Supplementary Material for Omniscient Video Super-Resolution

Peng Yi[1], Zhongyuan Wang*[1], Kui Jiang[1], Junjun Jiang[2,3], Tao Lu[4], Xin Tian[5], and Jiayi Ma[5]

[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University [2]School of Computer Science and Technology, Harbin Institute of Technology [3]Peng Cheng Laboratory [4]School of Computer Science and Engineering, Wuhan Institute of Technology
[5]Electronic Information School, Wuhan University
{yipeng, kuijiang, xin.tian}@whu.edu.cn, {wzy_hope, junjun0595}@163.com,
{lutxyl, jyma2010}@gmail.com

1. Experiments

1.1. HVSR with 4 LR frames

Our LOVSR actually utilizes 4 LR frames through the future hidden states H_{t+1}, and thus we have conducted an additional experiment of HVSR adopting 4 LR frames $(I_{t-1}, I_t, I_{t+1}, I_{t+2})$. As shown in Figure 1 and Table 1, HVSR^3 and HVSR^4 denote HVSR with 3 and 4 input LR frames respectively, where HVSR^4 surpasses HVSR^3 about 0.05 dB. Still, with the same 4 LR frames, our LOVSR outperforms HVSR^4 about 0.13 dB, and our GOVSR surpasses HVSR^4 about 0.37 dB with the assist of all LR frames in a video sequence. Note that these models have a similar number of parameters and calculation costs, which again proves the robustness of our OVSR framework.

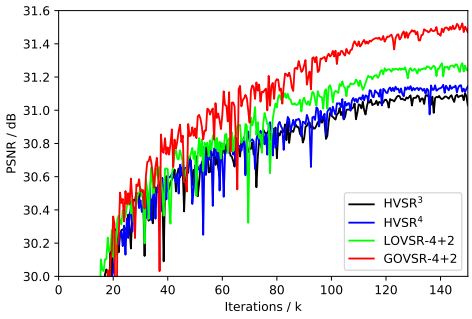


Figure 1: Training curves of different models.

1.2. Influences of Input Information

We further investigate the influence of each input information on the performance, as demonstrated in Table 2, we present the performances of different models by removing part of the input information.

Table 1: PSNR (dB) of different models.

Model	HVSR^3	HVSR^4	LOVSR-4+2	GOVSR-4+2
PSNR (dB)	31.10	31.15	31.28	31.52

Obviously, all models drop seriously without the center LR frame I_t, which accords with common sense because I_t contains the most basic and significant source information. Interestingly, IVSR, RVSR, and HVSR models all achieve 26.61 dB without I_t, and this is almost the same as the Bicubic magnified center frame I_t^{Bic} actually. Because generating the Bicubic magnified center frame I_t^{Bic} does not require learnable parameters, the IVSR, RVSR, and HVSR models all drop to I_t^{Bic} without I_t. Nevertheless, our OVSR models add the SR frames by Net_p and Net_s for reconstruction refinement (discussed in Equation (3) and Section 4.3 in the original paper), which requires the cooperation of two learnable networks and thus is more unstable than the Bicubic magnification, where they basically drop more than 10 dB by removing the center frame I_t.

IVSR drops nearly the same by removing I_{t-1} or I_{t+1}, which means I_{t-1} and I_{t+1} contribute almost equally to the result, and thus it is unwise for RVSR to overlook the subsequent frame I_{t+1}.

Excluding I_{t-1} or H_{t-1}, RVSR drops a lot, which proves that the hidden states can provide some beneficial information.

Moreover, HVSR drops more by removing I_{t+1} compared to removing I_{t-1}, and we reckon that this is due to the assist from H_{t-1}. This phenomenon further confirms that the hidden states contribute to the VSR indeed, and

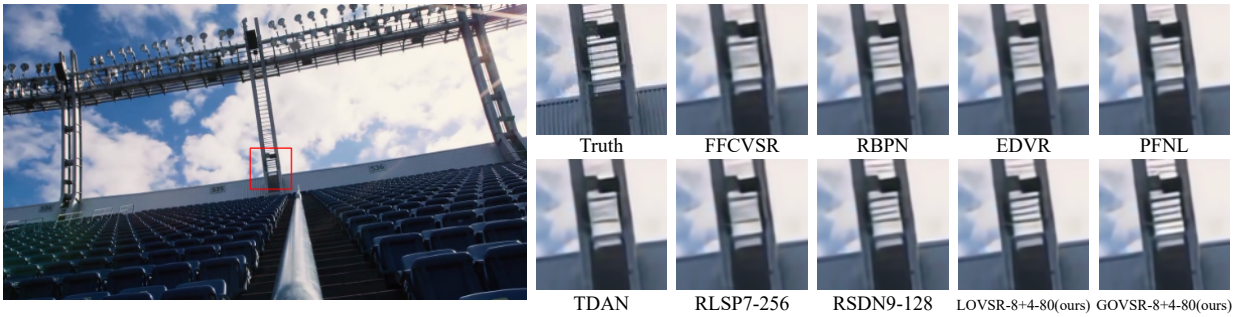
*Corresponding author. Code: https://github.com/psychopa4/OVSR.

Table 2: PSNR (dB) of different models, where 'Full' denotes the original model with all input information, and 'w/o' means without.

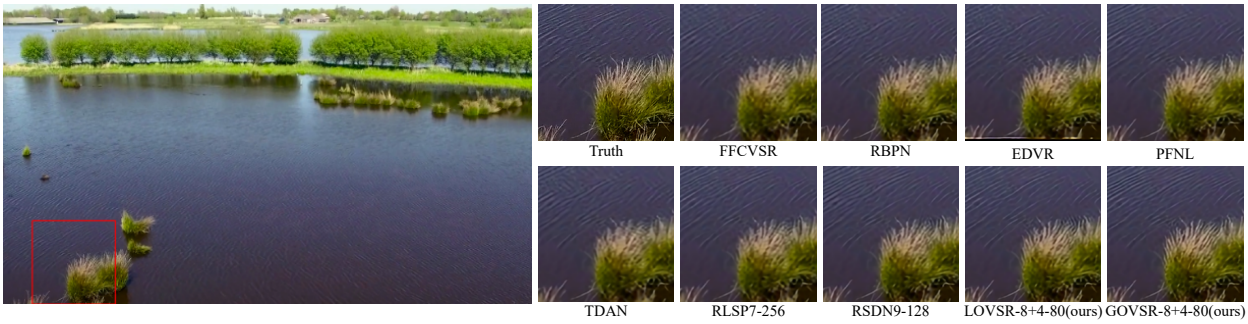
Model	IVSR	RVSR	HVSR	LOVSR-4+2			GOVSR-4+2		
Network	G	G	G	Net_p	Net_s	Both	Net_p	Net_s	Both
Full	30.66	30.60	31.10	31.28	31.28	31.28	31.52	31.52	31.52
w/o I_{t-1}	29.07 (-1.59)	28.95 (-1.65)	29.18 (-1.92)	29.27 (-2.01)	31.29 (+0.01)	29.27 (-2.01)	28.92 (-2.60)	31.52 (-0.00)	28.91 (-2.61)
w/o I_t	26.61 (-4.05)	26.61 (-3.99)	26.61 (-4.49)	21.18 (-10.10)	16.78 (-14.50)	8.75 (-22.53)	22.16 (-9.36)	13.70 (-17.82)	8.75 (-22.77)
w/o I_{t+1}	29.10 (-1.56)	-	28.91 (-2.19)	28.94 (-2.34)	30.66 (-0.62)	28.60 (-2.68)	28.98 (-2.54)	30.90 (-0.62)	28.87 (-2.65)
w/o H_{t-1}	-	29.35 (-1.25)	29.81 (-1.29)	29.74 (-1.54)	31.01 (-0.27)	29.53 (-1.75)	29.59 (-1.93)	30.85 (-0.67)	29.39 (-2.13)
w/o H_t	-	-	-	-	11.18 (-20.10)	-	-	12.44 (-19.08)	-
w/o H_{t+1}	-	-	-	-	30.52 (-0.76)	-	-	30.99 (-0.53)	-

Table 3: PSNR (dB) / SSIM of different video SR methods on Vimeo-90K testing dataset [5] by the upscaling factor of 4. Red and blue respectively indicate the best and second-best results. The * denotes the results reported in the original papers.

Methods	Vimeo-Slow	Vimeo-Medium	Vimeo-Fast	Vimeo-All
RBPN* [2]	34.18 / 0.9200	37.28 / 0.9470	40.03 / 0.9600	37.07 / 0.9435
EDVR* [4]	-	-	-	37.61 / 0.9489
FFCVSR [6]	33.59 / 0.9130	36.51 / 0.9416	38.68 / 0.9481	36.24 / 0.9367
RLSP7-256 [1]	33.86 / 0.9173	36.97 / 0.9463	39.20 / 0.9535	36.67 / 0.9415
RSDN9-128 [3]	33.91 / 0.9179	37.03 / 0.9466	39.41 / 0.9555	36.76 / 0.9421
LOVSR-8+4-80 (ours)	34.51 / 0.9256	37.84 / 0.9538	40.23 / 0.9615	37.53 / 0.9492
GOVSR-8+4-80 (ours)	34.60 / 0.9270	37.95 / 0.9548	40.32 / 0.9624	37.63 / 0.9503



(a) Frame 022 of auditorium from UDM10.

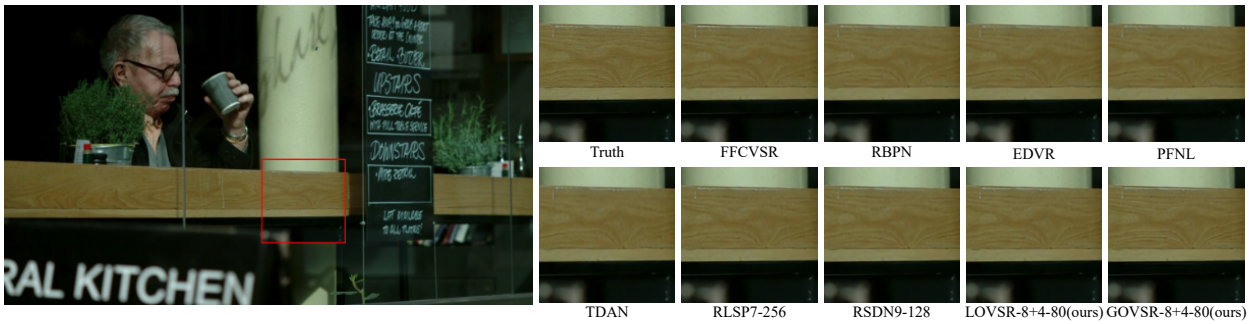


(b) Frame 015 of lake from UDM10.

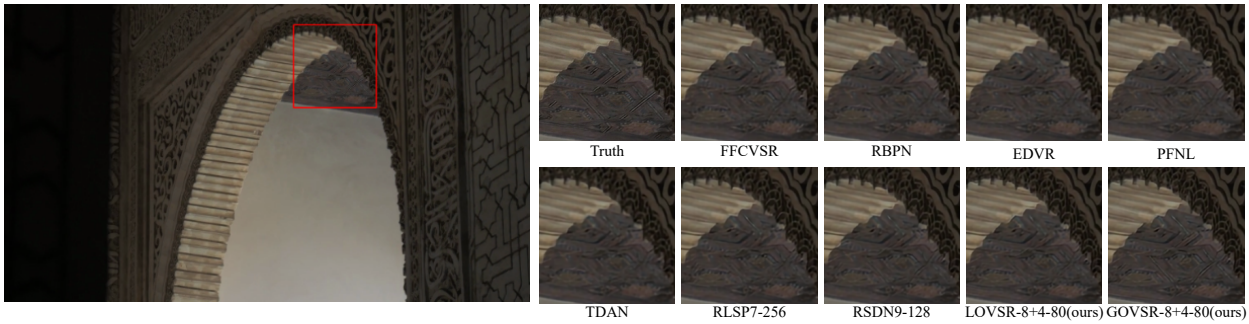
Figure 2: Visual comparisons of different methods.

thus it makes sense to further adopt the hidden states from the present and future to help VSR. Similar phenomena can also be observed in our models LOVSR and GOVSR.

Amazingly, by removing I_{t-1} in Net_s, our GOVSR s-



(a) Frame 015 of caffe from UDM10.



(b) Frame 015 of archwall from UDM10.

Figure 3: Visual comparisons of different methods.

tays the same but LOVSR increases 0.01 dB in PSNR instead, and we owe it to the OVSR framework, which leverages hidden states from the past, present and future to help VSR, through which it basically does not need I_{t-1} in Net_s anymore. Last but not least, compared to removing the LR frames I, our models LOVSR and GOVSR deteriorate more seriously or comparably by removing the corresponding hidden states H in Net_s, which again confirms that our models indeed make good use of the hidden states from the past, present and future.

2. Result on Vimeo-90K Dataset

We further conduct experiments on another public training dataset Vimeo-90K [5], and test the models on its testing dataset Vimeo-90K-T. According to the average motion flow magnitude, Vimeo-90K-T is divided into 3 categories: slow, medium, and fast [2], where there are 1616, 4983, and 1225 sequences in each category. Albeit Vimeo-90K contains tens of thousands of video sequences, each of which consists of only 7 frames, and the HR frames are at the fixed resolution 448×256, which is quite low. Besides, during testing, other methods only calculate the PSNR/SSIM of the center SR frame, and consequently, we do not think this dataset is suitable for VSR. Still, we retrain as many methods as we can on Vimeo-90K in a limited time for a more comprehensive comparison, under the same training settings. The PSNR and SSIM values are calculated only on the luminance channel of YCbCr colorspace, focusing on the center frame and eliminating 8 pixels on four borders.

As shown in Table 3, our model GOVSR-8+4-80 still achieves the best performance. It is worth mentioning that compared to GOVSR-8+4-80, LOVSR-8+4-80 can not utilize all 7 frames to rebuild the center frame, and thus it behaves a little worse.

3. Visual Comparisons

We show more visual comparisons, as illustrated in Figure 2 and Figure 3, most methods can only recover the low-frequency contour of the objects, which seems smooth and blurry. Our models can recover the right textures with more realistic details.

References

[1] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485, 2019.

[2] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[3] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *European Conference on Computer Vision (ECCV)*, pages 645–660, 2020.

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

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

[6] Bo Yan, Chuming Lin, and Weimin Tan. Frame and feature-context video super-resolution. In *AAAI Conference on Artificial Intelligence*, pages 5597–5604, 2019.