Block Based Enhancement using Deep Learning for Conversion of Low Resolution AVS Video to High Resolution HEVC Video

Rizwan Qureshi $^{1,1}$ and Mehmood Nawaz $^{2}$

$^1$National University of Computer and Emerging Sciences
$^2$Affiliation not available

November 8, 2023

Abstract

Conversion of one video bitstream to another video bitstream is a challenging task in the heterogeneous transcoder due to different video formats. In this paper, a region of interest (ROI) based super resolution technique is used to convert the low-resolution AVS (audio video standard) video to high definition HEVC (high efficiency video coding) video. Firstly, we classify a low-resolution video frame into small blocks by using visual characteristics, transform coefficients, and motion vector (MV) of a video. These blocks are further classified as blocks of most interest (BOMI), blocks of less interest (BOLI) and blocks of noninterest (BONI). The BONI blocks are considered as background blocks due to less interest in video and remains unchanged during SR process. Secondly, we apply deep learning based super resolution method on low resolution BOMI, and BOLI blocks to enhance the visual quality. The BOMI and BOLI blocks have high attention due to ROI that include some motion and contrast of the objects. The proposed method saves 20% to 30% computational time and obtained appreciable results as compared with full frame based super resolution method. We have tested our method on different official video sequences with resolution of 1K, 2K, and 4K. Our proposed method has an efficient visual performance in contrast to the full frame-based super resolution method.
Block Based Enhancement using Deep Learning for Conversion of Low Resolution AVS Video to High Resolution HEVC Video

Mehmood Nawaz¹, Member, IEEE, M.A.A Abdelgawad³, Mansoor Ali Teevno³, Rizwan Qureshi², Muhammad waqas⁴, Ali Raza Shahid³ Member, IEEE

¹Department of Diagnostic Radiology, The University of Hong Kong, Hong Kong
²National University of Computer and Emerging Sciences, Karachi Campus, Pakistan
³Department of Electrical Engineering, City University of Hong Kong, Hong Kong
⁴Department of Computer Sciences, University College of Zhob, Baluchistan University of IT, Engineering, and Management Sciences, Quetta, Pakistan

Abstract—Conversion of one video bitstream to another video bitstream is a challenging task in the heterogeneous transcoding due to different video formats. In this paper, a region of interest (ROI) based super resolution technique is used to convert the low-resolution AVS (audio video standard) video to high definition HEVC (high efficiency video coding) video. Firstly, we classify a low-resolution video frame into small blocks by using visual characteristics, transform coefficients, and motion vector (MV) of a video. These blocks are further classified as blocks of most interest (BOMI), blocks of less interest (BOLI) and blocks of non-interest (BONI). The BONI blocks are considered as background blocks due to less interest in video and remains unchanged. Secondly, we apply deep learning based super resolution method on low resolution BOMI, and BOLI blocks to enhance the visual quality. The BOMI and BOLI blocks have high attention due to ROI that includes some motion and contrast of the objects. The proposed method saves 20% to 30% computational time and obtained appreciable results as compared with full frame based super resolution method. We have tested our method on different official video sequences with resolution of 1K, 2K, and 4K. Our proposed method has an efficient visual performance in contrast to the full frame-based super resolution method.

Index Terms—Video coding standard, High efficiency video coding, Super-resolution, Block based enhancement

Impact Statement

We have proposed a new deep learning-based method to convert low-resolution video into high-resolution video. The proposed method is compared against existing methods on several benchmark datasets, and it is found that the method is superior in terms of quality, as well as computational complexity. The work has many applications, including, multimedia industry, security, and healthcare.

I. INTRODUCTION

A. Background

Nowadays, the ultra-high definition videos have become a future demand in the multimedia industry. There are various work has been proposed to achieve this objective. The high-efficiency video coding (HEVC) has been introduced by the joint collaborative team for video coding (JCT-VC). On the other hand, the audio video coding (AVC) standard working china group has developed a new audio-video coding standard (AVS) to fulfill the requirement of multimedia industry [2]–[4]. AVS has similar visual qualities to H.264/AVC with different computational cost. AVS presents a valuable substitute for H.264 within the video entertainment market in China. Two coding groups are very famous among many videos coding groups due to their effortless contribution in video trans-coding industry. One of them is VCEG (video coding expert group), and the second is the moving picture expert group (MPEG). The main valuable effort of these two group teams is HEVC. Changing the AVS format to the HEVC format is a valuable step in video coding industry. The requirement of HEVC based devices and applications are increasing due to the fast growing internet technology. As per CISCO data traffic prediction, 80% to 90% global internet circulation will contain high definition video contents in future. The fast progress in multimedia contents, the demand for better quality visualization and high-resolution videos is increasing [5].

Therefore, it is required to cut down large video data via different video compression coding standards like HEVC [17, 19, 20, 21, 22, 24 and 25] and H.264/AVS. The high resolution videos have many applications in different fields, such as healthcare [6], computer vision [7], and wireless multimedia transmissions [8].

The main goal of HEVC is to minimize the bitrate by up to 50% while giving comparable visual quality compared to its equivalent video coding standard AVC/H264. HEVC has many new prominent features; one of them is macro block sizes extension. By extending these blocks, we can easily improve the visual quality of low-resolution video. Different types of transcoders have been used to change the format of one bitstream to another bitstream [9]–[11]. The main purpose of transcoding is to fill the gap caused by the rapid changes in the digital world and the increasing capacity of video content in different formats such as AVI, VP8, MP4 and VP9, etc. In the past, the video transcoder was limited to convert the video...
heterogeneous transcoder requires a syntax conversion unit and changing the directionality of motion vectors (MV) and the picture characteristics such as the resolution, type, and rate. Therefore the implementation of this technique has many challenges [15]. Due to the output sequence having a different encoding format and the spatial temporal subsampling, the motion compensation loops of encoder and decoder in heterogeneous transcoder are very complex. Quadtree structure size is also a challenge in heterogeneous transcoding. Quadtree structure in HEVC uses a larger coding unit size \((64 \times 64)\) for estimating the object movement in the video frame, while H.264 / AVS uses \((16 \times 16)\) unit size [16]. While upsampling of low resolution frame into high resolution frame is also a well-known challenge in video transcoder.

II. LITERATURE REVIEW

There are various techniques have been proposed in the last two decades to overcome such types of problems [17]–[21]. Video enhancement has many application, such as in medical [22], underwater video enhancement [23], entertainment [24], and crime investigation [25]. A low-light weight MBLLEN network is proposed for image or video enhancement using convolutional neural networks in [26]. The main concept is extracting rich features at different levels so that we can apply augmentation via many subnets and then construct the output image via fusion of multi-branch. By this technique, the image quality is increased in various ways. Huang et al. [27] proposed a method, which adds high frequency information in videos frames. This technique did not perform well when noise ratio increased in the video frame. Chow-Sing et al. [14] proposed a fast intra transcoding based on discrete cosine transform (DCT) and prediction modes for converting from H.264/AVC to high-efficiency video coding. This method uses coefficients of DCT and the intra-prediction information embedded in H.264 bitstream to predict the coding depth map for depth limitation. Linwei et al. [16] proposed a fast H.264/AVC to HEVC transcoder based on machine learning by exploiting the similarity of block partition. This method uses a feature selection technique for predicting the quadtree coding unit partition in HEVC. Johan et al. [28] proposed a spatially misaligned HEVC transcoder with computational scalability, which minimizes the complexity of composition process and reduce some information from original bitstream. Mamoon et al. [29] proposed a transparent encryption technique with scalable video communication at lower latency. This method represents the gains in reducing delay and increasing distortion arising from a transparent encryption. Chia-Hung et al. [1] proposed a coding unit complexity for an efficient HEVC to SHVC transcoding with quality scalability based on coding unit depth and mode predictions. A survey of video compression is also given in [30]. A comprehensive survey for video enhancement methods using deep learning is presented in [31].

Assuncao et al. [15] proposed an open loop transcoder that cascades the decoder and encoder directly. The transcoding architecture does not contain a feedback loop for the drift error. The aim of this transcoding is to minimize the complexity and adjust DCT coefficients to reduce the bit
rate. Keesman et al. [32] proposed a close loop transcoder, which has the feedback loop in structure for minimizing the transcoding distortion by compensating the drift error in the transcoder. Wee-svsie et al. [10] proposed a spatial domain video transcoder to perform the dynamic bit rate adaption through rate control. Eletheriads et al. [33] proposed a frequency domain transcoder that used variable-length decoding (VLD) and inverse quantization to extract each block’s DCT value at the decoder end. In this technique, the residual errors of motion compensation are encoded through variable length coding (VLC) and re-quantization [2]. [34]. A scalable omni-directional video coding scheme for virtual reality applications is presented in [35]. Xie et al. [29] proposed a hybrid domain transcoder, which provides the trade-off between the video quality and computational cost. In [36] the end-to-end convolutional neural network is proposed for video compression.

The above disused methods have limitations in terms of computational complexity and data visualization. In this paper, the downsample AVS bitstream is classified based on the visual characteristics, motion vector (MV) features, and transform coefficients into many sub-blocks [37]. Blocks of most interest (BOMI) have large motions in the video, the blocks of less interest (BOLI) have slight motion, and the blocks of non-interest (BONI) are almost stable in the video. After classifying these blocks, we apply a super-resolution deep learning-based approach on BOLI and BOMI blocks, which enhances the visual quality of low resolution blocks. The BONI blocks have no interest in the video, so they are considered as background blocks. The enhancement of region of interest area in video, saves the transcoding time as well as super resolution complexity as compared to full frame based enhancement. The main contributions of our work are summarized as:

- We proposed a block detection and classification technique that uses the visual characteristics, transform coefficients, as well as motion vector (MV) features of a video. This technique classifies all blocks into three types: block of most interest, block of less interest, and block of non-interest.
- A deep learning based enhancement technique is used to convert low resolution to high resolution video. This technique enhances the region of interest area in a video instead of full frame and saves the computational time of the system.
- The multi scale super resolution technique is progressive in both design and training during the upsampling of frames. To make training process smoothly, we used curriculum learning in the proposed framework.
- We have carried out detailed experiments on six different benchmarks to evaluate the efficiency of our proposed method.

The rest of the paper is organized as follows. Section III shows the classification of blocks. The super resolution framework on different blocks is presented in section IV. In section V, the Simulation results are presented. Finally, the conclusion is discussed in section VI.

### III. Deep Learning based Super Resolution Framework

#### A. Classification of Blocks

In fact, humans are more concerned about the area of motion in video content. The sensitivity with respect to the eyes of the visual system is significant for the classification of blocks in the video frames. Mostly richer texture blocks or moving blocks get more attention of the viewers in the video. These blocks are known as the block of most interest (BOMI). The blocks, which are very near to the block of most interest and have low motion in video, are called blocks of less interest (BOLI). Frequently, background objects have no motion and are less appealing to the viewers, known as blocks of non-interest (BONI). The super-resolution algorithm can well preserve the video quality of BOMI and BOLI. The regions for newly incorporated objects and camera motion or fast-moving objects can be determined by motion vectors (MVs) and transforming coefficients. The regions of interest (ROI) can be detected in the video frame using coding parameters.

In AVS, the macroblocks (MBs) are used to encode the video frames, where each macroblock has a fixed dimension of $16 \times 16$. We will evaluate coding information of macroblocks for the classification of each coding tree unit (CTU). Although each macro block of the frame is used to estimate either frame is intra or inter. In the blocks detection, each video frame is divided into three block types of as shown in Fig. 1. The number of MBs have encoded with the intra mode in a frame is used to decide the intra or inter mode further. Let $h_{a_i}$ is used to indicate the $i^{th}$ MB in one frame.

$$a_i = \begin{cases} 0, & \text{intra mode} \\ 1, & \text{inter mode} \end{cases} \quad (1)$$

$$\sum_{i=1}^{16} a_i \geq T, \quad (2)$$

Eq. (2) is used as a criterion for classification of blocks. It also classifies a CTU as an intra/inter coded, where $T$ is used as a threshold, and its optimized value is 8. The motion vectors of current frame are calculated as follows:

$$FMV_{avg}(i) = \frac{1}{N} \sum_{i=1}^{M} B \times MV_{MB_k}(k), \quad (3)$$

where $FMV_{avg}$ is the frame average motion vectors, $i$ is the current frame index and $MB_k$ presents the $k^{th}$ macro block...
Fig. 3: Blocks detection and classification: (a) represents the original video sequences of Cactus, Jockey and Bosphorus video, (b) shows the AVS encoding information results: grey regions show skip macro blocks, purple color region shows intera MBs and blue region shows inter MBs, and (c) shows the region of interest: where white color shows the blocks of the most interest, gray color shows the blocks of less interest and black color shows the blocks of non-interest (Best viewed in color).

of the current frame. $BMV_{MB_k}(k)$ represents the magnitude of motion vectors, that is defined as:

$$BMV_{MB_k}(k) = \sqrt{BMV_{a}^2(k) + BMV_{b}^2(k)}, \quad (4)$$

where $BMV_{a}^2(k)$ and $BMV_{b}^2(k)$ are the motion vector (MV) components. These motion vector components are used horizontally and vertically in a frame. The current macro blocks and average motion vectors are calculated as:

$$MBMV_{avg}(i) = \frac{1}{N} \sum_{k=1}^{NF} \sum_{l=1}^{M} B \times MV_{MB_k}(k), \quad (5)$$

where $NF$ shows the macro blocks total number and $m$ shows the total number of motion vectors in the frame. The background macro blocks are determined by transform coefficient of each frame value. Therefore, the average value of transform coefficient and current macro block can be estimated by these equations:

$$FTC_{avg}(i) = \frac{1}{N} \sum_{k=1}^{NF} \sum_{l=1}^{M} B \times lkTC_{MB_k}(k). \quad (6)$$

$$MBTC_{avg}(i) = \frac{1}{N} \sum_{k=1}^{NF} \sum_{l=1}^{M} B \times lkTC_{MB_k}(k). \quad (7)$$

where $FTC_{avg}$ shows the average motion vectors, $i$ is the current frame index and $MB_k$ presents the $k^{th}$ macro block in the frame. $BTC_{MB_k}(k)$ shows transform coefficients of $k^{th}$ macro block in the current frame. According to the given parameters, we can easily detect the required blocks, which are shown in Fig. 1.

1) Blocks of most interest (BOMI)

$$\sum_{i=1}^{16} a_i \geq T \parallel (FMV_{avg}(i) > MBMV_{avg}(i)) \quad (8)$$

$$\delta \delta FTC_{avg}(i) \leq MBTC_{avg}(i)$$

2) Blocks of less interest (BOLI)


\[
\sum_{i=1}^{16} a_i < T \left( FMV_{\text{avg}}(i) > MBMV_{\text{avg}}(i) \right) \quad (9)
\]

\[
\delta FTC_{\text{avg}}(i) \leq MBTC_{\text{avg}}(i)
\]

Or

\[
\sum_{i=1}^{16} a_i < T\delta \left( FMV_{\text{avg}}(i) \leq MBMV_{\text{avg}}(i) \right) \quad (10)
\]

\[
\delta FTC_{\text{avg}}(i) > MBTC_{\text{avg}}(i)
\]

3) **Blocks of non-interest (BONI)**

\[
\sum_{i=1}^{16} a_i < T\delta \left( FMV_{\text{avg}}(i) \leq MBMV_{\text{avg}}(i) \right) \quad (11)
\]

\[
\delta FTC_{\text{avg}}(i) \leq MBTC_{\text{avg}}(i)
\]

These conditions are very useful to separate the region of interest block. The detection and classification of three video sequences, *Cactus*, *Jockey*, and *Bosphorus* is shown in Fig. 3. The resolution of each sequence is 1920 × 1080. Fig. 3(a) shows the AVS encoding information, which is used to classify the blocks into three gray, purple, and black colors. In Fig. 3(c), the white color blocks are region of most interest, gray color blocks are considered as less region of interest, and black colored blocks are mostly stable and have less interest to the viewers, which are known as blocks of non-interest.

**IV. SUPER RESOLUTION OF DIFFERENT BLOCKS**

We implemented multi scale super-resolution method to enhance the visual quality of low-resolution both MBOI and LBOI blocks of video frame. According to the visual features and motion vectors, the down sample frames of a video are divided into three blocks: MBOI, LBOI and BONI, as described in Section III. In Fig. 3, AVS decoded frame is divided into three types of blocks and then we applied multi-scale SR method to obtain the high quality blocks. The resolution of BONI blocks remains unchanged due to less interest as compared to MOBI and LBOI. The proposed technique mainly focuses on foreground area of the frame which obtained from blocks detection technique as shown in Fig. 3.

**A. Multi-Scale Super Resolution**

We presented a GAN-SR approach that is more progressive in the design as well as training. In intermediate phases, the network upsamples an image, while the curriculum learning is used for the learning process to order the process from simple to difficult. We designed a generative adversarial network (GAN) called Gan-SR that uses the same progressive multi-scale design idea to achieve more credible outcomes. This not only able to measure well against large sampling factors, but also a multi-scale technique that simultaneously raises the quality of the reconstruction of each upsampling factor. This method can work on all image channels (chrominance and luminance, etc.). Here we use luminance channel of each block to increase the spatial resolution of the frame, but other channels remained unchanged.

Suppose we have a set of N low resolution (LR) as an input frames with the corresponding target frames that are high resolution frames \( (x_1, y_1), \ldots, (x_n, y_n) \). The upscale function \( U : X \rightarrow Y \), with X low resolution and Y high resolution frames. It is difficult to find the parameters that are appropriate for the upsampling function \( U \) for high resolution ratios: the bigger the ratio, the more complicated the class of functions is needed. To this purpose, the upsampling function \( U \) is a progressive solution. We apply multi-scale super resolution pyramidal network on low resolution frames that are discussed below.

1) **Pyramidal Division**: A pyramidal division model is used which divides a upsampling factor \( U \) into a number of simpler functions \( (U_0, \ldots, U_n) \). The task of each function is to perform the upsampling and the refinement of the features of its own input. The cascade model consist of the compact compression units (CCUs), and a sub-pixel convolution layer, is used at each pyramid level. The asymmetrical structure is achieved by assigning of additional CCUs in the lower pyramid level as shown in Fig. 5. The high computational power in lower pyramid decreases the memory consumption and rises the receptive field to the original frame, which exceeds the symmetrical variation in terms of quality and execution. The upscale function \( U \) is divided across the pyramid levels. We additionally utilize two sub scaling-network which are represented by \( v_s \) and \( r_s \). These networks translate the individual transformation between the scale and standardized feature space of a frame. Figure 5 shows a schematic depiction of our progressive upsample network. This network uses a fixed upsampling of input \( (x) \), e.g. the bicubic interpolation to make the learning process easy to get a residual as an output.

\[
R_s(x) = (r_s \circ u_s \circ \ldots u_0 \circ v_s)(x) \quad (12)
\]

The estimated HR frame with a certain scaling factor \( s \) can be calculated as:

\[
\bar{Z} = R_s(x) + \phi_s(x) \quad (13)
\]

In particular, the Laplacian pyramid [38] concept is not followed in this network. In the Laplacian pyramid based network, intermediate sub-network outputs are neither supervised nor monitored as a basic frame. This network perform well over the Laplacian alternative, which simplifies the backpass and minimizes the optimization challenge. Moreover, the ground truth does not downsample for interim supervision. This network is very helpful to prevent artifacts which may occur from sub-sampling.

2) **Compact Compression Units (CCUs)**: A previously proposed DenseNet architecture [39] is used to construct each pyramid level. Similarly to skip connections, we use the compact connections that increase the flow of gradients and reduce disappearance in gradients [40]. A compact compression unit (CCU) is the central component of each level of the pyramid and includes a modified compact-connected block by 1 × 1 convolution \( CONV(1,1) \). The BN – ReLU – \( CONV(1,1) \)-BN – ReLU – \( CONV(3,3) \) is used in the initial compact layer. We eliminate all batch normalizations in super-resolution due to recent practice in SR methods [11], [33].
Fig. 4: The deep learning based super resolution method architecture where compact compression units (CCUs) are allocated in the lower pyramid level to reduce memory consumption and improve the reconstruction accuracy of high quality blocks in video frame.

Fig. 5: Flow chart of the proposed video transcoder with deep learning based super resolution.

However, as the characteristics of prior layers may differ, the first ReLU to rescale CONV(1, 1) characteristics can also be removed. This result leads to composition of a dense layer: \( CONV(1, 1) - RELU CONV(3, 3) \).

In Compact Net, the compact connection is broken at the end of each CCU with a compression layer \( CONV(1, 1) \) that quickly reassembles the information compact connections and increase the efficiency gain in results. A pyramid and local residual connections are used to improve the gradient propagation in this deep model, as illustrated in Fig. 5.

3) Progressive GAN: Generative adversarial networks [41] developed as an effective way to improve the perceptiveness of upscaled frames in SR. Training GANs is notoriously hard and the success of using GANs to SR is sparse to a single sample in very small targets. The proposed transcoder network is based on multi-scale SR network which is similar to the generator adversarial network suggested in the second part of Figure 5. The reverse pyramid structure \( u_2, u_1, u_0 \) are indicated in the second part of Fig. 6, where each level diminishes the input frame spatially using \( AvgPOOLING \) progressively.

A scale-specific frame transformation layers \( v_s \) is used before each pyramid that is similar to the generator. The network is entirely converted and produces a tiny features patch that is similar to PatchGAN [42] to handle multifaceted outputs from the generator. The discriminator works on residual of bicubic up sampled frame and the original frame is similar to the generator network. This permits both the generator and the discriminator to solely focus on the major variation source which is a upsampling procedure. This may also be seen as the subtraction from the discriminator of a database that contributes to a reduction of variance. We employ the less square loss rather than the original cross-entropy loss as our training target. The discriminator and the generator loss for the training of scales \( (s) \) may be described by means of the anticipated real residual and residual, which represented as \( r \) and \( \tilde{r} \):

\[
\mathcal{L}_{D_s} = (D(\tilde{r}_s))^2 + (D(r_s) - 1)^2,
\]

(14)
\[ \mathcal{L}_{R_s}^i = (D(R_s(x_s^i) - 1)^2 + \sum_{k \in \{2, 4\}} \| p_k(Z_t) - p_k(Z_i) \|^2, \]  

where \( p_k \) shows the \( k^{th} \) pooling layer of VGG-16 model [12].

4) Curriculum Learning: Curriculum learning [43] is a training approach which increases the complexity of the learning task progressively. It is widely used to forecast sequences and make decisions sequentially, where huge speeds may be achieved in the training period and improve performance. The pyramidal division of \( U \) makes it possible for us to naturally implement the curriculum learning. The loss for a scalable exemplar \( (x_s^i, z_i) \) can be set to:

\[ \mathcal{L}_{R_s}^i = \| R_s(x_s^i) + \theta_s(x_s^i) - z_i \|_1, \]  

where \( x_s^i \) coresponds to \( s \times \) downsampled version of \( z_i \) and the target at scale \( s \) is calculated as:

\[ \theta_s = \text{argmin}_{\theta_s} \sum_{s' \leq s} \sum_{i} \mathcal{L}_{R_{s'}}^i, \]  

where the value \( \theta_s \) parameterizes all functions in and below the current scale \( \{u_0, v_0, r_0, \ldots, u_s, v_s, r_s\} \) according to the pyramid network that is described in Fig. 5.

Our method merely starts with a \( 2 \times \) network instruction. As we enter a new phase (e.g. to \( 4 \times \)), the new pyramid level is progressively included to decrease its influence on the previously learned levels as shown in Fig. 6. The predicted \( \hat{z}_s \) of the generator at scale \( s \) is a linear combination of \( s \) and \( s-1 \) level outputs. For the discriminator, the new pyramid’s output features are combined with the scale-specific input of the preceding level vscale1 before entering the trained pyramids \( \{u_{s-1}, \ldots, u_0\} \).

The bilinear interpolation and AvgPOOL are employed before merging to match spatial dimensions. In both situations, \( \beta \) determines the effect of the new pyramid, which changes from 0 to 1 during the mixing process. As a consequence, we add the training pairs of next scale that scale gradually. Finally, in order to construct the batches we have chosen one of the scales \( s \) randomly to prevent the batch statistics from being mixed as recommended in [2].

This progressive training strategy considerably reduces overall training time in comparison to basic multi-scale training when training examples from multiple scales are provided to the network at the same time. In comparison to single scale and basic multi-scale training, it offers a further efficiency improvement on all included scales and relieves instabilities in GAN training.

V. EXPERIMENTS

The proposed SR technique is implemented by using Python and MATLAB 2018a. All experiments are performed at Intel (R) Core™ i5-CPU 4590 @ 3.3GHZ with 8 GB RAM with a spatial resolution of 3840 \times 2160 for evaluating the performance. The frames are encoded with the low delay-P-main configuration file. The performance of proposed method is evaluated based on different metrics such as the image quality (\( \Delta \) PSNR), the bit rate (\( \Delta \) Bit-rate), and the computational complexity (\( \Delta \) T). The parameters are calculated using the equations given below:

\[ \Delta \text{PSNR} = \text{PSNR}_{(\text{proposed})} - \text{PSNR}_{(\text{anchor})} \]  

\[ \Delta \text{Bitrate} = \frac{\text{Bitrate}_{(\text{proposed})} - \text{Bitrate}_{(\text{anchor})}}{\text{Bitrate}_{(\text{anchor})}} \times 100\% \]  

\[ \Delta T = \frac{T_{(\text{proposed})} - T_{(\text{anchor})}}{T_{(\text{anchor})}} \times 100\% \]

where \( \text{PSNR}_{(\text{proposed})}, \text{Bitrate}_{(\text{proposed})}, \text{and Bitrate}_{(\text{proposed})} \) are the proposed algorithm PSNR, encoding time and bitrate, respectively and \( \text{PSNR}_{(\text{anchor})}, \text{T}_{(\text{anchor})} \) and \( \text{Bitrate}_{(\text{anchor})} \) and represent the PSNR, encoding time, and bitrate of anchor software HM10.1, respectively.

In order to evaluate the performance of our proposed method, we used six different classes of videos. Each video belongs to a particular class. For example, Class A has PeopleOnStreet and Traffic video, Class B has Cactus and Kimono video, Class-C has BasketballDrill and PartyScene video, Class-D has BlowingBubbles and BQSquare video, Class-E has FourPeople and KristenAndSara video, and Class-F has ChinaSpeed and SlideEditing video. The experimental results of block-based super-resolution method and frame based method are shown in Table I and Table II. The difference between the frame based and block based results is shown in Table III. Honeybee up-sampling results show that our proposed method saves 37% time compared to the frame based method with similar picture quality PSNR (dB) and SSIM (index). In block based, Bosphorus saves 26% of the time during 1K to 4K upsampling. For 1K to 2K upsampling, Kimono takes 0.006% more SR time. The results of our experiments show that the proposed method has achieved good results in 2K and 4K video with low computational time. The negative sign in the Table-III indicates a decrease, and the positive sign indicates an increase in value. In Table III, the Honeybee PSNR drops by 0.0091 dB compared to the frame based value and the Cactus PSNR decreases by 0.0022 dB over the frame based result. As to compare bitrate results, BQTerrace obtains a maximum bit rate of 1.77% over the frame based value, and the Bosphorus achieves a minimum bit rate reduction of 0.745% based on the frame based value. Fig. 7 shows the HEVC encoding time analysis for six different videos with different resolutions. In Fig. 7, it can be seen that Bosphorus video consumes more HEVC encoding time in 2K and 1K resolutions.

Fig. 7(a) shows the HEVC encoding time of six different videos. The experiments show that the all block based videos have high HEVC encoding time. On average, the block based results take 15.98% more HEVC encoding time as compared to frame based. The average difference between block based and frames based in PeopleOnStreet video frame is -45.80 seconds, which indicates that frame based transcoder has a
TABLE I: Block Based Result of Proposed Method

| Video Sequence | Initial Resolution | Final Resolution | PSNR (dB) | SSIM (index) | SR Time (sec) | Bitrate (kb) | HEVC Encoding Time |
|---------------|--------------------|------------------|-----------|--------------|--------------|--------------|-------------------|
| HoneyBee      | 540 x 1080         | 1080 x 1080      | 37.8406   | 0.8801       | 1228.84      | 3831.21      | 8049.17           |
| Bosphorus     | 540 x 1080         | 1080 x 1080      | 37.8653   | 0.9469       | 1313.90      | 3021.42      | 8802.40           |
| Jockey        | 540 x 1080         | 1080 x 1080      | 37.8653   | 0.9469       | 1313.90      | 3021.42      | 8802.40           |
| BQTerrace     | 540 x 1080         | 1080 x 1080      | 37.8653   | 0.9469       | 1313.90      | 3021.42      | 8802.40           |
| Cactus        | 540 x 1080         | 1080 x 1080      | 37.8653   | 0.9469       | 1313.90      | 3021.42      | 8802.40           |
| Kimono        | 540 x 1080         | 1080 x 1080      | 37.8653   | 0.9469       | 1313.90      | 3021.42      | 8802.40           |

TABLE II: Frame Based Result of Proposed Method

| Video Sequence | Initial Resolution | Final Resolution | PSNR (dB) | SSIM (index) | SR Time (sec) | Bitrate (kb) | HEVC Encoding Time |
|---------------|--------------------|------------------|-----------|--------------|--------------|--------------|-------------------|
| HoneyBee      | 540 x 1080         | 1080 x 1080      | 37.8406   | 0.8801       | 1228.84      | 3831.21      | 8049.17           |
| Bosphorus     | 540 x 1080         | 1080 x 1080      | 37.8653   | 0.9469       | 1313.90      | 3021.42      | 8802.40           |
| Jockey        | 540 x 1080         | 1080 x 1080      | 37.8653   | 0.9469       | 1313.90      | 3021.42      | 8802.40           |
| BQTerrace     | 540 x 1080         | 1080 x 1080      | 37.8653   | 0.9469       | 1313.90      | 3021.42      | 8802.40           |
| Cactus        | 540 x 1080         | 1080 x 1080      | 37.8653   | 0.9469       | 1313.90      | 3021.42      | 8802.40           |
| Kimono        | 540 x 1080         | 1080 x 1080      | 37.8653   | 0.9469       | 1313.90      | 3021.42      | 8802.40           |

TABLE III: Difference between Block Based and Frame Based Result

| Video Sequence | Initial Resolution | Final Resolution | PSNR (dB) | SSIM (index) | SR Time (sec) | Bitrate (kb) | HEVC Encoding Time |
|---------------|--------------------|------------------|-----------|--------------|--------------|--------------|------------------|
| HoneyBee      | 540 x 1080         | 1080 x 1080      | -0.0091  | 0.8801       | 1228.84      | 3831.21      | 8049.17           |
| Bosphorus     | 540 x 1080         | 1080 x 1080      | -0.0079  | 0.8801       | 1228.84      | 3831.21      | 8049.17           |
| Jockey        | 540 x 1080         | 1080 x 1080      | -0.0087  | 0.8801       | 1228.84      | 3831.21      | 8049.17           |
| BQTerrace     | 540 x 1080         | 1080 x 1080      | -0.0091  | 0.8801       | 1228.84      | 3831.21      | 8049.17           |
| Cactus        | 540 x 1080         | 1080 x 1080      | -0.0091  | 0.8801       | 1228.84      | 3831.21      | 8049.17           |
| Kimono        | 540 x 1080         | 1080 x 1080      | -0.0091  | 0.8801       | 1228.84      | 3831.21      | 8049.17           |

TABLE IV: Image quality (PSNR) and Bitrate (kb) comparison among proposed method and other methods

| Video Sequence | TMM PSNR(14) | TMM Bitrate (14) | CSVT PSNR(45) | CSVT Bitrate (45) | Proposed PSNR(45) | Proposed Bitrate (45) |
|---------------|--------------|-----------------|---------------|------------------|------------------|----------------------|
| HoneyBee      | ——           | ——              | ——            | ——               | ——               | ——                   |
| Bosphorus     | ——           | ——              | ——            | ——               | ——               | ——                   |
| Jockey        | -0.1180      | 2.646           | -0.0115       | 2.4980           | -0.0030          | -0.4829              |
| BQTerrace     | -0.0790      | 1.5180          | -0.1450       | 2.8940           | -0.0030          | -0.4829              |
| Cactus        | -0.1250      | 3.0780          | -0.0660       | 1.6430           | -0.0022          | -1.2870              |
| Kimono        | -0.024       | 0.5520          | -0.0570       | 1.3080           | -0.0040          | 0.0854               |
| BQTerrace     | -0.1260      | 3.4610          | -0.1130       | 3.1000           | -0.0264          | 1.7741               |

high computational cost. High definition videos are more time consuming, due to high pixel density. Fig. 7(b) shows the super-resolution complexity between frame-based and block-based SR results. The SR time varies from video to video due to different resolution. Different types of up-sampling is also the part of experiments. For example, 1K low resolution (1280 x 800) images of PeopleOnStreet, and Traffic are up-sampled to 2K resolution (2560 x 1600) by using proposed SR method same as the low resolution (960 x 544) images of Cactus, and Kimono are up-sampled to 1K high resolution (1920 x 1080) frames using same SR technique.

Fig. 7(c) illustrates the visual quality of frame-based and block-based SR approaches. Here, we will contemplate the results of six different videos, which are used in the experiment. In most cases, the maximum signal to noise ratio (PSNR) of both frame based and block based videos is similar. The visual quality difference raises gradually in low resolution videos. In 4K videos, the visual quality of both block based and frame based videos is mostly comparable. During up-sampling 1K to 2K or 2K to 4K the number of pixels density decreases, which leads to decrease in the visual quality of all video frames. Fig. 7(d) shows a comparison of bitrates between frame-based and block-based transcoding methods. The bitrate is the size of video file per second of data. It is usually represented by kilobits (kbs) or megabits per second (mbs). The bitrate in video streaming quality is a vital factor.
Fig. 7: (a) shows the ROI based and full frame-based HEVC-encoding Time (sec), (b) shows the ROI based and full frame-based SR Time (sec), (c) shows the ROI based and full frame-based PSNR (dB), and (d) shows the ROI based and full frame-based Bitrate (kb).

Streaming media codec uses “lossy” compression to compress the video file. The high compression rate on video affects the video quality in transcoder. Overall, the results of proposed method are efficient and reliable in terms of SR time, bitrate as well as HEVC encoding time and PSNR. Fig. 8 shows the visual efficiency of ROI based super resolution and full frame based super resolution.

VI. CONCLUSION

This paper introduces a novel approach, where a deep learning based enhancement technique is used to convert low resolution to high resolution video. This technique enhances the region of interest area in a video instead of full frame and reduces the 20% to 30% SR complexity of a system. The results validate the efficiency of this approach. The proposed block based transcoder performs better than frame based super-resolution method with comparable visual quality in SSIM (index) and PSNR (dB).

In future, we can extend this transcoder for other video formats like H.264 to MPEG transcoder, H.264 to MPEG-4 transcorder, MPEG-4 to MPEG-2 transcorder, etc. This field is growing rapidly and continuous effort will bring significant improvement in video compression and quality.

REFERENCES

[1] C.-H. Yeh, W.-Y. Tseng, L.-W. Kang, C.-W. Lee, K. Mucthar, and M.-J. Chen, “Coding unit complexity-based predictions of coding unit depth and prediction unit mode for efficient hevc-to-shvc transcoding with quality scalability,” Journal of Visual Communication and Image Representation, vol. 55, pp. 342–351, 2018.

[2] I. Ahmad, X. Wei, Y. Sun, and Y.-Q. Zhang, “Video transcoding: an overview of various techniques and research issues,” IEEE Transactions on multimedia, vol. 7, no. 5, pp. 793–804, 2005.

[3] A. Vetro, C. Christopoulos, and H. Sun, “Video transcoding architectures and techniques: an overview,” IEEE Signal processing magazine, vol. 20, no. 2, pp. 18–29, 2003.

[4] J. Zhang, C. Jia, M. Lei, S. Wang, S. Ma, and W. Gao, “Recent development of avs video coding standard: Avs3,” in 2019 Picture Coding Symposium (PCS), pp. 1–5, IEEE, 2019.

[5] X. Zhao, S. Liu, L. Zhao, X. Xu, B. Zhu, and X. Li, “A comparative study of hevc, vvc, vp9, av1 and avs3 video codecs,” in Applications of Digital Image Processing XLIII, vol. 11510, p. 1151011, International Society for Optics and Photonics, 2020.

[6] G. Abbas, M. J. Khan, R. Qureshi, and K. Khurshid, “Scope of video magnification in human pulse rate estimation,” in 2017 International Conference on Machine Vision and Information Technology (CMVIT), pp. 69–75, IEEE, 2017.

[7] A. R. Shahid, S. Khan, and H. Yan, “Contour and region harmonic features for sub-local facial expression recognition,” Journal of Visual Communication and Image Representation, vol. 73, p. 102949, 2020.

[8] H. Hassan, M. N. Khan, S. O. Gilani, M. Jamil, H. Maqbool, A. W. Malik, and I. Ahmad, “H. 264 encoder parameter optimization for encoded wireless multimedia transmissions,” IEEE Access, vol. 6, pp. 22046–22053, 2018.

[9] F. Bossen, B. Bross, K. Suhring, and D. Flynn, “Hevc complexity and implementation analysis,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1685–1696, 2012.

[10] S. J. Wee, J. G. Apostolopoulos, and N. Feamster, “Field-to-frame transcoding with spatial and temporal downsampling,” in Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348), vol. 4, pp. 271–275, IEEE, 1999.

[11] M. Pantoja and N. Ling, “Transcoding with resolution conversion using super-resolution and irregular sampling,” Journal of Signal Processing Systems, vol. 60, no. 3, pp. 305–313, 2010.

[12] D. Zhang, B. Li, J. Xu, and H. Li, “Fast transcoding from h. 264 avc to high efficiency video coding,” in 2012 IEEE International Conference on Multimedia and Expo, pp. 651–656, IEEE, 2012.

[13] E. Peixoto, T. Shanableh, and E. Izquierdo, “H. 264/avc to hevc video transcoder based on dynamic thresholding and content modeling,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 24, no. 1, pp. 99–112, 2013.

[14] C.-S. Lin, W.-j. Yang, and C.-W. Su, “Fits: Fast intra transcoding from h. 264/avc to high efficiency video coding based on dct coefficients
and prediction modes,” *Journal of Visual Communication and Image Representation*, vol. 38, pp. 130–140, 2016.

[15] P. Assuncao and M. Ghanbari, “Transcoding of single-layer mpeg video into lower rates,” *IEEE Proceedings-Vision, Image and Signal Processing*, vol. 144, no. 6, pp. 377–383, 1997.

[16] L. Zhu, Y. Zhang, N. Li, G. Jiang, and S. Kwong, “Machine learning based fast h. 264/avc to hevc transcoding exploiting block partition similarity,” *Journal of Visual Communication and Image Representation*, vol. 38, pp. 824–837, 2016.

[17] L. Pengyu and J. Kebin, “Fast extraction method for video motion region of interest combining visual characteristics,” in *2011 International Conference on Computer Science and Service System (CSSS)*, pp. 3939–3944, IEEE, 2011.

[18] S.-F. Huang, M.-I. Chen, and M.-S. Li, “Region-of-interest segmentation based on bayesian theorem for h. 264 video transcoding,” in *2011 Visual Communications and Image Processing (VCIP)*, pp. 1–4, IEEE, 2011.

[19] P. Xing, Y. Tian, X. Zhang, Y. Wang, and T. Huang, “A coding unit classification based avc-to-hevc transcoding with background modeling for surveillance videos,” in *2013 Visual Communications and Image Processing (VCIP)*, pp. 1–6, IEEE, 2013.

[20] N. Bjork and C. Christopoulos, “Transcoder architectures for video coding,” *IEEE Transactions on Consumer Electronics*, vol. 44, no. 1, pp. 88–98, 1998.

[21] S. Wang, F. Luo, and S. Ma, “Overview of the second generation avs video coding standard (avs2),” *ZTE Communications*, vol. 14, no. 1, pp. 3–11, 2019.

[22] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, “Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 151–160, 2019.

[23] C. Li, S. Anwar, and F. Porikli, “Underwater scene prior inspired deep underwater image and video enhancement,” *Pattern Recognition*, vol. 98, p. 107038, 2020.
Fig. 8: shows the qualitative results of 1K, 2K, and 4K videos. (a) shows the results of 4K Bosphorus, Jockey, HoneyBee and ShakeNDry video. There is no any significant difference between picture quality of both frame based and block based method, (b) shows the results of class A videos. In this class, the visual quality of PeopleOnStreet and Traffic videos is similar, (c) illustrates the results of class B videos (BasketballDrive, BQTerrace, Cactus, and Kimono), (d) shows the visual quality of PartyScene and RaceHorses videos, which belong to class C, (e) shows the visual quality of BlowingBubbles and BQSquare video of class D, (f) shows the visual quality of FourPeople and KristenAndSara videos, which belong to class E, and (g) shows the results of ChinaSpeed and SlideEditing video of class F.

[24] K. Bito, I. Siio, Y. Ishiguro, and K. Takeda, “Automatic generation of road trip summary video for reminiscence and entertainment using dashcam video,” in 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 181–190, 2021.

[25] J. E. Jacob and S. Saritha, “Video enhancement and low-resolution facial image reconstruction for crime investigation,” in Intelligent Data Communication Technologies and Internet of Things, pp. 773–788, Springer, 2021.

[26] F. Lv, F. Lu, J. Wu, and C. Lim, “Mbllen: Low-light image/video enhancement using cnns,” in BMVC, p. 220, 2018.

[27] X. Huang, J. Søgaard, and S. Forchhammer, “No-reference pixel based video quality assessment for hevc decoded video,” Journal of Visual Communication and Image Representation, vol. 43, pp. 173–184, 2017.

[28] J. De Praeter, G. Van Wallendael, T. Vermeir, J. Slowack, and P. Lambert, “Spatially misaligned hevc transcoding with computational-complexity scalability,” Journal of Visual Communication and Image Representation, vol. 40, pp. 149–158, 2016.

[29] R. Xie, J. Liu, and X. Wang, “Efficient mpeg-2 to mpeg-4 compressed video transcoding,” in Visual Communications and Image Processing 2002, vol. 4671, pp. 192–201, International Society for Optics and Photonics, 2002.

[30] O. Stankiewicz, K. Wegner, and M. Domański, “Study of 3d video compression using nonlinear depth representation,” IEEE Access, vol. 7, pp. 31110–31122, 2019.

[31] C. Li, C. Guo, L.-H. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, “Low-light image and video enhancement using deep learning: A
survey.” IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 01, pp. 1–1, 2021.

[32] G. Keesman, R. Hellinghuizen, F. Hoeksema, and G. Heideman, “Transcoding of mpeg bitstreams,” Signal processing: Image communication, vol. 8, no. 6, pp. 481–500, 1996.

[33] A. Eleftheriadis and D. Anastassiou, Dynamic rate shaping of compressed digital video. PhD thesis, Columbia University New York, NY, 1995.

[34] W. Jiang, Y. Chen, and X. Tian, “Fast transcoding from h.264 to hevc based on region feature analysis,” Multimedia tools and applications, vol. 73, no. 3, pp. 2179–2200, 2014.

[35] D. Liu, P. An, R. Ma, W. Zhan, and L. Ai, “Scalable omnidirectional video coding for real-time virtual reality applications,” IEEE Access, vol. 6, pp. 56323–56332, 2018.

[36] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, “Dvc: An end-to-end deep video compression framework,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11006–11015, 2019.

[37] R. Luo, R. Xie, and L. Zhang, “Fast avs to hevc transcoding based on roi detection using visual characteristics,” in 2014 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, pp. 1–6, IEEE, 2014.

[38] P. J. Burt and E. H. Adelson, “The laplacian pyramid as a compact image code,” in Readings in computer vision, pp. 671–679, Elsevier, 1987.

[39] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and P. J. Burt, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440, 2016.

[40] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in Proceedings of the 26th annual international conference on Machine learning, pp. 41–48, 2009.

[41] M.A.A Abdelgawad received his B.Sc. degree in Electrical Engineering from Minia University, Minia, Egypt in 2017. He is currently pursuing his Ph.D. degree in Electrical Engineering at City University of Hong Kong, Hong Kong SAR. His research interests include signal and image processing, Machine Learning, and tensor computing.

[42] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[43] Rizwan Qureshi is a professor of Computer Science, at School of Computing, National University of Computer and Emerging Sciences, Karachi Campus, Pakistan. He received his PhD from the department of Electrical Engineering, City University of Hong Kong, Hong Kong in 2021. Before joining City University, he was a lecturer in electrical engineering department at Mehran University of Engineering Technology, Jamshoro Pakistan. His research interests include signal and image processing, Machine Learning, Energy Management and Spectral Imaging.

[44] MANSOOR ALI TEEVNO received his BEng in electronic engineering and Master of Engineering in Electronic System Engineering from Mehran University of Engineering Technology, Jamshoro Pakistan in 2012 and 2017 respectively. Currently, he is working towards his PhD degree in electrical engineering at City University of Hong Kong. Before joining City University, he was working as lecturer in electrical engineering department at Mehran University of Engineering Technology, Jamshoro Pakistan. His research interests include signal and image processing, machine learning and Embedded Systems.

[45] Muhammad Waqas is currently working towards his PhD degree in Computer Science at National University of Computer and Emerging Sciences, Karachi, Pakistan. He is also a lecturer at the department of Computer Sciences, University College of Zhob, Baluchistan University of Information Technology, Engineering, and Management Sciences, Quetta, Pakistan. His research interests include multiple instance learning, image processing, and computer vision.
Ali Raza Shahid (Graduate student member, IEEE) received an MSc degree in electrical engineering from COMSATS University Islamabad, Pakistan, in 2015. He is currently pursuing the Ph.D. degree in Electrical Engineering at City University of Hong Kong. He also served as lecturer in COMSATS University Islamabad, Pakistan. His research interests include magnetic resonance imaging and computer vision.