Two-Stream Action Recognition-Oriented Video Super-Resolution

Haochen Zhang, Dong Liu, Zhiwei Xiong
University of Science and Technology of China, Hefei, Anhui 230027, China
Email: zhc12345@ustc.edu.cn
National Engineering Laboratory for Brain-inspired Intelligence Technology and Application

Introduction

➢ **Background**

CNNs have been applied to action recognition task and obtained state-of-the-art performance. However, well-trained CNNs cannot be directly applied on LR video because of the existence of FC layers.

Solutions:

- ✕ Retraining a new classifier ………..highly cost (Dataset, Time, Storage)
- ✕ Simply re-scale the input …………lead to the absence of some details
- √ Super-resolution ………………………increase the resolution & recover some details

➢ **Motivation**

Most SR methods pursue higher PSNR or better visual quality. However, it is not clear whether PSNR or visual quality determines the quality of visual analytics results, for example, action recognition accuracy.

➢ **Contribution**

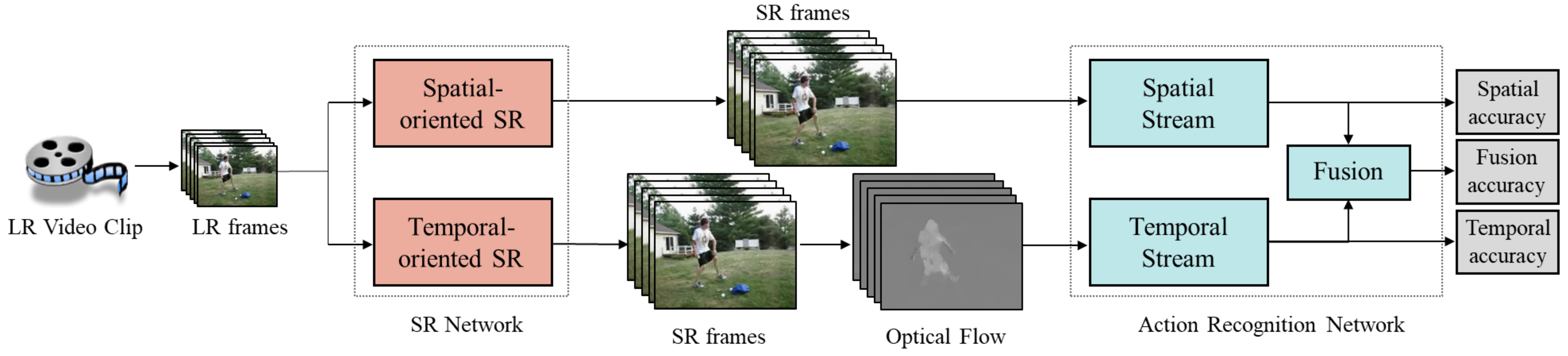
Investigated state-of-the-art image and video SR methods from the view of facilitating action recognition.

Tailored for two-stream action recognition framework:

- ✓ For the spatial stream, we propose an optical flow weighted MSE loss to guide our SoSR in paying more attention to regions with motion.
- ✓ For the temporal stream, we propose ToSR which enhances the consecutive frames together to achieve temporal consistency.

Verified the effectiveness of our methods by experimenting with two different recognition networks on two widely used datasets.

Method



SoSR

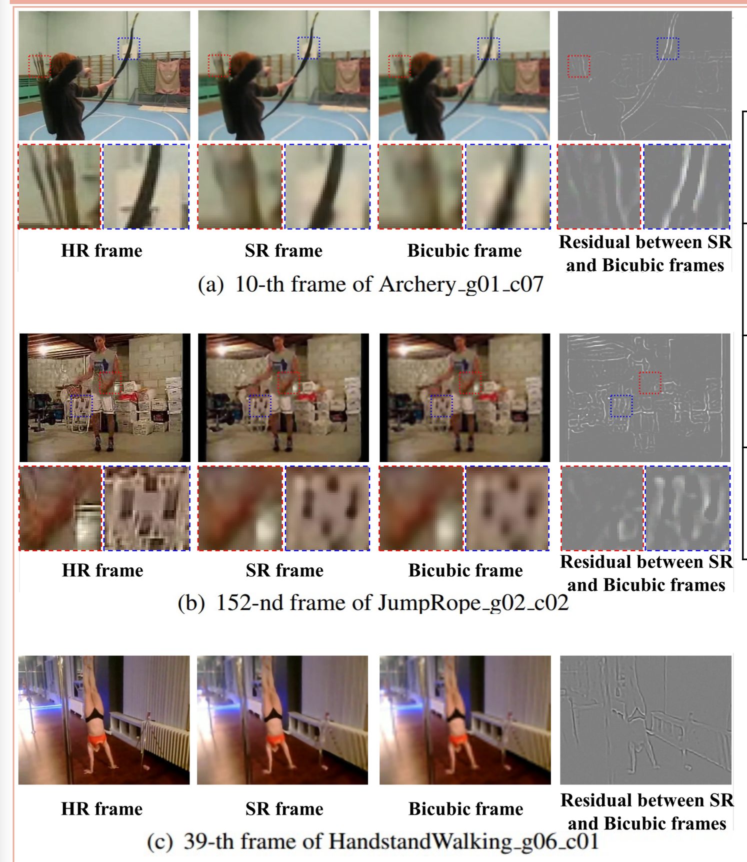


Table1. Different cases in recognition accuracy for the classes in UCF101 using the TSN network.

Case	Class	Recognition Accuracy (%)		
		HR	Bicubic	VDSR
a	Archery	**82.93**	36.59	70.73
	PlayingFlute	**97.92**	72.92	79.17
b	JumpRope	39.47	**42.11**	7.89
	SalsaSpin	79.07	**83.72**	53.49
c	FrontCrawl	64.86	32.43	**78.38**
	HandstandWalking	35.29	29.41	**41.18**

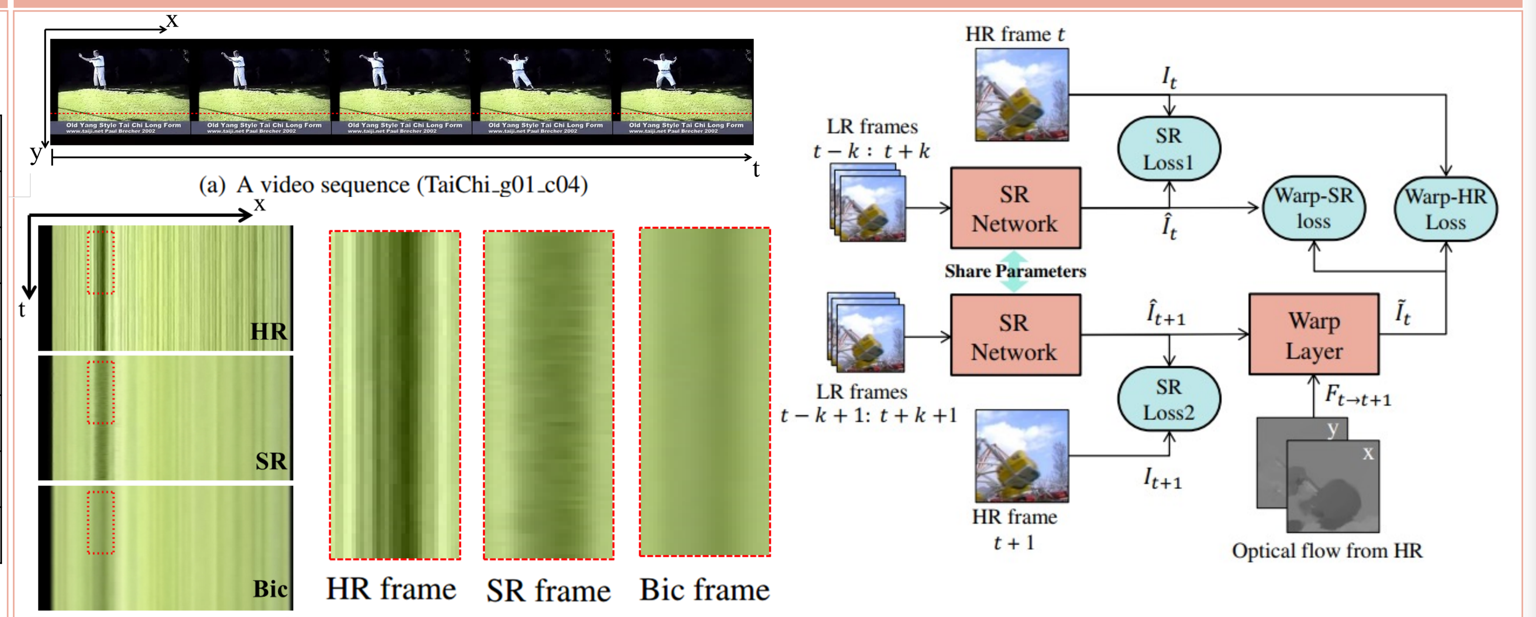
Conclusion

SR method should selectively enhance the image regions that are highly related to action recognition.

$$\mathrm{WMSE} = \frac{1}{N}\sum_{p=1}^{N}\left\|I(p) - \hat{I}(p)\right\|^2 \cdot \sqrt{u^2(p) + v^2(p)}$$

$$\mathcal{L}_{\mathrm{SoSR}} = \alpha\mathcal{L}_{\mathrm{WMSE}} + \beta\mathcal{L}_{\mathrm{Feature}} + \gamma\mathcal{L}_{\mathrm{Adversarial}}$$

ToSR



(b) Temporal profiles of different results

Conclusion

The unnatural connection between consecutive frames generates flicking artifact which affects the quality of optical flow and further incurs drop in recognition accuracy.

$$\mathcal{L}_{\mathrm{ToSR}} = \alpha\mathcal{L}_{\mathrm{SR}} + \beta\mathcal{L}_{\text{warp-SR}} + \gamma\mathcal{L}_{\text{warp-HR}}$$

$$\begin{cases} \mathcal{L}_{\mathrm{SR}} = \|I_t - \hat{I}_t\|_F^2 + \|I_{t+1} - \hat{I}_{t+1}\|_F^2 \\ \mathcal{L}_{\text{warp-SR}} = \|\hat{I}_t - \tilde{I}_t\|_F^2 \\ \mathcal{L}_{\text{warp-HR}} = \|I_t - \tilde{I}_t\|_F^2 \end{cases}$$

Two-Stream framework

Two-stream: One for exploiting spatial information from individual frames; The other for using temporal information from optical flow

Late Fusion: **TSN** uses a weighted average; **ST-ResNet** trains a fusion sub-network

Performance

- ➢ **Training Dataset**: CDVL-134 (Collected by ourselves) **Testing Dataset**: HMDB51 and UCF101
- ➢ **HR** = Original resolution video **LR** = 4X bicubic down-sampled video

Table2. Recognition accuracy (%) of 4x super-resolved video from UCF101 and HMDB51 dataset using two action recognition network, TSN and ST-ResNet. Number of VSR-DUF [14] indicates number of layers. Accuracy of HR video is provided for reference. (Please refer to the supplementary material for more results.)

Method	HMDB51						UCF101					
	TSN			ST-ResNet			TSN			ST-ResNet		
	Spatial	Temporal	Fusion	Spatial	Temporal	Fusion	Spatial	Temporal	Fusion	Spatial	Temporal	Fusion
Bicubic	42.81	56.54	63.53	43.59	53.76	59.48	71.25	81.08	87.87	72.01	78.28	84.62
VDSR [17]	46.6	55.1	63.59	49.18	54.44	60.2	67.09	79.81	86.84	72.27	79.43	84.48
RCAN [44]	48.76	56.8	66.21	51.76	55.72	62.61	67.18	82.12	88	72.23	80.52	85.01
SRGAN [20]	48.82	49.87	63.01	51.41	47.22	60.85	81.33	75.45	87.55	83.31	70.16	86.97
ESRGAN [40]	52.48	51.5	63.4	53.79	49.72	61.83	82.97	75.32	87.75	83.81	70.64	86.62
SoSR	**53.59**	50.26	64.51	**54.77**	48.27	63.01	**83.11**	74.1	86.63	**83.92**	69.68	85.77
SPMC [34]	48.95	56.41	64.31	53.14	53.53	63.66	70.42	80.19	87.15	74.45	77.44	84.09
VSR-DUF-16 [14]	48.37	59.48	66.08	50.62	55.07	61.11	68.56	84.89	89.36	72.11	80.06	83.9
VSR-DUF-52 [14]	48.5	60.52	66.86	52.84	57.61	65.23	70.54	85.09	89.85	74.49	80.16	84.88
ToSR	47.45	**61.5**	66.08	51.54	**58.92**	64.77	64.79	**85.29**	88.46	70.88	**81.07**	83.82
SoSR+ToSR	/	/	68.3	/	/	**67.32**	/	/	**92.13**	/	/	**90.19**
HR	*54.58*	*62.16*	*69.28*	*56.01*	*59.41*	*68.1*	*86.02*	*87.63*	*93.49*	*88.01*	*85.71*	*92.94*

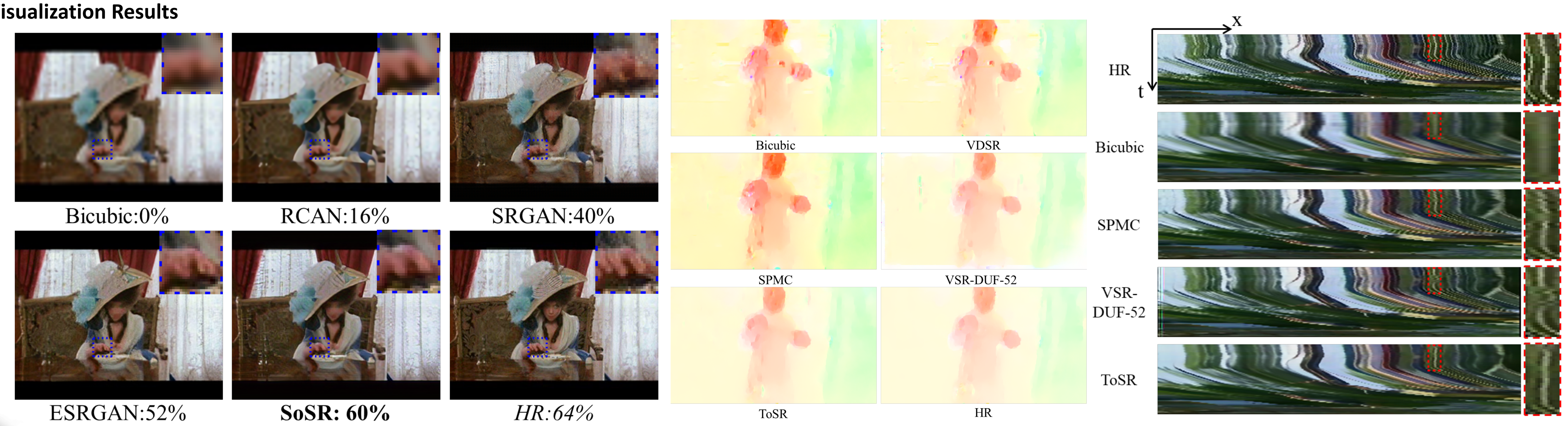
Table3. Ablation study for SoSR using different network structures and different loss functions, with TSN [38] on HMDB51 dataset.

Structure	MSE/WMSE	Feature	Adversarial	Accuracy
VDSR	MSE	-	-	46.6%
VDSR	WMSE	-	-	47.91%
VDSR	WMSE	√	-	50.39%
ESRGAN	WMSE	√	-	52.55%
ESRGAN	MSE	√	√	52.48%
ESRGAN	WMSE	√	√	**53.59%**

Table4. Ablation study for ToSR using different network structures and different loss functions, with TSN [38] on HMDB51 dataset.

Structure	Warp loss	Accuracy
VDSR	-	55.1%
VDSR	√	58.76%
VSR-DUF-16	-	59.48%
VSR-DUF-16	√	**61.5%**

Visualization Results



FAQ

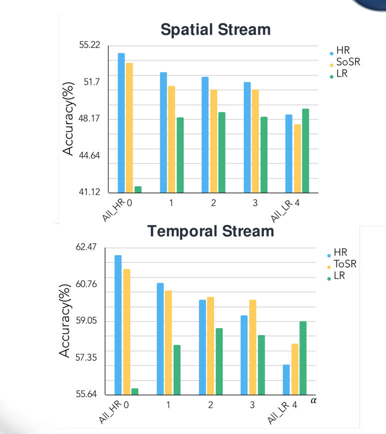


Table5. PSNR and SSIM of Y channel of 4X super-resolved video from UCF101 and HMDB51 dataset.

Method	UCF101		HMDB51	
	PSNR	SSIM	PSNR	SSIM
RCAN	30.9208	0.6983	33.0629	0.6826
ESRGAN	29.558	0.5959	31.2243	0.5711
SoSR	28.7279	0.5493	29.6327	0.5464
VSR-DUF-52	**31.9657**	**0.7297**	**33.7269**	**0.7067**
ToSR	30.8365	0.6935	32.8421	0.6718

Reference

[14] Y. Jo, et al. Deep video super-resolution network using dynamic up-sampling filters without explicit motion compensation. In *CVPR*, 2018.
[17] J. Kim, et al. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016.
[20] C. Ledig, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.
[34] X. Tao , et al. Detail-revealing deep video super-resolution. In *ICCV*, 2017.
[38] L. Wang , et al. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.
[40] X. Wang , et al. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018.
[44] Y. Zhang, et al. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018.