Supplementary Material for Two-Stream Action Recognition-Oriented Video Super-Resolution

Haochen Zhang, Dong Liu, Zhiwei Xiong
CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System,
University of Science and Technology of China, Hefei 230027, China
zhc12345@mail.ustc.edu.cn, {dongeliu, zwxiong}@ustc.edu.cn

This supplementary document consists of the following results:

- Table a presents the recognition accuracy results of combining different methods for spatial and temporal streams.
- Table b presents the PSNR and SSIM results of different methods, including our proposed SoSR and ToSR.
- Figure a and Table c show a case where visual quality and recognition accuracy are not consistent.
- Figure b and Figure c show the comparison of optical flow maps and temporal profiles, respectively.

References

[JAFF16] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016, pp. 694–711.

Spatial	Temporal	TSN	ST-Resnet
RCAN	VSR-DUF-52	66.27%	64.44%
ESRGAN	VSR-DUF-52	67.49%	66.34%
ESRGAN	ToSR	67.58%	66.99%
SoSR	VSR-DUF-16	68.24%	66.41%
SoSR	ToSR	**68.30%**	**67.32%**

Table a. Recognition accuracy of 4× super-resolved video from HMDB51 dataset using two action recognition network, TSN and ST-Resnet.

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.

Dataset	Recognition Accuracy				
	Conv1_2	Conv2_2	Conv3_3	Conv4_3	Conv5_3
HMDB51	48.1	48.69	49.48	49.87	50.39
UCF101	72.77	73.4	76.88	77.06	79.15

Table c. Recognition accuracy (%) of super-resolved video enhanced by VDSR networks that are trained with feature loss based on different levels of VGG-16 network. We can find the accuracy increases consistently as the level of the feature deepens. Please refer to Figure a for visual inspection.

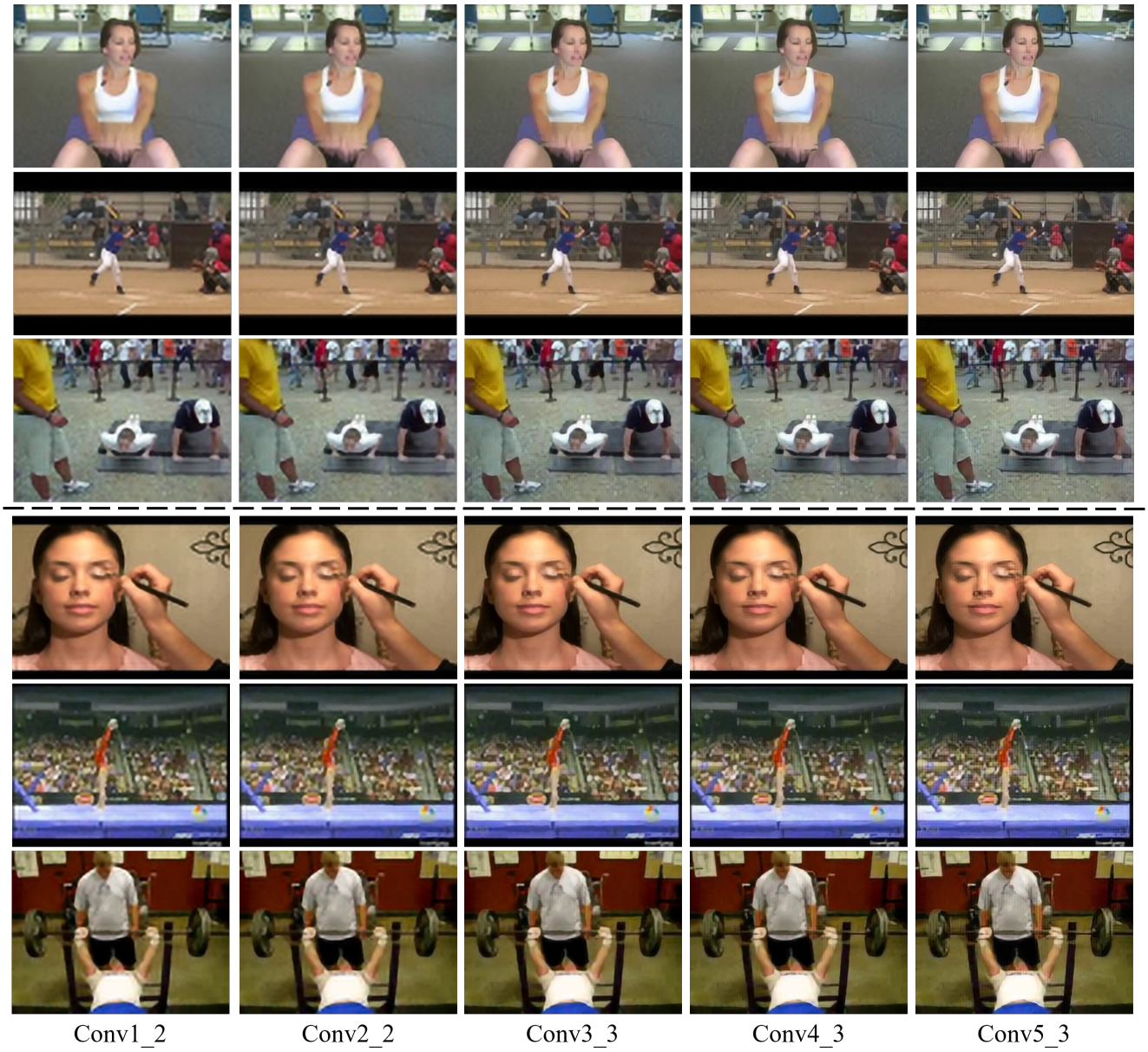
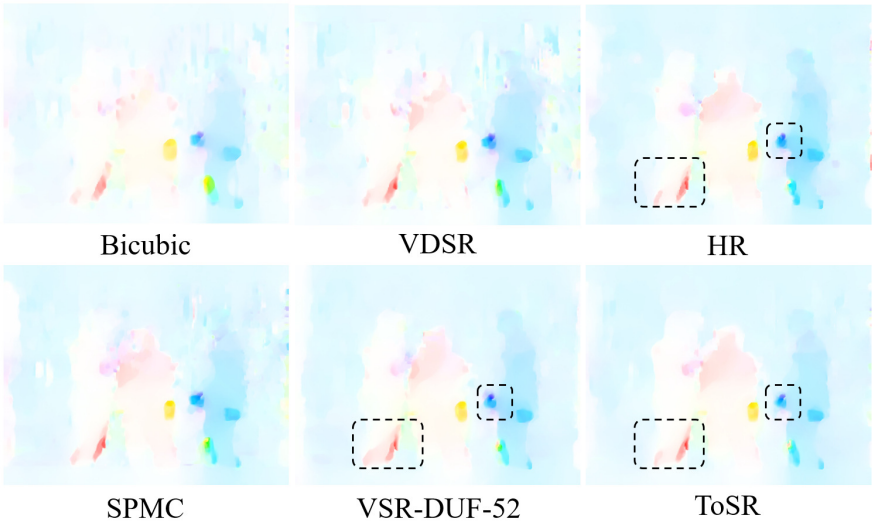
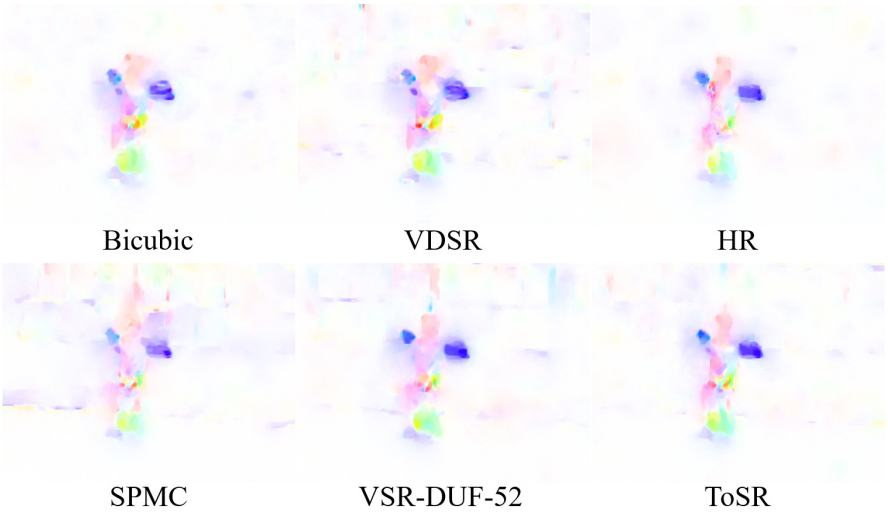


Figure a. Visual quality comparison (example frames above the dotted line come from HMDB51 dataset and frames below the dotted line come from UCF101 dataset), please refer to Table c for corresponding recognition accuracy of TSN. When we use feature loss to train VDSR networks, we could observe grid-like artifacts on the resulting images and the higher level features we use, the more visible these artifacts are. These artifacts are also reported in [JAFF16] and Johnson *et al*. recommend "Conv2_2" or "Conv3_3" for the best visual quality. However, the results in Table c indicate that using "Conv5_3" achieves the best recognition performance.



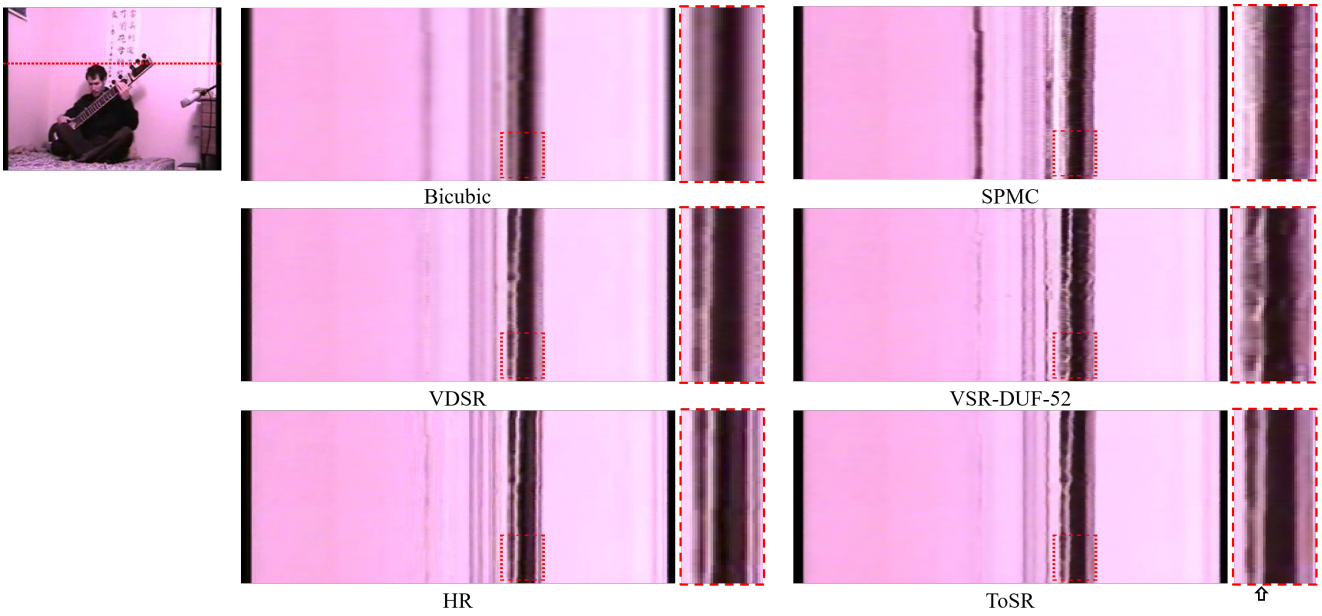
(a) An example frame for the `ShakeHands` class from HMDB51.

(b) An example frame for the `Punch` class from HMDB51.

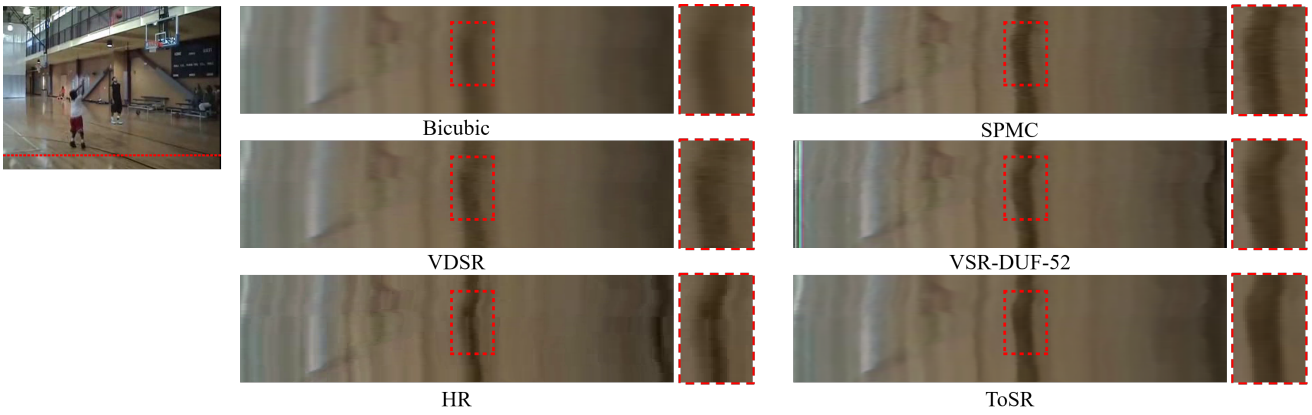


(c) An example frame for the `TaiChi` class from UCF101.

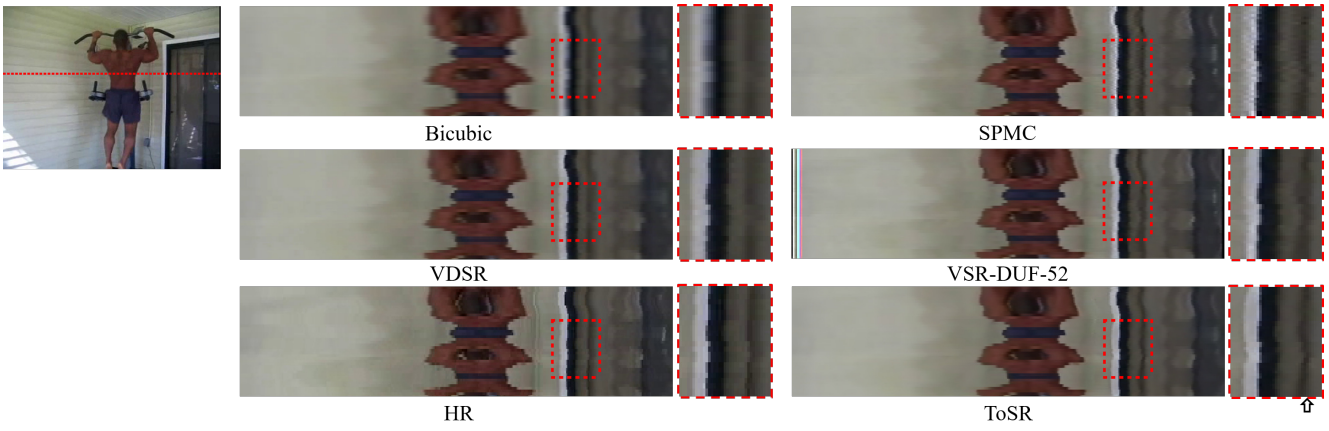
Figure b. Optical flow maps calculated from resulting frames of different SR methods. Our ToSR results have the optical flow maps most similar to HR.



(a) Example video for the `PlayingSitar` class from UCF101.



(b) Example video for the `Shoot` class from HMDB51.



(c) Example video for the `PullUp` class from HMDB51.

Figure c. Temporal profiles of different SR results, showing our ToSR results have the least flicking artifacts.