Supplementary Material

1. Additional Visual Comparison

In the main paper, we show that our method can generate superior super-resolution results for parkour videos captured by egocentric sports cameras. In addition to the parkour videos, we collect real-world video clips (See Fig. 3) from commonly seen categories: animation, movie, music video (MV) and vlog. The goal of this experiment is to test the **robustness** and **generalization** of our method in various scenarios. Note that the appearances of these videos are significantly different from the videos in our training set (Vimeo90K [7]). In this supplementary material, the results show that our method generalizes better than comparison methods EDVR [6], TOFlow [7] and TGA [1] which are also trained on Vimeo90K [7]. We now discuss the examples shown in Fig. 3.

1) Animation. Animation is challenging to frame alignment for its lack of textures and low frame rate. We show a 2D animation example (a) and 3D animation examples (b) and (c) in Fig. 3. In example (a), it is difficult to recover the facial details by simply fusing neighbor frames since the information is completely corrupted in the low-resolution frames (see bicubic and comparison results). Our method can recover the facial details thanks to the memory-augmented attention. In example (b), aligning the strings on a moving guitar is difficult for EDVR [6], TOFlow [7], DBVSR [3] and TGA [1]. Our cross-frame non-local attention can effectively avoid the artifacts caused by the misaligned frames. This mechanism also works better for still repetitive pattern regions like example (c), which are often recognized as moving patterns and shifted (EDVR, TOFlow, DBVSR and TGA). Image super-resolution method CSNLN [2] also fails due to the erroneous non-local attention.

2) Movie. Super-resolving movies are of interest in online video streaming services. Movies are also challenging for their large motion and low illumination. Our method can generate very small scale sharp details like the star on the shield (examples (d)), wrinkles on the face (example (e)) and eagle eye/feathers (example (f)), which are difficult to reconstruct using either frame alignment (EDVR, TOFlow, DBVSR and TGA) and regular non-local attention (PFNL and CSNLN).

3) MV. Music video (MV) is one of the most-watched video categories online. Music video usually focuses on close-up shots of dancing humans, making the reconstruction of details on clothes important. In examples (g) and (h), our cross-frame non-local attention module enables recovering the fine details under the presence of motion blur. The comparison non-local attention based methods PFNL [8] and CSNLN [2] cannot achieve results comparable to ours since in the regular non-local attention, equally treating the pixels in the entire video frame has a negative effect on the overall performance.

4) Vlog. Vlog is another type of daily video captured by hand-held devices. This category also covers video chat, which is also common in daily life. Due to the instability of hand-held devices, these videos are extremely shaky and difficult to super-resolve. Existing video super-resolution methods TOFlow [7], TGA [1] and PFNL [8] fail in example (j) and (k). EDVR [6] can recover some details in examples (i), (j) and (k), but their results are blurry in general due to the inaccurate frame alignment.

	PSNR↑	SSIM↑	LPIPS↓		PSNR↑	SSIM↑	LPIPS↓
Bicubic	31.60	0.9126	0.2135	DBVSR	34.41	0.9472	0.0993
MANA	**37.44**	**0.9626**	**0.0681**	TGA	37.12	0.9601	0.0707
EDVR	34.48	0.9456	0.1150	PFNL	37.04	0.9673	0.0819
TOFlow	35.58	0.9531	0.0968	CSNLN	36.09	0.9545	0.0844

Table 1. Quantitative comparison on the videos shown in Fig. 3. Larger numbers indicate better results for PSNR and SSIM, smaller numbers indicate better results for LPIPS.

	PSNR↑	SSIM↑	LPIPS↓		PSNR↑	SSIM↑	LPIPS↓
Bicubic	22.34	0.6131	0.5186	DBVSR	24.64	0.7547	0.3096
MANA	**25.22**	**0.7816**	**0.2842**	TGA	25.36	0.7949	0.2834
EDVR	25.79	0.8063	0.2489	PFNL	25.01	0.7788	0.3204
TOFlow	24.41	0.7435	0.3340	CSNLN	24.09	0.7202	0.3425

Table 2. Quantitative comparison on Vid4 [4], which consists of only 4 test videos. Larger numbers indicate better results for PSNR and SSIM, smaller numbers indicate better results for LPIPS.

2. Additional Quantitative Comparison

We show the average quantitative values for example videos in Fig. 3 in Table 1. We also provide quantitative result on Vid4 [4] dataset, consisting of 4 videos only, in Table 2. Although this small dataset is less diverse and representative, our method still achieves comparable results. More comprehensive comparisons in Table 1 and the main paper have shown that our method works consistently better than the existing state-of-the-art methods in various categories of real-world videos outside the domain of the Vimeo90K training set. This proves the robustness of our method, which is important for real applications.

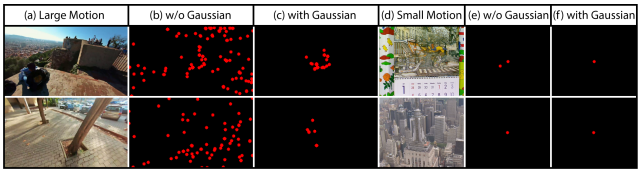

Figure 1. Visualization of correlation map between the center pixel and the frame 3 time steps away.

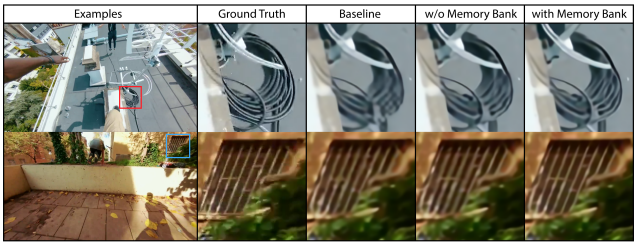

Figure 2. Visual effect of the memory bank.

Structure	PSNR↑	SSIM↑	LPIPS↓
Baseline (Encoder+Decoder)	32.42	0.9209	0.1393
+Conventional NLA	33.31	0.9338	0.1231
+Gaussian NLA	33.57	0.9367	0.1208
+Memory Module (Complete Model)	33.81	0.9397	0.1159
Baseline+Optical Flow	32.49	0.9230	0.1398

Table 3. Additional ablation experiment.

3. Additional Ablation Study

A complete ablation experiment demonstrating the effectiveness of each network module is shown in Table 3. Starting from only encoder and decoder (referred to as baseline), we gradually add conventional non-local attention (NLA), Gaussian weighted NLA and memory module and evaluate the performance on the large motion Parkour dataset. To show the necessity of NLA in the large motion video super-resolution, we also include a version by replacing the NLA with optical flow and evaluate its performance.

Effectiveness of Gaussian Non-Local Attention To make our Gaussian weighted Non-Local Attention approach more intuitive, we visualize the effects of Gaussian in Fig. 1. The red dots are the pixels having a correlation value greater than 10% of the maximum value to the center query pixel. Without Gaussian (Fig. 1(b)), the large motions may cause bad correspondences in conventional NLA that distributed across the entire frame. The distant correspondence is less reliable in general, so our model learns a single Gaussian centered at the query pixel to re-scale the correlation map. Note that the re-scaling won't completely zero out all distant matches. The intuition of Gaussian weighting is that it can filter out the erroneous correspondences with small correlations, but keep those truly helping the reconstruction of the query pixel. For small motion (Fig. 1(d)), Gaussian weighting has an insignificant effect on correlation. The standard deviation of the Gaussian ($\sigma = 12.1804$) is learned as a parameter so that the overall performance is optimized.

We also show additional ablation studies on with and without Gaussian in Table 3. The performance of Gaussian NLA has a 0.26dB PSNR gain compared to conventional NLA in large motion Parkour videos. This is because the learned Gaussian re-scales the correlation according to the distance to the query pixel, and effectively filters out the erroneous correspondences. This argument is further supported by the last row of Table 3, where we replace the NLA with optical flow (RAFT [5]). In the large motion videos, due to the difficulty of finding accurate correspondence among neighbor frames, its performance is inferior to both the conventional NLA and Gaussian NLA, which contribute 0.89dB and 0.26dB to the PSNR gain respectively.

Effectiveness of Memory Module Qualitative result on our model with and without memory module is shown in Fig. 2. Memory module improves the overall sharpness of local details in the results, e.g. the cable in the first example and the grid in the second example. Also note that the memory module contributes 0.24dB more PSNR gain compared to solely using Gaussian NLA, as shown in Table 3.

References

[1] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In *CVPR*, 2020. 1

[2] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S. Huang, and Humphrey Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR*, 2020. 1

[3] Jinshan Pan, Haoran Bai, Jiangxin Dong, Jiawei Zhang, and Jinhui Tang. Deep blind video super-resolution. In *ICCV*, 2021. 1

[4] Mehdi Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, 2018. 1

[5] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 2

[6] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPR*, 2019. 1

[7] 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. 1

[8] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, 2019. 1

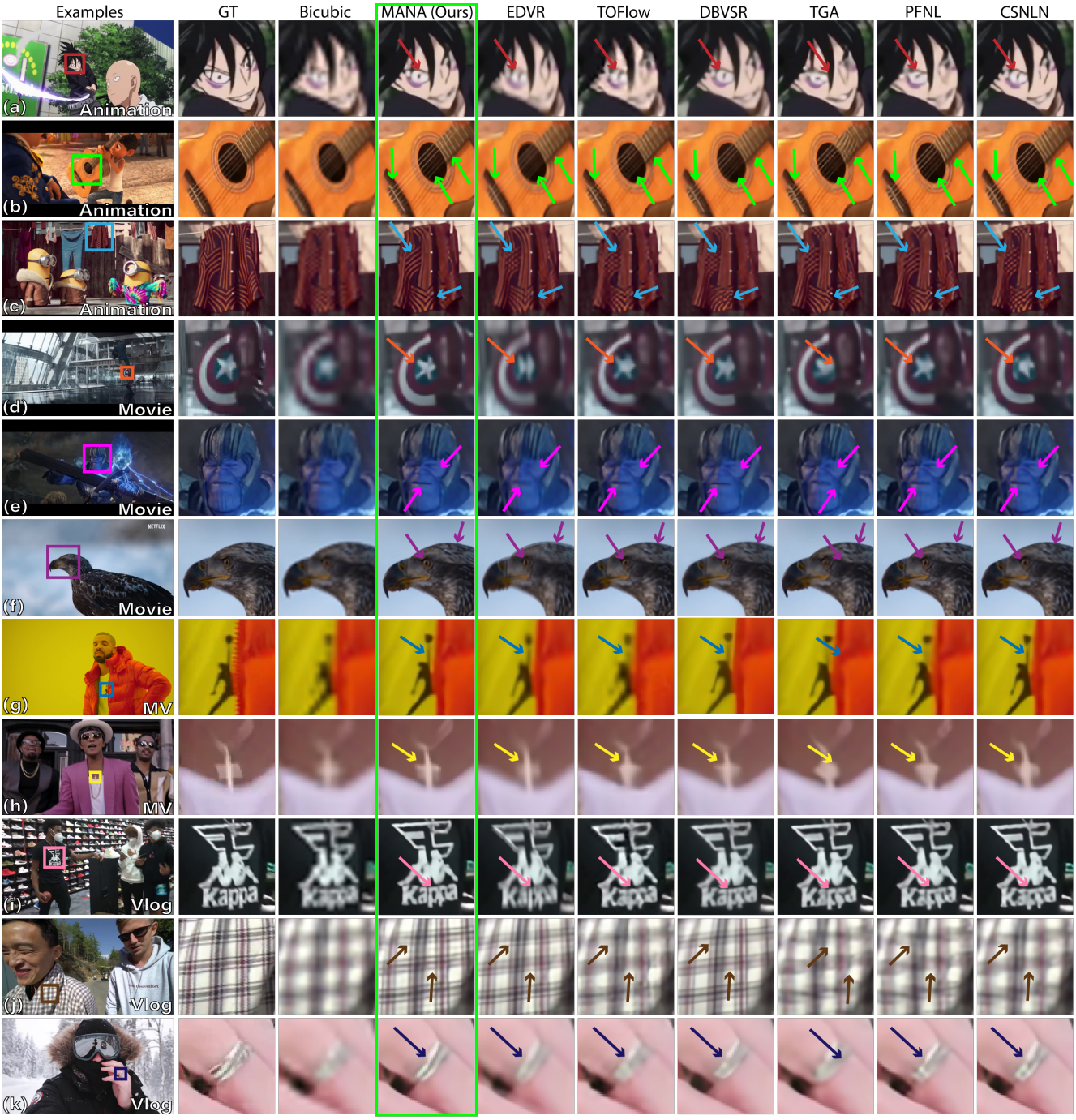

Figure 3. Additional visual comparison on examples from daily videos including animations (examples (a), (b) and (c)), movies (examples (d), (e) and (f)), MVs (examples (g) and (h)), and vlogs (examples (i), (j) and (k)). Our method works consistently better in common types of real-world video, indicating the robustness of our method.