Dual Attention with the Self-Attention Alignment for Efficient Video Super-resolution

Chu, Yuezhong; Qiao, Yunan; Liu, Heng; Han, Jungong

Published in:
Cognitive Computation

DOI:
10.1007/s12559-021-09874-1

Publication date:
2022

Citation for published version (APA):
Chu, Y., Qiao, Y., Liu, H., & Han, J. (2022). Dual Attention with the Self-Attention Alignment for Efficient Video Super-resolution. *Cognitive Computation*, *14*(3), 1140–1151. https://doi.org/10.1007/s12559-021-09874-1

Dual Attention with the Self-Attention Alignment for Efficient Video Super-resolution

Yuezhong Chu[1], Yunan Qiao[1], Heng Liu[1, *], Jungong Han[2]

[1] School of Computer Science and Technology, Anhui University of Technology, 243032, China

[2] Department of Computer Science, Aberystwyth University, SY23 3DB, U. K.

[*] Corresponding author: hengliusky@aliyun.com

Conflict of Interest: Yuezhong Chu declares that he has no conflict of interest. Yunan Qiao declares that she has no conflict of interest. Heng Liu declares that he has no conflict of interest. Jungong Han declares that he has no conflict of interest.

Abstract:

1) Background/Introduction

By selectively enhancing the features extracted from convolution networks, the attention mechanism has shown its effectiveness for low-level visual tasks, especially for image super-resolution (SR). However, due to the spatiotemporal continuity of video sequences, simply applying image attention to a video ~~cannot~~does not seem to obtain good SR results. At present, there is still a lack of suitable attention structure to achieve efficient video SR.

2) Methods

In this work, building upon~~based on the correlation exploration for~~ the dual attention, i.e., ~~–~~position attention and channel attention, we proposed deep dual attention, underpinned by ~~equipped with~~ self-attention alignment (DASAA), for video SR. Specifically, ~~Firstly,~~ we start by constructing a dual attention module (DAM) to strengthen the acquired spatiotemporal features and adopt a self-attention structure with the morphological mask to achieve attention alignment. Then, on top of ~~based on~~ the attention features, we utilize the up-sampling operation to reconstruct the super-resolved video images, and introduce the LSTM (long short time memory) network to guarantee the coherent consistency of the generated video frames both ~~in temporal and spatial domains~~ temporally and spatially.

3) Results

Experimental results and comparisons on the actual Youku-VESR dataset and the typical benchmark dataset- Vimeo-90k demonstrate that our proposed approach ~~not only~~ achieves the best video SR effect while ~~but also~~ taking~~s~~ the least amount of computation. Specifically, in the Youku-VESR dataset, our proposed approach achieves a test PSNR of ~~over PSNR/SSIM metrics is~~ 35.290db and a SSIM of 0.939, respectively. In the Vimeo-90k dataset, the PSNR/SSIM indexes of our approach are ~~individually~~ 32.878db and 0.774. Moreover, the FLOPS (float-point operations per second) of our approach is as low as 6.39G.

4) Conclusions

The proposed DASAA method surpasses all video SR algorithms in the comparison ~~outperforms all the compared algorithms for video SR~~. It is also revealed~~proved~~ that there is no linear relationship between positional attention and channel attention. It

suggests that our DASAA with LSTM coherent consistency architecture may have great potential for many low-level vision video applications.

Keywords: Video super-resolution, Dual attention, Self-attention alignment, FLOPS

1. Introduction

With the emergence of massive deep image SR works, Video SR is slowly coming into focus [3,7,9,10], especially in the field of high-definition video applications. A vast majority of existing video SR methods advocate the use of 3D convolution to extract temporal and spatial information to preserve the sequence characteristics of the video. Despite its promising preliminary results, 3D convolution limits the depth of the network, and the performance gains usually come at the cost of more parameters and heavy computation.

Alternatively, some other video SR methods opt ~~to~~for processing the video frame by frame [9,18] and thus perform SR based on the single image SR method. However, in this case, the consistency between video frames is not guaranteed, ~~and even~~ so that the local and the global feature dependencies cannot be well integrated. Recent works [19,22] attempt to utilize optical flow-based motion compensation ways to alleviate the issue of inter-frame inconsistency. But this will undoubtedly increase a burden on the computation of the entire model.

On the other hand, current deep image SR models always choose residual connections to convey features. The residual connections can reduce the attenuation of the features when they feedforward along the depth direction of the network so that the features can be expressed to any position of the network. Despite its convenience on feature transfer, the residual connection does not fully mine the feature information of different layers across different layers. Thus, instead of a simple residual skip connection, ~~some~~ complex residual variants are proposed, such as DRRN [12], RDN [1], etc. Here, RDN (Residual Dense Network) is the representative of such kinds of variant networks. It uses not only local dense residual learning but also global residual learning to extract and adaptively fuse local features and global features from all observed layers. Since RDN makes full use of multiple hierarchical features in LR images, it does improve the performance of image SR. However, these residual networks only ensure the application of features of different layers without~~, not~~ considering the relationship between different feature channel maps of the same layer. To remedy this situation~~So~~, Zhang et al. [2] proposed a

very deep residual channel attention network (RCAN) for image SR, which demonstrates a pretty good reconstruction effect. For video SR, Wang et al. [3] integrated ~~recommend integrating~~ temporal attention and spatial attention to ensure the continuity of the super-resolved video both in time and space. Moreover, Fu et al. [6] proposed to use two kinds of attention-channel attention and position attention for semantic segmentation, retaining rich contextual information from a global perspective. However, none of the above dual-channel works reveals the correlation between these two channels. Thus, how to use dual attention mechanisms and align them for better video SR performance remains challenging ~~is still a challenge~~.

Fig.1. The details in the super-resolved (4×) frame image generated by the proposed DASAA (rightmost column) are clearer than the ones produced by VDSR [15] (third column) and VSRnet [23] (fourth column); the LR and ground-truth frame images are shown in the left two columns.

~~In view of these issues,~~ ~~I~~In this paper, we first~~ly~~ investigate the correlation between the dual attentions. Based on the investigation, we propose an efficient dual attention mechanism equipped with a self-attention alignment deep model, namely DASAA, for video SR. An example of a super-resolved frame image recovered by the proposed DASAA model for 4× down-sampling is illustrated and compared with VDSR [15] and VSRnet [23] in Fig. 1. Given a video sequence, our model first~~ly~~ uses several residual dense blocks (RDBs) [1] to fully extract the hierarchical features of the sequence. Then, we send these extracted feature maps to the dual attentions—position attention and channel attention module for further feature selection and weighting enhancement. The position attention branch (PAB) can assign weights to different positions of the channel features, and therefore help~~ful~~ to stick out the feature information on certain important positions. While, channel attention branch (CAB) can select and highlight some significant channels according to criteria, thus enhancing the expressive ability of the network. Also, in order to align and fuse these two branch attentions, we take a self-attention structure, which can play the role of interactive learning. Finally, after an up-sampling operation, we use LSTM [17, 24, 25] to capture the motion variation of the inter-frames to guarantee the coherence or consistency of the super-resolved video.

The main contributions of our work are summarized as follows:

- We explore, for the first time, the correlation of the position attention and the channel attention ~~for the first time~~ and propose dual ~~the~~ attentions with self-attention alignment deep network (named DASAA) for video SR. The network ~~uses RDBs to~~ gets the rich hierarchical feature maps from video sequences by means of RDBs and adopts dual attention branches to select and enhance features from position and channel levels to obtain good video SR performance.
- We propose to use the self-attention structure with dual attention morphological mask to effectively align and fuse two branch attentions and simplify the motion compensation with the LSTM structure to guarantee the consistency for video SR.
- We verify and evaluate our DASAA model on the widely recognized public dataset not only with objective image measures but also with computation load index. It is proved that by dual attention with self-attention alignment, our approach can produce competitive results against the state-of-the-art video SR methods on benchmark videos and the actual data. More importantly, our method requires the least amount of floating-point computing pressure, compared with state-of-the-art competitors.

The remainder of this work is organized as follows. Section 2 provides an overview of the related work. Section 3 describes the proposed dual attentions with self-attention (DASAA) deep video SR network in detail. Section 4 presents extensive experimental results with comparative analysis and ablation discussions. Finally, Section 5 concludes the work.

2. Related Works

2.1 Super-resolution

Although image SR is a classic low-level vision task, there are still many new methods emerging in recent years, especially many deep-learning-based methods. Since the advent of SRCNN [14], many early proposed deep SR models adopt the ~~follow its~~ SR process of feature extraction, nonlinear mapping, and image reconstruction. Due to the strong ability of feature passing, many later image SR deep networks integrate the various residual structures [11] to fuse different levels of features. Lim et al. [13] enhanced the conventional deep residual network (EDSR) by removing its unnecessary modules for

image SR. At the same time, Tai et al.[12] presented recursive deep residual network (DRRN) to mine the features belonging to different layers. Recently, Zhang et al. [1] also proposed a residual dense network (RDN) to make use of local dense residual learning with global residual learning for image SR. What's more, Liu et al. [31] introduced the phase congruency edge prior and proposed a multi-scale deep encoder-decoder model for single image SR.

Video is a continuous image frame sequence that contains spatio-temporal information at the same time. Sajjad et al. [9] ~~directly~~ presented a frame-recurrent network, and Haris et al. [10] further utilized a recurrent back-projection network (RBPN) for video SR. Yan et al. [30] proposed a network containing local network and feature-context network, which makes full use of multi-frame LR information and HR frame context information to improve SR performance. Noting that spatio-temporal information is important for video SR, Caballero, et al. [22] proposed a spatio-temporal deep network to mine the spatial and temporal features for real-time video SR. Furthermore, Zhang et al. [16] utilized 3D normalized convolution operations to extract spatio-temporal features from input video clips. Moreover, Li et al. [8] combined several 3D spatiotemporal residual blocks with LR and cross-space residual connection to achieve fast video SR, and Tian et al. [29] used deformable convolutions [27] for temporal alignment, thus ensuring the accuracy of video super-resolution.

2.2 Attention

In addition to feature transfer by residual connection, the attention mechanism is widely used for feature preservation and enhancement in many SR models [2, 3, 6, 7]. Zhang et al. [2] proposed a residual channel attention network (RCAN), which utilized channel attention with residual blocks to adjust the task adaptability of channel features and strengthen their expression ability. Besides, Wang et al. [3] put forward the way of temporal and spatial attention (TSA) fusion in their EDVR model to emphasize the important functions for subsequent video restoration.

The self-attention mechanism is an important improvement ~~of~~ over general attention. Self-attention can reduce the dependence on external information and is better at capturing remote dependencies and internal correlation of data or elements. Want et al. [4] formalized self-attention as a non-local operation to model the spatial-temporal dependencies in video sequences. Inspired by this, Zhang et al. [5] proposed a self-attention GAN to better learn the dependence of global features. Meanwhile, Fu et al. [6] introduced the dual self-attention in the scene segmentation task to adaptively integrate

local features and global dependencies. For stereo image SR, Wang et al. [7] introduced a parallax-attention structure with a global receptive field along the epi-polar line to deal with large parallax changing.

3. Methods

In this section~~Here~~, we will discuss the relationship between position attention and channel attention from the perspective of experimental observations, and then describe the detailed structure of the proposed DASAA model. For ease~~the convenience~~ of expression, we will ~~introduce~~ begin with our network architecture, followed by the study of ~~at first and then analyze~~ the relationship between position attention and channel attention.

3.1 Network Architecture

Our overall network architecture ~~contains~~ consists of five parts: feature extraction, dual attention module (DAM), upscale layer, LSTM layer, and reconstruction part, which is illustrated in Fig. 2. We denote the input low-resolution (LR) video sequence as $\{I_{t-n}^{LR},\ldots,I_t^{LR},\ldots,I_{t+n}^{LR}\}$ that is a LR video input sequence, and the size of each frame image is $M_L \times N_L$; the output of DASAA (I_t^{SR}) is the high-resolution (HR) version of I_t^{LR} , and the size of each frame image of I_t^{SR} is $M_H \times N_H$. Note that, these symbols apply to all the following statements.

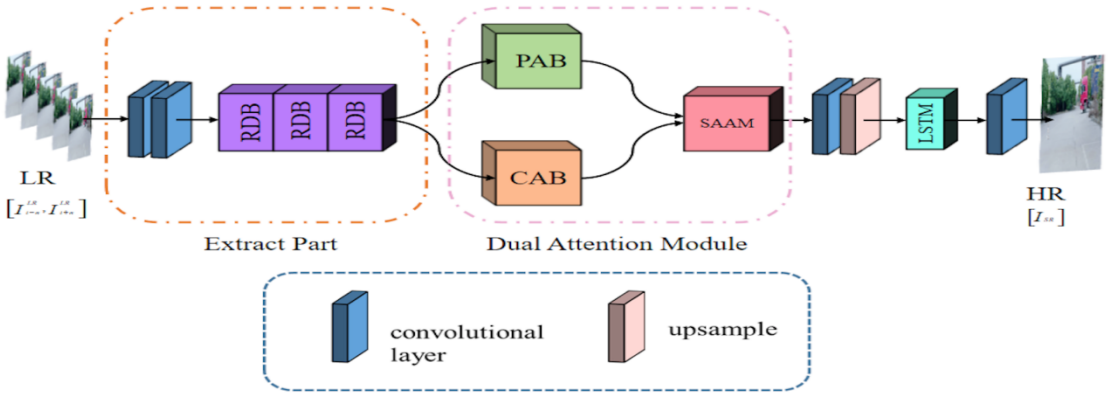


Fig. 2. The architecture of the proposed dual attention with self-attention alignment network for video SR

In the feature extraction part, two convolution layers are used to extract features from the LR input sequence, which can be described as:

$$F=H_C(I^{LR}), \tag{1}$$

where $H_C(\cdot)$ denotes two convolution operations, F represents the extracted features. Then, the features F are sent to residual dense blocks (RDBs) to extract dense feature maps. These dense features extraction can be denoted as:

$$F_D = H_{RDB}(F), \tag{2}$$

where $H_{RDB}(\cdot)$ refers to ~~denotes~~ the operations of residual dense blocks. Then, the dual attention module (DAM) accepts the output feature maps from the feature extraction part and produces the two branch attention outputs, which can be described as:

$$F_{DA} = H_{DAM}(F_D), \tag{3}$$

where F_{DA} is the output of DAM. The attention maps F_{DA} are sent to an upscale layer H_{UP} for resolution improvement and then follow with LSTM structure H_{LSTM} to capture temporal coherence between sequence maps. Finally, the reconstruction result is obtained via a convolutional layer H_R. All the above operations for SR can be described as:

$$I^{SR} = H_R\left(H_{LSTM}\left(H_{UP}\left(H_{DAM}\left(H_{RDB}\left(H_C(I^{LR})\right)\right)\right)\right)\right). \tag{4}$$

From Eq. (4), it is easy to find that our model is concise and compact, which may indicate its computation efficiency. Moreover, for the loss function of the proposed DASAA model, instead of MSE (mean square error) loss, we choose L_1 loss. Given a frame sequence training set $\{I_i^{LR}, I_i^{HR}\}_{i=1}^{N}$, which contains N LR frames and the HR counterparts, the goal of training DASAA is to minimize the L_1 loss function as:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|H_{DASAA,\theta}\left(I_i^{LR}\right) - I_i^{HR}\right\|_1, \tag{5}$$

where θ denotes the parameter set of our network. More details of training are ~~is~~ shown in Section 4.2.

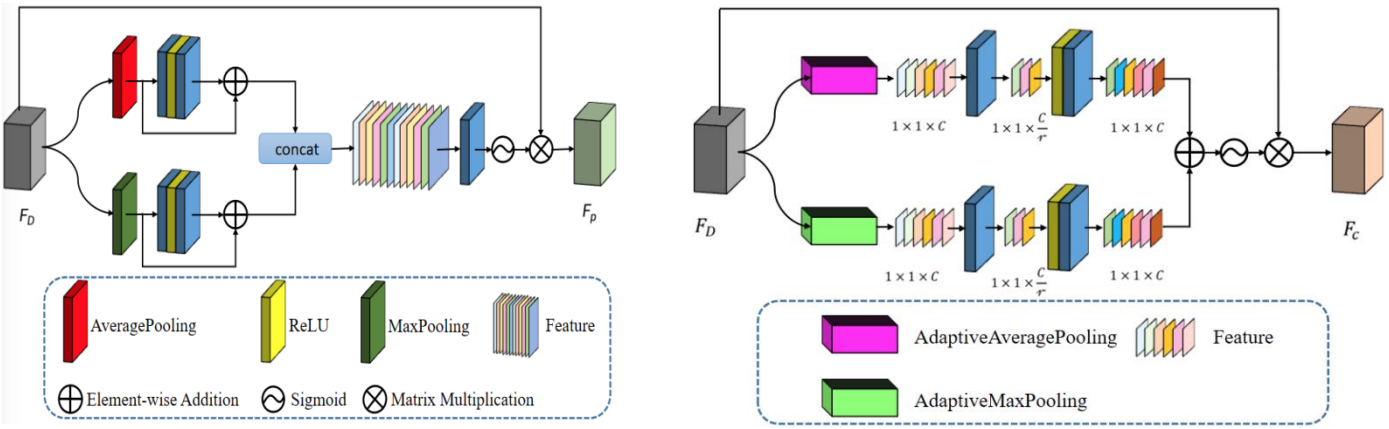
3.2 Position Attention Branch (PAB)

The detailed structure of this part is shown in Fig. 3(a). First, we perform average pooling and maximum pooling on the extracted feature F_D along the channel, and then pass through the residual block, respectively. Such pooling operations can reduce the error of the estimated value and the estimated mean due to the inappropriate convolution parameters, ~~and~~ while retaining more texture and background information. Besides, the residual blocks can convert these features from position-wise to channel-wise. Note that the convolution weights in the two branches' residual blocks are mutually shared. Due to the semantic similarity, a concatenation operation is used to merge the features of the two pooling branches. After a convolution

operation followed by a sigmoid activation, a position attention map is obtained. Then the final position attention output F_P will be achieved by multiplying the attention map with the features input F_D. Thus, the position attention module can be described as:

$$F_p = \mathcal{S}[\mathcal{W}(X_a; X_m)] \cdot F_D, \tag{6}$$

where $\mathcal{S}$ denotes the sigmoid function, $\mathcal{W}$ is the convolution fusion weight of two attention branches outputs X_a and X_m.



(a) (b)

Fig. 3. The detailed architecture of PAB and CAB. (a) PAB: the blue boxes represent filters with a convolution kernel-, the yellow boxes ReLU, the red box average pooling, and the green box max pooling. (b) CAB: the magenta box denotes adaptive average pooling, and the lime box adaptive max pooling, respectively. The signal "⊕" represents element-wise summation, "⊗" matrix multiplication, and "⊘" softmax/sigmoid activation function.

3.3 Channel Attention Branch (CAB)

The detailed structure of the CAB is shown in Fig. 3(b). The input features F_D are firstly sent to the adaptive average pooling layer and the adaptive max-pooling layer, respectively, for feature aggregation along with the spatial location. After adaptive pooling operations, the size of the output becomes $1 \times 1 \times C$. Then motivated by RCAN [2], the channel number C of the features is reduced to $\frac{C}{r}$ (r is reduction scale) first and then rescaled to C by different convolution operations. Obviously, the channel features are selected and weighted. Then, the two branch attention features are directly added to ensure integrity. Finally, the added features pass through the sigmoid function to obtain the channel attention map, which can

be continually multiplied by F_D to get the final channel attention output F_c. Thus, the channel attention module can be described as:

$$F_c = \mathcal{S}(F_a + F_m) \cdot F_D \,, \tag{7}$$

where $\mathcal{S}$ denotes the sigmoid function as same as Eq. (6), F_a and F_m are the two attention branches outputs.

3.4 Relationship analysis on PAB and CAB

To explore whether there is some relationship between the position attention features and the channel attention ones, we design a compact and symmetrical structure for the relationship visualization observation, as shown in Fig. 4. In this structure, PAB and CAB have switched positions because of the different multiplying results of their outputs. We visualize the outputs in different key positions and then find out the relationship between PAB and CAB by observing and comparing the coherence and the difference in these outputs. According to these visualization observations, it is believed ~~we recognize~~ that at least there is no linear correlation between PAB and CAB. For a more detailed discussion, see Section 4.2 (5 position observations).

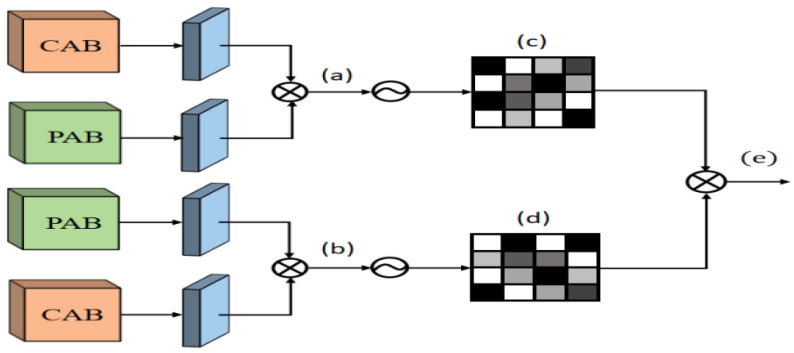


Fig. 4. The structure designed for relationship observation of PAB and CAB

3.5 Self-attention alignment model (SAAM)

Recognizing that PAB and CAB obtain different attention maps for the same input features, and motivated by the relationship analysis results in section 3.4, we use a more complex self-attention structure to align and fuse the two attention maps to acquire multi-faced consistent attention information, which is more conducive to video reconstruction. The detailed SAAM structure is illustrated in Fig. 5.

Similar to the structure in Fig. 4, the architecture of SAAM also contains the upper part and the lower part. In the upper part, the PAB output F_p and the CAB output F_c are individually passed through a 1×1 convolution layer to get the thin feature matrix *P* and *Q*. By a softmax activation, the result of *P* and *Q* multiplication turns to be the attention map M_{pc}. Also, the CAB output is sent to another 1×1 convolution operation to get another feature matrix *S*. When multiplying it with M_{pc} , we get the reference feature map *R*. In the lower part, the positions of PAB and CAB are swapped and their outputs undergo ~~are performed~~ similar operations. And finally, another attention map M_{cp} is acquired.

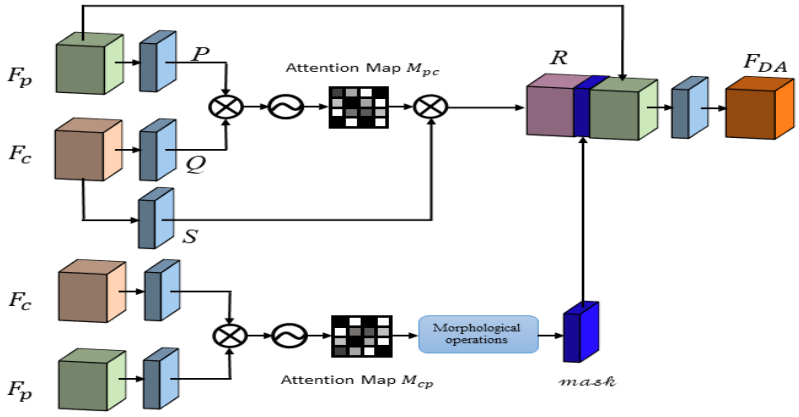


Fig. 5. The detail architecture of SAAM

However, since the pixels of the position attention maps and the pixels in the channel attention maps are difficult to correspond, we take the morphological operation to deal with this problem, which. ~~Which~~ can guide the attention features fusion and play the role of calibration. This is the attention alignment process. Finally, F_p , *R* and the mask are concatenated and convolved to get the output of DAM.

3.6 LSTM inter-frame consistency

As is well known, the recurrent structure can be used to extract and fuse inter-frame information and play a good role in ensuring sequence data coherence and consistency. However, traditional recurrent neural networks (RNN) are always confronted with the problems of gradient disappearance when processing long-time sequence data. Thus, to guarantee the consistency of the super-resolved video frames, we choose LSTM to undertake this task by up-sampling the dual attention fused features. Finally, the output of LSTM will be passed through a convolution layer to get the final reconstructed result.

4. Experimental results and analysis

4.1 Dataset and training details

The video dataset (https://tianchi.aliyun.com/dataset) used in our experiments is borrowed~~provided by~~ Youku 2019 super-resolution competition, consisting of~~. The data set includes~~ scenes of people, animals, landscapes, etc. It is subdivided into the training set, validation set, and test set. Among them, the training set contains 5.25k~~0~~ frames of video images, the validation set about 2.5k frames, and the test set 25k frames. The size of each frame image is 480×270.

Another Vimeo-90k [20] dataset is a public dataset widely used for training and test of SR task. In our experiments, ~~not only~~both the Youku test dataset and ~~but also~~ Vimeo-90k ~~is~~are fully utilized as evaluation data. Besides, PSNR and SSIM measures are treated as the evaluation criteria. During training, five consecutive frames of each video clip are taken as input to the model. The learning rate is set to 1×10^{-4}. The batch size is set to 64. And the PyTorch framework is used to implement the experiments with an RTX 2080Ti GPU. The source code of the work is available at https://github.com/qynan/DASAA.

4.2 Visualization experiments

As mentioned in Section 3.4, we are concerning whether there is some relationship between the position attention features and the channel attention ones and whether this relationship would affect the attention fusion maps, and further affect the performance of the entire SR. Thus, to find out whether there is a mutual influence between such two attentions, we set five observation spots in the route of the relationship analysis structure (see Fig. 4). The visualization results are illustrated in Fig. 6. Seen f~~F~~rom the results, there is no obvious linear relationship between the two attentions. When used in parallel, the contribution of each attention is unique and does not affect each other.

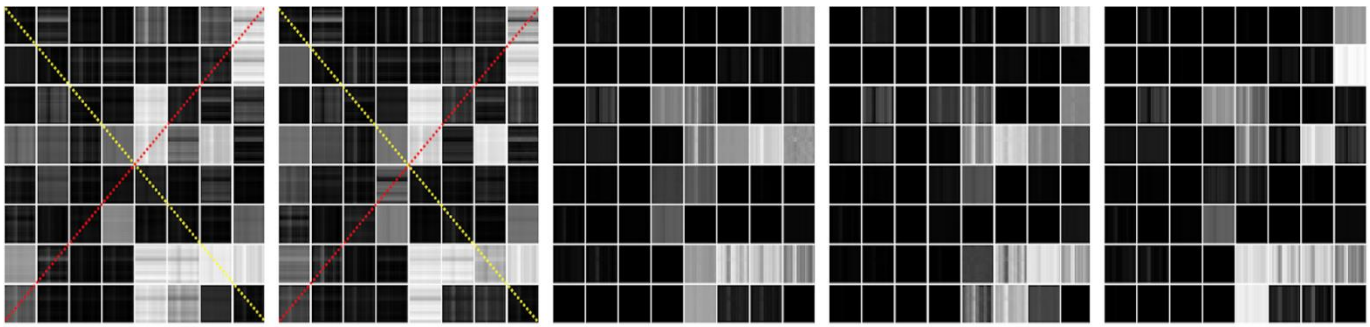
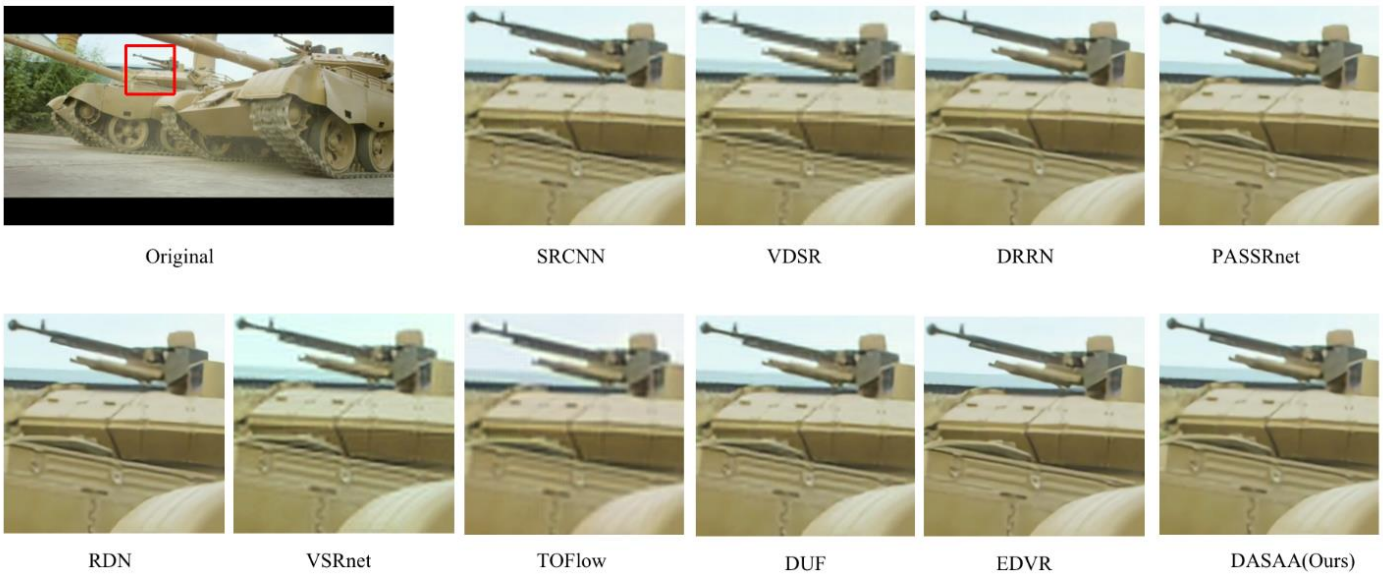


Fig. 6. Visualization results of five observation spots. The subfigures of (a), (b), (c), (d), and (e) correspond to those markers located in Fig. 4. In each Fig., the red dotted line in the positive diagonal indicates a positive correlation and the yellow dotted line in the negative diagonal reflects a negative correlation. It can be seen clearly from these figures that there is no obvious correlation relationship between such two attentions.

4.3 Comparisons and discussions

We compared the proposed method with different state-of-the-art SR methods, both qualitatively and quantitatively, including some image SR methods, such as SRCNN [14], VDSR [15], RDN [1], DRRN [12], and video SR methods, such as VSRnet [23], EDVR [3]. The visual qualitative results from the Youku-VESR dataset and Vimeo-90k dataset are shown~~compared~~ in Fig. 7 and Fig. 8, which display the original frame and the SR results at x4 magnification. Besides, we choose the PSNR and SSIM metrics as our evaluation indices. The corresponding quantitative comparisons under the Youku-VESR dataset and the Vimeo-90K dataset are listed in Table I and Table II, respectively. From Fig. 7, Fig. 8, it is clear that our proposed DASAA approach can produce a breezy improvement in restoring the edge and texture details. Meanwhile, according to Table I and Table II, the PSNR and SSIM indexes of our proposed approach are the best or the second-best compared with other cutting-edge methods, which highlights the remarkable performance of our method.



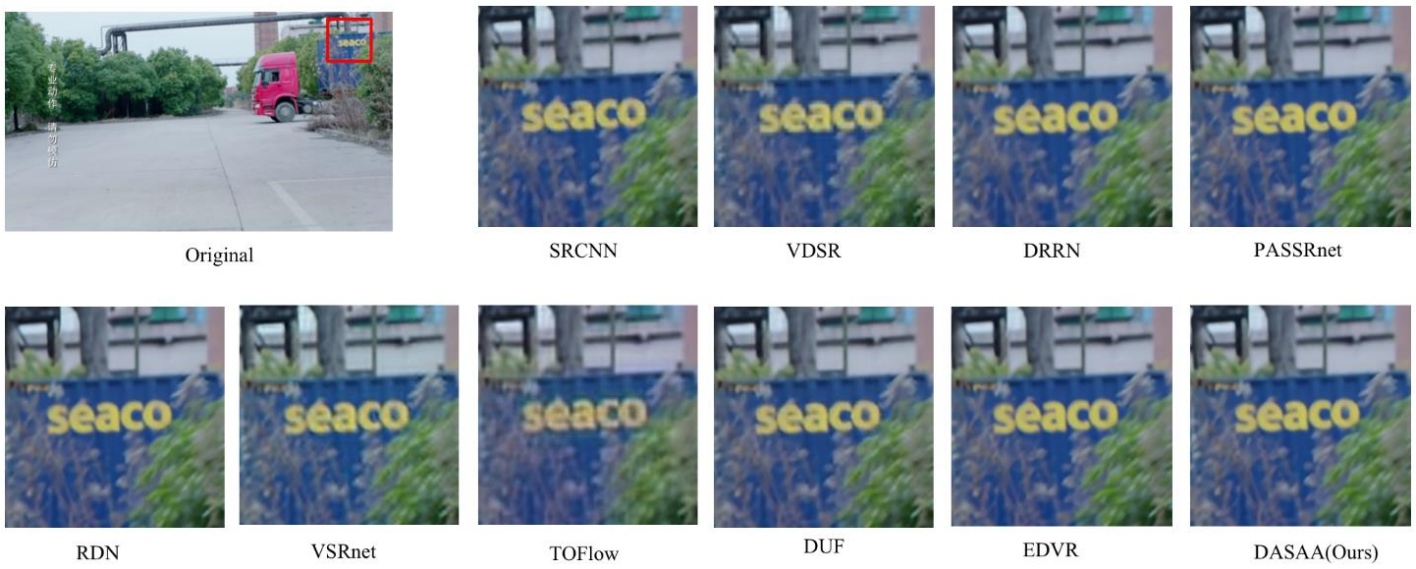


Fig. 7. The qualitative visual comparisons for 4x SR on Youku-VESR dataset

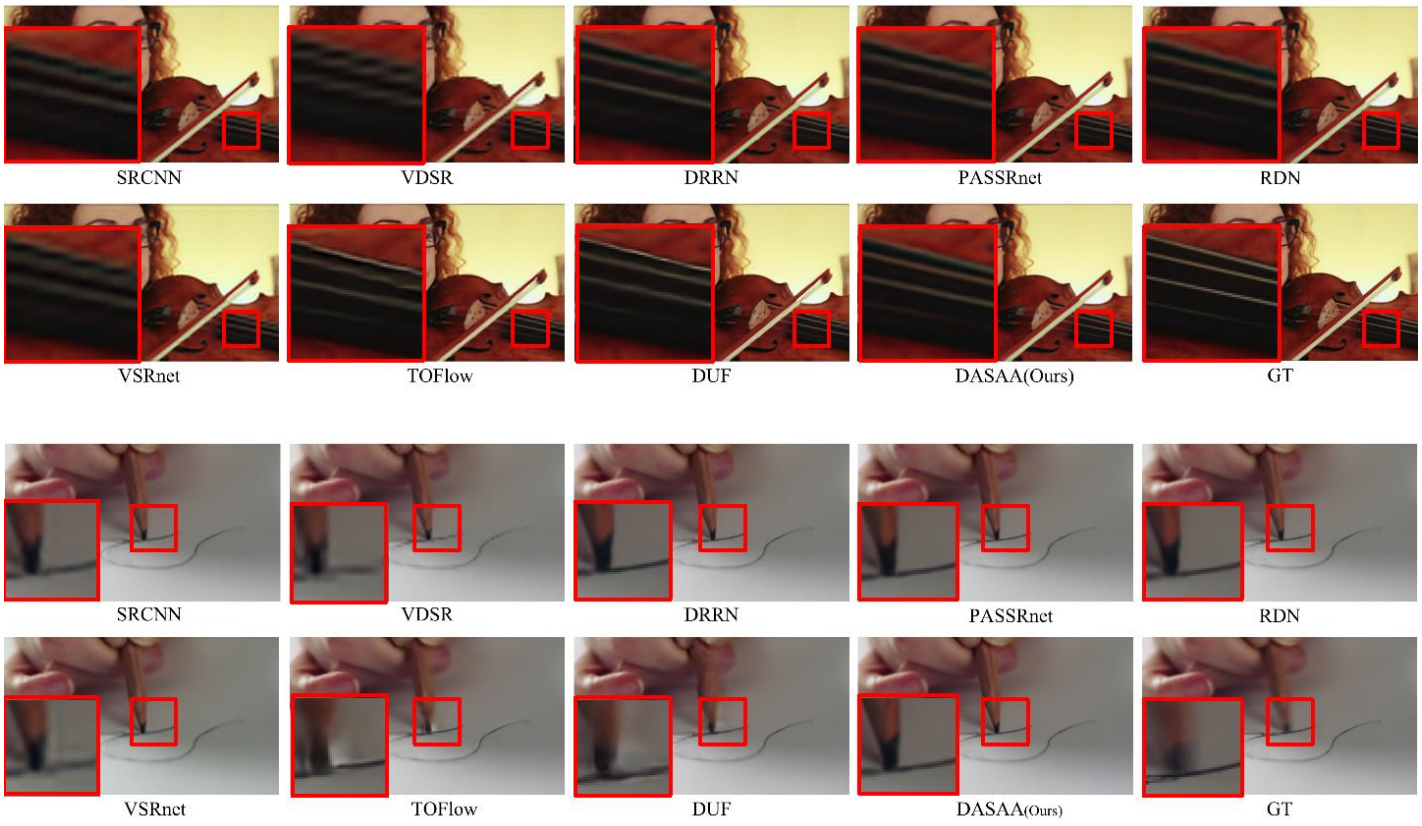


Fig. 8. The qualitative visual comparisons for 4x SR on Vimeo-90k dataset

Table I: The PSNR and SSIM comparisons for diverse SR methods on the Youku-VESR dataset. The best and the second best are indicated with bold red and bold blue, respectively.

Method	Kitchen	Violin	Pencil	Scissors	Hair	Girl	Average
SRCNN	32.804/0.807	32.956/0.814	35.370/0.927	33.755/0.885	33.048/0.833	32.442/0.792	32.682/0.765
VDSR	32.644/0.783	32.938/0.808	35.156/0.930	33.591/0.878	33.038/0.833	32.456/0.792	32.615/0.759
DRRN	33.171/0.833	33.130/0.819	35.813/0.939	33.854/0.887	33.111/0.834	32.457/0.794	32.814/0.771
RDN	33.207/0.807	33.153/0.820	35.812/0.939	33.850/0.887	33.149/0.832	**32.489**/0.792	32.852/0.772
PASSRnet	33.215/0.839	33.150/0.822	**35.847**/0.940	33.836/0.887	33.153/0.832	32.488/0.792	32.852/0.772
VSRnet	32.497/0.789	32.610/0.804	34.738/0.928	33.336/0.881	32.934/0.828	32.274/0.792	32.319/0.755
TOFlow	32.818/0.807	32.962/0.815	35.374/0.928	33.741/0.885	33.056/0.832	32.448/0.792	32.704/0.768
DUF	33.182/0.834	33.160/0.820	35.772/0.933	33.853/0.888	33.121/**0.834**	32.471/**0.794**	32.806/0.770
EDVR	**33.468/0.860**	33.240/**0.821**	35.807/**0.941**	**33.875**/**0.887**	**33.153/0.833**	32.443/0.785	**32.897**/**0.772**
OURS	**33.271/0.843**	**33.260**/**0.833**	**35.848**/**0.940**	33.862/**0.888**	**33.169**/**0.833**	**32.495/0.803**	**32.878**/**0.774**

Table II: The PSNR and SSIM comparisons for diverse SR methods on six typical selected images from Vimeo-90k dataset. The best and the second best are indicated with bold red and bold blue, respectively.

Model name	**PSNR**	**SSIM**
SRCNN[14]	33.247	0.924
VDSR[15]	31.414	0.898
RDN[1]	35.021	0.938
DRRN[12]	34.707	0.936
PASSRnet[7]	35.120	0.939
VSRnet[23]	30.224	0.891
TOFlow[20]	33.355	0.930
DUF[28]	34.318	0.935
EDVR[3]	**35.772**	**0.939**
Ours	**35.290**	**0.939**

It should be noted that in Table I, for the Youku-VESR dataset the PSNR of our method surpasses most of the methods but is a little lower than EDVR [3] method. While in Table II, for Vimeo-90K images, the PSNR of our methods are almost defeating other methods. EDVR [3] combines the deeper residual pyramid structure with the alignment module which takes more complex deformable convolution [27] operations. Although its performance is slightly improved ~~a little~~, the

introduction of the residual pyramid and the deformable convolution undoubtedly increases the amount of computation. Based on this, we introduce another objective computation load measure: the number of parameters and the floating-point operations per second (FLOPS). These comparison results of some related methods (their PSNR is close to ours) are shown in Table III. From this table, we can know that both the amount of the parameters and the computation load of our DASAA method are much less than that of the EDVR method, which reveals the high efficiency of our method. Thus, from Table I, Table II, and Table III, we conclude that our DASAA method can not only generate very good super-resolution results but also be beneficial for efficient computation.

Table III: The comparisons on the parameters and the FLOPS for different models

Model Name	RDN	PASSRnet	TOFlow	DUF	EDVR	Ours
Params	1.48M	1.36M	1.4M	5.8M	20.7M	1.8M
FLOPS	3.09G	2.47G	270.48G	204.8G	177.11G	6.39G

4.4 Ablation study

We perform ablation studies to analyze the contribution of each component in the proposed DASAA model. There are five main components in our model and each component corresponds to one variant model, which can be named as model 1, model 2, etc. Table IV displays the details. Due to the particularity of the dual attention structure, removing one attention branch will lead to the failure of subsequent network operations. Thus, we use the cloning strategy to duplicate the remaining attention branch to the one that has been removed. Such models are denoted as Model 4 and Model 5, respectively. Besides this, we also design an addition model–Model 6, which can change the positions of PAB and CAB mentioned in Section 3.4, to comprehensively analyze the attention relationship. The comparison between PSNR and SSIM on the Youku-VESR dataset of all variant models is shown in Table IV. Moreover, to show the role of each component more clearly, we also present the PSNR changing curve of each variant model when training on Fig. 9.

Table IV: The PSNR and SSIM comparisons for variant models under Youku-VESR dataset. The best is indicated with bold.

Component	RDB	mask	LSTM	PAB	CAB	PSNR	SSIM
Model 1	×	√	√	√	√	33.405	0.925
Model 2	√	×	√	√	√	35.270	0.939
Model 3	√	√	×	√	√	34.041	0.938
Model 4	√	√	√	×	√	35.165	0.939

Model 5	√	√	√	√	×	35.241	0.939
Model 6	√	√	√	√(CAB)	√(PAB)	**35.307**	0.939
DASAA	√	√	√	√	√	35.290	**0.939**

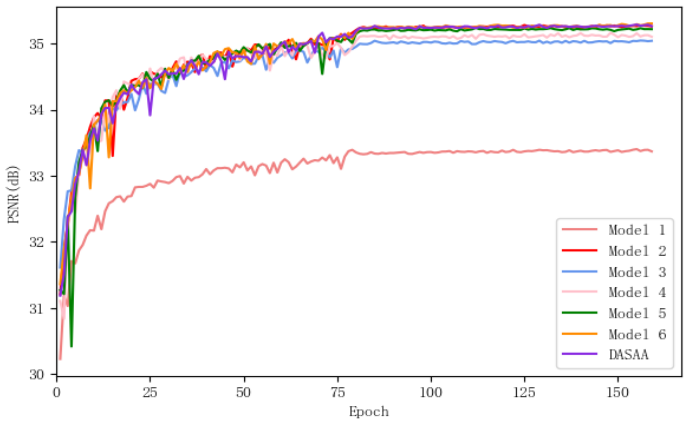


Fig. 9. PSNR comparisons during training for all variant models

From Table IV and Fig. 9, it is clear that the most three important components are RDB (light coral curve in Fig. 9), LSTM (cornflower blue curve in Fig. 9), and PAB (pink curve in Fig. 9). Also, based on the PSNR value and the curve trend of Model 6, exchanging two attention positions has no obvious effect on the performance compared with the original model. This also indicates that such two attention branches–position attention and channel attention are independent and unrelated.

4.5 Temporal consistency

The temporal profile is usually generated by obtaining several horizontal pixel rows from the video frames and stacking them vertically. Considering the scene changes in most videos may lead to deviations to the results and produce flicker artifacts, we specifically choose a test video with 21 frames that contains only one moving object without lens switching to obtain the temporal profile. The result is shown in Fig. 10. In this Fig., the temporal profile of each method represents ×4 results from the position marked by the yellow box in the original temporal profile.

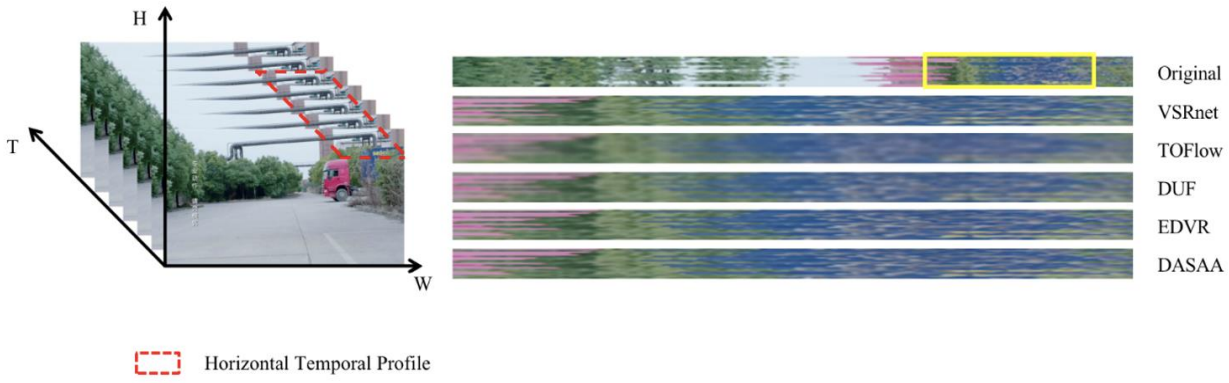


Fig. 10. Temporal profile of video super-resolution methods.

5. Conclusions

In this work, we have presented ~~propose~~ a concise and efficient dual attention with the self-attention alignment-based deep network for video super-resolution. First, we use RDB to make full use of the features from all related layers. After that, two kinds of attention–position attention and channel attention are introduced to excavate the channel and spatial information simultaneously and improve the expressive ability of the network. Furthermore, the self-attention structure is utilized to make reliable alignment and contextual learning of such two attention maps, which can improve the performance of SR. Finally, the LSTM structure is used to guarantee the consistency of the super-resolved video frames. Moreover, through experimental visualization observations, we find that such two attentions do not affect each other when they are introduced concurrently. Experimental results and comparisons show that our proposed network outperforms ~~has better performance than~~ the state-of-the-art network. Our upcoming work is to apply our proposed method to other video-related tasks, such as video background subtraction [32], to observe and analyze the results.

Compliance with Ethical Standards:

Funding: This work was funded in part by the National Natural Science Foundation of China under (grant No. 61971004), the Natural Science Foundation of Anhui Province (Grant No. 2008085MF190), and the Key Project of Natural Science of Anhui Provincial Department of Education (grant No. KJ2019A0083).

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2472-2481(2018).

[2] 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), pp. 286-301(2018).

[3] 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 Conference on Computer Vision and Pattern Recognition Workshops, pp. 0-0(2019).

[4] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794-7803 (2018).

[5] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318(2018).

[6] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146-3154 (2019).

[7] Wang, L., Wang, Y., Liang, Z., Lin, Z., Yang, J., An, W., Guo, Y.: Learning parallax attention for stereo image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12250-12259(2019).

[8] 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 Conference on Computer Vision and Pattern Recognition, pp. 10522-10531(2019).

[9] Sajjadi, M. S., Vemulapalli, R., Brown, M.: Frame-recurrent video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6626-6634 (2018).

[10] Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3897-3906(2019).

[11] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778(2016).

[12] 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, pp. 3147-3155(2017).

[13] 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, pp. 136-144(2017).

[14] Dong, C., Loy, C. C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence, 38(2), 295-307(2015).

[15] Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1646-1654(2016).

[16] Zhang, K., Mu, G., Yuan, Y., Gao, X., Tao, D.: Video super-resolution with 3D adaptive normalized convolution. Neurocomputing, 94, 140-151(2012).

[17] Gers, F. A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with LSTM (1999).

[18] Huang, Y., Wang, W., Wang, L.: Bidirectional recurrent convolutional networks for multi-frame super-resolution. In: Advances in Neural Information Processing Systems, pp. 235-243(2015).

[19] Wang, T. C., Liu, M. Y., Zhu, J. Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. arXiv preprint arXiv:1808.06601(2018).

[20] Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W. T.: Video enhancement with task-oriented flow. International Journal of Computer Vision, 127(8), 1106-1125(2019).

[21] Li, Y., Tsiminaki, V., Timofte, R., Pollefeys, M., Gool, L. V.: 3D appearance super-resolution with deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9671-9680(2019).

[22] 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, pp. 4778-4787(2017).

[23] Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A. K.: Video super-resolution with convolutional neural networks. IEEE Transactions on Computational Imaging, 2(2), 109-122(2016).

[24] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation, 9(8), 1735-1780(1997).

[25] Feng, Y., Ma, L., Liu, W., Luo, J.: Spatio-temporal video re-localization by warp lstm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1288-1297(2019).

[26] 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, pp. 4489-4497(2015).

[27] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp. 764-773 (2017).

[28] Jo, Y., Oh, S. W., Kang, J., & Kim, S. J.: Deep video super-resolution network using dynamic up-sampling filters without explicit motion compensation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3224-3232(2018).

[29] Tian, Y., Zhang, Y., Fu, Y., Xu, C.: TDAN: Temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3360 – 3369 (2020).

[30] Yan, B., Lin, C., Tan, W.: Frame and feature context video super-resolution. In: The Thirty-Third AAAI Conference on Artificial Intelligence, pp.5597-5604 (2019).

[31] Liu, H., Fu, Z., Han, J., Shao, L., Hou, S. and Chu, Y.: Single image super-resolution using multi-scale deep encoder–decoder with phase congruency edge map guidance. Information Sciences, 473, 44-58 (2019).

[32] Sakkos, D., Liu, H., Han, J. and Shao, L.: End-to-end video background subtraction with 3d convolutional neural networks. Multimedia Tools and Applications, 77 (17), 3023-23041 (2018).