Adapting Image Super-Resolution State-of-the-arts and Learning Multi-model Ensemble for Video Super-Resolution

Chao Li, Dongliang He, Xiao Liu, Yukang Ding, Shilei Wen
Department of Computer Vision Technology (VIS), Baidu Inc.
{lichao40,hedongliang01,liuxiao12,dingyukang,wenshilei}@baidu.com

Abstract

Recently, image super-resolution has been widely studied and achieved significant progress by leveraging the power of deep convolutional neural networks. However, there has been limited advancement in video super-resolution (VSR) due to the complex temporal patterns in videos. In this paper, we investigate how to adapt state-of-the-art methods of image super-resolution for video super-resolution. The proposed adapting method is straightforward. The information among successive frames is well exploited, while the overhead on the original image super-resolution method is negligible. Furthermore, we propose a learning-based method to ensemble the outputs from multiple super-resolution models. Our methods show superior performance and rank second in the NTIRE2019 Video Super-Resolution Challenge Track 1.

1. Introduction

Image super-resolution (ISR) has drawn extensive attention in recent decades. Taking a low-resolution image as input, ISR aims to generate a high-resolution image with more visual details. It has many applications in various fields, such as security surveillance, medical imaging. By exploiting deep convolutional neural networks, ISR has gained remarkable success. Dong et al. proposed a three-layer convolutional neural network for ISR and achieved significant improvement over conventional methods [22, 27, 26]. Kim et al. developed a much deeper network in VDSR [9] and DRCN [10] by using gradient clipping, skip connection, or recursive-supervision to handle the difficulty of training deep network. Lim et al. employed enhanced residual blocks and residual scaling to build a very wide network EDSR [13]. In [29] a residual dense network (RDN) was proposed to fully utilize the all the hierarchical features by leveraging dense connection and residual learning. In [28], Zhang et al. designed a residual channel attention network (RCAN) to adaptively rescale channel-wise features by considering interdependencies among channels.

Although the performance of image super-resolution has been significantly advanced by these above techniques, video super-resolution still needs to be promoted because of the high demand for VSR in real applications (e.g. video on demand services and live-broadcasting platforms) and its unsatisfactory performance with regard to visual quality and computation complexity. For video super-resolution, to exploit the plentiful temporal information among successive frames, several methods has been proposed. Tao et al. [21] used motion compensation transformer module for the motion estimation, and proposed a sub-pixel motion compensation layer for simultaneous motion compensation and upsampling. Sajjadi et al. [18] extended the conventional VSR model to a frame-recurrent VSR framework by using the previously inferred high-resolution frame to super-resolve the subsequent frame. The above methods heavily rely on the accuracy of flow estimation and motion compensation which also bring extra computation cost. Jo et al. [8] introduced a fundamentally different framework for VSR. They proposed an end-to-end deep neural network that generates dynamic upsampling filters and a residual image, which were computed depending on the local spatio-temporal neighborhood of each pixel to avoid explicit motion compensation. However, they employed 3D convolution kernels to model the temporal information among successive frames, which are slow to compute comparing to using 2D convolution kernels.

To address the aforementioned issues, we propose to adapt state-of-the-art ISR methods for VSR while keep the computation cost staying in the original level. A demonstration of our framework for VSR is in Figure 1. Firstly, we optimize state-of-the-art ISR methods by analyzing and modifying their network components to enable them to handle VSR task. The adapted methods significantly outperform the original methods which process a video frame-by-frame. Secondly, we propose a learning-based method to ensemble the outputs from multiple super-resolution models which again boosts the final performance remarkably. Leveraging the adapted super-resolution methods and

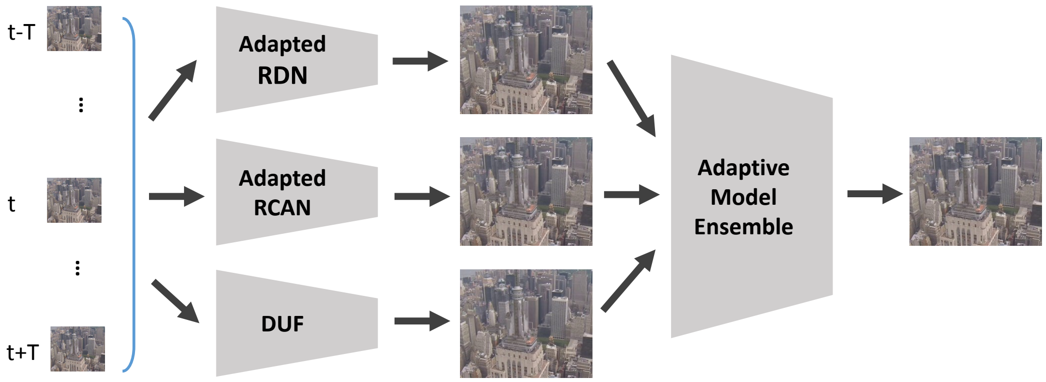

Figure 1. Demonstration of the Workflow of our framework for VSR. We adapt state-of-the-art image super-resolution frameworks, i.e., RCAN [28] and RDN [29] for video super-resolution. Besides, the 3D convolution based method DUF [8] is also used as one of our baseline model. The results generated by these models are ensembled via a novel adaptive model ensemble module to produce the final result.

the learning-based ensemble method, we ranked second in the NTIRE2019 Video Super-Resolution Challenge Track 1 [16].

2. Related Work

2.1. Image Super-resolution

Conventional image super-resolution methods use interpolation techniques based on sampling theory [11, 27], and several previous studies [20, 19] adopted natural image statistics to restore high-resolution images. However, these methods can not reconstruct satisfactory details and realistic textures. By utilizing techniques, such as neighbor embedding and sparse coding [1, 3, 6, 25, 23], some learning-based methods attempt to learn mapping functions between low-resolution and high-resolution image pairs.

Recently, with the development of deep convolutional neural networks (CNN), ISR gains dramatic improvements. Dong et al. [4, 5] propose a CNN-based super-resolution method. Afterwards, various CNN architectures have been investigated for ISR. Kim et al. [9, 10] employs the residual network for training much deeper network and achieved superior performance. In particular, they use skip-connection and recursive convolution to alleviate the burden of carrying identity information in the super-resolution network. In [14] Mao et al. utilize encoder-decoder networks and symmetric skip connections [17] to achieve fast and improved convergence. Futhermore, Lim et al. employed enhanced residual blocks and residual scaling to build a very wide network EDSR [13]. In [29] a residual dense network (RDN) was proposed to fully utilize the all the hierarchical features by leveraging dense connection and residual learning. In [28], Zhang et al. designed a residual channel attention network (RCAN) to adaptively rescale channel-wise features by considering interdependencies among channels. DBPN [7] exploits iterative up and down sampling layers to build an error feedback mechanism for errors projection at each stage. The authors construct mutually connected up and down sampling stages each of which represents different types of image degradation and high-resolution components.

2.2. Video Super-resolution

For video super-resolution, previous methods usually adopted two-stage approaches that are based on optical flow. In the first stage, motion estimation is conducted by computing optical flow. In the second stage, the estimated motion fields are used to perform image wrapping and motion compensation. For example, Liao et al. [12] used classical optical flow methods to generate high-resolution image drafts, and then predicted the final high-resolution frame by a deep draft-ensemble network. However, the classical optical flow algorithms are independent of the frame reconstruction CNN and much slower than the flow CNN during inference. To tackle this issue, Caballero et al. [2] design an end-to-end VSR network to jointly trains flow estimation and spatio-temporal networks. Tao et al. [21] compute low-resolution motion field based on optical flow network and develop a new layer to utilize sub-pixel information from motion and achieve sub-pixel motion compensation (SPMC) and resolution enhancement simultaneously. Xue et al. [24] exploit task-oriented flow to get better VSR results than fixed flow algorithms. Sajjadi et al. [18] propose a frame-recurrent VSR framework which uses the previously inferred high-resolution frame to super-resolve the subsequent frame. However, it is non-trivial to obtain high-quality motion estimation even with state-of-the-art optical flow estimation networks. Even with accurate motion fields, the image-wrapping based motion compensation will bring artifacts to super-resolved images, which could be propagated into the reconstructed high-resolution frames. Instead of explicitly computing and compensating motions between input frames, Jo et al. [8] proposed to utilize the motion information implicitly via generating dynamic upsampling filters. With the generated upsampling filters, the high-resolution frame is constructed by local filtering to the input center frame. However, they employ 3D convolution kernels to model the temporal information among successive frames, which are slow to compute comparing to using 2D convolution kernels.

3. The Proposed Framework

In this section, we describe our method for video super-resolution. The task is to map low resolution frames of a video into up-sampled high resolution ones. Different from image super-resolution, video frames not only demonstrate spatial correlation in each frame, but also contain temporal information among neighbouring frames. How to effectively and efficiently learn spatial-temporal feature to produce a super-resolved video frame is the key to VSR. Generally speaking, we propose to adapt multiple state-of-the-art image super-resolution methods for video super-resolution. Furthermore, for model ensemble, a CNN model is designed to adaptively ensemble results from multiple models.

3.1. Adapting Image Super-resolution State-of-the-arts for Video Super-resolution

We propose to learn deep spatial-temporal features for up-sampling video frames by adapting multiple state-of-the-art image super-resolution methods. Our insight is that, image super-resolution methods, which use 2D convolutional kernels to capture spatial features from images, can efficiently be adopted for spatial-temporal feature modeling by applying 2D convolutional filtering on an "super image" consists of several successive video frames. The image super-resolution methods adapted by us for video super-resolution including RCAN [28] and RDN [29]. To up-sample the t-th frame F_t, firstly, we early fuse successive frames of $[F_{t-T}, ..., F_{t-1}, F_t, F_{t+1}, ..., F_{t+T}]$ by concatenating them of along the channel dimension, such that the input "super-image" is a $3(2T+1) \times H_{lr} \times W_{lr}$ tensor, where H_{lr} and W_{lr} denote the height and width, respectively. Then the first convolution layer of RDN or RCAN is manipulated to accept "super-image" as input by changing its input channel number from 3 to $3(2T+1)$. The advantage of this adaptation schema is to largely reduce the computation cost compared with using 3D based solutions, such as DUF proposed in [8], and its effectiveness is also verified by our experimental results.

3.2. Learning Multi-model Ensemble

We design a learning based adaptive ensemble method to leverage the diversity among different models. We notice that even with the same model, in a super-resolved frame, the visual quality of different patches vary a lot. These priors motivate us to ensemble multiple models in an adaptive way, namely, the fusion weights of different models must be conditioned on the frames generated by these models in an image patch granularity.

The framework of our adaptive model ensemble solution is depicted in Figure 2. We use a 2D ConvNet to model the patch-level diversities among different models. The 2D ConvNet contains 3 Conv-BN-ReLU layers, the number of output channels of the 3 layers are 16, 32 and 64 respectively. The kernel size and the stride of all the 3 convolution layers are both set to 2, and no padding is used in this model. After the 3 Conv-BN-ReLU layers, a 1x1 convolution layer with output channel size of 1 is attached to generate a $1 \times \frac{H}{8} \times \frac{W}{8}$ score map for each input image. To ensemble images of N models needs to produce N score maps, we apply softmax to each spatial location of the score maps along the N model dimension and then $8\times$ up-sampling is used. Therefore, every non-overlapped 8×8 patch shares a same score and the scores of the N maps sum to 1 at every spatial location. These final score maps are treated as fusion weights for each model. We can see that the proposed ensemble method can be trained to adaptively ensemble the results of multiple models at 8×8 patch granularity yet at very limited cost.

4. Experiments

4.1. Dataset

We conduct experiments on a recently released novel dataset of low and high resolution blur-free videos (REDS [15]) obtained in diverse environments. It enables an accurate benchmarking of the performances achieved by the video super-resolution methods. In the REDS dataset, there are 300 pairs of low and corresponding high resolution blur-free video sequences and each video sequence is 100 frames long. For the NTIRE19 competition, REDS is divided into 240 video pairs for training and 30 for validation and 30 for testing.

The groundtruth of the 30 testing videos are not publicly available, therefore our experiments are performed on the training and validation videos. During the development phase of the NTIRE2019 competition, we split the 240 training video pairs into two parts, i.e., 216 for training and 24 (Val24) for model selection, and the 30 validation videos

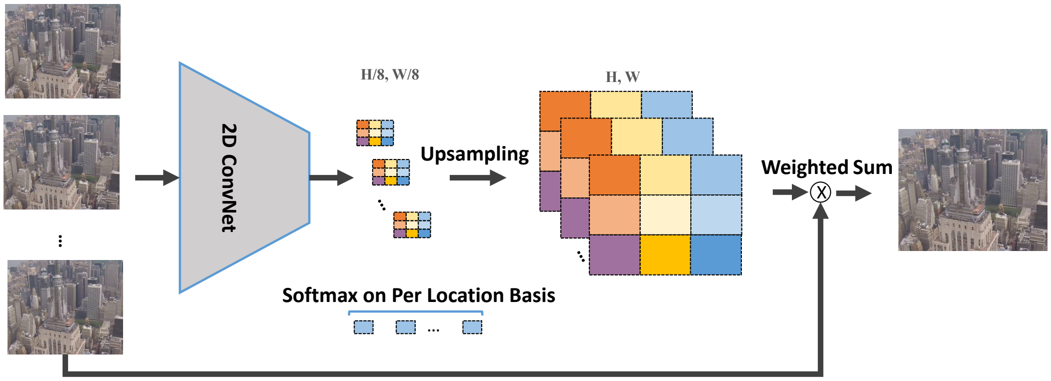

Figure 2. Illustration of the proposed adaptive model ensemble framework. The inputs are the frames generated by N models. A 2D ConvNet with stride of 8 is employed to extract a $1 \times H/8 \times W/8$ feature map for each image. Then, a softmax activation is applied to each spatial location of these feature maps, and afterwards the activation score map is up-sampled by a factor of 8. In this way, the fusion weights of all models sum to 1 at every pixel.

(Val30) are used for testing the performance of our model. During the final phase of the NTIRE2019 competition, besides the 216 training videos and 30 validation videos, we further split 12 video from the Val24 set for training, finally we got 258 videos for model training and 12 videos (Val12) for model selection.

4.2. Implementation Details

We investigate and adapt three ISR state-of-the-arts, i.e., EDSR [13], RDN [29], RCAN [28] for VSR. For each method, we adopt the configuration corresponding to the best performance reported in the orginal papers. L1-Loss is used to train the EDSR, RDN and RCAN models, and For DBPN, MSE loss is used. We employ ADAM optimizer and set initial learning rate to $1e-4$. We decrease the learning rate to $5e-5$, $3e-5$ and $1e-5$ sequentially when it meets a converge point. To validate the effectiveness of our ensemble method. We also trained a DUF model with L1-Loss and SGD, its learning rate is initially set to 0.1 and is 10× decreased every 50 passes. The adaptive ensemble CNN model are trained on the Val12 set which is reserved by us for model selection. Its learning strategy is the same as DUF. All the models are trained from scratch. Four kinds of augmentation are employed, i.e. 90 degree rotation, vertical flip, horizontal flip and temporal flip. For a training sample, each augmentation is adopted with probability of 0.5. In the training phase, 96×96 patches are random cropped.

For testing, following [13], we use self-ensemble to test each single model. In our setting, there are 16 augmentations in total by combining temporal flipping, vertical flip, horizontal flip and spatial rotation. In testing, the whole low-resolution "super image" is fed into each single model to produce up-sampling result with 16× self-ensemble. Temporal reflection padding is utilized for the t-th input "super-image" if any element in [$t-T$,...,$t-1$, t, $t+1$,..., $t+T$] is out of range, $t = 0, 1, ..., 99$ The proposed adaptive ensemble model is employed to ensemble the results from each single model. The PSNR metric is used for performance evaluation.

Method	Image SR	Adapted Video SR
EDSR	28.52	29.55
RCAN	28.63	29.93
RDN	28.52	29.71

Table 1. Comparison of original Image SR methods and the corresponding adapted methods for Video SR on Val24. All these models are trained on 216 videos and tested on Val24. No self-ensemble is used.

4.3. Evaluation of Adapted Methods

To evaluated the proposed adapting method, we train the original ISR methods, i.e., EDSR [13], RDN [29], RCAN [28], and their corresponding adapted variants, on the 216 training videos. The evaluation results on Val24 set is reported in Table 1, and no self-ensemble is used for each method in testing. The length of the input "super image" for each model is set to 5, i.e., the T as mentioned in Section 4.2 is set to 2. As shown in the table, All the three adapted VSR methods outperform their original ISR version significantly. It proves the effectiveness of the proposed adapting method for VSR. We also provide qualitative results in Figure 3. The adapted models successfully reconstruct the detailed textures and edges in the high-resolution images and present super-resolved outputs with better visual quality than the original ISR methods.

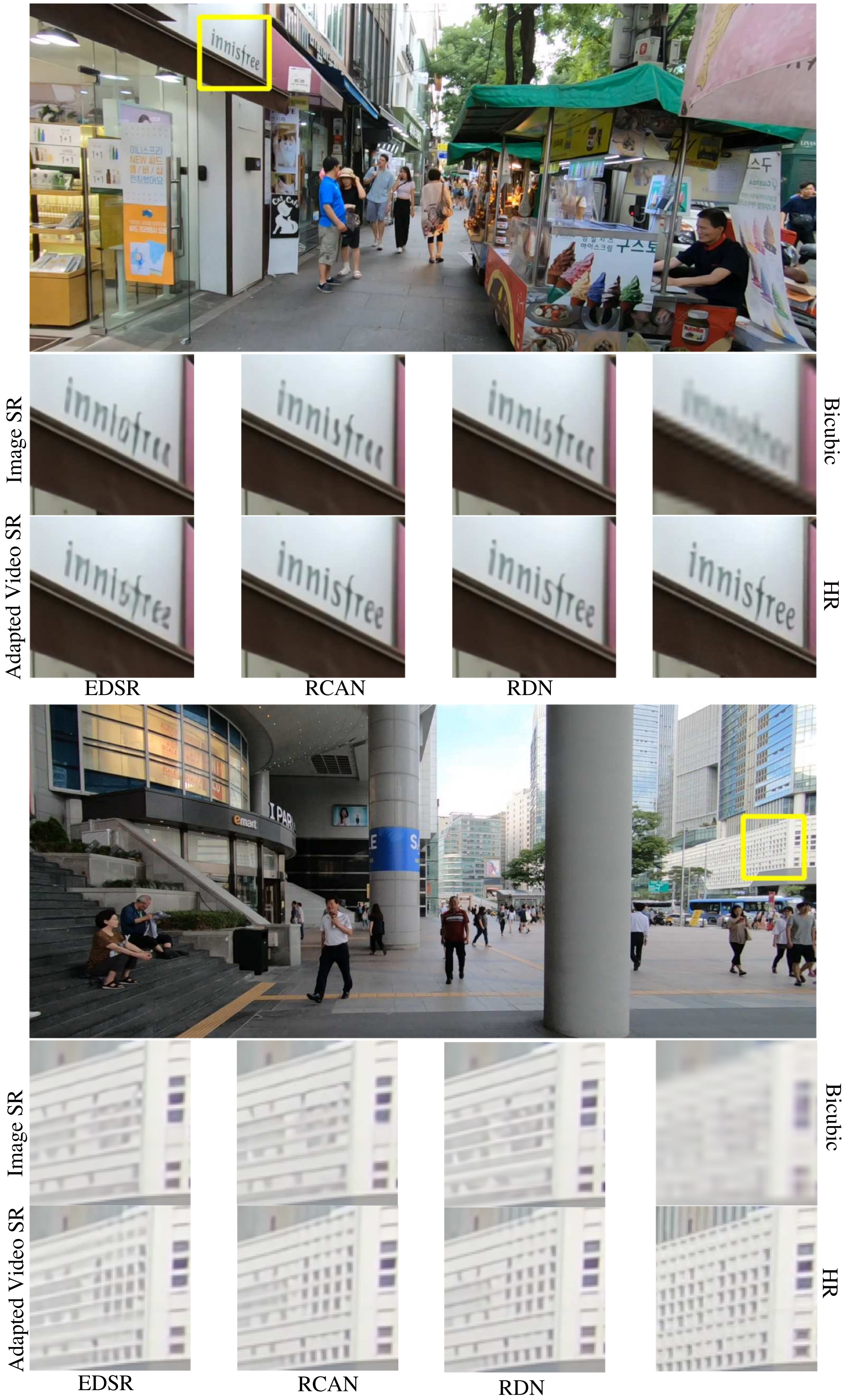

Figure 3. Qualitative comparison of our adapted Video SR models with the corresponding original Image SR models.

Method	Val30	Val12
RCAN	30.13	29.97
RDN	30.14	30.01
RDN-MSE	30.11	29.94
RDN-Deeper	×	30.03
RDN-Extradata	×	30.03
RDN-Bicubic	×	29.91
DUF	29.78	29.58
Avg Ensemble	30.20	30.05
Adaptive Ensemble	30.26	30.07

Table 2. Quantitative Results of different methods in terms of PSNR in dB. Val30 is the official validation set, on which the results in the developing phase are reported. Val12 is our own division on which the results trained on 258 videos in the final phase are reported. × means not used.

4.4. Learning Multi-model Ensemble

In the final phase of the NTIRE2019 VSR competition, five adatped RDN variants are trained, i.e., the adatped RDN network with L1 Loss (RDN), the adatped RDN network with MSE loss (RDN-MSE), the adapted RDN network with more dense blocks with L1 Loss (RDN-Deeper), the adapted RDN network augmented by adding a bicubic $4\times$ up-sampled residual connection (RDN-Bicubic) and the adapted RDN trained with extra data (RDN-Extradata). All the above RDN model are configured with the best setting in the original paper [29], except that we increase the channel number from 64 to 128. For the RCAN [28] model, we find that enlarge the channel numbers gains no performance improvement. We analyze the reason is that, RCAN utilizes channel attention mechanism which effectively exploits the information among each channel and the original channel number is already enough for fitting the REDS dataset. However, we still improve the performance by adding more blocks to RCAN. To validate the effectiveness of our ensemble method, besides the above adapted VSR methods, we also trained a DUF model as described in Section 4.2. To increase the model diversity, we use different T settings for the above models. The T settings of RCAN, RDN, RDN-MSE, RDN-Deeper, RDN-Extradat, RDN-Bicubic and DUF are 2,3,1,3,3,2 and 2 respectively.

For training the RDN-Extradata model, we manually select 9 videos from Youtube website and 30 video clips, each of which with length of 4 seconds, are randomly trimmed from them. Then we downscale the video clips with a factor of 4 and decode each of them into 100 frames. The video ids of the 9 videos are *44ZIHUIybM8, 9AX1_eEmRMU, AK80adiQVPw, dpgHqCrCmwY, ITpBDhc6p-0, L5FKHsFQ0IA, mjbUEuZtZ08, OU5ipJnApOY and zW_UqsYuhbY*. These extra data are used for finetuning our models in the test phase.

We report the experimental results in Table 2. The results on Val30 are from models trained on the 216 training videos and the results on Val12 are produced by models trained on the 258 traing videos, as mentioned in Section 4.1 As shown in the table, the learned adaptive ensemble outperforms the average ensemble by a remarkable margin, which evidences the effectiveness of the proposed ensemble method. The proposed adaptive ensemble method is data driven and can be trained to adaptively assign fusion weight to each single model. Compared to single method, such fusion method can better leverage model diversity to boost the performance.

5. Conclusion

In this paper we propose to adapt image super-resolution state-of-the-arts for video super-resolution. The proposed adapting method is straightforward to implement and the extra adapting overhead is negligible. Different from existing video super-resolution methods, our framework does not rely on flow estimation nor motion compensation, which are slow to compute and may propagate errors to high-resolution image reconstruction. Besides, our method only involves 2D convolution kernels rather than 3D convolution kernels as in DUF [8]. Compared to 3D Convolution, our method is also an efficient way to learn deep spatial-temporal features from videos. Therefore, our adapting method can handle video super-resolution task while the computation cost is the same as image super-resolution. The experimental results on the recent benchmark dataset for video super resolution, i.e., REDS, have verified the effectiveness of the proposed adapting method.

Furthermore, we propose an adaptive model ensemble framework, which can effectively exploit the diversity of the results from every single model. It is lightweight and can boost the performance remarkably. The experimental results on REDS also evidences its effectivenee.

References

[1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, pages 1–10, 2012.

[2] Jose Caballero, Christian Ledig, Andrew P. Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2848–2857, 2017.

[3] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), with CD-ROM, 27 June - 2 July 2004, Washington, DC, USA*, pages 275–282, 2004.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, pages 184–199, 2014.

[5] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, pages 391–407, 2016.

[6] Xinbo Gao, Kaibing Zhang, Dacheng Tao, and Xuelong Li. Image super-resolution with sparse neighbor embedding. *IEEE Trans. Image Processing*, 21(7):3194–3205, 2012.

[7] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1664–1673, 2018.

[8] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3224–3232, 2018.

[9] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1646–1654, 2016.

[10] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1637–1645, 2016.

[11] Xin Li and Michael T. Orchard. New edge-directed interpolation. *IEEE Trans. Image Processing*, 10(10):1521–1527, 2001.

[12] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 531–539, 2015.

[13] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1132–1140, 2017.

[14] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2802–2810, 2016.

[15] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenges on video deblurring and super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[16] Seungjun Nah, Radu Timofte, Shuhang Gu, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Kyoung Mu Lee, Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, Chen Change Loy, Yuchen Fan, Jiahui Yu, Ding Liu, Thomas S. Huang, Xiao Liu, Chao Li, Dongliang He, Yukang Ding, Shilei Wen, Fatih Porikli, Ratheesh Kalarot, Muhammad Haris, Greg Shakhnarovich, Norimichi Ukita, Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Jiayi Ma, Hang Dong, Xinyi Zhang, Zhe Hu, Kwanyoung Kim, Dong Un Kang, Se Young Chun, Kuldeep Purohit, A. N. Rajagopalan, Yapeng Tian, Yulun Zhang, Yun Fu, Chenliang Xu, A. Murat Tekalp, M. Akin Yilmaz, Cansu Korkmaz, Manoj Sharma, Megh Makwana, Anuj Badhwar, Ajay Pratap Singh, Avinash Upadhyay, Rudrabha Mukhopadhyay, Ankit Shukla, Dheeraj Khanna, A.S. Mandal, Santanu Chaudhury, Si Miao, Yongxin Zhu, and Xiao Huo. Ntire 2019 challenge on video super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, pages 234–241, 2015.

[18] Mehdi S. M. Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6626–6634, 2018.

[19] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*, 2008.

[20] Yu-Wing Tai, Shuaicheng Liu, Michael S. Brown, and Stephen Lin. Super resolution using edge prior and single image detail synthesis. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 2400–2407, 2010.

[21] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4482–4490, 2017.

[22] Radu Timofte, Vincent De Smet, and Luc J. Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1920–1927, 2013.

[23] Radu Timofte, Vincent De Smet, and Luc J. Van Gool. A+: adjusted anchored neighborhood regression for fast super-resolution. In *Computer Vision - ACCV 2014 - 12th*

Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV, pages 111–126, 2014.

[24] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T. Freeman. Video enhancement with task-oriented flow. *CoRR*, abs/1711.09078, 2017.

[25] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Trans. Image Processing*, 19(11):2861–2873, 2010.

[26] Kaibing Zhang, Xinbo Gao, Dacheng Tao, and Xuelong Li. Single image super-resolution with non-local means and steering kernel regression. *IEEE Trans. Image Processing*, 21(11):4544–4556, 2012.

[27] Lei Zhang and Xiaolin Wu. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE Trans. Image Processing*, 15(8):2226–2238, 2006.

[28] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, pages 294–310, 2018.

[29] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2472–2481, 2018.