Supplementary: Deep Space-Time Video Upsampling Networks

Jaeyeon Kang[1 [0000−0003−2172−0026]], Younghyun Jo[1 [0000−0002−8530−9802]], Seoung Wug Oh[1 [0000−0002−8498−0864]], Peter Vajda[2 [0000−0002−2031−4678]], and Seon Joo Kim[1,2 [0000−0001−8512−216X]]

[1]Yonsei University, [2]Facebook

1 Comparisons on VSR

Table 1 shows the quantitative comparison with other state-of-the-art VSR methods. We extract only the VSR part of our network and generate the results. Our results show higher performance than other methods on Vid4 testset. It shows that STVUN can serve as a good VSR method as well, although our method is trained to perform the joint space-time upsampling.

Table 1: Comparison of PSNR and SSIM on the Vid4 testset for the scaling factor $r = 4$. We only use our space upsampled frames to compare with other VSR methods. '*' means the values are taken from their papers. Except for '*', we compute evaluation metrics except for border frames. The best is shown in bold and the second is shown in underline.

Dataset	Bicubic	FRVSR*[3]	DUF[2]	RBPN*[1]	EDVR[4]	STVUN (Ours)
Calendar	20.39/0.5720	-	24.10/0.8132	23.99/0.807	24.07/0.8151	**24.59/0.8355**
City	25.06/0.6028	-	28.36/0.8344	27.73/0.803	27.99/0.8108	**28.57/0.8434**
Foliage	23.47/0.5666	-	26.41/0.7712	26.22/0.757	26.32/0.7626	**26.49**/0.7780
Walk	26.10/0.7974	-	30.62/0.9142	30.70/0.909	**31.01**/0.9144	30.95/**0.9169**
Avearage	23.78/0.6347	26.69/0.822	27.37/0.8332	27.12/0.8180	27.35/0.8258	**27.65/0.8434**

More comparison on various datasets is shown in Table 2. Vimeo-90K testset (Vimeo-T) and REDS validation (REDS-V) dataset are additionally used. As we observe dataset bias, we additionally train only VSR part on different training dataset. Ours-V and Ours-R are trained on Vimeo-90K, and REDS training dataset for each. Ours-Y is the VSR part extracted from our final version of STVUN which is trained on our YouTube 240fps dataset.

In our results, ours has higher value than other methods on Vid4, but on other testsets, we have lower value than EDVR. On the other hand, our method outperforms DUF and RBPN by a large margin on all testsets. Moreover, our method is at least 14 times faster than all other methods. Note that, although our final goal is to design space-time upsampling network, our VSR network is still competitive to the previous methods in terms of the performance and computation time.

Table 2: Comparison of average PSNR and SSIM on the various testsets. The best is shown in bold and the second is shown in underline. Ours-Y, Ours-V and Ours-R are our network trained on our Youtube 240fps, Vimeo-90K, REDS dataset respectively. The running time is measured when generating the results with the resolution 960×540.

Dataset	RBPN[1]	DUF[2]	EDVR[4]	Ours-Y	Ours-V	Ours-R
Vid4	27.12/0.8180	27.37/0.8332	27.35/0.8258	27.65/0.8434	**27.70/0.8446**	27.20/0.8310
Vimeo-T	37.07/0.9435	36.37/0.9387	**37.59/0.9487**	36.99/0.9361	37.31/0.9391	34.30/0.9373
STVT	31.12/0.9027	34.06/0.9327	**35.08/0.9425**	34.64/0.9393	34.71/0.9400	34.30/0.9373
REDS-V	30.45/0.8507	29.91/0.8254	**32.26/0.8797**	30.27/0.8367	30.35/0.8393	31.35/0.8648
#Params	12.7M	**5.8M**	20.6M	21.53M	21.53M	21.53M
Time	3.01s	0.61s	0.29s	**0.02s**	**0.02s**	**0.02s**

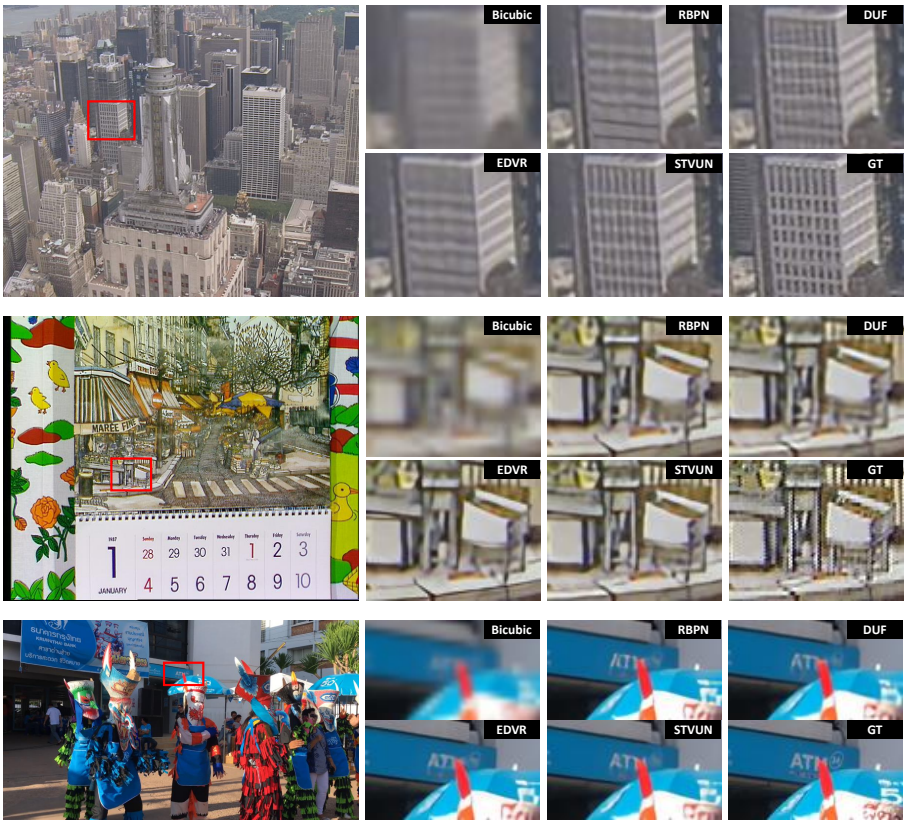


Fig. 1: Qualitative comparison of our method (Ours-Y) with other VSR methods. We only compare space upsampled frames in this figure. From top to bottom, *City*, *Calendar* in Vid4 and *Festival_1* in STVT are used. Zoom in to see better visualization.

Fig. 1 shows the visual comparison of our method with other algorithms. Our model can more improve the image details which are close to the detail of GT and shows more visually pleasing outputs.

2 Ablation studies on VSR

We additionally analyze how EFST affects the performance on our VSR network in Table 3. Like space-time upsampling ablation studies, we test our model on VSR without EFST (w/o EFST) and with alignment and fusion (w/ A&F). We use PCD and TSA modules from EDVR [4] for the alignment and fusion. The results show that using EFST improves the performance without the large difference in running time.

Table 3: Ablation studies on the EFST in VSR network. The whole models are trained on our YouTube 240fps training dataset. Vid4 testset is used for comparison. The running time is measured with the same environment in Table 2.

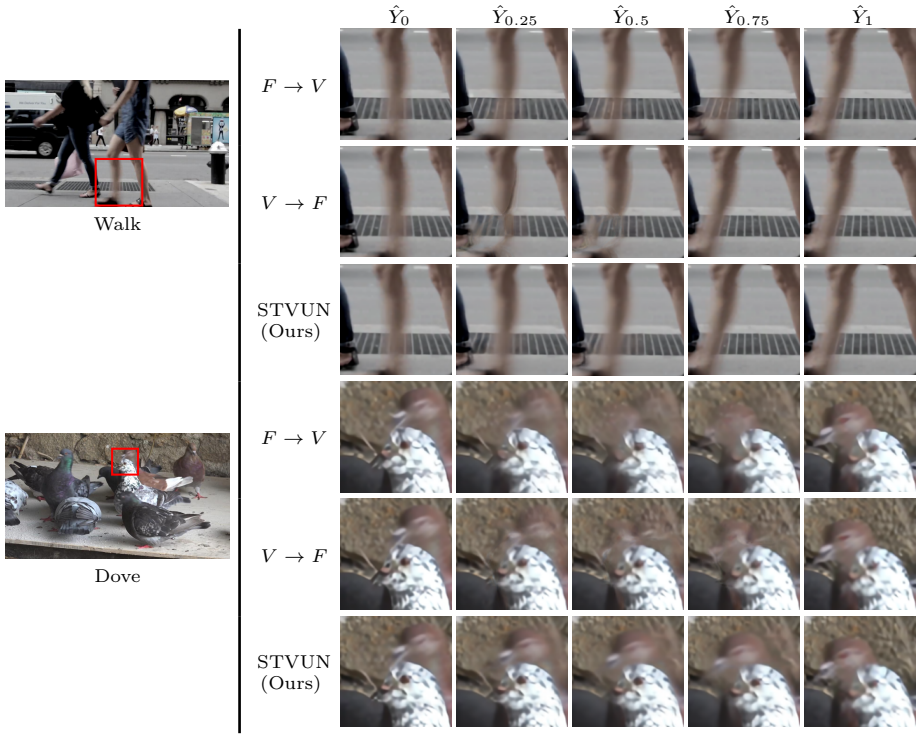
	w/o EFST	w/ A&F	Ours
PSNR/SSIM	27.45/0.8370	27.50/0.8387	**27.65/0.8434**
#Params	**21.19M**	23.03M	21.53M
Running Time	**0.016s**	0.40s	0.021s

3 More visual results on STVUN

We additionally generate more visual results in Fig. 2. It visually compares our method with the other methods. We recommend watching our demo video in the supplementary material.

References

1. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3897–3906 (2019) 1, 2
2. Jo, Y., Wug Oh, S., Kang, J., Joo Kim, S.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3224–3232 (2018) 1, 2
3. Sajjadi, M.S., Vemulapalli, R., Brown, M.: Frame-recurrent video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6626–6634 (2018) 1
4. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 1, 2, 3



(a) Comparison with baseline methods on STVT.

(b) Comparison with FISR on STVT, Sintel and Vid4.

Fig. 2: Visual comparisons of the space-time upsampling results. In (a), we generate a total of 5 frames that consist of 2 space upsampled and 3 intermediate frames. In (b), we generate one intermediate frame. Zoom in to see better visualization.