# Supplementary Material : Video Rescaling Networks with Joint Optimization Strategies for Downscaling and Upscaling

Yan-Cheng Huang[1,*] Yi-Hsin Chen[1,*] Cheng-You Lu[1]
Hui-Po Wang[2] Wen-Hsiao Peng[1] Ching-Chun Huang[1]
[1] National Yang Ming Chiao Tung University, Taiwan
[2] CISPA Helmholtz Center for Information Security
{s0756722.iie07g, yhchen.iie07g, johnny305.cs04}@nctu.edu.tw
hui.wang@cispa.saarland, {wpeng, chingchun}@cs.nctu.edu.tw

This supplementary document provides additional results to validate the proposed methods. These include (1) quantitative results based on the Video Multi-Method Assessment Fusion (VMAF) metric [2], a video quality metric that is shown to correlate highly with human perception; (2) an examination of temporal consistency on the downscaled and upscaled videos following the method in RSDN [1]; (3) an ablation study of the predictive module; and (4) more qualitative results of the downscaled and upscaled videos.

## 1. Quantitative results based on VMAF

This section presents quantitative results for upscaled and downscaled videos in Vid4 based on the Video Multi-Method Assessment Fusion (VMAF) [2] metric. VMAF is an objective video quality metric that is shown to correlate highly with human perception. It accepts as inputs a reference video and a distorted video. Its output is a score between 0 and 100. The higher the VMAF score, the more closely the two input videos match each other. Compared in Table A1 are the VMAF scores of the reconstructed high-resolution (HR) videos produced by a few baselines (whose code is available) and our schemes. The reference videos are the original HR videos. Likewise, Table A2 shows the results for the downscaled videos, where the reference videos are the bicubic-downscaled videos.

From Table A1, we observe that our LSTM-VRN, MIMO-VRN and MIMO-VRN-C outperform IRN [4] (image-based joint optimization scheme) consistently. As compared with EDVR-L [3], a traditional video super-resolution approach, the proposed methods show significantly improved VMAF scores. These observations are in line with the PSNR/SSIM results reported in Table 1 of the main paper. A similar observation can be made regarding the results of the downscaled LR videos in Table A2, which agrees with the PSNR/SSIM results in Table 3 of the main paper. Remarkably, all three proposed methods have VMAF scores very close to 100, suggesting that their LR videos are visually similar to bicubic-downscaled videos.

Table A1: Comparison of VMAF scores for ×4 upscaled HR videos on Vid4. The reference videos are the original HR videos. The higher the VMAP scores, the better the HR reconstruction quality. Red and green indicate the best and the second best performance, respectively.

| Method | Calendar | City | Foliage | Walk | Average |
|---|---|---|---|---|---|
| EDVR-L [3] | 64.11 | 47.22 | 77.12 | 89.86 | 69.58 |
| IRN [4] | 77.85 | 80.80 | 89.83 | 97.01 | 86.37 |
| LSTM-VRN | 78.03 | 85.16 | 91.14 | 97.81 | 88.04 |
| MIMO-VRN | 81.99 | 87.35 | 95.31 | 98.57 | 90.81 |
| MIMO-VRN-C | 80.44 | 85.47 | 94.24 | 98.42 | 89.64 |

Table A2: Comparison of VMAF scores for ×4 downscaled LR videos on Vid4. The reference videos are generated by the bicublic downscaling method. The higher the VMAP scores, the more closely the LR videos resemble the bicubic-downscaled ones. Red and green indicate the best and the second best performance, respectively.

| Method | Calendar | City | Foliage | Walk | Average |
|---|---|---|---|---|---|
| IRN [4] | 94.18 | 94.92 | 95.38 | 97.82 | 95.55 |
| LSTM-VRN | 95.86 | 95.31 | 96.52 | 99.47 | 96.79 |
| MIMO-VRN | 97.37 | 96.84 | 97.88 | 99.83 | 97.98 |
| MIMO-VRN-C | 97.94 | 97.17 | 98.34 | 99.87 | 98.33 |

## 2. Temporal consistency

In this section, the temporal consistency of the downscaled and upscaled videos is examined. We follow RSDN [1] to extract a row or a column of pixels at the

[*] Both authors contributed equally to this work.

Table A3: Ablation study of the predictive module. IRN_Ret is trained using the same dataset as LSTM-VRN. MIMO-VRN-C-Zero is an implementation of MIMO-VRN-C without the predictive module. All the presented results are evaluated on Vid4 and they show that the proposed predictive module can improve the quality of reconstructed HR videos.

| Method | Predictive Module | HR (PSNR-Y / SSIM-Y / VMAF) |
|---|---|---|
| IRN_Ret | | 30.72 / 0.9087 / 86.37 |
| LSTM-VRN | √ | 32.24 / 0.9369 / 88.04 |
| MIMO-VRN-C-Zero | | 33.04 / 0.9575 / 89.00 |
| MIMO-VRN-C | √ | 33.40 / 0.9609 / 89.64 |

co-located positions in consecutive video frames. We then stitch vertically (or horizontally) these extracted rows (or columns) of pixels to form an image, in order to visualize their variations in the temporal dimension. From Fig. A1, we are not aware of any noticeable inconsistency across reconstructed HR video frames. Moreover, the resulting images of our methods resemble closely the ground-truths.

In Fig. A2, IRN [4] and LSTM-VRN result in slight temporal inconsistency in some areas of the low-resolution (LR) videos, particularly the Calendar and City sequences. This is evidenced by the aliasing artifact especially appeared on the alphabet of Calendar sequence and the building of City sequence. Nevertheless, such inconsistency does not appear in the LR videos produced by MIMO-VRN and MIMO-VRN-C.

## 3. The effectiveness of the predictive module

Table A3 provides quantitative results to justify the effectiveness of the proposed predictive module. Recall that the predictive module forms a prediction $\hat{z}$ of the missing high-frequency component $z$ from the LR video frames $\hat{y}$. This new feature distinguishes our schemes from IRN [4], the image-based joint optimization scheme, which uses a Gaussian noise for $\hat{z}$. The upper section of Table A3 compares the results of LSTM-VRN with IRN_Ret (the retrained IRN that uses the same training dataset as LSTM-VRN), validating that the predictive module is effective for reconstructing better HR videos. The lower section of Table A3 conducts the same analysis for MIMO-VRN-C, where we replace $\hat{z}$ produced by the predictive module with a fixed zero tensor, a scheme termed MIMO-VRN-C-Zero. MIMO-VRN-C-Zero is trained in the same way as MIMO-VRN-C. We see that the predictive module is still effective, even though the gain is less significant than the case in LSTM-VRN.

## 4. More qualitative results

Figs. A3 and A4 provide more qualitative results, comparing the reconstructed HR videos of different models. They again suggest that our models can recover fine details and sharp edges.

Fig. A5 presents a frame-by-frame qualitative comparison between MIMO-VRN and MIMO-VRN-C, with a GoF size of 5. MIMO-VRN-C comes with an additional center loss to ensure temporal consistency. It is seen that both MIMO-VRN and MIMO-VRN-C can successfully reconstruct image details. Comparing MIMO-VRN with MIMO-VRN-C, we are not aware of any significant quality variation in the temporal dimension, even though Fig. 7 of the main paper suggests that the HR quality of MIMO-VRN may fluctuate more significantly than MIMO-VRN-C. Fig. A6 displays more LR video frames from Vid4, showing that our models offer comparable visual quality to the bicubic-downscaled videos.

## References

[1] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[2] R. Rassool. Vmaf reproducibility: Validating a perceptual practical video quality metric. In *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2017.

[3] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[4] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

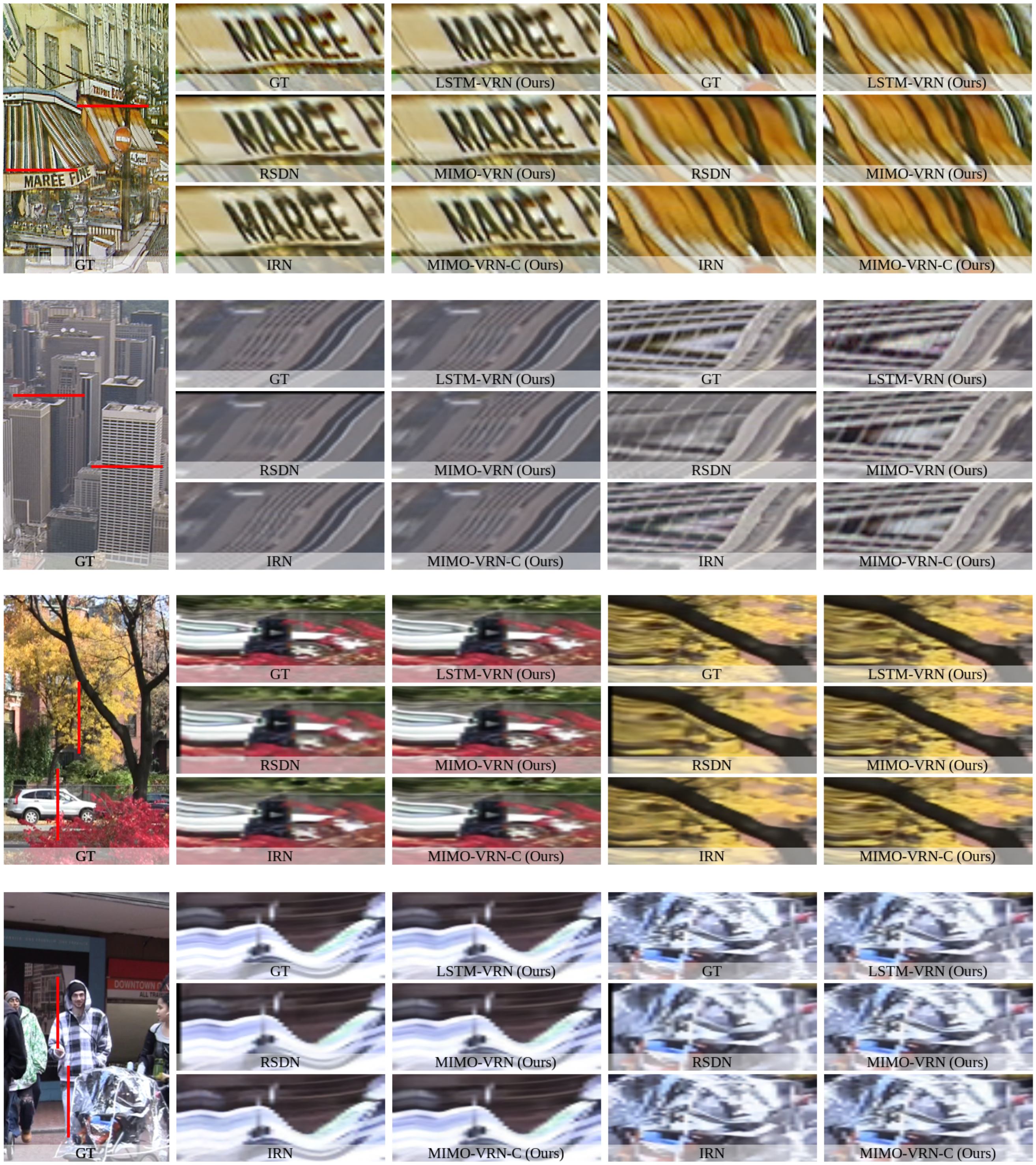

Figure A1: Visualization of temporal consistency across the reconstructed HR video frames. The images shown are formed by vertically (or horizontally) stitching rows (or columns) of pixels extracted separately from consecutive video frames at co-located positions (indicated by the red lines).

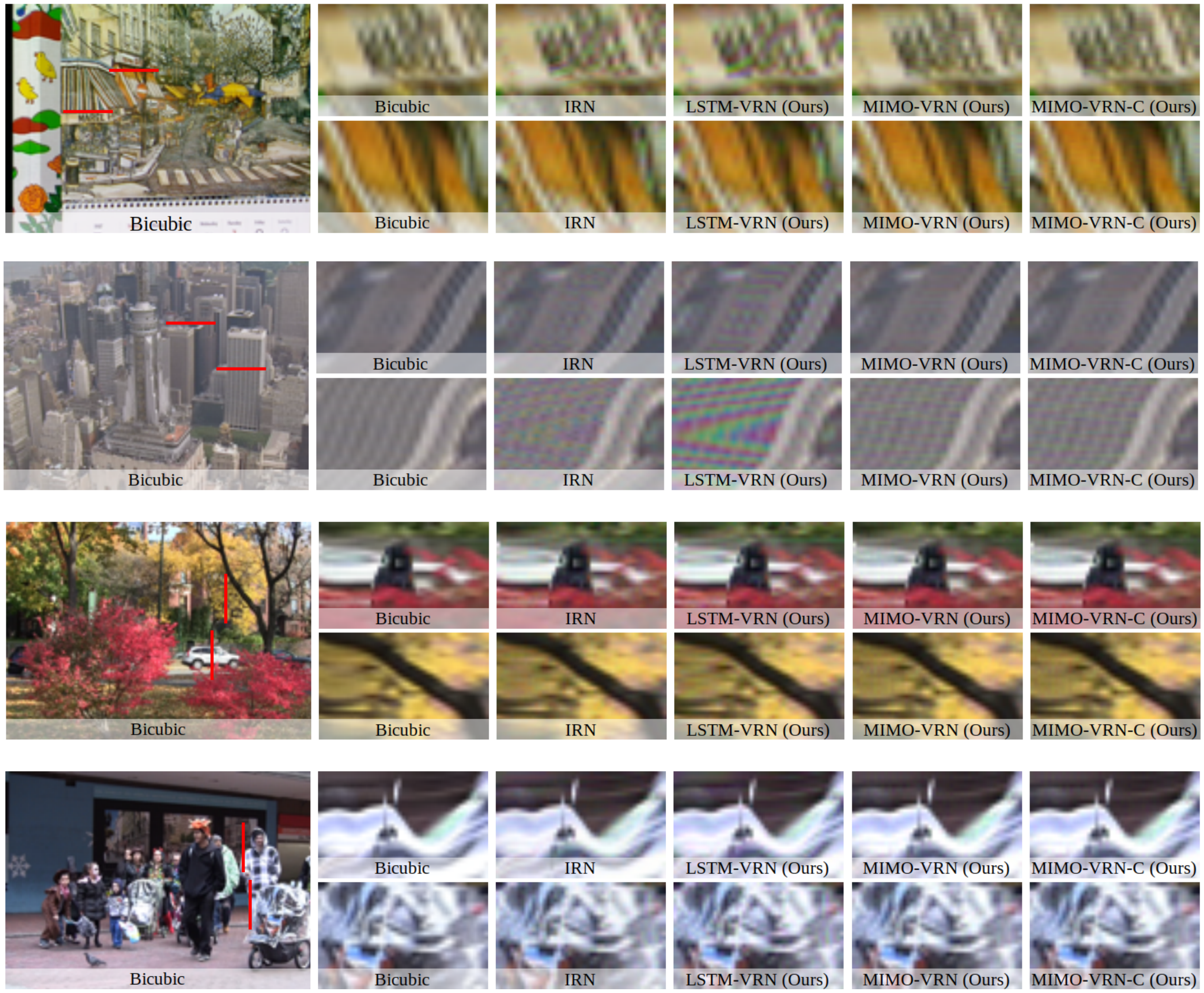

Figure A2: Visualization of temporal consistency across the downscaled LR video frames. The images shown are formed by vertically (or horizontally) stitching rows (or columns) of pixels extracted separately from consecutive video frames at co-located positions (indicated by the red lines).

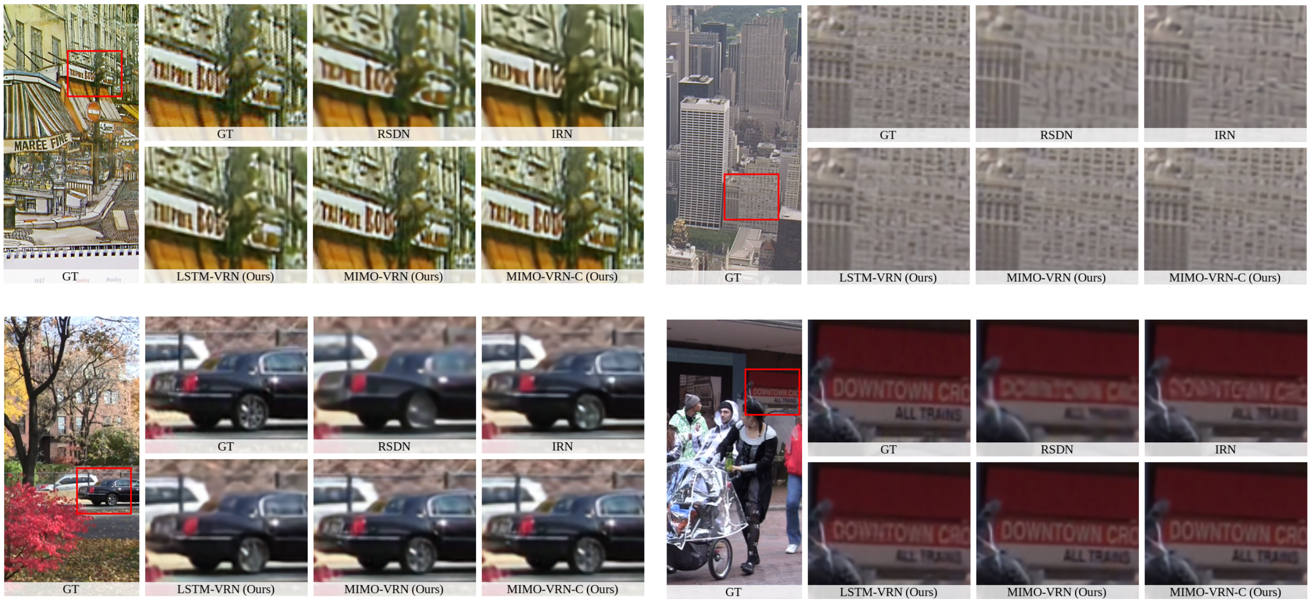

Figure A3: Qualitative comparison on Vid4 for ×4 upscaling. Zoom in for better visualization.

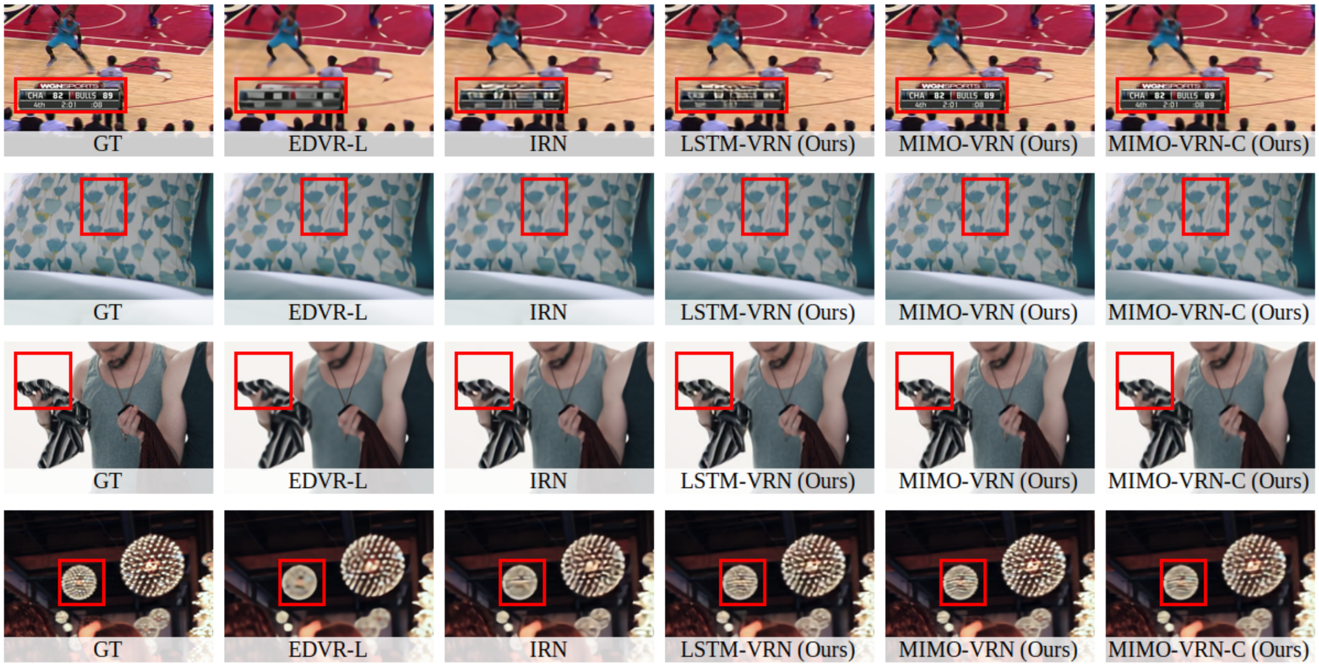

Figure A4: Qualitative comparison on Vimeo-90K-T for ×4 upscaling. Zoom in for better visualization.

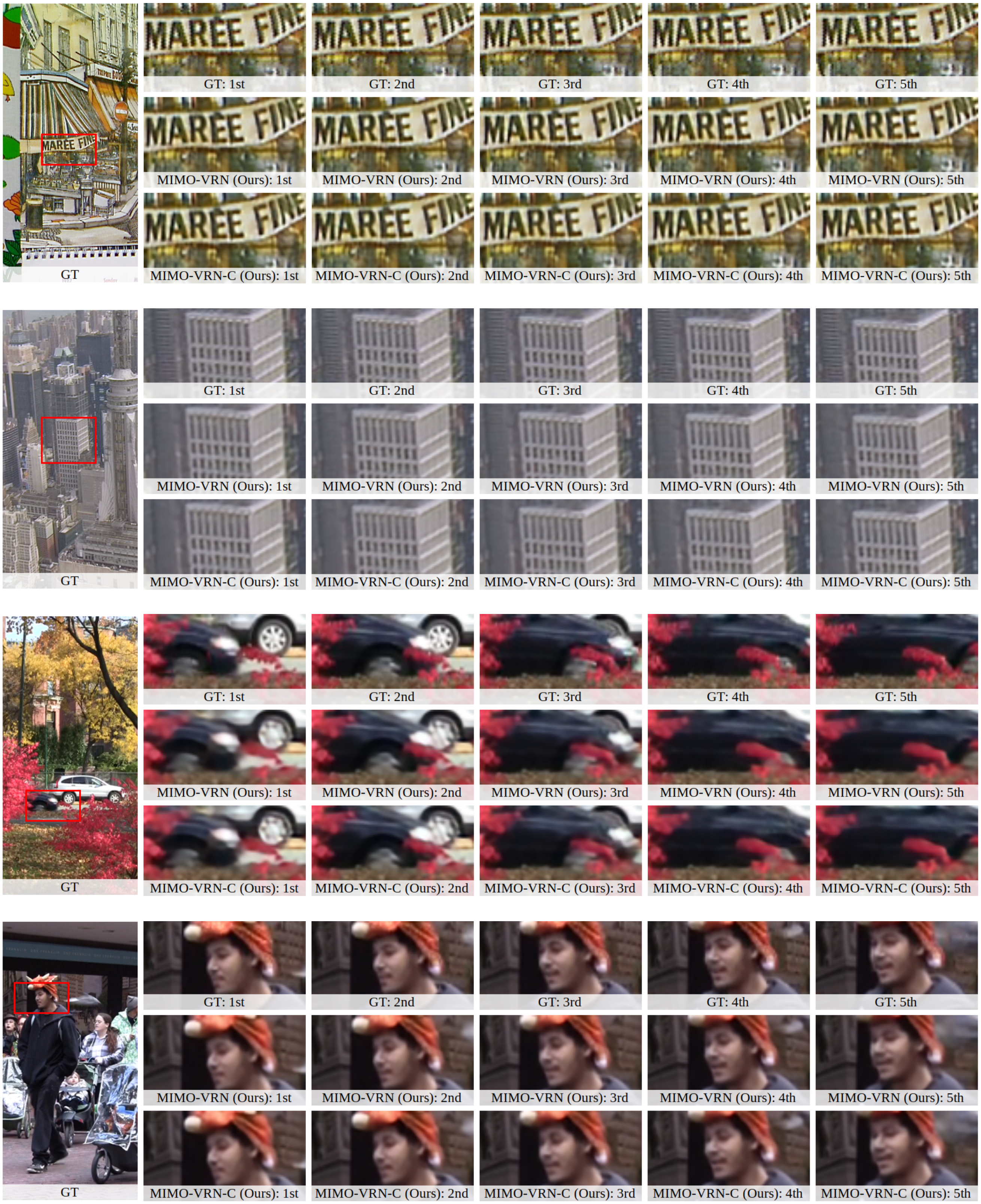

Figure A5: Qualitative frame-by-frame comparison between MIMO-VRN and MIMO-VRN-C, with a GoF size of 5. The images shown are ×4 upscaled HR video frames of Vid4.

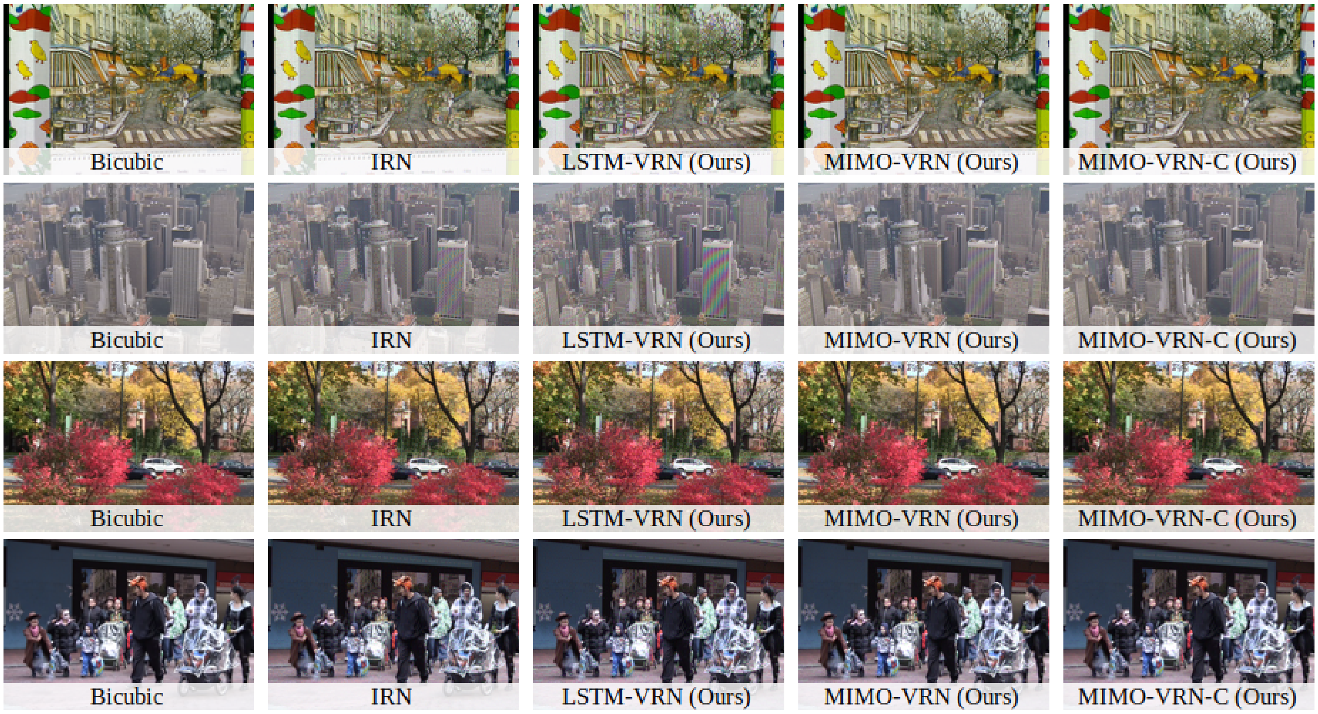

Figure A6: Sample LR video frames from Vid4. Our models show comparable visual quality to the bicubic method.