MuCAN: Multi-Correspondence Aggregation Network for Video Super-Resolution (Supplementary Material)

Wenbo Li[1], Xin Tao[2], Taian Guo[3], Lu Qi[1], Jiangbo Lu[4], and Jiaya Jia[1,4]

[1]The Chinese University of Hong Kong
[2]Kuaishou Technology [3]Tsinghua University [4]Smartmore Technology
{wenboli,luqi,leojia}@cse.cuhk.edu.hk jiangsutx@gmail.com
gta17@mails.tsinghua.edu.cn jiangbo@smartmore.com

1 More Ablation Study Results

Under the same network and training setting as ablation study, we further explore how the maximum displacements and patch size affect the performance of TM-CAM. By adjusting the displacements from $\{3, 5, 7\}$ to $\{3, 3, 3\}$, the PSNR/SSIM drops from 30.31dB/0.8648 to 29.98dB/0.8576, which demonstrates the importance of keeping large enough searching regions. Besides, models with patch size 1×1 and 5×5 obtain 30.15dB/0.8612 and 30.31dB/0.8653, respectively. Compared with default patch size 3×3 yielding 30.31dB/0.8648, although patch size 5×5 results in small improvement on SSIM, it needs much more parameters and nearly double FLOPS. Considering all these facts, we set the default displacements to $\{3, 5, 7\}$ and patch size to 3×3.

2 Results on the Vid4 Dataset

Table 1. Comparisons of results on the Y channel on the Vid4 dataset for $\times 4$ setting.

Method	Frames	PSNR(dB) / SSIM
Bicubic	1	23.78 / 0.6347
RCAN [9]	1	25.46 / 0.7395
VESPCN [1]	3	25.35 / 0.7557
SPMC [6]	3	25.88 / 0.7752
TOFlow [8]	7	25.89 / 0.7651
FRVSR [5]	recurrent	26.69 / 0.822
RBPN [2]	7	27.12 / 0.818
EDVR [7]	7	27.35 / 0.8264
MuCAN(Ours)	7	27.26 / 0.8215

Since Vid4 only contains 4 testing sequences without the training ones, most existing methods train their models with different training datasets and then

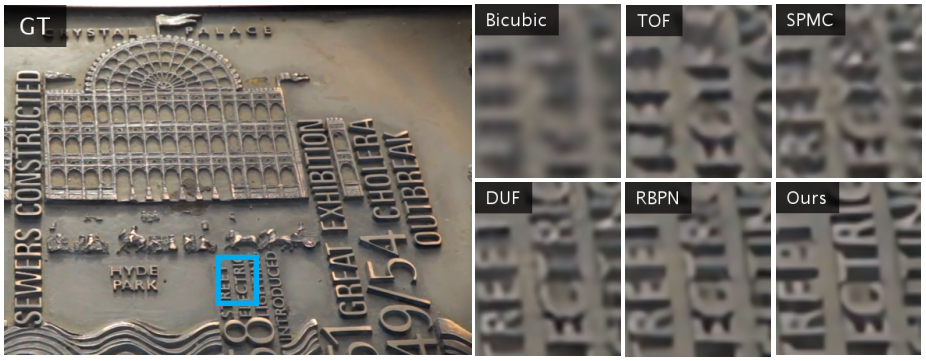

Fig. 1. Examples of the clip "LDVTG_009" on the SPMCS dataset. We compare our method with TOF [8], SPMC [6], DUF [4] and RBPN [3].

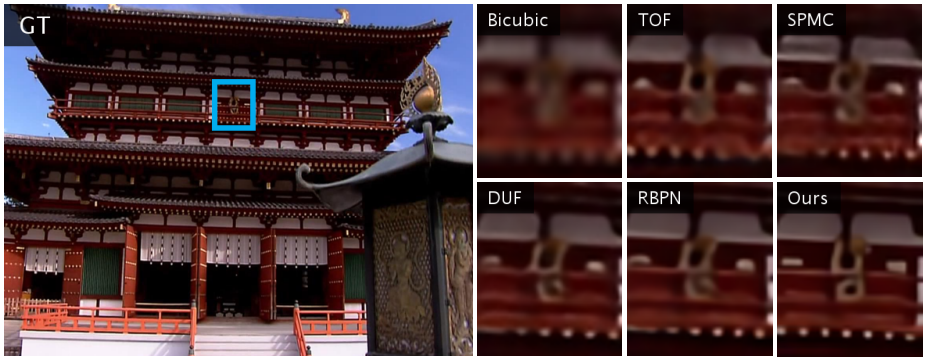

Fig. 2. Examples of the clip "jvc_009_001" on the SPMCS dataset.

test on Vid4. As shown in Table 1,our model trained on Vimeo-90K obtains PSNR/SSIM of 27.26dB/0.8215, which is slightly below (0.09dB) the SOTA result by EDVR. Compared with EDVR, our model is trained with only half of iterations.

3 Additional Qualitative Results on the SPMCS Dataset

Here, we compare our proposed MuCAN model trained on the Vimeo-90K dataset [8] with other state-of-the-art methods [8, 6, 4, 3] on the SPMCS validation dataset [6]. Some visual examples are shown in Figure 1, 2, 3, 4. Our MuCAN method is good at recovering more frame details compared with other methods. In addition, our MuCAN method shows robustness on various scenes and motion categories, and introduces less unpleasant visual artifacts.

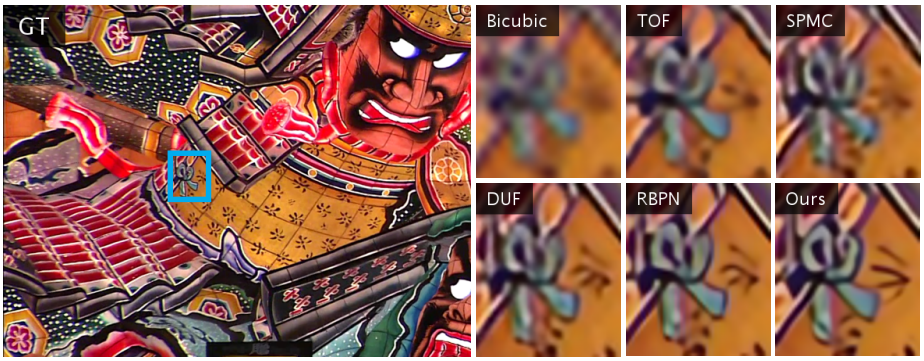

Fig. 3. Examples of the clip "hitachi_isee5_001" on the SPMCS dataset.

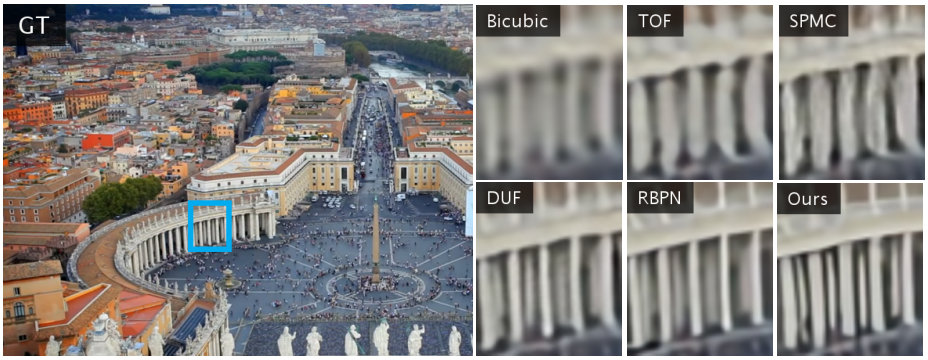

Fig. 4. Examples of the clip "RMVTG_024" on the SPMCS dataset.

Specifically, as shown in Figure 1, our model generates much clearer English letters. In Figure 2, details on the hangings and roofs are better recovered and few blur artifacts are generated by our MuCAN. From Figure 3, it is clear that our MuCAN recovers sharper and more accurate edges on the petal area while other methods fail. RBPN [3] also recovers some details but produces wrong edge directions on the petals. In Figure 4, sharper pillar details are recovered by our MuCAN while others either fail to restore the inner pillars (e.g., DUF [4] and RBPN [3], or generate distortions with artifacts (e.g., TOF [8] and SPMC [6]).

4 Generalization Analysis

To evaluate the generality of our method, we apply our model trained on the REDS dataset to test video frames in the wild. In addition, we test EDVR

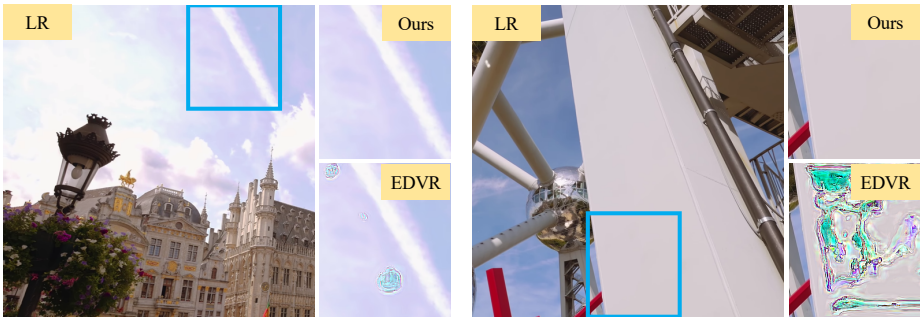

Fig. 5. Visualization of video frames in the wild for EDVR [7] and our MuCAN.

with the author-released model[1] on the REDS dataset. Some visual results are shown in Figure 5. We remark that EDVR may generate visual artifacts in some cases due to the variance of data distributions between training and testing. In contrast, our MuCAN shows its decent generality in the real world setting.

5 Video SR Demos of Our Proposed MuCAN Method

We also show 4 demo videos generated by our MuCAN, which are named "car.mp4", "hitachi.mp4", "jvc.mp4" and "veni.mp4". It can be seen that our MuCAN generates visually coherent HR videos with distinct details and rare artifacts, which proves the superiority and practicability of our proposed method.

[1] https://github.com/xinntao/EDVR

References

1. Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., Shi, W.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4778–4787 (2017)
2. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1664–1673 (2018)
3. 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)
4. 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)
5. 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)
6. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4472–4480 (2017)
7. 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)
8. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision **127**(8), 1106–1125 (2019)
9. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 286–301 (2018)