Two-Stream Action Recognition-Oriented Video Super-Resolution

Haochen Zhang, Dong Liu, Zhiwei Xiong

National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application

Background

CNNs have been applied to action recognition task and obtained state-of-the-art performance.

These well-trained CNNs cannot be directly applied on LR video because of the FC layers.

Solution:

- Retraining a new classifier ..... highly cost
- Simply re-scale the input ..... lose some details
- **Super-resolution** ..... resize the input & recover some details

Motivation

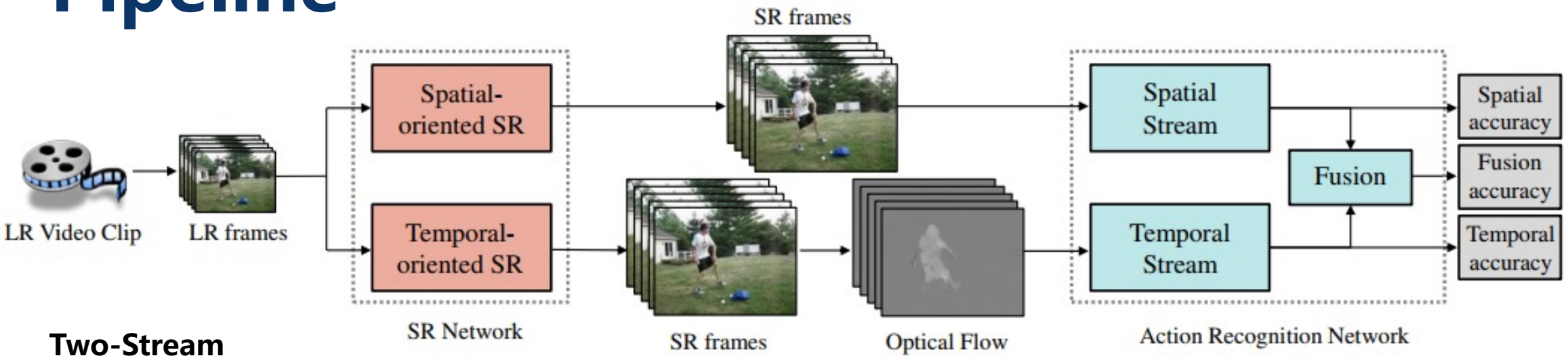
Super-resolution

Q1: What do most SR methods pursue? ..... PSNR and visual quality

Q2: Does PSNR or visual quality determine the quality of visual analytics results, e.g. action recognition accuracy? .....Not clear

This motivates us to investigate the video SR problem aiming to facilitate action recognition accuracy, instead of visual quality.

Pipeline



Two-Stream

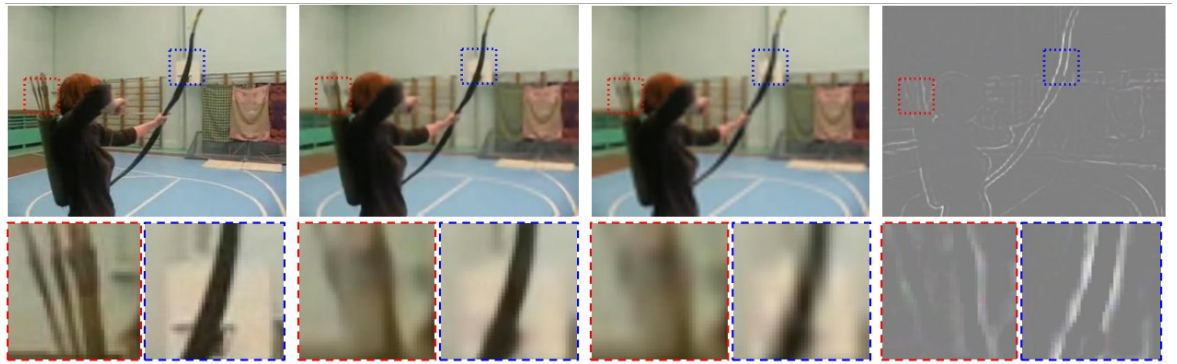
Two separate networks

- One for exploiting spatial information from individual frames----------SoSR
- The other for using temporal information from optical flow ----------ToSR

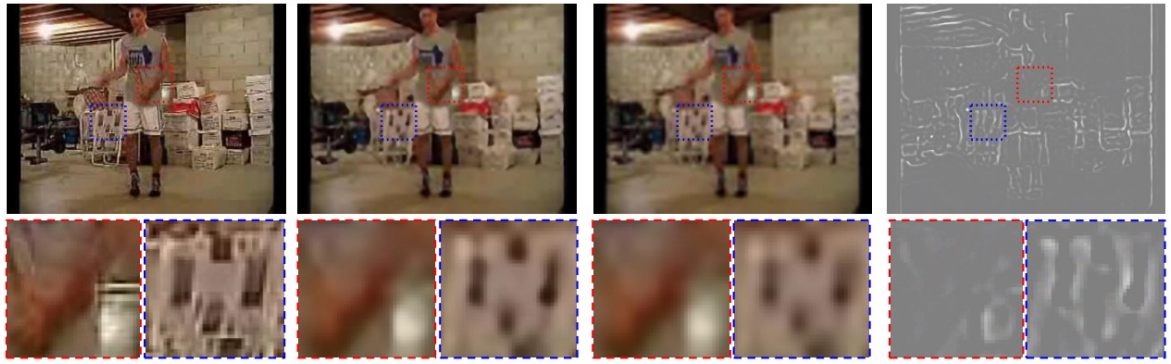
Late fusion

- **TSN** uses a weighted average
- **ST-Resnet** trains a fusion sub-network

SoSR Analysis



(a) 10-th frame of Archery_g01_c07



(b) 152-nd frame of JumpRope_g02_c02

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.

SoSR Method

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

weighted

- The optical flow is calculated offline from the HR video
- Guide the network in a pixel-wise manner
- Combine with feature loss and adversarial loss

Ablation Study

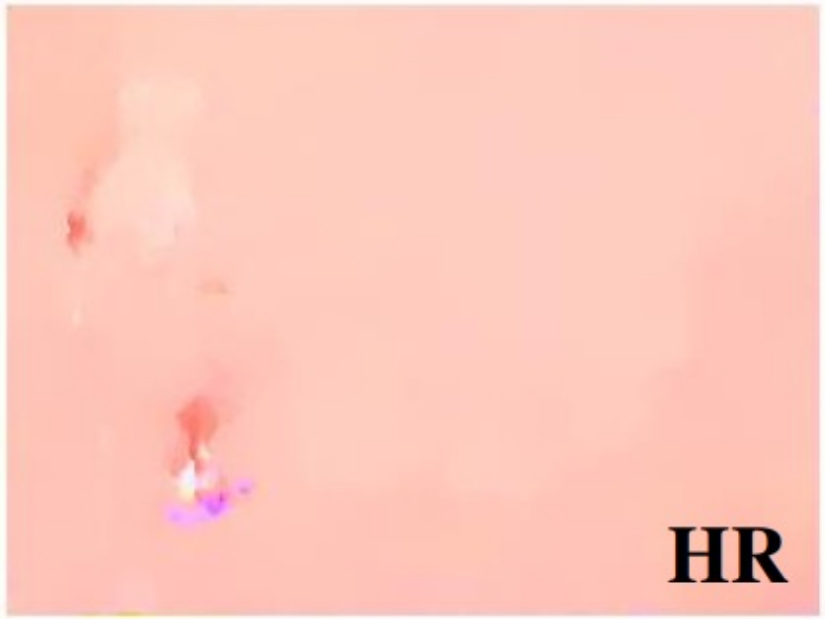
Structure	MSE/WMSE	Feature	Adversarial	Accuracy
VDSR	MSE	-	-	46.6%
VDSR	WMSE	-	-	47.91%
VDSR	WMSE	✓	-	50.39%
ESRGAN	WMSE	✓	-	52.55%
ESRGAN	MSE	✓	✓	52.48%
ESRGAN	WMSE	✓	✓	**53.59%**

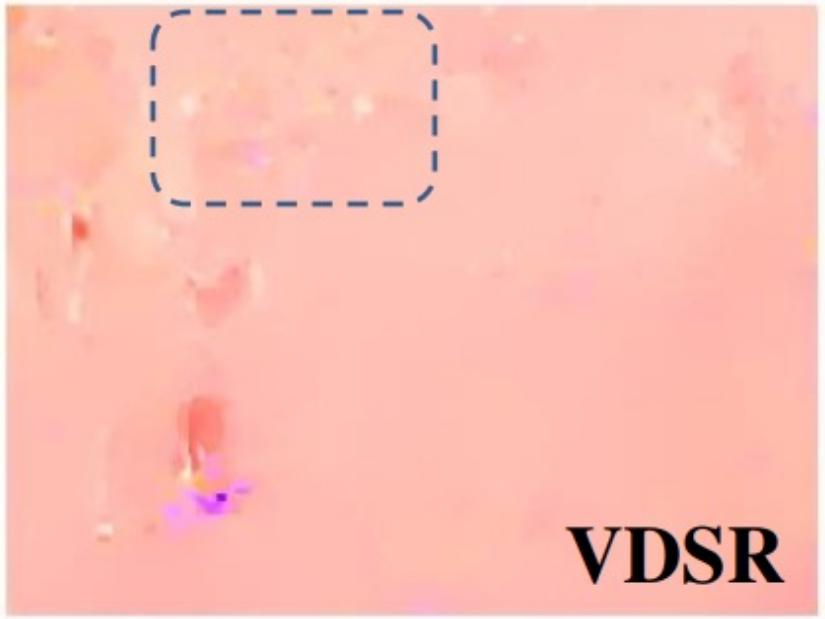
Table 2. Ablation study for SoSR using different network structures and different loss functions, with TSN [38] on HMDB51 dataset.

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

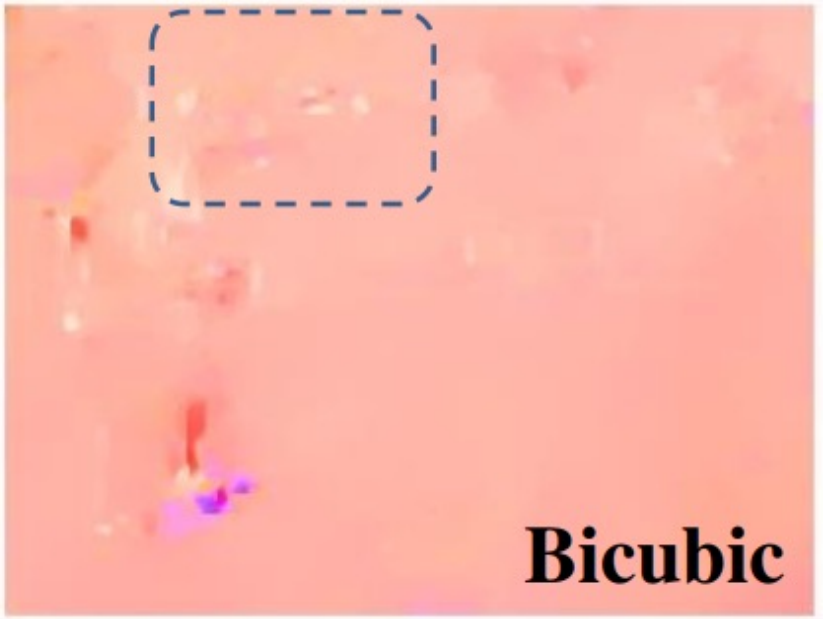
ToSR Analysis

Optical Flow



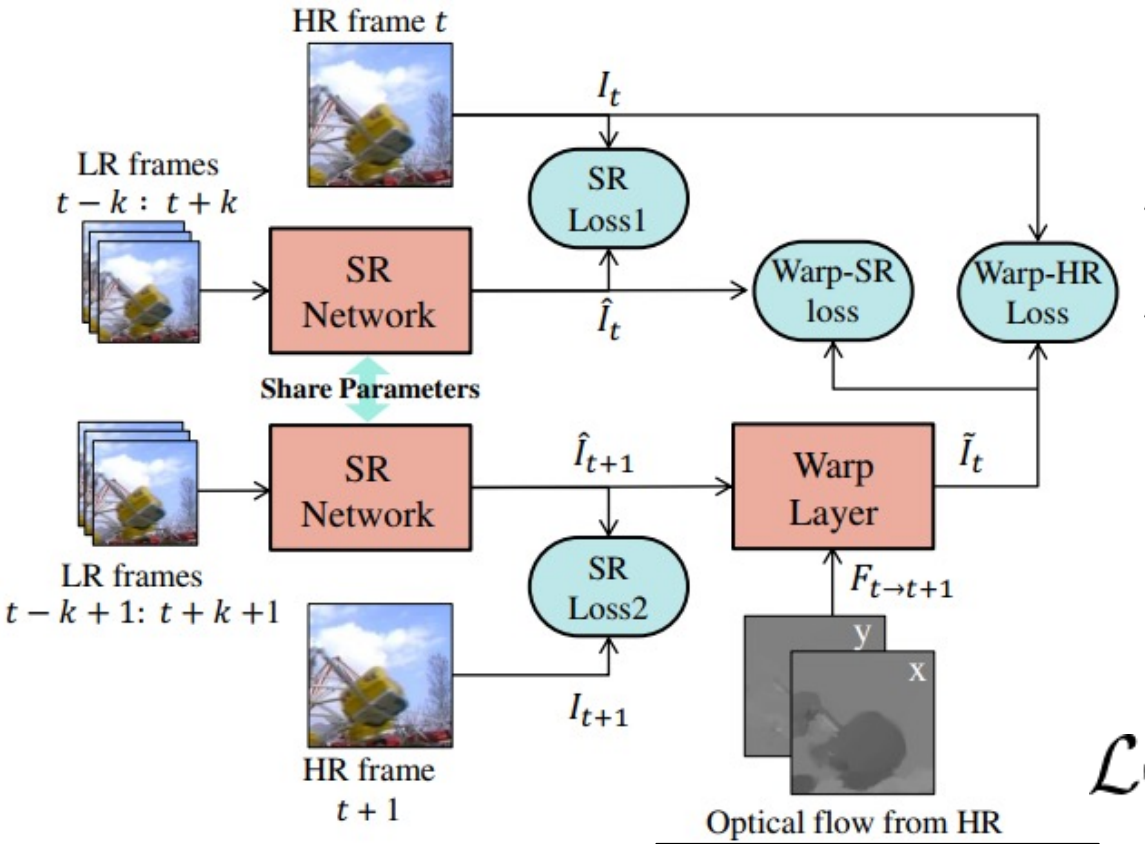


More Artifacts



Artifacts

ToSR Method



Calculated offline

$$\mathcal{L}_{\text{SR}} = \|I_t - \hat{I}_t\|_F^2 + \|I_{t+1} - \hat{I}_{t+1}\|_F^2$$

$$\mathcal{L}_{\text{warp-SR}} = \|\hat{I}_t - \tilde{I}_t\|_F^2$$

$$\mathcal{L}_{\text{warp-HR}} = \|I_t - \tilde{I}_t\|_F^2$$

Ablation Study

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

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

3 | RESULTS & DISCUSSION

Training Dataset: CDVL-134 (Collected by ourselves)

Testing Dataset: HMDB51 and UCF101

HR = Original resolution video
LR = 4X bicubic downsampled video

	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
SISR	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
SISR	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
SISR	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
SISR	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
SISR	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
VSR	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	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	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
VSR	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*

Table 4. Recognition accuracy (%) of 4× 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 PSNR and SSIM results of different methods.)

Spatial

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

1. Compared with optimizing MSE only, using perceptual loss could improve the recognition performance.
2. Our WMSE makes further gain.

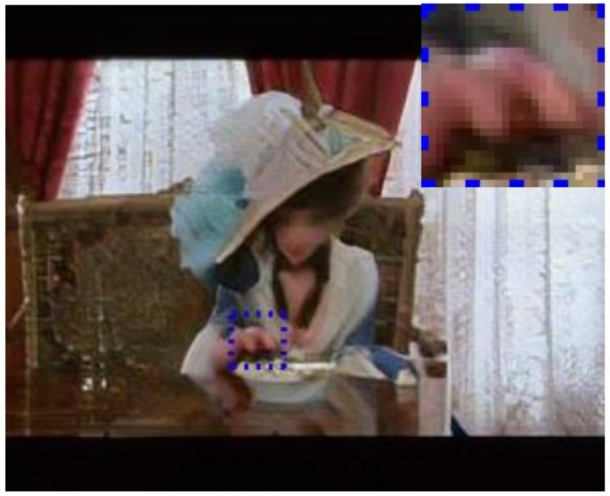
❸ | RESULTS & DISCUSSION

Bicubic:0%

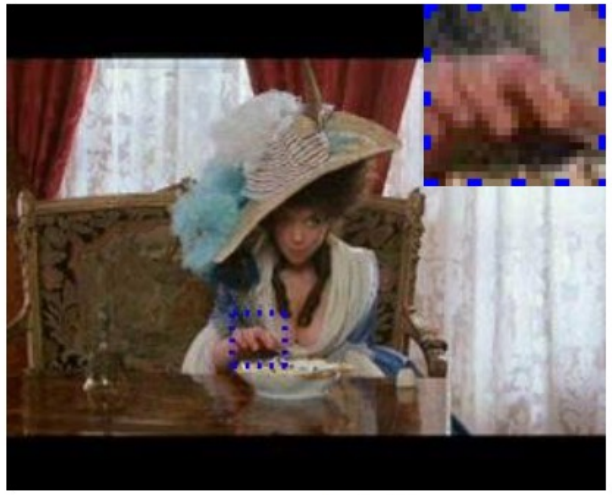
RCAN:16%

SRGAN:40%

ESRGAN:52%



SoSR: 60%



HR:64%

Visual quality:

SoSR<ESRGAN

but

Accuracy :

SoSR>ESRGAN

Temporal

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

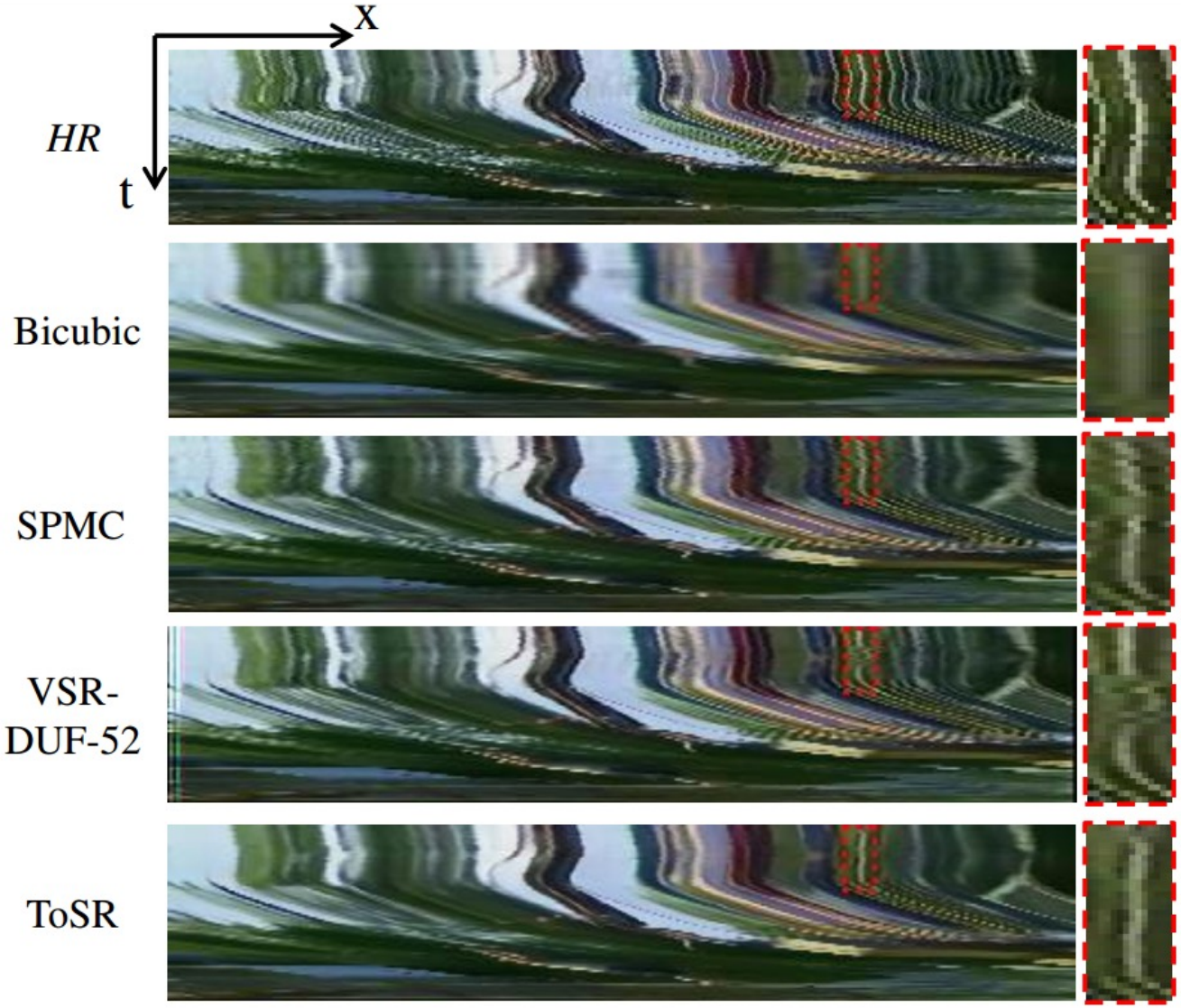
Perceptual loss worsen the performance of temporal recognition.

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

SPMC performs explicit motion compensation with optical flow estimated from LR frames
VSR-DUF extracts feature directly from video using 3D convolution

Our warp loss makes further gain.

3 | RESULTS & DISCUSSION



best temporal consistency

More visualization please refer to our supplementary material

4 | CONCLUSION

Have done

We consider the video SR problem for facilitating action recognition accuracy.

Tailored for two-stream action recognition networks, we propose:

✓ SoSR with a optical flow guided weighted MSE loss

✓ ToSR with a siamese network to emphasize temporal consistency.

Experimental results demonstrate the advantages of our proposed SoSR and ToSR methods.

To do

In the future, we plan to

- Combine SoSR and ToSR into a single step
- Study the tradeoff between visual quality and recognition accuracy.

Thank you for your listening

Welcome to my poster: 116 on Nov. 1 morning

NEL-BITA
类脑智能技术及应用国家工程实验室

End-to-End Optimization

Q:How about training the SR network with the recognition accuracy as the objective.

A:This involves a specific action recognition model.

Our empirical results indicate that, training the SR with a specific action recognition model (e.g. TSN), and testing with another model (e.g. STResnet), leads to much worse results.

However, as a low-level restoration, SR should benefit recognition generally. So it motivates us to design specific loss functions for the SR training.

Retraining the classifier

Q: What if the networks are trained with LR video?

A: we train several TSN models using mixed HR and LR video

The network trained on LR video only performs the best on LR video input, but performs the worst on HR video input. Thus, it should not be a good choice in practice.

Retraining different models is a straightforward solution for different resolutions, but has several limitations.

- It needs to train and maintain multiple models that can be costly.
- How to select the appropriate model to match the resolution for a given input video is a problem.

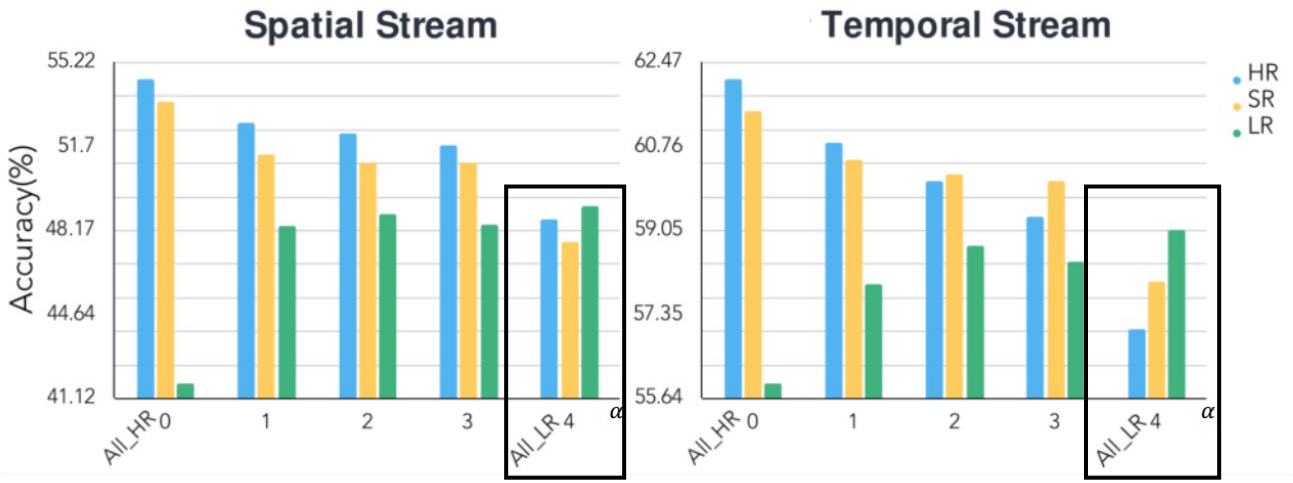


Figure 8. Recognition accuracy of models trained with different data augmentation configurations (denoted by α) and tested on HR, SR, LR video respectively.

PSNR comparison of different SR

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

Table b. PSNR and SSIM of Y channel of 4× super-resolved video from UCF101 and HMDB51 dataset.

Indeed, SoSR and ToSR make gain in recognition accuracy at cost of PSNR.