Indonesian Journal of Electrical Engineering and Computer Science
Vol. 8, No. 1, October 2017, pp. 43 ~ 58
DOI: 10.11591/ijeecs.v8.i1.pp43-58

An Adaptive Scheme to Achieve Fine Grained Video Scaling

S. Safinaz*[1], A.V.Ravi Kumar[2]
[1]sir.mvit, VTU, Bangalore, India
[2]Sjbit, VTU, Bangalore, India
*Corresponding author, e-mail: safinaz.mvit@rediffmail.com

Abstract

A robust Adaptive Reconstruction Error Minimization Convolution Neural Network (ARemCNN) architecture introduced to provide high reconstruction quality from low resolution using parallel configuration. Our proposed model can easily train the bulky datasets such as YUV21 and Videoset4.Our experimental results shows that our model outperforms many existing techniques in terms of PSNR, SSIM and reconstruction quality. The experimental results shows that our average PSNR result is 39.81 considering upscale-2, 35.56 for upscale-3 and 33.77 for upscale-4 for Videoset4 dataset which is very high in contrast to other existing techniques. Similarly, the experimental results shows that our average PSNR result is 38.71 considering upscale-2, 34.58 for upscale-3 and 33.047 for upscale-4 for YUV21 dataset.

Keywords*: PSNR, ARemCNN, YUV21, Reconstruction Quality, CNN*

1. Introduction

High-end digital devices like 3D TV, UHD (Ultra High Definition) devices with 8K resolution (7680×4320), smartphones and iPads with 4k/2k resolution has gained immense popularity in recent years due to its extreme high-resolution quality and true color. These High Definition (HD) devices provide better visual appearance. Therefore, there is a need of ultra-high resolution videos, which can provide compatibility to high-end digital devices. However, there are abundant volume of low-resolution videos are exist in the practical world. Therefore, there is a one way to up sample the low-resolution videos to high-resolution videos using super-resolution technique.

Super-resolution is a modern technique, which provides one of the most encouraging services in the field of video applications (video capturing and displaying). Super-resolution technique helps to produce UHD content from the lower resolution videos. These UHD videos widely used in many applications such as medical [1], satellite imaging [2], and face recognition [3], stereoscopic video processing [4], video coding/decoding [5] and surveillance [6]. Therefore, scalability in super-resolution technique is one of the most vital approach, which helps to up sample the low-resolution videos to high resolution. Super resolution can remove the blurriness, motion, blue kernel and noise present in the low-resolution videos to get high quality [7]. Therefore; super resolution approach requires high accuracy and lightning speed to handle processing of low-resolution video frames and images. However, existing techniques are not capable of handling the processing of large video datasets due to the blurriness, ringing and blocking effects present in low-resolution videos. These led to poor resolution quality in existing systems. Therefore, they are not capable of providing high quality resolution.

In [8], a unified super resolution technique adopted to estimate depth of asymmetric stereoscopic videos. This method helps reconstruct the low-resolution videos by enhancing its poor quality. However, it consists of high computational complexity and implementation can be much faster. In [9], a super resolution technique presented for video quality enhancement using interpolation based virtual view. However, it consists of compression distortion which can degrade its efficiency. In [10], a fast super-resolution technique presented based on key frames to reconstruct low-resolution frames. However, this method is not robust enough to handle reconstruction errors and distortion. In [11], a novel 3D super resolution technique presented to

Received June 1, 2017; Revised August 3, 2017; Accepted August 20, 2017

reconstruct low resolution video frames for high density molecules. However, the image reconstruction quality using this method is not satisfactory and can be improve more.

Blurriness in the lower resolution frames, high computational complexity, reconstruction error and distortion, poor quality of low resolution videos, slower implementation, and redundancy in the pixels/frames, bandwidth utilization are the issues present in the existing video super resolution (scaling) techniques such as Lanczos, bilinear, and bi-spline. Therefore, to handle such type of issues there is a need of a robust technique which can precisely eliminate these drawbacks and can scale up low resolution frames to very high (HD/UHD) resolution frames. Convolution Neutral Network (CNN) has become one of the most vital technique to encounter these issues over a last decade due to its lightning speed, faster implementation, easy training and can handle large datasets with ease by GPU parallel computing [12]. However, still there are very few techniques which can provide high quality frames from low resolution videos using CNN.

Therefore, to further eliminate noise and blurriness from the low-resolution frames without compromising required high quality we have introduced a robust Adaptive Reconstruction Error Minimization Convolution Neural Network ($ARemCNN$) architecture which helps to eliminate the drawbacks of existing algorithms and provides high quality features by up scaling low-resolution quality. Video Up-Scaling consists of many properties like de-interlacing, scaling, quality video reconstruction, frame conversion due to its post processing pipelining structure. It also enhances the sharpness, contrast and color of low resolution videos. Therefore, our video up-scaling technique based on $ARemCNN$ architecture helps to achieve real-time HD/UHD video processing. Video scaling converts source resolution to higher/lower display resolution without compromising the quality of video. These attributes of video scaling makes it one of the most vital and convincing technique in the field of video processing [13].

Our proposed Adaptive Reconstruction Error Minimization Convolution Neural Network ($ARemCNN$) architecture creates a connection between feed-forward neural methods and adaptive filters which helps to enhance visual appearance. Our proposed $ARemCNN$ architecture generates powerful image transformations by optimizing convolutional neural networks. However, in existing algorithms the quality of reconstructed frames decreases whenever upscaling factor increased. Therefore, to eliminate this drawback here we apply parallel GPU computing using CAFFE framework. Here, we have presented an adaptive sparse coding reconstruction algorithm to eliminate the errors generated after every feature extraction from each frame which computes all the features simultaneously and apply simultaneous parallel sparse coding algorithm to each frame. Therefore, it can save large amount of computational time and complexity and makes our model faster than any other existing algorithm. Therefore, it enhances its video frame quality to a large extent. To define non-linearity we have utilized Adaptive Sparse Rectified Linear Unit ($ASReLU$).To prove a significant relationship between input low resolution and output high resolution frames an Adaptive Sparse Coding Based Architecture (SCA) considered. Our proposed Adaptive Reconstruction Error Minimization Convolution Neural Network ($ARemCNN$) architecture for high quality video scaling model outperforms the existing algorithms in terms of resolution quality, high quality image reconstruction, noise reduction and scaling issues which is verified by experimental outcomes of our proposed model.

This paper is organize in following sections, which are as follows. In section 2, we describe about the video scaling issues and how they can eliminate by our proposed model. In section 3, we described our proposed methodology. In section 4, experimental results, evaluation shown, and section 5 concludes our paper.

2. Video Scalling Issues

Over last decade, the distribution of HD/UHD displays has taken drastic growth in electronic market due to its high quality visibility. Therefore, a variety of high resolution video formats (2k/4k/8k) are generated over the years which becomes the mandatory requirements for high end digital devices. However, there are huge amount of low resolution videos available in real world. Therefore, there is a need of scaling to reconstruct the low resolution frames into high resolution frames to gain compatibility with high end devices. However, there are many issues present in the existing algorithms related to video scaling such as blurriness, ringing and blocking effects in low-resolution frames, high computational complexity, reconstruction error and distortion, poor quality of low resolution videos, slower implementation, and redundancy in the

pixels/frames, bandwidth utilization etc. Here, literature survey related to our proposed model discussed below.

In [14], a CNN (Convolutional Neural Network) technique presented for 3D Super Resolution to eliminate redundancy and improve quality of fusion video frames/images and to handle large datasets. In [15], deep CNN (convolution neural network) adopted to recognize facial expressions. However, in this model performance degrades for large databases. In [16], a 3D CNN (Convolution Neural Network) technique presented based on saliency features with LSTM (long-short term memory) to recognize video actions. In this model, importance of foreground presented by integrating C3D net, LSTM and time pooling algorithms. However, it increases computational complexity due to integration of multiple algorithm. In [17], a high performance up-scaling architecture presented to get high quality of 4K videos based on LaGrange interpolation and image sharpening. This model increases the visual quality to a high extent and sharpens the image boundary using sharpening filter. However, performance of reconstructed image can degrade whenever upscaling factor increased to a high extent. In [18], energy efficient technique adopted based on deep convolutional neural networks to high speed visual attention in mobile applications. However, it consumes more time due to its slow implementation. In [19], a real time video super resolution technique presented based on efficient sub-pixel Convolutional Neural Network architecture. In this model, a global ill-posed reconstruction error occurs which degrades the performance of the system. In [20], a video super resolution technique presented based on patch similarity and non-linear mapping. However, ill poised problem and noise in low resolution frames degrades its performance. In [21], a high quality super-resolution algorithm adopted to enhance the resolution quality based on parallel GPU computing. In the model to maintain the 4K video processing quality is the biggest challenge. In [22], a robust super resolution technique presented based on deep convolution neural network with sparse prior. However, this model consists of reconstruction and inverse problems.

In this paper, we have introduced a robust Adaptive Reconstruction Error Minimization Convolution Neural Network ($ARemCNN$) architecture which helps to eliminate the drawbacks of existing algorithms and our proposed $ARemCNN$ model compared to many recent conventional algorithms. Our $ARemCNN$ model consists of multiple phases like adaptive shrinking, adaptive mapping with adaptive sparse coding last layer which helps to get better visual quality by removing global ill-poised problem and video frame reconstruction drawbacks present in the existing algorithms. Our model is tested on Videoset4 and YUV21 dataset and experimental results demonstrated that it can reconstruct the low-resolution video frames into high resolution frames in a much faster way than the existing algorithms. Our $ARemCNN$ model, recognize highly complex classes of video features, provide appearance information and remove non-linear deformations.

3. Quality Video Scaling Using Adaptive CNN Architecture

Convolution Neural Network (CNN) is the one of the most vital strategy in the video processing to handle large datasets such as Myanmar [23], ImageNet [24], videoset4 [7], yuv21 [25] and can easily train with these large datasets. Therefore, CNN Architecture can help to achieve high-resolution quality from the low-resolution frames. Medical, satellite imaging, face recognition, stereoscopic video processing, video coding/decoding and surveillance [26-29] are the some fields where CNN can be widely used due to its fast computation and high quality training. However, still there are very few techniques which can provide high quality frames from low resolution videos using CNN. Therefore, to further eliminate noise and blurriness from the low-resolution frames without compromising required high quality, here, we have introduced a robust Adaptive Reconstruction Error Minimization Convolution Neural Network ($ARemCNN$) architecture which helps to eliminate the drawbacks of existing algorithms. There are some special features which helps $ARemCNN$ to form an architecture which provides fast computation, precise implementation, easier training and link between low-resolution and high resolution frames. These special features are connection between feed-forward neural methods and adaptive filters which helps to enhance visual appearance, use of parallel GPU computing on CAFFE and use of Adaptive Sparse Rectified Linear Unit ($ASReLU$) [30] design to make $ARemCNN$ robust and test for different designs.

3.1. Image Reconstruction Architecture

Due to availability of abundant high-end devices in last few years there is a need of high definition/ultra-high definition videos. However, a large section of videos present in the low resolution. Therefore, to convert low-resolution frames into high resolution frames, video scaling is very essential. A healthy amount of work has been done in the field of video processing and scaling. However, very few techniques emerges to work in real world environment due to their blurriness and blocking effects in low-resolution frames, high computational complexity, reconstruction error and distortion, poor quality of low resolution videos, slower implementation and redundancy in the pixel issues. Therefore to eliminate these issues and provide high definition resolution from lower resolution frames we have implemented a robust Adaptive Reconstruction Error Minimization Convolution Neural Network ($ARemCNN$) architecture along with sparse coding structure. Parallel GPU computing helps in to provide lightening speed to handle large training datasets which is based on CAFFE framework.

Our model $ARemCNN$ helps to achieve better visual quality and can reconstruct low resolution frame into high resolution frames using $ASReLU$ structure. Figure 1 describes the architectural diagram of sparse coding reconstruction algorithm which shows working of reconstruction of an image/video frame. Consider a low-resolution frame of a video. In our model $ARemCNN$ for each frame in a video patch based features extracted. Then, each video frame is simultaneously down-sampled to get intermediate frames and increase the visual similarity of low resolution frames. Then, every individual frame is up sampled to preferred size. To get high quality image, the variation between up sample frame and down-sample frame is fed to adaptive sparse coding image /video frame reconstruction simultaneously for each frame. In our proposed model $ARemCNN$, all the frames of a video parallel processed at the same time to get original high quality image and error occurred in each video frame is eliminated using sparse coding reconstruction. This model not just extract high visual features from low resolution frames but also saves large computational time due to parallel implementation of sparse coding architecture. Thus, it consists of low computational complexity and high computational speed to handle large datasets. In this way our model $ARemCNN$ outperforms all the existing algorithms in terms of computational complexity, speed and high visual quality. There are five phases to reconstruct a low resolution video into high definition video in our model $ARemCNN$ shown in Figure 1.

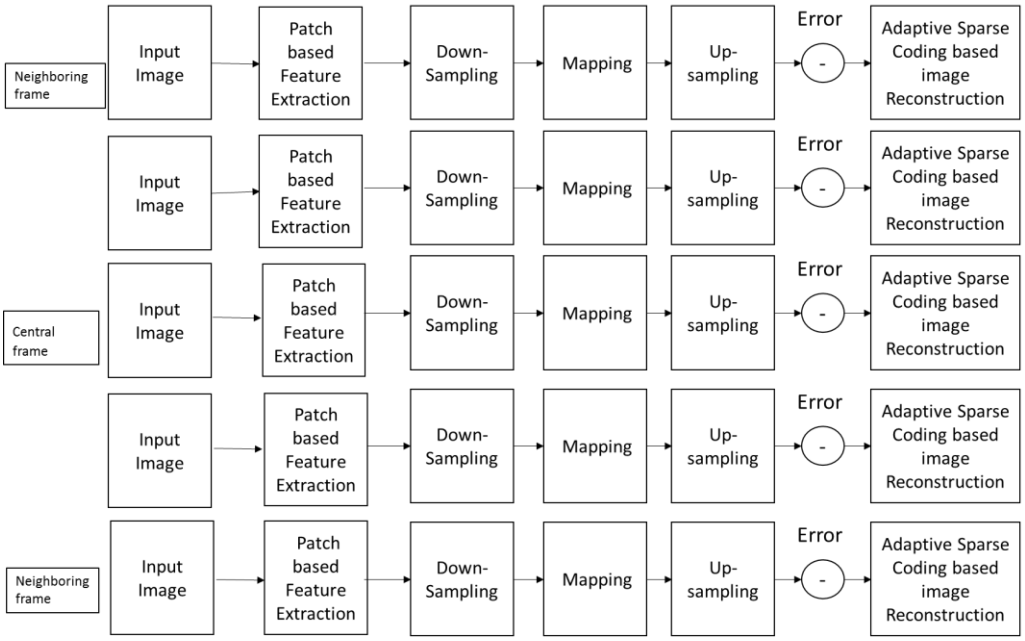

Figure 1. Adaptive sparse coding based image reconstruction architecture

There are five phases in $ARemCNN$ model to reconstruct the low resolution image which are as follows:

a. Adaptive Patch-Wise Feature Extraction

Patch wise feature extracted on every frame of a video set without using any type of interpolation using $ARemCNN$ model. Consider, here, input frame is represented as $\mathbb{X}_D$. To extract high resolution feature vector for every frame $\mathbb{X}_D$.convoluted to a set of filters. In $ARemCNN$ model filter set represented as D_1, L_1, q_1. In $ARemCNN$ model, the size of the first filter sheet D_1 can be represented as $D_1 = 5$ and the feature extraction on every video frame performed directly on original video frames. The channel used in our model $ARemCNN$ is Y out of three $YCbCr$ channel. Therefore, total number of channels present in our model $ARemCNN$ is $q_1 = 1$.In our model $ARemCNN$, L_1represents feature dimension number of low-resolution frames. Here, $Conv(5, m, 1)$ represent first layer feature dimension and m defined as first sensitive parameter.

b. Adaptive Down-Sampling

Here, after feature extraction for each frame we down sample each frame in parallel configuration to get high visual quality. However, existing techniques prefer mapping just after the feature extraction and cannot eliminate the error occurred during feature extraction. They try to convert extracted high dimension features into high-resolution features directly, which does not provide required quality precision. In fact, it enhances computational overhead to a high extent and performance affected due to larger size of n.Thus, down sampling introduced to remove this problem present in the conventional approaches on the extracted features. High-level vision schemes also apply these type of approaches to reduce computational cost.

Thus, to reduce the feature dimension size n largely we have individually down sampled each frame of a video simultaneously. Here, D_2 represent size of second layer filter, which kept as 1 to provide feature linearity. The dimensions of feature L_2 should be kept as $n \ll m$.This shows dimensions of feature decreased from m to n.Here, n represents the sensitivity of second variable which calculates volume of downsampling. Here, $Conv(1, g, s)$ represents dimension of second layer feature whose size taken as (1×1) to reduce dimensionality of features to a high extent.

c. Adaptive Mapping

Adaptive Mapping is one of most important stage of our model $ARemCNN$.Adaptive Mapping is a non-liner mapping stage whose accuracy decides the performance of the system. There are two parameters depth and width which are highly affected while mapping in our model $ARemCNN$. Here, depth shows the number of layers present in our model and width shows how many filters can be placed in a layer. Here, these two factors width and depth helps to perform non-linear mapping on every high dimensional feature extracted from adaptive patch wise feature extraction after down sampling simultaneously on all the frames. In conventional techniques, mapping layer is not introduced for large networks hence our model $ARemCNN$ signifies that this adaptive mapping layer can work efficiently for large networks as well. To handle large networks here we have adaptively mapped all the frames at the same time which helps to enhance performance by faster implementation and saves a lot of computational time. To implement this scheme we consider here a medium size filter $D_3 = 3$. Then, for performance enhancement we multiple layers are considered such as 3×3 layers. A sensitive variable s determines computational complexity and accuracy of the system. Every mapping layer for each frame have same number of filters as $L_3 = n$.Here, $Conv(3, n, n)$ represents non-linear mapping.

d. Adaptive Up-Sampling

Here, adaptive up-sampling used so that error difference of down sampling and up sampling is fed to the sparse coding reconstruction for each frame. In our model $ARemCNN$, downsampling introduced to lessen computational complexity so that performance of image reconstruction enhanced. Here, up-sampling used to create a high quality image for each frame. To maintain synchronization between adaptive up-sampling and down-sampling we perform operations on the layer whose dimensions are 1×1 layers. To increase the efficiency of the

system, up sampling layer can be described as $Conv(1, m, n)$ which is contradictory to downsampling layer.

f. Adaptive Sparse Coding Based Image Reconstruction

The difference between up sampling and down sampling stages is the error which produced during reconstruct of an image and is fed to adaptive sparse coding based image reconstruction. This block reduces this error for each frame simultaneously hence reconstruction quality enhanced. Then the resultant weight factor is a high quality reconstructed image. The sparse coding reconstructed layer described as $SparseCode(9,1,m)$ which shows it consists of 9×9 filter layers.

g. Adaptive Sparse Rectified Linear Unit ($\boldsymbol{ASReLU}$)

To get activation function, a conventional Rectified Linear Unit ($ReLU$) used for each layer of every frame. Here, we have introduced $ASReLU$ (Adaptive Sparse Rectified Linear Unit) instead of conventional $ReLU$.The activation function using $ASReLU$ can be described as:

$$f(z_k) = \max(z_k, 0) + \mathbb{B}_k \min(0, z_k) \quad (1)$$

The input for activation function f can be defined as z_k, $\mathbb{B}_k$ defined as negative phase coefficient of channel k. In our model, $\mathbb{B}_k$ can be either zero or any other value or it can be user defined. However, conventional techniques always prefer $\mathbb{B}_k$ as zero. The difference between the conventional $ReLU$ and our proposed $ASReLU$ is that our model can remove the dead features (features which are of no use) which produces in $ReLU$ due to zero gradient vectors [30]. Here, $ASReLU$ unit can helps to test every parameter of a multilevel network simultaneously for each frame for variety of designs. Our experimental results determine that our $ASReLU$ model can be more effective and accurate than conventional techniques. It also saves computational time to a large extent due to its parallel configuration.

h. Modelling to Deduct Computational Complexity and Cost Function

1. Computational Complexity

Computational complexity is one of the most important factor which decides efficiency of the system. Existing systems consists of high computational complexity which can affect their performance. This is due to drawbacks in the design of conventional $ReLU$. Computational complexity for conventional techniques can be described as,

$$O\{(D_1^2 L_1 + L_1 D_2^2 L_2 + L_2 D_3^2) D_{hr}\} \quad (2)$$

Our proposed model $ARemCNN$, uses the efficient design architecture and $ASReLU$ to handle large complexities produced in the reconstruction of video frames. Our design provide high speed implementation and reduces processing time by discarding dead features produced in the frame reconstruction for each frame simultaneously. The computational complexity can be expressed using our proposed model $ARemCNN$ as,

$$O\{(25m + nm + 9mn^2 + mn + 81n) D_{lr}\} = O\{(9mn^2)\} \quad (3)$$

2. Cost Function

Mean Square Root function (MSE) determines the cost function for our $ARemCNN$ model. Existing techniques uses cost function as:

$$\min_{\vartheta} \frac{1}{L} \sum_{l=1}^{L} \| L(I_n^l; \phi) - J^l \|_2^2 \quad (4)$$

Here, l^{th} low and high resolution reconstructed image pair can be defined as I_n^l and J^l while training of video frames. Resultant system function $L(I_n^l; \phi)$ consists of a ϕ parameter. A typical back propagation technique using stochastic gradient used for the effectiveness of these parameters.

3. Adaptive Sparse Coding Reconstruction Architecture

In our $ARemCNN$ model, the relationship between low resolution input video frames and high resolution output frames can be described using Adaptive Sparse Coding Design (ASCD). This design gives error free precise high resolution reconstructed frames. Our Adaptive Sparse Coding Design (ASCD) synchronized with convolutional neural networks to form a high quality video frames with the help of Modified Learned Iterative Shrinkage and Thresholding (MLIST) technique [31]. Figure 2 shows the architectural diagram of Adaptive Sparse Coding Based Design ($ASCD$).

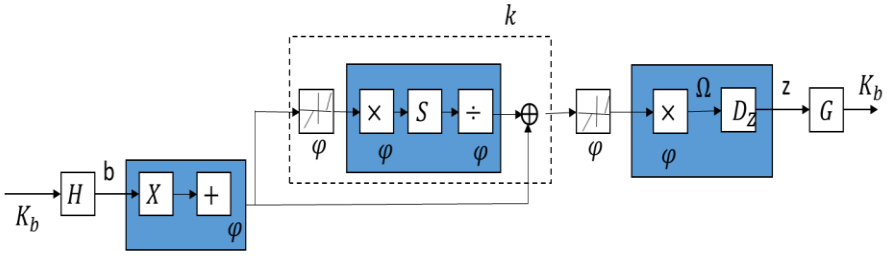

Figure 2. Structural diagram of Adaptive Sparse Coding Based Design ($ASCD$)

Here, ASCD based MLIST technique applied for every frame of a video simultaneously in coordination with convolutional neural networks to get high selectivity. In our $ARemCNN$ model, non-linearity can be explained with the help of $ASReLU$. Our model $ARemCNN$ in coordination with ASCD can work very efficiently by deducting computational complexity to a large extent. A weight map produced for each high resolution video frame simultaneously which rely upon their pixels. For each resultant frame the produced weight map multiplied with its corresponding pixels. Then every product of video frame simultaneously summed up to form high quality video frames. The resultant reformed video frames can be expressed as:

$$L(\mathbb{B};\ominus) = \sum_{l}^{q} \mathbb{X}_l(\mathbb{B};\phi_\omega) \odot L_{M_l}(\mathbb{B};\phi_{M_l}) \tag{5}$$

Where, low resolution input video frame can be explained as $\mathbb{B}$. Equation (5) consists of two functions in which function $\mathbb{X}_l(\mathbb{B};\phi_\omega)$ shows the behavior of produced weight maps and second function $L_{M_l}(\mathbb{B};\phi_{M_l})$ shows the reconstruction of resultant high resolution video frame M_l. Here, Equation (5) shows the point multiplication of weighted map for every reconstructed resultant video frame with their pixels. Here, equation (6) shows the reduction in loss between input and output frame while training.

$$\min_{\ominus} \sum_{g} \| L(\mathbb{B}_g;\ominus) - b_g \|_2^2 \tag{6}$$

Where, resultant output video frames can be represented by function $L(\mathbb{B}_g;\ominus)$, $\mathbb{B}_g$ is a high resolution video frame of g^{th} type and b_g shows the equivalent low resolution video frame. In our model $ARemCNN$, a set of parameters grouped together using $\ominus$ expression. In our model $ARemCNN$, cost function is described as combination of equation (5) and (6),

$$\min_{\phi_\omega,\{\phi_{M_l}\}_{l=1}^{q}} \sum_{l} \| \sum_{l}^{q} \mathbb{X}_l(\mathbb{B}_g;\phi_\omega) \odot L_{M_l}(\mathbb{B}_g;\phi_{M_l}) - b_g \|_2^2 \tag{7}$$

4. Performance Evaluation

We compute our outcomes with the similar dataset (Videoset4 and YUV21) as used in [7-26] to compare the performance and efficiency of our model to the existing techniques discussed in the related work. Our model is trained on different large datasets like Videoset4 and YUV21 [7-26]. Testing results shows that our model outperforms most of the existing techniques in terms of PSNR and reconstruction efficiency. We have tested our model for different up scaling factors (2, 3 and 4). Our results show accuracy and reconstruction efficiency increment to a large extent. Our model needs less amount of execution time to provide effective video scaling. Our model implemented on 64-bit windows 10 OS with 16 GB RAM which consists on INTEL (R) [31] core (TM) i5-4460 processor. It consists of 3.20 GHz CPU. We have compared our model with Enhancer [32], Bayesian [33] and Bayesian-MB [34], VSRnet AMC [25], VSRnet MC [25] and many other existing techniques.

4.1. Implementation Details

We have implemented our extensive experiments on large video datasets like Videoset4 and YUV21. In modern era, the availability of high-end devices is highly increased. However, in real world excess of low-resolution videos are available. Therefore, there is a huge demand of converting low resolution videos to high-resolution videos in market. These upgradation of high resolution videos can be possible using upscaling factor. Therefore we have used different upscaling factors to achieve these objectives by measuring performance and accuracy of the model for upscaling factor 2, 3 and 4. Videoset4 dataset contains total 4 videos such as walk, foliage, city and calendar. They all are of different resolutions as 720 ×480,720 ×576,704 ×576. In this paper, we use one more dataset as YUV21 to verify accuracy and efficiency of our model ARemCNN for different upscaling factors. The YUV21 dataset contains total 21 videos of similar size as 352×288 resolution. Here, we have considered BUS video to verify our model from the YUV21 dataset. All the experiments are undertaken on the MATLAB 16b framework in configuration with CAFFE.

4.2. Comparative Study

4.2.1. Videoset4 Dataset

Here, we have taken all 4 videos from the Videoset4 dataset considering upscale 2, 3 and 4 and compared it with existing algorithms. All the videos are compared to nine most popular existing techniques. Here, we have taken Videoset4 dataset which consists of total 4 videos such as city, walk, foliage and calendar. All the videos consists of different number of frames and resolution. All 4 videos walk, foliage, calendar and city has resolution as $720 \times 480, 720 \times 480, 720 \times 576, 704 \times 576$ respectively. Here, calendar has 41 frames, city has 34 frames, walk has 47 frames and foliage consists of total 49 frames. Here, we have evaluated the performance of our model $ARemCNN$ based on Average PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity Index) using upscale-2, 3 and 4. Our experimental outcomes shows that our model outperforms all the existing techniques verified in Table 1 in terms of PSNR and Table 2 in terms of SSIM. Our model $ARemCNN$ has average PSNR for upscale 2 is $39.81dB$, upscale 3 is $35.56\ dB$ and upscale 4 is $33.77dB$ which is very high from other existing techniques. Similarly, Our model $ARemCNN$ has average SSIM for upscale 2 is 0.9549, upscale 3 is 0.8879 and upscale 4 is 0.8351 which is very high from other existing techniques. Table 3 and 4 shows the PSNR and SSIM values for all four videos city, calendar, walk and foliage from the videoset4 dataset using upscale 2, 3 and 4.

Table 1. Average PSNR for videoset4 considering upscale 2, 3 and 4

Scale	Bicubic	A+ [33]	SRCNN [35]	ANN [36]	Bayesian [7]	Bayesian-MB [34]	Enhancer [32]	VSRnet [25]	MCResNet [37]	Our $\boldsymbol{ARemCNN}$
2	28.43	30.53	30.70	29.04	29.69	30.63	30.40	31.30	32.28	39.81
3	25.28	26.36	26.51	25.94	25.82	26.43	26.34	26.79	27.54	35.56
4	23.79	24.59	24.69	23.97	25.06	24.14	24.55	24.84	25.45	33.77

Table 2. Average SSIM for videoset4 considering upscale 2, 3 and 4

Scale	Bicubic	A+ [33]	SRCNN [35]	ANN [36]	Bayesian [7]	Bayesian -MB [34]	Enhancer [32]	VSRnet [25]	MCResNet [37]	Our $ARemCNN$
2	0.8676	0.9154	0.9172	0.8835	0.9055	0.9226	0.9151	0.9278	0.9433	0.9549
3	0.7329	0.7904	0.7933	0.7705	0.8323	0.8071	0.7948	0.8098	0.8448	0.8879
4	0.6332	0.6889	0.6918	0.6437	0.7466	0.6864	0.6877	0.7049	0.7467	0.8351

Table 3. PSNR for videoset4 (all 4 videos) considering upscale 2, 3 and 4

Videoset4	UPSCALE-2	UPSCALE-3	UPSCALE-4
CITY	35.72140739	31.677685	30.29367416
WALK	40.55026076	35.568857	34.47049427
FOLIAGE	39.07009192	35.782224	33.71167137
CALENDER	43.90874099	39.24466	36.61122388
Average	39.81262527	35.568357	33.77176592

Table 4. SSIM for videoset4 (all 4 videos) considering upscale 2, 3 and 4

Videoset4	UPSCALE-2	UPSCALE-3	UPSCALE-4
CITY	0.959551753	0.878086806	0.835809511
WALK	0.983058431	0.953960826	0.919574307
FOLIAGE	0.958292675	0.885953635	0.819836075
CALENDER	0.918922166	0.833916866	0.765302394
Average	0.954956256	0.887979533	0.835130572

4.2.2. YUV21 Dataset

Here, YUV21 dataset consists of total 21 videos of similar resolution 352×288 and out of which we have taken BUS video from the YUV21 dataset considering upscale 2, 3 and 4 and compared it with existing algorithms. All the videos are compared to nine most popular existing techniques. Here, BUS video consists of total 150 frames. Here, we have evaluated the performance of our model $ARemCNN$ based on Average PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity Index) using upscale-2, 3 and 4. Our experimental outcomes shows that our model outperforms all the existing techniques verified in Table 5 in terms of PSNR and Table 6 in terms of SSIM. Our model $ARemCNN$ has average PSNR for upscale 2 is $38.71\ dB$, upscale 3 is $34.58\ dB$ and upscale 4 is $33.047dB$ which is very high from other existing techniques. Similarly, Our model $ARemCNN$ has average SSIM for upscale 2 is 0.949, upscale 3 is 0.870 and upscale 4 is 0.819 which is very high from other existing techniques.

Table 5. Average PSNR for YUV21 considering upscale 2, 3 and 4

Scale	Bicubic	A+ [33]	SRCNN [35]	VDSR [38]	Bayesian [7]	Bayesian -MB [34]	Enhancer [32]	VSRnet [25]	MCResNet [37]	Our $ARemCNN$
2	30.58	33.09	33.39	34.16	31.99	31.63	31.78	33.44	34.37	38.71
3	27.71	29.34	29.51	30.34	28.62	28.54	28.33	29.55	30.11	34.58
4	26.29	27.50	27.66	28.39	26.14	26.48	26.74	27.63	28.08	33.047

Table 6. Average SSIM for YUV21 considering upscale 2, 3 and 4

Scale	Bicubic	A+ [33]	SRCNN [35]	VDSR [38]	Bayesian [7]	Bayesian -MB [34]	Enhancer [32]	VSRnet [25]	MCResNet [37]	Our $ARemCNN$
2	0.8752	0.9168	0.9186	0.9266	0.8999	0.8952	0.9018	0.9207	0.9338	0.949
3	0.7727	0.8195	0.8211	0.8392	0.8271	0.8033	0.7963	0.8246	0.8452	0.87
4	0.7063	0.7499	0.7529	0.7741	0.7339	0.7289	0.7248	0.7539	0.7746	0.819

5. Image Reconstruction Comparison

5.1. Image Reconstruction Comparison

Here, we have demonstrated 350th frame of city and walk as used in all the other existing techniques. The original Videoset4 video dataset contains total 4 different videos of different size and its original resolution is $720 \times 480, 720 \times 480, 720 \times 576, 704 \times 576$. We have shown

PSNR and image reconstruction quality comparison with all the conventional techniques. The PSNR result $39.81\ dB$ outperforms all the existing state-of- the- art techniques for upscale 2, similarly for upscale 3 and 4 is $35.56\ dB$ and $33.77dB$. From our experimental results it is clearly visible that our reconstruct frame has better reconstruction quality than any other recent existing techniques.

Ground Truth

Bicubic- $29.95\ dB$

SRCNN- $32.42\ dB$

VDSR- $33.26\ dB$

DRCN- $33.80\ dB$

$VSRnetAMC\ -\ 33.01\ dB$

$VSRnet\ MC - 34.09\ dB$

$MCResNet\ - 34.98\ dB$

$ARemCNN\ - 35.721dB$

Figure 3. City video frame in the Videoset4 dataset for scale factor 2 using different methods

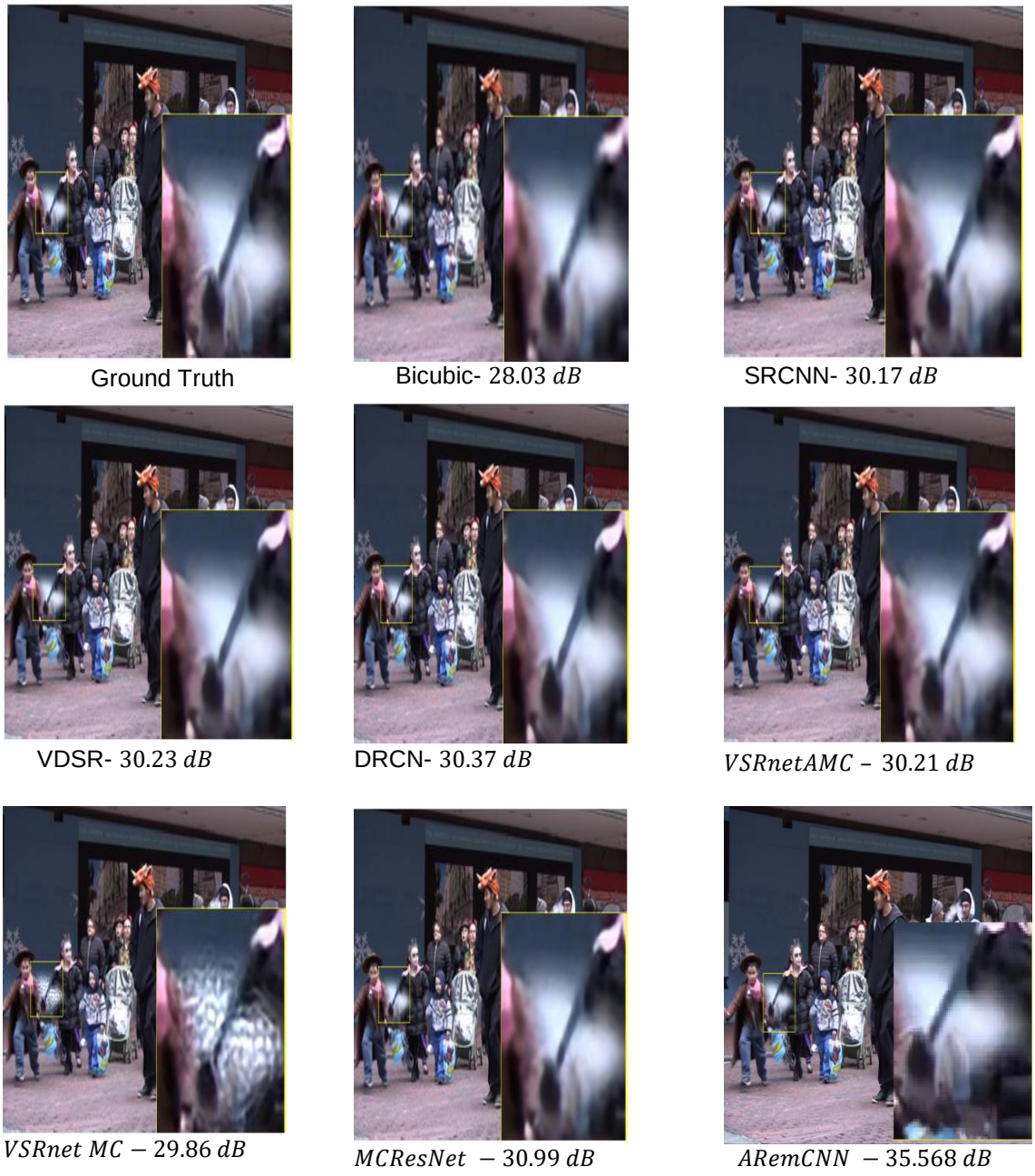

Figure 4. Walk video frame in the Videoset4 dataset for scale factor 3 using different methods

5.2. YUV21 Dataset

Here, we have demonstrated 350th frame of BUS as used in all the other existing techniques. The original YUV21 video dataset contains total 21 different videos of same size and its original resolution is 352×288. We have shown PSNR and image reconstruction quality comparison with all the conventional techniques. The PSNR result 38.71 dB outperforms all the existing state-of- the- art techniques for upscale 2, similarly for upscale 3 and 4 is 34.58 dB and 33.047 dB. From our experimental results it is clearly visible that our reconstruct frame has better reconstruction quality than any other recent existing techniques.

Figure 5. Bus video frame in the YUV21 dataset for scale factor 2 using different methods

5.3. Graphical Analysis

5.3.1. Videoset4 Dataset

The following graphs shows the comparison between our proposed model and existing approaches $VSRnetMC$ [25] and $MCResNet$[37] for upscale 2, 3 and 4 considering Videoset4 dataset. Figure 6 shows average PSNR comparison considering upscale -2, 3 and 4 for all the 4 videos city, calendar, foliage and walk with existing techniques. Figure 7 shows average SSIM comparison considering upscale 2, 3 and 4 for all the 4 videos city, calendar, foliage and walk with existing techniques. Figure 8 demonstrates PSNR comparison considering upscale-2, 3 and 4 for all the 4 videos city, calendar, foliage and walk. PSNR for upscale-2 using our proposed $ARemCNN$ technique is $39.81\ dB$, with upscale-3 is $35.56\ dB$ and from upscale-4 is $33.77\ dB$. Similarly, figure 9 shows SSIM comparison considering upscale-2, 3 and 4 for all the 4 videos city, calendar, foliage and walk. SSIM for upscale-2 using our proposed $ARemCNN$ technique is 0.9549 , with upscale-3 is 0.8879 and from upscale-4 is 0.8351.

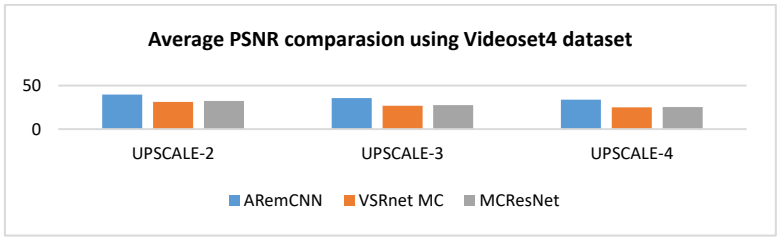

Figure 6. Average PSNR comparison using VIDEOSET4 dataset for upscaling 2, 3 and 4

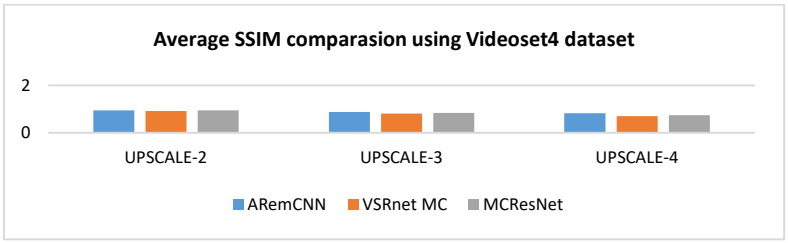

Figure 7. Average SSIM comparison using VIDEOSET4 dataset for upscaling 2, 3 and 4

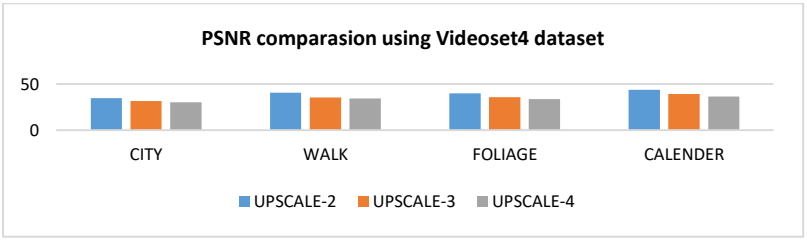

Figure 8. PSNR comparison using VIDEOSET4 dataset for upscaling 2, 3 and 4 for all 4 videos

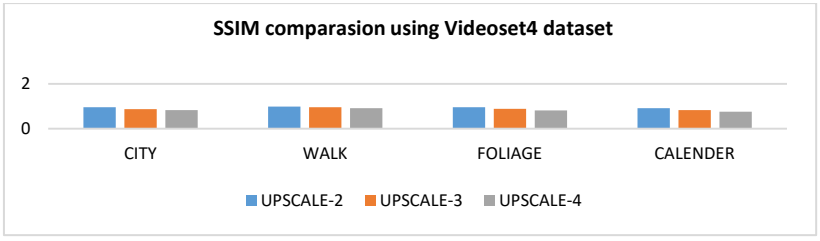

Figure 9. SSIM comparison using VIDEOSET4 dataset for upscaling 2, 3 and 4 for all 4 videos

5.3.2 YUV21 Dataset

The following graphs shows the comparison between our proposed model and existing approaches $VSRnetMC$ [25] and $MCResNet$[35] for upscale 2, 3 and 4 considering YUV21 dataset. Figure 10 shows average PSNR comparison considering upscale -2, 3 and 4 for BUS video with existing techniques. Figure 11 shows average SSIM comparison considering upscale 2, 3 and 4 for BUS video with existing techniques. PSNR for upscale-2 using our proposed $ARemCNN$ technique is $38.71\ dB$, with upscale-3 is $34.58\ dB$ and from upscale-4 is $33.047\ dB$.

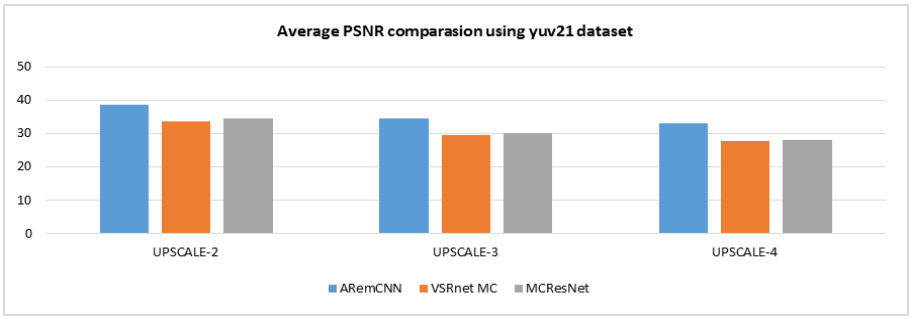

Figure 10. Average PSNR comparison using yuv21 dataset for upscaling 2, 3 and 4

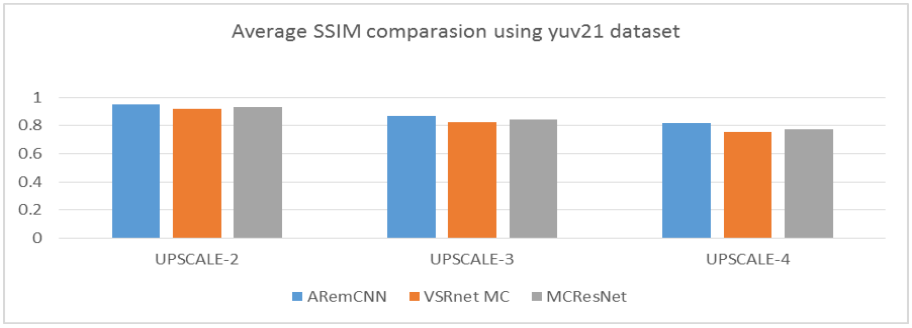

Figure 11. Average SSIM comparison using YUV21 dataset for upscaling 2, 3 and 4

6. Conclusion

The significance and complexities in designing real time error minimization adaptive model discussed. Drawbacks of current systems and its variants is presented. To provide ease of accessibility of high resolution videos to the subscribers we have introduced robust Adaptive Reconstruction Error Minimization Convolution Neural Network ($ARemCNN$) architecture which helps to eliminate the drawbacks of existing algorithms and provides high quality features with parallel configuration to save time and provide high speed. Convolution Neutral Network (CNN) provides lightning speed, faster implementation, easy training and can handle large datasets with ease by GPU parallel computing. Our proposed model can easily train the bulky datasets such as YUV21 and Videoset4.Our experimental results shows that our model outperforms many existing techniques in terms of PSNR, SSIM and reconstruction quality. The experimental results shows that our average PSNR result is 39.81 considering upscale-2, 35.56 for upscale-3 and 33.77 for upscale-4 for Videoset4 dataset which is very high in contrast to other existing techniques. Similarly, the experimental results shows that our average PSNR result is 38.71 considering upscale-2, 34.58 for upscale-3 and 33.047 for upscale-4 for YUV21 dataset which is very high in contrast to other existing techniques. Similarly, SSIM results for both datasets are very high compare to existing techniques.

This results proves our proposed model $ARemCNN$ robustness, high efficiency and better performance. Our proposed model can be effectively used in the applications such as medical, satellite imaging, surveillance, HDTV, video coding or decoding, stereoscopic video processing, and face recognition for future purpose to reconstruct efficient images or video frames.

References

[1] W Shi, J Caballero, C Ledig, X Zhuang, W Bai, K Bhatia, A Marvao, T Dawes, D ORegan, D Rueckert. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, editors, *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 2013; 8151 of *LNCS*: 9–16.

[2] MW Thornton, PM Atkinson, Da Holland. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. *International Journal of Remote Sensing*. 2006; 27(3): 473–491

[3] C Szegedy, W Liu, Y Jia, P Sermanet, S Reed, D Anguelov, D Erhan, V Vanhoucke, and A Rabinovich.Going deeper with convolutions. *arXiv preprint:1409.4842*, 2014.

[4] J Zhang, Y Cao, ZJ Zha, Z Zheng, CW Chen, Z Wang. A unified scheme for super-resolution and depth estimation from asymmetric stereoscopic video. *IEEE Trans. Circuits Syst. Video Technol.*, 2016; 26(3): 479–493.

[5] BC Song, S-C Jeong, Y Choi.Video super-resolution algorithm using bi-directional overlapped block motion compensation and onthe-fly dictionary training. *IEEE Trans. Circuits Syst. Video Technol*. 2011; 21(3): 274–285.

[6] L. Zhang, H. Zhang, H. Shen, and P. Li. A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*. 2010; 90(3): 848–859.

[7] C. Liu, D. Sun.A bayesian approach to adaptive video super resolution. Proc. of IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR). 2011: 209–216.

[8] J Zhang, Y. Cao, Z. J. Zha, Z. Zheng, C. W. Chen and Z. Wang.A Unified Scheme for Super-Resolution and Depth Estimation From Asymmetric Stereoscopic Video. *IEEE Transactions on Circuits and Systems for Video Technology*. 2016; 26(3): 479-493.

[9] Z. Jin, T. Tillo, C. Yao, J. Xiao and Y. Zhao.Virtual-View-Assisted Video Super-Resolution and Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*. 2016; 26(3): 467-478.

[10] H Zhang, Q. Yuan. A fast video super-resolution framework derived from key-frame model. *2016 3rd International Conference on Systems and Informatics (ICSAI)*, Shanghai, 2016, pp. 942-948

[11] J. Huang; M. Sun; J. Ma; Y. Chi.Super-Resolution Image Reconstruction for High-Density 3D Single-Molecule Microscopy. *IEEE Transactions on Computational Imaging*. 99: 1-1

[12] A Krizhevsky, I Sutskever, and G E Hinton. Imagenet classification with deep convolutional neural network. *In NIPS*. 2017: 1097–1105.

[13] Z. Shi, S. Yao, Y. Zhao. *A novel video image scaling algorithm based on morphological edge interpolation*. IEEE International Conference on Neural Networks and Signal Processing. 2008.

[14] Y. Xie, J. Xiao, T. Tillo, Y. Wei, Y. Zhao. 3D video super-resolution using fully convolutional neural networks. *IEEE International Conference on Multimedia and Expo (ICME)*, Seattle, WA, 2016: 1-6.

[15] A. Mollahosseini, D. Chan and M. H. Mahoor.Going deeper in facial expression recognition using deep neural networks. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid, NY. 2016: 1-10

[16] X. Wang, L. Gao, J. Song and H. Shen.Beyond Frame-level CNN: Saliency-Aware 3-D CNN With LSTM for Video Action Recognition. *IEEE Signal Processing Letters*. 2017; 24(4): 510-514.

[17] J. Lee and I. C. Park.High-Performance Low-Area Video Up-Scaling Architecture for 4-K UHD Video. in *IEEE Transactions on Circuits and Systems II: Express Briefs*. 2017; 64(4): 437-441

[18] S. Park, I. Hong, J. Park and H. J. Yoo.An Energy-Efficient Embedded Deep Neural Network Processor for High Speed Visual Attention in Mobile Vision Recognition SoC. *IEEE Journal of Solid-State Circuits*. 2016; 51(10): 2380-2388.

[19] W. Shi *et al*. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV. 2016: 1874-1883.

[20] L. Li, J. Du, M. Liang, J. M. Lee and L. Meng.Video super-resolution based on nonlinear mapping and patch similarity. *4th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, Beijing. 2016: 48-56.

[21] Z. Zhao, L. Song, R. Xie, X. Yang. GPU accelerated high-quality video/image super-resolution. *2016 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Nara. 2016: 1-4.

[22] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han and T. S. Huang. Robust Single Image Super-Resolution via Deep Networks With Sparse Prior. *IEEE Transactions on Image Processing*. 2016; 25(7): 3194-3207.

[23] Harmonic Inc. http://www.harmonicinc.com/resources/videos/4kvideo-clip-center. 2014.

[24] J. Deng, W Dong, R Socher, L Li. *A large-scale hierarchical image database*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2009; 248–255.

[25] http://www.codersvoice.com/a/webbase/video/08/152014/130.html.

[26] Ouyang W, Wang X. *Joint deep learning for pedestrian detection*. IEEE International Conference on ComputerVision. 2013; 2056–2063.

[27] Sun Y, Chen Y, Wang X, Tang X. *Deep learning face representation by joint identification-verification*. In: Advances inNeural Information Processing Systems. 2014; 1988–1996.

[28] Szegedy C, Reed S, Erhan D, Anguelov D. Scalable, highquality object detection. arXiv preprint arXiv:1412.1441. 2014.

[29] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: ECCV, Springer. 2014: 818–833.
[30] Wang Z, Liu D, Yang J, Han W, Huang T. *Deep networks for image super- resolution with sparse prior*. Proceedings of the IEEE International Conference on Computer Vision. 2015; 370-378
[31] R. Timofte, V. De Smet, L. Van Gool. *A+: Adjusted Anchored Neighborhood Regression for Fast Super-Resolution.* IEEE Asian Conference on Computer Vision. 2014; 1920–1927.
[32] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, and E. Wu. *Handling motion blur in multi-frame super-resolution*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5224–5232.
[33] A. Kappeler, S. Yoo, Q. Dai, A. K. Katsaggelos.Video Super-Resolution With Convolutional Neural Networks. *IEEE Transactions on Computational Imaging*. 2016; 2(2): 109-122. doi: 10.1109/TCI.2016.2532323
[34] D. Li; Z. Wang.Video Super-Resolution via Motion Compensation and Deep Residual Learning. in *IEEE Transactions on Computational Imaging*. 99: 1-1 doi: 10.1109/TCI.2017.2671360
[35] M. Cheng, N. Lin, K. Hwang, and J. Jeng. *Fast video super-resolution using artificial neural networks*. 2012 8th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP). 2012; 1–4.
[36] C. Dong, C. C. Loy, K. He, X. Tang. *Learning a Deep Convolutional Network for Image Super-Resolution.* in Proceedings of the IEEE European Conference on Computer Vision, 2014.
[37] J Kim, JK Lee, KM Lee. *Accurate image super-resolution using very deep convolutional networks*. in Proc. IEEE Conf. Comput. Vis. Pattern Recog. 2016; 1646–1654.