A. Implementation Details

All models included in experiments are trained from scratch to perform fair comparisons.

For video frame interpolation methods incorporated in the experiments (*i.e.* Super SloMo [3], QVI [8], and DAIN [1]), we train them on Adobe240 dataset [5]. We keep all the training settings the same as proposed in their original papers, including the optimizer, initial learning rate, learning rate decay strategy, and the number of training epochs. For data settings, 9 consecutive frames are selected from video clips for training in every iteration. Networks take the first and last frames as inputs and generate intermediate 7 frames. We calculate loss between generated frames and the original ground-truth frames. Each video frame is resized to have a shorter spatial dimension of 360, and a random crop of 352×352 is performed.

Zooming SlowMo [6] and TMNet [7] are two STVSR models included in our experiments. Zooming SlowMo only supports fixed frame interpolation, and the interpolation time is set to $0, 0.5, 1$ in the original paper. Following their settings, we also train the model from scratch to interpolate the fixed time instances. To ensure that the input video frames of all models are of the same frame rate, we extract 9 consecutive frames from video clips and take the 1^{st} and 9^{th} frames as inputs. We then down-sample the input frames via Bicubic interpolation by a factor of 4 and use the network to predict the high-resolution versions of the 1^{st}, 5^{th} and 9^{th} frames. TMNet supports arbitrary frame interpolation. In its paper, the authors mention that TMNet needs a two-stage training process for convergence, and we follow their suggestions. In the first stage, we pre-train the network on the Vimeo90K dataset [9]. The Vimeo90K dataset consists of 7-frame video sequences. We use the 1^{st}, 3^{rd}, 5^{th}, and 7^{th} frames after down-sampling as the network inputs and predict the high-resolution results of all the 7 frames, which means that the interpolation time is set to $0, 0.5, 1$ in this stage. In the second stage, we select 9 consecutive frames from video clips, and the 1^{st} and 9^{th} frames are taken as inputs. After down-sampling, we use the network to generate high-resolution predictions of all 9 frames and calculate the loss value with the original high-resolution frames. TMNet is trained with more data, which may lead to advantages in the experiments.

For the training of VideoINR, we select 9 consecutive frames and down-sample the 1^{st} and 9^{th} frames as model inputs. In each iteration, We randomly select three frames from the 9-frame video sequence and use the network to generate high-resolution predictions at the time instances of the three selected frames.

We keep the training settings unchanged for Zooming SlowMo, TMNet, and VideoINR. All three models are optimized with the Charbonnier loss function [4].

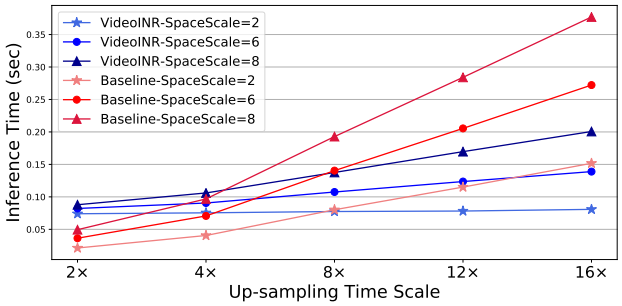

Figure 1. Inference time comparisons on different space scales.

Figure 2. Failure case. Left is the overlay of frames at t=0 and t=1. Right is the interpolated frame at t=0.5

B. Efficiency on Different Scales

To evaluate the efficiency of VideoINR on different up-sampling space scales, we provide more inference time comparisons in Figure 1. We select the two-stage method composed of SuperSlomo and LIIF as the baseline, as it supports arbitrary up-sampling scales on both space and time.

C. Limitations

In some challenging cases, large motion and occlusion result in errors on the motion flow field, leading to blurred results with unclear boundaries. We show a failure cases of VideoINR in Figure 2.

D. Additional Qualitative Results

We provide more qualitative results in Figures 3,4,5,6. We compare VideoINR with two STVSR methods, DAIN [1] + BasicVSR [2] and TMNet [7]. The up-sampling space scale is set to 4 for all examples. In Figure 3, 4, we set the time scale for interpolation to 8, which is in our training distribution. We observe that DAIN + BasicVSR and TMNet tend to generate blurry regions or artifacts. In contrast, the results of VideoINR are consistent and aligned across two input frames, with sharp edges and clear details. In Figure 5, 6, we set the time scale to 12 and 16, which are out of the training distribution. We find that VideoINR well recovers objects with large motion and preserves better textural information compared with other methods. In summary, VideoINR shows the advantages of using implicit function to represent continuous videos and address the space-time video super-resolution task. More visualization results can be found in the provided video.

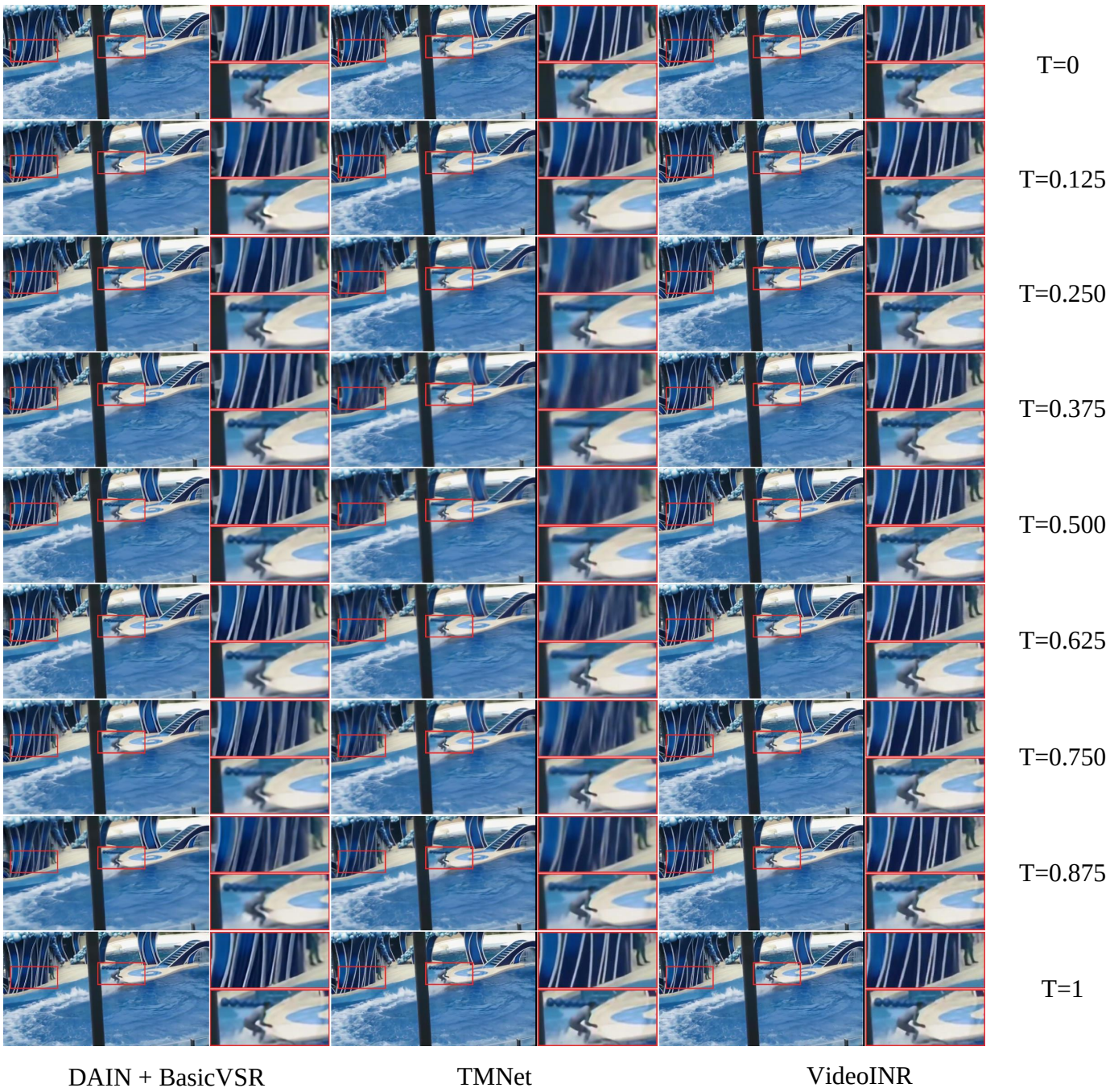

Figure 3. Qualitative comparisons of different STVSR methods on in-distribution time scale. Best zoom in for better visualization.

References

[1] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019. 1

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021. 1

[3] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018. 1

Figure 4. Qualitative comparisons of different STVSR methods on in-distribution time scale. Best zoom in for better visualization.

[4] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 1

[5] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1279–1288, 2017. 1

[6] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. Zooming slow-mo: Fast and

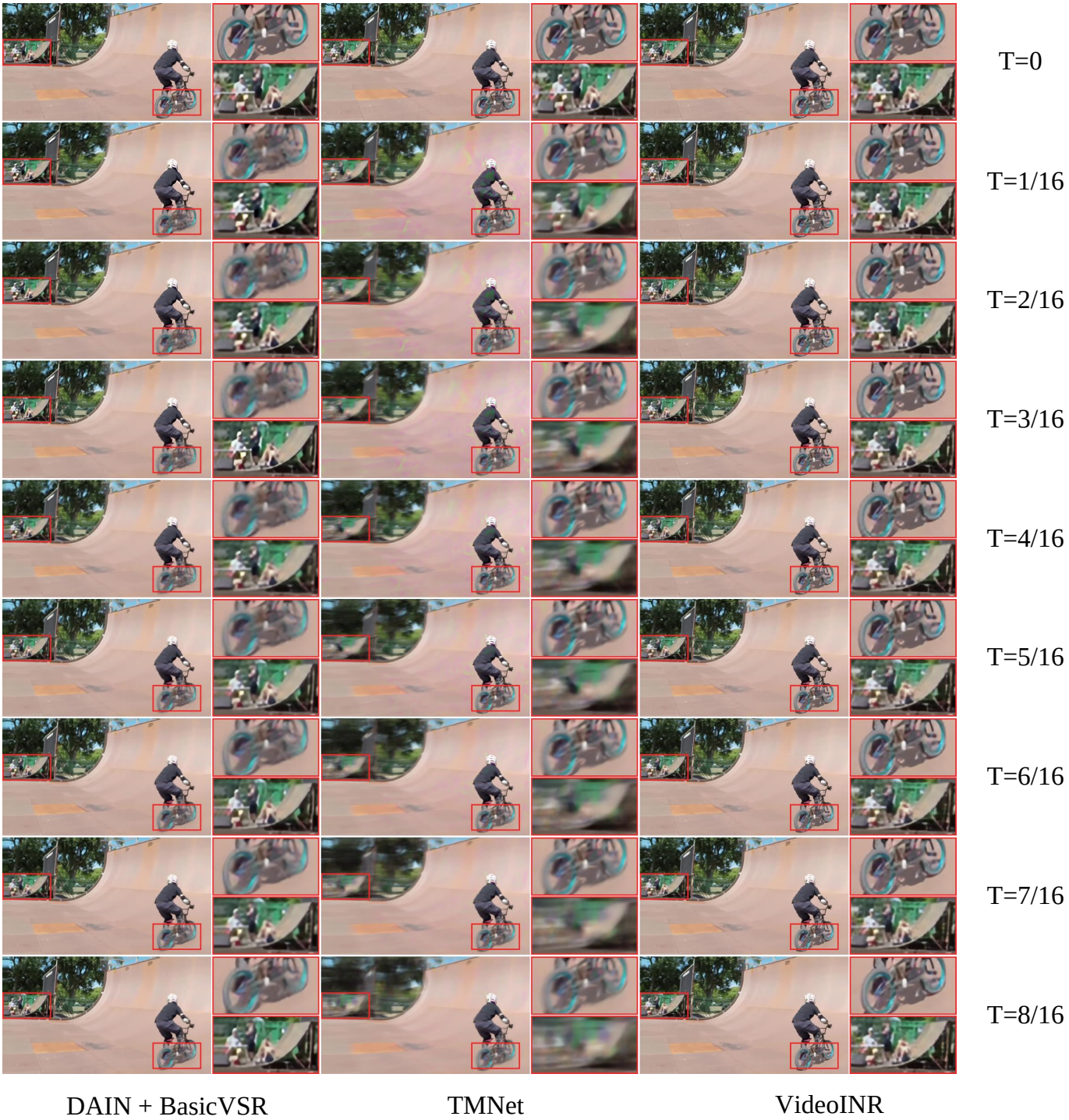

Figure 5. Qualitative comparisons of different STVSR methods on out-of-distribution time scale. Best zoom in for better visualization.

accurate one-stage space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3370–3379, 2020. 1

[7] Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. Temporal modulation network for controllable space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6388–6397, 2021. 1

[8] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. *Advances in Neural Information Processing Systems*, 32:1647–1656, 2019. 1

[9] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and

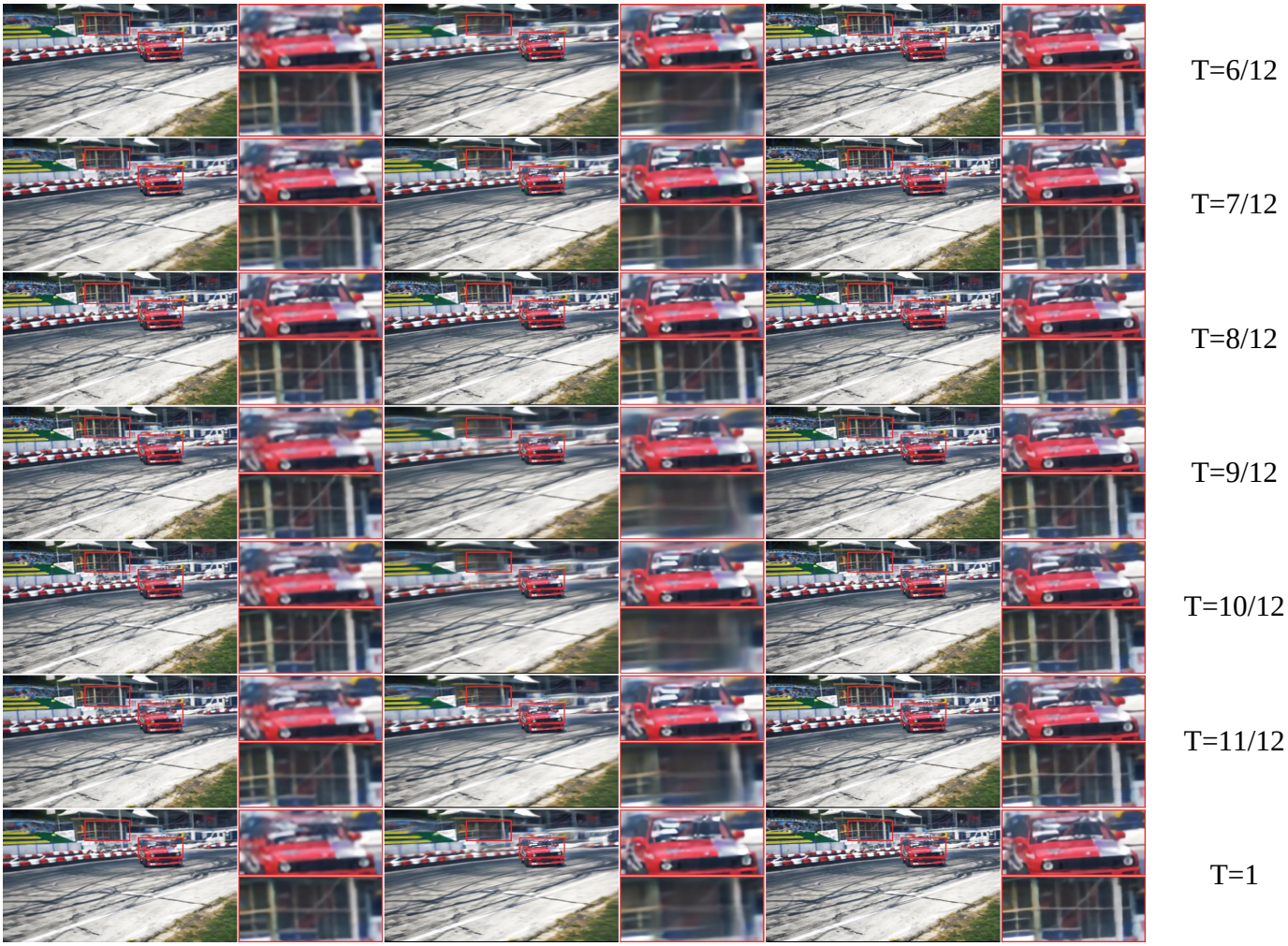

Figure 6. Qualitative comparisons of different STVSR methods on out-of-distribution time scale. Best zoom in for better visualization.

William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 1