93	1.68
8	8	0.99	1.77
16		**1.02**	**1.81**
	4	0.95	1.71
8	7	0.94	1.72
	10	0.96	1.72

Table 5: Ablation study on the number of PSCF blocks and the window size of Swin transformer layer with $QP = 37$.

4.4 Ablation Study

The effect of fused compensation (PSCF block). According to the results in Table 4, the Full PSCF method outperforms all methods that only consider motion compensation (Row 1) or global context (Row 2). To better explain the benefits of fusion compensation, we display the enhanced frames and learned residuals in Figure 6. We find that the residual learned by fusion compensation can capture more details. The compressed frame enhanced by fusion compensation has fewer compression artifacts and restores the texture clearer. This phenomenon further proves the effectiveness of fusing motion compensation and global context. As shown in Table 4 (Row 3-4), two fusion schemes have been applied to fuse the compensations. We select the Concat&Conv scheme to merge the motion compensation and global context for a better result.

The number of PSCF blocks. The number of PSCF blocks N is a vital configuration for our proposed model.

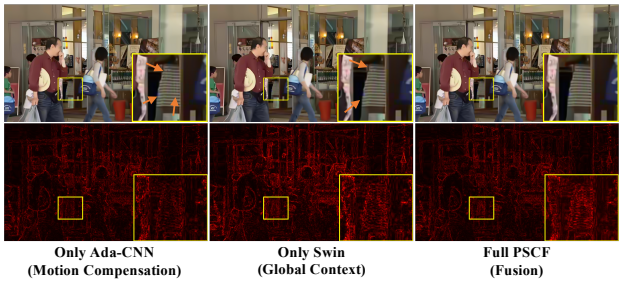

Figure 6: Visualization of the enhanced frames and learned residuals with different models.

We vary the N from 4 to 16 with a fixed window size of 8. The results are shown in Table 5. We observe that increasing N from 4 to 16 brings impressive improvements to ΔPSNR and ΔSSIM. It reflects the superiority of the PSCF block.

The window size of Swin transformer layer. The Swin transformer layer is applied in the PSCF block as part of temporal self-attention to capture the global context compensation. The window size M controls the performance of the Swin transformer. We tune M from 4 to 10 and show the results in Table 5 (Row 2 and 4-6). The ΔPSNR and ΔSSIM change slightly for different M, except $M = 8$. We select $M = 8$ as the default window size for the optimal result.

5 Conclusion

This paper proposes a novel Spatio-Temporal Compensation Fusion framework for removing video compression artifacts by learning and fusing the motion compensation and global context. In addition, a Swin-based temporal self-attention approximation strategy is introduced for efficiently capturing the global context. Extensive experiments demonstrate that our method can improve the quality of compressed videos considerably, reduce artifacts effectively, and outperform all the existing methods. Moreover, the Parallel Swin-CNN Fusion block can be easily adapted to the existing multi-frame methods of video-related tasks. Thus, we plan to extend our method to other low-level video tasks.

Acknowledgments

This work is supported in part by State Grid Corporation of China (Grant No. 5500-202011091A-0-0-00) and the National Natural Science Foundation of China (Grant No. 62072469).

References

Apostolidis, E.; Balaouras, G.; Mezaris, V.; and Patras, I. 2021. Combining global and local attention with positional encoding for video summarization. In *2021 IEEE International Symposium on Multimedia (ISM)*, 226–234. IEEE.

Bossen, F. 2011. 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*.

Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5972–5981.

Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; and Barlaud, M. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, volume 2, 168–172. IEEE.

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

Deng, J.; Wang, L.; Pu, S.; and Zhuo, C. 2020. Spatio-temporal deformable convolution for compressed video quality enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 10696–10703.

Ding, Q.; Shen, L.; Yu, L.; Yang, H.; and Xu, M. 2021. Patch-wise spatial-temporal quality enhancement for HEVC compressed video. *IEEE Transactions on Image Processing*, 30: 6459–6472.

Dong, C.; Deng, Y.; Loy, C. C.; and Tang, X. 2015. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE international conference on computer vision*, 576–584.

Guan, Z.; Xing, Q.; Xu, M.; Yang, R.; Liu, T.; and Wang, Z. 2019. 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.

Hoang, T. M.; and Zhou, J. 2021. Recent trending on learning based video compression: A survey. *Cognitive Robotics*, 1: 145–158.

Hou, D.; Zhao, Y.; and Wang, R. 2021. Video Compression Artifacts Removal with Efficient Non-local Block. In *2021 3rd International Conference on Advances in Computer Technology, Information Science and Communication (CTISC)*, 232–237. IEEE.

Jin, Z.; An, P.; Yang, C.; and Shen, L. 2018. Quality Enhancement for Intra Frame Coding Via Cnns: An Adversarial Approach. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1368–1372. IEEE.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1833–1844.

Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Cheng, Y.; Wei, W.; Xu, Z.; and Xie, Y. 2021a. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11235–11244.

Liu, D.; Wen, B.; Fan, Y.; Loy, C. C.; and Huang, T. S. 2018. Non-local recurrent network for image restoration. *Advances in neural information processing systems*, 31.

Liu, X.; Wu, X.; Zhou, J.; and Zhao, D. 2016. Data-driven soft decoding of compressed images in dual transform-pixel domain. *IEEE Transactions on Image Processing*, 25(4): 1649–1659.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Lu, G.; Zhang, X.; Ouyang, W.; Xu, D.; Chen, L.; and Gao, Z. 2019. Deep non-local kalman network for video compression artifact reduction. *IEEE Transactions on Image Processing*, 29: 1725–1737.

Mu, J.; Xiong, R.; Fan, X.; Liu, D.; Wu, F.; and Gao, W. 2020. Graph-based non-convex low-rank regularization for image compression artifact reduction. *IEEE Transactions on Image Processing*, 29: 5374–5385.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

Song, Q.; Xiong, R.; Fan, X.; Liu, D.; Wu, F.; Huang, T.; and Gao, W. 2020. Compressed image restoration via artifacts-free PCA basis learning and adaptive sparse modeling. *IEEE Transactions on Image Processing*, 29: 7399–7413.

Sullivan, G. J.; Ohm, J.-R.; Han, W.-J.; 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.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, T.; Chen, M.; and Chao, H. 2017. A novel deep learning-based method of improving coding efficiency from the decoder-end for HEVC. In *2017 Data Compression Conference (DCC)*, 410–419. IEEE.

Wang, Z.; 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.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17683–17693.

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.

Xu, Y.; Gao, L.; Tian, K.; Zhou, S.; and Sun, H. 2019. Non-local convlstm for video compression artifact reduction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7043–7052.

Xu, Y.; Zhao, M.; Liu, J.; Zhang, X.; Gao, L.; Zhou, S.; and Sun, H. 2021. Boosting the performance of video compression artifact reduction with reference frame proposals and frequency domain information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 213–222.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8): 1106–1125.

Yang, R. 2021. NTIRE 2021 challenge on quality enhancement of compressed video: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 647–666.

Yang, R. 2022. NTIRE 2022 Challenge on Super-Resolution and Quality Enhancement of Compressed Video: Dataset, Methods and Results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yang, R.; Xu, M.; Liu, T.; Wang, Z.; and Guan, Z. 2018a. Enhancing quality for HEVC compressed videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7): 2039–2054.

Yang, R.; Xu, M.; and Wang, Z. 2017. Decoder-side HEVC quality enhancement with scalable convolutional neural network. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 817–822. IEEE.

Yang, R.; Xu, M.; Wang, Z.; and Li, T. 2018b. Multi-frame quality enhancement for compressed video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6664–6673.

Yu, J.; Fan, Y.; Yang, J.; Xu, N.; Wang, Z.; Wang, X.; and Huang, T. 2018. Wide activation for efficient and accurate image super-resolution. *arXiv preprint arXiv:1808.08718*.

Zeng, K.; Zhao, T.; Rehman, A.; and Wang, Z. 2014. Characterizing perceptual artifacts in compressed video streams. In *Human Vision and Electronic Imaging XIX*, volume 9014, 173–182. SPIE.

Zhang, K.; Li, Y.; Liang, J.; Cao, J.; Zhang, Y.; Tang, H.; Timofte, R.; and Van Gool, L. 2022. Practical blind denoising via swin-conv-unet and data synthesis. *arXiv preprint arXiv:2203.13278*.

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, S.; Peng, H.; Fu, J.; and Luo, J. 2020a. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12870–12877.

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.

Zhang, Y.; Li, K.; Zhong, B.; and Fu, Y. 2019. Residual Non-local Attention Networks for Image Restoration. In *International Conference on Learning Representations*.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2020b. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7): 2480–2495.

Zhao, M.; Xu, Y.; and Zhou, S. 2021. Recursive fusion and deformable spatiotemporal attention for video compression artifact reduction. In *Proceedings of the 29th ACM International Conference on Multimedia*, 5646–5654.

Zhu, W.; Lu, J.; Li, J.; and Zhou, J. 2020. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30: 948–962.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9308–9316.