UDC: 004.93 https://doi.org/10.15407/jai2023.03.047

M. Maksymiv[1], T. Rak[2]
[1,2] Lviv Polytechnic National University, Ukraine
12, Bandera Str, Lviv, 79013
[1]mykolamaksymivua@gmail.com
[2]rak.taras74@gmail.com
[1]https://orcid.org/0009-0004-4915-6265
[2]https://orcid.org/0000-0003-0744-2883

METHODS OF VIDEO QUALITY-IMPROVING

Abstract. Video content has become integral to our daily lives, but poor video quality can significantly reduce viewers' experience and engagement. Various super-resolution methods are used to correct this, thereby reconstructing high-resolution videos from low-resolution ones. Two main categories of super-resolution methods exist traditional image processing and deep learning-based techniques. Deep learning-based techniques, such as Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Recurrent Neural Networks (RNNs), have shown great promise in enhancing video quality. The article discusses multiple adaptations of contemporary deep learning models to enhance video resolution. It also briefly explains the framework's design and implementation aspects. Lastly, the paper presents an overview and comparative analysis of the VSR techniques' efficiency on various benchmark datasets. At the same time, the paper describes potential challenges when choosing training sets; performance metrics, which can be used to compare different algorithms quantitatively.

This work does not describe absolutely all existing VSR methods, but it is expected to contribute to the development of recent research in this field and potentially deepen our understanding of deep learning-based VSR methods, as well as stimulate further research in this area. In this work, new solutions for improving the performance of the methods are proposed, in particular, new quality metrics and datasets for model training. Overall, AI-based methods for VSR are becoming increasingly crucial with the rising demand for high-quality video content.

Keywords: Video quality, super-resolution, deep learning, single-image super-resolution, multi-image super-resolution.

Introduction

Video super-resolution technology is a crucial aspect of enhancing and repairing videos. Before the emergence of deep learning, traditional image processing methods such as up-conversion were used to attain video resolution. In the 1960s, Harris and Goodman 7 developed a technique for restoring image information beyond the limit frequency of the optical system modulation transfer function (MTF) [2-4] through spectral extrapolation, which served as the foundation for image and video super-resolution algorithms.

Today, commonly used restoration methods include bilinear interpolation, local adaptive amplification interpolation, and cubic spline interpolation. To use image processing techniques for video, the video must first be parsed into individual frames, then processed with an image super-resolution [2] algorithm. However, this basic image interpolation technique neglects the timing dimension of the video, and fails to account for the correlation between frames before and after the video or the blurring resulting [3] from video transitions and rapid movement. While this method is quick, the reconstructed high-resolution video may exhibit incoherent, fuzzy, or degraded effects, leading to an unsatisfactory subjective outcome.

Problem statement

Video content has become a crucial part of our daily lives, from entertainment to education and advertising communication. However, poor video quality can significantly reduce the viewers' experience and engagement with the content. Video quality refers to the image and sound quality of the video, including the resolution, frame rate, bit rate, color depth, contrast, and brightness. Low-quality videos with low resolution and frame rate can cause blurred images, choppy motion, and pixelation, making it difficult to see the details and follow the action. The sound quality also plays an essential role in the overall video quality, as poor sound can disrupt the viewers' immersion and comprehension.

One of the most common problems with digital videos is poor quality and low

resolution. This can result in blurry images, distorted colors, and pixelated videos. Many factors contribute to this problem includes inadequate lighting, low-quality camera sensors, and compression techniques used during storage and transmission.

Another issue is the loss of detail in dark or bright areas, commonly known as shadow or highlight clipping. This occurs when the camera's dynamic range, or the range of brightness levels that can be captured in a single image is insufficient to capture both the brightest and darkest areas of the scene. This leads to a loss of detail in these areas, resulting in an unnatural or unappealing look.

Finally, motion blur can also be a significant problem in digital videos. This occurs when the scene has important movement, resulting in a loss of sharpness and detail. This can be particularly challenging when recording fast-moving objects or shooting handheld footage.

Given these challenges, **the purpose of this work** is to explore methods for improving the quality and resolution of digital videos. The following sections will discuss the most effective techniques for achieving this goal.

Super-resolution algorithms have become increasingly vital in contemporary times, owing to their usefulness in serving humanity. The execution time of these algorithms is crucial. These algorithms are utilized in a broad spectrum of humanitarian applications such as security, face detection, self-driving cars, computer-aided detection systems, and robot-assisted surgery systems, where image-, video-quality, low-cost, and real-time processing are crucial factors.

However, improving video quality can be challenging, requiring advanced technologies and techniques to capture, process, and display high-quality video content. The following sections will explore methods to improve video quality and resolution, including upscaling, denoising [8], compression, and color grading.

The following sections of this work consider improving the quality and increasing the resolution of video images. At the end of the work, the results of the research on super-resolution methods based on neural networks, given in the previous sections, are described. Based on these studies, the criteria for selecting the super-resolution method for the video were formed.

Analysis of recent research and publications

First, we need to define what Super-resolution is. The process of obtaining a high-resolution image or series of pictures from a set of low-resolution observations is known as super-resolution. Figure 1 provides a visual representation of the super-resolution concept. By enhancing appearance or video quality, super-resolution offers greater scene detail, which is crucial for precise analysis.

Two main categories of super-resolution methods exist traditional image processing and deep learning-based methods [10]. Traditional methods include interpolation-based techniques that estimate high-frequency details by interpolating the known low-frequency information and reconstruction-based methods that minimize an objective function to penalize the difference between the estimated high-resolution image and the observed low-resolution image.

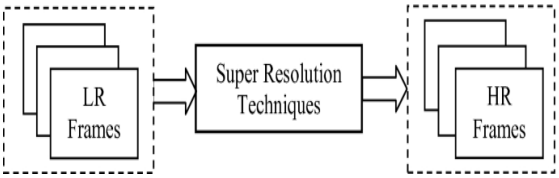

Fig. 1. A general representation of the super-resolution concept

Deep learning-based methods use deep neural networks to learn a mapping between low-resolution and high-resolution image spaces. These methods have achieved state-of-the-art performance in super-resolution and are widely used in practical applications. But deep learning-based methods will be discussed in the following sections.

Video Super-resolution methods can be divided into two main categories (Figure 2): single-image super-resolution (SISR) and multiple-image super-resolution (MISR). SISR strategies [13] aim to enhance the resolution of a single image, while MISR methods [14] use multiple low-resolution photos to produce a higher-resolution output.

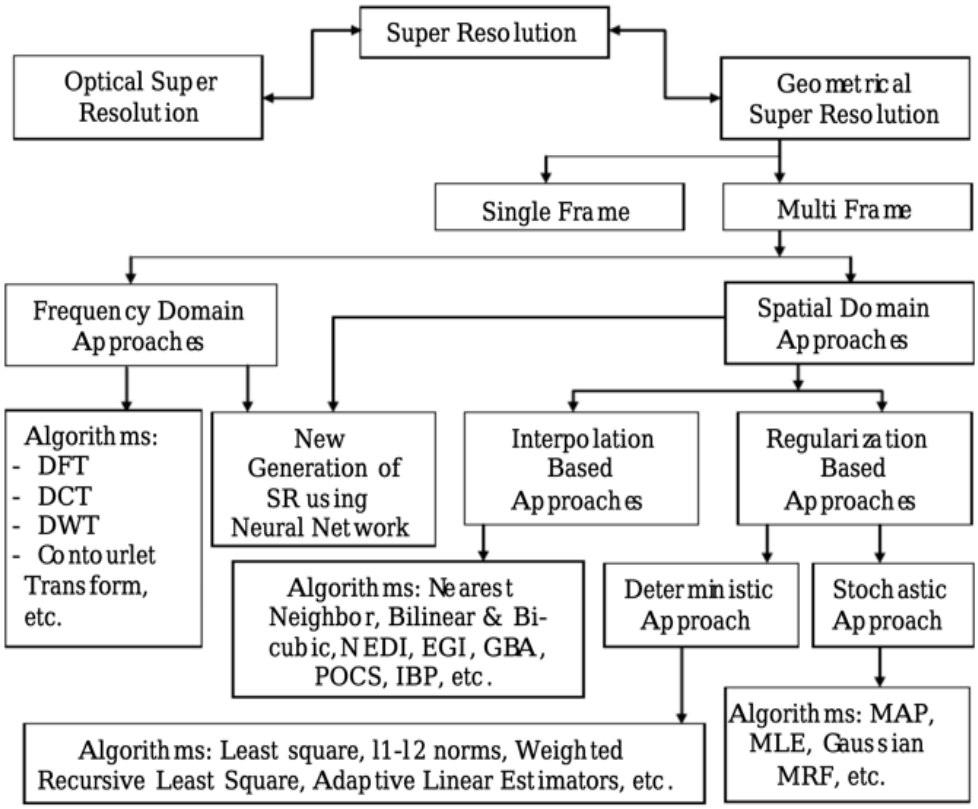

Fig. 2. Classification of Video Super-resolution methods by charts [15]

Single-image Super-resolution Methods

SISR methods can be further categorized into interpolation-based and reconstruction-based [15]. Interpolation-based processes utilize pixel interpolation techniques, such as bicubic [5] or Lanczos interpolation [5], to estimate high-resolution image. On the other hand, reconstruction-based methods employ machine learning algorithms to learn the mapping between low-resolution and high-resolution images. Some popular reconstruction-based methods are Sparse Coding-based Super-Resolution (SCSR) [9-10] and Example-based Super-Resolution (ESR) [6]. SISR methods for video upscaling can be classified into spatial-domain and frequency-domain approaches.

Spatial-domain approaches use pixel-wise interpolation to estimate the high-resolution frames in a video. These methods include bicubic interpolation, Lanczos interpolation [16], and nearest-neighbor interpolation. The most common spatial-domain approach is bicubic interpolation [5], widely used in video upscaling applications due to its simplicity and computational efficiency.

Frequency-domain Approaches of SISR Methods

Frequency-domain approaches use the discrete Fourier transform (DFT) [17] or discrete cosine transform (DCT) to decompose the video frames into frequency components, which are then manipulated to increase the resolution. These methods include the Lanczos-windowed sinc function (LanczosSinc) [16], frequency domain super-resolution (FDSR) [18], and non-local means (NLM) in the frequency domain [18].

They are first cropped into uniform-sized pixel blocks to obtain frequency-domain information from video images (Figure 3). These blocks are then passed through a Discrete Cosine Transform (DCT) module. The DCT module works by projecting the image block onto a collection of cosine components representing different frequencies of 2D signals. Essentially, it breaks down the

image block into its component frequencies.

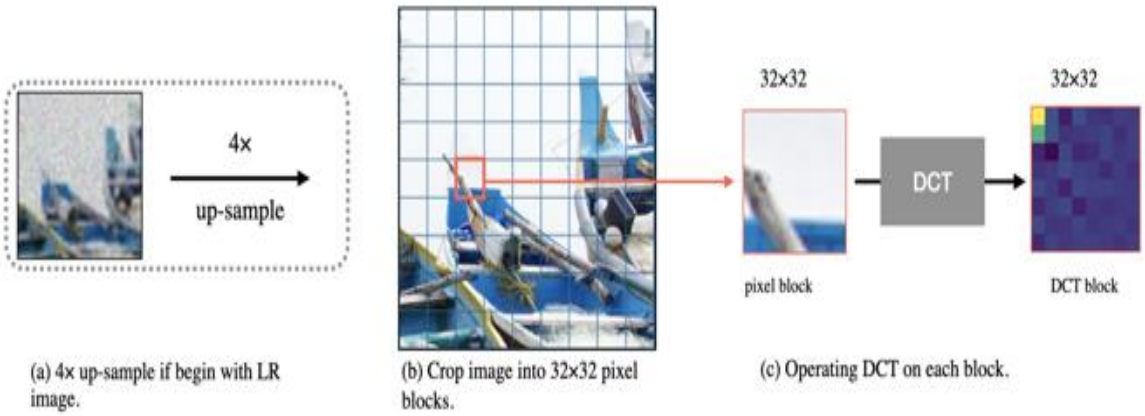

Fig. 3. Converting pixel blocks to DCT blocks

A parameter determines the size of the image block being transformed called the block size, denoted by M. When a 2D image block of size M × M pixels, designated as P, is passed through the DCT module; it is transformed into an M × M DCT block. This DCT block contains information about the frequencies in the original image block, which can be used for super-resolution.

One example of a spatial-domain approach for video upscaling is using bicubic interpolation followed by deblurring and sharpening techniques [7] to produce high-quality upscaled videos. An example of a frequency-domain system is FDSR [18], which applies a convolutional neural network to high-frequency components of the video frames to create a high-resolution output.

Multi-image Super-resolution Methods

MISR methods can be classified into two types: fusion-based and reconstruction-based. Fusion-based methods combine multiple low-resolution images into a single high-resolution image using averaging or weighted sum. Reconstruction-based methods, on the other hand, utilize machine learning algorithms to learn the mapping between low-resolution and high-resolution images using multiple input images. Some popular reconstruction-based methods include Multi-frame Super-Resolution (MFSR) and Recursive Super-Resolution (RSR) [12].

Multiple-image super-resolution (MISR) methods generate high-resolution images of the same scene by utilizing multiple low-resolution images. These methods can be broadly classified into two categories: spatial domain approaches and frequency domain approaches.

Spatial Domain Approaches of MISR Methods.

Spatial domain approaches rely on the alignment of the input images to generate a high-resolution image. One popular approach is multi-frame super-resolution (MFSR), which aligns multiple low-resolution images to create a high-resolution image using interpolation or averaging (Figure 4). The alignment can be achieved through feature-based registration or global motion estimation. The MFSR equation can be expressed as:

$$I_{hr} = \sum_{k=0}^{n} I_{lr_k} \qquad (1)$$

where I_{hr} is the high-resolution image, and I_{lr_k} are the low-resolution input frames [14].

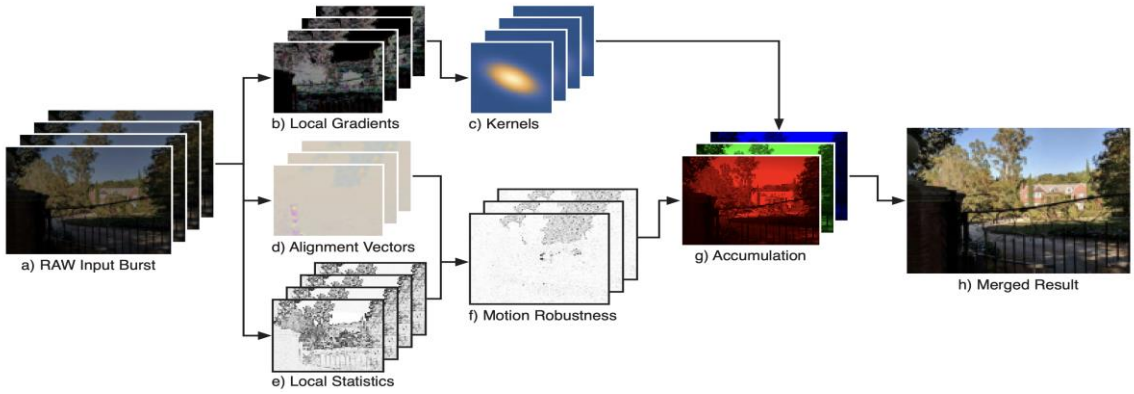

Fig. 4. General overview of variant of usage multi-frame super-resolution (MFSR) algorithm [19] - it takes a burst of raw images (a) as input

The first step is to align every frame locally (d) to a reference frame, which we call the base frame. Kernel regression is used to estimate the contribution of each frame at each pixel. These contributions are accumulated separately for each color channel (g). To adapt the kernel shapes (c), utilize the estimated local gradients (b), and the sample contributions are weighted based on a robustness model (f). This model calculates a weight for every frame per pixel using the alignment field (d) and local statistics (e) from the surrounding area of each pixel. The merged RGB image (h) is the final result of normalizing the accumulated values per channel. As depicted in (b)-(g), the merge refers to combining the individual frames using these steps.

Frequency Domain Approaches of MISR Methods

Frequency domain approaches transform the input images to the frequency domain, apply super-resolution techniques, and transform the output back to the spatial domain. Frequency domain approaches are typically more computationally efficient and can handle images with non-uniform motion blur or missing data.

One popular frequency domain approach is Recursive Super Resolution (RSR) [21], which recursively applies a high-pass filter to the low-resolution images to estimate the high-resolution image. The high-frequency details are then added to the previous estimation to produce a refined output [20-21].

MISR methods have been utilized in various applications, such as medical imaging, satellite imaging, and video surveillance. For example, MISR methods have been used in medical imaging [22] to improve the resolution of magnetic resonance imaging (MRI) and computed tomography (CT) images. MISR methods have been used in satellite imaging to generate high-resolution images of the Earth's surface [14]. MISR methods have been used in video surveillance to enhance the resolution of low-quality surveillance videos [14].

In summary, MISR methods are powerful tools for generating high-resolution images by combining multiple low-resolution photos. Spatial domain approaches, such as MFSR, and frequency domain approaches, such as RSR, are two popular categories of MISR methods extensively studied and utilized in various applications.

Artificial Intelligence-based Methods for Super-resolution

Enhancing video quality using Artificial Intelligence (AI) has become increasingly crucial, given the growing demand for high-quality video content, particularly with the rising popularity of high-resolution displays such as 4K and 8K. As a result, deep learning techniques, such as Convolutional Neural Networks (CNNs) [24], Generative Adversarial Networks (GANs) [25], and Recurrent Neural Networks (RNNs) [12, 23], have emerged as powerful tools for video super-resolution. These neural architectures have shown great promise in enhancing video quality by generating visually realistic and consistent high-resolution frames that align with the content of the original video.

When evaluating super-resolution deep learning methods, specific metrics can be used to assess their strengths and weaknesses. Metrics commonly used include Peak Signal-to-Noise Ratio (PSNR) [27-29], Structural Similarity Index (SSIM) [27-29], Perceptual Index (PI), and Reconstruction Time [29].

Famous examples of each architecture, like SRCNN, VDSR, and EDSR [24] are widely used for super-resolution tasks because they can learn complex features from input images. These models typically achieve good PSNR and SSIM scores but may not perform as well in perceptual quality. However, newer architectures like SRGAN and ESRGAN [25] have addressed this issue by incorporating adversarial loss in the training process.

On the other hand, GANs like SRGAN and ESRGAN have shown impressive results in generating visually appealing images with high perceptual quality. However, these models may be difficult to train and suffer from instability issues, resulting in higher reconstruction times than CNN-based models.

RNNs like LRCN and RSDN have demonstrated promising results in video upscaling applications by modeling temporal dependencies in video sequences and generating smooth, high-quality frames. However, these models can be computationally expensive and require more extended training than CNNs.

Another type of classification of deep learning methods is frame alignment. Methods for video upscaling can be divided into two main groups: those that involve frame alignment (Figure 5) and those that do not.

Methods that involve frame alignment use techniques such as optical flow estimation (for example, Motion Estimation and Motion Compensation, MEMC [30]) or feature matching to align frames before super-resolution. This results in improved super-resolution by capturing the relationships between pixels across frames. Such methods include BasicVSR [26], MuCAN [36], EDVR [32] and others.

On the other hand, ways that do not involve alignment rely on the relationships between pixels within each frame to perform super-resolution. Examples of these methods include the RSDN methodology. While techniques that involve alignment tend to serve better, ways that do not have alignment can still achieve good results and are more straightforward and faster to implement.

It's worth noting that the performance of these architectures can vary depending on the dataset, task, and implementation details. Therefore, evaluating the models using multiple metrics and on different datasets is recommended to understand their strengths and weaknesses comprehensively. Next, we will analyze some specific methods of different architectures.

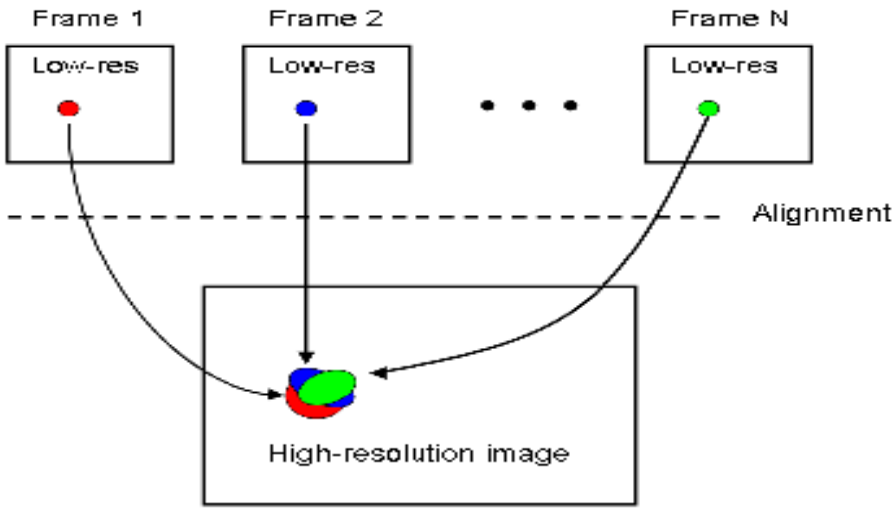

Fig. 5. Creating a super-resolution frame from sibling frames with low resolution

BasicVSR/IconVSR Methods

BasicVSR is a state-of-the-art deep learning-based super-resolution algorithm for video upscaling proposed by the Shanghai Jiao Tong University research team in 2020. It aims to generate high-quality and temporally coherent high-resolution video frames from low-resolution input frames.

The proposed video super-resolution framework in Figure 6 is called BasicVSR [26]. It is a bidirectional recurrent network with three modules: the backward (B) module, the forward (F) module, and the upsampling (U) module. The B module takes the output of the following B module, current frame, and following frame, while the F module takes the output of the previous F module, current frame, and preceding frame. These modules' outputs are fused through the U module to generate the existing structure, which is repeated until all edges are super-resolved. The B/F module comprises generic components, including motion estimation, spatial warping, and residual blocks.

The authors propose two processing mechanisms, information-refill and coupled propagation, to improve BasicVSR, which consists of the IconVSR [27] algorithm. The information-refill mechanism addresses misalignment issues by fusing frames in the selected keyframe set and sending aligned results directly to the residual block without fusion otherwise. The coupled propagation mechanism achieves information interaction between forward and backward processing by using the output of back propagation as input to forward propagation.

While techniques with frame alignment MEMC [30] are commonly used for video super-resolution, including for BasicVSR, they cannot guarantee motion estimation accuracy when lighting changes dramatically or there are large motions in videos, resulting in performance degradation. In situations where videos contain complex motions and varying illumination, the accuracy of motion estimation based on optical flow methods may be compromised as it may violate the assumptions of brightness consistency, small motion, and spatial coherence. This results in inaccurate estimation and the emergence of errors, which can cause artifacts and blurring. The proposed solution uses the EDVR super-resolution method based on Deformable Convolution Models to overcome this issue.

Also, a new model of BasicVSR appeared relatively recently. It is BasicVSR++ [39], which consists of two effective modifications for improving propagation and alignment. The proposed second-order grid propagation and flow-guided deformable alignment allow BasicVSR++ to significantly outperform the existing state of the arts with comparable runtime.

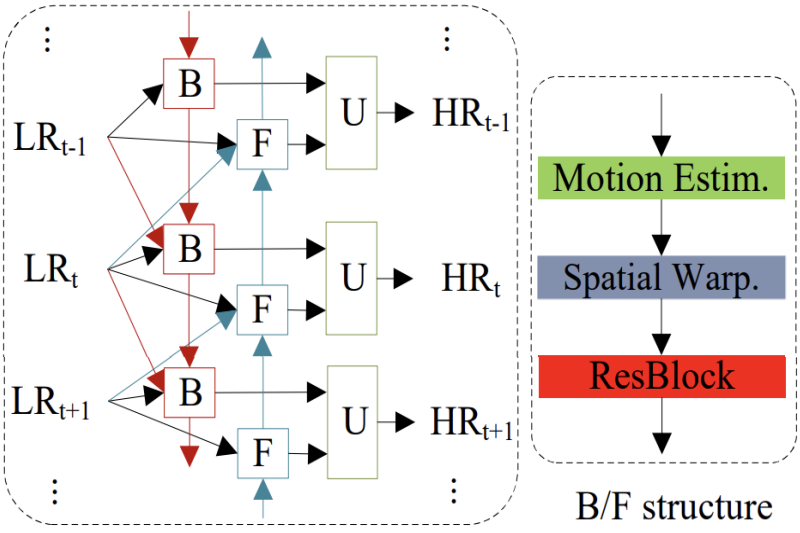

Fig. 6. The network architecture of BasicVSR [26]

Enhanced Deformable Video Restoration Method for Super-resolution

Enhanced deformable video restoration (EDVR) [32] is a state-of-the-art video super-resolution method based on deformable convolution methods.

Deformable Convolution Networks (DCNs) are an extension of traditional convolutional neural networks (CNNs) that address spatial misalignment problems in image and video processing tasks. Unlike regular convolutional layers [32], which use fixed and regular sampling grids, DCNs learn a spatial transformation that adapts to the input data, allowing them to handle large spatial displacements and deformations. The deformable convolution operation involves a learnable offset vector that is added to the regular sampling grid to control the spatial sampling positions of the convolution filter. The offset vector is typically learned from data using backpropagation during training.

EDVR method in Figure 7 uses the Pyramid, Cascading, and Deformable (PCD) alignment module and the Temporal-Spatial Attention (TSA) fusion module to address large motions in videos and to combine multiple frames effectively. The architecture of EDVR consists of four main parts: a PCD alignment module, a TSA fusion module, a reconstruction module, and an upsampling module that employs a sub-pixel convolutional layer. Initially, the input frames undergo alignment using the PCD alignment module. Subsequently, the aligned frames are fused through the TSA fusion module, refined by the reconstruction module, and then upsampled to produce a high-resolution residual image. The final output is attained by adding the residual image to a direct upsampling target frame. To enhance performance, EDVR implements a two-phase approach, where the second phase is comparable to the first phase but with a shallower network depth.

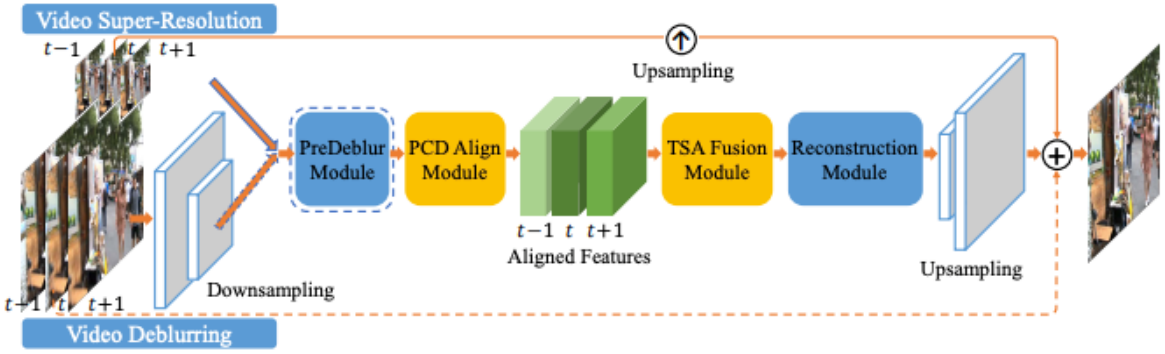

Fig. 7. The network architecture of EDVR [32]

Task-oriented Flow Method for Super-resolution

The architecture of the task-oriented flow (TOFlow) [34] is shown in Figure 8. The TOFlow method is an architecture designed for optical flow estimation aimed at specific tasks such as video interpolation and video super-resolution. It uses a flow field estimator based on a U-Net architecture [34] that predicts dense visual flow fields between two input frames. The U-Net architecture consists of an encoder and decoder network, where the encoder network extracts feature maps from input frames while the decoder network generates dense optical flow fields.

A task-specific module that takes the extracted feature maps and predicts task-specific output, such as interpolated or super-resolved frames, is incorporated. The task-specific module also uses the indicated flow field to warp the input frames for generating the output. The model is trained in a supervised manner using ground-truth data, including the flow field and the task-specific work. The TOFlow method has achieved state-of-the-art results in various video-related tasks, such as video interpolation and video super-resolution.

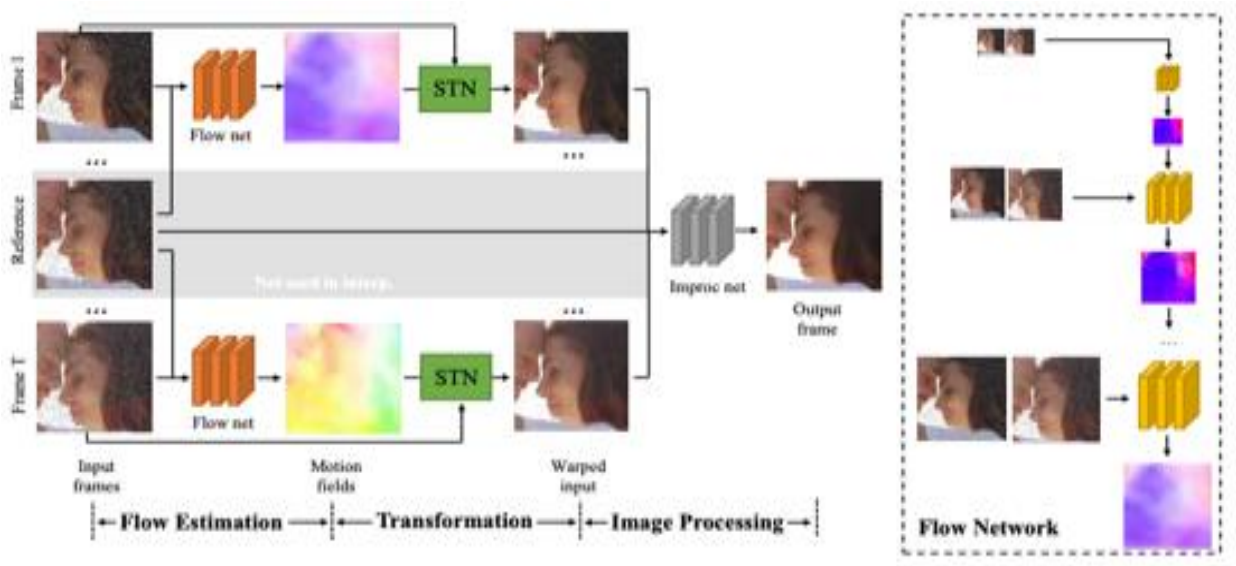

Fig. 8. The general overview of TOFlow [34] architecture, where STN is a spatial transformer network

The Recurrent Structure-Detail Network (RSDN) for Super-resolution

The Recurrent Structure-Detail Network (RSDN) [35] has two main parts: a spatial-temporal recurrent network and a structure-detail fusion module (Figure 9). The recurrent network, composed of three layers, is responsible for extracting features from the input video frames and capturing the temporal information. On the other hand, the structure-detail fusion module is responsible for fusing the features' high-frequency details and low-frequency structures. It consists of two sub-modules: the structure refinement module and the detail refinement module. The structure refinement module is responsible for reconstructing the low-frequency structures of the input frames, while the detail refinement module focuses on rebuilding the high-frequency details.

One of the advantages of RSDN is its ability to capture long-term temporal dependencies by using a spatial-temporal recurrent network. This allows RSDN to generate high-quality super-resolved frames even for videos with complex motion. Another advantage is the structure-detail fusion module, which helps RSDN to produce sharp and detailed results. However, RSDN also has some limitations. It requires many computational resources due to its complex architecture, which can be a problem for real-time applications. Additionally, the performance of RSDN may be limited when dealing with videos that have low-quality or noisy input frames.

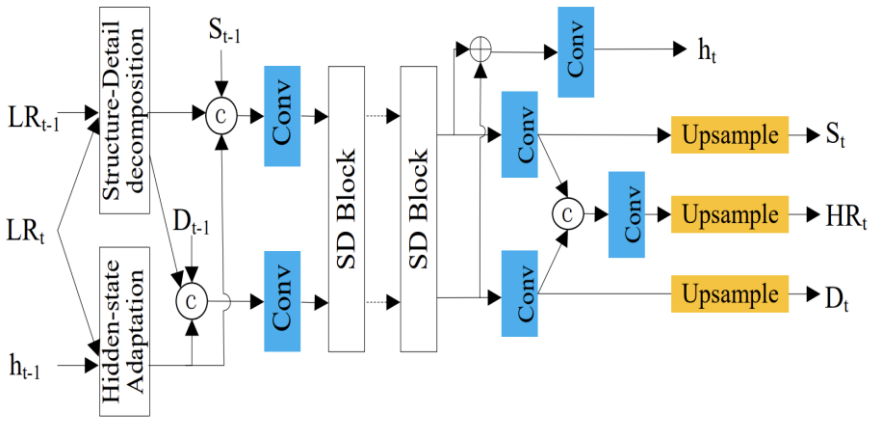

Fig. 9. Architecture overview of Recurrent Structure-Detail Network [35]

Multi-Correspondence Aggregation Network (MuCAN) for Super-resolution

Multi-Correspondence Aggregation Network (MuCAN) [36] is a complete end-to-end video super-resolution network consisting of three major modules (Figure 10): temporal multi-correspondence aggregation module (TM-CAM), cross-scale non-local-correspondence aggregation module (CN-CAM), and a reconstruction module. TM-CAM encodes two adjacent LR frames to lower-resolution features to achieve stability and robustness to noise. An aggregation unit (AU) aggregates multiple patches using a patch-based matching strategy in LR feature space to compensate for significant motion while moving up progressively to low-level/high-resolution stages for sub-pixel shift. In CN-CAM, a pyramid structure based on AvgPool is used for spatio-temporal non-local attention, and coarse-to-fine spatial awareness is executed. Finally, the results are aggregated and sent to the reconstruction module to yield the final HR result.

MuCAN can handle significant motion while maintaining structural information due to the patch-based matching strategy compared to other video super-resolution methods. The pyramid structure based on AvgPool in CN-CAM provides spatio-temporal non-local attention and coarse-to-fine spatial attention, improving the output's spatial resolution. However, the method is computationally expensive due to the multiple aggregation units used in TM-CAM, which can be a disadvantage in some scenarios.

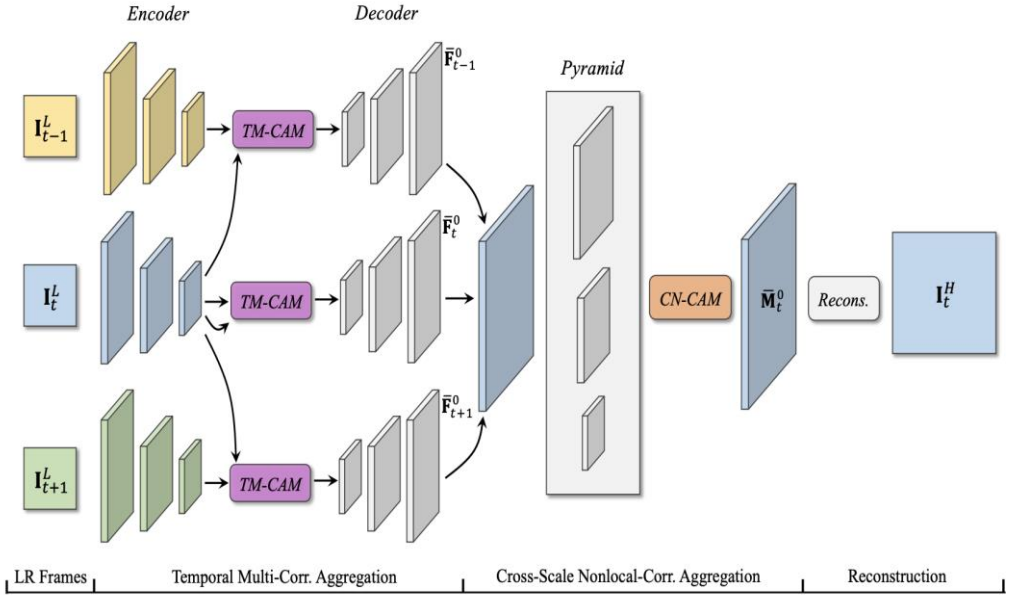

Fig. 10. Multi-Correspondence Aggregation Network Architecture

Performance Metrics

Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) were used to compare and evaluate the performance of super-resolution techniques.

Peak Signal-to-Noise Ratio

PSNR - measures the difference between the original and predicted images in terms of the mean squared error (MSE), defined as:

$$PSNR = 20log_{10}\left(\frac{MAXi}{\sqrt{MSE}}\right) \qquad (2)$$

where $MAXi$ represents the maximum range of color value, which is usually 255, and the mean squared error (MSE) is defined as:

$$MSE = \frac{1}{mn}\sum_{i-0}^{m-1}\sum_{j-0}^{n-1}|I(i,j) - K(i,j)|^2 \qquad (3)$$

where $I(i,j)$ is original video frame, $K(i,j)$ is the super-resolution video frame.

Structural Similarity Index

The Structural Similarity Index (SSIM) is a widely used metric for evaluating image quality. It is based on the idea that human perception of image quality is related to the similarity between local

image structures. SSIM compares two images' luminance, contrast, and design to calculate a similarity score between 0 and 1. SSIM is defined as:

$$SSIM(I,Y) = \frac{2u_I u_y + k_1}{u_I^2 + u_Y^2 + k_1} * \frac{2\sigma_{IY} + k_2}{\sigma_I^2 + \sigma_Y^2 + k_2} \quad (4)$$

Model Training Datasets

Vimeo-9K/Vimeo-9K-T, VID4, and REDs datasets were used for training and testing. The Vimeo-9K dataset and Vimeo-9K-T dataset are two widely used benchmark datasets for video super-resolution, introduced in 2017 [37]. The Vimeo-9K dataset contains 9,120 video sequences, divided into three subsets for training, validation, and testing, with different video content such as natural scenes, animation, and sports. The Vimeo-9K-T dataset is a temporal super-resolution extension of Vimeo-9K, consisting of 2,933 LR-HR video triplets of the same content.

The REDS dataset is a recent benchmark introduced in 2019, consisting of a large-scale collection of diverse real-world video sequences. The dataset contains multiple resolutions and frame rates, providing a more challenging testbed for super-resolution algorithms. The VID4 dataset is another benchmark dataset, consisting of four video sequences with different content, resolutions, and frame rates, that is often used for evaluating the performance of video super-resolution algorithms.

Overall, these benchmark datasets are essential for evaluating the performance of video super-resolution algorithms, enabling researchers to compare and contrast different methods and identify areas for improvement.

Comparison

The dataset contains two upscaling methods: BI and "BD". "BI" stands for bilinear interpolation, which involves averaging neighboring pixels to produce an upscaled image. On the other hand, "BD" stands for bicubic interpolation, a more advanced method that considers more neighboring pixels to produce a smoother upscaled image. Each video clip in the Vimeo-9K dataset is provided in three versions: the original low-resolution video, a bicubic upscaled version (BD), and a bilinear upscaled version (BI).

Note: that part of the PSNR and SSIM are from their original works. And a simple comparison on the performance may not be fair, since the training data, the pre-processing, and the cropped area in videos are likely totally different in the methods.

The dataset also uses the Y/RGB channels to represent the colors. In digital video, colors are typically represented in the RGB color space, which separates colors into red, green, and blue channels. However, many video super-resolution methods operate only on the luminance (Y) channel, representing the image's brightness. In the table, Vid4 and Vimeo-9K-T represent Y channel, REDs represent RGB color images.

Table 1 shows the best results in video super-resolution Vimeo-9K-T datasets as follows. Best results on Vimeo-9K-T presented by PSNR (for both BI/BD metrics): EDVR(BI: 37.61, BD: 37.81), BasicVSR++(BI: 37.59, BD: 38.31), IconVSR(BI: 37.47, BD: 37.84), BasicVSR(BI: 37.20, BD: 37.55). EDVR has the best PSNR result with the bicubic degradation model, while IconVSR has a significant PSNR result in BD models.

In Vid4 datasets, which are known to contain more high-frequency details, the best result from the point of view of PSNR value are the following methods: BasicVSR++(BI: 27.79, BD: 29.04), IconVSR(BI: 27.39, BD: 28.04), EDVR(BI: 27.35, BD: 27.85), BasicVSR(BI: 27.27, BD: 27.98). IconVSR has the best result in both BI/BD sections.

Table 1. Comparison of different video methods based on neural networks for different Test sets

Method/ Test Set	Params size (MB)	BI PSNR/ SSIM	BD PSNR/ SSIM
BasicVSR/ REDs	8.7	31.41/0.8909	-/-
BasicVSR/ Vimeo-9K-T	8.7	37.20/0.9451	37.55/0.9499
BasicVSR/ Vid4	8.7	27.27/0.8248	27.98/0.8556
IconVSR/ REDs	8.64	31.69/0.8951	-/-
IconVSR/ Vimeo-9K-T	8.65	37.47/0.9476	37.84/0.9524
IconVSR/ Vid4	8.65	27.39/0.8279	28.04/0.8570
BasicVSR++/ REDs	9.61	32.39/0.9069	-/-
BasicVSR++/ Vimeo-9K-T	9.61	37.59/0.9476	38.31/0.9550
BasicVSR++/ Vid4	9.61	27.79/0.8401	29.04/0.8735
MuCAN/ REDs	25.7	30.98 0.8750	-/-
MuCAN/ Vimeo-9K-T	19.9	37.32/0.9465	-/-
RSDN/ Vimeo-9K-T	6.19	-/-	37.23/0.9471
RSDN/Vid4	6.19	-/-	27.92/0.8505
TOFlow/ Vimeo-9K-T	1.37	33.08 0.9417	-/-
TOFlow/Vid4	1.37	23.54/0.8070	-/-
EDVR/ REDs	20.76	30.34/0.8664	28.88/0.8361
EDVR/Vimeo-9K-T	18.23	37.61/0.9489	37.81/0.9523
EDVR/Vid4	18.45	27.35/0.8264	27.85/0.8503

In a single training dataset REDs, that corresponds to the RGB color specter, the best results are:
BasicVSR++(32.39/0.9069),
IconVSR(31.69/0.8951),
BasicVSR(31.41/0.8909),
MuCAN(30.98/0.8750).

To summarize, the results provide readers with guidelines for selecting different models based on the results presented in Table 1. To achieve super-resolution videos with realistic textures and rich details without large motions, we recommend using the following methods as prime candidates: BasicVSR++, IconVSR, BasicVSR, and EDVR. These methods are ordered based on their PSNR values on the Vid4 dataset.

Another important indicator is the size of the input parameters for training the network. The parameter size of a training dataset in video super-resolution algorithms refers to the total number of learnable parameters, such as weights and biases, used in the neural network during training. It is an essential factor to consider when training deep learning models because it can impact the computational complexity and accuracy of the algorithm. If we go back to the results in Table 1, we can spectate, that EDVR and MuCAN(for both RGB and Y color channels) have almost 20MB or more of input parameter size, while others have less than 10MB. TOFlow methods accept only 1.37MB of input params for training, it could be advantageous in mobile or embedding systems with tight GPU Memory.

On the three datasets, BasicVSR++ exhibits exceptional performance. IconVSR and EDVR also have great success in VSR tasks. BasicVSR++ leverages optical flow to align features, a bidirectional recurrent network for temporal feature propagation, and an information-refill mechanism for feature refinement, resulting in superior performance compared to other methods in some instances. Additionally, it achieves more significant performance gains with BD degradation than

BI degradation on Vimeo-90K-T and Vid4. On the other hand, EDVR employs cascaded multi-scale deformable convolutions for alignment and TSA to fuse multiple frames.

Video Super-resolution solutions

Unsupervised VSR methods

Most state-of-the-art approaches to video super-resolution (VSR) rely on deep neural networks trained in a supervised manner, but these methods require large datasets of low-resolution (LR) and high-resolution (HR) video frame pairs, which can be difficult and costly to acquire. Moreover, when input video frames have poor resolution, these super-resolution models may not perform well. Additionally, current VSR models trained on labeled datasets can only learn the inverse process of a predefined degradation, which may be too simplistic to represent real-world scenarios.

One potential solution is using unsupervised VSR methods, which work well on unpaired LR and HR video sets. Generative adversarial network (GAN) models [25], in particular, have shown promise for unsupervised VSR, as they can learn to generate realistic HR frames from LR inputs without the need for explicit training on paired LR-HR data. An excellent example of using this architecture is the GAN model with Edge Enhancement [40]. The authors propose the edge enhancement function, which uses the Laplacian edge module to perform edge enhancement on the intermediate result, which helps further improve the results.

By exploring and developing unsupervised methods like GANs, it may be possible to overcome the limitations of supervised training on labeled datasets and improve the performance of VSR models on real-world video data.

Outdated Evaluation Metrics

While deep learning methods have shown great promise in improving VSR performance, traditional evaluation metrics such as peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) may not accurately capture the perceptual quality of the generated video.

Therefore, new evaluation metrics are needed to better assess the perceptual quality of VSR methods based on deep learning. This is because PSNR and SSIM are based on pixel-wise differences that may not correspond well to human perception, and deep learning-based VSR methods may introduce new artifacts that are not captured by these metrics.

Examples of potential evaluation metrics that are better suited for VSR methods based on deep learning include the Perceptual Index (PI), Learned Perceptual Image Patch Similarity (LPIPS), and Fréchet Video Distance (FVD) [28]. These metrics are designed to measure the perceptual quality of the generated video by comparing it to the ground truth video using pre-trained deep neural networks.

By developing and using more appropriate evaluation metrics, researchers and practitioners can more accurately evaluate and compare different VSR methods, leading to further improvements in VSR performance.

Synthetic Datasets

Another challenge in VSR methods is the availability of appropriate training datasets. Many current VSR models are trained on synthetic datasets, which may not accurately represent real-world scenarios. To overcome this, we need to explore new large-scale video datasets such as YouTube-8M [41] and Vimeo-90K-T to capture the natural variability of video content better. However, these datasets also present unique challenges, such as noisy and incomplete data and the need for scalable methods.

Conclusions

Many researchers are exploring the concept of super-resolution and devising possible solutions for its associated challenges. The spatial domain method, which uses pixels for processing, can lead to computational complexity and high memory requirements, reducing the feasibility of real-time systems. Therefore, developers increasingly consider image or frame features rather than the image itself, leading to more attraction toward frequency or wavelet domain techniques. With the emergence of a new generation of resolution enhancement techniques that reformulate spatial and frequency domain

techniques using neural networks, fast and parallel computation of features is now possible.

Different methods based on neural networks were considered separately in this article, synthetic indicators of the quality of each algorithm were given, and the data were analyzed as result. IconVSR, BasicVSR, and EDVR achieve the best results. EDVR has the best PSNR result with the bicubic degradation model, while IconVSR has a significant PSNR result in BD models. IconVSR and EDVR exhibit exceptional performance on the three datasets. The parameter size of the training dataset is an essential factor to consider when training deep learning models, impacting the computational complexity and accuracy of the algorithm. The TOFlow method has a small input parameter size, which could be advantageous in mobile or embedding systems with tight GPU memory.

Based on the conducted analysis, it can be concluded that it is expedient to conduct more detailed research using deep learning methods, especially with using a new deep learning technique based on BasicVSR++ model. There is a need to investigate the performance of these algorithms for each method separately or in combination with new datasets with different characteristics. The Vimeo-9K-T/Vid4 datasets used in this study are known to contain high-frequency details, but there are other datasets with different factors, such as the REDS dataset used for training the MuCAN algorithm. Evaluating these algorithms' performance on various datasets could provide more comprehensive insights into their effectiveness.

And the same time, we should start using new evaluation metrics to better assess the perceptual quality of VSR methods based on deep learning. These metrics can help researchers and practitioners develop more effective VSR systems.

Overall, while significant progress has been made in VSR using deep learning-based methods, challenges still need to be addressed. Improving training datasets to capture the natural variability of video content better is one potential solution, along with developing more effective methods for handling motion artifacts and avoiding introducing new artifacts. There is an opportunity to explore using more advanced techniques, such as adversarial training or attention mechanisms, in video super-resolution algorithms. These techniques have shown promising results in other computer vision tasks and could be applied to video super-resolution to further improve upscaled videos' quality.

References

1. Harris, J. L. (1964). Diffraction and resolving power Journal of the Optical Society of America, 54 (7), 931-933.
2. Suresh, S., Babu, R. V. and Kim, H. J. (2008). No-reference image quality assessment using modified extreme learning machine classifier Applied Soft Computing Journal, 9(2), 541-552.
3. Criminisi, A., Perez, P. and Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting IEEE Transactions on Image Processing, 13 (9), 1200-1212.
4. Freeman, W. T., Pasztor, E. C. and Carmichael, O. T. (2000). Learning low-level vision International Journal of Computer, Vision 40 (1), 25-47.
5. Upscaling: R. Keys, "Cubic convolution interpolation for digital image processing," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 6, pp. 1153-1160, Dec. 1981.
6. Super Resolution: W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," IEEE Computer Graphics and Applications, vol. 22, no. 2, pp. 56-65, Mar./Apr. 2002. doi: 10.1109/38.988747
7. Deblurring: L. Yuan, J. Sun, L. Quan, and H.-Y. Shum, "Image deblurring with blurred/noisy image pairs," ACM Transactions on Graphics, vol. 26, no. 3, pp. 1-10, Aug. 2007. doi: 10.1145/1276377.1276390
8. Denoising: D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," Biometrika, vol. 81, no. 3, pp. 425-455, Sep. 1994. doi: 10.1093/biomet/81.3.425
9. Yang, J., Wright, J., Huang, T., & Ma, Y. (2010). Image super-resolution via sparse representation. IEEE Transactions on Image Processing, 19(11), 2861-2873. doi: 10.1109/TIP.2010.2050625
10. Dong, C., Loy, C. C., He, K., & Tang, X. (2016). Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2), 295-307. doi: 10.1109/TPAMI.2015.2439281
11. Timofte, R., Agustsson, E., Van Gool, L., Yang, M. H., Zhang, L., Lim, B., ... & Lee, K. M. (2018). NTIRE 2017 challenge on single image super-resolution: Methods and results. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 1110-1121). doi: .1109/CVPRW.2017.146
12. Wang, Y., Fan, Y., Yang, J., & Liu, Y. (2019). Deep recursive residual network for image super-resolution. In Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition (pp. 3158-3166). doi: 1109/CVPR.2019.00329

13. Huang, J. B., Singh, A., Ahuja, N., & Yang, M. H. (2015). Single image super-resolution from transformed self-exemplars. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5197-5206).
doi: 10.1109/CVPR.2015.7299187

14. Hu, H., Wang, R., Xu, J., Sun, Z., & Zhang X. (2014). A survey on multi-image super-resolution algorithms. Signal Processing, 93(10), 2876-2894.

15. Daithankar, Mrunmayee & Ruikar, Dr. Sachin. (2020). Video Super Resolution by Neural Network: A Theoretical Aspect. Journal of Computational and Theoretical Nanoscience, 17, 4202-4206.
doi: 10.1166/jctn.2020.9045.

16. The interpolation algorithm: Lanczos Interpolation, 2023.
URL:https://ww2.lacan.upc.edu/doc/intel/ipp/ipp_manual/IPPI/ippi_appendices/ippi_appB_LanczosInterpolation.htm

17. Dr. Steve Arar. An Introduction to the Discrete Fourier Transform, July 20, 2017.
URL:https://www.allaboutcircuits.com/technical-articles/an-introduction-to-the-discrete-fourier-transform/

18. Runyuan Cai, Yue Ding, Hongtao Lu (2021). FreqNet: A Frequency-domain Image Super-Resolution Network with Dicrete Cosine Transform.
URL: https://arxiv.org/abs/2111.10800v1

19. Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-kai Liang, Marc Levoy, and Peyman Milanfar (2021). Handheld Multi-Frame Super-Resolution.
URL: https://arxiv.org/pdf/1905.03277v2.pdf

20. Zhao, H., Wang, Y., & Cai, J. (2014). Recursive filtering based super-resolution for medical images. Journal of Medical Imaging and Health Informatics, 4(2), 254-261. doi: 10.1166/jmihi.2014.1222

21. Udupa, J. K., & Herman, G. T. (1986). A recursive algorithm for nonlinear digital filtering. IEEE Transactions on Acoustics, Speech, and Signal Processing, 34(2), 310-321.
doi: 10.1109/TASSP.1986.1164786

22. Dong, C., Loy, C. C., He, K., & Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In European Conference on Computer Vision (pp. 184-199). Springer, Cham.

23. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 806-814). doi: 10.1109/CVPRW.2017.29

24. Kim, J., Kwon Lee, J., & Mu Lee, K. (2016). Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1646-1654).

25. Ledig, C., Theis, L., Huszr, F., Caballero, J., Cunningham, A., Acosta, A., ... & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4681-4690).
doi: 10.1109/CVPR.2017.19

26. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C. (2021b). BasicVSR: The search for essential components in video super-resolution and beyond. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (pp 4947–4956).

27. Wang, X., Jiang, Y., Ma, S., Yang, Y., Sun, X., & Zhang, Q. (2021). IconVSR: A Generic Framework for Video Super-Resolution with Incomplete Supervision. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(7), 2405-2420.
doi: 10.1109/TPAMI.2020.3042388

28. Zhang, R., Isola, P., & Efros, A. A. (2018). "Learning a perception-based distance metric for image restoration." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7606-7615. URL: https://arxiv.org/abs/1801.03924

29. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). "Image quality assessment: From error visibility to structural similarity." IEEE Transactions on Image Processing, 13(4), 600-612.
doi: 10.1109/TIP.2003.819861

30. Wang, X., Chan, K.C.K., Yu, K., Dong, C., Loy, C.C. (2019). EDVR: Video restoration with enhanced deformable convolutional networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, (pp 1954–1963).

31. Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In Proceedings of the 7th International Joint Conference on Artificial Intelligence-Volume 2 (pp. 674-679).

32. Jifeng, Dai., Haozhi, Qi., Yuwen, Xiong., Yi, Li., Guodong, Zhang., Han, Hu., Yichen, Wei. (2017). Deformable Convolutional Networks. Computer Vision and Pattern Recognition.
URL: https://doi.org/10.48550/arXiv.1703.06211

33. Wang, L., Guo, Y., Lin, Z., Deng, X., An, W. (2019). Learning for video super-resolution through HR optical flow estimation. In: Proc. Asian Conf. Comput. Vis., (pp. 514–529).

34. Xue, T., Chen, B., Wu, J., Wei, D., (2019). Freeman WT. Video enhancement with task-oriented flow. Int J Comput Vis 127(8),1106–1125.

35. Isobe, T., Jia, X., Gu, S., Li, S., Wang, S., Tian, Q. (2020). Video super-resolution with recurrent structure dedetail network. In: Eur. Conf. Comput. Vis., (pp. 645–660).

36. Li, W., Tao, X., Guo, T., Qi, L., Lu, J., Jia, J. (2020). MuCAN: Multi-correspondence aggregation network for video super-resolution. In: Eur. Conf. Comput. Vis., (pp. 335–351).

37. Renjie, Liao, Xin, Tao, Ruiyu, Li, Ziyang, Ma, Jiaya, Jia (2015). Video Super-Resolution via Deep Draft-Ensemble Learning IEEE International Conference on Computer Vision.
doi: 10.1109/ICCV.2015.68

38. MMEditing, C. (2022). MMEditing: OpenMMLab Image and Video Editing Toolbox (Version 0.13.0) [Computer software].
URL: https://github.com/open-mmlab/mmediting

39. Chan, Kelvin C.K. and Zhou, Shangchen and Xu, Xiangyu and Loy, Chen Change (2021). BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment.
URL: https://doi.org/10.48550/arXiv.2104.13371

40. Wang, Jialu, Teng, Guowei & An, Ping. (2021). Video Super-Resolution Based on Generative Adversarial Network and Edge Enhancement. Electronics. 10. 459. doi: 10.3390/electronics10040459.

41. YouTube-8M Segments training dataset (2023). URL: https://research.google.com/youtube8m

The article has been sent to the editors 01.10.23.
After processing 20.10.23.
Submitted for printing 30.11.23.