International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

Video Super Resolution Techniques: A Survey

[1] Neeboy Nogueira
Computer Engineering
SRIEIT, Goa University
Shiroda, India

[2] Shawnon Guedes
Computer Engineering
SRIEIT, Goa University
Shiroda, India

[3] Vaishnavi Mardolker
Computer Engineering
SRIEIT, Goa University
Shiroda, India

[4] Amar Parab
Computer Engineering
SRIEIT, Goa University
Shiroda, India

[5] Shailendra Aswale
Computer Engineering
SRIEIT, Goa University
Shiroda, India

[6] Pratiksha Shetgaonkar
Computer Engineering
SRIEIT, Goa University
Shiroda, India

***Abstract*:- Security plays a critical role in our lives. To protect homes, offices, valuable property we need to install security cameras. Most of the time criminals get away due to bad video quality. To overcome this problem, video super resolution technique is applied onto these feeds which can then store the enhanced video footage in turn increasing the chances of any criminal going unaccounted. High resolution video streaming devices can be replaced using the technology to upscale the already existing videos to a higher resolution for a better user experience. Video super resolution is achieved by recovering a high-resolution video from a low-resolution video. In this paper different existing techniques of video super resolution are surveyed and compared. It is found that deep learning technique Convolutional Neural Network (CNN) is a promising solution to achieve video super resolution. In addition, novel video enhanced technique is proposed to enhance the live security feed from a low-resolution security camera.**

Keywords— Video Super Resolution, Deep learning, CNN, RNN, Streaming, Security.

I. INTRODUCTION

A low-quality video has nothing to vouch for, as everyone adequately understands that higher the quality, the richer the experience and better detail visibility. This is sufficiently justified to the fact that High Resolution (HR) frames undoubtedly possess more data when compared to its lower resolution counterpart. Conversion of a low quality video into a higher quality is possible because of super resolution. Super resolution converts the low-level videos into high level by enhancing the video on a frame-by-frame basis, individually up scaling each frame. This can be achieved by either the traditional methods or by employing the deep learning methods. Gaining the upper hand in this case, super resolution genuinely enjoys the recent advances. Video super resolution is typically needed in sectors like medical, satellite, microscopy, astrological studies, surveillance and many more [1] [2] [3].

Primary focus is in the surveillance sector as more and more security cameras are being installed on a daily basis. The data accumulated by the same is in large quantities, but most of it is in low resolution form. As for a HR video a more capable or advance camera is naturally needed, which comes at a high price. Enhancing a video with the use of software is far better and can more enormously increase the quality of a more capable camera. The effective method we tentatively propose employs the unique Convolutional Neural Network (CNN) approach as it is more computational friendly and typically requires less data compared to other techniques.

The results of recent technological development in surveillance video represent the fundamental methods of maintaining public security. Many of the proofs or evidences are being collected by the police by the source of the surveillance video and many grave crimes are being observed by the surveillance video as well. For unraveling criminal cases, surveillance video plays a vital role. It serves the respective departments to reach the end results of particular cases. In addition, the drawbacks of surveillance video would be Low Resolution (LR) because the video footages are probably being captured in low resolution. Most of the surveillance cameras are explicitly used for this purpose. Frequently the users set to have highly compressed surveillance video to utilize minimum storage space, which results in blurred video. Since the video footages were captured in LR form or retain undesirable visual quality, the police authorities find it complicated to deal with the situations. Therefore, new technologies need to be incorporated that would optimally restore the LR videos into HR. This paper presents a survey on numerous video and image super resolution methods. The comparison is based on the various classifiers being used and the final resultant values.

The organization of the paper is as follows: Section II adequately provides literature review of different video super resolution techniques in brief. Section III adequately portrays our proposed methodology. Section IV concludes the paper.

II. LITERATURE REVIEW

Video super resolution represents a domain growing at tremendous speed. Many effective techniques are found which performs this up-scaling of a video from a lower resolution one. The recent techniques are based on learning algorithms which have increased both the efficiency and accountability of the results. This survey is based on the methods used and classifies into three categories: CNN based, Recursive Neural Network (RNN) based and other which did not fit within these is considered as miscellaneous.

A. CNN

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

CNN is a fully look ahead deep learning neural network. It represents a gigantic step-forward in image and video recognition as well as image processing. It has been expressed as constructive network architecture for studying image details, segmentation, etc. due to its accuracy. Moreover, it has been sufficiently proved that CNN learns precisely interpretable image features. The methods utilizing this approach can be observed directly or indirectly in the following papers.

In the VSRnet method, the CNN model is trained on both spatial and temporal resolution of a video to enhance the video [4]. In this, model is pre-trained so that while computing even a small database it enables to compete with state-of-the-art methods. Because of less database requirement, the computational complexity is also reduced along with the need for storage.

An Ordinary Differential Equation (ODE) inspired design scheme is adopted for Single Image Super Resolution (SISR) [5]. This modern method was resorted to on single images and

TABLE I. CNN BASED TECHNIQUES

Ref.	Method	Dataset	Advantages	Disadvantages	PSNR Value	PSNR/SSIM value
[4]	VSRnet	Converted images and videos into the YCbCr colorspace and only used the luminance channel (Y) for training	• much faster than traditional approaches. • reducesd training time of VSRnet by almost 20% • adaptive motion compensation • less computational complexity	overlapping patches are used which leads to a considerable computational overhead	34.33	NA
[5]	OISR-RK3	DIV2K [22]	• combines sub-pixel convolution with spatio-temporal networks & motion compensation. • input is sequence of consecutive frames • enhances the performance and reduces computation overhead	Does not use batch normalization layers	34.67	NA
[6]	ESPCN	Timofte data [23]	• reduces the memory complexity	Increase the computational complexity.	28.09	NA
[7]	CNN based SR	NA	• sharper super resolved video sequence. • less storage space.	loss of sound quality.	38.02	NA
[8]	VDSR	Own Dataset	• easily applicable to other image restoration problems such as denoising and compression artifact removal • good accuracy and visual improvements	output image has the same size as the input image by padding zeros every layer same learning rates for all layers	NA	37.53/ 0.9587
[9]	FSTRN	25 YUV video sequences dataset	• high computational efficiency • presents a novel fast spatio-temporal residual network	Interpolated upscaled input images.	NA	29.95 / 0.87
[10]	DRRN	Yang et al. [24] Berkeley Segmentation Dataset [25]	• advances the SR performance with a deeper yet concise network. • can be easily trained even with 52 convolutional layers • can improve accuracy by increasing depth without adding any weight parameters.	NA	33.99	NA
[11]	VESPCN	CDVL database[28]	• combines sub-pixel convolution with spatio-temporal networks & motion compensation. • Input is sequence of consecutive frames	• not compatible with recurrence, residual connections or training networks • Slower than 29ms on GPU for each frame of 512 × 383 size	27.25	NA
[12]	FRVSR	Vimeo.com	• system is end-to-end trainable and does not requireany pre-training stages. • framework canpropagate information over a large temporal range withoutincreasing computations.	• computationally expensive • generating each output frame separately reduces the system's ability to produce temporally consistent frames, resulting in unpleasing flickering artifacts..	26.63	26.69/0.822

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

[13]	Resolution Network via Exploiting Non	Video Dataset[26]	• The proposed network is able to outperform the state-of-the-art methods with fewer parameters and faster speed.	NA	33.20	NA
[14]	Temporal Group Attention	Vimeo-90k[27]	• Fast spatial alignment method to handle videos with large motion	NA	37.59	NA
[15]	Dynamic up sampling filters (16L-52L)	Own Dataset	• deep network can implicitly handle the motion • explicit motion estimation and compensation	• long training times	NA	27.34/0.8327
[16]	Detail Fusion and SPMC	Own Dataset	• accomplish high-quality results both qualitatively and quantitatively • flexible regarding scaling factors and numbers of input frames	• very computational heavy	NA	29.69 / 0.87
[17]	EVSR (enhanced video SR network)	Myanmar and India Bulidings [28]	• residual blocks and long skip-connection with DAL are introduced for restoring high-frequency components. • The combination of long skip-connection with DAL and residual blocks improves the performance of video SR.	NA	27.99	NA

shows potential for video super resolution as they cast numerical schemes of ODE's as a blueprint and possess two types of network structures. LF-block and RK-block, which correspond respectively to the Leapfrog method and Runge-Kutta method in numerical ordinary differential equations. This method enhances the performance and reduces computation overhead.

Another work proposed to perform the feature extraction stages in the LR space instead of HR space [6]. It uses a novel sub-pixel convolution layer which is capable of super-resolving LR data into HR space with minimum computational cost. This CNN model is capable of SR HD videos in real time on a single GPU.

A novel SR technique uses CNN and intensity-based image registrations. Initially, each frame is super resolved followed by up-sampling of each frame using deep learning-based patch mapping. Intensity based image registration is applied for intensifying the sharpness of the up-sampled frame. The proposed technique resulted in enhancing the quality of the super-resolved video sequence [7].

Another research work uses very deep CNN for video super resolution [8]. It uses residual-learning, adjustable gradient clipping and extremely high learning rates to optimize quickly a very deep network. Convergence speed is maximized and it uses gradient clipping to ensure the training stability. This method outperforms the existing methods in accuracy and visual improvements.

In research work Fast Spatio-Temporal Residual Network (FSTRN) is proposed for video super resolution [9]. This method enhances the performance while maintaining a low computational load. Cross-space residual learning is implemented that directly links the high-resolution space and low-resolution space that relieves the computational burden on the feature fusion and up-scaling parts. FSTRN significantly outperforms other state-of-the-art super resolution methods.

Deep Recursive Residual Network (DRRN) with 52 convolutional layers adopts residual learning in local and global ways [10]. It mitigates the problem of training very deep networks. To control the model parameters while increasing the depth, recursive learning is used. This method is deep, concise, and superior model for SISR.

The VESPCN methodology was introducing a spatio-temporal sub-pixel convolution network which remarkably exploits temporal redundancies and helps improve reconstruction accuracy. Use of early fusion, slow fusion and 3D convolutions for the joint processing of multiple consecutive video frames were done. This provided high accuracy and temporally more consistent videos [11].

FRVSR was proposing a flexible end-to-end trainable framework for video super-resolution that was able to generate higher quality results while being more efficient than existing sliding window approaches. Using Gaussian blur in a recurrent system to train clips of varied length. In an extensive set of experiments, it was revealed that the model outperforms competing baselines in various different settings. The proposed model also significantly outperforms state-of-the-art video super-resolution approaches both quantitatively and qualitatively on a standard benchmark dataset maintaining temporal consistency [12].

Adopting a non-local block for Low Resolution (LR) frame processing. Then, after one 5×5 convolution layer, a series of progressive fusion residual blocks (PFRBs) were added to the network, supposed to make full extraction of both intra-frame spatial correlations and inter-frame temporal correlations among multiple LR frames. Ultimately, it then merged the information from all channels in PFRB and enlarges it to obtain one residual HR (high resolution) image, which is added to the bicubically magnified image to obtain the HR estimate. This proposed network is able to outperform the most recent methods with fewer parameters and higher speeds [13].

In this work, innovative proposal of a novel deep neural network which hierarchically integrates temporal information in an implicit manner was made. To effectively leverage complementary information across frames, the input sequence and is reorganized into several groups of subsequences with different frame rates. The grouping allows

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

to extract spatio-temporal information in a hierarchical manner, followed by an intra-group fusion module and inter-group fusion module. The intra-group fusion module extracts feature within each group, while the inter-group fusion module borrows complementary information adaptively from different groups. Furthermore, a fast spatial alignment is proposed to deal with videos in case of significant motion. The proposed method was capable to reconstruct high-quality HR frames and also maintain the temporal consistency. Extensive experiments on several benchmark datasets demonstrate the effectiveness of the method [14].

Introducing a new deep learning-based framework for VSR that learns to output dynamic upsampling filters and the residual learning simultaneously using temporal augmentation. Achieving the state-of-the-art performance with this new framework when compared with other associated works and recover sharp HR frames and also maintain the temporal consistency. It was also revealed that the model's deep network can implicitly handle the motion without explicit motion estimation and compensation using quantitative evaluation and qualitative comparisons [15].

Enhanced Video SR network with residual blocks (EVSR) recovers HR output frames from multiple LR input frames [17]. To achieve better performance residual blocks and dimension Adjustment layer are introduced. The proposed network completely utilizes spatio-temporal information and efficiently learns non-liner mappings between HR and LR frames.

Comparison of all this research work is summarized in Table I.

B. RNN

Recently, to achieve video super-resolution, deep learning-based methods employing 3-Dimensional (3D) approach show promising performance. However, using 3D convolutions requires high computational demand which restricts the depth of video super resolution models and thus undermine the performance. In RNN shortcut connections are used to skip a few stacked layers in CNN and same set of weights are used recursively resulting in fewer number of parameters. It has been mainly incorporated for video captioning due to its encoder-decoder approach. It has also been employed for video segmentation, video super resolution, etc.

A novel hidden state for the recurrent network, which achieves the best performance among all temporal modeling methods is proposed [18]. In this proposed hidden state, the identity branch carries rich image details from the previous layers to the next layers and helps to avoid gradient vanishing in recurrent training.

Integrating Single Image Super Resolution (SISR) and Multi Image super Resolution (MISR) in a unified Video Super resolution (VSR) framework. Here SISR and MISR extract missing details from diverse sources. Iterative SISR extracts various feature maps representing the precise details of a target frame. MISR adequately provides multiple sets of feature maps from other frames. These various sources are iteratively updated in temporal order through RNN for VSR. The cognitive development for recurrent encoder-decoder mechanism for seamlessly incorporating details extracted in SISR and MISR paths through the back projection was carried out. Here, the network was able to understand the large gap since each context is calculated separately, rather than jointly as in previous work, this separate context plays an important role in Deep Back-Projection Networks (RBPN) [19].

Comparison of RRN methods is summarized in Table II.

C. MISCELLANEOUS

The other Deep learning or traditional methods with more diversity are classified into this category. These represent methods which employed neither RNN nor CNN. Here a mixture of techniques can be referred to in the following papers.

Using Temporally Deformable Alignment Network (TDAN) and SR Reconstruction Network. The TDAN model could compatibly align reference frame and each supporting frame together without determining optical flow. The model would utilize distinctive features from reference frame and each supporting frame to predict the offsets of sampling convolution kernels according to which the model could align the supporting frame with the corresponding reference frame. The Reconstruction Network could accurately predict and restore the HR video frames. For training the entire model Loss functions were used [20].

TABLE II. RNN BASED TECHNIQUES

Ref.	Method	Dataset	Advantages	Disadvantages	PSNR Value	PSNR/SSIM value
[18]	RNN (Recurrent Residual Network)	Vimeo-90k [27]	• RRN is highly computational efficiency and produces temporal consistent VSR results with finer details • Highly efficient and effective	NA	38.7	NA
[19]	RBPN (Recurrent Back-Projection Network)	Vimeo-90k[27]	• combining ideas from single and multiple-frame super resolution	NA	30.10	NA

TABLE III. MISCELLANEOUS TECHNIQUES

Ref.	Method	Dataset	Advantages	Disadvantages	PSNR Value	PSNR/ SSIM value
[20]	TDAN	Vimeo-90k[27]	• strong visual context exploration capability with dynamic sampling.	• can't recover fine image structure and details.	26.58	NA
[21]	Deep Learning Algorithm: SISR	NA	• Iteratively refining HR feature maps representing missing details by up- and down-sampling processes	NA	31.78	NA

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

Focusing primarily on two specific areas that are efficient neural network architectures designed for Single image super resolution (SISR) and Effective optimization objectives for Deep learning (DL) based SISR learning. The practical purpose for this classification is that when applying DL algorithms to tackle a specified task, it is best to consider both the universal DL strategies and the specific domain knowledge [21]. Above discussed technique in this subsection are summarized in table III.

III. PROPOSED METHODOLOGY

Video Super Resolution is a technique where lower quality images are enhanced to give a better resolution image with slightly more precise detail. The architecture is divided into nine phases: To start with the surveillance data is fed live to the module in sequence and the appropriate scene is selected. In the selected scene the most important frames are carefully selected and supplied to the given CNN model where the frames are being worked on by the algorithm, followed by frame mapping and interpolation so that the general stability and truth of the video remains genuine. The resultant video is a higher-resolution video which is then followed by the quality assessment and storage of the HR video. As shown in fig. 1.

IV. CONCLUSION

In this survey paper we exhaustively examined the different video super resolution techniques with or without machine learning. We satisfactorily performed comparative analysis of the major algorithms intentionally used like CNN, KNN, VDSR, SISR. It was scrupulously observed that CNN is more used than RNN used and both outperform most other methods, it was also discovered that CNN is preferred over RNN for fairish more extraordinary precision and remarkable accuracy for videos with intense motion. We subsequently implemented the CNN algorithm in our method to upscale the live video feed from the security cameras and to store the same. A better-quality feed was directly available in case any need for using it arises. This helped in better accurately identifying the principal culprits and better understanding the considered crimes.

In the future our aim is to more significantly improve the sound quality of the videos as more and more techniques for sound enhancement are being generated and also to expand our method in other fields like Medical Video Processing, satellite Video Processing, Microscopy Video Processing, Astrological Studies, Multimedia Industry.

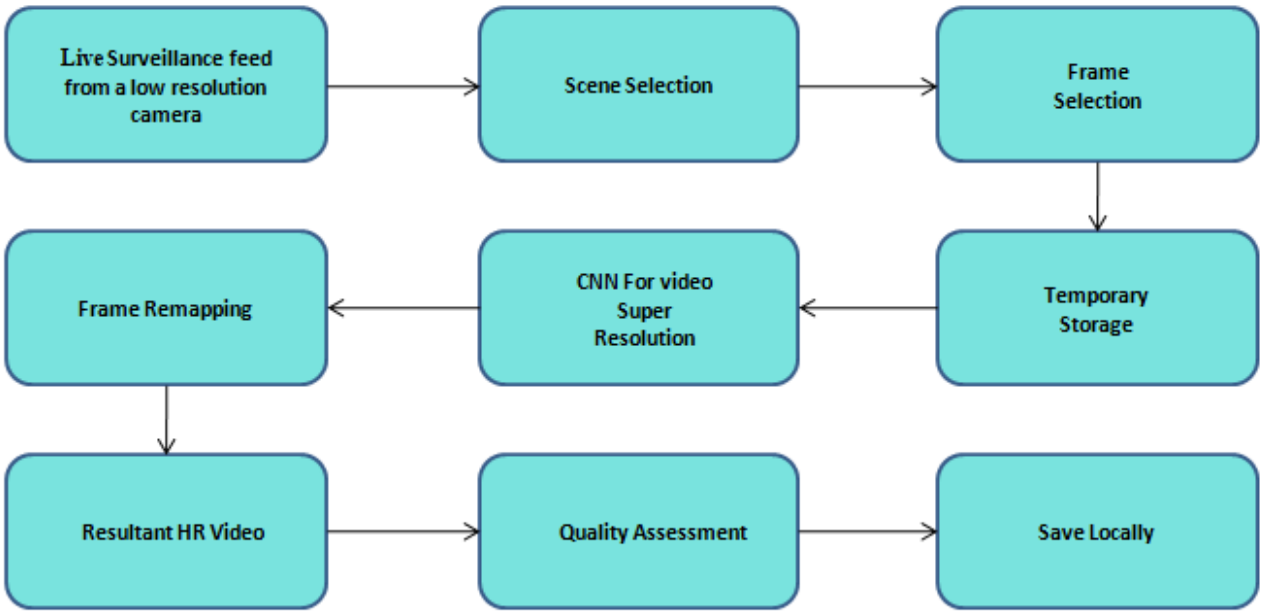

Fig. 1. Proposed Approach.

REFERENCES

[1] H. Greenspan. “Super-resolution in medical imaging”. The Computer Journal, 52(1):43–63, 2009.)
[2] H. Demirel and G. Anbarjafari. “Discrete wavelet transformbased satellite image resolution enhancement”. IEEE Transactions on Geoscience and Remote Sensing, 49(6):1997– 2004, 2011
[3] L. Zhang, H. Zhang, H. Shen, and P. Li. “A super-resolution reconstruction algorithm for surveillance images. Signal Processing”, 90(3):848–859, 2010.
[4] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K. Katsaggelos.“Video Super-Resolution With Convolutional Neural Networks”.IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING, VOL. 2, NO. 2, JUNE 2016
[5] Xiangyu He, et al. “ODE-inspired Network Design for Single Image Super-Resolution”.Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1732-1741
[6] Wenzhe Shi1 , et al. “Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network”.Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1874-1883
[7] Gholamreza Anbarjafari.”Video resolution enhancement using deep neural networks and intensity based registrations”.International Journal of Innovative Computing, Information and Control ICIC International 2018 ISSN 1349-4198 Volume 14, Number 5, October 2018 pp. 1969–1976

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

[8] Jiwon Kim, Jung Kwon Lee, Kyoung Mu Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1646-1654

[9] Sheng Li, et al."Fast Spatio-Temporal Residual Network for Video Super-Resolution".

[10] Ying Tai, Jian Yang, Xiaoming Liu. "Image Super-Resolution via Deep Recursive Residual Network".Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3147-3155.

[11] Jose Caballero, et al. " Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation".Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4778-4787.

[12] Mehdi S. M. Sajjadi, Raviteja Vemulapalli, Matthew Brown. "Frame-Recurrent Video Super-Resolution".Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6626-6634.

[13] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Jiayi Ma."Progressive Fusion Video Super,".Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3106-3115.

[14] Takashi Isobe1, et al. "Video Super-resolution with Temporal Group Attention." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8008-8017

[15] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, Seon Joo Kim. "Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3224-3232

[16] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, Jiaya Jia."Detail-Revealing Deep Video Super-Resolution." Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4472-4480.

[17] WenjunWang,ChaoRen,XiaohaiHe,HonggangChen,LinboQing. "Video Super-Resolution via Residual Learning".10.1109/ACCESS.2018.2829908

[18] Takashi Isobe, Fang Zhu, Xu Jia, Shengjin Wang. "Revisiting Temporal Modeling for Video Super-resolution".

[19] Muhammad Haris, Gregory Shakhnarovich, Norimichi Ukita. "Recurrent Back-Projection Network for Video Super-Resolution." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3897-3906

[20] Yapeng Tian, Yulun Zhang, Yun Fu, Chenliang Xu. "TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution".Proceedings of the IEEE/CVF Conference onComputer Vision and Pattern Recognition (CVPR), 2020, pp. 3360-3369

[21] Wenming Yang, et al. "Deep Learning for Single Image Super-Resolution: A Brief Review"

[22] Radu Timofte, et al. NTIRE 2017 challenge on single image super-resolution: Methods and results. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Honolulu, HI, USA, July 21-26, 2017, pages 1110–1121, 2017.

[23] R. Timofte, V. De Smet, and L. Van Gool. A+: "Adjusted anchored neighborhood regression for fast super-resolution." In Asian Conference on Computer Vision (ACCV), pages 111–126. Springer, 2014

[24] J. Yang, J. Wright, T. Huang, and Y. Ma. "Image superresolution via sparse representation." IEEE Transactions on image processing, 19(11):2861–2873, 2010.

[25] D. Martin, C. Fowlkes, D. Tal, and J. Malik. "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics." In ICCV, 2001.

[26] Zhongyuan Wang, et al. "Multi-memory convolutional neural network for video super-resolution." IEEE Transactions on Image Processing, 28(5):2530–2544,2019.

[27] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. "Video enhancement with task-oriented flow." International Journal of Computer Vision, 127 (8):1106–1125, 2019.

[28] https://www.harmonicinc.com/free-4k-demo-footage/.

[29] ITS. Consumer Digital Video Library, accessed on 08/2016 at http://www.cdvl.org/.