Frame and Feature-Context Video Super-Resolution

Bo Yan*, Chuming Lin, Weimin Tan
School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
{byan, cmlin17, wmtan14}@fudan.edu.cn

Abstract

For video super-resolution, current state-of-the-art approaches either process multiple low-resolution (LR) frames to produce each output high-resolution (HR) frame separately in a sliding window fashion or recurrently exploit the previously estimated HR frames to super-resolve the following frame. The main weaknesses of these approaches are: 1) separately generating each output frame may obtain high-quality HR estimates while resulting in unsatisfactory flickering artifacts, and 2) combining previously generated HR frames can produce temporally consistent results in the case of short information flow, but it will cause significant jitter and jagged artifacts because the previous super-resolving errors are constantly accumulated to the subsequent frames.

In this paper, we propose a fully end-to-end trainable frame and feature-context video super-resolution (**FFCVSR**) network that consists of two key sub-networks: local network and context network, where the first one explicitly utilizes a sequence of consecutive LR frames to generate local feature and local SR frame, and the other combines the outputs of local network and the previously estimated HR frames and features to super-resolve the subsequent frame. Our approach takes full advantage of the inter-frame information from multiple LR frames and the context information from previously predicted HR frames, producing temporally consistent high-quality results while maintaining real-time speed by directly reusing previous features and frames. Extensive evaluations and comparisons demonstrate that our approach produces state-of-the-art results on a standard benchmark dataset, with advantages in terms of accuracy, efficiency, and visual quality over the existing approaches.

The goal in image and video super-resolution (SR) is to reconstruct a high-resolution (HR) image or video from its down-sampled low-resolution (LR) version. Super-resolution approaches commonly serve as an important step for a variety of computer vision applications including image and video compression (Li et al. 2017; Kappeler et al. 2016a), medical imaging (Yang et al. 2012), object recognition (Yang et al. 2018), satellite imaging (Demirel and Anbarjafari 2011), face recognition (Gunturk et al. 2003), *etc*. To recover high-frequency details, single image super-resolution needs to fully exploit spatial statistics, while temporal correlations from multiple input frames are required to be exploited in order to improve reconstruction in the case of video super-resolution. Therefore, how to effectively exploit temporal redundancies becomes the key issue for video super-resolution.

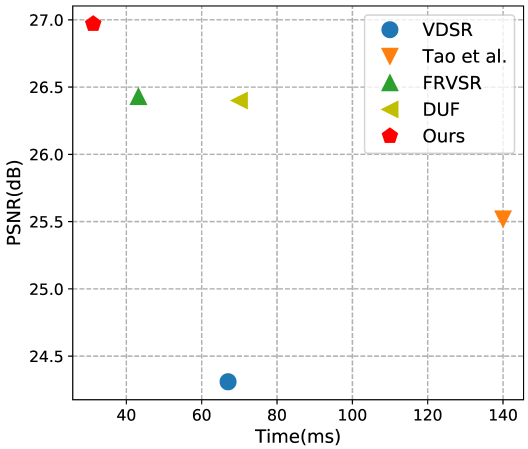

Figure 1: The proposed approach consistently outperforms state-of-the-art video super-resolution methods in terms of reconstruction quality and efficiency (x4 SR on Vid4).

Recent advances in video super-resolution are remarkable, benefiting mostly from the successful application of Deep Convolutional Neural Networks (DCNNs). However, there is still a large room for improvement over the DCNN based video super-resolution (SR) models that do not consider the super-resolution quality and temporal consistency simultaneously. The latest state-of-the-art approaches (Dong, Chen, and Tang 2016; Kim, Lee, and Lee 2016; Liu and Sun 2011; Liao et al. 2015; Kappeler et al. 2016b; Caballero et al. 2017; Tao et al. 2017; Jo et al. 2018) formulate the task of video super-resolution as a great deal of separate multi-frame super-resolution subtasks. They exploit a sequence of consecutive LR frames to generate a single HR estimate, focusing on obtaining high-quality reconstruction results for each single frame. However, the way of separately generating each HR estimate results in temporally inconsistent frames, producing unsatisfactory flickering artifacts. In

*This work was supported by NSFC (Grant No.: 61772137; 61522202).

addition, such approaches have high computational cost because each input frame is processed several times.

To address the above issues, (Sajjadi, Vemulapalli, and Brown 2018) proposed a frame-recurrent video super-resolution model that recurrently exploits the previously estimated HR frames to super-resolve the subsequent frame. This approach is able to generate temporally consistent result because the current super-resolving frame will refer to those previously HR estimates. However, only referring to previously inferred HR frames will produce significant jitter and jagged artifacts because the previous super-resolving errors are constantly accumulated to the subsequent frames.

The above issues motivate us to develop a new approach for video super-resolution by introducing frame and feature as context to simultaneously improve the super-resolution quality and temporal consistency. Specifically, we design a novel end-to-end trainable video super-resolution framework that consists of two key sub-networks: local network and context network. The local network explicitly utilizes a batch of LR frames to generate local feature and local SR frame. Then, the context network combines the outputs of local network and the previously estimated HR frames and features to super-resolve the subsequent frame, which guides the network learning alignment between frames to maintain consistency. This framework offers three advantages:

- The inter-frame information from multiple LR frames can be effectively exploited by local network to generate high-quality SR frames (local SR frame) and reference features (local feature) that provides the context network higher-quality data to work with.
- By utilizing the context information from previously predicted HR frames and features and the outputs of local network, our framework naturally encourages the video super-resolution model to generate temporally consistent results, making it to learn alignment between SR frames.
- It has low computational cost due to its recurrent nature of using previous frames and features and no motion compensation block.

Benefiting from the property of combining context information from previous frames and features, the resulting architecture produces the most consistent results while containing finer details in each SR frame. Our model is fully convolutional and no other prior information such as optical flow estimation and motion compensation. To demonstrate the effectiveness of the proposed framework, we conduct ablation study for analyzing the importance of each component of our model. Besides, we compare our FFCVSR with several latest video super-resolution approaches and show that it produces state-of-the-art results on a standard benchmark dataset, with advantages in terms of accuracy, speed, and visual quality over the existing algorithms (see Fig. 1). Furthermore, based on the characteristics of our framework, we propose a suppression-updating algorithm to effectively solve the problem of error accumulation of high frequency information. Finally, we also apply our trained model to real scenes to demonstrate its good abilities of generalization and practicability.

Related Works

Over the past decades, a large number of image and video super-resolution approaches have been developed, ranging from traditional image processing methods such as Bilinear and Bicubic interpolation to example-based frameworks (Timofte, De, and Gool 2014; Jeong, Yoon, and Paik 2015; Xiong et al. 2013; Freedman and Fattal 2011), self-similarity methods (Huang, Singh, and Ahuja 2015; Yang, Huang, and Yang 2010), and dictionary learning (Perezpellitero et al. 2016). Some efforts have devoted to study different loss functions for high-quality resolution enhancement (Sajjadi, Scholkopf, and Hirsch 2017). A complete survey of these approaches is beyond the scope of this work. Readers can refer to a recent survey (Walha et al. 2016; Agustsson and Timofte 2017) on super-resolution approaches for details. Here, we focus on discussing recent video super-resolution approaches based on deep network.

Benefiting from the explosive development of convolutional neural network (CNN) in deep learning, CNN based approaches have refreshed the previous super-resolution state-of-the-art records. Since (Dong et al. 2014) uses a simple and shallow CNN to implement single super-resolution and achieves state-of-the-art results, following this fashion, numerous works have proposed various deep network architectures. Most of the existing CNN based video super-resolution approaches regard the task of video super-resolution as a large number of separate multi-frame super-resolution subtasks. They exploit a sequence of consecutive LR frames to generate a single HR estimate. (Kappeler et al. 2016b) uses an optical flow method to warp video frames LR_{t-1} and LR_{t+1} onto the frame LR_t. Then, these three frames are combined to feed into a CNN model that outputs the HR frame SR_t. Similar to (Kappeler et al. 2016b), (Caballero et al. 2017) uses a trainable motion compensation network to replace the optical flow method in (Kappeler et al. 2016b). Following this fashion, Tao *et al.* (Tao et al. 2017) propose a network comprising motion estimation, motion compensation, and detail fusion to process a batch of LR frames and output HR estimate.

Different from the above mentioned approaches, (Sajjadi, Vemulapalli, and Brown 2018) proposes a frame recurrent video super-resolution (FRVSR) framework that combines the previous HR estimates to generate subsequent frame. This method warps the SR_{t-1} frame onto the SR_t based on the optical flow information estimated from LR_{t-1} and LR_t. Then, it uses a trainable super-resolution network to fuse the warped SR_{t-1} and LR_t, yielding the SR_t frame. Therefore, there are two loss items in their loss function, the mean squared error between SR_t and HR_t, and the warped LR_{t-1} and LR_t. The FRVSR has advantage of producing temporally consistent results in the case of short information flow, but it will cause jitter and jagged artifacts because the previous super-resolving errors are constantly accumulated to the subsequent frames.

Though significant progress have been achieved by these studies in recent years, there is still a large room for improvement over the CNN based video super-resolution approaches that do not consider the super-resolution quality and temporal consistency simultaneously.

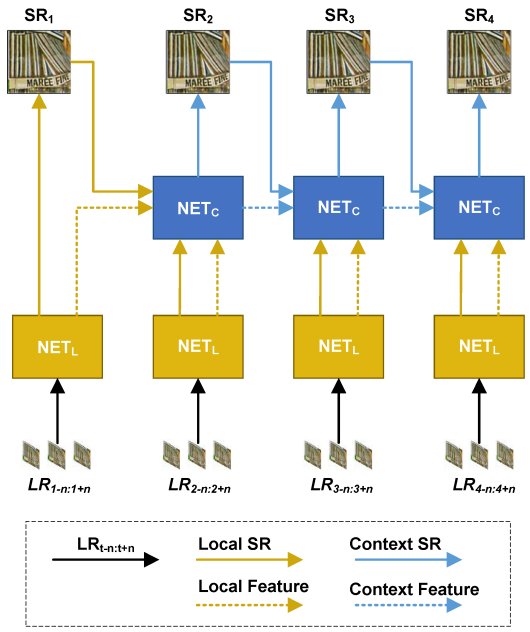

Figure 2: Overview of the proposed FFCVSR framework. It consists of two trainable components: local network NET_L (shown in yellow) and context network NET_C (shown in blue). The NET_L produces local frame SR_t^{Local} and local feature F_t^{Local} by processing a sequence of LR frames. Then, the NET_C outputs the super-resolved result SR_t and an additional output F_t. During training, the loss is applied on the output of NET_L and NET_C, and back-propagated through both NET_L and NET_C for jointly training them.

Method

To overcome the aforementioned problems, we propose a frame and feature-context video super-resolution (**FFCVSR**) approach to reasonably combine both previous frames and features for accurate and fast video super-resolution. We will dedicate to state the proposed approach in detail in following subsections.

FFCVSR Framework

1. Overview of the Proposed FFCVSR Framework: To better understand our FFCVSR, we start out with the introduction of FFCVSR architecture, as illustrated in Fig. 2. It consists of two trainable components: local network NET_L (shown in yellow) and context network NET_C (shown in blue). Given a sequence of LR frames, the local network NET_L outputs local frame SR_t^{Local} and local feature F_t^{Local} by exploiting inherent inter-frame information in the form of local correlations, helping the following context network NET_C to recover lost high-frequency details. Considering the super-resolved results should maintain temporal consistency, the context network NET_C not only exploits the local frame SR_t^{Local} and previous SR frame SR_t but also combines the local feature F_t^{Local} and previous SR feature F_t, yielding visually pleasing and temporally consistent results. We will investigate the importance of each component by performing ablation study in the experiment, providing several insights for further designing better video super-resolution approach. Note that our FFCVSR framework has no motion compensation module commonly used in previous methods, which has additional advantage of reducing the computational cost. This processing flow is summarized in Algorithms 1.

Algorithm 1 Frame and Feature-Context Video Super-Resolution

Input: A sequence of consecutive LR frames, $LR_{t-n:t+n}$. T is the updating step. T = 50 in our experiment.
Output: Estimated high-resolution frame, SR_t.

```
for t = 1 → VideoLen do
  if t == 1 then
    SR_t ← SR_t^Local
    F_t ← F_t^Local
  else
    (SR_t, F_t) ← NET_C(SR_{t-1}, F_{t-1}, SR_t^Local, F_t^Local)
  end if
end for
% Suppression updating algorithm
if t mod T == 0 then
  SR_{t-1} ← SR_t^Local
  F_{t-1} ← F_t^Local
else
  SR_{t-1} ← SR_t
  F_{t-1} ← F_t
end if
```

2. Architecture of Local Network: The proposed local network NET_L is shown in Fig. 4. It exploits inherent inter-frame information in the form of local correlations and outputs local frame and feature by processing a sequence of LR frames. For demonstration convenience, we only show three consecutive LR frames including current frame that needs to be super-resolved. Our simple NET_L consists of 5 convolutions (kernel size=3 $\times$ 3, stride=1), 1 deconvolution (kernel size=8 $\times$ 8, stride=4), and 8 ResBlocks (Lim et al. 2017). The ResBlock in purple (shown on the right side of Fig. 3) contains two convolutions with skip connection. We use the sum of the deconvolution result and the Bicubic interpolation result of LR_t as the output SR_t^{Local}. The output F_t^{Local} is produced by adding a new side output with two convolution operations. Let $T = (LR_t, HR_t), t = 1, \ldots, N$ denotes the training data set, where LR_t is the input LR frame and HR_t denotes the corresponding ground truth high-resolution frame. We use W_L to denote the collection of all network layer parameters in NET_L. Thus, the local frame and feature can be given by:

$$SR_t^{Local}, F_t^{Local} = NET_L(LR_{t-1}, LR_t, LR_{t+1}; W_L). \quad (1)$$

3. Architecture of Context Network: The proposed context network NET_C is shown in Fig. 3. It produces the HR estimate SR_t and feature F_t by exploiting the context information from previously predicted HR frames and features (SR_{t-1}, F_{t-1}) and the outputs of local network

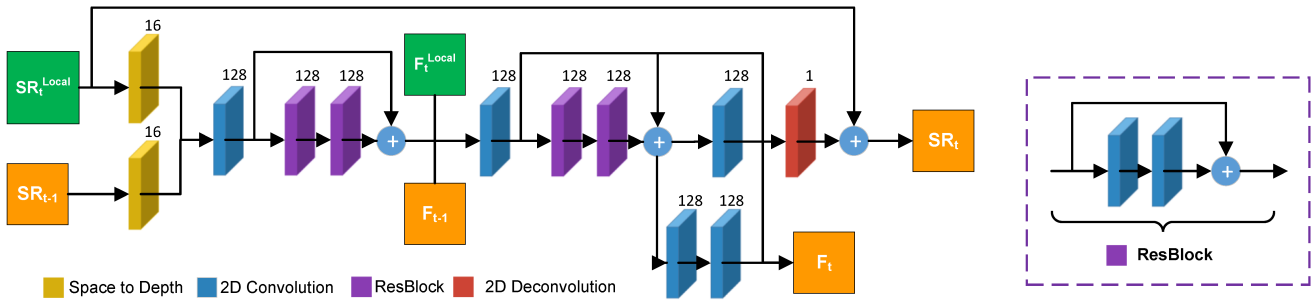

Figure 3: Architecture of the proposed NET_C. It produces the HR frame SR_t and feature F_t by exploiting the context information from previously predicted HR frames and features ($i.e.$, SR_{t-1}, F_{t-1}) and the outputs of NET_L ($i.e.$, $SR_t^{Local}, F_t^{Local}$).

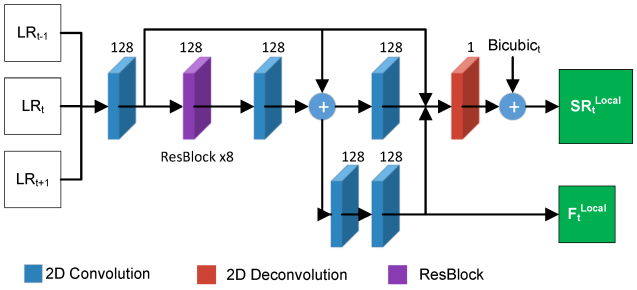

Figure 4: Architecture of the proposed local network NET_L. It processes a sequence of LR frames to output local frame SR_t^{Local} and local feature F_t^{Local}.

$(SR_t^{Local}, F_t^{Local})$, where the context information means that generating HR estimate will refer to previous HR frames and features to maintain temporal consistency. Our NET_C consists of 5 convolutions (kernel size=3×3, stride=1), 1 deconvolution (kernel size=8 × 8, stride=4), 4 ResBlocks, and 2 space-to-depth transformations (shown in yellow) (Sajjadi, Vemulapalli, and Brown 2018). Here, we use space-to-depth transformation to reduce the computational cost. We use the sum of the deconvolution result and local frame SR_t^{Local} as the final output SR_t. We also provide another output of feature F_t for super-resolving subsequent frame by adding a new side output with two convolution operations. We use W_C to denote the collection of all network layer parameters in NET_C. Thus, the estimated HR frame and feature can be given by:

$$SR_t, F_t = NET_C(SR_{t-1}, F_{t-1}, SR_t^{Local}, F_t^{Local}; W_C). \tag{2}$$

Loss Function

The proposed local network NET_L and context network NET_C in our FFCVSR framework are seamlessly combined and jointly trained with the loss function defined as:

$$Loss(W_L, W_C) = \left\|SR_t^{Local} - HR_t\right\|_2^2 + \|SR_t - HR_t\|_2^2. \tag{3}$$

The loss is applied on the output of NET_L and NET_C, and back-propagated through both NET_L and NET_C. Note that there is no need for defining additional loss function to constrain the output of features F_t^{Local} and F_t, because as training progresses, both of them gradually provide high-quality data required by the NET_C network.

Suppression Updating Algorithm

There is a key observation that the super-resolved video has significant jitter and jagged artifacts when using the previously inferred HR frames as reference information to generate subsequent frame, because the previous super-resolving errors are constantly accumulated to the subsequent frames. The Fig. 10 provides intuitive image examples for showing this observation. To overcome this problem, based on the characteristics of our FFCVSR framework, we propose a simple suppression-updating algorithm to effectively solve the problem of error accumulation of high frequency information. Specifically, we replace the SR_{t-1} and F_{t-1} outputted by NET_C with SR_t^{Local} and F_t^{Local} outputted by NET_L at each interval of T frames, respectively (see also Algorithms 1), because after several iterations, the outputs of NET_C have accumulated a considerable amount of super-resolving error while the outputs of NET_L still maintain accurate information from current LR frame without introducing accumulative error from previous SR frames. In the experiment, we observe that $T = 50$ can produce favorable results.

Training and Inference

Our training dataset consists of 2 high-resolution videos (4k, 60fps): *Venice* and *Myanmar* downloaded from harmonic[1]. The lengths of these two videos are 1,077 seconds and 527 seconds, respectively. We select them as training set because they contain more than 140 different scenes including human, natural scene, building, traffic, $etc.$ To produce HR videos, we firstly downscale the original videos by factors of 4 (960 × 540), 6 (640 × 360), 8 (480 × 270), 12 (320 × 180), and 16 (240×135) to obtain the high-resolution ground truth with a variety of receptive fields. Then, we extract patches of size 128 × 128 to produce the HR videos. To produce the

[1]https://www.harmonicinc.com/free-4k-demo-footage/

input LR videos, we downsample them to the original 1/4 size using bilinear interpolation.

During training, we extract clips of 10 consecutive frames from the videos. We avoid the clips containing keyframes that have large scene changes. The extracted LR patches are randomly flipped horizontally and vertically for data augmentation. Besides, the order of sequences is also randomly reversed. We employ the brightness channel y to train the proposed model. The parameters are updated with initial learning rate of 10^{-4} before 300K iteration steps and changed to 10^{-5} at the following 50K. The loss is minimized using Adam optimizer (Kingma and Ba 2015) and back-propagated through both networks NET_L and NET_C as well as through time. After repeatedly minimizing the loss on the training data, the resulting network is capable of directly producing the full video frames, without needing any additional post-processing operations.

When super-resolving the first frame $SR_{t=1}$ in each clip, the local network NET_L upsamples it at both training and testing time. At the same time, we regard the local frame and feature as the previously inferred frame and feature and feed them into the the context network NET_C to produce SR_t and F_t. This simple technique that reuses the outputs of NET_L to deal with the first frame without prior information can encourage the network to exploit local information from LR frames during early training instead of only depending on the previously inferred HR estimates. Our architecture is fully end-to-end trainable and does not require pre-training sub-networks.

During inference, the trained model can process videos with arbitrary length and size due to the fully convolutional property of the networks. We can obtain the enhanced video by performing a single feedforward inference over frame by frame. In the following section, we report the reconstruction accuracy, efficiency, and visual quality of the model.

Experiments and Analyses

In this section, we introduce compared methods and utilized dataset, and report the performance of our proposed approach. Firstly, we conduct ablation experiments to investigate the importance of each component of our approach, providing insight into how the performance of our FFCVSR varies with context information. Then, we compare our approach with current state-of-the-art methods on the standard **Vid4** benchmark dataset (Liu and Sun 2011) in terms of visual quality, objective metric, temporal consistency, and computational cost. Following (Caballero et al. 2017), the evaluation metrics of Peak signal-to-noise ratio (**PSNR**) and structural similarity (**SSIM**) are computed on the brightness channel on the 4-video dataset Vid4. Thirdly, we detail the suppression-updating algorithm for depressing iteration error of high-frequency information. Finally, real-world examples are provided to verify the effectiveness of our approach. All experiments are carried out for 4x upscaling.

We conduct our experiments on a machine with an Intel i7-7700k CPU and an Nvidia GTX 1080Ti GPU. Our framework is implemented on the TensorFlow platform.

Table 1: Experimental result of ablation study. Average video PSNR of different architectures on Vid4.

Methods	Calendar	City	Walk	Foliage	average
SISR	21.789	25.926	27.742	24.429	24.971
only local network	23.206	27.166	29.465	25.713	26.387
w/o feature context	23.601	27.393	29.908	26.079	26.745
w/o feature context+ with optical flow	23.528	27.405	29.842	26.044	26.705
Full model	**23.828**	**27.564**	**30.172**	**26.296**	**26.965**

Ablation Analysis

Our architecture consists of two key components: local network NET_L and context network NET_C. We experiment with different design options to illustrate the contribution of each component to the video super-resolution result in terms of objective metric and visual quality. To explore the performance of local network in the proposed architecture shown in Fig. 2, we remove the context network and use "only local network" to denote the resulting model, $i.e.$, blue arrows are removed in Fig. 2. Besides, we also explore the effectiveness of utilizing features including local feature (blue dashed arrow in Fig. 2) and context feature (yellow dashed arrow in Fig. 2) in the proposed framework. Thus, we remove the local feature and context feature from the architecture and use "w/o feature context" to denote it. Finally, we use "Full model" to denote our complete model including NET_L and NET_C. Since optical flow method (Sajjadi, Vemulapalli, and Brown 2018) is widely employed in prior art, we incorporate it into our architecture to test whether it improves the recovering ability of our model. Here, we use "w/o feature context + with optical flow" to denote the model that removes feature information and introduces optical flow method. We also compare our model with a single image super-resolution (SISR) baseline, which is obtained by only feeding the current frame LR_t into the local network.

Quantitative Comparison: The quantitative results are reported in Table 1. The "only local network" model that exploits temporal information from input consecutive frames outperforms SISR baseline, which demonstrates that exploiting temporal redundancies is helpful to recover high-frequency details for video super-resolution. The "w/o feature context" model utilizing previously estimated HR frames further improves the performance of "only local network". This result implies that propagating information from previous HR frames to the following step helps the model to recover lost fine details. The complete model "Full model", simultaneously exploiting previously inferred frames and features, obtains the best results on all videos from Vid4. The result well demonstrates the effectiveness of introducing local and context features.

Compared with "w/o feature context", the "w/o feature context + with optical flow" method incorporating optical flow component leads to a slight decrease in PSNR. The possible reason is that the convolutional kernels are better at learning motion information from consecutive frames than optical flow method because of small motion of pixels in consecutive frames. We observe that directly using convolutional kernels to learn motion information instead of optical

Table 2: Quantitative comparison with state-of-the-art approaches. Values marked with a star are referenced from the corresponding publications. Obviously, our approach outperforms other methods in terms of PSNR, SSIM, and computational cost.

Methods	BayesSR*	DESR*	VSRNet*	VESPCN*	VDSR*	Tao et al.*	FRVSR	DUF	**Ours**
x4 PSNR	24.42	23.50	22.81	25.35	24.31	25.52	26.43	26.40	**26.97**
x4 SSIM	0.72	0.67	0.65	0.76	0.67	0.76	0.80	0.80	**0.83**
time (ms)	-	-	-	-	73.2	140	43.2	70	**31.2**

Figure 5: A visual comparison of oblation study. The "Full model" produces the best result having fine details.

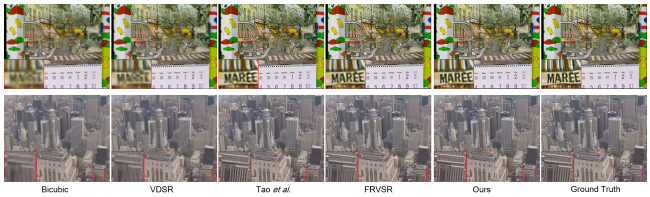

Figure 6: Visual comparison with state-of-the-art approaches (x4 SR).

flow method not only improves the reconstruction accuracy of the model, but also decreases its computational cost. All the above experiments show that the proposed architecture for video super-resolution is reasonable and appropriate.

Visual Comparison: Figure 5 shows a visual comparison of oblation study with different design options for our framework. We can observe that the "Full model" is capable of recovering finer details and generating visually satisfactory results. Compared with "only local network" and "w/o feature context" methods, the recovered result produced by "Full model" is both sharper and closer to the ground truth, as shown in the white snow in Fig. 5.

Comparison with prior art

We compare the proposed approach with various state-of-the-art video super-resolution methods, including **VDSR** (Kim, Lee, and Lee 2016), **BayesSR** (Liu and Sun 2011), **DESR** (Liao et al. 2015), **VSRNet** (Kappeler et al. 2016b), **VESPCN** (Caballero et al. 2017), **Tao *et al.*** (Tao et al. 2017), **FRVSR** (Sajjadi, Vemulapalli, and Brown 2018), and **DUF** with 16 layer (Jo et al. 2018) on the Vid4 benchmark

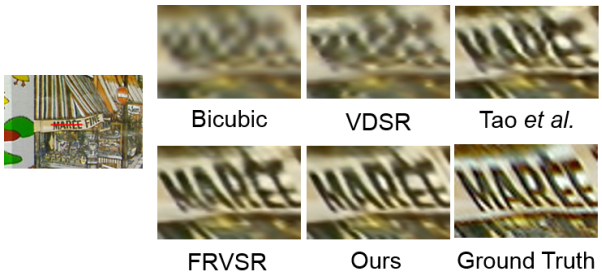

Figure 7: Demonstration of temporal profiles for comparing temporal consistency of different approaches.

dataset in terms of PSNR and SSIM. For all competing approaches except FRVSR and DUF, the PSNR and SSIM values are directly referenced from the corresponding publications by authors. Since FRVSR and DUF are the newest approaches, we implement them on the TensorFlow platform. For fair comparison, these two implements are trained and tested on the identical dataset used by our approach.

Quantitative Comparison: Table 2 reports the PSNR and SSIM produced by our approach and previous state-of-the-art methods on Vid4. It is obvious that our proposed approach substantially outperforms the current state-of-the-art methods by a large margin in terms of reconstruction accuracy and efficiency. Comparing with the current best results, our approach surpasses them by more than 0.5 dB in PSNR and 0.03 score in SSIM. This implies that our approach produces the most accurate result and our architecture is reasonable and appropriate for video super-resolution.

Quality Comparison: Fig. 6 demonstrates quality comparison of different approaches. From the close-up images, we observe that the proposed approach produces better structural detail than other competing methods. This result indicates that our strategy of exploiting previously inferred information in terms of frame and feature is essential such that the resultant SR images look much closer to the ground truth.

Temporal Consistency

To compare temporal consistency of different approaches, following (Caballero et al. 2017), we use temporal profile to show the result on paper. Fig. 7 reports a temporal profile on the row highlighted by a red line across a number of frames. While (Tao et al. 2017) generates better results than VDSR method, it still contains considerable flickering artifacts due to separately estimating each output frame. By referring previous frames, FRVSR has improved a lot in the temporal consistency, but it has some blurs compared with the ground

Figure 8: Real-world examples to evaluate the practical ability of different approaches.

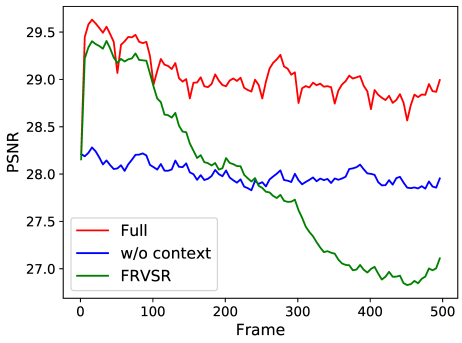

Figure 9: Performance of FRVSR, our full model, and "only local network" on *HongKong* as a function of the number of previous frames processed. Our suppression-updating algorithm can effectively depress iteration error of high-frequency information from previous frames processed.

truth. In contrast, our approach produces the most temporal coherence result that looks much closer to the ground truth.

Computational Efficiency

Figure 1 and Table 2 illustrate the comparison result of computational efficiency. Note that the running times of compared approaches including BayesSR (Liu and Sun 2011), DESR (Liao et al. 2015), VSRNet (Kappeler et al. 2016b), VESPCN (Caballero et al. 2017) are not listed in Table 2, because their running times are not stated in corresponding publications. The result clearly shows that our model is much more efficient than other approaches. It averagely takes 31.2 ms with our unoptimized TensorFlow implementation on an Nvidia GTX 1080Ti when running on Vid4 to generate a single HR image for 4x upsampling. Benefiting from directly taking advantage of previous features and frames, our approach is able to maintain real-time speed while producing high-quality temporal-coherency result.

RealWorld Examples

To evaluate the performance of our approach on real-world data, following (Caballero et al. 2017), a visual comparison result is reported in Fig. 8. From the close-up images, we observe that our approach is able to recover the fine details and

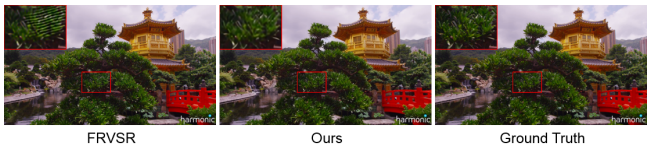

Figure 10: Illustration of iteration error of high-frequency information.

remove the blur artifacts, even though the model is trained on a set of LR-HR frame pairs, where the LR frames are obtained by performing bicubic down-sampling.

Suppressing Iteration Error of High-Frequency Information

Because the previous super-resolving errors are constantly accumulated to the subsequent frames, the super-resolved video has significant jitter and jagged artifacts when using previously inferred HR frames. Fig. 9 illustrates the performance of FRVSR, our full model, and "only local network" (without context network) on *HongKong*[2] as a function of the number of previous frames processed. It shows that the reconstruction accuracy of FRVSR approach is high in the early stage and decreased slightly in the low range of information flow (less than 100 frames), but it decreases dramatically when the number of previous frames processed is over 100, even worse than our "only local network". In contrast, benefiting from the proposed suppression-updating algorithm, our full model and "only local network" are not affected by the number of previous frames processed and both achieve stable performance. Interestingly, the "full model" outperforms "only local network" method in all frames, which intuitively demonstrates the key contribution of the context network NET_C. Fig. 10 shows a visual comparison of iteration error of high-frequency information. Our approach effectively removes the unpleasing flickering artifacts existed in FRVSR method.

Conclusion

In this paper, we presented a frame and feature-context video super-resolution approach. Instead of only exploiting multiple LR frames to separately generate each output frame, we propose a fully end-to-end trainable framework consisting of local network and context network to simultaneously utilize previously inferred frames and features. Furthermore, based on the characteristics of our framework, we propose a suppression-updating algorithm to effectively solve the problem of error accumulation of high frequency information. Extensive experiments including ablation study demonstrate that our approach significantly advances the state-of-the-art on a standard benchmark dataset and is capable of efficiently producing high-quality temporal-consistency video resolution enhancement.

[2]https://www.harmonicinc.com/free-4k-demo-footage/

References

Agustsson, E., and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'2017)*, 1122–1131.

Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2017)*, 2848–2857.

Demirel, H., and Anbarjafari, G. 2011. Discrete wavelet transform-based satellite image resolution enhancement. *IEEE Transactions on Geoscience and Remote Sensing* 49(6):1997–2004.

Dong, C.; Chen, C. L.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *Proceedings of European Conference on Computer Vision (ECCV'2014)*, 184–199.

Dong, C.; Chen, C. L.; and Tang, X. 2016. Accelerating the super-resolution convolutional neural network. In *Proceedings of European Conference on Computer Vision (ECCV'2016)*, 391–407.

Freedman, G., and Fattal, R. 2011. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG'2011)* 30(2):12:1–12:11.

Gunturk, B. K.; Batur, A. U.; Altunbasak, Y.; and Mersereau, R. M. 2003. Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing (TIP'2003)* 12(5):597.

Huang, J. B.; Singh, A.; and Ahuja, N. 2015. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2015)*, 5197–5206.

Jeong, S.; Yoon, I.; and Paik, J. 2015. Multi-frame example-based super-resolution using locally directional self-similarity. *IEEE Transactions on Consumer Electronics* 61(3):353–358.

Jo, Y.; Oh, S. W.; Kang, J.; and Kim, S. J. 2018. Deep video super-resolution network using dynamic upsampling filterswithout explicit motion compensation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2018)*.

Kappeler, A.; Yoo, S.; Dai, Q.; and Katsaggelos, A. K. 2016a. Super-resolution of compressed videos using convolutional neural networks. In *IEEE International Conference on Image Processing (ICIP'2016)*, 1150–1154.

Kappeler, A.; Yoo, S.; Dai, Q.; and Katsaggelos, A. K. 2016b. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging* 2(2):109–122.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2016)*, 1646–1654.

Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR'2015)*.

Li, T.; He, X.; Qing, L.; Teng, Q.; and Chen, H. 2017. An iterative framework of cascaded deblocking and super-resolution for compressed images. *IEEE Transactions on Multimedia (TMM'2017)* PP(99):1–1.

Liao, R.; Tao, X.; Li, R.; Ma, Z.; and Jia, J. 2015. Video super-resolution via deep draft-ensemble learning. In *IEEE International Conference on Computer Vision (ICCV'2015)*, 531–539.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Lee, K. M. 2017. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR'2017)*, 1132–1140.

Liu, C., and Sun, D. 2011. A bayesian approach to adaptive video super resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2011)*, 209–216.

Perezpellitero, E.; Salvador, J.; Ruizhidalgo, J.; and Rosenhahn, B. 2016. Psyco: Manifold span reduction for super resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2016)*, 1837–1845.

Sajjadi, M. S. M.; Scholkopf, B.; and Hirsch, M. 2017. Enhancenet: Single image super-resolution through automated texture synthesis. In *IEEE International Conference on Computer Vision (ICCV'2017)*, 4501–4510.

Sajjadi, M.; Vemulapalli, R.; and Brown, M. 2018. Frame-recurrent video super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2018)*.

Tao, X.; Gao, H.; Liao, R.; Wang, J.; and Jia, J. 2017. Detail-revealing deep video super-resolution. In *IEEE International Conference on Computer Vision (ICCV'2017)*, 4482–4490.

Timofte, R.; De, V.; and Gool, L. V. 2014. Anchored neighborhood regression for fast example-based super-resolution. In *IEEE International Conference on Computer Vision*, 1920–1927.

Walha, R.; Drira, F.; Lebourgeois, F.; Alimi, A. M.; and Garcia, C. 2016. Resolution enhancement of textual images: a survey of single image-based methods. *IET Image Processing* 10(4):325–337.

Xiong, Z.; Xu, D.; Sun, X.; and Wu, F. 2013. Example-based super-resolution with soft information and decision. *IEEE Transactions on Multimedia (TMM'2013)* 15(6):1458–1465.

Yang, J.; Wang, Z.; Lin, Z.; Cohen, S.; and Huang, T. 2012. Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing (TIP'2012)* 21(8):3467–78.

Yang, X.; Wu, W.; Liu, K.; Kim, P. W.; Sangaiah, A. K.; and Jeon, G. 2018. Long-distance object recognition with image super resolution: A comparative study. *IEEE Access* PP(99):1–1.

Yang, C. Y.; Huang, J. B.; and Yang, M. H. 2010. Exploiting self-similarities for single frame super-resolution. In *Asian Conference on Computer Vision (ACCV'2010)*, 497–510.