ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

Recent Advances on Video Super Resolution

Zirui Liu

Shang Hai Jiao Tong University, Shang Hai, China

Abstract. Super-resolution(SR) reconstruction is a technique which constructs high-resolution(HR) images/videos from low- resolution(LR) images/videos. Recently super resolution tech- niques has been prospering rapidly and being applied in various areas. In this article we look through recent works which made great contributions to SR, mainly on VSR. We will compare different works' effectiveness and their unique structures, dis- cussing their pros and cons, trying to find out the reasons why some SR approaches perform significantly better than the others and how they can be applied to help further improve the best performance we can achieve.

Keywords: Index Terms — Image Super-resolution, Video Super-resolution, Space-time Video Super-resolution.

1. Introduction

Super resolution can be of great use in many aspects of our lives, such as HDTV [28], medical imaging [29], [30], satellite imaging [31], face recognition [32], surveillance [33] and so on. Researches on image super resolution are mostly focused on improving the quality of SR pictures and at the same time reducing its computational cost to make it more practical in real life. Video super resolution on the other hand involves more consideration. For a video to be fluent enough for audience to watch, it should have a good real- time performance of at least 24 fps. We all may have seen videos after super-resolution reconstruction, but they are all made after long time of training and processing. So in the field of video SR, what we need to achieve is both effectiveness and efficiency. That is to say, we want to achieve better quality while at the same consuming less resources. Instead of running on several latest GPUs, we want video SR can be processed on small devices with little computing capability, like mobile phones. Video super-resolution, unlike image super-resolution, is definitely not simply overlap of every frame's restoration in the video. Both spatial and temporal dependency relationships should be fully exploited to make restored videos having a better performance. Through we haven't achieved a satisfying result for everyone, luckily many great researchers have made great contributions to the advance of this area.

As to image super-resolution, it is no exaggeration to say that it's the foundation of video super resolution. Currently many researches focus on making models for SISR(Single Image Super-Resolution) training small and takes less com- putation. There are a few very classic models [25]–[27] which guide many following newly invented models through darkness. MobileNet already has three big versions till now. MobileNet along with its all kinds of versions can be of further implementation in all kinds of tasks, like object detection or semantic segmentation. They can mainly work on mobile CPUs with great efficiency, proving their practicality in various works. [11] which improves its own structure to have bet- ter performance with significantly lower computational cost. There are also some methods [8] which further outperforms FSRCNN with the use of Collapsible Linear Blocks and residual connections.

On the other hand, video super-resolution generally is diffi- cult in that researchers need to extract complicated information from video frames which has not been aligned. There are typically two approaches for this challenge. The first one is sliding-window framework which makes use of short-time windows with several frames to reconstruct every frame in the video. The second one is recurrent framework. It is employed to discover the dependency relationships within a long time period. Although the recurrent models have made a more concrete model than those sliding window ways, their problems in long-term information transmissions and feature alignments can not be neglected.

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

As was proposed in [5], most existing video super-resolution approaches can be roughly divided into four parts, which is propagation, alignment, aggregation, and upsampling. These four parts BasicVSR has set made a basic pipeline for future video super-resolution works, since every aspect of this struc- ture can be split out to make improvements on performance. In this pipeline, propagation is used to propagate features temporally, alignment matters greatly on the spatial trans-formation applied to misaligned features. While aggregation is employed to to combine all the aligned features together, upsampling is the final step used to generate high resolution images with aggregated features. While BasicVSR++ [6] is a state-of-art method making significant improvements based on BasicVSR's work. BasicVSR++ made further improvements by applying second-order grid propagation and flow-guided deformable alignment to enhance recurrent network. With these improvements, BasciVSR++ improves BasicVSR by 0.82 dB in PSNR without adding complexity in computation.

TDAN [1] works on the feature level to align the refer- ence frame and the supporting frames, in conjunction with deformable convolutions [4] to dynamically predict sampling parameters, thus gaining the ability of handling videos with large motions that can't be dealt with previously. TDAN, due to its novel and effective architecture, became a model which inspires many later approaches, like EDVR [3] and Zoooming Slow-Mo [2].

EDVR made further improvements on TDAN. EDVR de- vises a Pyramid, Cascading and Deformable (PCD) alignment module and a new Temporal and Spatial Attention(TSA). fusion module. These two new modules work together to align features and emphasize them more effectively.

Zooming Slow-Mo [2] aims at generating high-resolution slow-motion video from low frame rate and low resolution video. In this approach, it also takes advantage of EDVR's PCD module for better alignment of features. Compared to current two-stage strategy for STVSR like DAIN [7] + EDVR, Zooming Slow-Mo is not only three times faster, but also better at handling large motions and therefore stores more information to restore more accurate images with sharper edges.

Table 1. quantitative comparison (psnr/ssim).

			BI degradation			BD degradation		
	Params (M)	Runtime (ms)	REDS4 [34]	Vimeo-90K-T [35]	Vid4 [36]	UDM10 [37]	Vimeo-90K-T [35]	Vid4 [36]
Bicubic	-	-	26.14/0.7292	31.32/0.8684	23.78/0.6347	28.47/0.8253	31.30/0.8687	21.80/0.5246
FRVSR [22]	5.1	137	-	-	-	37.09/0.9522	35.64/0.9319	26.69/0.8103
DUF [23]	5.8	974	28.63/0.8251	-	-	38.48/0.9605	36.87/0.9447	27.38/0.8329
RBPN [13]	12.2	1507	30.09/0.8590	37.07/0.9435	27.12/0.8180	38.66/0.9596	37.20/0.9458	-
TDAN [1]	-	-	-	-	26.42/0.789	-	-	26.86/0.814
EDVR-M [3]	3.3	118	30.53/0.8699	37.09/0.9446	27.10/0.8186	39.40/0.9663	37.33/0.9484	27.45/0.8406
EDVR [3]	20.6	378	31.09/0.8800	37.61/ 0.9489	27.35/0.8264	39.89/0.9686	37.81/0.9523	27.85/0.8503
MuCAN [19]	-	-	30.88/0.8750	37.32/0.9465	-	-	-	-
RSDN [12]	6.2	94	-	-	-	39.35/0.9653	37.23/0.9471	27.92/0.8505
RRN [24]	3.4	45	-	-	-	38.96/0.9644	-	27.69/0.8488
BasicVSR [5]	6.3	63	31.42/0.8909	37.18/0.9450	27.24/0.8251	39.96/ 0.9694	37.53/0.9498	27.96/0.8553
IconVSR [5]	8.7	70	31.67/ 0.8948	37.47/0.9476	27.39/ 0.8279	40.03/ 0.9694	37.84/ 0.9524	28.04/ 0.8570
BasicVSR++ [6]	7.3	77	32.39/ 0.9069	37.79/ 0.9500	27.79/ 0.8400	40.72/ 0.9722	38.21/ 0.9550	29.04/ 0.8753

All results are calculated on y-channel except reds4 [34] (rgb-channel). Red and blue colors indicate the best and the second -best performance , respectively. The runtime is computed on an lr size of 180)320. a 4) upsampling is performed following previous studies. Blanked entries correspond to results not reported in previous works .

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

Table 2. Components in current vsr methods. We categorize components based on their

		Propagation	ALignment	Aggregation	Upsampling
Sliding-Window	EDVR [3]	Local	Yes(DCN)	Concatenate + TSA	Pixel-Shuffle
	MuCAN [19]	Local	Yes(correlation)	Concatenate	Pixel-Shuffle
	TDAN [1]	Local	Yes(DCN)	Concatenate	Pixel-Shuffle
Recurrent	BRCN [?]	Bidirectional	No	Concatenate	Pixel-Shuffle
	FRVSR [22]	Unidirectional	Yes(flow)	Concatenate	Pixel-Shuffle
	RSDN [12]	Unidirectional	No	Concatenate	Pixel-Shuffle
	BasicVSR [5]	Bidirectional	Yes(flow)	Concatenate	Pixel-Shuffle
	IconVSR [5]	Bidirectional(coupled)	Yes(flow)	Concatenate+ Refill	Pixel-Shuffle
	BasicVSR++ [6]	Second-Order Grid Propagation	Yes(Flow-guided deformable alignment)	Concatenate	Pixel-Shuffle

functionalities: i) propagation refers to the way in which features are propagated temporally, ii) alignment concerns on the spatial transformation applied to misaligned images/features , iii) aggregation defines the steps to combine aligned features , and iv) upsampling describes the method to Transform the aggregated features to the final output image. Bolded texts correspond to designs that were reported to Achieve better performance in the literature.

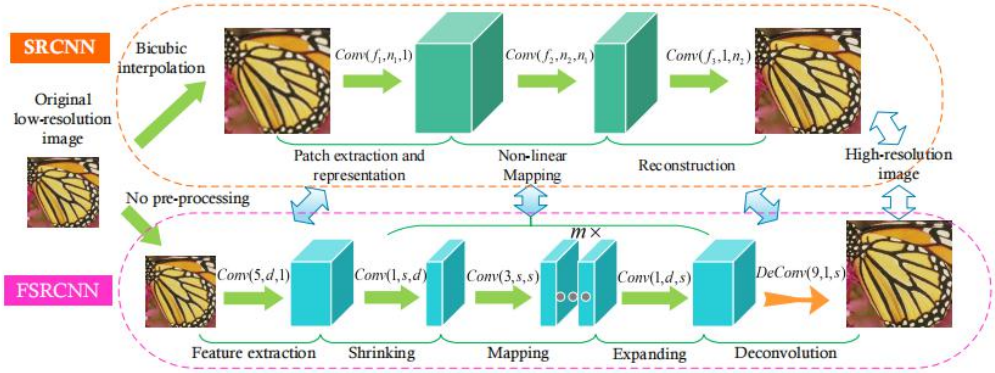

Fig.1 Original low-resolution image

2. Evaluations and comparisons

There are many judging quality metrics that is being used nowadays, like MSE, PSNR, MSSIM(SSIM), UQI and Sarnoff. But normally only the former three methods are used frequently.

2.1 Mse

Mse, known as the mean squared error, is the simplest and most widely used full-reference quality metric. MSE is computed through doing the average value of squared intensity differences of distorted and reference image pixels. The MSE for an image is computed in the following equation:

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

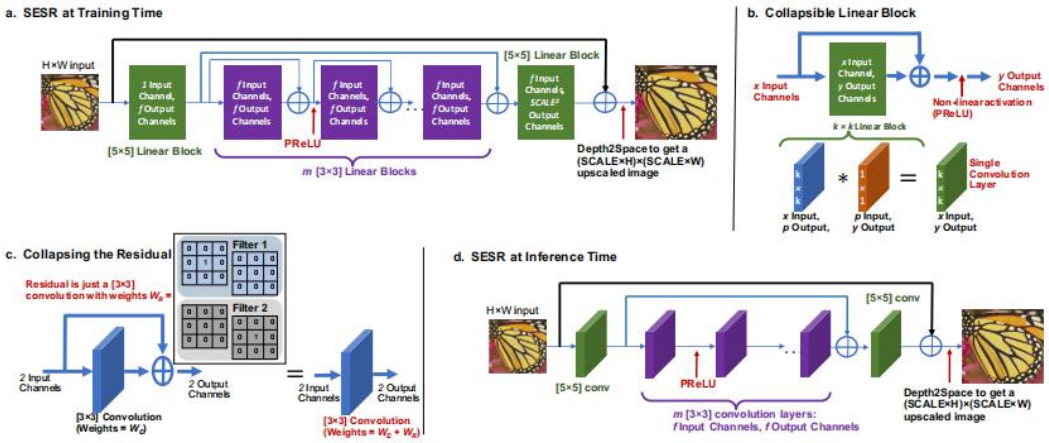

Fig. 2. (a) Proposed at training time contains two 5) 5 and m 3) 3 linear blocks. Two long residuals and several short residuals over 3) 3 linear blocks are applied. (b) A k) k linear block first makes use of a k) k convolution to project x input channels to p intermediate channels. Then these channels are projected back to y output channels via a 1) 1 convolution. (c) Short residuals can be collapsed into convolutions. (d) At inference time just contains two long residuals and m+2 narrow convolutions.

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

Though MSE is mathmatically easy to compute, but the effect of its evaluation is not good.

2.2 Psnr

Psnr, known as peak signal-to-noise ratio, is evaluated based on the misses between pixels. Due to its relative reliability compared to MSE and its convenience in computing, it has been to most widely used method in image quality evaluation. But the weakness of PSNR lies in that it doesn't take human visual system into account, so the result of PSNR sometimes are not matched to human sense visual quality.

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

2.3 Ssim

The inner principle of how evaluation works decides how it performs. Images we see in our daily life are basically highly structured since their pixels tend to have relative strong dependencies on near ones. These dependencies carry large amounts of information necessary for reconstructing images. But approaches like PSNR and MSE evaluate images are based on Minkowski error metrics, which focuses on calculating pixel-level signal differences, can miss so much when dis- cussing evaluation of structure similarity.

SSIM is proposed in order to solve the problem of incon- sistency between evaluation results and human visual sense. This approach is based on the assumption that human's visual system highly relies on extracted structural information from the viewing field.

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2\right)^{\frac{1}{2}} \tag{3}$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \tag{4}$$

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)\,(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

In equation (10), μx is the mean value of x, μy is the mean value of y;

So in the consideration of both accuracy and efficiency, we nowadays tend to combine PSNR with SSIM in evaluation.

3. Different architectures and their characters

3.1 Methods Used In Image Super-resolution

Image super-resolution, though quite not like video super- resolution, some of its technologies can be implemented in VSR for processing key frames. So it's still necessary to discuss some base stone methods being used. Due to SR- CNN's [9], [10] high computational cost, FSRCNN made some improvements in redesigning its structure. Firstly, they take the original low-resolution image as input instead of being processed by bicubic interpolation. Secondly, FSRCNN replaces SRCNN's non-linear mapping module by shrinking, mapping and expanding. Thirdly, FSRCNN takes a smaller filter size and a deeper network structure. Finally, They introduced a deconvolutional layer at the end of the network to do upsampling. Thus by improving the limitations of SRCNN, a more efficient model for SR is devised. By re-designing the model, FSRCNN achieves an acceleration of more than 40 times without weakening its restoration quality.

Another model which achieves great success is Collapsible Linear Blocks for Super-Efficient Super Resolution. [8]

The SESR model, making use of Collapsible Linear Blocks model, is proposed based on the observation that current state- of-art Single Image Super Resolution(SISR), which are mainly based on Convolutional Neural Networks(CNNs). But CNNs has a serious problem of being computationally very expen- sive. For example, the number of Multiply-Accumulate(MAC) operations needed to perform upscaling is typically large. Their experiments show that even FSRCNN, when being used in 100% utilazation, would achieve only 37 FPS on a 4- TOP/s NPU. While other large deep networks would lead to completely impractical situations which achieves no more than 3 FPS. To tackle these problems, SESR decides to make use of linear overparameterization blocks which have not been proposed on super-resolution problems beforehand.

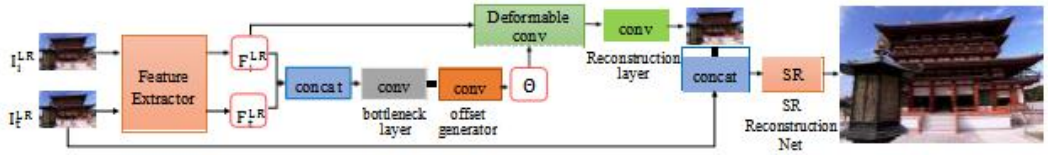

Fig. 3. TDAN framework

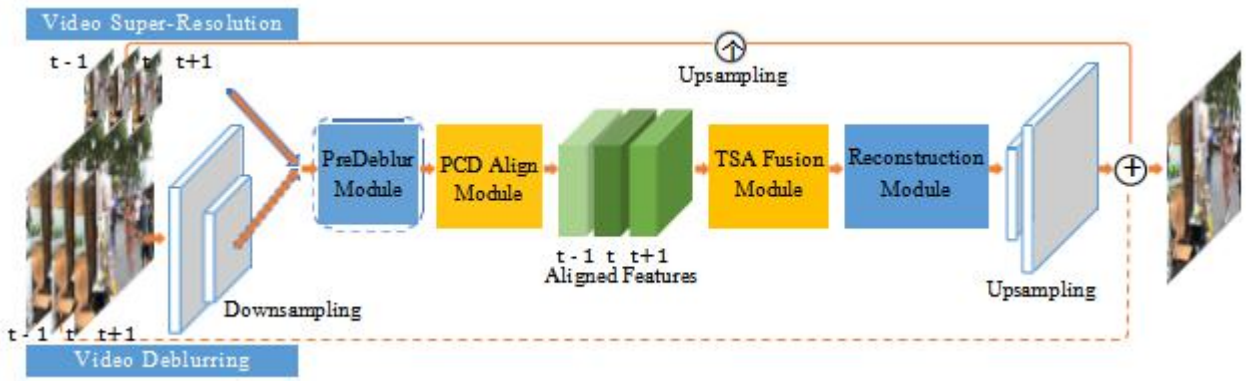

Fig. 4. EDVR framework

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

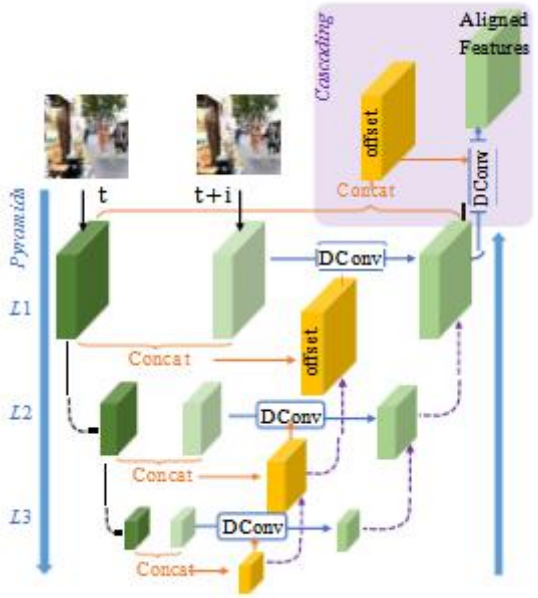

Fig. 5. PCD alignment for EDVR

.

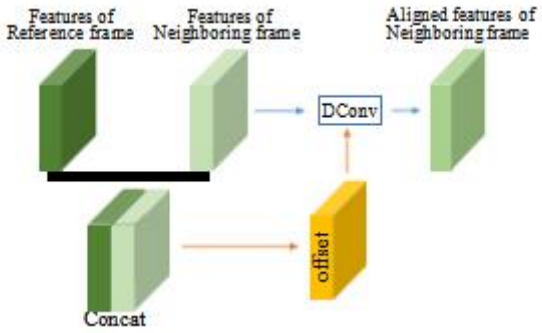

Fig. 6. DCN for alignment for EDVR

ExpandNets's [14] and ACNet's [15] success in overparam- eterizing a convolutional layer have shown the effectiveness of it in accelerating the train of convolution networks. So SESR develops Collapsible Linear Block to take effect of both overparameterization and residual connections for better convergence and restoration quality for SESR tasks.

Despite of the great performance SESR has achieved, there are still many model compression approaches can be imple- mented on SESR. There are many effective methods which will further reduce the resources needed for computation on small devices like (1) [42]–[46]which explores attention mechanism to find the most informative region for the best-quality reconstruction; (2) Knowledge distillation [38] used to transfer knowledge from big teacher networks to small student network [39]; (3) Combining lightweight residual blocks with variants of group convolution [47]–[49]. These methods, since they are orthogonal to SESR's compact structure, can be implemented jointly to improve the performance.

These proposed superior models can be of important usage in real-time video super-resolution, also inspiring new models in this area.

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

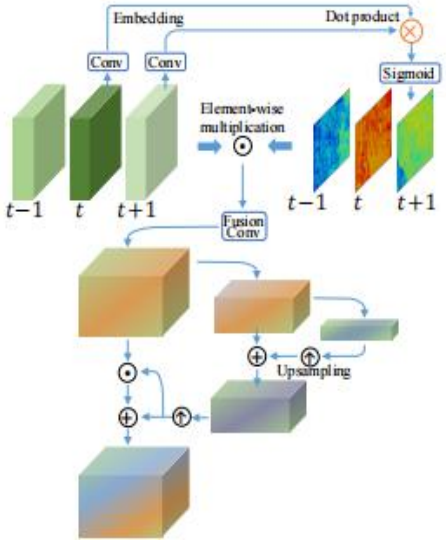

Fig.7. TSA fusion for EDVR

3.2 Methods Used In Video Super-resolution

3.2.1 Methods that employ sliding-windowframeworks

Previ- ously, sliding-window approaches typically utilize optical flow between the reference frame and every supporting frame and warp the supporting frames to achieve temporal alignment. But a severe problem lies in this procedure. Due to the accuracy of common processing procedure, both inaccurate optical-flow and warping strategy may lead to serious problems within the warped supporting frames. On the other hand, precise optical- flow estimation is quite time-consuming and performs badly in front of REDS. So later some more sophisticated designs were invented for implicit alignment instead of traditional optical flow alignment. An outstanding model is TDAN [1]. Its advancement not only exists in the fact that it achieves a new state-of-the-art result in video super-resolution, but also inspires many approaches coming later, such as EDVR [3] and Zooming-Slow-Mo [2].

TDAN works on the feature level to adaptively align the reference frame and the supporting frames, abandoning old optical flow-based approaches.It also got inspired by de- formable convolutions [4] to dynamically predict sampling parameters, thus gaining the ability of handling situations with large motions.

As the figure has proposed, TDAN mainly is a two level framework. The first level is a temporally-deformable align- ment network(TDAN) aiming at aligning every supporting frames with the reference frame. The second level is a super- resolution reconstruction network to generate HR frame. The first level also composed of three modules: feature extraction, deformable alignment and aligned frame reconstruction. As experiments have showed, the adaptively-learned offsets can implicitly capture motion information and explore neighboring features for better alignment within the same image structure. The SR reconstruction network also contains three levels, tem- poral fusion, nonlinear mapping, and HR frame reconstruction respectively. They also employ a sub-pixel convolution [18] as upscaling layer.

Though TDAN has so much encouraging results, it's limita- tion lies in its light-weight architecture with only 1.9 million parameters. Thus TDAN could fail in some cases where very deep SISR networks like RCAN can fully do the job.

Enlightened by TDAN's outcome, an approach called EDVR [3] made some improvements based on the former one's work. EDVR won the champion of NTIRE19 Chal- lenge and outperforms the second place with huge gap. This challenging benchmark is named REDS. Two new aspects of difficulties are raised: the first one is alignment of frames with large motions, the second one is fusion of different frames with diverse motion and blur. To deal with these problems, EDVR on one hand devises a Pyramid, Cascading and Deformable (PCD) alignment module and

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

take advantage of deformable convolutions to laign features in a course-to- fine manner. On the hand, EDVR propose a new Temporal and Spatial Attention(TSA) fusion module to focus attention both temporally and spatially, achieving emphasis of important features for restoration.

It is worth mentioning that EDVR can perform in several video restoration tasks, including super-resolution, deblurring, denoising, deblocking and so on. Take video SR as an exam- ple, EDVR takes 2N+1 low-resolution frames as inputs and outputs a high-resolution output. Each neighboring frame is aligned to the reference frame at the feature level by PCD alignment module. The TSA fusion module will then fuses image information of different frames. The fused features will then pass through a reconstruction module, which is a cascade of residual blocks. The upsampling operation is performed at the end of the network to increase the spatial size. The upsampling method EDVR adopted is also sub-pixel shuffling. Finally SR residual images are directly added to upsampled images to get high-resolution frame. To further boost the performance of their model, they adopt a two-stage strategy, cascading another EDVR with shallower depth for the refinement of the ouput frameworks of the first stage.

The application of deformable convolution for alignment is nearly the same as those in TDAN. But EDVR made further improvement on TDAN for better alignment in handling large motions, with a more sophisticated design of PCD module. EDVR also applies bilinear interpolation to do — 2 upsampling. To reduce computational complexity, the actual implement uses three-level pyramid, each layer showed as in the figure for DCN alignment. We can also notice that a subsequent deformable alignment is cascaded to improve the module from course to fine. This PCD module can work without supervision or pretraining like optical flow.

The Fusion with Temporal and Spatial Attention module is raised due to the fact that: 1) different neighbour frames are not equally informative 2 misalignment and unalignment made a bad impact on reconstruction.As a result, dynamic aggrega- tion of neighbouring frames in pixel-level is irreplaceable for effective fusion.

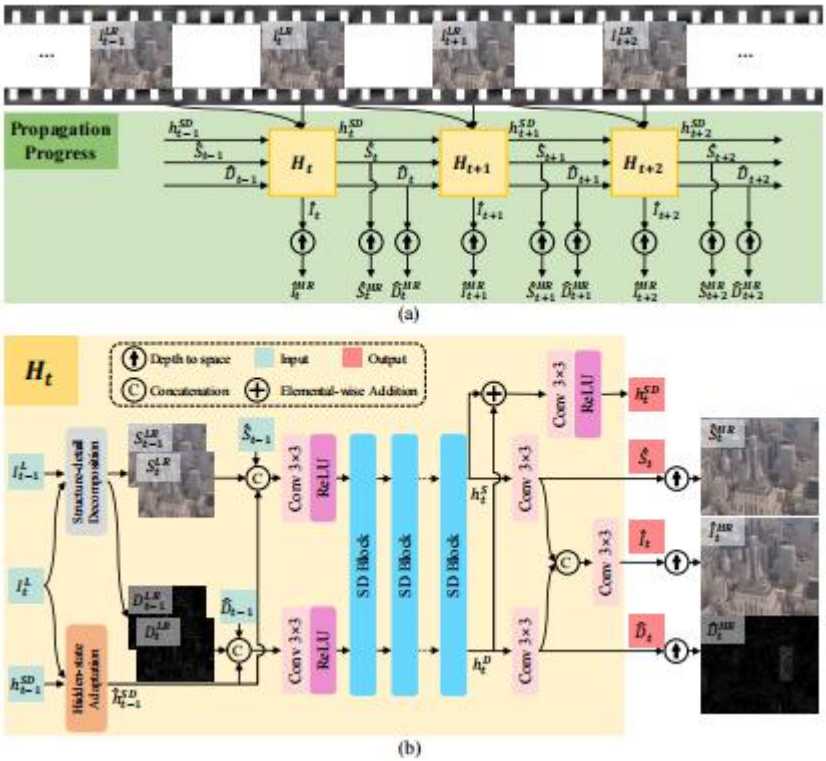

Fig. 8. (a) Pipeline for RSDN; (b)recurrent structure-detail unit

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

3.2.2 Methods that employ recurrent frameworks

Compared to temporal sliding-window approaches, recurrent methods tend to have a more compact structure, which is more efficient in exploring temporal information. (whether recurrent-based methods are more effective still needs to be judged) One method that needs to be speak of is RSDN [12]. It achieves even better performance than EDVR on dataset like Vid4 and UDM10, though EDVR still performs better on Vimeo.

There are some novel structures which blow my mind. Instead of inputting the whole frame to a recurrent network, RSDN disintegrate each frame into two components such as structure component and detail component. Then these two components are processed in structure-detail block module, which is able to ease the problem of vanishing gradient. They also make use of hidden states to adapt to current situation since they think hidden state at time t for example, would conclude previous information, making it better able to describe the relationship between the motion of a certain scene and its temporal switches.

RSDN's design of recurrent structure-detail network is its most shining spot. They think since the structure component contains mostly low-frequency information and motion be- tween frames, while the detail component mainly consists of high-frequency information and changes in temporal appear- ance. These two components are quite different in reconstruc-tion process, while still interact with each other, thus needed to be processed in different ways.

Despite RSDN's some quite amazing new designs,Ba- sicVSR and its extension IconVSR, made further improvement in the model's effectiveness. Based on the fact that current

VSR methods vary greatly and turn to increasing complexity, a baseline should be set to reduce harming on reproducibility and improvement. BasicVSR therefore made a pipeline which made VSR's various methods having something in common. It summarizes existing methods of VSR into four steps: Propagation, Alignment, Aggregation and Upsampling.

BasicVSR made following researches on detailed designs of VSR. About propagation, local-propagation, like sliding- window based approaches, would lose some performance due to lack of consideration of further frames. While in Unidi-rectional propagation, the imbalanced information received in different frames would result in artifacts in propagation. BasicVSR thus adopted Bidirectional Propagation, which is able to solve upper problems. About alignment, BasicVSR adopts optical flow of spatial alignment by warping the images on the feature level. Recently, some researches like [40], [41] have shown the improvement with moving the alignment from image level to feature level. About aggregation and upsampling module, BasicVSR adopted common components of feature concatenation and pixel-shuffle.

Using BasicVSR as backbone, IconVSR further introduced two new components: Information-refill mechanism and cou- pled propagation. IconVSR proposed an additional feature.extractor to extract deep

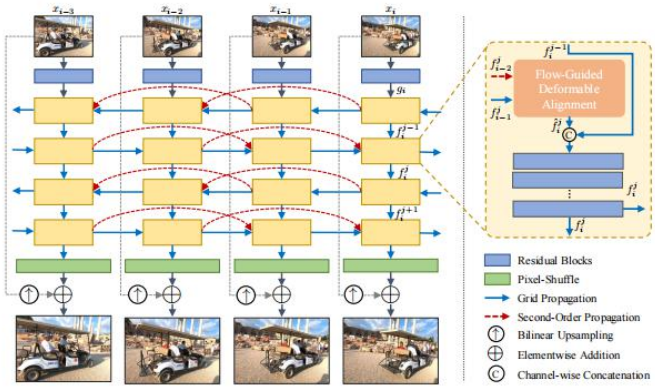

Fig. 9. Overview of BasicVSR++

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

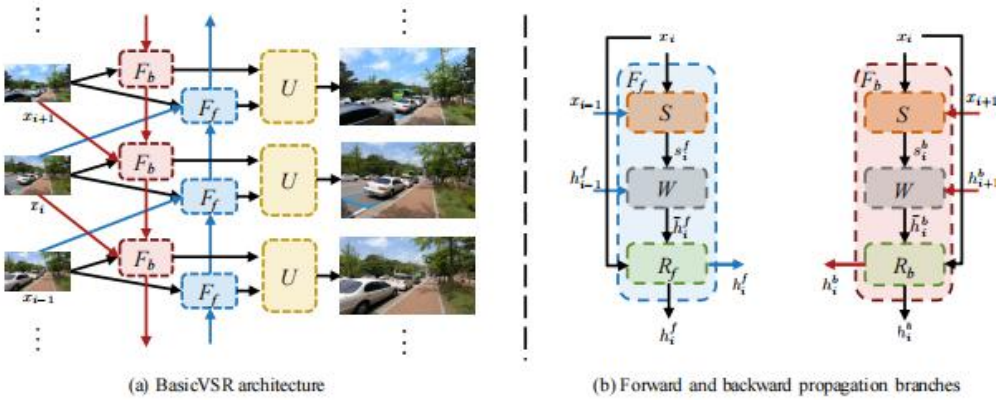

Fig. 10. (a) Structure for BasicVSR; (b) Forward and backward propagation branches for IconVSR

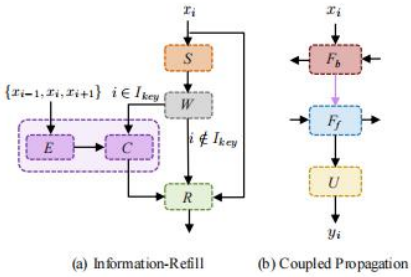

Fig. 11. (a) Information-refill module for IconVSR; (b) Coupled Propagation for IconVSR

information from selected keyframes and fuse into the aligned features $\bar{h}$i :

$$e_i = E(x_{i-1}, x_i, x_{i+1}),$$
$$\hat{h}_i^{\{b,f\}} = \begin{cases} C\left(e_i, \bar{h}_i^{\{b,f\}}\right) & \text{if } i \in I_{key}, \\ \bar{h}_i^{\{b,f\}} & \text{otherwise,} \end{cases} \tag{6}$$

In this equation, E and C refer to the feature extractor and convolution layer respectively. After this process, The refined features will then be passed on to the residual blocks for further process:

$$h_i^{\{b,f\}} = R_{\{b,f\}}(x_i, \hat{h}_i^{\{b,f\}}). \tag{7}$$

The Coupled Propagation is raised due to the reason that propagation modules are inter-connected. So taking backward propagated information as input in forward propagation would intuitively lead to better performance(also confirmed in exper- iments).

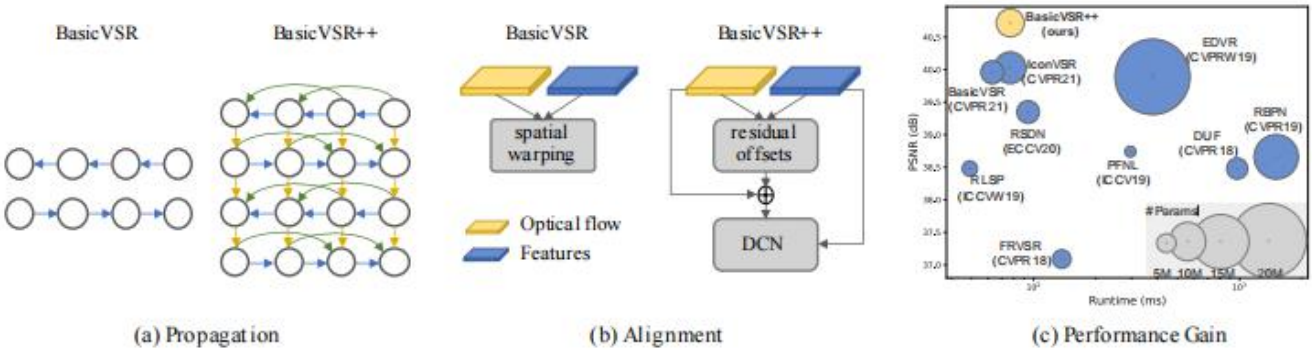

Fig. 12. BasicVSR++'s improvements over BasicVSR

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

$$\begin{aligned} h_i^b &= F_b(x_i, x_{i+1}, h_{i+1}^b), \\ h_i^f &= F_f(x_i, x_{i-1}, h_i^b, h_{i-1}^f), \\ y_i &= U(h_i^f). \end{aligned} \tag{8}$$

BasicVSR's work did paid back. After several months, BasicVSR++ made further improvements by applying second -order grid propagation and flow-guided deformable alignment to enhance recurrent network. With these few improvements, BasicVSR++ surpasses BasicVSR by 0.82 dB in PSNR without changing the order of magnitude of parameters. The normal steps for BasicVSR++ are: For an input video, residual blocks will firstly be implemented to extract features from every frame. Then the second-order grid propagation will propagate features which will be aligned by flow-guided deformable alignment module later. Lastly these aligned fea- tures will be processed in convolution and pixel-shuffling to generate the final images.

Detailed information about BasicVSR++'s two improve- ments will be shown in the following passage: BasicVSR++ got the motivation from the good performance of bidirec- tional propagation, so they noticed the importance of repeated refinement in propagation. To further improve the effective- ness of backward and forward features used in propagation rather than simply taking backward propagated information as input to forward propagation, they devised grid propagation to repeatedly extracts information, thus improving feature effectiveness a lot. This design, shows improving robustness and effectiveness in every sense.

For example, we make xi to be one input image, gi be the extracted feature from xi , and fi be the feature computed at the i-th timestep in the j-th propagation branch. To compute the feature fi, we first align fi-1 and fi-2

$$\hat{f}_i^j = \mathcal{A}\left(g_i, f_{i-1}^j, f_{i-2}^j, s_{i\to i-1}, s_{i\to i-2}\right), \tag{9}$$

where si →i -1, si →i -2 denote the optical flows from i-th frame to the (i-1)-th and (i-2)-th frames, respectively, and A represents flow-guided deformable alignment[1] . Then the features will be passed on to residual blocks:

$$f_i^j = \hat{f}_i^j + \mathcal{R}\left(c\left(f_i^{j-1}, \hat{f}_i^j\right)\right), \tag{10}$$

[1] s0 → -1=s0 → -2=s1 → -1=f-1=f_2=0.

where f_i^0 = gi , 2 denotes the residual blocks, and c denotes concatenation process.

For the Flow-Guided Deformable alignment module, unlike previous methods that directly compute deformable convolu- tional offsets, BasicVSR++ employs flow-guided deformable alignment to take advantage of the offset diversity created from deformable convolution. Although deformable alignment has proved to be significantly more effective than optical flow-based alignment, deformable alignment module may be difficult to train since offset overflow is resulted all the time. This flow-guided deformable alignment module in conclusion have following advantages:

First, by pre-aligning the features using optical flow, the learning offsets of CNN can be assisted.

Second, through optical flow-guided training approach, the burden in handling typical deformable alignment's training instability reduces greatly.

Finally, the modulation masks in deformable convolution network can act as attention maps to weigh the contributions of different pixels, providing additional flexibility.

By ablation study, we can prove that both of these two modules can bring significant improvement in restoration of details. And optical flow-guided deformable alignment can preserve more detailed information. Besides, comparing to sliding-window approaches, recurrent structures have better temporal consistency. All in all, these improvements made BasicVSR++ the new state-of-art method.

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

3.2.3 A new combining way in solving STVSR

There exists another method applied to space-time video super-resolution, which is called Zooming Slow-Mo [2], aiming at generating high-resolution slow-motion video from low frame rate and low resolution video. In this approach, it also takes PCD module in EDVR for deformable alignment. Compared to current two-stage strategy for STVSR like DAIN + EDVR, Zooming Slow-Mo is not only three times faster, but also better at handling large motions and therefore stores more information to restore more accurate images with sharper edges. This module mainly composed of four modules: feature extractor, frame feature temporal interpolation, deformable ConvLSTM and HR frame reconstruction.

They first use the feature extractor to get feature maps as input for the next stage. Then the frame feature temporal interpolation module is based on deformable sampling. The generated F¶ is meant to predict the corresponding HR frame, so it will implicitly let the offsets to capture accurate local temporal information to cope with large motions.

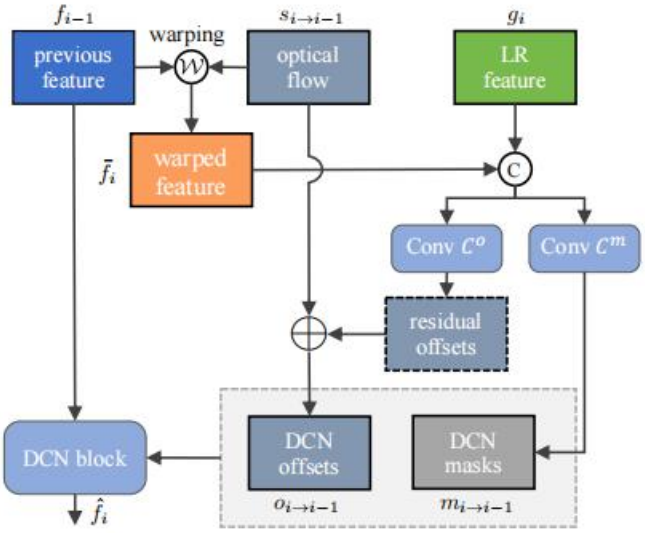

Fig. 13. Flow-guided deformable alignment for BasicVSR++

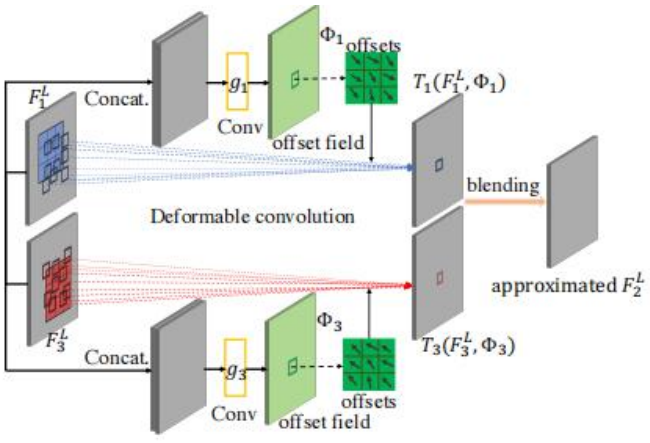

Fig. 14. Frame Feature temporal interpolation for Zooming Slow-Mo

The deformable ConvLSTM is designed based on the idea that except optical flow-based methods, most of the current methods like [3] employs many-to-one architecture for tem- poral alignment. That means they need to process a batch of LR frames to predict only one HR frame. So they come up with the idea that ConvLSTM can effectively ease sequence- to-sequence learning. But vanilla ConvLSTM have the artifact that due to lack of explicit temporal alignment, it will fail in handling large motions. So to tackle this problem, they explicitly add deformable alignment into this ConvLSTM to enhance its ability. Finally, they reconstruct the HR Slow-Mo video sequence

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119
from aggregated feature maps. Totally, this one-stage method for STVSR has four times smaller model size, making it easier and faster to train than other two-stage methods.

4. Future improvements

We can conclude from the upper approaches to find out that in video super-resolution, methods that achieve a high performance tends to employ deformable convolution in han-dling large motions.All the

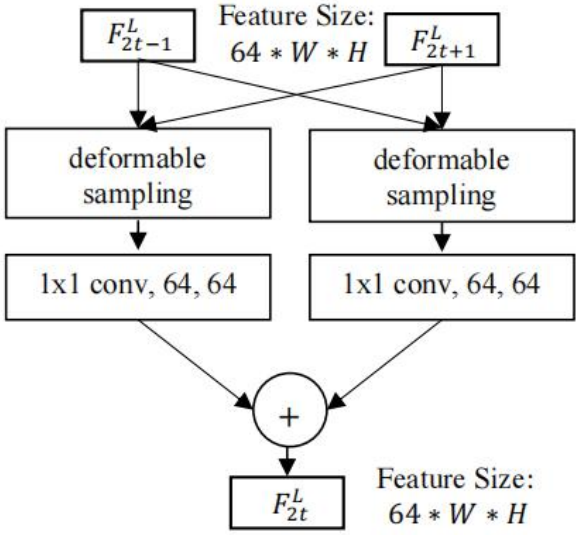

Fig. 15. Feature temporal interpolation for intermediate LR frames in Zooming Slow-Mo

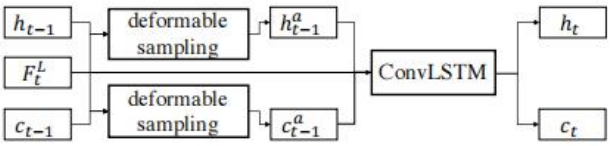

Fig. 16. Deformable ConvLSTM for Zooming Slow-Mo

listed methods that make use of DCN achieve better results than those that don't. But with the introduction of DCN networks, the calculation that needs to be done increases therewith. The DCN based models also can be unstable in training. So when deploying DCN networks, we should trade-off between its pros and cons.

And bidirectional propagation, have proved to be more effective in several works. Furthermore, the works in Ba- sicVSR++ improved propagation by re-devising second-grid propagation, showing that both forward and backward infor-mation are necessary to be made good use of in reconstructing high-quality video frames.

On the aggregation and upsampling side, we recently have not made so much progress. These two aspects may be the breakthroughs in the future.

The future work of super resolution should lies in achiev- ing a high-performance balance between effectiveness and efficiency. Under the same design, networks that contains larger sizes and more parameters usually tend to make better performance. In consequence, the cost of calculation grows rapidly. But in the real world application of super resolution, we normally can't afford a very big price in calculating since practically we want the super-resolution work to be done on the edge side, on traditionally small devices like mobile phones. Online super-resolution is also becoming more and more popular. Deploying online super-resolution instead of off-line super-resolution can better suit the future application of VSR.

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119

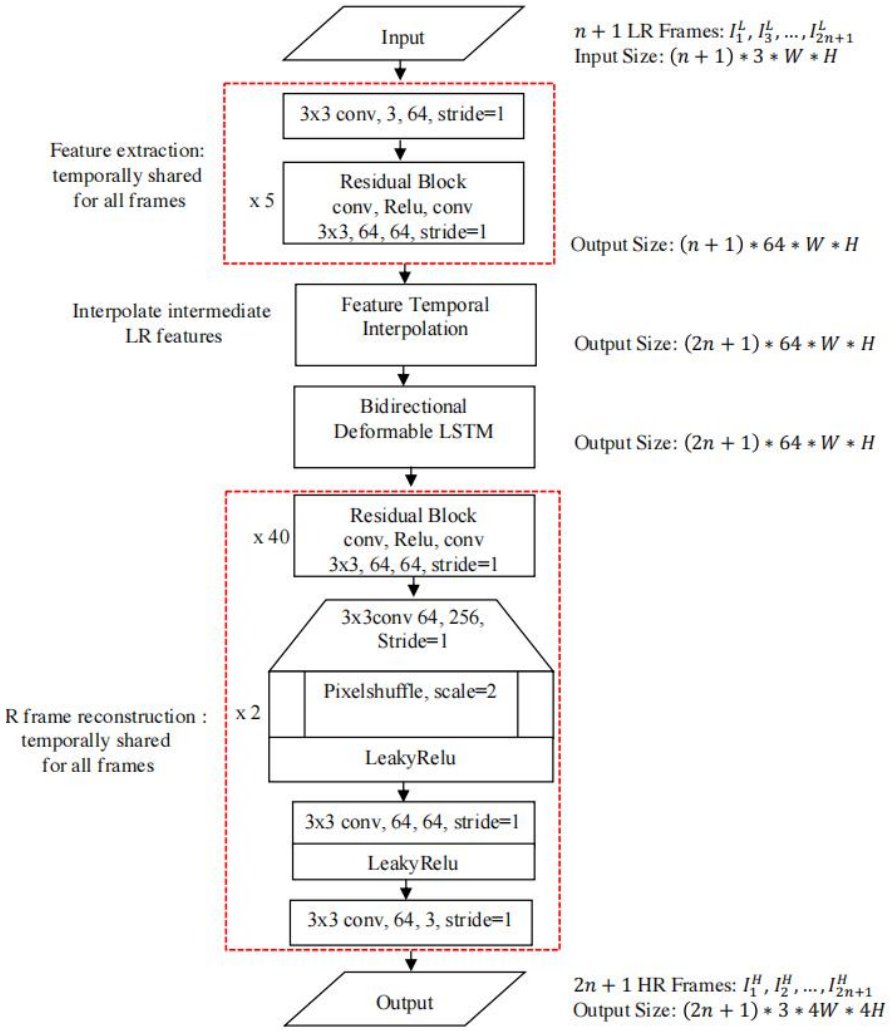

Fig. 17. Flowchart for Zooming Slow-Mo

References

[1] Tian, Yapeng et al. “TDAN: Temporally-Deformable Alignment Net-work for Video Super-Resolution.” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020): 3357-3366.

[2] Xiang, Xiaoyu et al. “Zooming Slow-Mo: Fast and Accurate One-Stage,Space-Time Video Super-Resolution.” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020): 3367-3376.

[3] Wang, Xintao et al. “EDVR: Video Restoration With Enhanced De-formable Convolutional Networks.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019): 1954-1963.

[4] Chan, Kelvin C. K. et al. “Understanding Deformable Alignment inVideo Super-Resolution.” AAAI (2021).

[5] Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen ChangeLoy. BasicVSR: The search for essential components in video super- resolution and beyond. In CVPR, 2021.

[6] Kelvin C. K. Chan et al. ”BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment” ¡i¿arXiv e- prints¡/i¿, 2021.

[7] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao,and Ming-Hsuan Yang. Depth-aware video frame interpolation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3703– 3712, 2019.

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119
[8] Bhardwaj, K., “Collapsible Linear Blocks for Super-Efficient Super Resolution”, ¡i¿arXiv e-prints¡/i¿, 2021.

[9] Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV. (2014) 184–199

[10] Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. TPAMI 38(2) (2015) 295–307

[11] Dong, C, Change Loy, C, and Tang, X., “Accelerating the Super-Resolution Convolutional Neural Network”, ECCV(2016)

[12] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In ECCV, 2020.

[13] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super- resolution. In CVPR, 2019.

[14] Shuxuan Guo, Jose M. Alvarez, and Mathieu Salzmann. Expandnets:Linear over-parameterization to train compact convolutional networks. In Advances in Neural Information Processing Systems, volume 33, pages 1298–1310, 2020.

[15] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet:Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.

[16] Zhou Wang and Alan C. et al. ”Image Quality Assessment: From Error Visibility to Structural Similarity”, IEEE Transl vol. 13, pp. 600–612,2004

[17] Lin, M., Chen, Q., and Yan, S., “Network In Network”, ¡i¿arXiv e-prints¡/i¿, 2013.

[18] Shi, W., “Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network”, CVPR(2016)

[19] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia.MuCAN: Multi-correspondence aggregation network for video super- resolution. In ECCV, 2020.

[20] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convo-lutional networks for multi-frame super- resolution. In NeurIPS, 2015.

[21] Yan Huang, Wei Wang, and Liang Wang. Video super- resolution viabidirectional recurrent convolutional net- works. TPAMI, 2018.

[22] Mehdi S M Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In CVPR, 2018.

[23] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim.Deep video super-resolution network using dy- namic upsampling filters without explicit motion compensa- tion. In CVPR, 2018.

[24] Takashi Isobe, Fang Zhu, and Shengjin Wang. Revisiting temporal modeling for video super-resolution. In BMVC, 2020.

[25] Howard, A. G., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”, ¡i¿arXiv e-prints¡/i¿, 2017.

[26] Sandler, M., Howard, A. Zhu, M.,Zhmoginov, A, and Chen,L.-C., “MobileNetV2: Inverted Residuals and Linear Bottlenecks”, CVPR(2018)

[27] Howard, A., “Searching for MobileNetV3”, ICCV(2019)

[28] T. Goto, T. Fukuoka, F. Nagashima, S. Hirano, and M. Sakurai. Super-resolution System for 4K-HDTV. 2014 22nd International Conference on Pattern Recognition, pages 4453–4458, 2014.

[29] S. Peled and Y. Yeshurun. Superresolution in MRI: application to human white matter fiber tract visualization by diffusion tensor imaging. Magnetic resonance inmedicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine, 45(1):29–35, 2001.

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

ISSN:2790-1688 DOI: 10.56028/aetr.1.1.119
[31] M. W. Thornton, P. M. Atkinson, and D. a. Holland. Sub-pixel map- ping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. International Journal of Remote Sensing, 27(3):473–491, 2006.

[32] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau. Eigenface-domain super-resolution for face recognition. IEEE Transactions on Image Processing, 12(5):597–606, 2003.

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

[34] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon,Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenge on video deblurring and super- resolution: Dataset and study. In CVPRW, 2019.

[35] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. IJCV, 2019.

[36] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution.TPAMI, 2014.

[37] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma.Progressive fusion video super-resolution net- work via exploiting non- local spatio-temporal correlations. In ICCV, 2019.

[38] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the References knowledge in a neural network, 2015.

[39] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In Proceed- ings of the IEEE conference on computer vision and pattern recognition, 2018.

[40] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu,and Yichen Wei. Deformable convolutional networks. In ICCV, 2017.

[41] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. De-formable convnets v2: More deformable, better results. In CVPR, 2019.

[42] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong,and Yun Fu. Image super-resolution using very deep residual channel attention networks. In European Confer-ence on Computer Vision (ECCV), 2018.

[43] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cui- hua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In European Conference on Computer Vision (ECCV), 2020.

[44] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong,Yu Qiao, and Chen Change Loy. Esrgan: En- hanced super-resolution generative adversarial networks. In Proceedings of the European Con- ference on Computer Vi- sion (ECCV) Workshops, pages 0–0, 2018.

[45] Abdul Muqeet, Jiwon Hwang, Subin Yang, Jung Heum Kang, Yongwoo Kim, and Sung-Ho Bae. Multi-attention based ultra lightweight image super-resolution. In European Conference on Computer Vision (ECCV)2020 Workshops, 2020.

[46] Hengyuan Zhao, Xiangtao Kong, Jingwen He, Yu Qiao, and Chao Dong.Efficient image super-resolution using pixel at- tention, 2020.

[47] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate,and lightweight super-resolution with cascading residual network. In Proceedings of the European Conference on Computer Vision (ECCV), pages 252–268, 2018.

[48] Xin Liu, Yuang Li, Josh Fromm, Yuntao Wang, Ziheng Jiang, Alex Mariakakis, and Shwetak N. Patel. Splitsr: An end-to-end approach to super-resolution on mobile devices. CoRR, abs/2101.07996, 2021.

[49] Ying Nie, Kai Han, Zhenhua Liu, An Xiao, Yiping Deng, Chunjing Xu,and Yunhe Wang. Ghostsr: Learning ghost features for efficient image super-resolution. CoRR, abs/2101.08525, 202.