PRELIMINARY PREPRINT VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference.

Transcoded Video Restoration by Temporal Spatial Auxiliary Network

Li Xu[1], Gang He[1,2*], Jinjia Zhou[3], Jie Lei[1], Weiying Xie[1], Yunsong Li[1], Yu-Wing Tai[2]

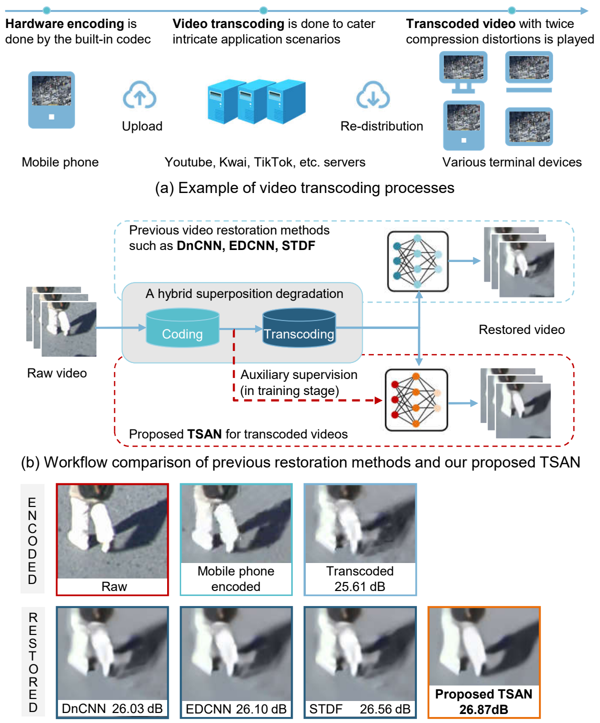
[1]Xidian University, China [2]Kuaishou Technology, China [3]Hosei University, Japan
cherylxu@stu.xidian.edu.cn, {ghe, wyxie}@xidian.edu.cn, zhou@hosei.ac.jp,
{jielei, ysli}@mail.xidian.edu.cn, yuwing@gmail.com

Abstract

In most video platforms, such as Youtube, Kwai, and TikTok, the played videos usually have undergone multiple video encodings such as hardware encoding by recording devices, software encoding by video editing apps, and single/multiple video transcoding by video application servers. Previous works in compressed video restoration typically assume the compression artifacts are caused by one-time encoding. Thus, the derived solution usually does not work very well in practice. In this paper, we propose a new method, temporal spatial auxiliary network (TSAN), for transcoded video restoration. Our method considers the unique traits between video encoding and transcoding, and we consider the initial shallow encoded videos as the intermediate labels to assist the network to conduct self-supervised attention training. In addition, we employ adjacent multi-frame information and propose the temporal deformable alignment and pyramidal spatial fusion for transcoded video restoration. The experimental results demonstrate that the performance of the proposed method is superior to that of the previous techniques. The code is available at https://github.com/icecherylXuli/TSAN.

1 Introduction

Massive uncompressed video data is a substantial burden for hardware storage and transmission. Without proper encoding, it is almost impossible to transmit a high resolution video through the widespread 4G/5G network in real-time. Over the past decades, many video coding algorithms have been emerged to exploit video contents' spatial and temporal redundancy while pursuing an acceptable visual quality. The classic video coding standards, such as H.262/MPEG-2 (Rec 1994), H.263 (Recommendation 1998), H.264/AVC (Telecom et al. 2003), H.265/HEVC (Sullivan et al. 2012), have been developed rapidly. Indeed, due to the limitations of transmission conditions and various mobile devices, almost all the videos we encounter on the Internet were transcoded to meet the target bitrates for efficient transmission. For example, a video taken directly using our mobile phone needs to be compressed at least twice (hardware encoding in mobile phone before uploading, and the transcoding in video server before re-distribution) before it can be shared on Youtube, Kwai or TikTok, etc., as illustrated in Fig. 1 (a). Intuitively, transcoded videos that end users viewed usually suffer from single/multiple transcoding degradation.


Hardware encoding is done by the built-in codec
Video transcoding is done to cater intricate application scenarios
Transcoded video with twice compression distortions is played
Mobile phone
Upload
Youtube, Kwai, TikTok, etc. servers
Re-distribution
Various terminal devices
Previous video restoration methods such as **DnCNN, EDCNN, STDF**
A hybrid superposition degradation
Coding
Transcoding
Raw video
Restored video
Auxiliary supervision (in training stage)
Proposed **TSAN** for transcoded videos
ENCODED
Raw
Mobile phone encoded
Transcoded 25.61 dB
RESTORED
DnCNN 26.03 dB
EDCNN 26.10 dB
STDF 26.56 dB
Proposed TSAN 26.87dB

(a) Example of video transcoding processes

(b) Workflow comparison of previous restoration methods and our proposed TSAN

(c) Example of subjective quality comparison

Figure 1: (a) Example of video transcoding processes. (b) Workflow comparison of previous restoration methods and our proposed TSAN. (c) Example of subjective quality comparison, including our TSAN, DnCNN (Zhang et al. 2017), EDCNN (Pan et al. 2020), and STDF (Deng et al. 2020).

Recently, some learning-based methods have been explored to eliminate the artifacts of compressed images and videos. However, they focused on improving the quality of encoded videos without transcoding. The integral distortion by video encoding and transcoding is not a simple superposition of two coding distortions, as transcoding will further introduce different degrees of distortions in the degraded regions with severe or slight artifacts. In Fig.1 (c), only the

[*]Corresponding author.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

road texture is lost in the mobile phone encoded image, but severe artifacts appear on the character's lower body and its shadow in the transcoded image. This makes the transcoded video restoration task difficult. To the best of our knowledge, there is currently no learning-based works dedicated to improving the quality of transcoded videos. A straightforward solution is to retrain these previous algorithms and reuse them to transcoded videos. Fig. 1 (b) shows the workflow comparison of previous methods and our TSAN for transcoded videos. As demonstrated in Fig.1 (c), the performance of previous methods is limited because they are specifically designed for the video whose contents only suffer from single compression. Considering that the distortions of transcoded videos are a hybrid superposition of twice or more encoding distortions, we propose a temporal spatial auxiliary network to enhance transcoded videos.

The main contributions can be summarized as follows:

- This study is the first exploration on transcoded video restoration with deep neural networks. We reveal most videos suffer from transcoded deterioration and verify these previous learning-based restoration algorithms for video encoding are fragile for video transcoding.
- We propose a network paradigm that uses the initial encoding information as an auxiliary supervised label to assist the network training. More specifically, we design auxiliary supervised attention and global supervised reconstruction modules to improve the algorithm performance in a coarse-to-fine manner.
- We design a temporal deformable module capable of offsetting the motion in a progressive motion-learning manner while extracting abundant valuable features. Furthermore, a pyramidal spatial fusion adopting four diverse downsampling filters is developed to capture more lossy details at multiple spatial scaling levels.
- We quantitatively and qualitatively demonstrate our proposed method is superior to that of the previous methods.

2 Related Work

2.1 Encoded Image/Video Restoration

Inspired by the success of deep learning, a large number of recent works (Liu et al. 2020; Dong et al. 2015; Zhang et al. 2017; Dai, Liu, and Wu 2017; Yang et al. 2018a; He et al. 2018; Ding et al. 2019; Xue et al. 2019; Yang et al. 2018b; Guan et al. 2019; Deng et al. 2020) have shown that convolutional neural network (CNN) achieves excellent performance in enhancing quality of compressed images and videos. ARCNN (Dong et al. 2015) is the first work to leverage CNN to reduce artifacts caused by JPEG. Next, benefitting from its robustness, DnCNN (Zhang et al. 2017) is often used as an image restoration baseline, including denoising, artifacts reduction, and so forth. As a learning-based post-processing method, VRCNN (Dai, Liu, and Wu 2017) was developed to promote the HEVC intra coding frames' quality. Later, quality enhancement convolutional network (QECNN) was proposed to improve the quality for I frames and P/B frames of HEVC separately. Analyzing the characteristics of video coding, a partition-masked CNN (He et al. 2018) was designed, which employed the decoded frames' partition information to instruct the network to improve performance. Meanwhile, Pan et al. (2020) designed an enhanced deep convolutional neural network (EDCNN) as an efficient in-loop filter to remove annoying artifacts and achieve a better quality of experience.

Due to the lack of effective use of adjacent information which is capable of providing supplementary details, these single-image encoded restoration works can improve the quality of damaged frames, but their improvement abilities are limited. Multi-image encoded restoration algorithms are becoming a prevalent trend. In task-oriented flow (TOFlow) (Xue et al. 2019), the learnable motion estimation component is self-supervised to facilitate video restoration. Later, MFQE (Yang et al. 2018b) and its extended version (Guan et al. 2019) developed a PQF detector to search for the highest quality reference frame, thus improving the damaged videos' quality. Moreover, to handle motion relationships efficiently, a spatio-temporal deformable fusion scheme (Deng et al. 2020) is proposed to aggregate temporal information so as to eliminate undesirable distortions.

2.2 Dilated Convolution

Dilated convolutions (Yu and Koltun 2015), also called atrous convolutions, can expand receptive fields while keeping the same resolution of feature maps. It is widely used in semantic segmentation (Chen et al. 2017a,b), image classification (Yu, Koltun, and Funkhouser 2017), image restoration (Yu et al. 2018; Guo et al. 2021), etc. Atrous spatial pyramid pooling(Chen et al. 2017a,b) employed atrous convolution in cascaded or parallel by adopting multiple atrous rates and was used to handle the problem of semantic segmentation at different spatial scales. In image classification, dilated residual network (Yu, Koltun, and Funkhouser 2017) was developed to improve the accuracy of downstream applications, and it outperforms its non-dilated counterparts. In (Yu et al. 2018), a generative inpainting method utilized dilated convolution to rough out the missing contents. Guo et al. (2021) proposed EfficientDeRain with a pixel-wise dilation filtering to predict multi-scale kernels for each pixel.

2.3 Deformable Convolution

Standard convolution is innately constrained in establishing geometric transformations on account of the invariable local receptive fields. The variant, deformable convolution (DConv) (Dai et al. 2017), was developed to obtain learnable spatial information with the guidance of the additional offset. Temporally deformable alignment network (Tian et al. 2018) first applied DConv to align the neighboring frames instead of optical flow to predict high-resolution videos. Inspired by (Tian et al. 2018), EDVR (Wang et al. 2019) elaborated on a pyramid manner to estimate offset more precisely. Interlayer restoration network (He et al. 2021) for scalable high efficiency video coding employed DConv to compensate the compression degradation difference through the multi-scale learnable offsets. In particular, Chan et al. (2021) revealed the relation between deformable alignment and flow-based alignment and proposed an offset-fidelity loss to alleviate the training instability.

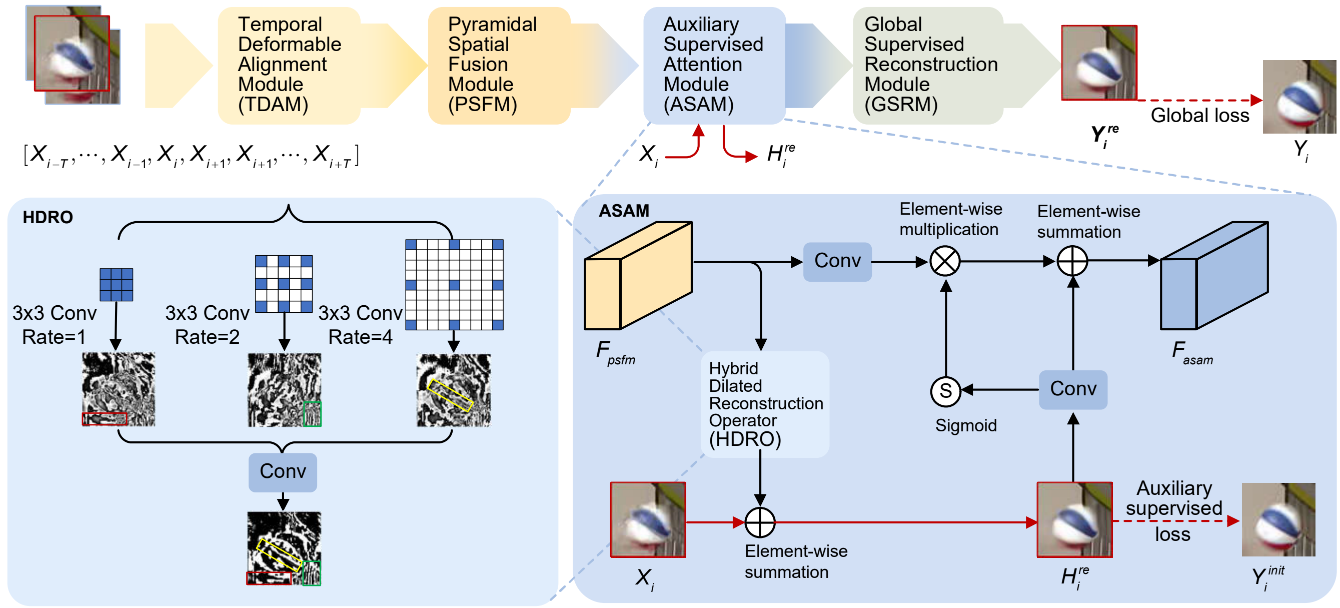

Temporal Deformable Alignment Module (TDAM)
Pyramidal Spatial Fusion Module (PSFM)
Auxiliary Supervised Attention Module (ASAM)
Global Supervised Reconstruction Module (GSRM)
$[X_{i-T}, \cdots, X_{i-1}, X_i, X_{i+1}, X_{i+1}, \cdots, X_{i+T}]$
X_i
H_i^{re}
Y_i^{re}
Global loss
Y_i
HDRO
3x3 Conv Rate=1
3x3 Conv Rate=2
3x3 Conv Rate=4
Conv
ASAM
F_{psfm}
Conv
Element-wise multiplication
Element-wise summation
F_{asam}
Hybrid Dilated Reconstruction Operator (HDRO)
S
Sigmoid
Conv
X_i
Element-wise summation
H_i^{re}
Auxiliary supervised loss
Y_i^{init}

Figure 2: Architecture of our temporal spatial auxiliary network (TSAN) and structure of auxiliary supervised attention module.

3 Methodology

3.1 Preliminary

Transcoded videos usually suffer from multiple encoding degradations and are hard to be restored due to the complicated unknown process. Table 1 summarizes the performance of the classic restoration method (Deng et al. 2020) on one-time video encoding and transcoding in three test sequences. Note that the DNN-based method was retrained on the dataset which covers the same videos but suffers from different degradation. The typical STDF was originally designed for improving the quality of the distorted videos compressed by the H.265/HEVC reference software HM16.5. Here we also follow the same configuration to prepare the training dataset. The procedure of preparing transcoding dataset will be depicted in Sec. 4.1. In Table 1, although the qualities (peak signal-to-noise, PSNR) of the three listed sequences suffering one-time encoding and transcoding fluctuate, they are similar. However, the improvement of video transcoding by STDF decreased from 1.051 dB to 0.646 dB in sequence *BQMall*. Likewise, when applying STDF to video transcoding, different degrees of degradations are observed in all the listed sequences. These results validate that *the previous learning-based methods for video encoding are fragile for video transcoding*.

3.2 Network Architecture Overall

As discussed above, the artifacts of compressed videos are mostly a combination of video encoding and transcoding distortions. Hence, we design a progressive restoration network that divides the transcoded video restoration task into two parts and restores the transcoded videos in a coarse-to-fine manner. The first part focuses on removing the distortion introducing by transcoding at a lower bitrate. Then, the second part is inclined to eliminate the artifacts caused by the initial video encoding. For this purpose, *we employ the initial encoded video with a high bitrate as an additional auxiliary supervised label to assist network at training stage.* We denote our method, temporal spatial auxiliary network (TSAN).

Sequences	Status	One-time Encoding (QP 37)	Transcoding (1 or 0.5 Mbps)	Δ
Basketball-Drill	Before	31.591	31.747	-
	After	32.409	32.303	-
	ΔPSNR	0.818	0.556	**-0.262**
BQMall	Before	31.297	31.949	-
	After	32.340	32.595	-
	ΔPSNR	1.051	0.646	**-0.405**
BQTerrace	Before	31.247	31.565	-
	After	31.894	31.954	-
	ΔPSNR	0.647	0.389	**-0.258**

Table 1: Comparison of restoration on one-time encoding and transcoding in terms of PSNR (dB). “Before” status denotes the sequence has been compressed but not enhanced. “After” denotes the sequence has been enhanced by the retrained STDF (Deng et al. 2020).

Given $2T+1$ consecutive low-quality frames $X_{[i-T:t+T]}$, we denote the center frame X_i as the target frame need to be restored and the other frames as the reference frames. The input of the network is the target frame X_i and the $2T$ neighboring frames, and the output is the enhanced target frame Y_i^{re}. The objective function can be formulated as follows,

$$Y_i^{re} = \mathbf{Net_{TSAN}}(X), \tag{1}$$

where $\mathbf{Net_{TSAN}}$ is our proposed temporal spatial auxiliary network and X is a stack of transcoded frames which is defined as

$$X = [X_{i-T}, \cdots, X_{i-1}, X_i, X_{i+1}, \cdots, X_{t+T}], \tag{2}$$

where i denotes the frame number and T is the maximum number of reference frames.

The architecture of TSAN is shown in Fig. 2. TSAN is devised to estimate a high quality output with the guidance of its consecutive transcoded frames. In the following subsection, we will give detailed analysis on the motivation and rationality of each module.

3.3 Auxiliary Supervised Attention

The temporal deformable alignment and pyramidal spatial fusion modules serve for the severe distortions where the contents have been degraded repeatedly. However, the lossy information is hard to be recovered due the hybrid transcoding degradation. Combining the traits of video encoding and transcoding, we proposed an auxiliary supervised attention module (ASAM) whose structure is illustrated in the blue part of Fig. 2. First, we use a hybrid dilation reconstruction operator (HDRO) to predict the high-frequency map. Specifically, we apply three dilated convolutions with different dilation rates ($r = 1, 2, 4$) to reconstruct the lossy frequency maps at different spatial scale. Intuitively, the 3×3, 5×5, and 7×7 receptive fields of each pixel are supported by these convolutions and the various highlighted texture, like the red, green, and yellow rectangular boxes, are generated. Note that the 1-dilated convolution is equivalent to the standard convolution. Following the parallel dilated convolutions, these frequency maps are sent into a 3×3 standard convolution to yield the integrated result. The process of HDRO is given by:

$$F_{HDRO} = Rec([D_1(F_{psfm}), D_2(F_{psfm}), D_4(F_{psfm})]), \tag{3}$$

where D_1, D_2, and D_4 denote dilated convolutions with 1, 2 and 4 dilation rates, respectively. $Rec(\cdot)$ is the 3×3 standard convolution.

After HDRO, the integrated frequency map is sum up to the low-quality target frame X_i to yield the initial restored one H_i^{re}. Up to now, the initial restoration stage is complete. Next, we feed H_i^{re} again into the convolution for providing excited features. The sigmoid activation function is used to restrict these features in [0, 1] and generate supervised attention maps. Then, these earliest input feature through a 3×3 transitional convolution and are refined by the supervised attention maps. Ultimately, these self-refined features are added to the excited features by H_i^{re} as the output of module.

In summary, ASAM plays an essential role in guiding the first part of our network to approach the intermediate lossy representation and establishing a connection between the transcoding degradation and initial encoding degradation. In the second part, we design a global supervised reconstruction module (GSRM) which consists of 10 residual blocks and the HDRO with a short-cut connection of target frame X_i. With the help of GSRM, the final restored output Y_i^{re} is reconstructed.

3.4 Temporal Deformable Alignment

Adjacent frames are essential for target frame restoration but they are not equally informative due to view angle, motion blocking, and video compression problems. Hence, we leverage the neighboring forward and backward frames to exploit more advantageous information. In Fig. 3, we employ optical flow estimation (OFE), i.e. SPyNet (Ranjan and Black 2017), to compute the forward and backward optical flows among the adjacent frames. Then, these estimated optical flows are regarded as the plain motion information and are warped with the input frames to yield the plain aligned frames. Here we describe the forward frame alignment in detail, and the backward frame alignment can be inferred similarly. The forward alignment at $i{-}t{-}1$ timestep is done,

$$X^A_{i\text{-}t\text{-}1 \to i\text{-}t} = warp(\mathbf{OFE}(X_{i-t-1}, X_{i-t}), X_{i-t-1}), \tag{4}$$

where $t \in [0, T)$ and $\mathbf{OFE}(\cdot)$ is optical flow estimation. We use bilinear interpolation to implement $warp(\cdot)$ function.

Later, these initial aligned frames X^A and the target frame X_i are concatenated together to send into feature excitation (FE) and generate the motion refinements on the basis of plain motion estimation. Three stacked plain 3×3 convolution layers are adopted to feature excitation. Integrating the plain motion transformed by a convolutional filter and the motion refinements, the more progressive refined motion information is generated, and it is regarded as the learnt predicted offset $\triangle P$ to help the explicit temporal deformable alignment. The mathematical equation is

$$F_{tdam}(p_0) = \sum_{p_k \in \mathcal{R}} \omega_k \cdot X(p_0 + p_k + \triangle p_k), \tag{5}$$

the deformable convolution will be operated on the deformed sampling locations $p_k + \triangle p_k$, where ω_k and p_k denote the weight and predicted offset for k-th location in $\mathcal{R} = \{(-1,-1), (-1,0), \cdots, (0,0), \cdots, (0,1), (1,1)\}$. Note that the $\triangle p_k \in \triangle P$. Finally, a stack of temporal deformable alignment features F_{tdam} are acquired by compensating the motion in a progressive motion-learning manner.

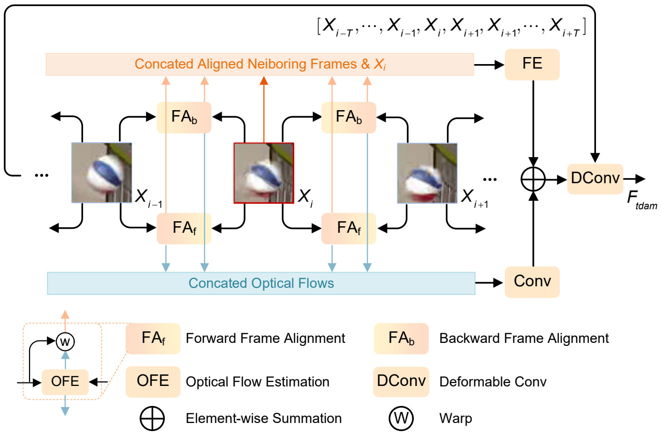


Figure 3: Structure of temporal deformable alignment.

3.5 Pyramidal Spatial Fusion

After obtaining the temporal deformable aligned features F_{tdam}, we design a pyramidal spatial fusion module (PSFM) based on an UNet (Ronneberger, Fischer, and Brox 2015)

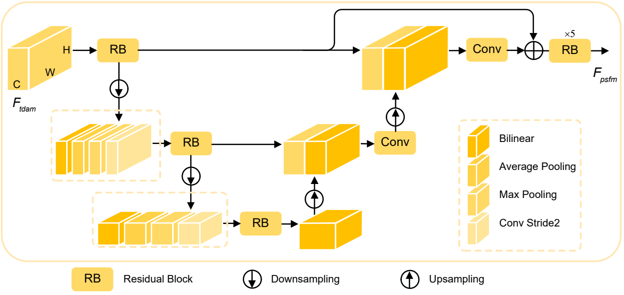


Figure 4: Structure of pyramidal spatial fusion.

structure to learn the contextual information at multiple spatial levels. The notable differences between our PSFM and UNet are two-fold.

First, four diverse downsampling filters, including the bilinear operator, average pooling, max pooling, and strided convolutional filters, are adopted instead of single downsampling method. Following this pyramidal downsampling manner, we enlarge the aligned features' receptive filed to merge the neighboring information from the temporal dimension to spatial dimension and capture more lossy details at multiple scaling levels. Second, residual learning with a single shortcut connection is introduced to replace the plain 3×3 convolutions. This variant is conducive to propagate the proceeding learnt valuable information from the shallower layers to the deeper ones, and ameliorate gradient vanishing and explosion. After that, we employ 5 residual blocks to generate enhanced features.

3.6 Loss Function

We develop a loss function which consists of auxiliary supervised loss function and global supervised loss function. The initial shallow encoded video Y_i^{init} with a high bitrate is regarded as the auxiliary supervised label, and the partial loss function is calculated as:

$$Loss_a = 1/N \sum_{i=1}^{N} \left\| Y_i^{init} - H_i^{re} \right\|^2, \tag{6}$$

where N is the batch size and MSE loss is adopted for optimization. Meanwhile, the global supervised loss function is calculated as:

$$Loss_g = 1/N \sum_{i=1}^{N} \| Y_i - Y_i^{re} \|^2, \tag{7}$$

where Y_i denotes the raw target frame without any compression and Y_i^{re} denotes the enhanced one.

Our network is an end-to-end method and the two loss functions are combined together as the final loss of the entire network for back propagation process. It is defined as:

$$Loss = \alpha \cdot Loss_a + \beta \cdot Loss_g, \tag{8}$$

where α and β are the weight factors. Validating by the related experiment, we set $\alpha = 0.2$ and $\beta = 0.8$.

Settings		Initial Encoding	One-time Transcoding
Bitrate	HR[1]	10 Mbps	1000 kbps
	LR[2]	10 Mbps	500 kbps
Coding Software		X265	X265
Preset		Medium	Medium
Rate Control Mode		Average Bitrate	Average Bitrate
Loop Filters		Deblock and SAO	Deblock and SAO
Group of Pictures		250	250
Max References		3	3

[1] HR denotes high resolution videos higher than 720p.
[2] LR denotes low resolution videos lower than or equal to 720p.

Table 2: Detailed settings of video initial encoding and one-time transcoding. The bitrate is set according to the resolution of videos and the other settings are same.

4 Experiments and Analyses

4.1 Experimental Settings

Training and Testing Dataset. To establish a training dataset for video transcoding restoration, we employed 108 sequences from Xiph.org (Xiph.org), VQEG (VQEG), and Joint Collaborative Team on Video Coding (JCT-VC) (Bossen et al. 2013). The resolutions of these sequences cover SIF, CIF, 4CIF, 360p, 480p,1080p, and 2k. We adopt all 18 standard test sequences from JCT-VC for testing.

Encoding and Transcoding Settings. All videos in the training and testing dataset have been processed by initial encoding and one-time transcoding. The detailed settings have been listed in Table 2. The video resolutions have been kept the same and the same encoding tool x265 (x265 Developers) has been adopted in both the initial encoding and one-time transcoding. Both x265 presets have been set as "medium" and others settings including rate control strategy, group of pictures (GOP) size, loop filters, etc. are all the same default ones except the bitrates for them. The initial encoding bitrate is set as the high bitrate 10 Mbps. This is because we calculated that the average bitrate of 100 videos taken by iPhone12 is close to 10 Mbps. Meanwhile, the transcoding bitrate is set as 500/1000 kbps according to different resolutions to simulate the cases of real practical applications such as TikTok. We have randomly downloaded about 200 hundred videos of TikTok and do the statistical bitrate data for them.

Implementation Details. We implement our TSAN with Pytorch 1.6.0 framework on a NVIDIA GeForce 2080Ti GPU and update it with Adam optimizer. The batch size is set to 16 and the learning rate is initialized as 1e-4. The network training stops after 300k iterations.

4.2 Transcoding Restoration Performance

Table 3 shows the performance of our method in transcoded videos, compared with DNN-based methods (Zhang et al. 2017; Yang et al. 2018a; Pan et al. 2020; Deng et al. 2020). The delta peak signal-to-noise ratio (△PSNR) and delta structural similarity index metric (△SSIM) are calculated.

The average PSNR gain is 0.782 dB when the resolutions of videos vary from 416×240 to 2560×1600, and the highest PSNR gain is 1.264 dB in *BQSquare*. It demonstrates that

Classes	Sequences	DnCNN (Zhang et al., TIP' 17)	QECNN (Yang et al., TCSVT' 18)	EDCNN (Pan et al., TIP' 20)	STDF (Deng et al., AAAI' 20)	Proposed TSAN
A	Traffic	0.309/0.005	0.238/0.004	0.292/0.005	0.392/0.009	0.554/0.008
	PeopleOnStreet	0.423/0.027	0.302/0.019	0.490/0.030	0.950/0.046	1.260/0.054
B	BasketballDrive	0.375/0.012	0.275/0.010	0.429/0.013	0.600/0.016	0.863/0.020
	BQTerrace	0.299/0.007	0.280/0.005	0.332/0.007	0.389/0.008	0.545/0.010
	Cactus	0.320/0.010	0.240/0.008	0.342/0.010	0.480/0.012	0.743/0.017
	Kimono	0.263/0.010	0.196/0.008	0.280/0.011	0.368/0.012	0.632/0.017
	ParkScene	0.171/0.008	0.132/0.006	0.177/0.007	0.254/0.010	0.447/0.015
C	BasketballDrill	0.448/0.011	0.339/0.008	0.484/0.011	0.556/0.013	0.796/0.017
	BQMall	0.462/0.011	0.360/0.009	0.468/0.011	0.646/0.014	0.985/0.019
	PartyScene	0.241/0.012	0.174/0.009	0.264/0.012	0.360/0.016	0.539/0.022
	RaceHorses	0.274/0.014	0.189/0.011	0.299/0.015	0.417/0.018	0.707/0.026
D	BQSquare	0.443/0.004	0.357/0.002	0.564/0.004	0.613/0.004	1.264/0.011
	BasketballPass	0.366/0.007	0.373/0.004	0.457/0.005	0.675/0.009	0.948/0.013
	BlowingBubbles	0.329/0.007	0.253/0.005	0.338/0.007	0.506/0.010	0.804/0.014
	RacesHorses	0.474/0.007	0.385/0.005	0.442/0.006	0.501/0.008	0.826/0.013
E	FourPeople	0.484/0.004	0.408/0.003	0.390/0.004	0.708/0.005	0.933/0.006
	Johnny	0.256/0.003	0.129/0.002	0.099/0.002	0.291/0.003	0.480/0.004
	KristenAndSara	0.358/0.003	0.293/0.003	0.155/0.003	0.521/0.004	0.756/0.005
	Average	0.350/0.009	0.274/0.007	0.350/0.009	0.513/0.012	0.782/0.016

Table 3: Improvement (**∆PSNR/∆SSIM**) of our TSAN and previous DNN-based methods in video transcoding.

Components	V1	V2	Proposed
Temporal Deformable Alignment	✓	✓	✓
Pyramidal Spatial Fusion	×	✓	✓
Auxiliary Supervised Attention	×	×	✓

Table 4: Compositions of different proposed networks.

the proposed network can substantially improve the quality of transcoded videos. To further validate the effectiveness of our TSAN, we reimplement the four representative video restoration DNN-based methods (Zhang et al. 2017; Yang et al. 2018a; Pan et al. 2020; Deng et al. 2020). Note that the presented results are generated by the networks trained with the transcoded dataset. As can be observed, with the help of these methods, the averages of △PSNR are 0.274∼0.513 dB on the 18 test videos suffered from video encoding and transcoding. It should be noted that these previous methods have an essential effect on the video restoration of single encoding. However, when applied to transcoded video restoration, their utilities are decreased. Comparing with these previous methods for one-time compression scenarios, our TSAN can improve the average △PSNR by 0.269∼0.508 dB. We deem that this is due to the different application scenarios and well-designed network. In terms of SSIM, the average gain of proposed TSAN is 0.016, which is about 0.004∼0.009 more than previous methods. The results verify that our proposal can generate a delightful perceptual quality improvement in comparison to previous DNN-based methods. The parameter of our method is 5.75M, while the parameter of the previous method is 0.56∼3.84M. Despite achieving higher quality performance, the efficiency of our algorithm should be optimized.

4.3 Ablation Study

To validate the contribution of each component, a baseline combining components gradually is presented in Table 4.

Sequences	V1	V2	Proposed
Traffic	0.434/0.012	0.462/0.007	0.554/0.008
PeopleOnStreet	0.913/0.042	1.084/0.048	1.260/0.054
BasketballDrive	0.633/0.016	0.687/0.017	0.863/0.020
BQTerrace	0.415/0.008	0.474/0.009	0.545/0.010
Cactus	0.533/0.012	0.605/0.014	0.743/0.017
Kimono	0.468/0.014	0.528/0.015	0.632/0.017
ParkScene	0.329/0.011	0.356/0.013	0.447/0.015
BasketballDrill	0.566/0.013	0.607/0.014	0.796/0.017
BQMall	0.689/0.014	0.783/0.016	0.985/0.019
PartyScene	0.353/0.016	0.401/0.018	0.539/0.022
RaceHorses	0.537/0.021	0.551/0.023	0.707/0.026
BQSquare	0.732/0.005	1.023/0.009	1.264/0.011
BasketballPass	0.608/0.007	0.861/0.012	0.948/0.013
BlowingBubbles	0.531/0.010	0.634/0.012	0.804/0.014
RacesHorses	0.487/0.007	0.760/0.012	0.826/0.013
FourPeople	0.703/0.005	0.825/0.006	0.933/0.006
Johnny	0.352/0.003	0.520/0.004	0.480/0.004
KristenAndSara	0.589/0.004	0.747/0.005	0.756/0.005
Average	0.548/0.012	0.662/0.014	0.782/0.016

Table 5: Ablation study in terms of improvement (**∆PSNR/∆SSIM**).

We develop three variants of TSAN: the first version (V1) only consists of a temporal deformable alignment module; the second version (V2) is extended by pyramidal spatial fusion module; the final version (Proposed) includes not only the proceeding parts and but also the auxiliary supervised attention and global supervised reconstruction modules.

In practice, we listed the corresponding improvement in terms of ∆PSNR and ∆SSIM in Table 5. As shown in it, the performance gains of the three versions increase gradually and steadily, and the highest performance is obtained by our final version which is well-designed for transcoding videos. Notably, the average PSNR gain of V1 is 0.548 dB, and it outperforms those of the previous DNN-based methods (0.274∼0.513 dB). Compared with previous methods men-

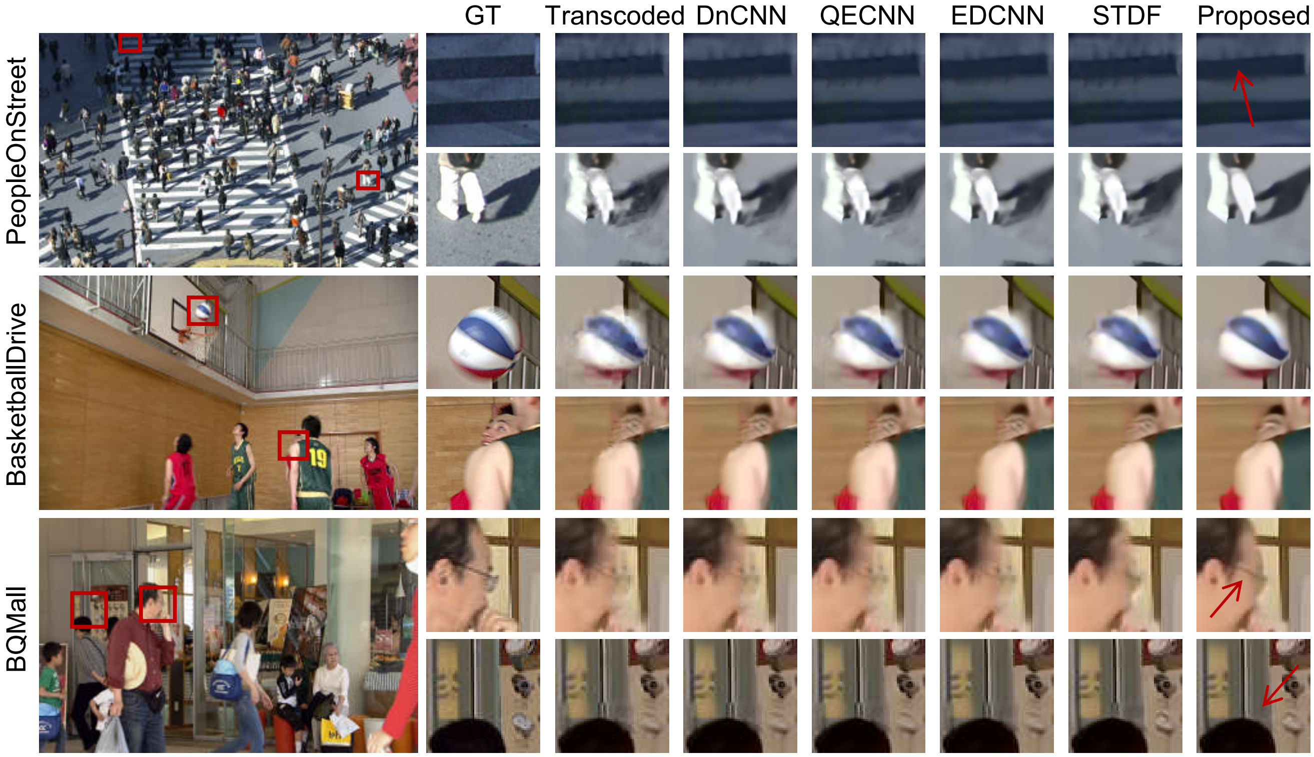


Figure 5: Subjective quality performance of our method and previous methods.

(α, β)	(0, 1)	(0.2, 0.8)	(0.5, 0.5)
ΔPSNR (dB)	0.586	0.757	0.703

Table 6: Comparison of different weight factors in TSAN loss function, Class C.

tioned above, the baseline of our proposal is more capable of eliminating the artifacts under the circumstance that the neighboring information is leveraged fully and more precise motion alignment is performed. Since the pyramidal spatial fusion scheme further explores the lossy contextual details, it brings a 20.8% increment compared with V1. It is apparent that the effectiveness of PSFM has been demonstrated. Furthermore, we verify the significant advantages of introducing auxiliary supervised attention. As mentioned above in Sec. 3.6, α and β are the weight factors that can control the proportion of auxiliary supervised loss function and global supervised loss function in the whole one. When $\alpha = 0$ and $\beta = 1$, the network removed ASAM, i.e., V2. From Table 6, we can observe that utilizing the initial encoded video with high bitrate is conducive to learning more lossy information and reduce annoying distortions. More specifically, a 1:4 combination of the auxiliary supervised loss function and global supervised one can achieve a higher result. Note that the weight factors are not optimal, because we deem that further optimization of the weight factors will bring limited benefits, and we do not conduct too many experiments to optimize it.

4.4 Subjective Performance

Fig. 5 shows the subjective quality performance of our TSAN and previous method (Zhang et al. 2017; Yang et al. 2018a; Pan et al. 2020; Deng et al. 2020) for transcoding restoration. Comparing the highlighted area of *PeopleOnStreet*, we can find that the severely distorted zebra crossing is restored and the artifacts of the shadow are removed by our method. Likewise, the texture and graininess including the basketball and basketball player's face in sequence *BasketballDrive* and the glasses frame and metal rod in sequence *BQMall* are restored to a great extent. According to the favorable subjective quality performance, we can conclude that our method can acquire not only substantial objective achievements but also pleasuring perceptual results.

5 Conclusion

In this paper, we first explore the connection and difference between one-time video encoding and transcoding. Then, we demonstrate that these previous learning-based restoration methods are not robust for video transcoding. Based on this, we proposed a network paradigm that take advantage of initial encoding information as a forepart label to instruct the network optimization. Specifically, we proposed a temporal spatial auxiliary network (TSAN) which including temporal deformable alignment, pyramidal spatial fusion, and auxiliary supervised attention mainly to improve transcoded videos. This work is the first time dedicated to transcoded video restoration and we believe that this work can arouse broad interest in video restoration community.

References

Bossen, F.; et al. 2013. Common test conditions and software reference configurations. *JCTVC-L1100*, 12(7).

Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. Understanding Deformable Alignment in Video Super-Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 973–981.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017b. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

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, D.; Kong, L.; Chen, G.; Liu, Z.; and Fang, Y. 2019. A Switchable Deep Learning Approach for In-loop Filtering in Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*.

Dong, C.; Deng, Y.; Change Loy, 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*.

Guo, Q.; Sun, J.; Juefei-Xu, F.; Ma, L.; Xie, X.; Feng, W.; Liu, Y.; and Zhao, J. 2021. EfficientDeRain: Learning Pixel-wise Dilation Filtering for High-Efficiency Single-Image Deraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1487–1495.

He, G.; Xu, L.; Lei, J.; Xie, W.; Li, Y.; Fan, Y.; and Zhou, J. 2021. Interlayer Restoration Deep Neural Network for Scalable High Efficiency Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*.

He, X.; Hu, Q.; Zhang, X.; Zhang, C.; Lin, W.; and Han, X. 2018. Enhancing HEVC compressed videos with a partition-masked convolutional neural network. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 216–220. IEEE.

Liu, D.; Li, Y.; Lin, J.; Li, H.; and Wu, F. 2020. Deep learning-based video coding: A review and a case study. *ACM Computing Surveys (CSUR)*, 53(1): 1–35.

Pan, Z.; Yi, X.; Zhang, Y.; Jeon, B.; and Kwong, S. 2020. Efficient In-Loop Filtering Based on Enhanced Deep Convolutional Neural Networks for HEVC. *IEEE Transactions on Image Processing*, 29: 5352–5366.

Ranjan, A.; and Black, M. J. 2017. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4161–4170.

Rec, I. 1994. H. 262 and ISO/IEC 13818-2 (MPEG-2 Video), Generic Coding of Moving Pictures and Associated Audio Information Part 2: Video. *ed: ITU, ISO Std., Rev.*

Recommendation, H. 1998. 263: Video coding for low bit rate communication. *ITU-T, February*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

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.

Telecom, I.; et al. 2003. Advanced video coding for generic audiovisual services. *ITU-T Recommendation H. 264*.

Tian, Y.; Zhang, Y.; Fu, Y.; and Xu, C. 2018. Tdan: Temporally deformable alignment network for video super-resolution. *arXiv preprint arXiv:1812.02898*.

VQEG. https://www.its.bldrdoc.gov/vqeg/video-datasets-and-organizations.aspx. Accessed: July 21, 2021.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

x265 Developers. https://www.x265.org. Accessed: July 24, 2021.

Xiph.org. https://media.xiph.org/video/derf/. Accessed: July 21, 2021.

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.; 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.; 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, F.; and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Yu, F.; Koltun, V.; and Funkhouser, T. 2017. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 472–480.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5505–5514.

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.