A Survey on Image and Video Super Resolution with Deep Learning Models

Athirasree Das
M-Tech Scholar
Dept. of Computer Science and Engg
College of Engineering ,Kidangoor
Kerala, India

Dr.K.S Angel Viji
Associate Professor
Dept. of Computer Science and Engg
College of Engineering ,Kidangoor
Kerala, India

Linda Sebastian
Assistant Professor
Dept. of Computer Science and Engg
College of Engineering ,Kidangoor
Kerala, India

***Abstract* - Super Resolution is the process of creating high resolution images from low resolution images with minimum reduction in image quality. There are Single Image Super Resolution (SISR) and Video Super Resolution (VSR) methods. A Single Image Super Resolution goal is to change low resolution image to high resolution image. A Video Super Resolution comes from image super resolution and goal is to restore low resolution videos to high resolution videos. Convolutional Neural Networks (CNN) are the special type of deep neural networks. There are many deep learning methods for images and video super resolution. This is the survey on deep learning, CNN models for image and video super resolution. In video Super Resolution, different frames of the video are temporally connected. It means that position of different objects change from frame to frame. It is very essential to align objects. For aligning frames in video, there are two common techniques MEMC (motion estimation and compensation) methods and DC (deformable convolution) methods. Now a days Deep learning techniques have been used for super resolution and a stack of CNN have been used for SISR and VSR. In this survey we studied about convolutional neural network based methods such as SRCNN, DDAN, LapSRN, SICNN and VSRnet. We have analyzed and summarized SR deep learning models.**

Keywords* - *single image super resolution; video super resolution; convolutional neural network; motion estimation and compensation; deformable convolution

I. INTRODUCTION

Super Resolution (SR) is the process of lower resolution input image up scaled to higher resolution output image. Deep learning has made great improvements in image super resolution and video super resolution. There are Single Image Super Resolution (SISR) and Video Super Resolution(VSR). A Single Image Super Resolution goal is to upscale low resolution image to high resolution image. Based on the number of input images, image super resolution can be divided into single image super resolution (SISR) and multi-image super-resolution (MISR). SISR are mainly classified into interpolation-based methods, reconstruction-based methods and learning-based methods. Video Super Resolution is the combinations of SISR and goal is to upscale low resolution videos to high resolution videos. Super Resolution have many applications in medical image systems, satellite imaging, Astronomical imaging, Bio metric information identification etc..

Single image super-resolution and video super resolution that combines with deep convolutional neural network (CNNs) models are reviewed in this survey. Conv olutional neural network (CNN) are the special type of deep neural network. There are many deep learning methods for image and video super resolution such as Single Image Convolutional Neural Network (SICNN), Deep Dual Attention Network (DDAN), Video Super Resolution network (VSRnet), Laplacian Pyramid Super-Resolution Network (LapSRN) and Super-Resolution Convolutional Neural Network (SRCNN).

The main difference between video super-resolution and image super-resolution is in the processing of inter-frame information. In video Super Resolution, videos and different frames of the video are temporally connected. It means that position of different objects change from frame to frame. However the movement of objects becomes slow if different frames are not align properly. It is very essential to align objects. For aligning frames in video, there are two common techniques motion estimation and compensation methods (MEMC) and deformable convolution (DC) methods.

The motion estimation is the process to extract inter-frame information, and motion compensation is the process to perform the joining operation between frames according to inter-frame information and helps to make one frame align with another frame. MEMC is the common method used for solving alignment problem in the video super resolution. Deformable convolution used for feature alignment where target frame concatenating with the neighboring frames via additional convolutional layers.

Now a days videos are very helpful to exchangee the information and so high resolution are very important.

Deep learning techniques have been used for super resolution and a stack of CNN have been used for SISR and VSR. Peak signal-to-noise ratio (PSNR) and Structural similarity index (SSIM) are used as evaluation metrics to compare with image and video super resolution models.

PSNR is defined via the maximum pixel value and the mean squared error (MSE) between images. SSIM is proposed for measuring the structural similarity between images.

II. LITERATURE SURVEY

A. Deep Learning for Super Resolution

Deep learning (DL) is a branch of machine learning that aims to learning the hierarchical representations of data. Representation learning attempts to automatically learn good features and representations which helps to increase the complexity and abstractions. Single image super-resolution (SISR) has a challenging ill-posed problem. The drawbacks of SISR are clear with deep learning advanced techniques which helps to better performance. Metrics and Loss functions are used to measure the performance of deep learning models. So SR methods have achieved efficient improvements, both quantitatively and qualitatively by better representative metrics. There are different types of frame work design used in deep learning CNN network architecture such as Pre up-sampling, Post up- sampling, Progressive up-sampling, Iterative up and down Sampling. The Pre up-sampling SR framework is used to learn an end-to-end mapping from LR images to HR images. The HR images are up sampled from LR images with the same dimension using traditional methods like bicubic interpolation.

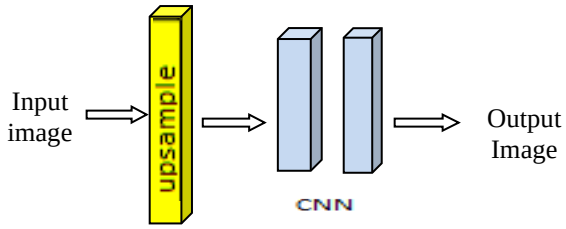

Fig 1: Pre-upsampling SR

In Post up- sampling method the lower resolution images passed through CNN's then up sampling is done on the last layer using the learnable layer. Here feature extraction is performed in the lower dimensional space, thus reducing the computational complexity. The model can be trained from end to end, using the learnable up sampling layer.

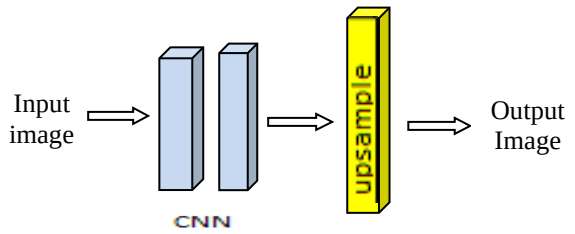

Fig 2 : Post-upsampling SR

In Progressively up- sampling method with reducing the computational complexity and obtain the better performance. Laplacian pyramid SR network use this model. The use of CNNs in this model progressively reconstruct high resolution image with scale factors at each steps. It helps to reduce the difficulty to improve final performance.

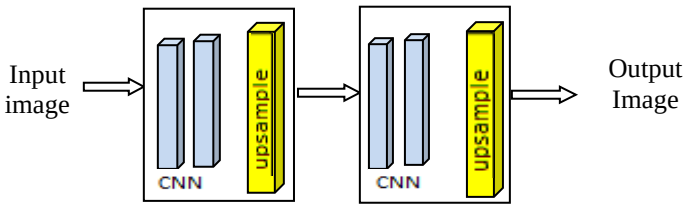

Fig 3: Progressively-upsampling SR

The framework, Iterative up and down Sampling SR is a popular model architecture using several hourglass structures in series. This method provides higher-quality reconstruction results because it maintains deep relation between the LR-HR image pairs.

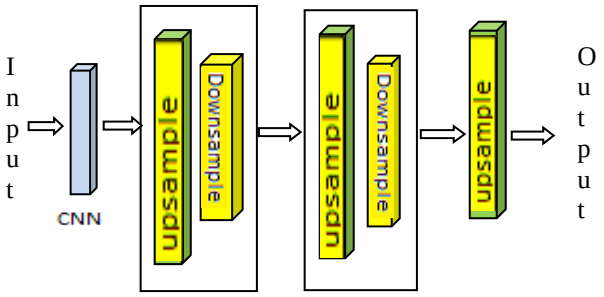

Fig 4 : Iterative up-and-down Sampling SR

B. Convolutional neural network for super resolution

Convolutional neural network is a special kind of multi-layer neural network applied in images. CNN for super resolution helps for efficient training implementation, easy to access data, for training larger models with faster and good quality. The CNN based video super-resolution methods mainly includes feature extraction and representation module, Non-linear mapping module and reconstruction module. These modules are helpful for better performance. Feature extraction and representation is the operation that extracts features and useful representation from low resolution images for better output. Non-linear mapping is the operation that feature maps non linearly at each high-dimensional vector of image onto another high-dimensional vector of output image. Reconstruction is the last stage that combines the predictions to estimates HR images.

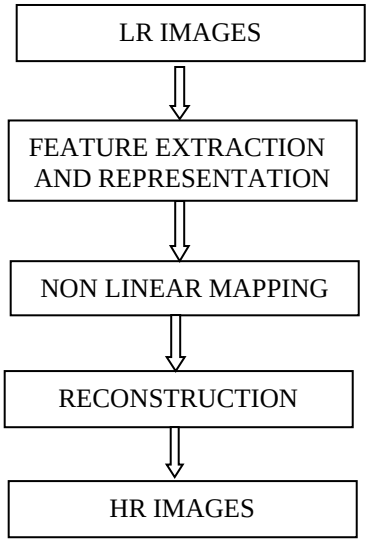

Fig 5 : CNN based video super-resolution methods

C. *Single and Multi-Image Super-Resolution*

A Single Image Super Resolution is to change low resolution input image to high resolution output image. Based on the number of input images, super resolution can be divided into single image super resolution (SISR) and multi-image super-resolution (MISR). SISR are mainly classified into interpolation-based methods, reconstruction-based methods and learning-based methods. Interpolation-based methods are very fast and straightforward but it has low accuracy problem. Reconstruction-based SR methods are very time-consuming. Learning based methods are fast computation and outstanding performance. Deep leaning based SISR algorithms are demonstrated with reconstruction-based and learning based methods. There are three types of multi-frame methods: Interpolation methods, Frequency-domain methods and Regularizationion methods. SRCNN is an example of SISR and so SRCNN has convolutional layers which has the advantage that the input images can be of any size. VSRnet is an example of MISR and so VSRnet uses motion compensated consecutive frames as input to a CNN.

D. *Image and Video super resolution models*

The goal of single image super resolution models is to make a low resolution image to high resolution image. Vvideos are the sequences of moving visual images and sound with frames. Many video super resolution methods are there based on convolutional neural network (CNN) such as DDANand VSRnet. Single image super resolution methods are there based on convolutional neural network (CNN) such as SRCNN, LapSRN and SICNN. The CNN based video super-resolution methods mainly includes alignment, feature extraction and fusion and reconstruction. In SISR and VSR quality is mainly evaluated by calculating matrices peak signal-noise ratio (PSNR) and structural similarity index (SSIM). These indices measure the difference in pixels and the similarity of the structures between the two images. In 2016, Armin Kappeler et.al [11] presented a video super resolution method with convolutional neural networks (VSRnet). In 2020 Feng Li et.al [1], proposed a deep dual attention network (DDAN), including a motion compensation network (MCNet) and a SR reconstruction network (ReconNet). In 2020 Jing Liu et.al [3] proposed a single image convolutional neural network (SICNN). The most popular data sets for testing with Vid4, Myanmar, Set 5 and Set 14.

E. *Methods with Alignment*

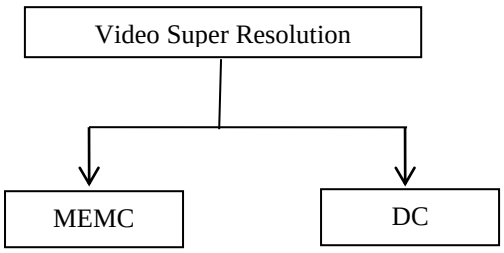

Fig 6 : Methods of alignment for VSR

In video super resolution, method of alignment greatly influences performance that indicates the use of information between frames arrange in proper and adequate manner. It helps to increases the effects of super resolution. Frames are aligning with the method of alignment and the techniques are Motion estimation and motion compensation (MEMC) or Deformable Convolution (DC). The process of motion estimation is to extract inter-frame motion information, and motion compensation is the process of performing the warping operation between frames and helps to align one frame with another. Some of the deep learning based methods followed the alignment techniques. For example, VSR net used motion estimation and motion compensation for improving the video resolution performance.

F. *Image Quality Assessment*

Image quality refers to visual attributes of images and focuses on the perceptual assessments of viewers. Image quality assessment (IQA) methods include subjective and objective types. Subjective methods based on human's perception and objective methods based on computational methods. IQA methods are classified into three types which are full - reference methods, reduced - reference methods and no-reference methods. Full-reference methods for performing assessment using reference images. Reduced - reference methods based on comparisons of extracted features. No-reference methods without any reference images.

Loss functions are used to measure the difference between input image and output image. It is an error

calculation function in a supervised learning model. There are pixel loss, content loss, texture loss and total variation loss. Pixel loss is a simplest loss function of pixels of output image as it compared with input image. The content loss function comparing the output image with input image with features of image and it is based on perceptual quality. Texture loss function is comparing the texture, color, contrast of output image with input image. Total variation loss function is used to calculate the noise generated in output images.

There are two metrics peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) for measuring the performance and most widely used evaluation criteria for SR models. PSNR is one of the most popular quality measurement for super resolution. PSNR is defined the maximum pixel value and the mean squared error (MSE) between images. Structural similarity index is used for measuring the structural similarity between images in terms of luminance, contrast and structures.

G. Super-Resolution Convolutional Neural Network (SRCNN)

Chao Dong et.al [9] proposed that SRCNN method is used for single image super resolution which is a classical problem in computer vision. SRCNN has several appealing properties. It provides best accuracy result as compared with another SR models. It achieves fast speed in various applications. It has restoration quality. The network explains an end-to-end mapping between low and high resolution images. There are three parts in SRCNN network such as patch extraction and representation, nonlinear mapping and reconstruction. Network patch extraction is the process that extract patches from low resolution image. Non-linear mapping is the process of non linearly maps at high dimensional vector into another high-dimensional vector. Reconstruction is the process of aggregation of high resolution patch-wise representations to create the ultimate high resolution image. SRCNN has achieved good performance and maintains a high and competitive speed as compared with other models. The advantages are simplicity and robustness.

H. Laplacian Pyramid Super-Resolution Network (LapSRN)

Wenming Yang et.al [4] says that Laplacean Pyramid Super-Resolution Network (LapsRN) is an advanced super resolution model that resolves low-resolution images with Laplacian pyramid framework. This method is fast and achieves superior performance. It has two branches of feature extraction and image reconstruction. In feature extraction stage that removes the noise from low resolution input images and extract the useful representation for better output. Image reconstruction up samples the low resolution images and takes the features of input image for reconstructing the output image. LapSRN architecture have convolutional layer, transposed convolutions for up sampling and element wise addition operators. These models are good choices for large scale factors which can achieve the best results. It reduces the computational complexity. LapSRN architecture includes convolutional layer like pyramid frame work and are represented in Red arrows, blue arrows and green arrows. Red arrows indicate the convolutional layer. Blue arrows indicate transposed convolutions for up sampling. Green arrows denote element wise addition operators. LapSRN used for fast and accurate image super-resolution and achieves superior performance in terms of run-time and image quality.

I. Video Super Resolution network (VSRnet)

Armin Kappeler et.al[11] suggested the Video SR network (VSRnet), to single frame and video SR algorithms. The information between the frames in video which greatly influence the performance on super resolution. Proper and adequate use of such information is helpful to increases the effects of super resolution. Video frames are aligned with the method of alignment and the techniques are called motion estimation and motion compensation.

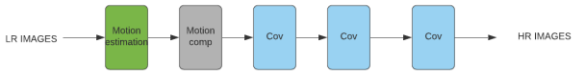

Fig 7 : Model of VSR network

VSRnet mainly consists of motion estimation and compensation techniques and three convolutional layers. Convolutional layer have a rectified linear unit (ReLU) except for the last layer. The main difference between VSRnet and SRCNN methods is the case of number of input frames. SRCNN has a single image as input, while VSRnet uses multiple images as input so which are called compensated frames. VSRNet also use Filter Symmetry Enforcement (FSE) Mechanism and Adaptive Motion Compensation Mechanism. Mechanism are used to accelerate training and reduce the impact of unreliable compensated frame for improving video super resolution performance.

J. Single Image Convolutional Neural Network (SICNN)

Jing Liu et.al [3] proposed that single image convolutional neural network (SICNN) method not requiring an external training data set for image SR reconstructing. Super resolution (SR) reconstruction is the process of restoring low resolution images to high resolution images. Super resolution reconstruction process can retain the image internal features clearly. There are interpolation based methods and rreconstruction based methods in SICNN. Interpolation based method helps to estimate the missing pixels in the high resolution image using neighborhood pixels in the input LR images. The complexity of output image by interpolation method is low. Learning based SR methods helps to learn the mapping relationship between

input and output images. Input LR image and output HR image pairs are used as external training data set and estimate its HR image from the test LR image using this mapping relationship. Learning based SR methods can be divided into four sub types such as sparse coding based, regression based, neighbor embedding based and deep learning based methods. Reconstruction based SR methods are used to reconstruct the HR image. The advantage of SICNN model is that it does not require external data sets. It works with internal data sets and extract the useful representation from input images and remove the noise and reconstructs its HR image by using capturing features. SICNN contains of 8 convolutional layers and 8 rectified linear units.

K. Deep Dual Attention Network (DDAN)

Feng Li et.al [1] proposed a deep dual attention network (DDAN), including a motion compensation network module (MCNet) and a SR reconstruction network module (ReconNet). It contains full advantages of spatial and temporal dependencies and it helps to learn meaningful information for accurate video super resolution. In motion compensation network module explains multi level optical flow representation between adjacent frames in a pyramid fashion. It is a motion compensation strategy between two frames and it takes the center frame and neighboring frame as input to produce motion compensated neighboring frame and it is given as input to reconstruction network to generate HR frames. Rreconstruction process in the networks is a method that integrates dual attention mechanisms along the channel and spatial dimensions to emphasize meaningful features for retrieving high frequency details of image. Vvideo quality is mainly evaluated by calculating peak signal noise ratio and structural similarity index. The network architecture of deep dual attention network (DDAN) method with spatio - temporal video SR. It contains a motion compensation network (MCNet) helps to synthesize the motion information across the neighboring frames and a SR reconstruction network (ReconNet) to generate accurate output frames. For comparing with other SR models with DDAN used different data sets to prove the robustness.

L. Comparison Table of SR models

Method	year	scale	data set	PSNR	SSIM
DDAN[1]	2020	2	Vid4	33.65	0.9517
	2020	2	Myanmar	40.84	0.9775
VSRnet[11]	2016	2	Vid4	31.30	0.9278
	2016	2	Myanmar	38.48	0.9679
SICNN[3]	2020	2	Set 5	33.76	0.9017
	2020	2	Set 14	33.23	0.8524
LapSRN[12]	2017	2	Set 5	37.52	0.959
	2017	2	Set 14	33.08	0.913
SRCNN[8]	2014	2	Set 5	36.66	0.9542
	2014	2	Set 14	32.45	0.9067

TABLE 1. THE BEST RESULT OF EACH METHOD

The performance of SR methods with scale factor 2 is analyzed quantitatively in terms of both PSNR and SSIM. The top 2 methods in the table DDAN, VSRnet are the VSR methods and which is compared with datasets Vid4, Myanmar. SICNN, LapSRN, SRCNN are the SISR methods and which are compared with datasets set5 and set 14. By the analysis The most popular data sets for testing are Vid4, Myanmar, Set 5 and Set 14. LapSRN outperforms consistently and achieved better value of PSNR and SSIM with scaling factor 2 and which is higher value best approach in SISR. DDAN obtains superior SR performance on public data sets for VSR.

III. CONCLUSION

The literature review presents the facts that there are large number of deep learning video and image resolution models. Deep learning is useful in the classical computer vision problem of super-resolution and can achieve good quality and speed. In this survey we have studied CNN based methods such as SRCNN, DDAN, VSRnet, LapSRN and SICNN. Video quality is mainly evaluated by calculating peak signal-noise ratio (PSNR) and structural similarity index (SSIM). The newest VSR models are DDAN, SICNN. SISR models are VSRnet, LapSRN, SRCNN. This is likely due to development of deeper and more complex network structures, and there is a improvement in computational ability of hardware. However, the performance of the latest methods still needs to be improved in the future. Ttherefore

we can conclude that the LapSRN has achieved superior performance in Image super resolution and DDAN in Video super resolution.

ACKNOWLEDGMENT

The author would like to thank Dr. Ojus Thomas Lee (HOD & Associate Prof. Dept. of CSE, CE Kidangoor), Mrs. Rekha K.S (Assistant Prof. Dept. of CSE, CE Kidangoor) for their valuable suggestions. The author is extremely grateful to guidance in improving the paper's quality by Dr. K.S Angel Viji (Associate Prof. Dept. of CSE, CE Kidangoor) and Mrs. Linda Sebastian (Assistant Prof. Dept. of CSE, CE Kidangoor).

REFERENCES

[1] F. Li, H. Bai and Y. Zhao, "Learning a Deep Dual Attention Network for Video Super Resolution," in IEEE Transactions on Image Processing, vol. 29, pp. 4474-4488, 2020, doi: 10.1109/TIP.2020.2972118.

[2] Z. Wang, J. Chen and S. C. H. Hoi, "Deep Learning for Image super Image : A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi:10.1109/TPAMI.2020.2982166.

[3] J. Liu, Y. Xue, S. Zhao, S. Li and X. Zhang, "A Convolutional Neural Network for Image Super-Resolution Using Internal Data set," in IEEE Access, vol. 8, pp. 201055-201070, 2020, doi: 10.1109/ACCESS.2020.3036155.

[4] W. Yang, X. Zhang, Y. Tian, W. Wang, J. Xue and Q. Liao, "Deep Learning for Single Image Super-Resolution: A Brief Review," in IEEE Transactions on Multimedia, vol. 21, no. 12, pp. 3106-3121, Dec. 2019, doi: 10.1109/TMM.2019.2919431.

[5] D. Liu et al., "Learning Temporal Dynamics for Video Super-Resolution: A Deep Learning Approach," in IEEE Transactions on Image Processing, vol. 27, no. 7, pp. 3432-3445, July 2018, doi: 10.1109/TIP.2018.2820807.

[6] https://www.research gate.net/publication/318009713 Super Resolution via Deep Learning

[7] W. Shi et al., "Real-Time Single Image and video super resolution Using an Efficient Sub-Pixel Convolutional Neural Network," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 1874-1883, doi: 10.1109/CVPR.2016.207.

[8] C. Dong, C. C. Loy, K. He and X. Tang, "Image Super-Resolution using Deep Convolutional Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295-307, 1 Feb. 2016, doi: 10.1109/TPAMI.2015.2439281.

[9] Chao Dong1, Chen Change Loy1, Kaiming He2, and Xiaoou Tang / publication/319770198 Learning a Deep Convolutional Network for Image Super-Resolution

[10] D. Glasner, S. Bagon and M. Irani, "Super-resolution from a single image," 2009 IEEE 712th International Conference on Computer Vision, Kyoto, Japan, 2009, pp. 349-356, doi: 10.1109/ICCV.2009.5459271.

[11] A. Kappeler, S. Yoo, Q. Dai and A. K. Katsaggelos, "Video Super Resolution With Convolutional Neural Networks," in IEEE Transactions on Computational Imaging, vol. 2, no. 2, pp. 109-122, June 2016, doi: 10.1109/TCI.2016.2532323.

[12] W. Lai, J. Huang, N. Ahuja and M. Yang, "Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5835-5843, doi: 10.1109/CVPR.2017.618.