Pattern Recognition 110 (2021) 107619

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Video super-resolution based on a spatio-temporal matching network

Xiaobin Zhu[a], Zhuangzi Li[b], Jungang Lou[c,*], Qing Shen[c]

[a] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
[b] School of Computer and Information Engineering, Beijing Technology and Business University, Beijing, China
[c] School of Information Engineering, Huzhou University, Huzhou, China

ARTICLE INFO

Article history:
Received 23 December 2019
Revised 19 August 2020
Accepted 29 August 2020
Available online 2 September 2020

Keywords:
Deep matching
Wavelet domain
Non-local matching
Residual learning
Video super-resolution

ABSTRACT

Deep spatio-temporal neural networks have shown promising performance for video super-resolution (VSR) in recent years. However, most of them heavily rely on accuracy motion estimations. In this paper, we propose a novel spatio-temporal matching network (STMN) for video super-resolution, which works on the wavelet domain to reduce dependence on motion estimations. Specifically, our STMN consists of three major components: a temporal fusion wavelet network (TFWN), a non-local matching network (NLMN), and a global wavelet domain residual connection (GWDRC). TFWN adaptively extracts temporal fusion wavelet maps via three 3d convolutional layers and a discrete wavelet transform (DWT) decomposition layer. The extracted temporal fusion wavelet maps are rich in spatial information and knowledge of different frequencies from consecutive frames, which are feed to NLMN for learning deep wavelet representations. NLMN integrates super-resolution and denoising into a unified module by pyramidally stacking non-local matching residual blocks (NLMRB). At last, GWDRC reconstructs the super-resolved frames from the deep wavelet representations by using global wavelet domain residual information. Consequently, our STMN can efficiently enhance reconstruction quality by capturing different frequencies wavelet representations in consecutive frames, and does not require any motion compensation. Extensive experiments conducted on benchmark datasets demonstrate the effectiveness of our method compared with state-of-the-art methods.

1. Introduction

Video super-resolution (VSR) aiming at recovering high-resolution (HR) frames from low-resolution (LR) frames attracts extensive attention in research and industrial communities. It can be used in many real-world applications, such as surveillance [8], medical image processing [9], recognition [10]. However, VSR is an inherent ill-posed problem since its one-to-many mapping nature. In another word, a variety of HR images can be mapped to the same LR image.

In the VSR task, the temporal motion information modeling is a key problem. Temporal motion information underlying consecutive frames can provide beneficial priors for reconstructing current frame and keeping visual consistency between super-resolved frames. As a typical graph matching problem [11], explicit motion estimation (e.g. optical flow) is commonly formulated in the state-of-the-art VSR methods. Kappeler *et al.* [1] adopted motion compensation with convolutional neural networks to build temporal relations of sequential frames. Caballero *et al.* [6] collaborated motion compensation and super-resolution network in an end-to-end manner. However, the precise motion estimation is challenging and time-consuming. Many motion estimation algorithms rely on the brightness constancy assumption, but they may fail due to lightness/pose variation and the presence of motion blurs and occlusions. In another aspect, some efforts [3,5,12] utilize recurrent architectures to build temporal dependence and show excellent efficiency. In [7], a fast spatio-temporal residual network was proposed, and they adopted three-dimensional convolutions to simultaneously exploit spatio-temporal relations. These VSR methods do not specially consider frames frequency information, and noise disturbances are not technically eliminated.

In this paper, we propose a novel spatio-temporal matching network (STMN) for video super-resolution in wavelet domain. It mainly contains three components: a temporal fusion wavelet network (TFWN), a non-local matching network (NLMN), and a global wavelet domain residual connection (GWDRC). TFWN is designed to aggregate consecutive frames for implicitly modeling motion information, instead of explicitly formulating complicated and time-consuming motion information as in motion compensation methods [1,6]. Then we transform the aggregated features into wavelet domain, which retains spatial information and meanwhile provides

* Corresponding author.
E-mail address: loujungang0210@hotmail.com (J. Lou).

https://doi.org/10.1016/j.patcog.2020.107619
0031-3203/

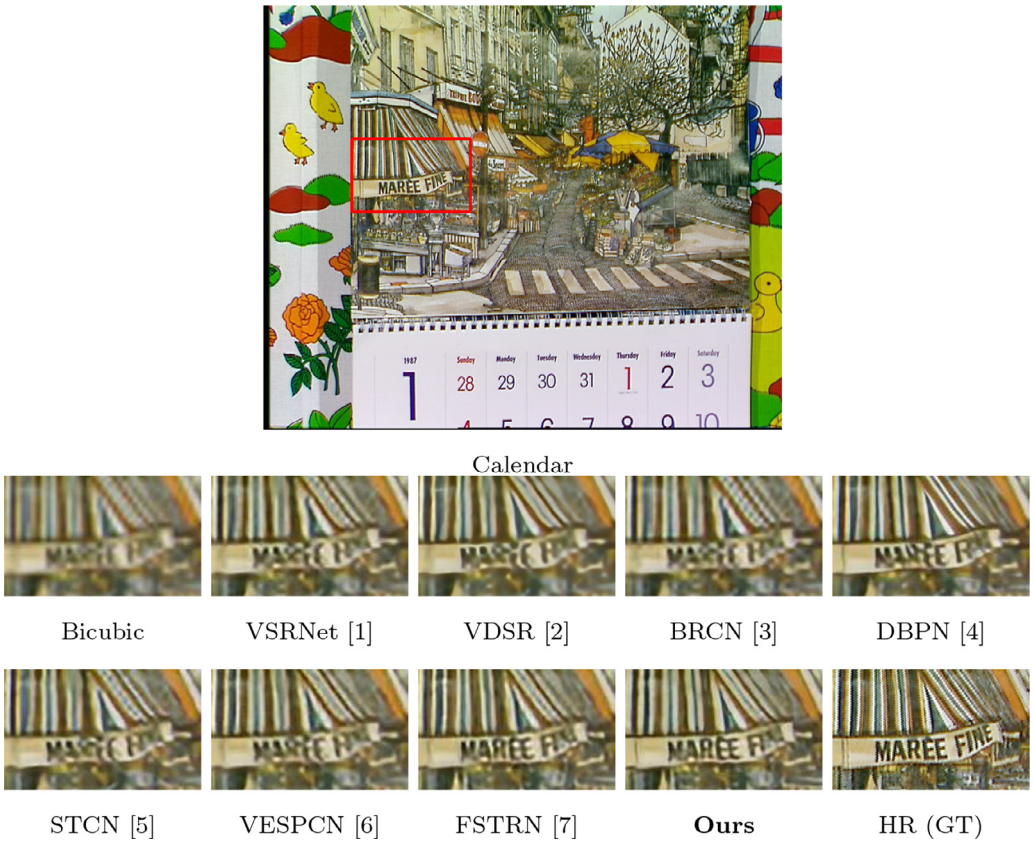


Fig. 1. Visual comparisons of different video super-resolution methods on the Calendar dataset with 4 × up-scale factors; our method can clearly recover the "words" and "lines".

knowledge of different frequencies. In NLMN, we adopt a pyramid design to regularly augment feature channels along with network going deeper. Besides, a novel non-local matching residual block (NLMRB) is elaborately designed in our NLMN. Our NLMRB is highly motivated by the non-local matching strategy, and can produce a denosing weight matrix [13], so that it can well integrate super-resolution and denosing process. Last but not least, we design the GWDRC which reconstructs the super-resolved frames with global wavelet domain residual information. We show an example of VSR results of different methods in Fig. 1, and our method showcases very promising performance compared with others.

In summary, the main contributions are four-fold:

- To the best of our knowledge, this paper presents one of the very first attempts towards wavelet domain deep video super-resolution.
- Proposing a temporal fusion wavelet network (TFWN) to adaptively produce temporal fusion wavelet maps so that avoids motion compensations.
- Constructing a non-local matching network (NLMN) with an innovative non-local matching residual block (NLMRB), which specially targets at integrating super-resolution and denoising on wavelet domain.
- Extensive experiments conducted on publicly available datasets demonstrate the state-of-the-art performance of our method.

The rest of our paper is organized as follows: In Section 2 we overview related work. Our method is illustrated in Section 3. In Section 4 the experimental results are shown and discussed. In Section 5, we conclude this paper.

2. Related work

2.1. Single image super-resolution

The pioneering single image super-resolution (SISR) methods often use direct interpolations, e.g. bilinear interpolation and bicubic interpolation, but these methods neglect prior knowledge and cannot recover fine-grained details. So, learning based methods are widely investigated [14,15], and these methods utilize external example pairs to supervise the learning process. Cruz et al.. [16] proposed a Wiener filter in similarity domain for super resolution, which formulates the SISR as a minimization problem of reconstruction error subjected to a sparse self-similarity prior. On the other hand, some the external example-based methods [14] learn a mapping function from low resolution patches to high resolution patches based on an external dataset. These methods pay particular attention to learn and optimize dictionaries [17] or build efficient mapping functions. However, their super-resolved results are usually unsatisfactory with large magnification factors.

With the flourish of deep learning, convolutional neural networks (CNNs) are successfully adopted for single image super-resolution [18,19]. Dong et al. [15] firstly presented a three-layer CNN based single image super resolution method and achieves better performance than sparse coding based method [14]. Afterwards, some deep residual networks are proposed to construct more powerful single image super-resolution models. In [4], a deep back-projection network is presented to achieve promising performance in large-scale image super-resolution by stacking down-sample and up-sample operations. In [18], an ordinary differential equation inspired scheme is designed for single image super-resolution, which brings a new understanding of the residual network (ResNet) originally used to image classification problems. He et al. [20] proposed

a residual network (ResNet) to conduct super-resolution for image classification. ResNet facilitates the training of networks that are substantially deeper than before. However, ResNets with random weight dropout can achieve improved performance, demonstrating that a great amount of redundancy can be found in residual networks. Wang et al. [21] adopted a general way through introducing the idea of ensemble into SR task. Also some works [22] try to accelerate the running speed by post setting up-sample strategies. Overall, in single image super-resolution, high-resolution image is recovered by a single low-resolution image without temporal dependencies, which is not fully suitable to video super-resolution. Because the simple consideration of single image losses abundant clues for the same scene are conveyed in consecutive frames.

2.2. Video super-resolution

VSR tasks not only require photo-realistic details spatially, but also demand details can coherently change with low-resolution content temporally. Although the spatial information can be well processed by powerful convolution networks, the temporal inconsistency is still a challenging problem. Existing VSR approaches [6,23–25] extract optical flows from consecutive frames for motion estimations in the first stage and utilize the estimated motion fields to perform motion compensation in the second stage. Some works improve the quality of motion compensation,e.g., Mitzel et al. [23] imposed a total variation regularity for the high-accuracy optical flow estimation. In [24], a Bayesian approach is proposed to estimate underlying motion, blur kernel and noise level. Kappeler et al. [1] combined the motion compensation step and CNN effectively for VSR. Liu et al. [25] designed a temporal matching network to learn temporal relations. Caballero et al. [6] utilized spatial transformers to encode optical flows for improve motion compensations. They also investigated three temporal fusion strategies, i.e. early, slow fusion and three-dimensional convolutions. In [26], a sub-pixel motion compensation method is proposed to accelerate the temporal modeling speed. However, the optical flow of [26] is artificially designed that cannot fully fuse into super-resolution networks, which restricts the feature representative ability, so that is harmful to he super-resolution reconstruction quality.

On the other hand, some recurrent neural network (RNN) based VSR approaches [3,5,12,27,28] establish temporal relations potentially, but they usually need large memory allocations and the recurrent network is hard to train. Li et al. [7] proposed a fast three-dimensional convolution network for VSR, and they adopted local residual learning and global residual learning strategies to ease the network training. M. Haris et al. [28] integrated spatial and temporal contexts from consecutive video frames by using a recurrent encoder-decoder module that fuses multi-frame information with more traditional, single frame super-resolution path for the target frame. Guo et al. [5] designed a spatio-temporal network for VSR, which consists of a CNN based spatial component, a RNN based temporal component, and a convolutional layer based reconstruction component. Liu et al. [29] effectively utilize temporal information by coherently combining a temporal adaptive neural network and a spatial alignment network. Besides, some works [3,27] directly build a light architecture for achieving a high efficiency. In our study, we aim to design a more simple and effective temporal feature aggregation technique to replace explicit motion compensation.

2.3. Wavelet domain super-resolution

Wavelet transform is an efficient and highly intuitive tool to represent and store multi-resolution images [30]. It can depict the contextual and textural information of an image at different levels of frequency. The Haar wavelet is one of the most conventional wavelet transform method. Given an image $\mathbf{I}$, the two-dimensional discrete wavelet transform f_{DWT} can decompose $\mathbf{I}$ to four sub-bands, including: average $\mathbf{I}_{LL}$, vertical $\mathbf{I}_{HL}$, horizontal $\mathbf{I}_{LH}$, and diagonal $\mathbf{I}_{HH}$ information. For single image super-resolution, the wavelet transform is mostly used in interpolation-based methods and statistic-based methods [31]. For example, H. Ji et al. [32] presented a wavelet-based iterative reconstruction algorithm for image super-resolution,in which noise is efficiently suppressed.

Recently, many deep learning based image super-resolution methods [30,33,34] adopt the wavelet transform technology. Guo et al. [33] directly solve image super-resolution in wavelet domain, which exhibits impressive efficiency and good performance. Further, Guo et al. [35] proposed an orthogonally regularized deep super-resolution network which takes advantage of image transform domain while adapts the design of transform basis to the training image set. In [34], a wavelet transform based up-sample module is proposed. Huang et al. [30] proposed a face hallucination approach, which shows the effectiveness of textures and details reconstruction in the wavelet domain. In summary, there are no attempts to build temporal and frequency relations in the existing deep learning based video super-resolution methods. Therefore, this situation motivates us to introduce wavelet transform into video super-resolution.

3. Our method

3.1. Overview

In order to make a well usage of temporal dependencies and enhance representation abilities, we propose a novel spatio-temporal matching network (STMN) video super-resolution in the wavelet domain, as shown in Fig. 2, which mainly consists of three components, i.e., a temporal fusion wavelet network (TFWN), a non-local matching network (NLMN), and a global wavelet domain residual connection (GWDRC). TFWN is designed to fuse input low-resolution frames into temporal fusion wavelet maps. NLMN is proposed to learn deep wavelet representations. GWDRC is designed to reconstruct the super-resolved frames with global wavelet domain residual information. In the following, we will elaborate those three components.

3.2. Temporal fusion wavelet network

Although many researches [30,33] have demonstrated the efficiency and capability of discrete wavelet transform (DWT) in representing and storing multi-resolution images, DWT itself has not been utilized in the video super-resolution task. For that, we propose a temporal fusion wavelet network (TFWN) to adaptively produce temporal fusion wavelet maps to reduce the dependence on precise motion estimations.

TFWN consists of three sequential $3d$ convolutional layers [37] and one DWT layer, as shown in Fig. 2. The first $3d$ convolutional layer captures a sequence of low-resolution frames $\mathbf{X}_0$ and generate 32-channel feature maps $\mathbf{X}_1$. The second $3d$ convolutional layer is the same as the first $3d$ convolutional layer with a 32-channel output $\mathbf{X}_2$, while the last $3d$ convolutional layer produces 16-channel temporal fusion feature maps $\mathbf{X}_3$. All the $3d$ convolutional layers include bias and are followed by rectified linear unit (ReLU) layers. All these layers adopt $3 \times 3 \times C_T$ convolutional kernels with zero padding, and C_T is the temporal radius.

Similar to the multi-scale face hallucination [30] method that applies wavelet transform on deep feature maps, we perform a DWT layer on temporal fusion feature maps, which decomposes the feature maps into sequences of wavelet coefficients with the

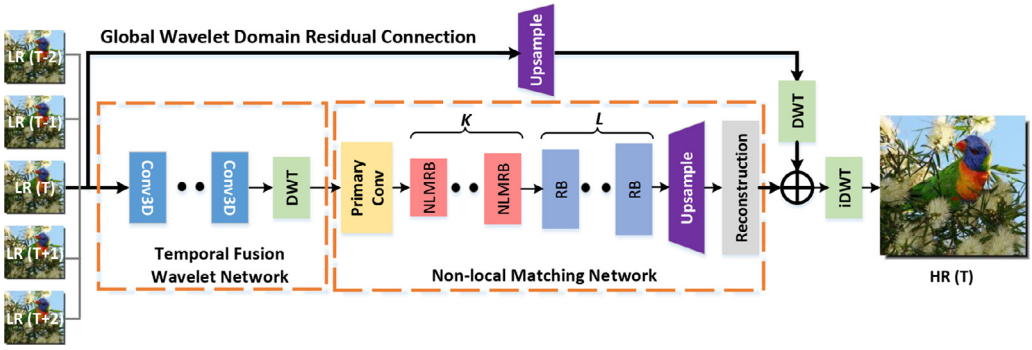


Fig. 2. Framework of our spatio-temporal matching network (STMN) video super-resolution in the wavelet domain; five consecutive frames are used as an input sequence; $\oplus$ denotes element-wise addition. NLMRB denotes a non-local matching residual block, as shown in Fig. 4. RB denotes a classical residual block [36].

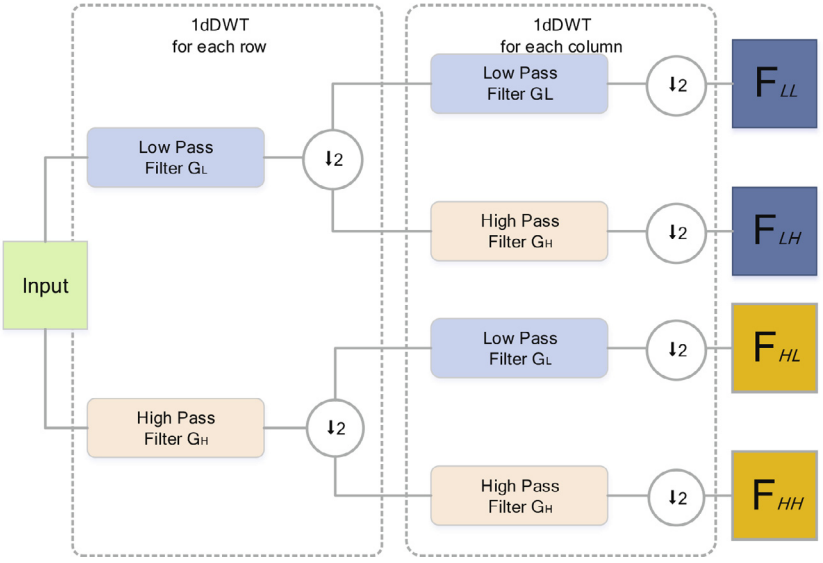


Fig. 3. Flowchart of 2dDWT decomposition.

same size, as follows:

$$\mathbf{F}_{LL}, \mathbf{F}_{LH}, \mathbf{F}_{HL}, \mathbf{F}_{HH} = f_{DWT}(\mathbf{X}_3), \tag{1}$$

where $\{\mathbf{F}_X\}$ ($X \in \{LL, LH, HL, HH\}$) are temporal fusion wavelet maps of different frequencies, and each one has 16 channels. Moreover, the spatial resolution of $\mathbf{F}_X$ decreases two times than $\mathbf{X}_3$. To present DWT more clearly, an example of 1-level 2*d*DWT decomposition with Haar kernels is shown in Fig. 3. In our implementation, we choose the Haar wavelet, for it is simple and enough to represent different frequency information. The right part of Fig. 3 is the notation of each sub-band of wavelet coefficients. It is clear that the 2*d*DWT captures the image details in four sub-bands: average (LL), vertical (HL), horizontal (LH) and diagonal (HH) information, which are corresponding to each wavelet sub-bands coefficients. Note that after 2*d*DWT decomposition, the combination of four sub-bands always has the same dimension as the original input image. The wavelet coefficients at different levels are computed by repeating the decomposition in Fig. 3 to each output coefficient iteratively. The 2*d* Inverse DWT (2*d*IDWT) can trace back the 2*d*DWT procedure by inverting the steps, allowing the prediction of wavelet coefficients to generate super-resolution results.

3.3. Non-local matching network

As shown in Fig. 2, the proposed non-local matching network (NLMN) mainly contains a primary convolutional layer, K non-local matching residual blocks (NLMRBs) and L residual blocks (RBs) [36]. Both NLMRBs and RBs are stacked in a pyramidal manner of increasing channels along with the network propagation. More details are described as follows.

Non-local matching residual block Existing super-resolution methods [28] tend to magnify noises during the high-frequency reconstruction. For denoising, we propose a simple yet effective non-local matching residual block, which is realized by adding a non-local matching [13] block after the first 1×1 convolution of the bottleneck residual block (RB) [36], as shown in Fig. 4. The non-local operation computes a denoising mapping by taking a weighted mean of features at all spatial locations. Given an input feature map $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, its non-local means are computed as follows:

$$\mathbf{y}_i = \frac{1}{\mathcal{N}(\mathbf{x})} \sum_{\forall j \in \mathcal{L}} f(\mathbf{x}_i, \mathbf{x}_j) \cdot g(\mathbf{x}_j), \tag{2}$$

where $g(\mathbf{x})$ denotes a linear embedding function which is implemented by a 1×1 convolution. $\mathcal{L}$ denotes all spatial locations, $f(\mathbf{x}_i, \mathbf{x}_j)$ is a feature-dependent weighting function, and $\mathcal{N}(\mathbf{x})$ is a normalization function. $\mathbf{y}_i$ denotes the ith value of the denoised feature map $\mathbf{y}$. In our model, the function $f(\mathbf{x}_i, \mathbf{x}_j)$ is defined as the Gaussian (softmax) form:

$$f(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\sqrt{C}} e^{\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)}, \tag{3}$$

where $\phi(\mathbf{x})$ and $\theta(\mathbf{x})$ are two embedding functions which are implemented as 1×1 convolutions, C denotes the channel number.

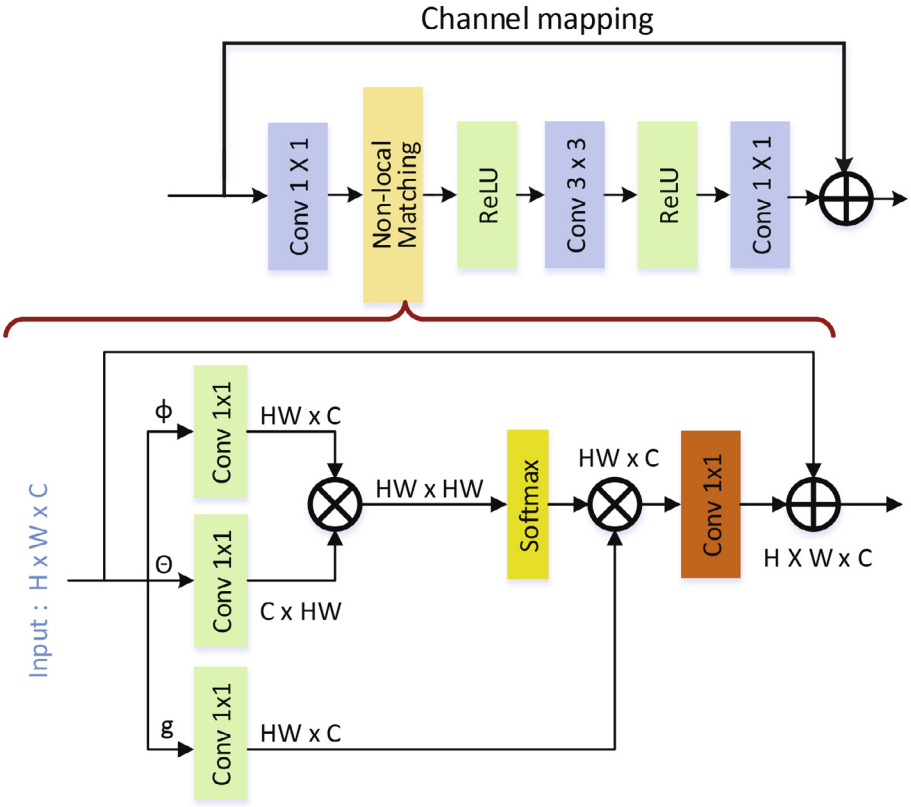


Fig. 4. The diagrammatic drawing of the proposed non-local matching residual block.

So the normalization function is computed as:

$$\mathcal{N}(\mathbf{x}) = \sum_{\forall j \in \mathcal{L}} f(\mathbf{x}_i, \mathbf{x}_j), \tag{4}$$

and the final output is computed as:

$$\mathbf{z} = \psi(\mathbf{y}) + \mathbf{x}, \tag{5}$$

where $\psi(\mathbf{y})$ is a 1×1 convolutional embedding function.

Pyramidal stacking manner According to the study of [38], increasing residual block channel numbers along with the network propagation is beneficial to improving network learning capability. For that, a pyramidal stacking way is proposed to linearly increase the channel numbers of NLMRBs and RBs during the non-local matching network propagation. Specifically, for NLMRB and RBs, the channel increasing rate R is set to $\frac{\alpha}{K+L}$, where $\alpha > 0$ a hyper-parameter and K and L are numbers of NLMRBs and RBs, respectively.

3.4. Global wavelet domain residual connection

Different from existing residual learning methods [2], a global wavelet domain residual connection (GWDRC) is proposed to reconstruct super-resolved frames in this paper, as shown in Fig. 2. In fact, GWDRC can be regarded as a cross-space residual learning method. In our GWDRC, a simple super-resolution mapping is directly adopted to map low-resolution frames to high-resolution frames by an interpolation based up-sample layer. Then, the high-resolution mapped frames are converted to wavelet domain, and are added to the NLMN results, forming a global residual learning in the wavelet domain. Specifically, for the Tth frame $\mathbf{I}^{LR}(T)$ of an input low-resolution sequence $\mathbf{I}^{LR}$, the GWDRC firstly applies a $4\times$ bicubic interpolation operation based up-sample layer to generate a high resolution frame $\mathbf{I}_b^{LR}(T)$. Then, the GWDRC performs a DWT layer on $\mathbf{I}_b^{LR}(T)$ and adds the corresponding DWT result with the non-local matching network's output to obtain the high resolution wavelet maps $\mathbf{I}^{SRW}(T)$, as follows:

$$\mathbf{I}^{SRW}(T) = f_{DWT}(\mathbf{I}_b^{LR}(T)) + NLMN(TFWN(\{\mathbf{I}^{LR}(t)\})), \tag{6}$$

where $NLMN(\cdot)$ and $TFWN(\cdot)$ denote the proposed NLMN and TFWN networks, respectively. In our implementation, we adopt five consecutive frames as the input of our method, which means that $\{\mathbf{I}^{LR}(t) | t \in \{T-2, T-1, T, T+1, T+2\}$. Finally, the inverse DWT (IDWT) is applied to transform the wavelet maps $\mathbf{I}^{SRW}(T)$ into Y-channel space to get final super-resolved frames.

4. Experiments

4.1. Dataset and metrics

Due to the lack of training video benchmark datasets, we collect 195 videos of 1080p (1920×1080) from the 699 pic[1] and vimeo[2] websites, which include different scenarios, such as nature, streetscape and daily life. The collected videos are 2 time down-sampled to 960×540 and randomly clipped to form 5,800 video sequences. For testing, the public available benchmark dataset of Vid4 is adopted [1,6]. Vid4 contains four video sequences, namely, Calendar, Foliage, City, and Walk. All experiments are performed by $4\times$ up-sample factor from low resolutions to high resolutions. According to previous studies, peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) are selected as evaluation metrics. PSNR and SSIM are all calculated on the individual Y-channel.

[1] http://699pic.com/
[2] https://vimeo.com/

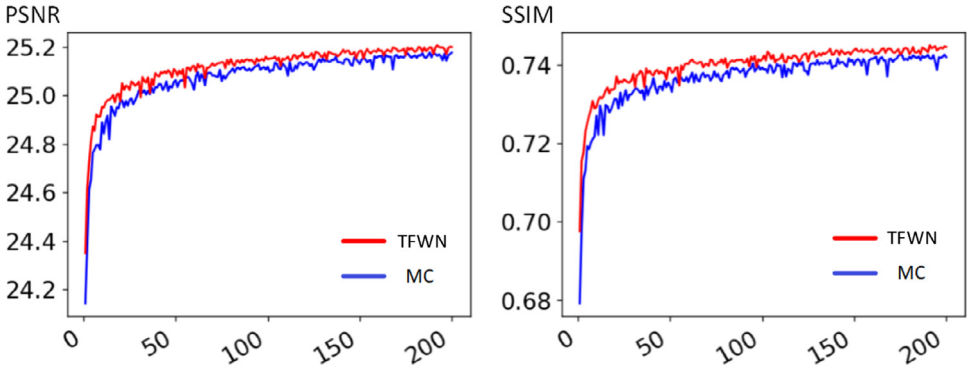


Fig. 5. Scores (PSNR and SSIM) across training iterations of MC and TFWN on Vid4.

Table 1
Ablation study on Vid4; the classification network is VGG-16.

Model	Calendar	Foliage	City	Walk
MC	21.90/0.7086	24.54/0.6887	26.09/0.6995	28.16/0.8707
TFWN	**21.91/0.7114**	**24.55/0.6895**	**26.14/0.7038**	**28.20/0.8720**

Table 2
Ablation study of NLMRB on Vid4; the backbone network is VGG-16.

Model	Calendar	Foliage	City	Walk
*K0L*10	21.91/0.7115	24.56/0.6895	26.14/0.7038	28.21/0.8720
*K1L*9	21.88/0.7106	24.57/0.6907	26.19/0.7065	28.21/0.8722
*K2L*8	**21.92/0.7121**	**24.65/0.6950**	**26.24/0.7107**	**28.25/0.8728**
*K3L*7	21.90/0.7101	24.55/0.6895	26.14/0.7041	28.20/0.8715
*K4L*6	21.90/0.7118	24.55/0.6900	26.17/0.7071	28.23/0.8724
*K8L*2	21.90/0.7105	24.56/0.6901	26.13/0.7037	28.23/0.8723

4.2. Temporal fusion wavelet network evaluation

In this section, we compare our temporal fusion wavelet network (TFWN) with a motion compensation (MC) based method. In the MC method, for efficiently extracting optical flows, we adopt a CNN based optical flow extraction method [39], and warp the flow into consecutive frames according to [6]. For a fair comparison, we use the same pyramidal residual super-resolution network, which sets $K = 0$ and $L = 10$, and three consecutive frames are utilized. The experimental results are shown in Table 1. Apparently, our TFWN consistently outperform MC. To be concrete, our TFWN outperforms MC by 0.04 dB PSNR and 0.0043 SSIM on "City" dataset, respectively. TFWN outperforms MC by 0.04 dB PSNR and 0.0013 SSIM on "Walk" dataset, respectively. In addition, TFWN defeats MC by 0.01 dB PSNR and 0.008 SSIM on the complex "Foliage" dataset. These comparisons adequately demonstrate the effectiveness of TFWN. Besides, the scores (PSNR and SSIM) across training iterations are also visualized, which is shown in Fig. 5. It can be seen that the red line (TFWN) is steadily higher than the blue line (MC), both on PSNR and SSIM. Experiments show that our TFWN can adaptively produce temporal fusion wavelet maps, and adequately capture temporal relations. Therefore, TFWN is both more effective and more efficient than the MC version.

4.3. Non-local matching residual block evaluation

In this section, we demonstrate the effectiveness of our NLMRB (non-local matching residual block). As shown in Fig. 2, we set K NLMRBs and L RBs (residual blocks). *K0L*10 denotes that there are 0 NLMRB and 10 RBs, while *K2L*8 denotes that there are 2 NLMRBs and 8 RBs. We set the total number of $K + L = 10$ to reduce computation cost. We adopt TFWN for frame fusion and the temporal radius is set to 3. For extensively investigating the effectiveness of NLMRB, we totally evaluate six different settings on the Vid4 dataset, i.e. *K0L*10, *K1L*9, *K2L*8, *K3L*7, *K4L*6, and *K8L*2. The corresponding experimental results are listed in Table 2. We can see that the performance of *K1L*9 is slightly worth than *K0L*10 on the "calendar" dataset, and *K1L*9 achieves higher scores on the "city" dataset and "Foliage" dataset. *K2L*8 achieves the best performance among all other methods. Notably, *K2L*8 outperforms *K1L*9 by 0.08 dB PSNR and 0.0043 SSIM on the "Foliage" dataset, and outperforms *K1L*9 by 0.05 dB PSNR and 0.0042 SSIM on the "City" dataset. *K2L*8 outperforms *K0L*10 by 0.09 dB PSNR and 0.0045 SSIM on the "Foliage" dataset, and outperforms the *K0L*10 by 0.1 dB PSNR and 0.0069 SSIM on the "City" dataset. Compared to *K2L*8, *K3L*7 has worse performance. Similarly, *K4L*6 and *K8L*2 achieve comparable performance, and their PSNR and SSIM are all lower than *K2L*8. The experiments show that NLMRB has the ability to denoise feature maps, and the denoised features are more helpful for super-resolution frames reconstruction. However, superfluous NLMRBs will put a lot of pressure on these layers and decrease detailed information. Consequently, we select *K2L*8 as our configuration.

4.4. Ablation study for global wavelet domain residual connection

In our work, the global wavelet domain residual connection (GWDRC) is proposed to reconstruct super-resolved frames. We visualize training curves of two conditions, i.e. a network with GWDRC and the other without GWDRC (no GWDRC). For simplicity, in the ablation study, we conduct experiments with three input consecutive frames. We adopt the architecture of *K2L*8 according to former experiments. As shown in Fig. 6, the training curves of "GWDRC" are consistently higher than "no GWDRC" in all sets. For super-resolution tasks, input and output are highly correlated, residual connection between the input and output is wildly investigated in the state-of-the-art methods. Previous works either perform residual learning on amplified inputs or perform residual connection directly on the input-output low-resolution space, none cross-space global residual connection on wavelet domain has been investigated. Our newly designed GWDRC can capture rich wavelet domain information directly from low-resolution frames, and strengthen the ability to extract global cross-space features by upsampling and DWT operations. It can be clearly seen that the network performs a faster convergence speed and achieves better performance with the help of GWDRC, which demonstrates the effectiveness of using GWDRC.

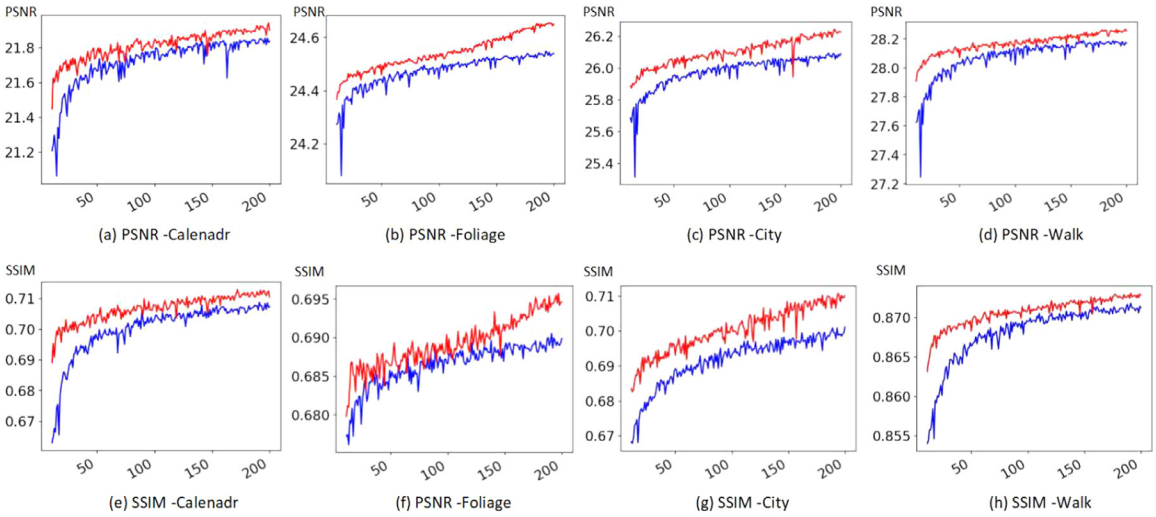


Fig. 6. Scores across training iterations of "GWDRC" and "no GWDRC" on Vid4. Red curves are results generated by "GWDRC" and blue cures are generated by "no GWDRC". (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Comparisons with the state-of-the-art methods by PSNR, SSIM, and MOS (4 ×). Scores in bold denote the highest values.

Testing dataset	Matrices	Bicubic (Base)	VSRNet [1]	BRCN [3]	VDSR [2]	DBPN [4]	STCN [5]	VESPCN [6]	FSTRN [7]	STMN (ours)
Calendar	PSNR	20.31	21.28	20.99	21.30	22.22	21.47	21.76	22.20	**22.60**
	SSIM	0.5862	0.6691	0.6461	0.6813	0.7353	0.6826	0.7036	0.7390	**0.7580**
Foliage	PSNR	23.45	24.31	24.12	24.28	24.69	24.33	24.61	24.90	**25.33**
	SSIM	0.5992	0.6766	0.6673	0.6657	0.6975	0.6775	0.6984	0.7272	**0.7425**
City	PSNR	25.16	25.61	25.47	25.71	25.83	25.77	26.12	26.51	**26.88**
	SSIM	0.6155	0.6617	0.6609	0.6631	0.7007	0.6756	0.7051	0.7413	**0.7648**
Walk	PSNR	26.03	27.54	27.07	27.73	28.56	27.85	28.12	27.95	**28.81**
	SSIM	0.8070	0.8522	0.8394	0.8492	0.8798	0.8610	0.8675	0.8701	**0.8858**
Average	PSNR	23.74	24.69	24.41	24.76	25.33	24.86	25.35	25.39	**25.90**
	SSIM	0.6520	0.7224	0.7034	0.7148	0.7533	0.7241	0.7437	0.7694	**0.7878**

4.5. Comparisons with the state-of-the-art methods

Quantitative evaluation Bicubic interpolation is selected as the baseline. VDSR (Super-Resolution Using Very Deep Convolutional Networks) [2] and DBPN (Deep Back-Projection Networks) [4] are single image super-resolution methods. VSRNet (Video Super-Resolution with Convolutional Neural Networks) [1] and VESPCN (Video Efficient Sub-Pixel Convolution Network) [6] are classical VSR methods, and they both adopt optical flow for motion compensation. BRCN (Bidirectional Recurrent Convolutional Network) [3] and STCN (Spatial-Temporal CNN) [5] are recurrent network based approaches. FSTRN (Fast Spatio-Temporal Residual Network) [7] is a 3*d* convolution based network. All these methods are re-implemented versions. The detail experimental results are listed in Table 3, and **STMN** denotes our **Spatio-Temporal Matching Network**.

From Table 3, we can see that our STMN outperforms another kind of 3*d* convolution based network, i.e. FSTRN, on SSIM by 0.4 dB, 0.43 dB, 0.37 dB and 0.86 dB in "Calendar", "Foliage", "City", "Walk", respectively. Our STMN outperforms FSTRN on the SSIM by 0.019 dB, 0.0153 dB, 0.0235 dB, and 0.0157 dB in "Calendar", "Foliage", "City", "Walk", respectively. In average, our STMN outperforms FSTRN with a great margin of 0.51 dB and 0.0184 dB on PSNR and SSIM. Besides, experiments also demonstrate STMN's powerful texture recovering capability. Its average SSIM surpasses VSRNet, STCN, and FSTRN more than 0.0654, 0.0637, and 0.0184, respectively. Comparing with the single image super-resolution method with the best performance in our experiments, i.e. DBPN, STMN outperforms it by 0.64 dB PSNR on the "Foliage" dataset and 1.05 dB PSNR on the "City" dataset. In average, our STMN outperforms DBPN with a great margin of 0.57 dB and 0.0345 on PSNR and SSIM. In summary, our STMN achieves the highest PSNR and SSIM scores on all datasets, which shows STMN's powerful texture recovering capability and temporal consistency keeping ability.

Qualitative evaluation Visual comparisons are provided in Fig. 7. We can see that our STMN method obtain better detail and texture reconstructions. On the "Foliage" dataset, we can recover the tree trunk more clearly. On the "City" dataset, our STMN achieves obvious better performance than others, it can well recover the textures of buildings. On the "Walk" dataset, our method can generate more distinct contours. Based on those quantitative and qualitative comparisons, we can conclude that our STMN achieve superior performance over the state-of-the-art methods, including higher PSNR/SSIM scores, and recovering the finest details and produces most pleasing results.

4.6. Implementation details

4.6.1. Network settings

Our final model is trained by five consecutive frames (in ablation study three consecutive frames). The temporal fusion wavelet network (TFWN) adopts three three-dimensional convolutional layers in total, and the output channel of the last convolutional layer is set to 16. Our f_{DWT} adopts the Haar wavelet, which can decompose the 16-channel features to 64-channel features. The primary convolution of non-local matching network (NLMN) is a 3 × 3 convolutional layer with a ReLU activation function, and its output channel number is set to 64. The number K of NLMRB is 2, and the

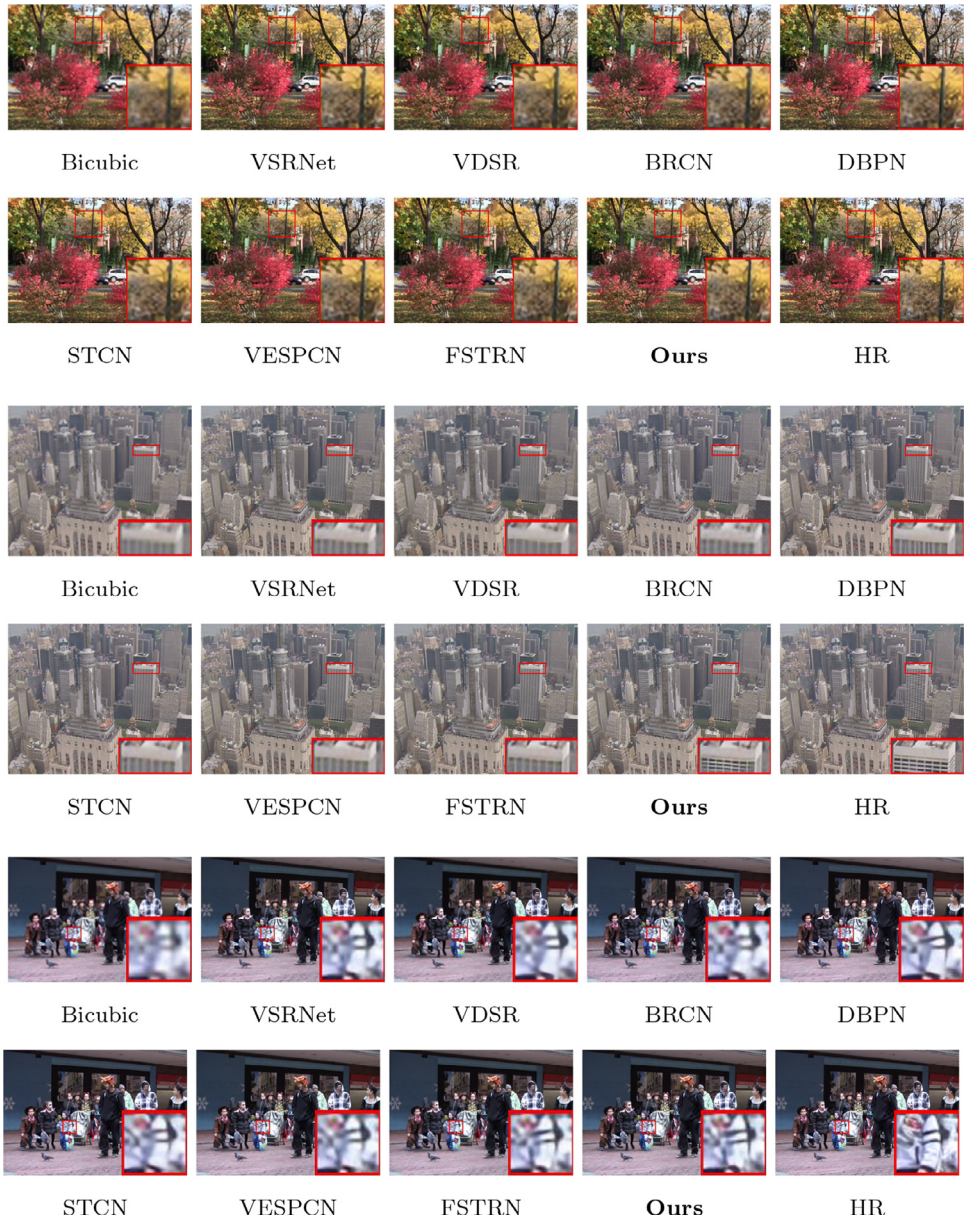


Fig. 7. Visual super-resolution comparisons of videos "Foliage", "City" and "Walk" on 4 × upscale factors.

L is set to 8. The input channel number of the up-sample layer is 548. The up-sample layer is composed by two stacked transposed convolutional layers with parametric rectified linear unit (PReLU) functions, and each transposed convolutional layer can up-sample 2 × and the output channel is set to 256. In the reconstruction layer, we convert the 256-channel features to wavelet space by a 3 × 3 convolution layer. Due to some existing methods VSRNet, BRCN, STCN and FSTRN have not published training codes yet, we re-implement their methods using Pytorch.

4.6.2. Training details

In our implementation, we randomly crop 320 × 320 patches in each frame as the ground-truth dataset, and down-sample the patches to 80 × 80 as the input LR patches for training. Hence, the output size of the temporal matching network is 40 × 40. Our STMN is optimized by Adam [40] with an initial learning rate 0.0002, and we decrease the learning rate with 0.95 scale after each 10 epochs. The training process is stopped at 500 epochs. For the training loss, we adopt the mean square error (MSE). We convert image to the YCbCr space and train our network on the individual Y, Cb and Cr channels are directly processed by a bicubic interpolation. Each epoch costs about 2.5 min, and the total training time of our network is 24 h. The average running time on Vid4 dataset is 15 ms. All experiments are performed on four NVIDIA Titan XP GPUs.

5. Conclusion

In this paper, a novel video super-resolution framework in wavelet domain is proposed. It contains a temporal fusion wavelet network (TFWN), a non-local matching network (NLMN), and a global wavelet domain residual connection (GWDRC). In TFWN, we replace motion compensations with three dimensional convolutional layers to extract temporal fusion wavelet maps. In NLMN,

a novel non-local matching residual block (NLMRB) is presented, which aims to combine super-resolution and denoising in a unified framework in wavelet domain. Besides, GWDRC is devised to force our network to learn fine details. Extensive experiments conducted on publicly available datasets demonstrate the state-of-the-art performance of our method. However, the optimization for each band is still time-consuming. Our future work is to train a network to predict the sub-bands in wavelet domain, so that the computational complexity may be decreased.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

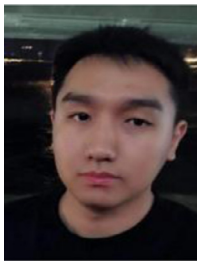
Acknowledgments

This work was supported by National Key R&D Program of China (2019YFB1405900), National Natural Science Foundation of China (61802123, 61806044, 61602517, and 61871378), and the Natural Science Foundation of Zhejiang Province (LR20F020002).

References

[1] A. Kappeler, S. Yoo, Q. Dai, A.K. Katsaggelos, Video super-resolution with convolutional neural networks, IEEE Trans. Comput. Imaging 2 (2) (2016) 109–122.
[2] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.
[3] Y. Huang, W. Wang, L. Wang, Video super-resolution via bidirectional recurrent convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 1015–1028.
[4] M. Haris, G. Shakhnarovich, N. Ukita, Deep back-projection networks for super-resolution, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1664–1673.
[5] J. Guo, H. Chao, Building an end-to-end spatial-temporal convolutional network for video super-resolution, in: AAAI, 2017, pp. 4053–4060.
[6] J. Caballero, C. Ledig, A.P. Aitken, A. Acosta, J. Totz, Z. Wang, W. Shi, Real-time video super-resolution with spatio-temporal networks and motion compensation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2848–2857.
[7] S. Li, F. He, B. Du, L. Zhang, Y. Xu, D. Tao, Fast spatio-temporal residual network for video super-resolution, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019. 1–1
[8] L. Zhang, H. Zhang, H. Shen, P. Li, A super-resolution reconstruction algorithm for surveillance images, Signal Process. 90 (3) (2010) 848–859.
[9] H. Okuhata, R. Imai, M. Ise, R.Y. Omaki, H. Nakamura, S. Hara, I. Shirakawa, Implementation of dynamic-range enhancement and super-resolution algorithms for medical image processing, in: IEEE International Conference on Consumer Electronics, 2014, pp. 181–184.
[10] Z. Qian, K. Huang, Q. Wang, J. Xiao, R. Zhang, Generative adversarial classifier for handwriting characters super-resolution, Pattern Recognit. 107 (2020) 1–12.
[11] Z. Jiang, T. Wang, J. Yan, Unifying offline and online multi-graph matching via finding shortest paths on supergraph, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (1) (2020). 1–1
[12] X. Zhu, Z. Li, X.-Y. Zhang, C. Li, Y. Liu, Z. Xue, Residual invertible spatio-temporal network for video super-resolution, in: AAAI, 2019, pp. 3897–3906.
[13] C. Xie, Y. Wu, L. van der Maaten, A.L. Yuille, K. He, Feature denoising for improving adversarial robustness, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 501–509.
[14] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.
[15] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2016) 295–307.
[16] C.a. Cruz, R. Mehta, V. Katkovnik, K.O. Egiazarian, Single image super-resolution based on wiener filter in similarity domain, IEEE Trans. Image Process. 27 (3) (2017) 1376–1389.
[17] X. Luo, Y. Xu, J. Yang, Multi-resolution dictionary learning for face recognition, Pattern Recognit. 93 (2019) 283–292.
[18] X. He, Z. Mo, P. Wang, Y. Liu, M. Yang, J. Cheng, Ode-inspired network design for single image super-resolution, in: CVPR, 2019, pp. 1732–1741.
[19] K. Nguyena, C. Fookes, S.S.M. Tistarelli, M. Nixonc, Super-resolution for biometrics: a comprehensive survey, Pattern Recognit. 78 (2018) 23–42.
[20] Y. Lu, Z. Lai, X. Li, D. Zhang, W.K. Wong, C. Yuan, Learning parts-based and global representation for image classification, IEEE Trans. Circuits Syst. Video Technol. 28 (12) (2017) 3345–3360.
[21] L. Wang, Z. Huang, Y. Gong, C. Pan, Ensemble based deep networks for image super-resolution, Pattern Recognit. 68 (2017) 191–198.
[22] W. Shi, J. Caballero, F. Huszar, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1874–1883.
[23] D. Mitzel, T. Pock, T. Schoenemann, D. Cremers, Video super resolution using duality based TV-L1 optical flow, in: Joint Pattern Recognition Symposium, 2009, pp. 432–441.
[24] C. Liu, D. Sun, A Bayesian approach to adaptive video super resolution, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 209–216.
[25] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, X. Wang, T.S. Huang, Learning temporal dynamics for video super-resolution: a deep learning approach, IEEE Trans. Image Process. 27 (7) (2018) 3432–3445.
[26] X. Tao, H. Gao, R. Liao, J. Wang, J. Jia, Detail-revealing deep video super-resolution, in: International Conference on Computer Vision, 2017, pp. 4482–4490.
[27] W. Yang, J. Feng, G.-S. Xie, J. Liu, Z. Guo, S. Yan, Video super-resolution based on spatial-temporal recurrent residual networks, Comput. Vis. Image Underst. 168 (2018) 79–92.
[28] M. Haris, G. Shakhnarovich, N. Ukita, Recurrent back-projection network for video super-resolution, in: IEEE Computer Society Conference on Computer Vision and Pattern, 2019, pp. 3897–3906.
[29] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, X. Wang, T.S. Huang, Learning temporal dynamics for video super-resolution: a deep learning approach, IEEE Trans. Image Process. 27 (7) (2018) 3432–3445.
[30] H. Huang, R. He, Z. Sun, T. Tan, Wavelet domain generative adversarial network for multi-scale face hallucination, Int. J. Comput. Vis. 127 (6–7) (2019) 763–784.
[31] S. Zhao, H. Han, S. Peng, Wavelet-domain HMT-based image super-resolution, in: International Conference on Image Processing, 2003, pp. 953–956.
[32] H. Ji, C. Fermüller, Robust wavelet-based super-resolution reconstruction theory and algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 31 (4) (2009) 649–660.
[33] T. Guo, H.S. Mousavi, T.H. Vu, V. Monga, Deep wavelet prediction for image super-resolution, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1100–1109.
[34] Z. Zhong, T. Shen, Y. Yang, Z. Lin, C. Zhang, Joint sub-bands learning with clique structures for wavelet domain super-resolution, in: Advances in Neural Information Processing Systems, 2018, pp. 165–175.
[35] T. Guo, H.S. Mousavi, V. Monga, Adaptive transform domain image super-resolution via orthogonally regularized deep networks, IEEE Trans. Image Process. 28 (9) (2019) 4685–4700.
[36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[37] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2012) 221–231.
[38] D. Han, J. Kim, J. Kim, Deep pyramidal residual networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6307–6315.
[39] A. Ranjan, M.J. Black, Optical flow estimation using a spatial pyramid network, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2720–2729.
[40] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2015. 1–1

Xiaobin Zhu received the M.Sc. degree from Beijing Normal University in 2006 and the Ph.D. degree from the Institute of Automation, Chinese cademy of Sciences, in 2013. He is currently an Associate Professor with the School of Computer and Communication Engineering, University of Science and Technology Beijing. His research interests include machine learning, image/video super-resolution, image content analysis, multimedia information indexing and retrieval, and so on.



Zhuangzi Li received the B.Sc. degree from the Institute of Disaster Prevention Science and Technology in 2016. He is pursuing M.Sc. degree in School of Computer and Information Engineering, Beijing Technology and Business University. His research interests include machine learning, network technique, pattern recognition, and image processing.

Jungang Lou (b. Feb 10, 1982) received the B.S. degree in Mathematics from Zhejiang Normal University, China, in 2003, and the M.S. degree in computational mathematics and the Ph.D. degree in computer science and technology from Tongji University, Shanghai, China, in 2006 and 2010, respectively. He is currently a Professor with the School of Information Engineering, Huzhou University, Huzhou, China. He also holds a postdoctoral position at the Institute of Cyber-Systemsand Control, School of Control Science and Engineering, Zhejiang University, Zhejiang, China. He was a Visiting Scholar with the department of Computer Science at The University of Texas at San Antonio between Nov. 2017 and May 2018 (advisor Professor Qi Tian, IEEE Fellow). His current research interests include artificial intelligence and information security. He has published over 80 papers in refereed international journals including Journal of Network and Computer Applications, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybnetics, IEEE Transactions on Reliability and so on.

Qing Shen (b. Apr9, 1982) received the B.S. degree in Computer science and technology and the M.S. degree in computer applications technology from North Universit of China, in 2004 and 2007, respectively. She is currently an associate professor with the School of Information Engineering, Huzhou University, Huzhou, China. Her current research interests include intelligent information processing and information security. She has published over 30 papers in refereed international journals including Journal of Network and Computer Applications, IEEE Transactions on Neural Networks and Learning Systems, Neurocomputing, Softcomputing and so on.