International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 5 (2018) pp. 2558-2568


# Deep–CNN Architecture for Error Minimisation in Video Scaling

**Safinaz S.[1] and Dr. Ravi Kumar AV[2]**

[1]*Department of Electronics and Electrical Engineering, Sir M.V.I.T, Bangalore Karnataka, 562 157, India.*
[1]*E-mail:* safinaz8383@gmail.com
[2]*Department of Electronics and Electrical Engineering, SJBIT, Bangalore, Karnataka, 560060, India.*
[2]*E-mail: avr187@gmail.com*

**Abstract**

People like to watch high-quality videos, so high-resolution videos are in more demand from few years. The techniques like DWT which are used to obtain the high-quality videos results with high distortions in videos which ends in low resolution. But there are numerous super-resolution techniques in signal processing to obtain the high-resolution frames from the multiple low resolution frames without using any external hardware. Super-resolution techniques offer very cheap and efficient ways to obtain high-resolution frames. Convolutional Neural Networks (CNN) technique is most widely used Deep-Learning technique for various application such as feature extraction, face detection, image classification, image scaling etc. Removing a noise from the frame is a very difficult task, so a CNN is introduced with super-resolution technique. Moreover, CNN technique can easily train bulky datasets, remove blurriness and can provide the end-to-end mapping between high and low-resolution patches. Therefore, here, we have introduced an efficient and robust Reconstruction Error Minimization Convolution Neural Network Architecture. Here, our model is highly efficient to handle large datasets and provide visually attractive results compared to existing state techniques using CNN architecture. The proposed CNN model has an additional unit of Pipelined structure to increase the processing speed operating on large datasets. Our experimental results verify that our model outperforms other existing state of-art-techniques in terms of Peak Signal to Noise Ratio-PSNR, Structural Similarity Index Matrices - SSIM and visual quality appearance.

**Keywords:** Video Scaling, Super-resolution, CNN- Deep-Learning Architecture, CAFFE

## INTRODUCTION

A major trend in mobile realm and current Internet communication augments voice conversation is with video [1], such as Video conferencing over IP (VoIP) but faces the problem of fluctuating network and heterogeneous conditions. Video scaling is very essential where individual wanted to watch video on high resolution full screen displaying devices regardless of high resolution video or a low resolution video. [2], [3], [4] Scaling can be of two types of: upscaling and downscaling. Upscaling is required to obtain a high quality video from an existing low resolution dataset. Therefore, here to obtain a high resolution images or videos from a multiple low resolution video frames, a technique called super resolution approach is used. [5]Super resolution technique is used to upscale the existing low resolution images or videos to the high resolution frames. This type of video scaling's are very useful in the fields such as face recognition, video coding or decoding, satellite imaging etc. In case of super resolution technique, it captures the video and displays the video in which it will helps in production of UHD content. [6]To obtain a noise free video or image with high quality is a very difficult task in recent years. To retrieve the high quality images or videos needs high amount of precision point and also needs high speed to access large amount of datasets in which the normal techniques cannot be able to handle because of the noise present in it and their less resolution standard. [7]Therefore to overcome this difficulty super resolution technique have been introduced. The normal existing techniques are not enough to decrease the noise effects in the images or videos and reconstruct a high quality videos. Super resolution approaches uses signal processing algorithm to recreate a high resolution images or videos from the multiple low resolution images or videos. [8]This technique uses no hardware to retrieve the high quality videos or images and also it is very cheap to develop and it decreases the noise effects like blurring effects, ring to the maximum extent. So this is the best and smart way to retrieve the high quality videos.

Super resolution approach also faces the problem of noise effect and blurring effects, so in recent times a lot of researchers have accepted this super resolution technique. Therefore to overcome this noise effects and blurring effects from the images or videos, super resolution technique uses the end-to-end deep learning framework which will gives the less noise high quality images or videos. [9]This deep learning methods are used to obtain an end to end mapping between the low-resolution and high-resolution patches. Since this deep learning approach also results in noise effects in the videos. So it has become a very difficult task to reduce all these noise effects, reconstruction distortion, less resolution quality. So for creating a noise and distortion free high quality images or frames there is a technique introduced which is called convolution neural network (CNN) which is one of the type of deep learning framework. [10]This technique helps in obtaining the effective and robust high quality features from the low resolution images. Here, CNN uses the deep learning framework approach to obtain a maximum noise free effective high quality images or videos.

In all types of deep learning frameworks, convolution neural network (CNN) is the best type which has been proved in recent decades because of its simpler training on large datasets, [11]compatibility with parallel GPU computing, rapid implementation, and the high speed. Therefore, here, we have introduced a robust and efficient Reconstruction Error

Minimization Convolution Neural Network Architecture. Also to make a processing much faster and efficient than the normal processing, [12] and [13] CNN uses an extra additional unit. An additional unit used is nothing but the pipeline structure where video scaling has consisted of a pipeline structure. Video scaling consists a pipeline structure which is useful in post-processing of large datasets and increasing sharpness, contrast, and color of the video frames. [14]Therefore, video scaling is the best technique for the use of CNN to overcome the video scaling issues.

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.

## RELATED WORK

Nowadays, lots of high dimensional displays with high-quality images or videos are in progress. So, for restoring a high-quality, high-resolution video frames from the several low-resolution frames is a challenging task in recent days. Therefore we have researched in this area to improve and obtain high-quality video frames. In [15] complex compression artifacts, particularly the blocking artifacts, ringing effects, and blurring are introduces in Lossy compression. Therefore, Existing algorithms either focus on removing blocking artifacts and produce blurred output or restores sharpened images by eliminating ringing effects.

Exhilarated and influenced by the deep convolutional networks (DCN) on super-resolution, researchers formulated a impermeable and proficient network for smooth, coherent, consistent, flawless attenuation of different compression antiquity and artifacts. In [16] Sparse coding has been widely applied to learning based single image super-resolution (SR) and has obtained zealous performance by jointly learning effective, potent and well founded representations and correspondences for low-resolution and high-resolution image patch pairs.

Nonetheless, the emanated HR images often go through and get deteriorate from ringing, jaggy, and blurring antiquity due to the assumptions that the LR image patch representation is equal to, and linear with support set that corresponds the HR image patch representation [17]. Prompted by the success of deep learning, researchers developed several coupled data-driven models. Latterly, [18] models based on deep neural networks have reached great success in terms of both reconstruction accuracy, exactitude and computational attainment for single image super resolution Before reconstruction a High resolution image is obtained from a low resolution (LR) input image, up scaled using a common bi-cubic interpolation technique. This implies that the super-resolution (SR) operation is performed in HR space, resulting in sub-optimal and computational complexity. The first convolutional neural network (CNN) is capable of real-time SR of 1080p videos on a single K2 GPU, for which, the researchers have prepared a different CNN architecture where the feature maps are extracted in the LR space. [19] In addition, this paper introduced an efficient sub-pixel convolution layer which learns an array of upscaling filters to upscale the final LR feature maps into the HR output.

By doing so, they efficaciously restored the home-grown bi-cubic filter in the SR pipeline with more complex upscaling filters precisely trained for each feature map, at the same time parsing down the computational complexity of the overall SR operation. [20] They evaluated the proposed approach using images and videos from publicly available datasets and show that it performs significantly better (+0.15dB on Images and +0.39dB on Videos) and is an order of magnitude faster than previous CNN-based methods.[21] Influenced by the recent approaches of image super resolution using convolutional neural network (CNN), a CNN-based block upsampling scheme for intra frame coding is proposed where a block is down-sampled before being compressed by normal intra coding, and then up-sampled to its original resolution. Being a unique approach from previous studies on down/up-sampling based coding, the up-sampling methods in this scheme have been designed by training CNN instead of hand-crafted. This modish CNN structure for up-sampling consists of features deconvolution of feature maps, multi-scale fusion, and residue learning, making the network both compact and efficient. [22]They also design different networks for the up-sampling of luma and chroma components, respectively, where the chroma up-sampling CNN utilizes the luma information to boost its performance. [23]This proposed network has two stages to perform up-sampling, the first stage is a block-by-block coding loop and the second stage is to process and refine block boundaries.

In this paper, we have introduced a robust and efficient Reconstruction Error Minimization Convolution Neural Network Architecture which helps to counter the existing problems of conventional algorithms. Our model helps to provide high reconstruction quality.

## DEEP-LEARNING CNN ARCHITECTURE BASED IMAGE SUPER-RESOLUTION

Convolutional Neural Network has been extensively passed down in deep learning frameworks with very promising achievements due to its capability of faster training on large datasets like Myanmar [24], ImageNet [25], videoset4 [8], yuv21 [26], Dash-Svc dataset [27] using various frameworks such as CAFFE. CNN Architecture can widely use in various pertinence such as Medical, satellite imaging, face recognition, stereoscopic video processing, video coding/decoding and surveillance [28]-[29]. Therefore, CNN Architecture can help to achieve high resolution quality from the low resolution frames. Here, we have presented a Convolution Neural Network architecture to reduce the noise and blurriness in low-quality frames and compared with other existing state-of-art scaling techniques. CNN architecture inheres multiple layers and every convolution layer described using its filter weights, evaluated during the training phase by an iterative update method. Our proposed model also reduces computational complexity with comparatively easier training and quick implementation. The visual appearance of various medical

images can be enhanced by providing a seam between feed-forward neural techniques and adaptive filters. In our model, CAFFE framework based parallel GPU computing used for faster computation on large datasets and train CNN. Also, the exertion of Adaptive Sparse Rectified Linear Unit makes our architecture more robust.

**CNN Architecture Modeling**

CNN architecture consists of assorted layers and every convolution layer described using its filter weights which can be evaluated in training phase by an iterative update method. Initially, these weights are initialized and then fine-tuned by back propagation method to depreciate the cost function. In the testing phase, all the weights become fixed. These weights play an eloquent role in emphasizing the input patches of the test images. Filter weights can work as a reference signal/ visual pattern to measure the correlation between input patches. For every filter weights, its correlation measured to compute the similarity between various filter weights and patches. Filter weights demonstrate a spectral decomposition of an input image. These weights are usually under complete and orthogonal. We divide a large image into various smaller patches and process them in parallel. CNN learns the correlation between input and output and collects the learned knowledge in their filter weights. CNN consists of various layers in its architecture such as convolution layer, a pooling layer, a normalization layer, a fully connected layer, and loss layer etc. The detailed expression for these layers in the architecture represented by equation 1.

$$I1 \rightarrow v1 \rightarrow I2 \rightarrow \cdots \rightarrow IM-1 \rightarrow vM-1 \rightarrow IM \rightarrow vM \rightarrow a \quad (1)$$

Where, equation 1 shows the working of CNN layer by layer while moving forward. Here, $i^1$ represents the input image. Then, input image $I^1$ passes to the first layer $v^1$. The first layer $v^1$ output and second layer input is denoted as $I^2$. This processing undergo till the last layer and the output of last layer is represented as $I^M$. An additional block, backward error propagation $v^M$ is added to eliminate errors and learning the good parameters. The last layer is called as loss layer. Let, $g$ is the respective ground- truth value for input image $I^1$. To evaluate the discrepancy between the CNN estimation $I^M$ and the ground- truth value $g$, a loss function $z$ can be introduced which is expressed as,

$$a = (2)^{-1}\|g - I^M\|^2 \quad (2)$$

Here, $g$ and $I^M$ are probability functions used for the evaluation of distance between them. CNN model learns all the parameters $v^1, v^2 \ldots\ldots\ldots v^{M-1}$ during implementation of training and these trained parameters helps to estimate output. Moreover, several phases need to be passed through for precise scaling of low resolution images using our proposed model based on CNN architecture. The various phases which is presented in figure 1 are as follows:

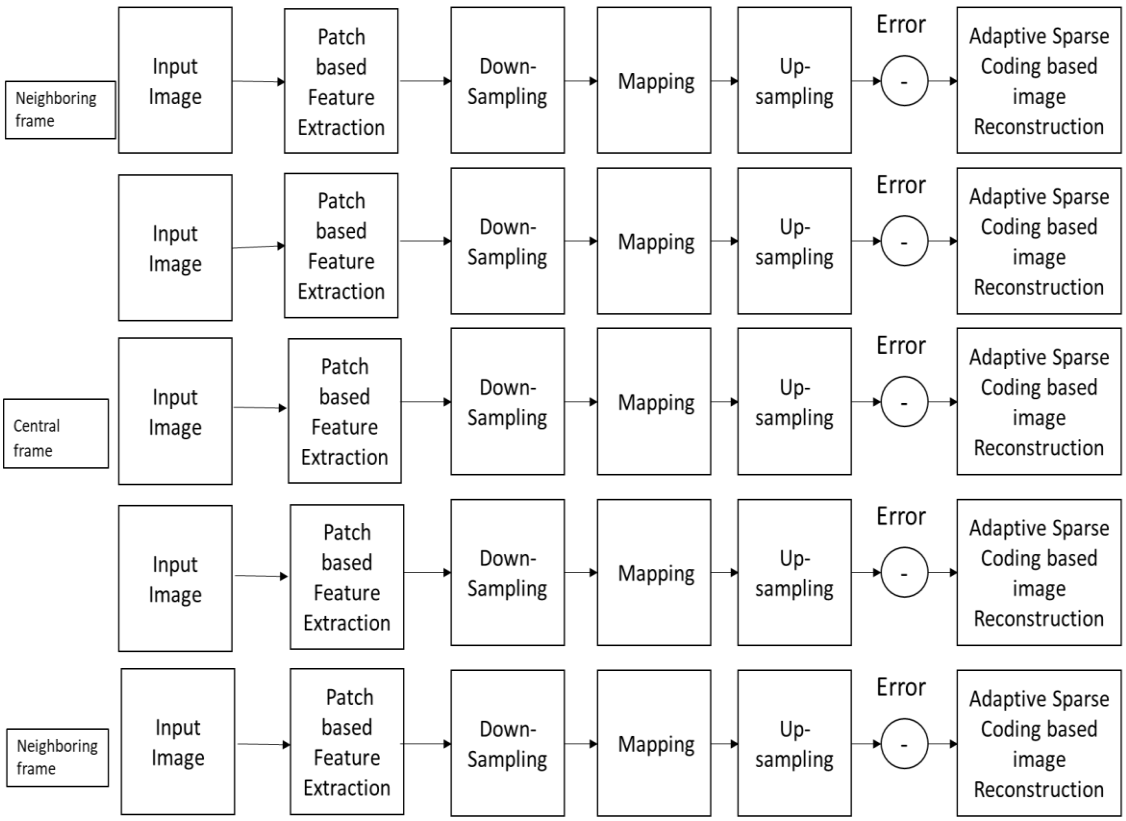



**Figure 1.** Architecture diagram of CNN based proposed method

### A. CNN based Feature Representation for LR frames

Super-resolution refers to converting a low resolution frame into a high resolution frame of a similar or different scene. There are various methods available to get high resolution images from low resolution images of similar scenes such as [3], [5], [8] etc. However, these techniques consist of multiple problems such as ill-poised, de-noising and de-blurriness problem, blocking effects in low-resolution frames, high computational complexity, reconstruction error and distortion, slower implementation, and redundancy in the pixels/frames. Therefore, to accord with these problems efficiently, we have introduced a CNN based architecture which can provide high resolution images from low-resolution images with faster implementation using a set of similar scenes for training purposes. CNN technique utilized to eliminate inverse problems such as upscaling, downscaling and de-convolution.

This section pacts with the issues of feature extraction of high and low resolution frames. Here, every input image is segmented into various patches of size $5 \times 5$. Feature vector points defines the every patches of an image. The low resolution ($LR$) and high resolution ($HR$) images can be described as $I_y$ and $J_y$. Here, for both low resolution ($LR$) and high resolution ($HR$) images similar number of patches can be extracted. Similarly, some other $LR$ and $HR$ images can be denoted as $I_x$ and $J_x$ of a similar training set. For $LR$ and $HR$ images $I_x$ and $J_x$ also similar number of patches can be extracted. The image patch sets for $I_x$ , $J_x$ and $I_y$ , $J_y$ can be denoted as $\{\mathbb{i}_x^r\}_{r=1}^{P_x}, \{\mathbb{j}_x^r\}_{r=1}^{P_x}$ and $\{\mathbb{i}_y^g\}_{g=1}^{P_y}, \{\mathbb{j}_y^g\}_{g=1}^{P_y}$ , respectively. Here, $P_x$ and

$P_y$ rely upon the size of patches and overlap degree between neighboring patches. Here, for different $LR$ and $HR$ images $I_x$ , $J_x$ and $I_y$ , $J_y$ the patches (feature vectors points) can be denoted as $\mathbb{i}_x^r, \mathbb{j}_x^r, \mathbb{i}_y^g$ and $\mathbb{j}_y^g$ respectively. Here, the corresponding low resolution image ($I_y$) and high resolution image ( $J_y$) patches are linked to each other. These low resolution image ($I_y$) patches also preserve inter-patch association with neighboring high resolution image ( $J_y$) patches. By reducing the local reconstruction artifacts in the neighboring low resolution image $I_x$ , we can evaluate weights for every patches in low resolution image $I_y$.To define the reconstruction weights in every patch of $I_y$, the local reconstruction error can be reduced by following equation (3),

$$\varphi^g = \left\| \mathbb{i}_y^g - \sum_{\mathbb{i}_x^r \in P_g} v_{gr} \mathbb{i}_x^r \right\|^2, \quad (3)$$

Where, equation (3) is referred as the squared distance between patch $\mathbb{i}_y{}^g$ and its reconstruction. Here, $v_{gr}$ represents weights. Here, $v_{gr}$ consist of two values which depends on the condition $\mathbb{i}^r{}_x$ demonstrated as,

$$\begin{cases} v_{gr} = 1, & when \sum_{\mathbb{i}_x^r \in P_g} v_{gr} \\ v_{gr} = 0, & when \sum_{\mathbb{i}_x^r \notin P_g} v_{gr} \end{cases}, \quad (4)$$

In equation (3), $\varphi^g$ led to least square problem and to eliminate this error a local patch matrix can be formed for patch $\mathbb{i}_y^g$ which is described as follows,

$$L_g = (\mathbb{i}_y{}^g 1^N - \mathbb{X})^T (\mathbb{i}_y{}^g 1^N - \mathbb{X}), \quad (5)$$

Where, $\mathbb{X}$ denotes the matrix $E \times Q$ and its columns are adjacent to the patch $\mathbb{i}_y{}^g$ and 1 represents a column vector of ones. Furthermore, we form a set of weights of the adjacent frames to build a $Q$- dimensional feature weight vector by rearranging $r$ from every weights $v_{gr}$.Thus, the least squared problem can be solved by the expression (6) as,

$$vg = L - g1.(1NL - g1.1) - 1, \quad (6)$$

Here, to reduce the least squared problem we can also use $L_g v_g = 1$ instead of equation (6) so that the feature weight vectors can be optimized as $\sum_{\mathbb{i}_x^r \in P_g} v_{gr} \mathbb{i}_x^r = 1$. By using equation (3), (4) and (5) for all the image patches in an image $I_y$ , the reconstruction weights can be used to form a feature weight matrix as,

$$\mathbb{V} = [vgr] Py \times Px. \quad (7)$$

### B. CNN based Mapping of Feature Vectors

In this section, we address the deburring effect on the high resolution images and retaining the lowresolution information and by recovering only the missing information in every patch of high resolution image, efficient mapping of feature vectors and quality image reconstruction can be achieved. Thus, here, patch based image reconstruction is presented. Assume an image patch can be denoted as $\mathbb{i}^r = K_e I$ whose size can be considered as $5 \times 5$ placed at the position $e$ and extracted from the input image $I$ of size $P$ using linear optimization factor $K_e$. The HR image patches can be estimated from the LR image patches as $\mathbb{j}_x^r = K_e J_x$ and $\mathbb{i}^r{}_x = K_e I_x$ respectively. After extracting all the $HR$ patch estimates, the $HR$ image can be reconstruct using the averaging method. To average the overlapping reconstructed patches, the averaging technique can be utilized on their overlaps. To select the required overlap size between the neighboring patches a balancing between quality image reconstruction and run-time can be formed. However, only one certain model cannot estimate all the $HR$ image information in all the overlapping patches. Therefore, to acquire only partial overlapping patches, we obtain patches only for central frames which are positioned at the sampling grid. Moreover, $LR$ patches also helps to removes sparse coding high-computational complexity. To represent the relationship between a $LR$ and $HR$ different sparse designs such as $-f_x = \{-1,1\}^{sx}$ and $-f_y = \{-1,1\}^{sy}$ respectively a model can be expressed as,

$$BR(fy|fx) = (A1)^{-1} \exp(\mathbb{h}Ny\, fy + fyN\mathbb{V}yxfx) = \prod_{\mathbb{j}=1} \emptyset\,((\mathbb{h}y,\mathbb{j} + vyxN\,,\mathbb{j}fx)fy,\mathbb{j}), \quad (8)$$

Where, $\mathbb{h}_y \in K^{sy}$ is a vector which represent pattern for $HR$ sparsity and $\mathbb{V}_{yx}$ is a interconnecting matrix to present relationship between patterns of $LR$ and $HR$ sparsity. Here, $\emptyset(\mathbb{z}) = \{1 + \exp(-2\mathbb{z})\}^{-1}$ represents a sigmoid function. Next,

using the sparsity pattern $f_y$ the $HR$ coefficients can be defined as $\delta_y$ and $\delta_x$ for the $LR$ sparsity $f_x$. To determine mapping in high resolution image $f_y$ using $\delta_y$ and $\delta_x$, a model can be defined as follows,

$$\delta_{y,\mathbb{j}} = \begin{cases} \mu_{\mathbb{j}}, & f_{y,\mathbb{j}} = 1 \\ 0, & f_{y,\mathbb{j}} = -1, \forall \mathbb{j} = 1 \dots \dots s_y \end{cases} \quad (9)$$

Where, $\mu$ represents a Gaussian distribution for the $\delta_x$ coefficient. So that $(\mu|\delta_x) \in P(Z_{yx}\ \delta_x\ , \sum yx)$ where $Z_{yx} \in K^{sy \times sx}$ and $\sum yx \in K^{sy \times sy}$. To achieve this coefficients a conditional expectations should be made which is as follows,

$$T\left[\left(\delta_{y,\mathbb{j}} \middle| f_{y,\mathbb{j}}\right) = 1, \delta_x\right] = z_{yx,\mathbb{j}}^{N} \delta_{x,} \qquad \forall \mathbb{j} = 1 \dots \dots s_y. \quad (10)$$

Where, equation (9) and (10) suggests overall mapping between low sparsity $\delta_x$ and high sparsity $\delta_y$ for every sparsity pattern $f_y$. However, $Z_{yx}$ matrix describes all the possible mappings $2^{sy}$.

### C. Quality Reconstruction of HR images from LR images

In this section, the high quality image reconstruction is discussed. First, in our reconstruction technique, the information from two adjacent frames are fused together to form a high resolution image which consists of detailed information of extracted feature vectors. The adjacent frames can be denoted as

$I_x$ and $I_{x+1}$. The $I_{x+1}$ output is up-scaled and fused with the $I_x$ output to join different scale information successively using element wise sum. This method can provide high quality resolution image $I_y$.

Down-scaling and up-scaling are two very essential phenomenon in image scaling process. Downscaling is used to resize the obtained feature vectors as demonstrated in figure 1.This technique can also be used to reduce computational cost of the model. To reduce the high dimensionality of feature vectors downscaling method can be used. This technique can decrease high amount of feature dimensions.

We also use up-scaling process as a deconvolution layer as shown in figure 1 in CAFFE framework. Upscaling is reverse process of down-scaling as presented in figure 1. An upscaling layer can be presented to develop a quality $HR$ image. Both down-scaling and Up-scaling procedures used for the successive frames of a video to get high resolution frames and to form a high resolution video. This techniques can rise the performance of the model.

### D. Sparse Rectified Linear Unit based Adaption

This section describes about the activation functions and the complex relationship between low and high resolution images such as $I_x$ and $I_y$. This technique can enhance accuracy in a high manner. This method can be incorporated in convolutional neural networks (CNN) to further increase the quality of an $HR$ image from the $LR$ image. The $SReLU$ activation function canbe defined as,

$$\mathbb{f}(t_{\mathbb{j}}) = \uparrow (t_{\mathbb{j}}, 0) + \mathbb{h}_{\mathbb{j}} \downarrow (0, t_{\mathbb{j}}) \quad (11)$$

Where, $\mathbb{f}$ is the activation function for the input $t_{\mathbb{j}}$, where $\mathbb{j}$ define channel and $\mathbb{h}_{\mathbb{j}}$ represents the negative coefficient. The value of $\mathbb{h}_{\mathbb{j}}$ can be user-defined to reduce the unused feature points which creates using zero gradient feature vectors. This process enhances accuracy and rapidness of the model. Therefore, our experimental results outperforms other existing state-of-art-models and proves our model is comparatively more superior and stable.

Here, if there is a difference in the reconstructed feature vector points and the output, then adaption method can be applied to make similar number of channels as in the final output. In adaption method two convolutional layers can be utilized of kernel size $3 \times 3$ as shown in figure 1 and a $SReLU$ layer is applied before every convolution layer. In this way, we can achieve high resolution frames ($HR$) from the low resolution frames ($LR$).

## PERFORMANCE EVALUATION

In this section, we discussed about the dataset and performance comparison of proposed model with other existing state-of-art-techniques. Here, we have implemented our model using DASH-SVC dataset [27], [30] to compare the efficiency of our proposed model in contrast to other existing techniques. DASH dataset contains total 4 different high definition videos such as Big Buck Bunny (BBB) [31], Elephants Dream [32], Tears of Steel [33] and Sintel [34]. Our model is an extention of our previous work in the field of convolutional networks for video scaling application [35] [36]. Our model is trained on various DASH-SVC dataset images using CNN technique in CAFFE framework. We have implemented our model on various up-scaling factors such as 2, 3 and 4. The testing outcomes verifies that our proposed model outperforms other state-of-art-algorithms in terms of quality reconstruction and visual appearance. Our model perform parallel computing using CAFFE framework for faster implementation and accurate results. Thus, it requires less computation time for the implementation of our model with less computational complexity. An INTEL (R) core (TM) i5-4460 processor with 64-bit windows 10 OS with 16 GB RAM used for the implementation of our model. It works on the 3.20 GHz CPU. The output of our proposed model is compared with Dynamic Adaptive Streaming over HTTP (DASH) and Scalable Video Coding (SVC) technique.

### Experimental Study

Our proposed model is employed using large High Definition (HD) videos of DASH [30] dataset. In recent time, the demand of high resolution videos extremely enhanced. Therefore, here, we have proposed a robust and efficient Reconstruction Error Minimization Convolution Neural Network Architecture ($RemCNN$) to convert low resolution videos to high resolution videos using upscaling factor. We have implemented our model for various upscaling factors such as 2, 3 and 4 to measure accuracy and efficiency of our model on DASH dataset. All the training and testing experiments are

commenced on the MATLAB 16b environment in configuration with CAFFE framework.

**Comparative Study**

In this section, we demonstrate the comparison of our proposed model and other state-of-art-techniques using DASH dataset which contains 4 videos such as Big Buck Bunny (BBB) [31], Elephants Dream [32], Tears of Steel [33] and Sintel [34]. Here, Big Buck Bunny (BBB) contains total of 14315 frames. Similarly, Elephants Dream contains 15691 frames, Tears of Steel contains 17620 frames and Sintel contains 21312 frames. Here, Big Buck Bunny (BBB) video has dimension of $854 \times 480$.Similarly, Elephants Dream (ED), Tears of Steel (TOS) and Sintel has the dimension of $480 \times 360$.We have tested our model for all four videos using different upscaling factors. Here, all four videos consists of framerate 24 fps. Here, all four video results are discussed using our proposed method in terms of Average Peak Signal to Noise Ratio (PSNR) and Structural Similarly Index Metrics (SSIM) considering different upscaling factors 2, 3 and 4. All four videos are compared with existing DSH-SVC technique in terms of Average PSNR, SSIM and visual appearance. For all videos objective and graphical analysis are discussed. Our experimental outcomes demonstrate the superiority of our proposed method in contrast to existing DASH-SVC results verified in PSNR table 1and SSIM table 2. Our model shows average PSNR value for Big Buck Bunny (BBB) video is $42.761\ dB$ considering upscale 2, $55.575\ dB$ considering upscale 3 and $51.57\ dB$ considering upscale 4.Similarly, Our model shows average PSNR value for Elephants Dream (ED) video is $48.059\ dB$ considering upscale 2, $45.955\ dB$ considering upscale 3 and $38.257\ dB$ considering upscale 4. Our model shows average PSNR value for Tears of Steel (TOS) video is $40.264\ dB$ considering upscale 2, $38.257\ dB$ considering upscale 3 and $36.43\ dB$ considering upscale 4. Our model shows average PSNR value for Sintel video is $48.647\ dB$ considering upscale 2, $45.820\ dB$ considering upscale 3 and $43.083\ dB$ considering upscale 4.Similarly, Our model shows average SSIM value for Big Buck Bunny (BBB) video is $0.959$ considering upscale 2, $0.9978$ considering upscale 3 and $0.994$ considering upscale 4.Similarly, Our model shows average SSIM value for Elephants Dream (ED) video is $0.957$ considering upscale 2, $0.974$ considering upscale 3 and $0.9597$ considering upscale 4. Our model shows average SSIM value for Tears of Steel (TOS) video is $0.9573$ considering upscale 2, $0.93$ considering upscale 3 and $0.8914$ considering upscale 4. Our model shows average SSIM value for Sintel video is $0.991$ considering upscale 2, $0.982$ considering upscale 3 and $0.971$ considering upscale 4.Our model PSNR results considering upscale 2 for all four videos are presented in table 1 and SSIM results in table 2. Similarly, our model PSNR results considering upscale 3 for all four videos are presented in table 3 and SSIM results in table 4. Similarly, our model PSNR results considering upscale 4 for all four videos are presented in table 5 and SSIM results in table 6.

**Table 1.** Comparison of proposed technique with existing DASH-SVC technique considering upscale-2 in terms of PSNR

| PSNR | Existing technique- DASH-SVC | | | | Proposed Technique | | | |
|---|---|---|---|---|---|---|---|---|
| | Y | U | V | AVG | Y | U | V | AVG |
| Big Buck Bunny (BBB) | 35.641 | 40.7263 | 41.727 | 36.46 | 46.591 | 41.642 | 40.050 | 42.7616 |
| Elephants Dream (ED) | 37.759 | 55.626 | 54.845 | 39.024 | 40.041 | 52.840 | 51.296 | 48.059 |
| Tears of Steel(TOS) | 33.4165 | 40.7362 | 40.3502 | 34.5798 | 33.803 | 44.396 | 43.672 | 40.624 |
| Sintel | - | - | - | - | 42.269 | 52.167 | 51.504 | 48.647 |

**Table 2.** Comparison of proposed technique with existing DASH-SVC technique considering upscale-2 in terms of SSIM

| SSIM | Existing technique - DASH-SVC | | | | Proposed Technique | | | |
|---|---|---|---|---|---|---|---|---|
| | Y | U | V | AVG | Y | U | V | AVG |
| Big Buck Bunny (BBB) | 0.9017 | 0.9164 | 0.933 | 0.9094 | 0.9829 | 0.9524 | 0.941 | 0.9590 |
| Elephants Dream (ED) | 0.898 | 0.969 | 0.967 | 0.922 | 0.9370 | 0.9671 | 0.9680 | 0.957 |
| Tears of Steel(TOS) | 0.8693 | 0.9206 | 0.9142 | 0.8853 | 0.9370 | 0.9671 | 0.9680 | 0.9573 |
| Sintel | - | - | - | - | 0.9842 | 0.994 | 0.993 | 0.991 |

**Table 3.** Comparison of proposed technique with existing DASH-SVC technique considering upscale-3 in terms of PSNR

| PSNR | Existing technique - DASH-SVC | | | | Proposed Technique | | | |
|---|---|---|---|---|---|---|---|---|
| | Y | U | V | AVG | Y | U | V | AVG |
| Big Buck Bunny (BBB) | 40.320 | 44.288 | 45.143 | 41.037 | 48.590 | 58.274 | 59.862 | 55.575 |
| Elephants Dream (ED) | 42.900 | 59.351 | 58.872 | 43.930 | 36.687 | 51.637 | 49.5422 | 45.955 |
| Tears of Steel(TOS) | 36.267 | 41.4014 | 41.998 | 37.236 | 30.6833 | 42.4626 | 41.6275 | 38.2578 |
| Sintel | - | - | - | - | 39.309 | 49.405 | 48.746 | 45.820 |

**Table 4.** Comparison of proposed technique with existing DASH-SVC technique considering upscale-3 in terms of SSIM

| SSIM | Existing technique - DASH-SVC | | | | Proposed Technique | | | |
|---|---|---|---|---|---|---|---|---|
| | Y | U | V | AVG | Y | U | V | AVG |
| Big Buck Bunny(BBB) | 0.9555 | 0.9547 | 0.9636 | 0.9567 | 0.9957 | 0.9988 | 0.9990 | 0.9978 |
| Elephants Dream (ED) | 0.9612 | 0.9853 | 0.9836 | 0.9689 | 0.9421 | 0.9922 | 0.9878 | 0.9740 |
| Tears of Steel(TOS) | 0.9160 | 0.9441 | 0.9411 | 0.9248 | 0.8974 | 0.9467 | 0.9460 | 0.9300 |
| Sintel | - | - | - | - | 0.969 | 0.990 | 0.988 | 0.982 |

**Table 5.** Comparison of proposed technique with existing DASH-SVC technique considering upscale-4 in terms of PSNR

| PSNR | Existing technique - DASH-SVC | | | | Proposed Technique | | | |
|---|---|---|---|---|---|---|---|---|
| | Y | U | V | AVG | Y | U | V | AVG |
| Big Buck Bunny (BBB) | 40.751 | 43.941 | 44.867 | 41.369 | 54.768 | 51.295 | 48.646 | 51.570 |
| Elephants Dream (ED) | 43.262 | 58.898 | 58.352 | 44.35 | 34.517 | 48.612 | 46.543 | 43.224 |
| Tears of Steel(TOS) | 16.1485 | 32.4062 | 31.8213 | 26.792 | 28.910 | 40.487 | 39.891 | 36.430 |
| Sintel | - | - | - | - | 36.074 | 46.877 | 46.297 | 43.083 |

**Table 6.** Comparison of proposed technique with existing DASH-SVC technique considering upscale-4 in terms of SSIM

| SSIM | Existing technique –DASH-SVC | | | | Proposed Technique | | | |
|---|---|---|---|---|---|---|---|---|
| | Y | U | V | AVG | Y | U | V | AVG |
| Big Buck Bunny (BBB) | 0.949 | 0.952 | 0.964 | 0.952 | 0.9978 | 0.995 | 0.991 | 0.994 |
| Elephants Dream (ED) | 0.9500 | 0.9855 | 0.9834 | 0.9615 | 0.9106 | 0.9873 | 0.9811 | 0.9597 |
| Tears of Steel(TOS) | 0.628 | 0.9288 | 0.9203 | 0.7269 | 0.8199 | 0.9285 | 0.9258 | 0.8914 |
| Sintel | - | - | - | - | 0.9489 | 0.984 | 0.9817 | 0.971 |

**Visual Representation**

In this section, we have presented our results to demonstrated high visual quality of images using our proposed method in comparison with DASH-SVC dataset. Here, we have used $347^{th}$ frame for Big Black Bunny, $554^{th}$ frame for Sintel video, $472^{nd}$ frame for Tears of Steel and $4124^{th}$ frame for Elephants Dream to demonstrate our visual quality. The Big Buck Bunny (BBB) video has dimension of $854 \times 480$.Similarly, Elephants Dream (ED), Tears of Steel (TOS) and Sintel has the dimension of $480 \times 360$.Our PSNR and SSIM results outperforms all the other state-ofart-techniques considering upscale 2, 3 and 4. From our experimental outcomes it is clearly observed that our reconstruct frame has better reconstruction quality than existing DASH-SVC technique which is shown in table 7 where table 7(a) represents Big Buck Bunny (BBB), table 7(b) represents Tears of Steel (TOS), table 7(c) represents Elephant Dream (ED) and table 7(d) represents Sintel.

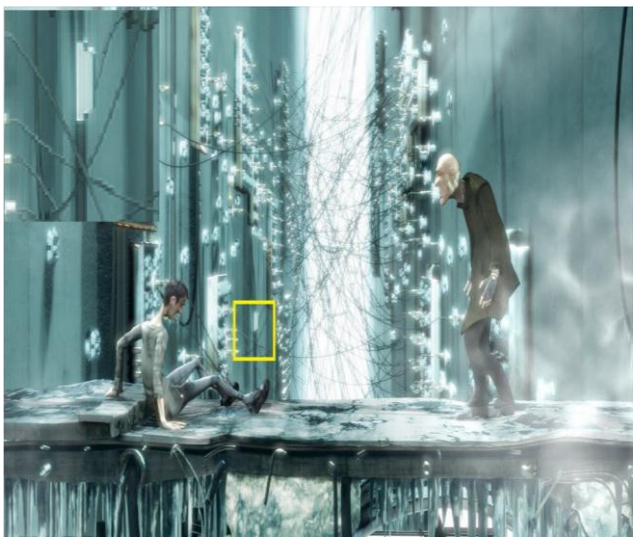
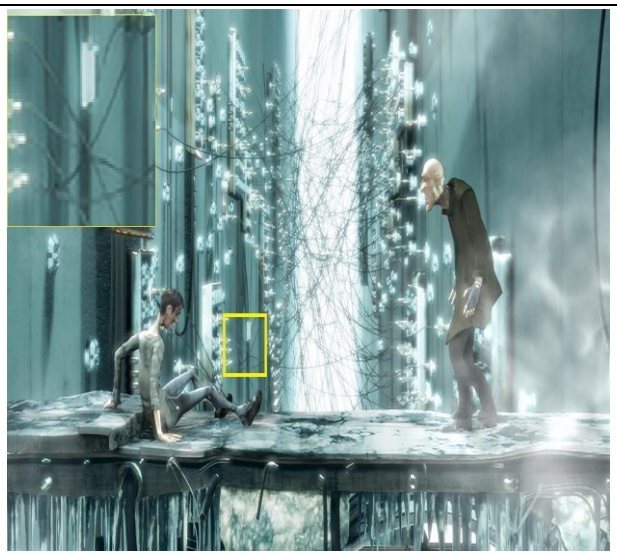
**Table 7.** Visual representation of all four videos using our proposed technique

<table>
<tr>
<td> 

7(a) Ground Truth</td>
<td> 

7(a) Proposed System</td>
</tr>
<tr>
<td> 

7(b) Ground Truth</td>
<td> 

7(b) Proposed System</td>
</tr>
<tr>
<td> 
7(c) Ground Truth</td>
<td> 
7(c) Proposed System</td>
</tr>
</table>

7(d) Ground Truth

7(d) Proposed System

**Graphical Representation**

In this section, we have demonstrated the graphical representation of our proposed model in comparison with DASH-SVC technique considering upscale 2, 3 and 4 for all videos of DASH dataset in terms of PSNR and SSIM. Here, figure 2 demonstrates average PSNR comparison with DASH-SVC technique considering upscale 2 for all 4 videos as Big Buck Bunny (BBB), Elephants Dream (ED), Tears of Steel (TOS). Similarly, figure 3 demonstrates average PSNR comparison with DASH-SVC technique considering upscale 3 and figure 4 shows PSNR comparison for upscale 4. Similarly, here, figure 5 demonstrates average SSIM comparison with DASH-SVC technique considering upscale 2 for all 3 videos as Big Buck Bunny (BBB), Elephants Dream (ED), Tears of Steel (TOS) whereas figure 6 demonstrates average SSIM comparison with DASH-SVC technique considering upscale 3 and figure 7 demonstrates average SSIM comparison with DASH-SVC technique considering upscale 4 for all 3 videos. PSNR considering upscale-2 using our proposed method for BBB video is 42.761 $dB$, for upscale-3 is 55.575 $dB$ and upscale 4 is 51.57 $dB$. Similarly, SSIM considering upscale 2 for BBB video is 0.9978, for upscale-3 is 0.974 and upscale 4 is 0.93. PSNR considering upscale-2 using our proposed method for TOS video is 40.62 $dB$, for upscale-3 is 38.25 $dB$ and upscale 4 is 36.43 $dB$. Similarly, SSIM considering upscale 2 for TOS video is 0.9573, for upscale-3 is 0.930 and upscale 4 is 0.891. PSNR considering upscale-2 using our proposed method for ED video is 45.955 $dB$, for upscale-3 is 43.224 $dB$ and upscale 4 is 36.43 $dB$. Similarly, SSIM considering upscale 2 for ED video is 0.957, for upscale-3 is 0.974 and upscale 4 is 0.9597.This results demonstrates that our proposed method outperforms existing DASH-SVC technique

**CONCLUSION**

The significance of our proposed model and design complexities in implementing a robust and efficient Reconstruction Error Minimization Convolution Neural Network Architecture ($RemCNN$) is presented. Drawbacks of existing video scaling techniques are discussed in the literature. Our model efficiently convert low resolution videos into high resolution to offer easy accessibility to the subscribers. The CNN architecture make use of GPU parallel computing using CAFFE framework to offer faster and easy training on large datasets. Our proposed technique is trained using bulky DASH dataset and compared with DASH-SVC technique. Our experimental outcomes demonstrates that our proposed technique shows superior results in terms of PSNR, SSIM and Visual appearance. The experimental study verifies that our average PSNR value for upscaling factor 2 is 42.7616, for upscaling factor 3 is 55.57 and for upscaling factor 4 is 51.57 for Big Black Buck video. Similarly, we have evaluated PSNR for Sintel, Elephant Dream and Tears of Steel considering upscale-2, 3 and 4 and compared with existing DASH-SVC technique. Similarly, the experimental study verifies that our average SSIM value for upscaling factor 2 is 0.959, for upscaling factor 3 is 0.9978 and for upscaling factor 4 is 0.994 for Big Black Buck video. Similarly, we have evaluated average SSIM for Sintel, Elephant Dream and Tears of Steel considering upscale-2, 3 and 4 and compared with existing DASH-SVC technique. The reconstruction quality of all the video frames are visually very high which is verified by the experimental results. In future, this technique can be used in real time for scaling of videos on HD TV, HD mobiles and laptops.

**REFERENCES**

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

[2] M. W. Thornton, P. M. Atkinson, and D. a. Holland, 2006 "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*, 27(3):473–491,

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

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

[5] J. Zhang, Y. Cao, Z. J. Zha, Z. Zheng, C. W. Chen, and Z. Wang, 2016 "A unified scheme for super-resolution and depth estimation from asymmetric stereoscopic video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 479–493.

[6] B. C. Song, S.-C. Jeong, and Y. Choi, 2011 "Video super-resolution algorithm using bidirectional overlapped block motion compensation and onthe- fly dictionary training," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 3, pp. 274–285.

[7] Pawar Ashwini Dilip,K Rameshbabu. November 2014, "Bilinear Interpolation Image Scaling Processor for VLSI Architecure". International Journal of Reconfigurable and Embedded Systems (IJRES). Vol.3, No.3, pp. 104~113.ISSN: 2089-4864.

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

[9] A. Punnappurath; T. M. Nimisha; A. N. Rajagopalan, "Multi-image blind superresolution of 3D scenes," in *IEEE Transactions on Image Processing* , vol.PP, no.99, pp.1-1

[10] M. Sharma, S. Chaudhury and B. Lall, 2017 "Deep learning based frameworks for image super-resolution and noise-resilient super-resolution," *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, pp. 744-751.

[11] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, 2013, " High-frequency restoration using deep belief nets for super-resolution"In Signal-Image Technology & Internet-Based Systems (SITIS), International Conference on. IEEE,

[12] Y. Wang, L. Wang, H. Wang, and P. Li, 2016 "End-to-end image superresolution via deep and shallow convolutional networks," arXiv preprint arXiv:1607.07680.

[13] A Krizhevsky, I Sutskever, and G E Hinton, 2012, "Imagenet classification with deep convolutional neural networks," in *NIPS*, pp. 1097–1105.

[14] G. Suryanarayana, Ravindra Dhuli, November 2015, "Sparse Representation Based SuperResolution Algorithm using Wavelet Domain Interpolation and Nonlocal Means", TELKOMNIKA Indonesian Journal of Electrical Engineering Vol. 16, No. 2, , pp. 296 ~ 302 DOI: 10.11591/telkomnika.v16i2.8816.

[15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, 2014 , " Learning a deep convolutional network for image super-resolution", Computer Vision–ECCV 2014. Springer.

[16] K. Zeng, J. Yu, R. Wang, C. Li, and D. Tao, 2015 " Coupled deep autoencoder for single image super-resolution", IEEE Transactions on Cybernetics, (99):1–11,

[17] Z. Wang, Y. Yang, Z. Wang, S. Chang, W. Han, J. Yang, and T. Huang. 2015, Self-tuned deep super resolution. In 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1–8.

[18] W. Shi, J. caballero, F. Huszr, J. totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, 2016, "Real-time single image and video superresolution using an efficient sub-pixel convolutional neural network," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883.

[19] Jyoti Mahajan, Kashyap Dhruve Devanand, May 2011 "REBEE-Reusability Based Effort Estimation Technique using Dynamic Neural Network," Global Journal of Computer Science and Technology.

[20] J. Y. Cheong and I. K. Park, Aug. 2017 "Deep CNN-Based Super-Resolution Using External and Internal Examples," in *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1252-1256.

[21] Y. Li; D. Liu; H. Li; L. Li; F. Wu; H. Zhang; H. Yang, "Convolutional Neural NetworkBased Block Up-sampling for Intra Frame Coding," in *IEEE Transactions on Circuits and Systems for Video Technology* , vol.PP, no.99, pp.1-1

[22] J. Lee and I. C. Park, April 2017, "High-Performance Low-Area Video Up-Scaling Architecture for 4-K UHD Video," in *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 4, pp. 437-441.

[23] C. M. Lee and H. F. Yeh, 2017"Adaptive band-based super-resolution reconstruction of video sequence," *2017 International Conference on Applied System Innovation (ICASI)*, Sapporo, pp. 288-291.

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

[25] J. Deng, W. Dong, R. Socher, and L. Li, 2009 "A large-scale hierarchical image database" Proceedings of the IEEE Conference on Computer Vision and

Pattern Recognition, pp. 248–255.

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

[27] DASH Dataset at ITEC/Alpen-Adria-Universität Klagenfurt, http://www-itec.uniklu.ac.at/dash/?page_id=207

[28] Ouyang, W., Wang, X, (2013) "Joint deep learning for pedestrian detection": IEEE International Conference on Computer Vision. pp. 2056–2063

[29] Sun, Y., Chen, Y., Wang, X., Tang, X,2014, "Deep learning face representation by joint identification-verification", Advances in Neural Information Processing Systems. pp. 1988–1996

[30] S. Lederer, C. Müller, and C. Timmerer, 2012, "Dynamic Adaptive Streaming over HTTP Dataset", Proceedings of the 3rd Multimedia Systems Conference, MMSys '12, pages 89–94, New York, NY, USA,. ACM.

[31] Blender Foundation. Big Buck Bunny. http://bigbuckbunny.org/.

[32] Blender Foundation. Elephants Dream. http://elephantsdream.org/.

[33] Blender Foundation. Tears of Steel, Mango Open Movie Project. http://tearsofsteel.org.

[34] Blender Foundation. Sintel, Durian Open Movie Project. http://sintel.org/.

[35] Safinaz S and Ravi Kumar A.V, 2017, "Real-Time Video Scaling Based on Convolution Neural Network Architecture", ICTACT Journal on Image and Video Processing, volume: 08, issue: 01. .pp1533-1542.

[36] Safinaz S and Ravi Kumar A.V, 2017, "An Adaptive Scheme to Achieve Fine Grained Video Scaling", Indonesian Journal of Electrical Engineering and Computer Science Vol. 8, No. 1.pp43-58.