Space-Time-Aware Multi-Resolution Video Enhancement —Supplementary Materials—

Muhammad Haris[1]*, Greg Shakhnarovich[2], and Norimichi Ukita[1]
[1]Toyota Technological Institute, Japan [2]Toyota Technological Institute at Chicago
muhammad.haris@bukalapak.com,greg@ttic.edu,ukita@toyota-ti.ac.jp

1. Network Design

Unlike video SR methods [6, 9] that accept input frames recurrently, a set of input frames (I_t^l and I_{t+1}^l) are fed into a network simultaneously for their ST-SR outputs (e.g., I_{t+n}^{sr}) in STAR as with [3, 7], as illustrated in Figure 3. The next inputs (I_{t+1}^l and I_{t+2}^l) are then fed into the network with no memory given by the last cycle that accepts I_t^l and I_{t+1}^l. This current implementation can be extended to the one with recurrent input frames.

2. Model Improvement

Multi-frame. STAR can be improved by accepting more input frames as video SR methods (e.g., RBPN) do. STAR uses RBPN as $\texttt{Net}_S$, so that $\texttt{Net}_S$ can be easily extended to accept a multi-frame input.

Interpolation Rate. STAR can interpolate one frame at any moment, $t + n(n \in [0, 1])$. Each STAR is trained for a specific n. For example, for T-SR by factor of 4, we have two options. (1) Three STARs are trained for $n = \frac{1}{4}, \frac{2}{4}, \frac{3}{4}$ independently and (2) STAR with $n = \frac{1}{2}$ is trained and used iteratively so that $n = \frac{1}{4}$ is generated from a pair of $n = 0$ and $n = \frac{1}{2}$. Note that we can also extend STAR so that it is trained for any n by following standard practice in frame interpolation literature. In this paper, we did not evaluate such regimes, to be consistent with previous work [1, 2, 8].

3. Details on Flow Computation

Flow Refinement Explanantion. $F_{t\rightarrow t+n}$ +$F_{t+n\rightarrow t+1}$ is elementwise sum of the two flow fields. The magnitude of this half-way flows is generally lower, which brings them to the same average scale as the $t \rightarrow t + 1$ flow. In practice, we find that this operation has a localized smoothing effect, improving the interpolation, as seen in Figure 5.

The computation of $F_{t\rightarrow t+n}$. During training, we have access to the ground truth of interpolated frames (I_{t+n}^l). So that, $F_{t\rightarrow t+n}$ is the function of optical flow computation which calculated flows from an input frame at t to this ground truth (i.e., from I_t^l to I_{t+n}^l).

*He is currently working at Bukalapak in Indonesia.

4. Additional Experimental Results

4.1. The Upper-Bound Performance of Flow Refinement Module

Further analysis of the Flow Refinement (FR) module is performed. Our STAR used bidirectional dense motion flow maps, $F_{t\rightarrow t+1}$ and $F_{t+1\rightarrow t}$ that are computed only by I_t^l and I_{t+1}^l as explained in the main manuscript. However, the flow noise exists when calculating large motions objects.

In order to further investigate (1) the bad impact of noisy flows on ST-SR in the test stage and (2) the effectiveness of the FR module for suppressing this bad impact, we compare the ST-SR performances in different conditions as follows.

The bad impact of the noisy flows can be confirmed if we have the ground-truth of the flow map for comparison. While this ground-truth is not available, we propose to use flow maps computed by GT in-between frames of I_t^l and I_{t+1}^l. We call this flow *GT flow*. Remember that the GT flow is used in our proposed STAR for training the FR module using L_{flow} (Eq. (15) in the main manuscript). The results have been proven in our experiments. On the other hand, the GT flow is unavailable in the test stage because the GT in-between frame is what STAR tries to predict. However, for further analysis, we can also assume that the upper-bound performance of the FR module is by giving access to GT flow also during testing. We show the upper-bound performance of our FR module in Table 1, specifically in the 4th rows.

In the 1st and 2nd rows, we compare our STAR w/ and w/o FR module. We can see that the FR module improves the performance by 0.11dB and 0.09dB on I_t^{sr} and I_{t+}^{sr}, respectively. The 3rd and 4th rows show the results obtained by the GT flow. We can also see that FR is able to improve the performance even using GT flow, but not substantial: 0.035dB and 0.027dB in I_t^{sr} and I_{t+}^{sr}, respectively.

The use of GT flow proves an improvement by comparing the 1st and 3rd rows: 0.055dB and 0.111dB in I_t^{sr} and

FR[1]	GT Flow[2]	I_t^{sr} PSNR	I_t^{sr} SSIM	I_{t+}^{sr} PSNR	I_{t+}^{sr} SSIM
-	-	32.257	0.937	30.617	0.927
✓	-	32.349	0.938	30.704	0.928
-	✓	32.312	0.938	30.728	0.928
✓	✓	32.347	0.938	30.755	0.928

Table 1. Analysis on FR module using STAR-ST with RBPN (L_r). The first column (FR[1]) indicates whether FR is used or not. The second column (GT flow[2]) indicates whether GT flow is used during testing or not.

I_{t+}^{sr}, respectively. The 2[rd] and 3[th] rows show an interesting comparison where FR has comparable performance to the one w/o FR but using GT flow. This shows that FR is able to generate semantic contents closely enough to the GT flow. Finally, we can see that the 2[nd] and 4[th] rows shows a small margin. This means that FR is able to improve the performance near to the upper-bound performance.

5. Additional Visual Results

5.1. Visual Results on ST-SR

More visual results on ST-SR are shown in Fig. 1. Here we can see that our STAR outperforms the other methods. In the first row, we can see that our STAR is able to predict the motion of the right arm without losing the pinkish texture on it, which is indicated by the red arrow in the images. In the second row, the main difference is that the drum stick has a clearer texture compared to the other methods. In the third row, our STAR is able to interpolate the motion of the horse's leg better than the other methods. In the fourth and fifth rows, we can see that the structure of the hoop is clearer in our STAR's results. Finally, in the last row, we can see that STAR has a better prediction on the ball motion.

5.2. Visual Results on T-SR

More visual results of T-SR are shown in Fig. 2. Mostly same as in ST-SR results, in the T-SR results, our STAR is able to construct better motion and texture compared to the other methods.

5.3. Visual Results on S-SR

More visual results of S-SR are shown in Fig. 3. Here also, our STAR shows sharper and better detail compared to the other methods, as shown by the red arrows.

5.4. Video Results

Here we show some samples of video frames as in Fig. 4. The full video results of STSR (4x2r) from the HD video dataset are available in the following link.

5.5. Failure cases

The performance improvement of STARnet builds upon the assumption that both tasks are mutually beneficial to each other. Here we show some results where S-SR task produce artifacts, then T-SR task also suffer from it as in Fig. 5.

References

[1] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *CVPR*, 2019.

[2] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *arXiv preprint arXiv:1810.08768*, 2018.

[3] 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 *CVPR*, 2017.

[4] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for single imaage super-resolution. *arXiv preprint arXiv:1904.05677*, 2019.

[5] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, 2019.

[6] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Advances in Neural Information Processing Systems*, pages 235–243, 2015.

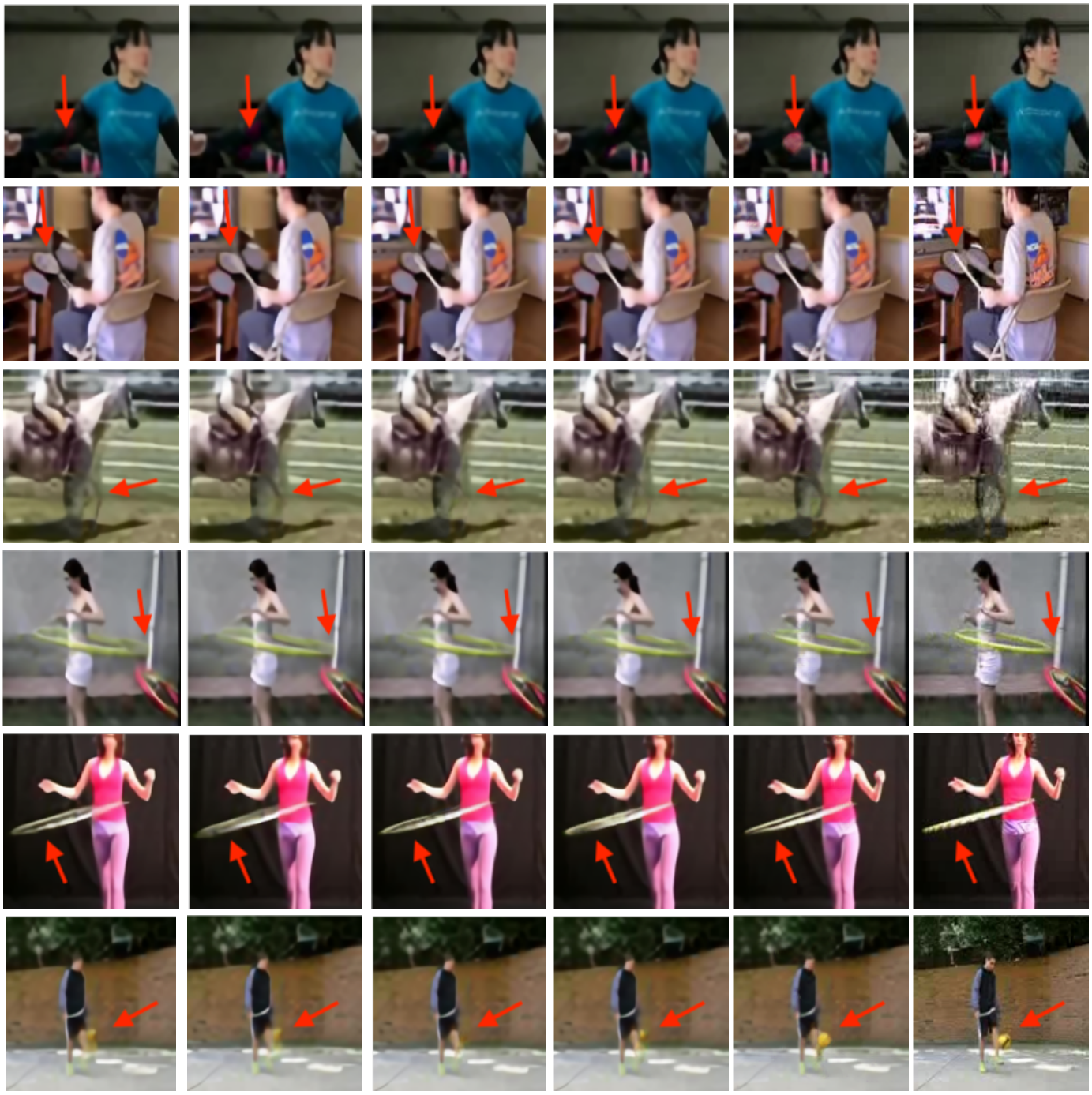
[7] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, pages 3224–3232, 2018.

[8] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *ICCV*, pages 261–270, 2017.

[9] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, pages 6626–6634, 2018.

[10] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[11] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127(8):1106–1125, 2019.



(a) ToFlow [11]→RBPN [5]

(b) DAIN [1]→RBPN [5]

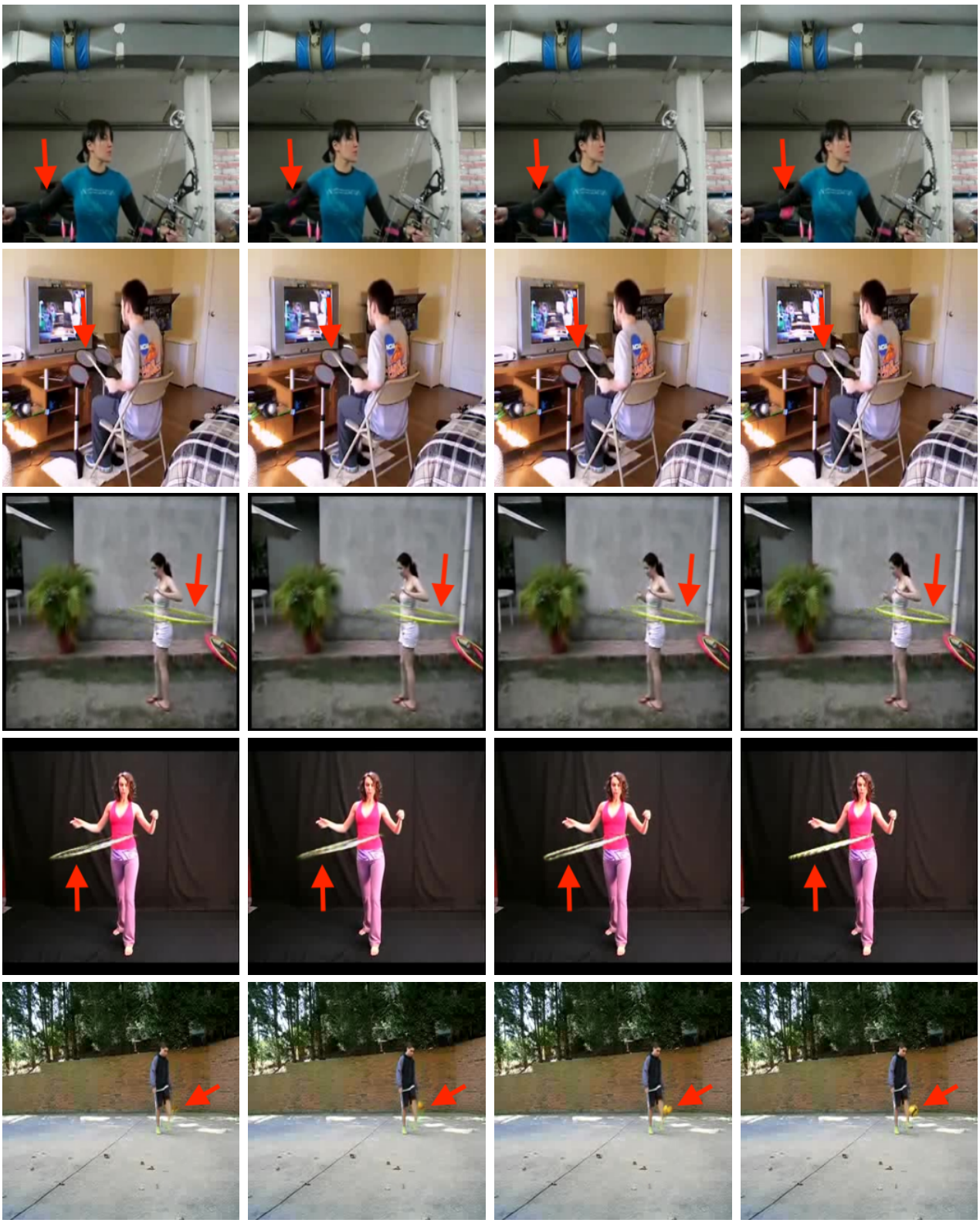
(c) RBPN [5]→ToFlow [11]

(d) RBPN [5]→DAIN [1]

(e) STAR-ST-FR

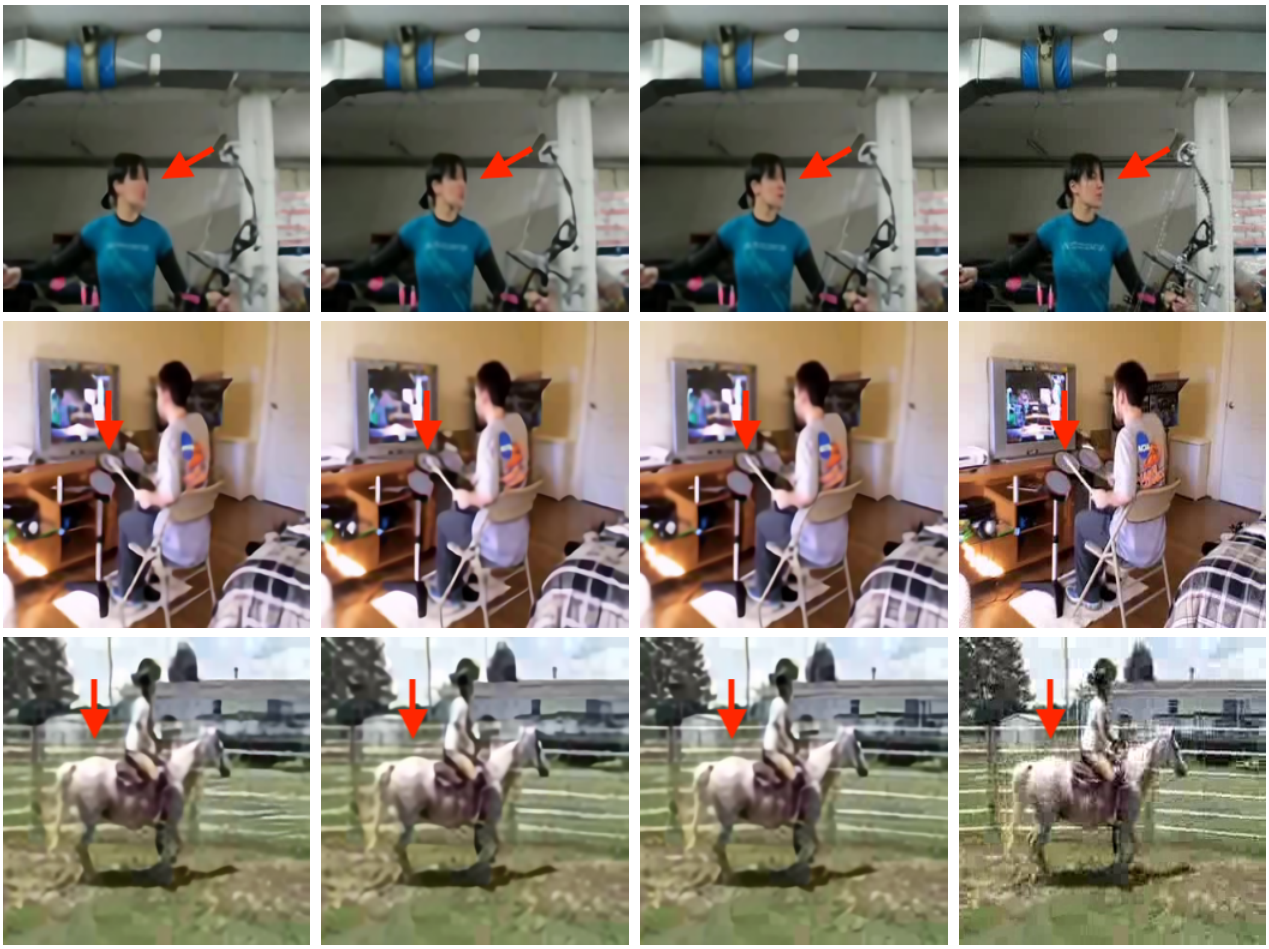
(f) GT

Figure 1. Additional visual results of ST-SR (I_{t+}^{sr}) on UCF101 [10]. Red arrows here and in the other figures indicate the highlighted area.



(a) ToFlow [11] (b) DAIN [1] (c) STAR-T-FR (d) GT

Figure 2. Additional visual results of T-SR on the original resolution of UCF101 [10].



(a) DBPN [4] (b) RBPN [5] (c) STAR-S-FR (d) GT

Figure 3. Additional visual results of S-SR on UCF101 [10].

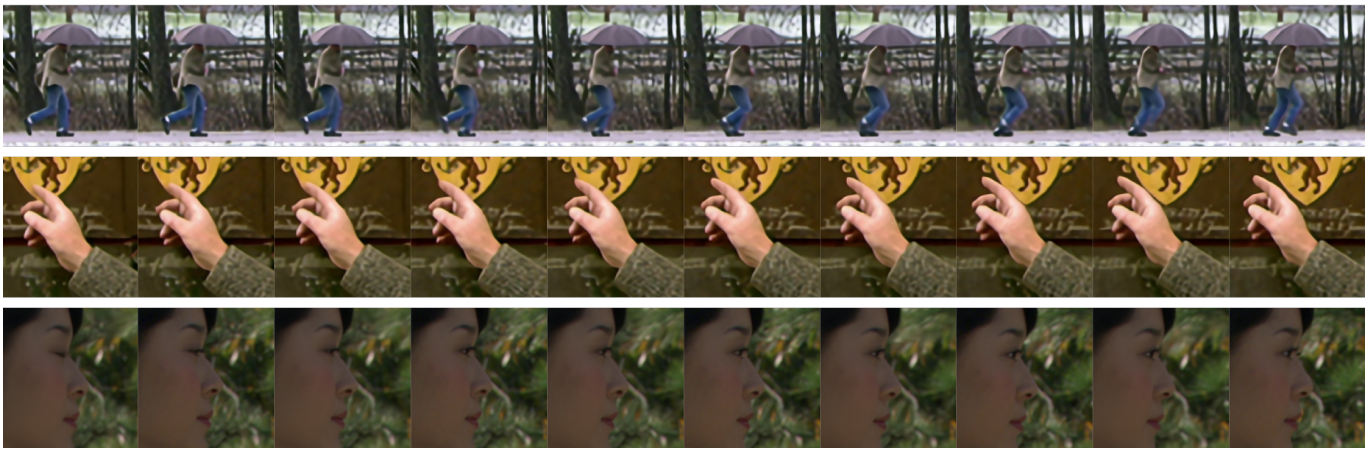


Figure 4. Sample of video results. Order are from left to right.

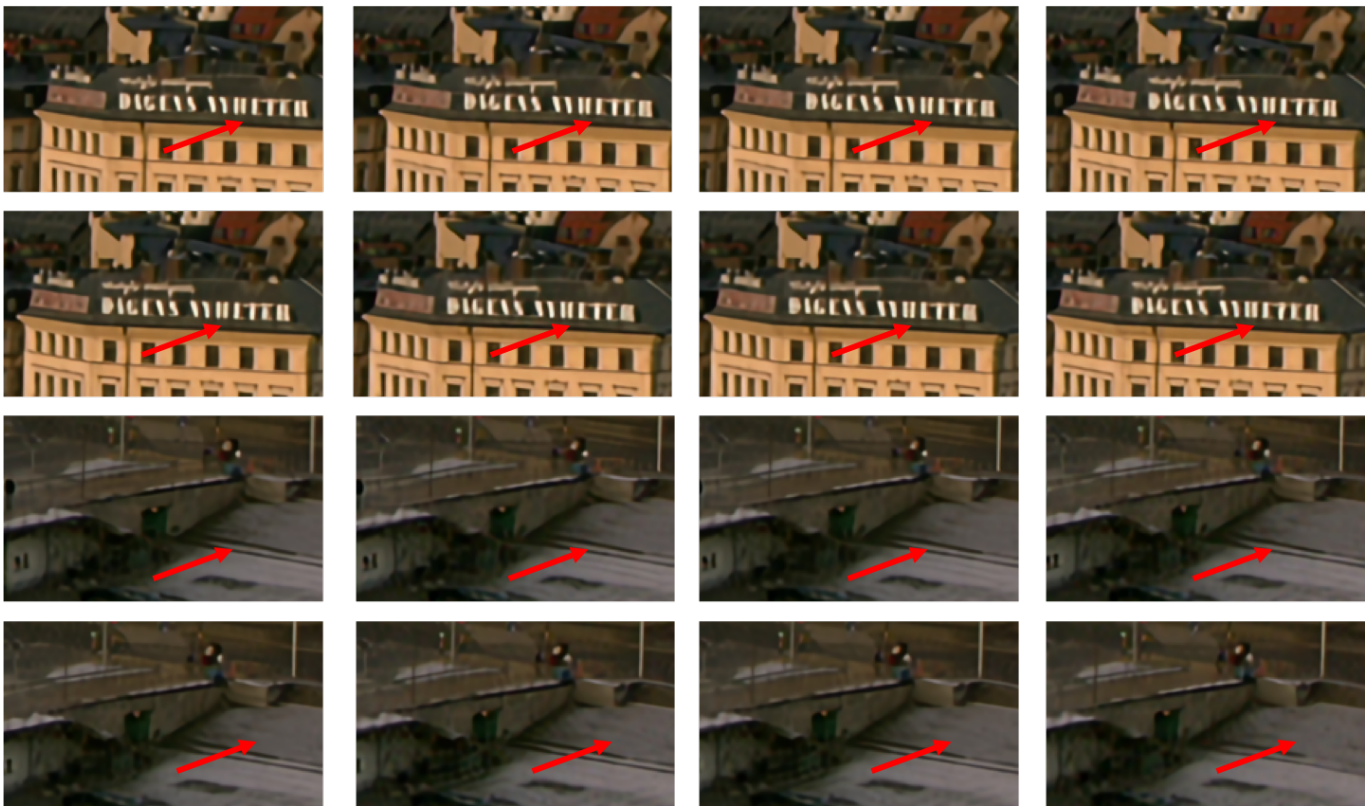


Figure 5. Samples of failure cases. Order are from left to right and top to down.