AIM 2019 Challenge on Video Extreme Super-Resolution: Methods and Results

Dario Fuoli Shuhang Gu Radu Timofte Xin Tao Wenbo Li Taian Guo
Zijun Deng Liying Lu Tao Dai Xiaoyong Shen Shutao Xia Yurong Dai
Jiaya Jia Peng Yi Zhongyuan Wang Kui Jiang Junjun Jiang Jiayi Ma
Zhiwei Zhong Chenyang Wang JunJun Jiang Xianming Liu

Abstract

This paper reviews the video extreme super-resolution challenge associated with the AIM 2019 workshop, with emphasis on submitted solutions and results. Video extreme super-resolution ($\times$16) is a highly challenging problem, because 256 pixels need to be estimated for each single pixel in the low-resolution (LR) input. Contrary to single image super-resolution (SISR), video provides temporal information, which can be additionally leveraged to restore the heavily downscaled videos and is imperative for any video super-resolution (VSR) method. The challenge is composed of two tracks, to find the best performing method for fully supervised VSR (track 1) and to find the solution which generates the perceptually best looking outputs (track 2). A new video dataset, called Vid3oC, is introduced together with the challenge.

1. Introduction

VSR describes the task of reconstructing the high-resolution (HR) video from its LR representation. Super-resolution is an ill-posed problem, as high-frequency information is inherently lost when downscaling an image or video, because of the lower Nyquist frequency in LR space. SISR methods usually restore this information by learning image priors through paired examples. For VSR, additional information is present in the temporal domain, which can help significantly improving restoration quality over SISR methods. SISR has been an active research for a long time [6, 18, 23, 25, 11, 34, 37, 32], while VSR has gained traction in recent years [30, 15, 33, 16, 7, 2, 40, 36, 31, 17], also due to the availability of more and faster computing resources. While there exists a lot of prior work on super-resolution factors $\times 2$, $\times 3$ and $\times 4$ with impressive results, attempts at higher factors are less common in the field [22]. Restoring such a large amount of pixels from severly limited information is a very challenging task. The aim of this challenge is therefore to find out, if super-resolution with such high downscaling ratios is still possible with acceptable performance. Two tracks are provided in this challenge. Track 1 is set up for fully supervised example-based VSR. The restoration quality is evaluated with the most prominent metrics in the field, Peak Signal-to-Noise Ratio (PSNR) and structural similarity index (SSIM). Because PSNR and SSIM are not always well correlated with human perception of quality, track 2 is aimed at judging the outputs according to how humans perceive quality. Track 2 is also example-based, however, the final scores are determined by a mean opinion score (MOS).

2. Related Work

With the recent success of deep learning and the introduction of convolutional neural networks (CNN) [21], learning based SISR super-resolution models have shown to be superior compared to classical methods. SRCNN, a lightweight network with only 3 layers proposed by [6], is one of the first convolutional network for super-resolution. VDSR [18] shows substantial improvements over SRCNN by designing a much deeper network and introducing residual learning. Photorealistic and more natural looking images can be obtained by methods like SRGAN [23], EnhanceNet [32], and [31, 4, 28] that introduce alternative loss functions [10] to super-resolution, which improve perceptual quality. These methods however, deviate from the accuracy to the ground truth to achieve perceptually more pleasing images. A general overview of SISR methods can be found in [37].

Traditionally, VSR problems are solved by formulating demanding optimization problems [1, 8, 26], which are very slow compared to recent learning based methods. Deep learning based methods often leverage temporal information by concatenating multiple low-resolution frames to produce a single HR estimate [24, 29, 5]. This strategy is followed by all competitors in the challenge. Ca-

D. Fuoli (dario.fuoli@vision.ee.ethz.ch, ETH Zurich), S. Gu, and R. Timofte are the AIM 2019 challenge organizers, while the other authors participated in the challenge.
Appendix A contains the authors teams and affiliations.
AIM webpage: `http://www.vision.ee.ethz.ch/aim19/`

2473-9944/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCVW.2019.00430

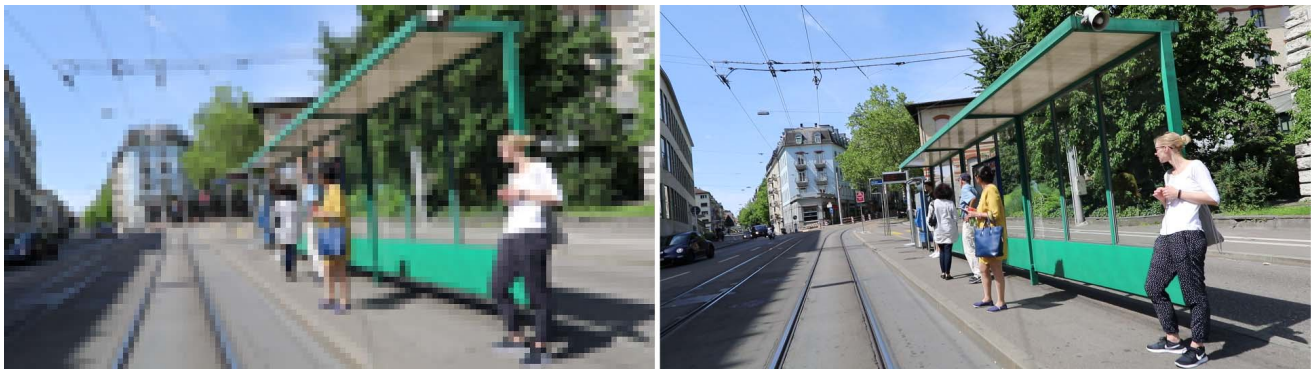

Figure 1. Downscaled frame with factor $\times 16$ (left) and corresponding HR frame (right), taken from the test set.

ballero *et al*. [3] warps adjacent frames towards the center frame. A spatio-temporal network is used to fuse the frames. The method also adopts 3D convolutions [38]. Liu *et al*. [27] calculate multiple high-resolution estimates in parallel branches. An additional temporal modulation branch computes weights, to balance the respective high-resolution estimates for final aggregation. Kappeler *et al*. [17] combine motion compensated adjacent frames, by calculating optical flow and warping. The combined frames are processed with 3 convolution layers. Jo *et al*. [16] does not rely on explicit motion estimation. Dynamic upsampling filters and residuals are calculated from adjacent LR frames with a single network. In a final step, the dynamic upsampling filters are used to process the center frame, which is then combined with the residuals.

Recurrent neural networks (RNN) have also been proposed for super-resolution. As a fixed number of input frames inherently limits the available information, they provide a potentially more powerful alternative to deal with super-resolution. Tao *et al*. [36] gathers 7 input frames. After computing the flow maps, a preliminary high-resolution estimate is prepared through a subpixel motion compensation layer. An encoder-decoder style network, with an intermediate convolutional LSTM [13] layer, processes the preliminary estimate to get the final HR estimate. Huang *et al*. [14] propose a bidirectional recurrent network. The network adopts 2D and 3D convolutions with recurrent connections and combines a forward and a backward pass to generate the HR frames. To leverage temporal information, Sajjadi *et al*. [33] propose a network which warps the previous HR output towards the current time step, according to the optical flow in LR space. The warped output is concatenated with the current low-resolution input frame and a super-resolution network produces the HR estimate. The network is trained end-to-end.

3. AIM 2019 Challenge

3.1. Dataset

In contrast to SISR, there are no standardized benchmarks that are widely accepted in the field for higher resolutions, longer duration and more realistic video material. The most frequently used benchmark for VSR is vid4 [26]. Vid4 is a dataset composed of 4 short sequences with a resolution between 480×704 and 576×704. In order to promote such a standard, a novel dataset, called Vid3oC [19] has been collected. It provides videos in high resolution taken by a cell phone camera (Huawei P20), a ZED stereo camera, which also records depth information, and a high quality DSLR camera (Canon 5D Mark IV). The cameras are aligned as close as possible on a rig and all videos are taken at the same time to ensure proper alignment. With 3 different cameras, stereo frames and depth information this dataset is extremely versatile and can be used for many video related tasks. For the video extreme super-resolution challenge only the high-quality DSLR videos were used, because they are ideal to serve as high quality ground truth.

3.2. Dataset preparation

The whole training set (encoded HR videos) was provided to the participants, so they could generate their training data according to their needs and to reduce data traffic. For the validation and test phase, two separate sets are provided of 16 sequences each, all composed of 120 LR frames in PNG format, to be processed and evaluated on the CodaLab servers. To generate the LR data from the FullHD (1920×1080) HR videos, the HR sequences are first cropped to 1920×1072, to be dividable by 16. The sequences were then downscaled by Matlab's *imresize* method with factor $\times16$ using the standard settings, resulting in sequences of 120×67. A single frame and a selection of crops, taken from the test set, is shown in Fig. 1 and Fig. 2 respectively. There is a severe loss in information for such a high downscaling, which results in heavily blurred out

Figure 2. Downscaled crops with factor $\times 16$ (top) and corresponding 100×100 HR crops (bottom), taken from the test set.

frames and clearly shows the difficulty of the task. Scripts were provided to the participants such that they could follow identical processing steps to generate the training data. Due to filesize limits on CodaLab, the participants were asked to only submit every 20th frame to the server.

3.3. Track 1: Supervised VSR

This track aims at restoring the HR videos as close to the ground truth as possible in terms of PSNR and SSIM. The winner can be objectively chosen according to these values.

3.4. Track 2: Perceptual VSR

PSNR/SSIM values do not always correspond to how humans perceive visual quality. Therefore, relying on direct losses to the ground truth, typically L1 or L2, alone does not guarantee visually pleasing results. To increase the perceived quality, other losses are applied. Common are losses evaluated in feature space from networks like VGG [35] or AlexNet [21] and GAN [10] losses, to find a good trade-off between accuracy with respect to the ground truth and perceived quality. In this challenge, the HR videos should be restored with the highest perceptual quality, decided by a MOS.

3.5. Evaluation protocol

Validation phase: During the validation phase, the LR inputs for the validation set were provided on CodaLab. The participants had no direct access to the validation ground truth, but could get feedback through the server on CodaLab. Due to the storage limits on the servers, only a subset of frames could be submitted and evaluated. We reported PSNR and SSIM for both tracks, even though track 2 is ultimately evaluated by a MOS. 10 submission were allowed per day, 20 submission in total for the whole validation phase.

Test phase: In the test phase, participants submitted their final results to the CodaLab test server. In contrast to the validation phase, no feedback was given in terms of PSNR/SSIM to prevent comparisons with other teams and overfitting to the test data. After the deadline, the participants were asked to provide the full set of frames, from which the final results were calculated.

4. Challenge Results

From initially 39 and 30 registered participants in track 1 and 2 respectively, three teams (fenglinglwb, NERCMS, and HIT-XLab) entered the final ranking and submitted their results, codes, executables and factsheets.

Team fenglinglwb and HIT-XLab provided the same solutions to both tracks, while NERCMS submitted a solution to the first track. According to PSNR/SSIM and MOS, the winner of both challenge tracks is fenglinglwb with $+0.17$ dB over team NERCMS and $+1.84$ dB over the Bicubic baseline. The final ranking, PSNR, SSIM, MOS, runtimes, platform and type of hardware is shown in Tab. 1

4.1. Runtime/efficiency

For real-time video processing, efficient algorithms are crucial. The limited time between frames, imposes hard limits on the runtime. It is therefore desired to design video processing algorithms with an emphasis on efficiency. Thus, the participants were also asked to report the runtime of their algorithms, even though the values did not affect the ranking. Interestingly, the runtimes of all methods are quite far apart. The fastest method, apart from the bicubic baseline, is the RLSP baseline which can process a single frame in only 95ms. However, the PSNR/SSIM value is below the two top teams. Fenglinglwb, the winning team in both tracks, achieves a good trade-off between performance and speed, aswell as team NERCMS, with 350ms and 510ms per frame respectively. Team HIT-XLab achieves the low-

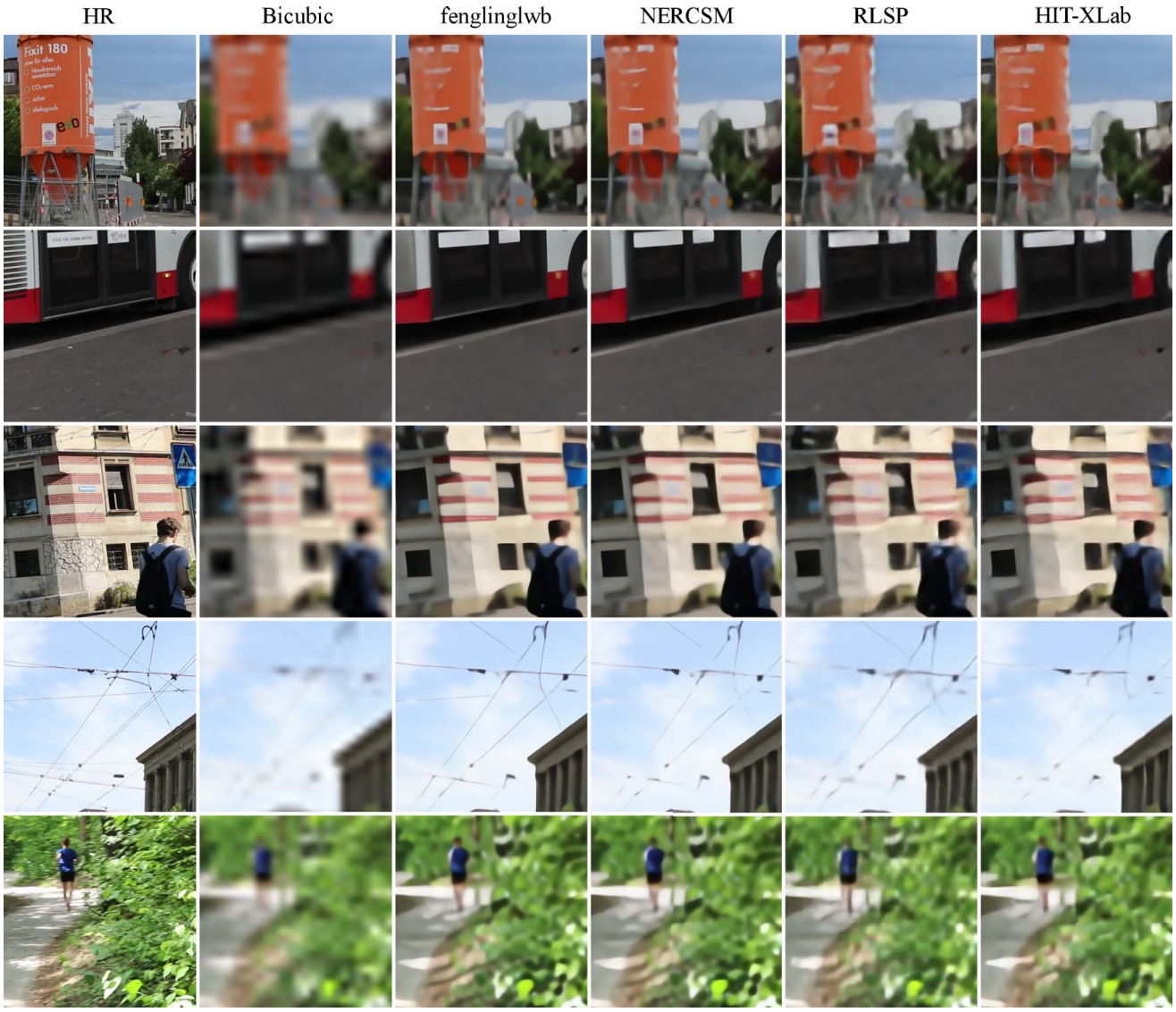

Figure 3. Results for all methods on the test set (400x400 crops).

est performance with the highest runtime (60s).

4.2. Use of temporal information and consistency

In SISR, the available information is limited to the single LR image. In video however, the temporal axis provides additional information for a single frame. It is therefore important to leverage this information, when trying to improve performance against a SISR method. All teams make use of temporal information by feeding a batch of adjacent frames to the network and try to leverage the correlations between the frames. The baseline method RLSP additionally introduces recurrent connections between time steps, which allows accumulation of long-term information.

In contrast to single independent images, frames in a video are ordered and correlated along the time axis. Temporal inconsistencies in videos can result in perceptually low quality, as humans can easily detect these artefacts. Typical losses used for VSR, like L1 or L2, are calculated per frame and therefore do not directly take into account the temporal evolution of content. Temporal profiles can help to asses the quality of temporal consistency visually. To produce such a profile, a single line of pixels is recorded for a sequence and stacked together. A sharp profile indicates good temporal consistency. Fig. 4 shows profiles for all methods. While all method's profiles are clearly sharper than the Bicubic baseline, no method is able to come close to the HR ground truth, which is expected for such an extreme scaling factor.

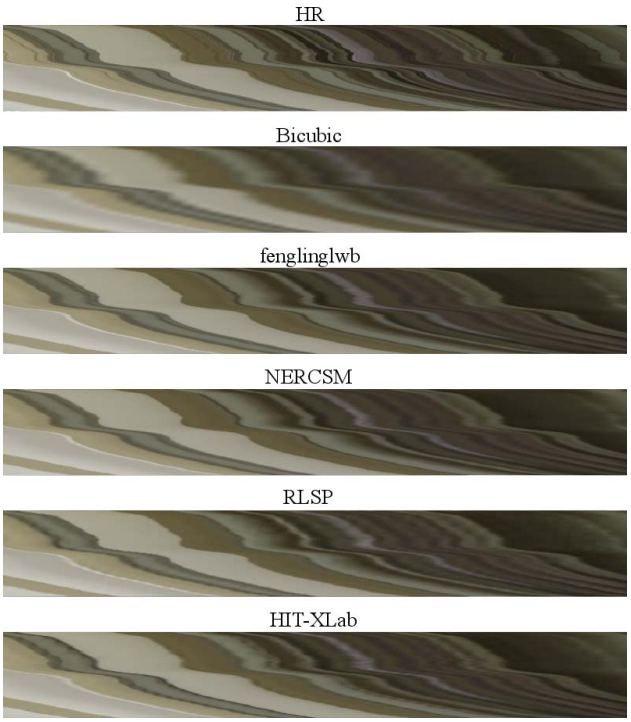

Figure 4. Temporal profiles of sequence 80 for all methods.

4.3. Ensembles

It is very common for challenge participants to use ensembling in order to boost the performance. Both team fenglinglwb and team NERCMS make use of this strategy. Team NERCMS achieves an improvement of 0.16dB for their ensembling strategy on the validation set. Team fenglinglwb reports a boost in performance of 0.05dB on the validation set.

4.4. Conclusions

With such a high downscaling factor, it is very hard to restore accurate details in the HR frames. Nevertheless, the top team still achieves impressive performance in such a challenging setting, as can be seen in Fig. 3.
Interestingly, the difference in PSNR/SSIM between the subsampled frames and the full set for team fenglinglwb does not change, for team NERCMS and both baselines the difference in PSNR is only 0.01. The results for team HIT-XLab are therefore still meaningful and using subsampling for evaluation during the validation phase provided accurate feedback to the participants.

5. Challenge Methods and Teams

5.1. fenglinglwb team

The team's solution is based on the network EDVR [39]. Additionally, an edge mask guided model is applied during learning. The ground truth frames are convolved with a Laplacian filter to extract edge information. The extracted edge areas are then weighted more when calculating the loss, which encourages the network to learn more refined edges. The method also incorporates a non-local module to make use of global information. Since the setting $\times 16$ aims to expand each pixel to 256 pixels, it is essential to take the context information into full play. Especially, regions with similar patterns could benefit each other. Therefore, the non-local module is designed to take advantage of global information based on similarities of regions.

To boost the final performance for the challenge, the team makes use of ensembling. The final ensemble model (see Fig. 5) is composed of two edge mask models of different sizes and a non-local model. The LR input frames are fed to all three models and are averaged in a final step to produce the HR estimate. Training is conducted on 4 Nvidia GeForce GTX TITAN X GPU's with a mini-batch size of 44 per GPU. The network is trained for 110k iterations with cosine annealing for the learning rate. On top of the global ensemble strategy, self-ensembling is also used for each individual module. The method can produce a single output frame in 0.35s. The framework can easily be adapted to other video super-resolution factors.

The team submitted identical results for both tracks 1 & 2.

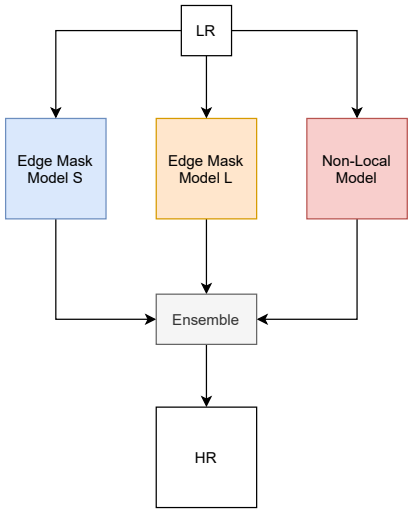

Figure 5. Architecture of team fenglinglwb.

5.2. NERCMS team

The NERCMS team follows a progressive fusion network by Yi *et al.* [41], which is composed of a series of progressive fusion residual blocks (PFRBs). These blocks are designed to leverage both spatial (intra-frame) and temporal (inter-frame) correlations between frames. Figure 6

Team	user	PSNR every 20th frame	SSIM every 20th frame	PSNR all frames	SSIM all frames	MOS Track 2	Runtime [s]	Platform	GPU/CPU
fenglinglwb	fenglinglwb	**22.53**	**0.64**	**22.53**	**0.64**	$\mathbf{1}^{st}$	0.35	PyTorch	4× Titan X
NERCMS	Mrobot0	22.35	0.63	22.36	0.63	-	0.51	PyTorch	2× 1080 Ti
RLSP	*baseline*	21.75	0.60	21.74	0.60	-	**0.09**	TensorFlow	Titan Xp
HIT-XLab	zhwzhong	21.45	0.60	-	-	2^{nd}	60.00	PyTorch	V100
Bicubic	*baseline*	20.68	0.58	20.69	0.58	-	0.01	Matlab	i7

Table 1. Results for the participating teams. We evaluated PSNR and SSIM for track 1 and conducted a MOS for track 2.

illustrates the case of adopting 5 frames as input, still, the PFRB can be easily extended to alternative frame numbers, e.g. 3, 7. In a PFRB, input frames are first convoluted separately to obtain rather self-independent features. Then, feature maps are fused into one part, which contains a large deal of temporal-correlated information. This aggregated feature map is distilled for more temporal-related and concise information. The distilled information is further concatenated to each feature map, which are convoluted to generate feature maps containing both self-independent spatial information and fully mixed temporal information. A residual learning scheme is adopted to ease the training difficulty.

The method gathers a batch of 7 consecutive frames and processes them with a non-local residual block. This module computes correlations between a single pixel and all other pixels in the batch to enable global feature extraction. Because this module costs huge GPU memory, it may not be able to process frames of large input size. The resulting feature maps are further convolved with a 5×5 kernel, which is followed by a series of PFRBs. At the end of the processing chain, the residuals are added to the input center frame, which is upscaled with bicubic interpolation.

The network consists of 20 PFRBs and the parameter number is about 3.4 M. During training, 16 LR patches of size $7 \times 48 \times 48 \times 3$ are randomly extracted from training data. The network is trained with L1 loss and Adam optimizer [20]. The learning rate is set to 5×10^{-4} in the beginning and is divided by 2 after every 5×10^4 iterations approximately, until to 10^{-5}. The model is trained with PyTorch on 2 GTX 1080 Ti GPUs. It takes about 0.07s to generate one frame without ensemble strategy. The team submitted results for track 1.

5.3. HIT-XLab team

The team's method uses EDSR [25] as the backbone and replaces all 2D convolutions with 3D convolutions [38]. To reduce memory usage, the number of residual blocks [12] is reduced to 8. During training, 64 LR patches of size 5×32×32×3 are randomly extracted from the training data. The network is trained with L1 loss and gradient loss with Adam optimizer [20]. The learning rate is set to 10^{-4} in the beginning and is divided by 2 after every 10 epochs. Because the method relies on 3D convolutions, it has to deal

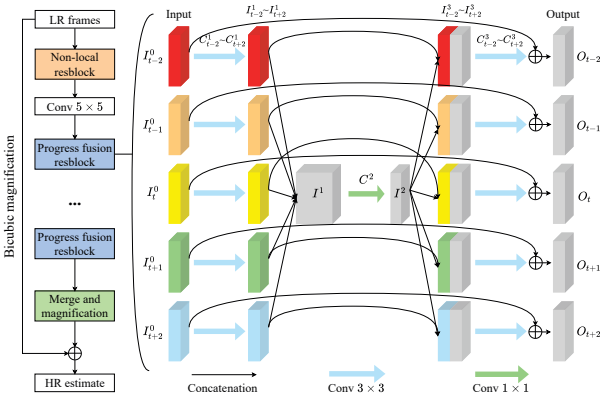

Figure 6. Architecture of team NERCMS.

with some form of temporal padding. The team chose to use no padding before processing the sequences, instead the missing frames at the temporal borders are upscaled using Bicubic interpolation. The model is trained with PyTorch on a V100 GPU. The training takes about 48 hours, evaluation speed is around 60s per frame. The team submitted identical results for both tracks 1 & 2.

5.4. RLSP (baseline)

Additionally to Bicubic interpolation, we provide another method as a baseline called RLSP [9]. The method proposes a fully convolutional RNN for VSR with upscaling factor 4. In addition to feeding back the previous output, which has been adopted by other network architectures [33], RLSP introduces high dimensional hidden states to enable implicit information propagation along time. In contrast to previous super-resolution networks, no optical flow and/or warping is used. Instead, accumulation and processing of temporal information is handled implicitly by the hidden states. Due to the recurrent nature, the method is extremely efficient at runtime. For this challenge, the RNN cell is adapted to enable upscaling factor ×16, see Fig. 7. Differently from the original implementation, the upscaling is divided into 4 stages of factor ×2, to handle the extreme super-resolution factor. The number of layers is set to 14, each convolution layer has 128 channels.

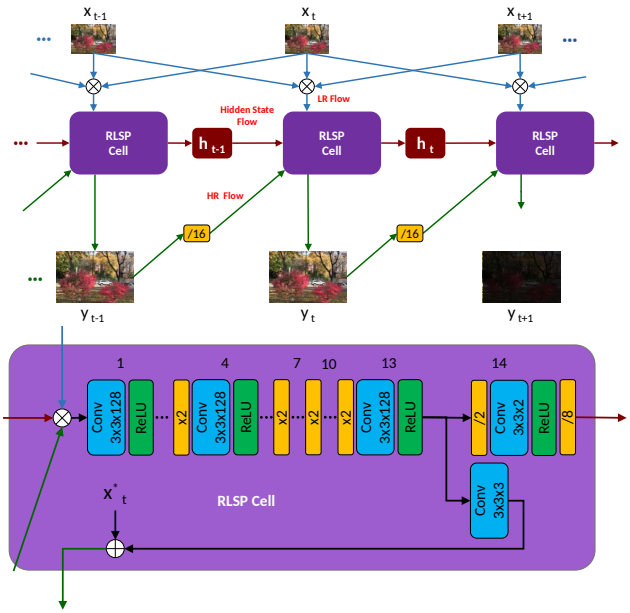

Figure 7. Architecture of RLSP.

Acknowledgments. We thank the AIM 2019 sponsors.

A. Teams and affiliations

AIM 2019 team

Title: AIM 2019 Challenge on Video Extreme Super-Resolution
Members: Dario Fuoli, Shuhang Gu, Radu Timofte
Affiliations:
Computer Vision Lab, ETH Zurich, Switzerland

fenglinglwb

Title: Generate High-Resolution Results with High Fidelity and Perceptual Quality
Members: Xin Tao, Wenbo Li, Taian Guo, Zijun Deng, Liying Lu, Tao Dai, Xiaoyong Shen, Shutao Xia, Yurong Dai, Jiaya Jia
Affiliations:
Tecent X-Lab, Shenzhen, Guangdong, China

NERCMS

Title: Progressive Fusion Video Super-Resolution Network via Exploiting Non-Local Spatio-Temporal Correlations
Members: Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Jiayi Ma
Affiliations:
National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, China

HIT-XLab

Title: Video Super-Resolution using 3D-ResNet
Members: Zhiwei Zhong, Chenyang Wang, JunJun Jiang, Xianming Liu
Affiliations:
Harbin Institute of Technology, Harbin 150001, China

References

[1] S. P. Belekos, N. P. Galatsanos, and A. K. Katsaggelos. Maximum a posteriori video super-resolution using a new multichannel image prior. *IEEE Transactions on Image Processing*, 19(6):1451–1464, June 2010.

[2] R. A. Borsoi, G. H. Costa, and J. C. M. Bermudez. A new adaptive video super-resolution algorithm with improved robustness to innovations. *IEEE Transactions on Image Processing*, 28(2):673–686, Feb 2019.

[3] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. Pixel recursive super resolution. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[5] Q. Dai, S. Yoo, A. Kappeler, and A. K. Katsaggelos. Sparse representation-based multiple frame video super-resolution. *IEEE Transactions on Image Processing*, 26(2):765–781, Feb 2017.

[6] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, Feb 2016.

[7] Salehe Erfanian Ebadi, Valia Guerra Ones, and Ebroul Izquierdo. Uhd video super-resolution using low-rank and sparse decomposition. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[8] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and Robust Multiframe Super Resolution. *IEEE Transactions on Image Processing*, 13:1327–1344, Oct. 2004.

[9] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *ICCV Workshops*, 2019.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. 2014.

[11] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 235–243, Cambridge, MA, USA, 2015. MIT Press.

[15] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 106–122, 2018.

[16] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. In *IEEE Transactions on Computational Imaging*, 2016.

[18] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] Sohyeong Kim, Guanju Li, Dario Fuoli, Martin Danelljan, Zhiwu Huang, Shuhang Gu, and Radu Timofte. The vid3oc and intvid datasets for video super resolution and quality mapping. In *ICCV Workshops*, 2019.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012.

[22] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[23] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[24] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017.

[26] Ce Liu and Deqing Sun. A bayesian approach to adaptive video super resolution. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 209–216, Washington, DC, USA, 2011. IEEE Computer Society.

[27] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[28] A. Lucas, S. Lopez Tapia, R. Molina, and A. K. Katsaggelos. Generative Adversarial Networks and Perceptual Losses for Video Super-Resolution. *arXiv e-prints*, June 2018.

[29] O. Makansi, E. Ilg, and T. Brox. End-to-End Learning of Video Super-Resolution with Motion Compensation. *arXiv e-prints*, July 2017.

[30] Seungjun Nah, Radu Timofte, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[31] E. Pérez-Pellitero, M. S. M. Sajjadi, M. Hirsch, and B. Schölkopf. Photorealistic Video Super Resolution. *arXiv e-prints*, July 2018.

[32] Mehdi S. M. Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[33] Mehdi S. M. Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-Recurrent Video Super-Resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[34] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[36] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[37] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[38] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[39] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[40] Z. Wang, P. Yi, K. Jiang, J. Jiang, Z. Han, T. Lu, and J. Ma. Multi-memory convolutional neural network for video super-resolution. *IEEE Transactions on Image Processing*, 28(5):2530–2544, May 2019.

[41] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.