




Article

DSTnet: Deformable Spatio-Temporal Convolutional Residual Network for Video Super-Resolution

Anusha Khan , Allah Bux Sargano *  and Zulfiqar Habib * 

Department of Computer Science, COMSATS University Islamabad, Lahore 54000, Pakistan; anooshyk@gmail.com

* Correspondence: allahbux@cuilahore.edu.pk (A.B.S.); drzhabib@cuilahore.edu.pk (Z.H.)

Abstract: Video super-resolution (VSR) aims at generating high-resolution (HR) video frames with plausible and temporally consistent details using their low-resolution (LR) counterparts, and neighboring frames. The key challenge for VSR lies in the effective exploitation of intra-frame spatial relation and temporal dependency between consecutive frames. Many existing techniques utilize spatial and temporal information separately and compensate motion via alignment. These methods cannot fully exploit the spatio-temporal information that significantly affects the quality of resultant HR videos. In this work, a novel deformable spatio-temporal convolutional residual network (DSTnet) is proposed to overcome the issues of separate motion estimation and compensation methods for VSR. The proposed framework consists of 3D convolutional residual blocks decomposed into spatial and temporal (2+1) D streams. This decomposition can simultaneously utilize input video's spatial and temporal features without a separate motion estimation and compensation module. Furthermore, the deformable convolution layers have been used in the proposed model that enhances its motion-awareness capability. Our contribution is twofold; firstly, the proposed approach can overcome the challenges in modeling complex motions by efficiently using spatio-temporal information. Secondly, the proposed model has fewer parameters to learn than state-of-the-art methods, making it a computationally lean and efficient framework for VSR. Experiments are conducted on a benchmark Vid4 dataset to evaluate the efficacy of the proposed approach. The results demonstrate that the proposed approach achieves superior quantitative and qualitative performance compared to the state-of-the-art methods.

Keywords: video super-resolution; deformable convolution; 3D convolution; spatio-temporal; residual neural network; deep learning



Citation: Khan, A.; Sargano, A.B.; Habib, Z. DSTnet: Deformable Spatio-Temporal Convolutional Residual Network for Video Super-Resolution. *Mathematics* **2021**, *9*, 2873. <https://doi.org/10.3390/math9222873>

Academic Editors: Akemi Galvez Tomida, Lihua You, Hassan Ugail, Andres Iglesias Prieto and Alexander Malyshev

Received: 10 October 2021

Accepted: 6 November 2021

Published: 12 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, image and video super-resolution have attracted a lot of attention due to their wide range of applications, including, but not limited to, medical image reconstruction, remote sensing, panorama video super-resolution, UAV surveillance, and high-definition television (HDTV) [1–3]. Because video is one of the most often used forms of multimedia in our everyday lives, super-resolution of low-resolution videos has become critical. In general, image super-resolution (ISR) algorithms process a single image at a time. In contrast, video super-resolution algorithms handle many successive video frames at a time to reconstruct the target HR frame using the relationship between the frames. Video super-resolution (VSR) may be considered a subset of image super-resolution since it can be processed frame by frame by image super-resolution methods. However, performance is not always satisfactory due to the artefacts and jams that may be introduced, resulting in unreliable temporal coherence across frames [4].

Earlier VSR methods estimated motion through separate multiple optical flow algorithms [5] and enabled an end-to-end trainable process for VSR. However, it made VSR methods highly dependent upon externally estimated optical flow between frames.

In addition to this, standard optical flow calculated by these algorithms is not an optimal motion representation for video restoration tasks, including VSR [6]. To address this issue, several methods proposed the use of local spatio-temporal information between frames to capture optical flow for motion estimation [7,8]. These methods employed spatial and temporal information separately to extract the feature from frames and estimate movement during the motion compensation step. However, these methods lacked in utilizing the discriminating spatio-temporal information of input LR frames efficiently, resulting in reduced coherence between reconstructed HR videos [9].

With the remarkable success of deep learning in various fields, super-resolution techniques based on deep learning have been explored intensively. Many video super-resolution approaches based on deep neural networks, such as convolutional neural network (CNN), generative adversarial network (GAN), and recurrent neural network (RNN), have been developed. A large number of low-resolution video sequences are fed into the neural network for feature extraction, inter-frame alignment, and feature fusion to produce high-resolution videos. The pipeline of these methods consists of three major components, frame alignment, features-fusion of aligned frames, and the reconstruction of HR frames [7,10]. The success of these methods depends on the inter-frame motion estimation and compensation methods. These methods usually use spatial and temporal information separately for feature extraction and motion estimation to effectively align the video frames. Hence, are unable to utilize the discriminative spatio-temporal information of input LR frames fully and efficiently [10]. In this regard, the use of 3D convolution (Conv3D) is one of the most straightforward ways for the simultaneous handling of spatial and temporal information. However, additional parameters introduced by the temporal dimension of Conv3D add computation complexities to the overall process and limit the depth of CNN models for VSR. Consequently, this restricts the capabilities of the CNN-based VSR method to learning and modeling the complex functions effectively. To handle this problem, a lightweight variant of Conv3D [11] was introduced but due to the fixed receptive fields of convolution kernels, this method was also not effective for modeling the complex motion and geometric transformation. Therefore, just increasing the depth of neural networks is neither an optimal nor an effective choice for developing an efficient model for VSR [12].

This paper proposes an end-to-end deformable spatio-temporal convolutional residual network (DSTnet) for VSR by adopting the ResNet as the underlying architecture. Unlike simple convolution kernels, deformable convolution (Dconv) kernel size and positions are also learnable, making it suitable for estimation and compensation of inter-frame motion. The proposed model consists of multiple novel spatio-temporal convolutional residual blocks (resST) where each resST consists of spatial and temporal convolutional blocks (2+1) D instead of plain 3D convolution layers. This decomposes each 3D filter into a 2D filter for learning spatial features and a 1D filter for learning temporal features. Unlike previous methods, a deformable convolution is introduced in a special module towards the end of the network for handling complex geometric transformations with reduced computational complexity. Thus, the proposed model simultaneously utilizes spatial and temporal information, effectively, handles inter-frame complex motion, and produces visually appealing results. The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 elaborates the proposed methodology. Section 4 discusses the experimental setup and results, and, finally, the paper is concluded in Section 5.

2. Related Work

During the last few years, several methods have been proposed for image and video super-resolution. These methods can be categorized into model-based and learning-based methods. Model-based methods usually try to model internal similarities of images using machine-learning techniques. In this regard, Freedman et al. [13] used the self-example approach to generate LR exemplar patches from the input image using a variant of the nearest neighbor search algorithm and attempt to generate HR patches using LR patches of similar pattern. Yang et al. [14] proposed the alternative to the nearest neighbor strategy

based on the regression method that further improved the LR and HR patches mapping and accuracy. However, all these methods required extensive pre and post-processing to reconstruct the HR image. On other hand, learning-based methods used low and high-resolution exemplar pairs to learn the mapping function. The sparse-coding-based method [15] is an example of the learning-based methods. Earlier learning-based SR methods extracted patches from input LR images to learn representation and reconstruct LR images in the same fashion. Hence, in all aforementioned methods patches were the focus of optimization. Different techniques were used to extract LR patches before the main SR process and aggregation of HR output, which lead to considerable computational overhead. With the introduction of deep learning, features can be learned directly from the raw data. Which eliminates the need for any hard-coded pre- or post-processing of input data. These deep representation-based approaches are generally called end-to-end learning-based methods. In the subsequent sections, state-of-the-art deep learning-based methods for image and video SR are discussed.

2.1. Single Image Super-Resolution (SSIR)

The single image super-resolution (SSIR) takes a single LR image as input to reconstruct the HR image. In this direction, Dong et al. proposed SRCNN [16], a CNN-based architecture to learn the non-linear mapping between LR and HR images for image super-resolution, and reported superior results. Then, Kim et al. [17] proposed a CNN-based architecture with increased network depth and residual learning as an extension to SRCNN. Later, Tai et al. [18] came up with a CNN-based network with residual connection and reported better efficiency. The major limitation of these methods was using a pre-processing step for up-scaling the LR images to the desired output size before training the model, which caused increased computational complexity and alteration in details of input LR frames. Shi et al. [19] avoid this problem by using an efficient sub-pixel CNN layer instead of a deconvolution layer to upscale the LR feature map to HR at the end of the network and achieved better results than the previous method for the SR. In the NTIRE 2017 challenge, Timofte et al. [20] provided a huge dataset of diverse 2K resolution images (DIV2K). This dataset enabled the researcher to develop deeper and more effective models for SR, such as EDSR [21], RCAN [22], and RDN [23].

2.2. Video Super-Resolution (VSR)

VSR methods generally divide the video super-resolution task into three steps; feature extraction and motion compensation, feature fusion, and SR reconstruction. In this direction, Kappeler et al. [24] extended SRCNN [16] for VSR and used multiple consecutive frames as input to predict HR frames. Later, Wang et al. [25] used optical flow information for inter-frame motion compensation and temporal alignment and reported a better temporal alignment method. Liao et al. [5] used two different classical optical flow algorithms for motion compensation and then used a CNN-based model to reconstruct SR video frames. Later, Liu et al. [26] proposed an improved optical-flow alignment method that generated HR frames in temporal scales by a temporal adaptive method. Caballero et al. [27] proposed an end-to-end efficient sub-pixel convolution neural network for video (VESPCN), comprised of three modules: a spatial-transformer for motion compensation and feature extraction, feature fusion, and HR reconstruction. Tao et al. [28] emphasized the importance of accurate inter-frame alignment and motion compensation for VSR. They used a sub-pixel motion compensation (SPMC) layer in their method to simultaneously achieve motion compensation and super-resolution. All these methods revealed that precise optical flow prediction is crucial for VSR, errors in the optical flow computation or the image-level wrapping operation can introduce artefacts in the resultant VSR.

Alternative techniques were proposed for VSR to capture temporal-relation without explicit motion compensation. For example, Huang et al. [29] proposed bidirectional recurrent convolution networks capture long-term spatio-temporal relations between frames for VSR. Another method, frame recurrent video super-resolution (FRVSR) [30]

recurrently used two deep CNN models by taking previously estimated HR feature value as input and reconstructed subsequent HR frames. Some other deep learning-based methods, such as feed-forward networks [31], generative adversarial networks (GANs) [32] were also proposed. Although these methods improved visual quality, these methods are slower than many CNN-based VSR methods. To overcome this issue, Zhang et al. [33] used pixel correlations extracted by compression algorithms to exploit dense representation of the network; by transferring the SR result between adjacent frames, they accelerated the VSR process by almost 15 times with little performance loss. Another method proposed by Xue et al. [6] was a turning point for the optical flow-based method for VSR. It concluded that traditional optical flow is not an ideal motion representation for video restoration tasks, including VSR. To circumvent the problem, Jo et al. [34] proposed an implicit motion compensation model that generated dynamic up-sampling filters using each pixel's local spatio-temporal neighborhood and HR residual image. Tain et al. [8] proposed a temporally deformable alignment network (TDAN) to avoid explicit motion compensation problems using a one-stage temporal alignment process at the feature level. This method used features from both the reference frame and neighboring frames using deformable convolution, and then applied these learned kernels to perform the frame alignment. More recently, Bao et al. [35] proposed an end-to-end trainable motion estimation and compensation network, by combining both kernel-based and flow-based methods for frame interpolation. They designed a unique adaptive warping layer that integrates both estimated optical flow and interpolation kernels to synthesize target HR frame pixels and achieved encouraging results for video enhancement tasks including VSR.

2.3. Deformable Convolution-Based Methods

Dia et al. [36] proposed the deformable convolution (Dconv) that enhanced the capability of traditional CNN-based methods to learn geometric transformation. Deformable convolution appends learned offset to the sampling grid of regular convolution kernel, which enables it to learn information away from its local neighborhood. Deformable convolutions are widely used in high-level computer vision tasks, such as action recognition, semantic video segmentation [37]. For example, Zhang et al. [38] proposed a deep deformable 3D convolutional neural network for task of gesture recognition, that not only achieved excellent accuracy but also met the demand of real-time processing. However, Tain et al. [8] was the first method to introduce a deformable convolution-based method for VSR and successfully achieved the frame alignment without explicitly computing optical flow. They reported superior results as compared to state-of-the-art VSR methods. Wang et al. [7] also used an enhanced deformable convolution network for video restoration tasks, including VSR. Their proposed architecture consists of two modules: (1) a pyramid and cascading alignment based on TDAN [8] (2) a temporal and spatial attention-based fusion model. Furthermore, the deformable convolution layers are proposed to integrate with the convolutional LSTM [39], and recurrent convolutional network [40] that enhanced the performance of VSR methods. Recently, Lpezatapia et al. [41] proposed gated recurrent neural networks for VSR that incorporate some of the key components of a gated recurrent unit and deformable convolution.

2.4. 3D Convolution-Based Methods

The most straightforward way to learn spatio-temporal information from the input video sequence is to employ 3D convolution (Conv3D). Furthermore, there are considerable similarities between LR and desired HR videos, so the residual connection is widely used in VSR methods. Li et al. [11] used residual connections and proposed a model termed as fast spatio-temporal residual network (FSTRN) for VSR by utilizing factorized Conv3D for learning spatio-temporal features. This method used spatial and temporal kernels in different layers and effectively reduced computational complexity at training time. Other methods, such as [11,42], used C3D [12] as their backbone architecture. Similarly, some other methods, such as [21,43], gained much success in image SR tasks by efficiently

using the ResNet [44] as a backbone architecture. However, these techniques are not fully explored and utilized for VSR [11].

3. Methodology

In this section, the detail of the proposed architecture is presented. As shown in Figure 1, the proposed architecture consists of four modules: (1) spatio-temporal convolutional residual blocks (*resST*), (2) deformable spatio-temporal convolutional residual blocks (Deformable *resST*), (3) features fusion, and (4) SR reconstruction.

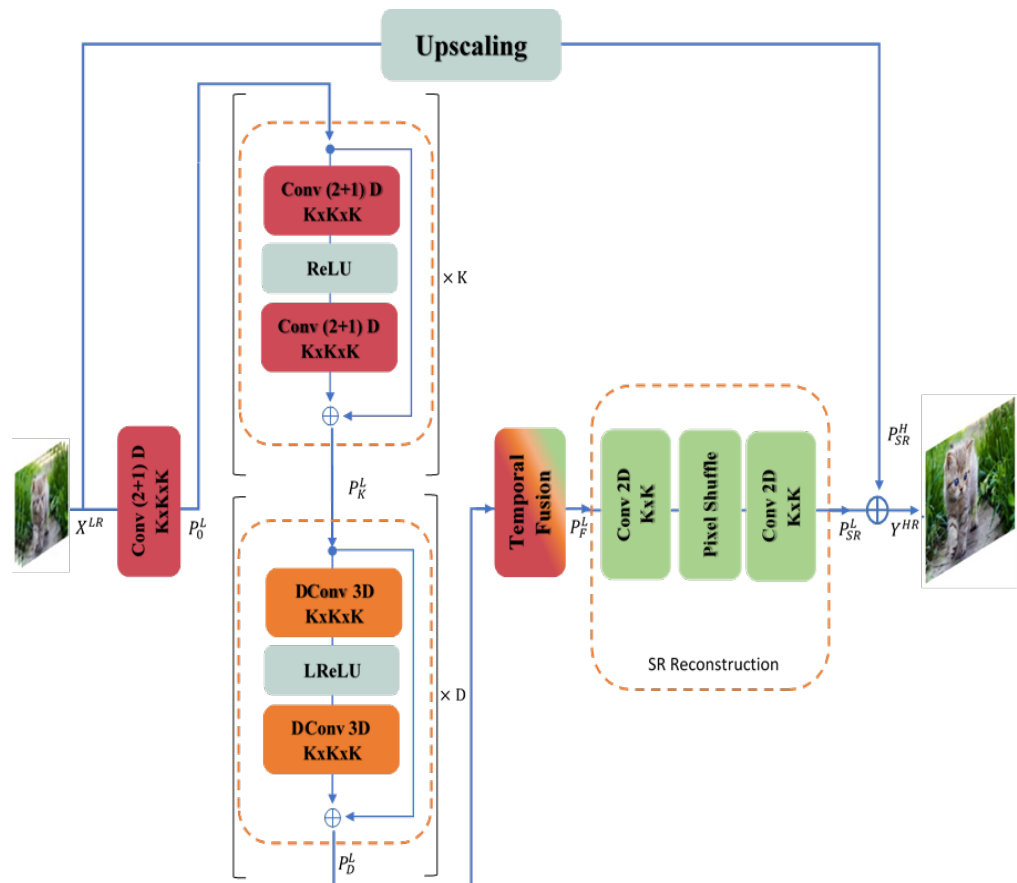


Figure 1. Proposed DSTnet architecture, the structure of a residual blocks is shown on left, and temporal fusion and reconstruction module shown on the right side of the figure.

Let us denote the input and output of the proposed DSTnet method $F_{(DSTnet)}$ by X_{LR} and Y_{HR} . First, N number of LR frames X_{LR} are fed to a spatio-temporal convolutional block (2+1) D with convolution kernel of size $3 \times 3 \times 3$ to extract features, which can be expressed as:

$$P_0^L = F_{((2+1)D)}(X_{LR}), \tag{1}$$

where $F_{((2+1)D)}$ represents (2+1) D convolution function to obtain initial feature map P_0^L . That later used as input for spatio-temporal convolutional (2+1) D residual (*resST*) blocks, to learn in-depth spatio-temporal features. Assume K number of *resST* blocks are used, the first residual block uses P_0^L as input. The next *resST* block further learns features using the previous residual block's output and so forth; more details about the *resST* blocks are presented in Section 3.1 The output of the Kth *resST* block P_K^L can then be obtained by:

$$P_K^L = F_{(resST,K)}(F_{(resST,K-1)}(F_{(resST,K-2)}(\dots(F_{(resST,1)}(P_0^L)\dots))), \tag{2}$$

where $F_{(resST,K)}$ denotes the operation of *resST* blocks. P_K^L is further used as input for the special deformable spatio-temporal convolutional residual blocks. These blocks are designed to enhance the proposed network’s ability to learn the complex motion. In this module each 2D spatial convolution layers in *resST* blocks are replaced with the deformable convolution layers [38]. More details will be presented in Section 3.2. Assume D such residual block has been used in model the output of D th block can be denoted by P_D^L . To combine estimated P_D^L across time-space, the temporal fusion module is used, detail of the module is presented in Section 3.3. The term that expresses this fused deep feature map is P_F^L in Figure 1, which is further used as input for the SR reconstruction module. The output SR feature map can be denoted by P_{SR}^L . Finally, the output of the network is composed of SR mapping from LR space termed as P_{SR}^L from SR reconstruction module and mapping of reference LR frame in HR space the obtained P_{SR}^H . The estimated HR frame Y_{HR} is obtained by performing concatenation of P_{SR}^L and P_{SR}^H using a global residual connection. The overall output of *DSTnet* is as following,

$$Y_{HR} = F_{DSTnet}(X_{LR}) = P_{SR}^L + P_{SR}^H \tag{3}$$

where F_{DSTnet} represents the overall operations performed by the proposed *DSTnet* to reconstruct the HR video frames Y_{HR} .

3.1. Spatio-Temporal Convolutional Residual Blocks

Residual blocks have gained much success in computer vision tasks by ensuring excellent performance [43,45]. Lim et al. [21] proposed the modified residual blocks for SR by removing the batch-normalization layer from residual blocks, as shown in Figure 2b. To apply residual neural networks in videos, generally, 2D convolutions are replaced with 3D convolutions to utilize spatial and temporal relations between frames. However, Li et al. [11] proposed to decompose 3D convolution into 2D convolution followed by a 1D convolution for the VSR task, as shown in Figure 2a. This section presents details about the proposed spatio-temporal convolutional residual block (*resST*), shown in Figure 2c.

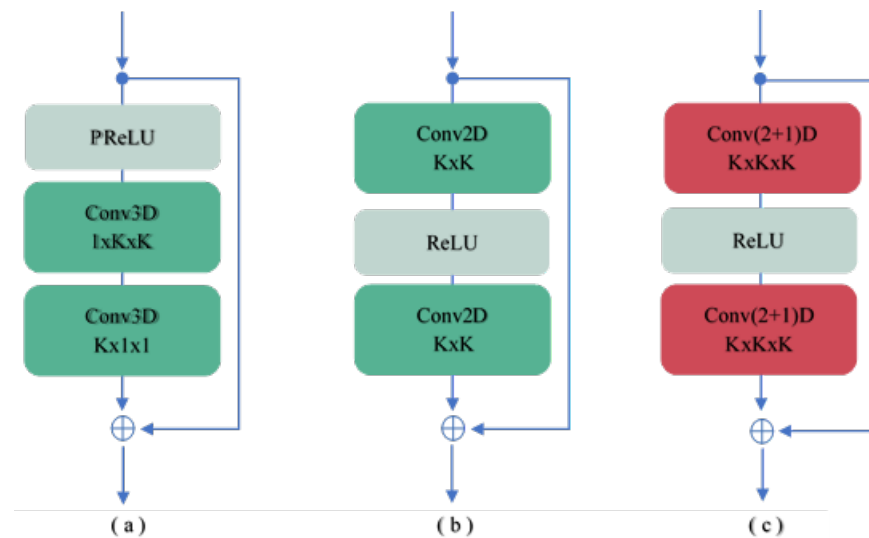


Figure 2. Comparison of (a) FSTRN [11] 3D residual block, (b) EDSR [21] 2D residual block, and (c) the proposed spatio-temporal convolutional residual block.

In the proposed residual block, each N number of 3D convolution filters with dimension $t \times s \times s$ are replaced by (2+1) D convolutional block consists of N number of spatial convolution filter of dimension $1 \times s \times s$. Followed by M temporal filters of dimension $1 \times 1 \times t$ where M determines the subspace dimensionality between both filters. The value of M is calculated in a way similar to [46]. This ensures that (2+1)D block learnable param-

eters are not more than that required for a 3D convolution kernel. The proposed residual module consists of two (2+1)D convolutional blocks with ReLU as the activation function. Additionally, the same-padding strategy is adopted in every spatio-temporal convolutional block to avoid dimensionality reduction. The use of (2+1) D convolutional blocks has two major advantages over 3D convolution. First, it can better model more complex functions due to additional non-linearity between 2D and 1D convolution in each layer. Second, it provides better optimization with superior performance at training and test time without additional computation.

3.2. Deformable Spatio-Temporal Convolutional Residual Block

Dai et al. [36] proposed a deformable convolution (Dconv) that achieved much success in the field of computer vision. Let us consider a simple convolution operation Y with stride = 1 summarized as follows:

$$Y(P_0) = \sum_{k=1}^K w_k \cdot X(P_0 + P_k), \quad (4)$$

where P_0 represents a location in the output feature and P_k represents the convolution-sampling grid. As depicted in the equation above, convolution is a weighted summation of sampled input features using a convolution kernel for a fixed location P_0 . In contrast, a deformable convolution kernel can augment the sampling grid by learning an additional offset P_k for each sampling location. Thus, it can enlarge the spatial receptive field. Deformable convolution shows superiority in many high-level computer vision tasks. Inspired by its success, it was recently used for temporal frame alignment in the state-of-the-art VSR method [8].

The proposed variant of 3D convolution can learn and model spatio-temporal relations simultaneously. However, convolutional neural networks (CNN) have an inherited limitation in learning complex geometric transformation and motion. To overcome this issue, in this work deformable convolution is used that enhances the model's capability to estimate and compensate complex motion between input LR frames. Although, in the proposed network, deformable convolution can easily be integrated within every spatio-temporal convolutional block of each *resST* block. However, it is evident from the literature that deformable convolution requires high-level semantic information to perform best [38]. Integrating it in every layer, especially using it in the starting layers of the network would only bring extra computational complexity, in the form of learned offset for each deformable convolution. Hence, to make model architecture robust and the training process optimized, deformable convolution layers are proposed to use only in the high-level layer of the network (i.e., towards the end of the network).

3.3. Temporal Fusion

The main objective of this module is to temporally combine spatial and temporal features of the output learned features map of the residual blocks. In the proposed feature fusion module a 2D convolution layer is used as a bottleneck layer to fused residual block's output feature maps temporally. The obtained HR features maps are passed to 2D convolution residual blocks to further fine-tune the fused features.

3.4. SR Reconstruction

The SR reconstruction module is used to obtain the estimated super-resolved video in HR space after efficiently extracting deep features in LR space. In this module, sub-pixel convolution layer proposed by Shi et al. [19] is used for HR frame reconstruction. This module consists of simple convolution layers followed by the rearrangement of the pixels by using pixel-shuffle operation. This upscale the feature map to desired resolution using values from all learned feature maps.

4. Experimentation and Results

4.1. Dataset

Vimeo-90K [6], and Vid4 [47] are two benchmark datasets explicitly developed for video super-resolution. The Vimeo-90K is comprised of 89,800 video clips. Each video clip presents different contents, diverse scenes, and motions. These videos were extracted from a popular video-sharing website, vimeo.com. Although Vid4 is comprised of four video sequences: city, calendar, foliage, and walk. Each video sequence has a different motion sequence and is recorded to varying resolutions under different circumstances.

The Vimeo-90k dataset is comparatively more challenging due to diverse camera and object motion recorded for high-quality videos in different circumstances. In this research work, the Vimeo-90k dataset has been used to train the proposed model, and the Vid4 benchmark dataset is used to evaluate the performance of the proposed VSR method. This strategy is in-line with state-of-the-art methods to ensure a fair comparison.

4.2. Experiment Details

The video super-resolution is a supervised learning task, aims to infer the degradation function between input LR video and corresponding HR ground truth. Following other VSR methods, we use MATLAB's *imresize* function as a degradation method to generate LR video frames using HR frames of both datasets. Hence, the final training dataset has 89,900 pairs of HR frames and their corresponding LR frames for training. For the evaluation of the proposed VSR method, the Vid4 dataset is used.

All experiments are conducted using Pytorch [48] with an NVIDIA 2080Ti GPU, with a batch size of 32. During training Adam [49] is used as an optimization algorithm with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The model initial learning rate was set to 1×10^{-4} and halved after every 30 epochs. LReLU [50] is chosen activation function. Furthermore, motivated by experimental results in some modules of DSTnet ReLU is selected as an activation function. Mean square error (MSE) is the loss function used during the proposed network training. A sequence of 7 frames is utilized as input to the proposed model. Following the previous methods [26], Ref. [6] we consider only luminance (Y) channel in YCbCr color space of input frames during training.

If not specified otherwise, each convolution layer has 64 filters, and kernel size is set to $3 \times 3 \times 3$. The proposed model consists of three spatio-temporal convolutional residual blocks followed by five residual blocks with integrated deformable convolution. Further detail of the proposed network's modules, output shape of each module, and their order in model architecture is given in Table 1. For quantitative comparison of the proposed method with existing methods, both PSNR and SSIM [51] are used as evaluation metrics.

4.3. Results

The proposed method is compared with several single image super-resolution methods and VSR methods, including VESPCN [27], RCAN [22], VSRnet [24], SOF-VSR [25], BRCN [29], DBPN [31], VSRResNet [32], TOFlow [6], and TDAN [8] on the benchmark VSR Vid4 [47]. Quantitative results are shown in Tables 2 and 3, first and last two frames of videos are not consider for evaluation following TDAN [8] for fair comparison. Additionally, note that most of the aforementioned methods are trained on different datasets and, the comparison is made on the results they all provide in their research work. However, the same training as reported in TOFlow [6] and TDAN [8] is used in this work. The proposed method achieved the highest Structural Similarities Index Measure (SSIM) value on a benchmark dataset and reported comparative PSNR.

Table 1. Detail of DSTnet architectures considered in experiments. Output shape dimension are batch size, channel, height, and width. Layers details is formatted as: filter shape, number of filters. The spatio-temporal convolutional blocks are shown in brackets, by the number of times these are repeated in each resST blocks.

Module	Output Shape	Layer's Detail
Feature Extraction	64, 7, 64, 112	$\left\{ \begin{array}{c} 1 \times 3 \times 3, 64 \\ ReLu \end{array} \right\} \times 1$
resST blocks	64, 7, 64, 112	$\left\{ \left\{ \begin{array}{c} 1 \times 3 \times 3, 64 \\ ReLu \end{array} \right\} \times 2 \right\} \times 3$
Deformable resST blocks	64, 7, 64, 112	$\left\{ \begin{array}{c} 3 \times 3 \times 3, 54 \\ LReLU \end{array} \right\} \times 5$
Temporal Fusion	64, 64, 112	$\left\{ \begin{array}{c} 3 \times 3, 64 \\ LReLU \end{array} \right\} \times 6$
SR Reconstruction	1, 256, 448	$\left\{ \begin{array}{c} 1 \times 1, 64 \\ PixelShuffle \\ 3 \times 3, 64 \end{array} \right\} \times 1$

Table 2. Comparison of PSNR (dB) with state-of-the-art methods on Vid4 dataset with scaling factor of 4. The best methods results are shown in boldface.

Method	Year	City	Walk	Calendar	Foliage	Average
Proposed	2021	27.08	29.56	23.14	25.77	26.39
TDAN [8]	2020	26.99	29.50	22.98	25.51	26.24
TOFlow [6]	2019	26.78	29.05	22.47	25.27	25.89
VSRResNet [32]	2019	–	–	–	–	25.51
SOF-VSR [25]	2018	–	–	–	–	26.02
RCAN [22]	2018	26.06	28.64	22.33	24.77	25.45
DBPN [31]	2018	25.80	28.64	22.29	24.73	25.37
VESPCN [27]	2017	26.17	28.31	21.98	24.91	25.34
VSRnet [24]	2016	25.62	27.54	21.34	24.41	24.73
BRCN [29]	2015	–	–	–	–	24.43

Qualitative results of all four videos of the Vid4 dataset are shown in Figures 3–6. It can be observed that the proposed method can produce more visually appealing results, with less blur and motion artefacts around objects. As shown in Figure 3, the name is more visible in proposed method generated results. In Figure 4, the texture and building detail is only reconstructed by the proposed model results. Similarly, in Figures 5 and 6, the texture and scene details of the roof of the car and bags, respectively, are preserved by the proposed methods while all other methods produced overly smoothed and distorted results.

Table 3. Comparison of SSIM with state-of-the-art methods on Vid4 dataset with scaling factor of 4. The best methods results are shown in boldface.

Method	Year	City	Walk	Calender	Foliage	Aaverage
Proposed	2021	0.779	0.899	0.769	0.733	0.795
TDAN [8]	2020	0.757	0.890	0.756	0.717	0.780
TOFlow [6]	2019	0.740	0.879	0.732	0.709	0.765
VSRResNet [32]	2019	–	–	–	–	0.753
SOF-VSR [25]	2018	–	–	–	–	0.771
RCAN [22]	2018	0.694	0.873	0.723	0.664	0.738
DBPN [31]	2018	0.682	0.872	0.715	0.661	0.732
VESPCN [27]	2017	0.696	0.861	0.691	0.673	0.730
VSRnet [24]	2016	0.654	0.844	0.644	0.645	0.697
BRCN [29]	2015	–	–	–	–	0.633

Conclusively, the proposed approach outperforms the state-of-the-art VSR method in terms of PSNR and SSIM on a benchmark dataset. The residual connection and spatio-temporal convolutional blocks have played an important role in learning deep generalized representations. DSTnet can also fully utilize and model spatio-temporal relation with the ability to model complex motion, using the novel deformable convolution module for super-resolution of video clips with dynamic scenes and motion. As illustrated in the results, the proposed approach outperforms benchmark datasets in terms of structural similarity (SSIM) and PSNR as compared to state-of-the-art approaches for VSR with reduced parameters.

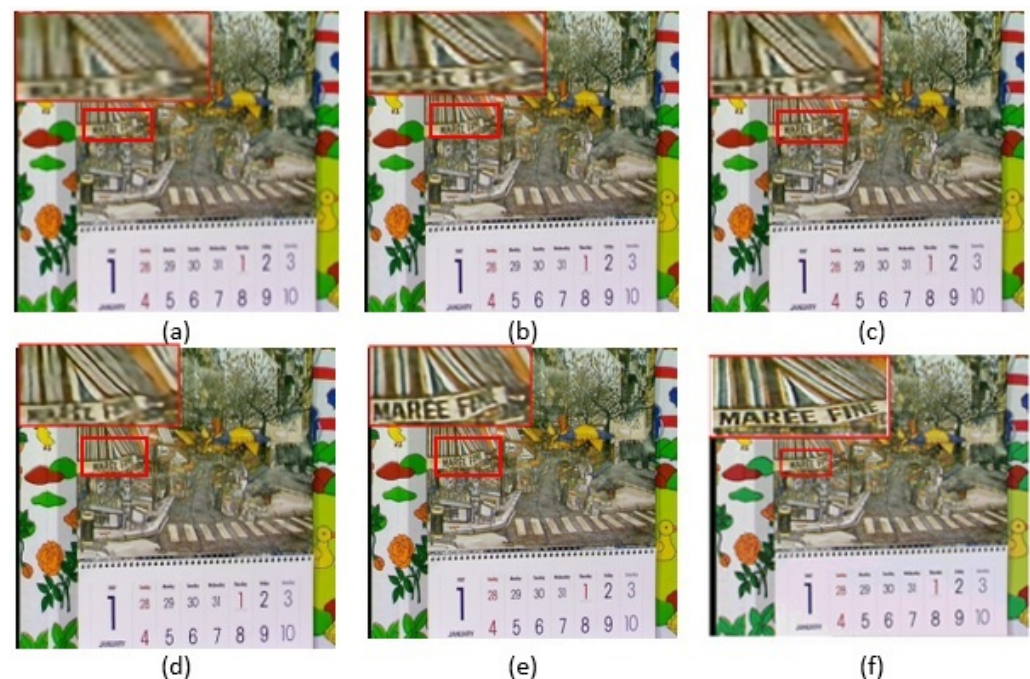


Figure 3. Qualitative comparison between VSR results obtained by (a) Bicubic, (b) VSRNET [24], (c) SRCNN [16], (d) VESPCN [27], (e) Proposed and (f) Ground-Truth on “calendar” from Vid4 with scaling factor of x4.

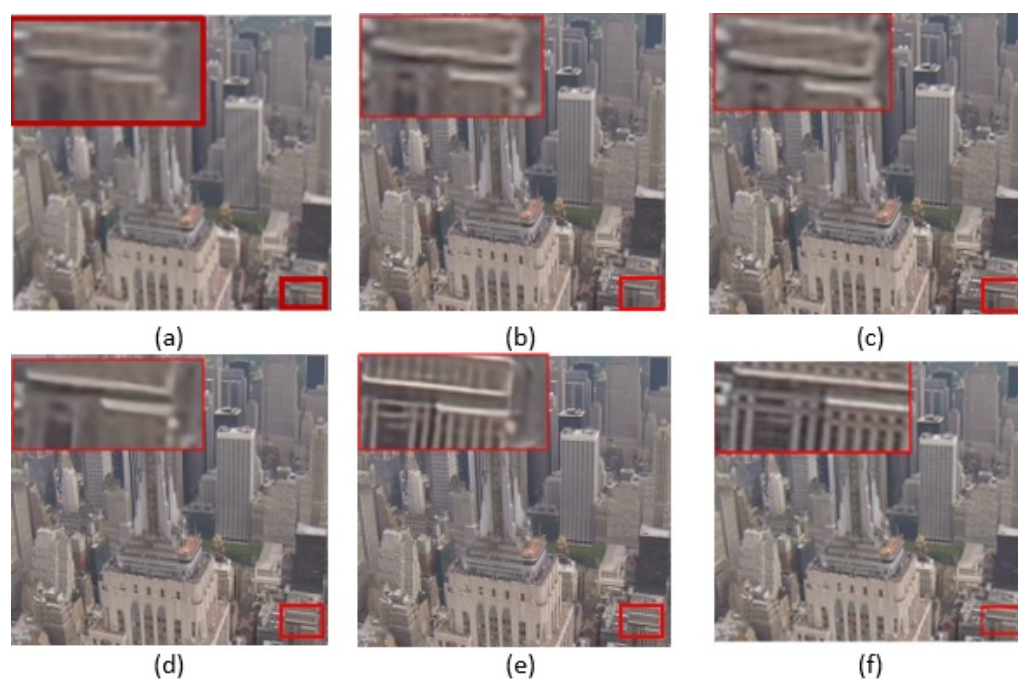


Figure 4. Qualitative comparison between VSR results obtained by (a) Bicubic, (b) VSRNET [24], (c) SRCNN [16], (d) VESPCN [27], (e) Proposed and (f) Ground-Truth on “city” from Vid4 with scaling factor of x4.

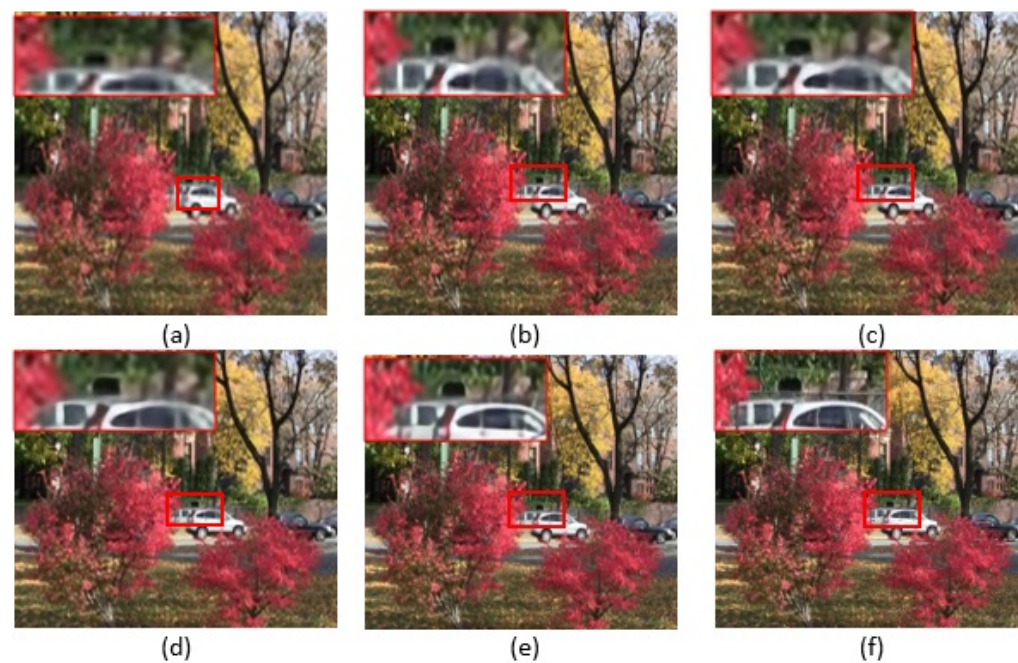


Figure 5. Qualitative comparison between VSR results obtained by (a) Bicubic, (b) VSRNET [24], (c) SRCNN [16], (d) VESPCN [27], (e) Proposed and (f) Ground-Truth on “foliage” from Vid4 with scaling factor of x4.

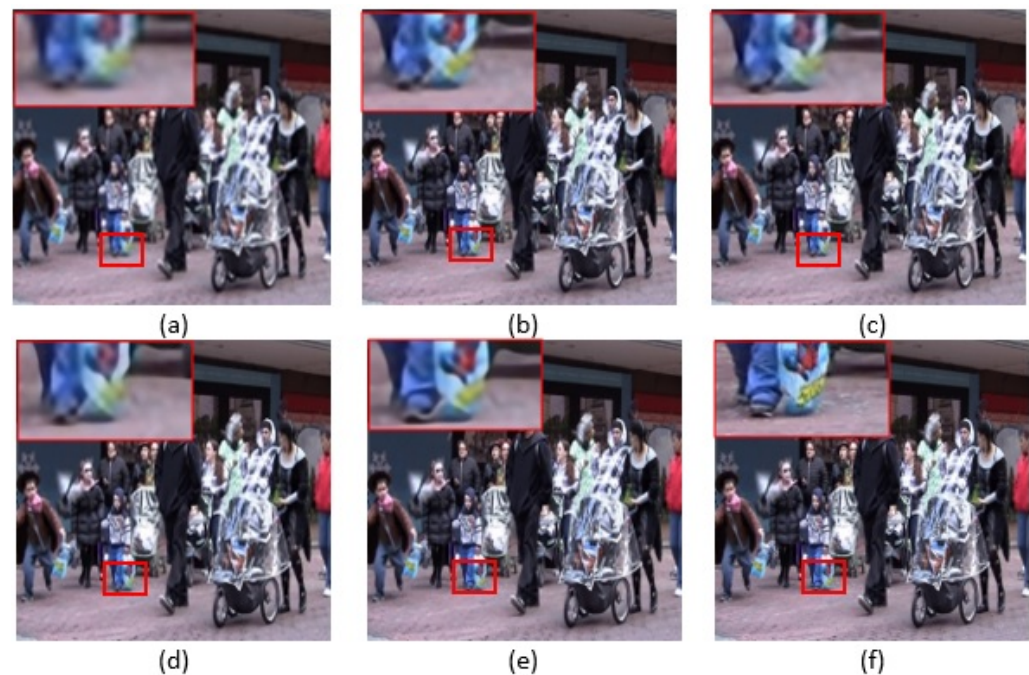


Figure 6. Qualitative comparison between VSR results obtained by (a) Bicubic, (b) VSRNET [24], (c) SRCNN [16], (d) VESPCN [27], (e) Proposed and (f) Ground-Truth on “walk” from Vid4 with scaling factor of $\times 4$.

Furthermore, our model has 3.91 million learnable parameters, which influence the model’s size. Table 4 shows the number of learnable parameters of several networks of VSR. It can be observed from Table 4 statistics that the proposed model sizes is small in comparison with the networks having leading VSR performance: ToFlow [6], RDN [23], and RCAN [22] except from TDAN [8]. This demonstrates that the proposed method is computationally efficient, due to the tiny model size. Even with such a lightweight model, the proposed model is deep in comparison with other deep-learning-based models using 3D convolution for VSR. However, the proposed method still achieves encouraging VSR performance and outperforms the state-of-the-art methods, as shown in Tables 2 and 3.

Table 4. Numbers of parameter in millions of different networks with leading VSR performance.

Methods	RCAN [22]	RDN [23]	TOFlow [6]	TDAN [8]	Proposed
Parameters (in millions)	15.50 M	22.30 M	6.20 M	1.97 M	3.91 M

5. Conclusions and Future Work

In this paper, a novel deformable spatio-temporal convolution residual network (DST-net) is proposed for video super-resolution. This method consists of spatio-temporal (2+1) D convolutional residual block with deformable convolution layers to simultaneously utilize spatial and temporal information. Experiments confirm that DSTnet can effectively capture and model complex motion between frames and outperform state-of-the-art methods on the benchmark Vid4 dataset. The proposed method is evaluated using two well known and widely used metrics for VSR methods, i.e., SSIM and PSNR. It achieves SSIM of 0.795 and PSNR of 26.39 dB, which are higher than state-of-the-art VSR methods. Moreover, the proposed method has fewer parameters to learn during training, making it computationally lean and proving its fast learning ability. As a future research direction, we would like to extend this method to handle various complex motions by improving the feature-fusion module of this method.

Author Contributions: A.K., A.B.S. and Z.H. conceived and designed the experiment; A.K. performed the experiments; A.B.S. and Z.H. analyzed the data; A.B.S. and A.K. contributed reagents/materials/analysis tools; A.K. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the PDE-GIR project which has received funding from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 778035.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rasti, P.; Uiboupin, T.; Escalera, S.; Anbarjafari, G. Convolutional neural network super resolution for face recognition in surveillance monitoring. In Proceedings of the International Conference on Articulated Motion and Deformable Objects, Palma de Mallorca, Spain, 13–15 July 2016; Springer: Cham, Switzerland, 2016; pp. 175–184.
2. Cao, L.; Ji, R.; Wang, C.; Li, J. Towards domain adaptive vehicle detection in satellite image by supervised super-resolution transfer. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
3. Gunturk, B.K.; Altunbasak, Y.; Mersereau, R.M. Multiframe resolution-enhancement methods for compressed video. *IEEE Signal Process. Lett.* **2002**, *9*, 170–174. [[CrossRef](#)]
4. Liu, H.; Ruan, Z.; Zhao, P.; Dong, C.; Shang, F.; Liu, Y.; Yang, L. Video Super Resolution Based on Deep Learning: A comprehensive survey. *arXiv* **2020**, arXiv:2007.12928.
5. Liao, R.; Tao, X.; Li, R.; Ma, Z.; Jia, J. Video super-resolution via deep draft-ensemble learning. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 531–539.
6. Xue, T.; Chen, B.; Wu, J.; Wei, D.; Freeman, W.T. Video enhancement with task-oriented flow. *Int. J. Comput. Vis.* **2019**, *127*, 1106–1125. [[CrossRef](#)]
7. Wang, X.; Chan, K.C.; Yu, K.; Dong, C.; Change Loy, C. Edvr: Video restoration with enhanced deformable convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–19 June 2019.
8. Tian, Y.; Zhang, Y.; Fu, Y.; Xu, C. Tdan: Temporally-deformable alignment network for video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3360–3369.
9. Kappeler, A.; Yoo, S.; Dai, Q.; Katsaggelos, A.K. Super-resolution of compressed videos using convolutional neural networks. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 1150–1154.
10. Hayat, K. Multimedia super-resolution via deep learning: A survey. *Digit. Signal Process.* **2018**, *81*, 198–217. [[CrossRef](#)]
11. Li, S.; He, F.; Du, B.; Zhang, L.; Xu, Y.; Tao, D. Fast spatio-temporal residual network for video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10522–10531.
12. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
13. Freedman, G.; Fattal, R. Image and video upscaling from local self-examples. *ACM Trans. Graph. (TOG)* **2011**, *30*, 1–11. [[CrossRef](#)]
14. Yang, J.; Lin, Z.; Cohen, S. Fast image super-resolution based on in-place example regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 1059–1066.
15. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image super-resolution via sparse representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [[CrossRef](#)] [[PubMed](#)]
16. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [[CrossRef](#)] [[PubMed](#)]
17. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
18. Tai, Y.; Yang, J.; Liu, X. Image super-resolution via deep recursive residual network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3147–3155.
19. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
20. Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.H.; Zhang, L. Ntire 2017 challenge on single image super-resolution: Methods and results. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 114–125.

21. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
22. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
23. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2472–2481.
24. Kappeler, A.; Yoo, S.; Dai, Q.; Katsaggelos, A.K. Video super-resolution with convolutional neural networks. *IEEE Trans. Comput. Imaging* **2016**, *2*, 109–122. [[CrossRef](#)]
25. Wang, L.; Guo, Y.; Lin, Z.; Deng, X.; An, W. Learning for video super-resolution through HR optical flow estimation. In Proceedings of the Asian Conference on Computer Vision, Perth, WA, Australia, 2–6 December 2018; Springer: Cham, Switzerland, 2018; pp. 514–529.
26. Liu, D.; Wang, Z.; Fan, Y.; Liu, X.; Wang, Z.; Chang, S.; Huang, T. Robust video super-resolution with learned temporal dynamics. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2507–2515.
27. Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; Shi, W. Real-time video super-resolution with spatio-temporal networks and motion compensation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4778–4787.
28. Tao, X.; Gao, H.; Liao, R.; Wang, J.; Jia, J. Detail-revealing deep video super-resolution. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4472–4480.
29. Huang, Y.; Wang, W.; Wang, L. Bidirectional recurrent convolutional networks for multi-frame super-resolution. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 235–243.
30. Sajjadi, M.S.; Vemulapalli, R.; Brown, M. Frame-recurrent video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6626–6634.
31. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep back-projection networks for super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1664–1673.
32. Lucas, A.; Lopez-Tapia, S.; Molina, R.; Katsaggelos, A.K. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Trans. Image Process.* **2019**, *28*, 3312–3327. [[CrossRef](#)] [[PubMed](#)]
33. Zhang, Z.; Sze, V. FAST: A framework to accelerate super-resolution processing on compressed videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 19–28.
34. Jo, Y.; Oh, S.W.; Kang, J.; Kim, S.J. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3224–3232.
35. Bao, W.; Lai, W.S.; Zhang, X.; Gao, Z.; Yang, M.H. MEMC-Net: Motion Estimation and Motion Compensation Driven Neural Network for Video Interpolation and Enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 933–948. [[CrossRef](#)] [[PubMed](#)]
36. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
37. Bertasius, G.; Torresani, L.; Shi, J. Object detection in video with spatiotemporal sampling networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 331–346.
38. Zhang, Y.; Shi, L.; Wu, Y.; Cheng, K.; Cheng, J.; Lu, H. Gesture recognition based on deep deformable 3D convolutional neural networks. *Pattern Recognit.* **2020**, *107*, 107416. [[CrossRef](#)]
39. Xiang, X.; Tian, Y.; Zhang, Y.; Fu, Y.; Allebach, J.P.; Xu, C. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3370–3379.
40. Wang, H.; Su, D.; Liu, C.; Jin, L.; Sun, X.; Peng, X. Deformable non-local network for video super-resolution. *IEEE Access* **2019**, *7*, 177734–177744. [[CrossRef](#)]
41. López-Tapia, S.; Lucas, A.; Molina, R.; Katsaggelos, A.K. Gated Recurrent Networks for Video Super Resolution. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 24–28 August 2020; pp. 700–704.
42. Huang, Y.; Wang, W.; Wang, L. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1015–1028. [[CrossRef](#)] [[PubMed](#)]
43. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
45. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [[CrossRef](#)]
46. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.

47. Liu, C.; Sun, D. On Bayesian adaptive video super resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 346–360. [[CrossRef](#)] [[PubMed](#)]
48. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In Proceedings of the NIPS-W, Long Beach, CA, USA, 4–9 December 2017.
49. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
50. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; Volume 30, p. 3.
51. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]