Supplementary Material for Recurrent Video Restoration Transformer with Guided Deformable Attention

In this supplementary material, we first give more details on training and testing datasets, as well as evaluation metrics. Then, we provide more visual comparisons of different methods.

1 More Details on Datasets and Evaluation Metrics

Training and testing datasets. For video super-resolution, we train the model on two different training datasets for scale factor 4. First, we generate low-resolution images by the MATLAB `imresize` function (*i.e.*, bicubic degradation) and train the model on REDS [8]. REDS4 [17] (*i.e.*, clip 000, 011, 015, 020) is used as the test set. Second, we train the model on Vimeo-90K [18] with two different degradations: bicubic and blur downsampling (Gaussian blur with $\sigma = 1.6$ followed by subsampling). The testing datasets include Vimeo-90K-T [18], Vid4 [7] and UDM10 [19]. On 8 Nvidia A100 GPUs, it takes about 17 days. For video deblurring, we train the model on two different datasets DVD [12] and GoPro [9]. The training time is about 10 days. We test it on their corresponding testing sets. For video denoising, we train the model on the DAVIS [4] and test it on the corresponding testing set and Set8 [14]. The training time is similar to deblurring. We train all models on 8 Nvidia A100 GPUs. It takes about 16.6 days for video SR and 9.7 days for video deblurring and denoising. For training memory cost, it consumes about 39GB and 29GB for video SR and other two tasks, respectively.

REDS [8] REDS is a newly-proposed high-quality (1280×720) video dataset for video restoration. It has 270 clips for training and validation. Following [17], we use REDS4 (4 selected representative clips, *i.e.*, 000, 011, 015 and 020) for evaluation and the rest 266 clips for training. This dataset is used for training bicubic video SR.

Vimeo-90K [18] Vimeo-90K is a widely-used middle-quality (448×256) dataset for video restoration. For video SR benchmarking, it uses 64,612 clips for training and 7,824 clips for testing (denoted as Vimeo-90K-T). This dataset is used for training bicubic and blur-downsampling video SR.

Vid4 [7] Vid4 is a classical testing dataset for video restoration. It contains 4 video clips (*i.e.*, calendar, city, foliage and walk). Each clip has at least 34 frames (720×480).

UDM10 [19] UDM10 is a recent proposed testing dataset for video super-resolution. It contains 4 video clips of various scenes, each of which has 32 frames (1272×720).

DVD [12] DVD is a widely-used high-quality (1280×720) dataset for video deblurring. Blurred images are generated from high fps videos. It has 61 videos (5,708 frames in total) for training and 10 videos (1,000 frames in total) for testing.

GoPro [9] GoPro is a popular high-quality (1280×720) for image and video deblurring. Similar to DVD [12], blurred images are synthesized based on high fps videos. It is consisted of 22 training clips (2,103 frames in total) and 11 testing clips (1,111 frames in total).

DAVIS [4] DAVIS-2017 is a popular middle-quality (854×480) dataset for video denoising. It consists of 90 videos for training and 30 videos for testing.

Set8 [15] Set8 consists of 8 middle quality (960×540) videos (*i.e.*, tractor, touchdown, park_joy, sunflower, hypersmooth, motorbike, rafting and snowboard). It is often used as a testing dataset in video denoising. Following [14–16], we only use the first 85 frames of each video.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Evaluation metrics. Following [17, 2, 5, 13, 14], we calculate the metrics on RGB channel for REDS4 [17], DVD testing set [12], GoPro testing set [9], DAVIS testing set [4] as well as Set8 [14], and on the Y channel for Vimeo-90K-T [18], Vid4 [7] and UDM10 [19].

2 More Visual Comparison

As shown in Fig. 1, Fig. 2 and Fig. 3, we provide more visual results to show the effectiveness of the proposed RVRT on video super-resolution, video deblurring and video denoising. RVRT generates visually pleasing frames with fine details and sharp edges.

References

[1] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021.

[2] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021.

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. *arXiv preprint arXiv:2104.13371*, 2021.

[4] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision*, pages 123–141, 2018.

[5] Dongxu Li, Chenchen Xu, Kaihao Zhang, Xin Yu, Yiran Zhong, Wenqi Ren, Hanna Suominen, and Hongdong Li. Arvo: Learning all-range volumetric correspondence for video deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7721–7731, 2021.

[6] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022.

[7] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):346–360, 2013.

[8] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1996–2005, 2019.

[9] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017.

[10] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3043–3051, 2020.

[11] Hyeongseok Son, Junyong Lee, Jonghyeop Lee, Sunghyun Cho, and Seungyong Lee. Recurrent video deblurring with blur-invariant motion estimation and pixel volumes. *ACM Transactions on Graphics*, 40(5):1–18, 2021.

[12] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1279–1288, 2017.

[13] Maitreya Suin and AN Rajagopalan. Gated spatio-temporal attention-guided video deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7802–7811, 2021.

[14] Matias Tassano, Julie Delon, and Thomas Veit. Dvdnet: A fast network for deep video denoising. In *IEEE International Conference on Image Processing*, pages 1805–1809, 2019.

[15] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1354–1363, 2020.

[16] Gregory Vaksman, Michael Elad, and Peyman Milanfar. Patch craft: Video denoising by deep modeling and patch matching. In *IEEE International Conference on Computer Vision*, pages 1759–1768, 2021.

[17] Xintao Wang, Kelvin CK 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*, pages 1954–1963, 2019.

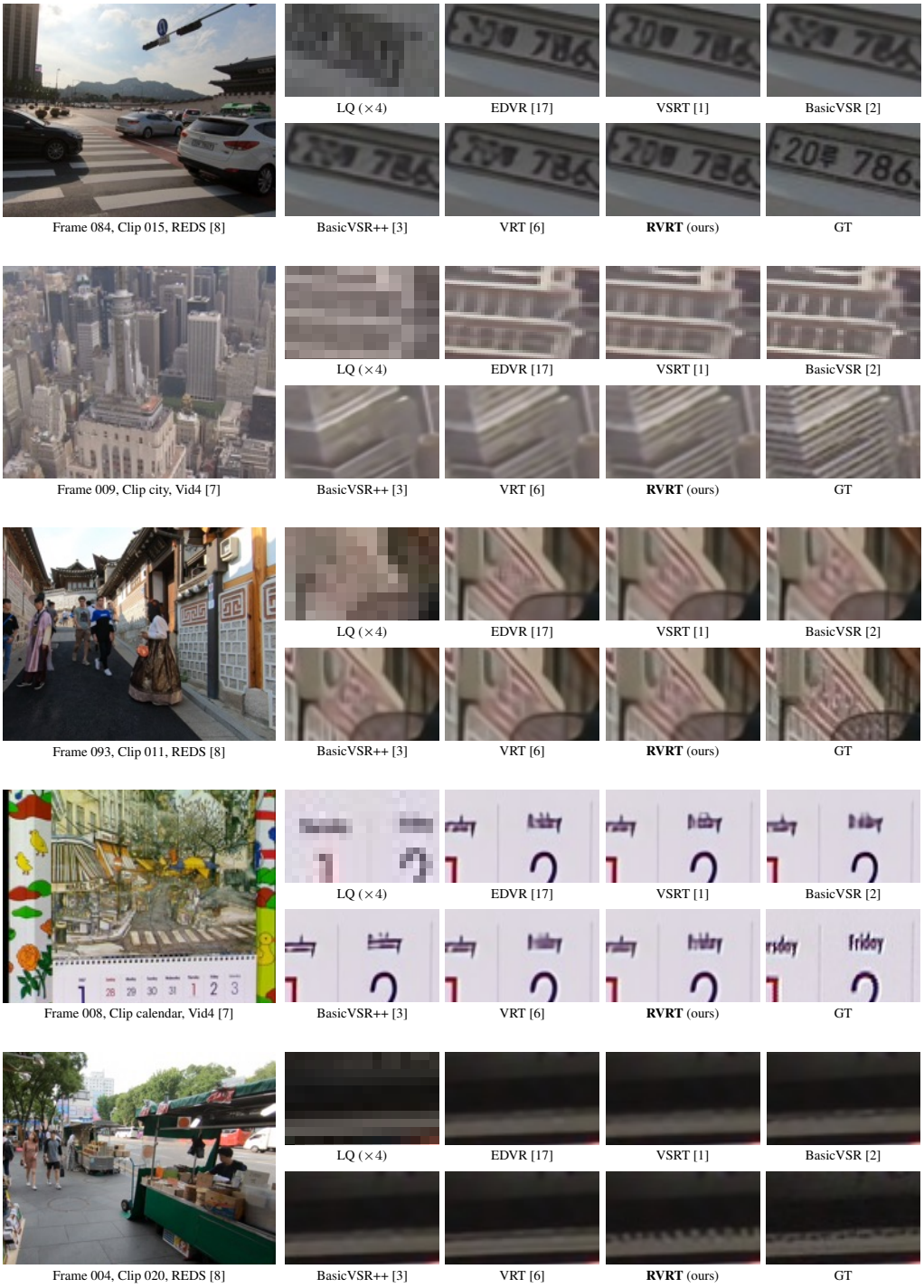

Figure 1: More visual comparison of **video super-resolution** ($\times 4$) methods.

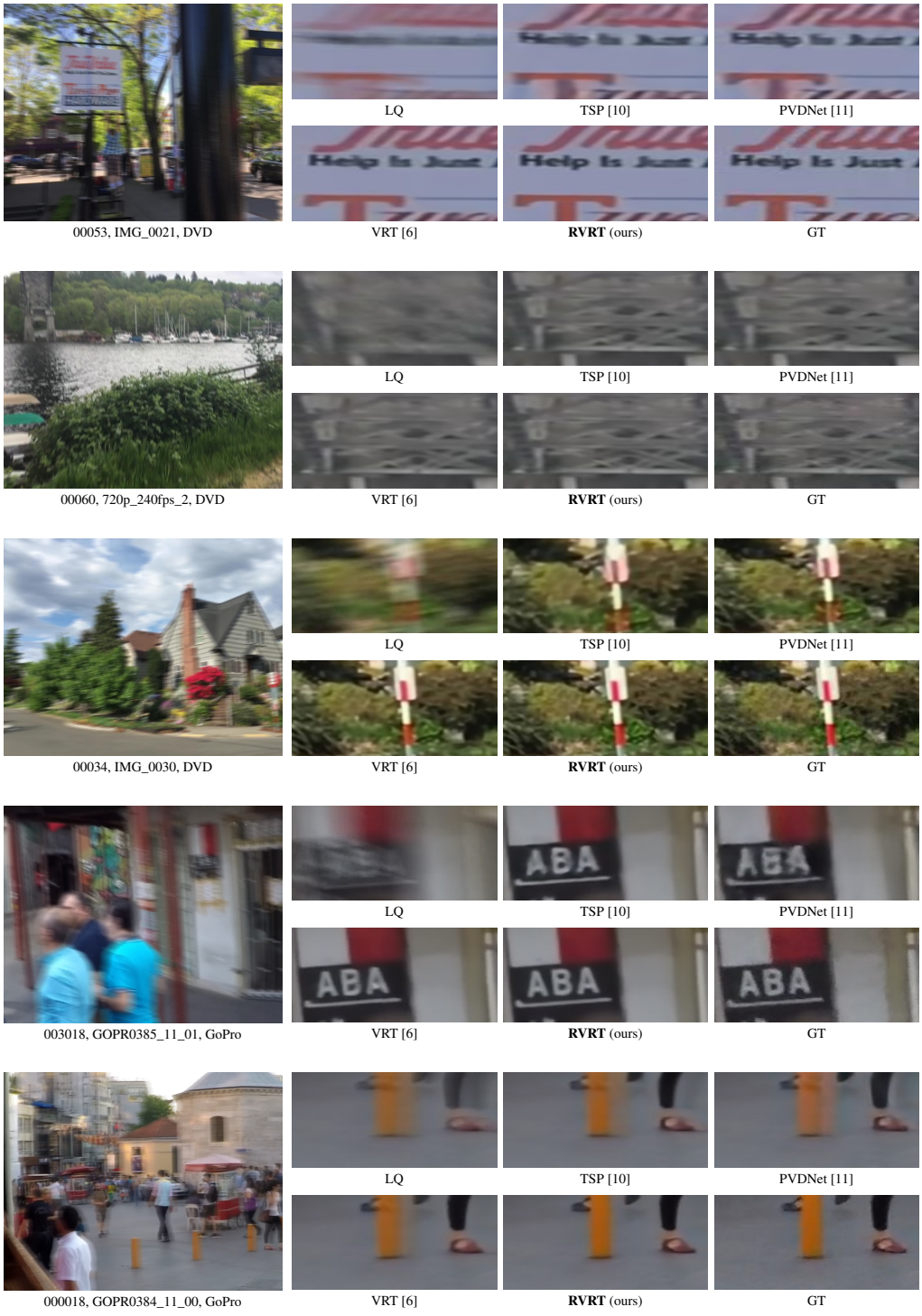

Figure 2: More visual comparison of **video deblurring** methods on DVD [12] and GoPro [9].

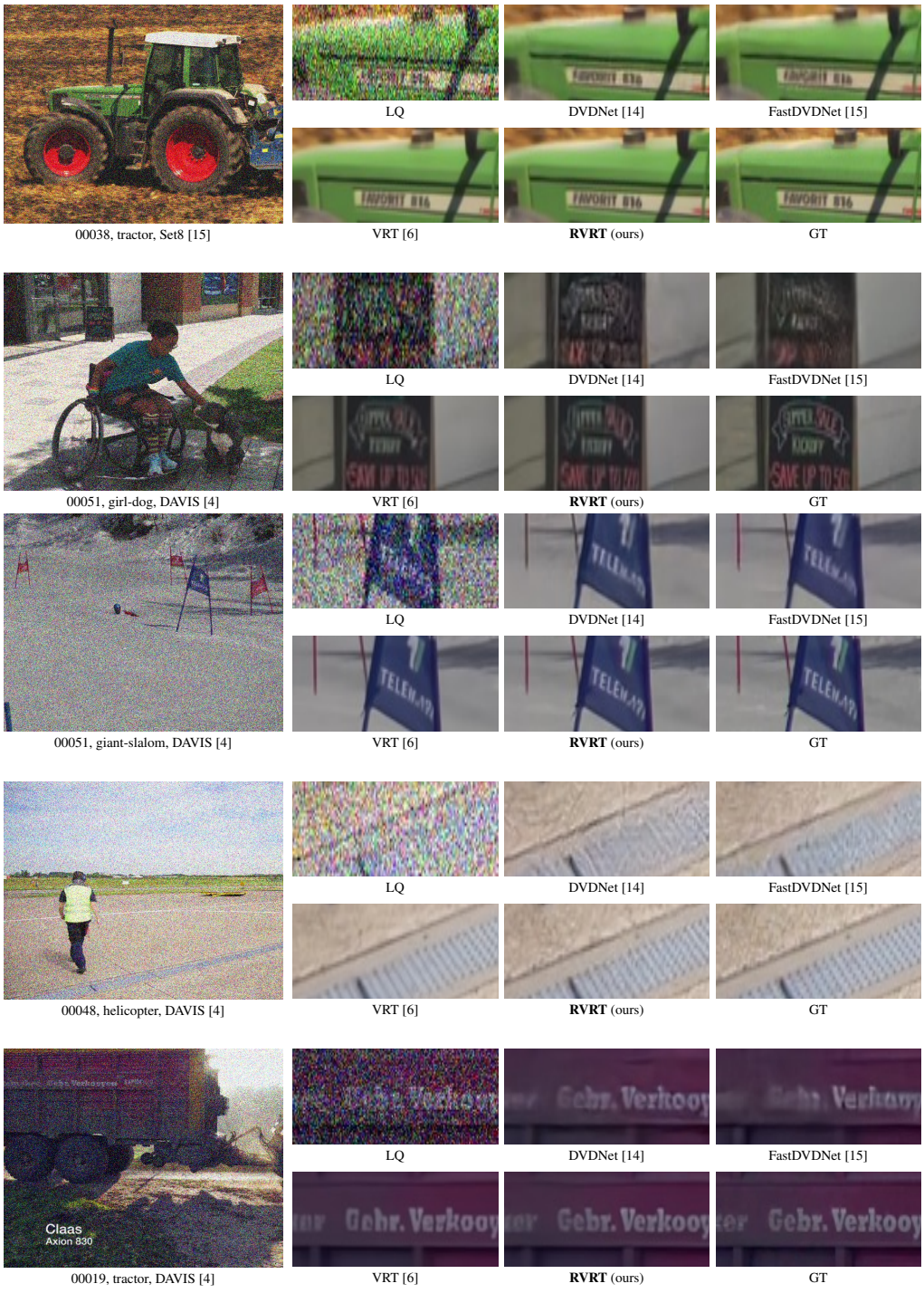

Figure 3: More visual comparison of **video denoising** ($\sigma = 50$) methods on DAVIS [4] and Set8 [15].

[18] 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.

[19] 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 *IEEE International Conference on Computer Vision*, pages 3106–3115, 2019.