Temporal Kernel Consistency for Blind Video Super-Resolution: Supplementary Material

Lichuan Xiang[1*], Royson Lee[2*], Mohamed S. Abdelfattah[3],
Nicholas D. Lane[2,3], Hongkai Wen[1,3]
[1]University of Warwick [2]University of Cambridge [3]Samsung AI Center, Cambridge
l.xiang.2@warwick.ac.uk

1. Implementation Details of KernelGAN & ZSSR.

We used the default settings and hyperparameters provided by KernelGAN [1] and ZSSR [5]. For KernelGAN, the estimated downscaling kernel size is set to 13×13 and the input image is cropped to 64×64 before kernel extraction. The kernel is extracted after 3000 iterations using the Adam [3] optimizer with learning rate set to 0.0002, β_1 set to 0.5 and β_2 set to 0.999. For ZSSR, the input LR image and the provided estimated kernel is used to generate a downsampled variant of the LR image. The resulting image pair is then used to train the model using the Adam optimizer starting with learning rate set to 0.001 with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For more details, please refer to the provided repository[1].

2. Architecture of DAN & EDVR

A more detail architecture of our model experimented in Sec. 5 of the main paper is shown in Fig. 1. Notably, the features of the HR frames are concatenated with the LR features in each CRB block of DAN [4] and we utilized the existing channel attention layer (CALayer) for temporal kernel estimation. During the last iteration, the LR features, which were conditioned on the input frames and their estimated kernel, were fed into the temporal blocks of EDVR [6] for temporal alignment, fusion, and restoration. In particular, the PCD module follows a pyramid cascading structure, which concatenates features of differing spatial sizes and uses deformable convolution at each respective pyramid level to the aligned features. The TSA module then fused these aligned features together through both temporal and spatial attention. Specifically, temporal attention maps are computed based on the aligned features and applied to these features through the dot product before concatenating and fusing them using a convolution layer. After which, the fused features are then used to compute the spatial attention maps which are then applied to these features. For more details, please refer to EDVR [6].

3. Implementation Details of DAN & EDVR.

For training, we used scaling factor $\times 4$, input patch size of 100×100, and set $N = 1$, i.e. considering sequences of 3. We set $N = 2$ to highlight the kernel mismatch on motion compensation as shown in Fig. 4. The batch size was set to 4, and all models were trained for 300 epochs, using the Adam optimizer [3] ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). The initial learning rate was set to 1×10^{-4}, and decayed with a factor of 0.5 at every 200 epochs. Following DAN [4], we ran for 4 iterations ($J = 4$) and used L1 loss for both kernel estimation and video restoration across all our models in every iteration. When multiple frames were utilized for temporal alignment, we applied a scaling factor of $1/2N$ to weight the loss from supporting frames. Following previous works [6, 8], PSNR and SSIM [7] were computed after converting each frame from RGB to Y channel and trimming the edges by the scale factor. All experiments were run on NVIDIA 1080Ti and 2080Ti GPUs. Temporal kernel consistency of our test benchmark, *REDS10*, is shown in Fig. 2. Similar to Fig. 1 in the main paper, we quantified kernel temporal change by measuring the sum of absolute difference between consecutive kernel PCAs. In particular, video sequences such as 016 and 018 have high temporal kernel consistency and sequences such as 000 and 014 have low temporal kernel consistency.

4. Additional Results

We provided additional results here due to space limitations in the main paper. Fig. 3 provides additional examples for Fig. 3 of the main paper, highlighting that using a fixed kernel to upscale all the frames in a video can result in inferior restoration outcomes as compared to using a per-frame kernel even without incorporating temporal frame information. Likewise, Fig. 4 shows the additional examples for

*Equal contributions.
[1]https://github.com/sefibk/KernelGAN

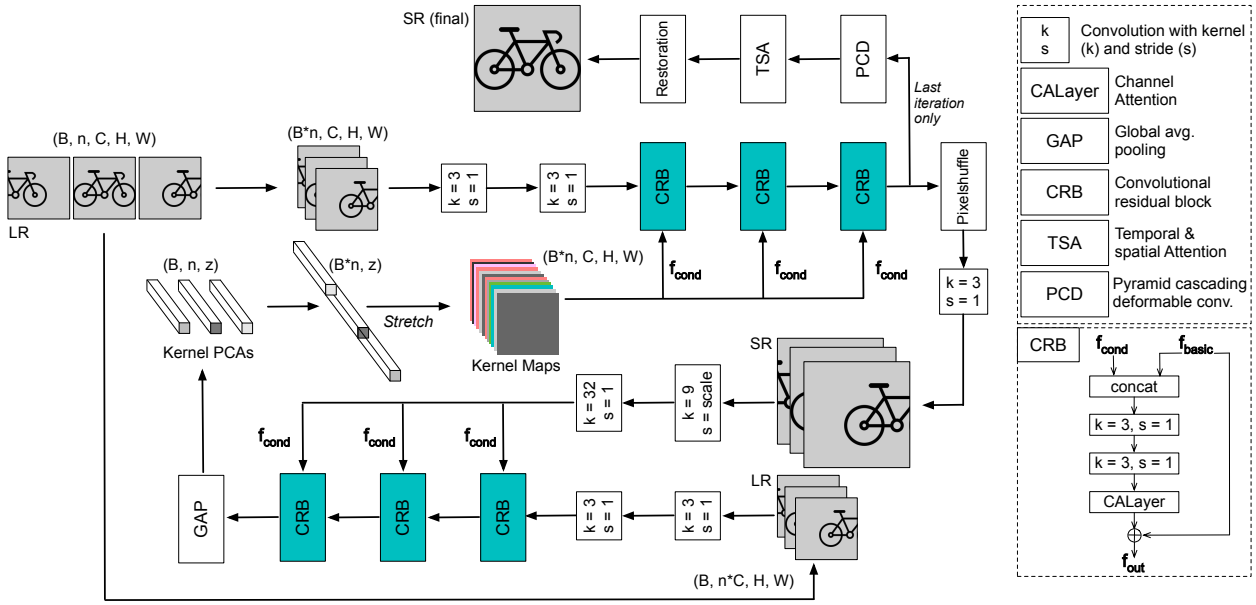

Figure 1: Detailed architecture of our model experimented in Sec. 5 of the main paper.

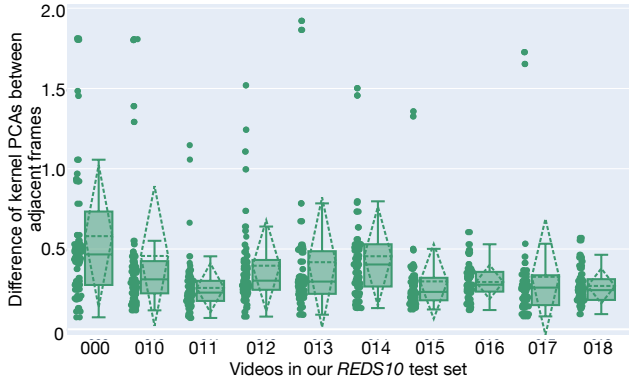

Figure 2: Temporal kernel consistency of videos in our *REDS10* benchmark, measured by kernel PCA changes for adjacent frames in the videos. Kernel changes are represented by solid dots while boxplots show distributions.

Fig. 4 of the main paper, underlining that the use of explicit motion compensation in previous works for temporal alignment results in more errors when applied to real-world videos. Fig. 5 and Fig. 6 provides additional qualitative examples comparing our multi-frame SR model with previous multi-frame SR and blind image-based SR models on *REDS10* and real-world videos respectively. Lastly, we provided a sample video along with this document.

References

[1] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In *Advances in Neural Information Processing Systems*. 2019. 1, 3

[2] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The Something Something Video Database for Learning and Evaluating Visual Common Sense. In *ICCV*, 2017. 3

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] Zhengxiong Luo, Y. Huang, Shang Li, Liang Wang, and Tieniu Tan. Unfolding the alternating optimization for blind super resolution. In *Advances in Neural Information Processing Systems*. 2020. 1

[5] Assaf Shocher, N. Cohen, and M. Irani. "zero-shot" super-resolution using deep internal learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3

[6] Xintao Wang, Kelvin C. K. Chan, K. Yu, C. Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 1

[7] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 1

[8] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *European Conference on Computer Vision (ECCV)*, 2018. 1

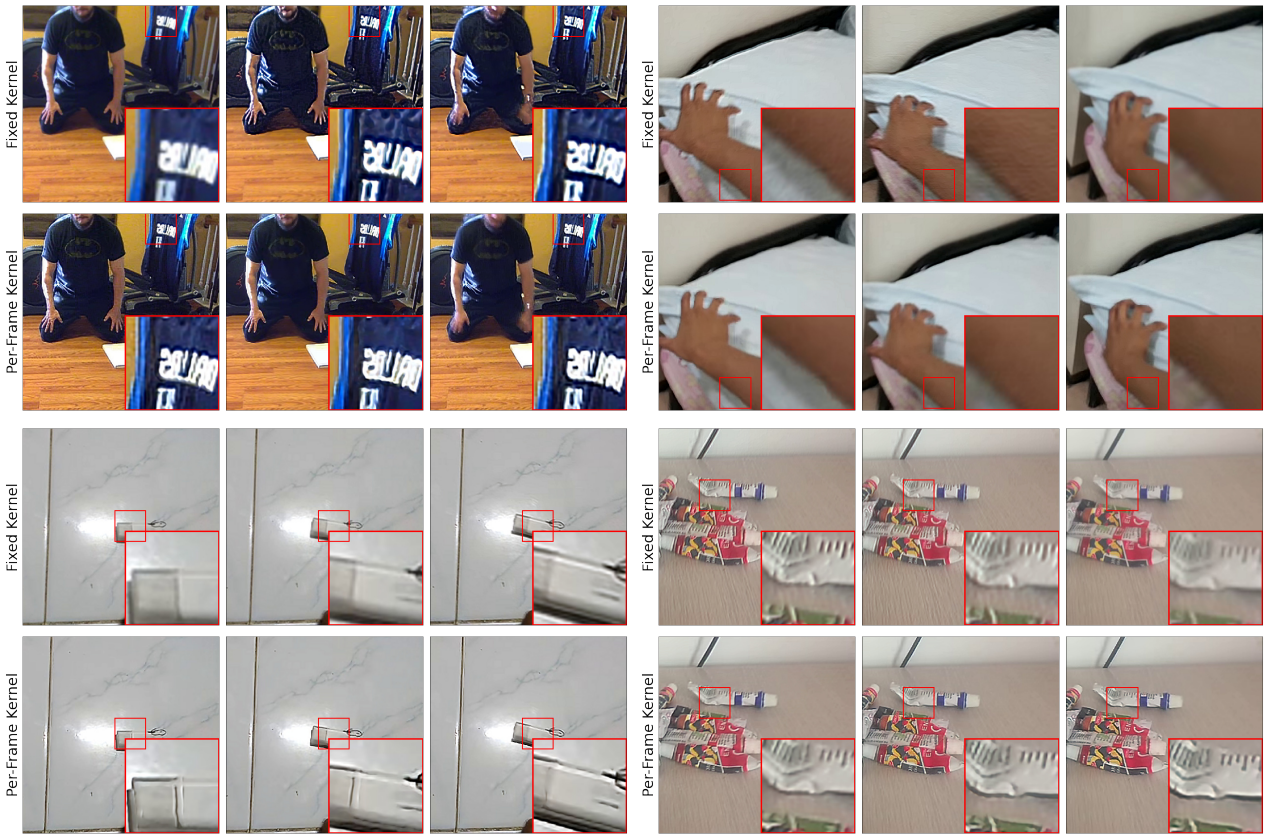

Figure 3: Additional Examples of consecutive frames in real-world videos taken from Something-Something [2] dataset upscaled using a fixed kernel (top in each example), and a different per-frame kernel (bottom in each example). Kernels are estimated using KernelGAN [1] and the frames are restored using ZSSR [5].

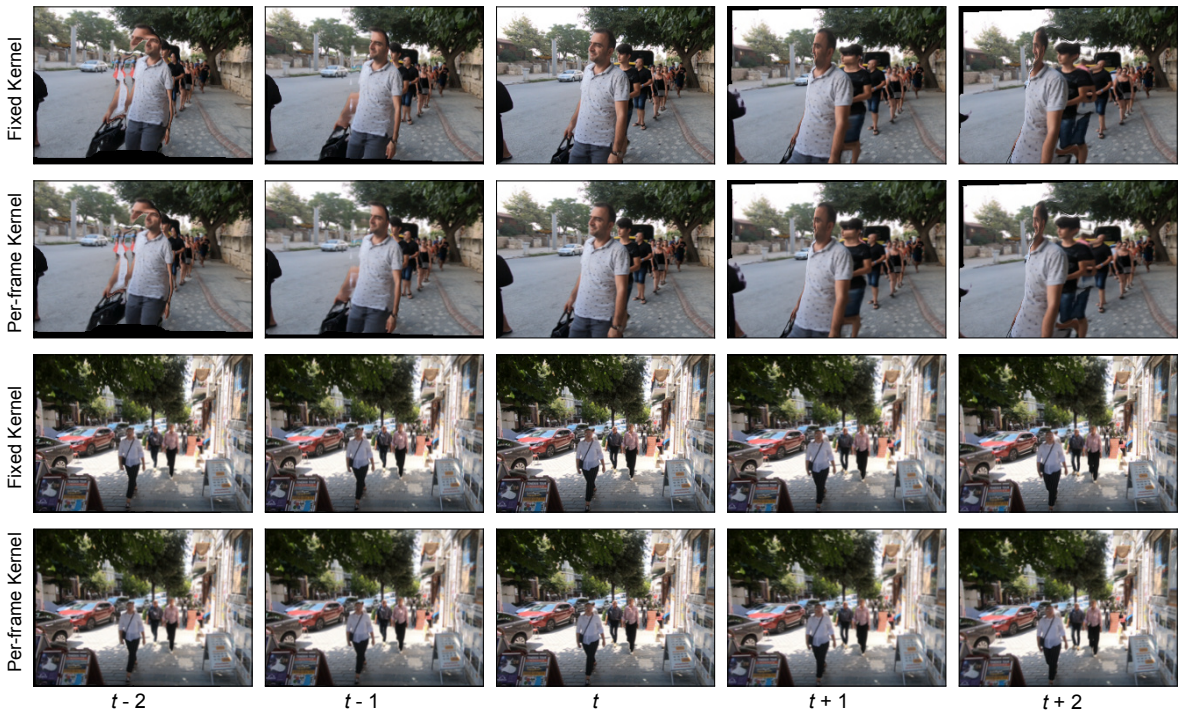

Figure 4: Additional example aligned frames at time step $t-2, t-1, t+1, t+2$ with their reference frame at time step t. The aligned frames are oversmoothed and blurred due to kernel mismatch for per-frame kernels found in real-world videos. In comparison, using a fixed downsampling kernel at every time step, which does not hold for real-world videos, leads to better motion compensation. Zoom in for best results.

Figure 5: Qualitative comparison among existing models, along with bicubic upscaling, on our benchmark test sequences. Zoom in for best results.

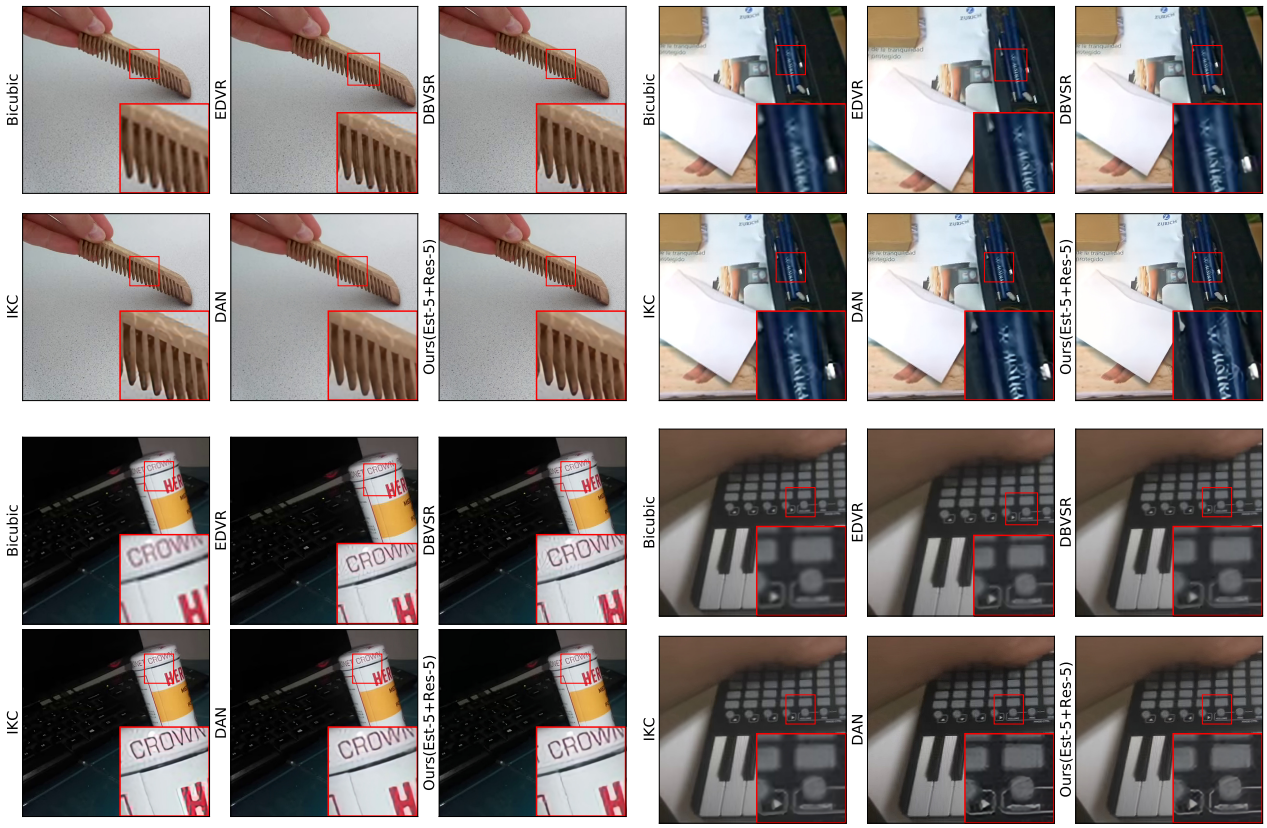

Figure 6: Real-world qualitative comparisons among existing models, along with bicubic upscaling. Zoom in for best results. Note that there is no ground-truth available.