Supplementary Material for
Reference-based Video Super-Resolution Using Multi-Camera Video Triplets

Junyong Lee Myeonghee Lee Sunghyun Cho Seungyong Lee
POSTECH
{junyonglee, myeonghee, s.cho, leesy}@postech.ac.kr

1. Analysis on Reference Video Types

While we assume that a triplet of ultra-wide, wide-angle, and telephoto videos is given, our approach uses only a wide-angle video as a reference to perform super-resolution of an ultra-wide video, and a telephoto video is used only as additional supervision in the adaptation stage. One may wonder why our approach does not utilize a telephoto video as a reference as it provides the highest-resolution details. To answer the question, this section provides an analysis of the effect of reference video types on the SR quality. Specifically, among different combinations of wide-angle and telephoto videos, we find the best combination for reference video(s) and for supervision to train our network. In the following, we use the terms 'input reference' and 'reference supervision' to indicate a reference video that is fed into the network, and a high-resolution video used for additional supervision, respectively. We denote them by I^{Ref} and $I^{Ref_{HR}}$, respectively, as done in the main paper.

To quantitatively analyze the effect of reference combinations, we prepare five models that are trained with *only* the pre-training stage (Table 1), where each model is trained with a different combination of wide-angle and telephoto videos for the input reference I^{Ref} and the reference supervision $I^{Ref_{HR}}$. We also prepare a model taking dual references, both wide-angle and telephoto videos. To this end, we modify the reference alignment and propagation module (Sec. 3.3 in the main paper) to separately obtain features of a wide-angle frame $\widetilde{h}_t^{Wide}$ and features of a telephoto frame $\widetilde{h}_t^{Tele}$ that are aligned to I_t^{LR}. Moreover, the propagative temporal fusion module is modified to take both $\widetilde{h}_t^{Wide}$ and $\widetilde{h}_t^{Tele}$, and utilizes confidence maps c_t^{Wide} and c_t^{Tele} computed by matching wide-angle and telephoto frames to an LR ultra-wide frame, respectively. The module also utilizes accumulated confidence maps $\tilde{c}_t^{Wide,\{f,b\}}$ and $\tilde{c}_t^{Tele,\{f,b\}}$ as guidance for temporal Ref features during the fusion (Fig. 1). For training the model taking dual references, we use both a wide-angle I_t^{Wide} and telephoto I_t^{Tele} frames for the proposed pre-training loss ℓ_{pre} (Eq. 10 in the main paper). Specifically, ℓ_{pre} is modified as follows:

$$\ell_{pre} = \ell_{rec}(I_t^{SR}, I_t^{HR}) + \lambda_{pre}\big(\ell_{Mfid}(I_t^{SR}, I_{t\in\Omega}^{Wide}) + \ell_{Mfid}(I_t^{SR}, I_{t\in\Omega}^{Tele})\big).$$

Code and dataset: https://github.com/codeslake/RefVSR

I^{Ref}	$I^{Ref_{HR}}$	Patch size	PSNR↑	SSIM↑	Params (M)
tele	tele	32×32	29.81	0.893	4.277
wide	tele	32×32	30.41	0.895	4.277
wide	wide	32×32	30.36	0.897	4.277
dual	dual	32×32	30.39	0.888	5.076
wide	wide	64×64	**31.68**	**0.914**	4.277

Table 1. Quantitative comparison on models trained with different reference video types. In the top row, I^{Ref}, $I^{Ref_{HR}}$, and patch size indicate the input reference, reference supervision, and patch size used for the pre-training stage, respectively. 'wide', 'tele', and 'dual' indicate wide-angle, telephoto, and both wide-angle and telephoto videos, respectively. Only the pre-training stage is used for training the models.

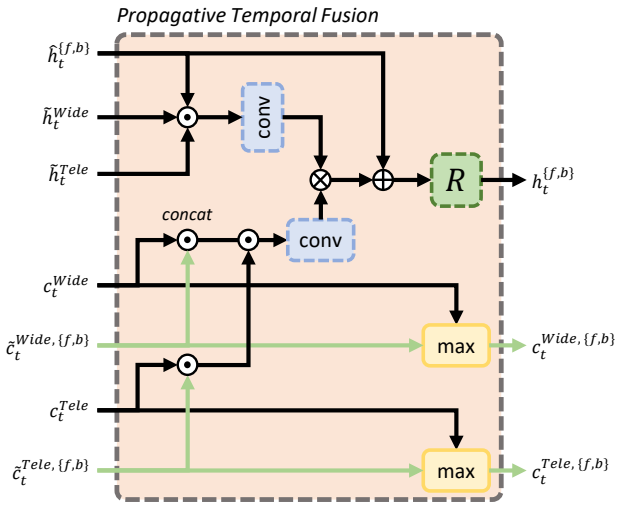

Figure 1. Modified propagative temporal fusion module for handling dual reference features.

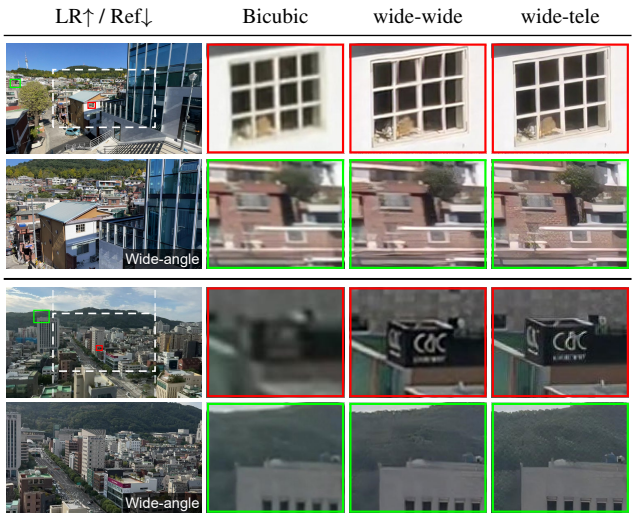

Figure 2. Qualitative comparison on 8K 4×VSR results from models trained with different reference video types for a supervision $I^{Ref_{HR}}$ in the adaptation stage. The first column shows LR and Ref real-world HD inputs. The other columns show zoomed-in cropped SR results of models taking wide-angle video as an input reference I^{Ref}, but trained with different videos for the reference supervision $I^{Ref_{HR}}$ (*e.g.*, 'wide-tele' indicates that wide-angle and telephoto videos are used for I^{Ref} and $I^{Ref_{HR}}$, respectively). Red and green boxes indicate inside and outside the overlapped FoV between LR and Ref frames, respectively.

We first verify that a wide-angle video is the best option for an input reference I^{Ref}. In Table 1, compared to the model utilizing a telephoto video as I^{Ref} (the first row of the table), the models using a wide-angle video (from second to fourth rows) show much better SR performance. This is mainly due to the larger matching coverage of a wide-angle frame on an ultra-wide frame (about 25%) than that of a telephoto frame (about 6.25%). Larger matching coverage allows fine details of reference frames to be widely transferred to a resulting SR frame, which contributes significantly in reconstructing high-quality results.

Moreover, we verify that a wide-angle video is also the best choice for reference supervision $I^{Ref_{HR}}$ needed for the pre-training stage. While it may look reasonable to use a telephoto video as $I^{Ref_{HR}}$ to transfer the resolution of a telephoto video, we found that it does not improve the SR quality much (the second row vs. the third row). This is because both wide-angle and telephoto frames lose details as they are downsampled 2× and 4×, respectively, for being used as supervision for the pre-training stage to match the scale of contents with a resulting SR frame.

One question that naturally follows is why not utilize both wide-angle and telephoto videos as dual references. However, the SR performance of the model taking dual references is almost the same as models taking a single wide-angle reference video (the third vs. the fourth rows). This indicates that it is not worth utilizing dual reference videos. A slight SR performance gain does not fully justify extra memory and computational costs (58.1T and 71.5T MACs[1] for the single and dual reference models, respectively) needed for processing additional reference video.

According to the analysis, we take advantage of the broad matching coverage of a wide-angle video and use it as the input reference I^{Ref} for both pre-training and adaptation stages. Moreover, we use wide-angle videos as the reference supervision $I^{Ref_{HR}}$ for the pre-training stage as we can have a larger training patch size (64×64 in practice) than the patch size possible when a telephoto video is used for the supervision (32×32 at maximum, due to small overlap between a telephoto and a downsampled LR frames), which boosts SR quality (the last row in the table).

In the adaptation stage, however, we can establish a large patch size even when a telephoto frame video is used as the reference supervision $I^{Ref_{HR}}$ because downsampling is not required in the real-world scenario. We thus directly use a telephoto video as reference supervision $I^{Ref_{HR}}$ for the adaptation stage to take advantage of their finest details in reconstructing SR results from a real-world HD video. The benefit of taking a telephoto video as supervision for the adaptation stage is qualitatively shown in Fig. 2. In the figure, compared to the model trained with a wide-angle video as reference supervision $I^{Ref_{HR}}$ (the third column in the figure), the model trained with a telephoto video as $I^{Ref_{HR}}$ shows sharper and finer details (the last column).

2. Effect of Propagative Temporal Fusion

In this section, we analyze the effect of the propagative temporal fusion module on the SR quality inside and outside the overlapped FoV between an LR frame and the corresponding Ref frame. The proposed propagative temporal fusion module performs fusion between Ref features $\widetilde{h}_t^{Ref}$ at the current time step and temporally aggregated features $\widehat{h}_t^{\{f,b\}}$ propagated from the previous step. During the fusion, the module utilizes the matching confidence c_t and the accumulated matching confidence $c_{t\pm1}^{\{f,b\}}$ as guidance for $\widetilde{h}_t^{\{f,b\}}$ and $\widehat{h}_t^{\{f,b\}}$, respectively.

The proposed propagative temporal fusion module improves the SR quality of both regions inside and outside the overlapped FoV, as the module utilizes the accumulated matching confidence $c_{t\pm1}^{\{f,b\}}$ during the fusion. This is because $c_{t\pm1}^{\{f,b\}}$ provides a cue for the temporal propagative fusion module to select well-matched *temporal* Ref features aggregated in temporally aggregated features $\widehat{h}_t^{\{f,b\}}$.

Fig. 3 qualitatively demonstrates the effect of the propagative temporal fusion module. For the evaluation, we prepare models with and without the propagative temporal fusion module. For the model without the propagative tempo-

[1] Computational costs are measured as the number of multiply-accumulate operations (MACs) computed on 1920×1080 frames.

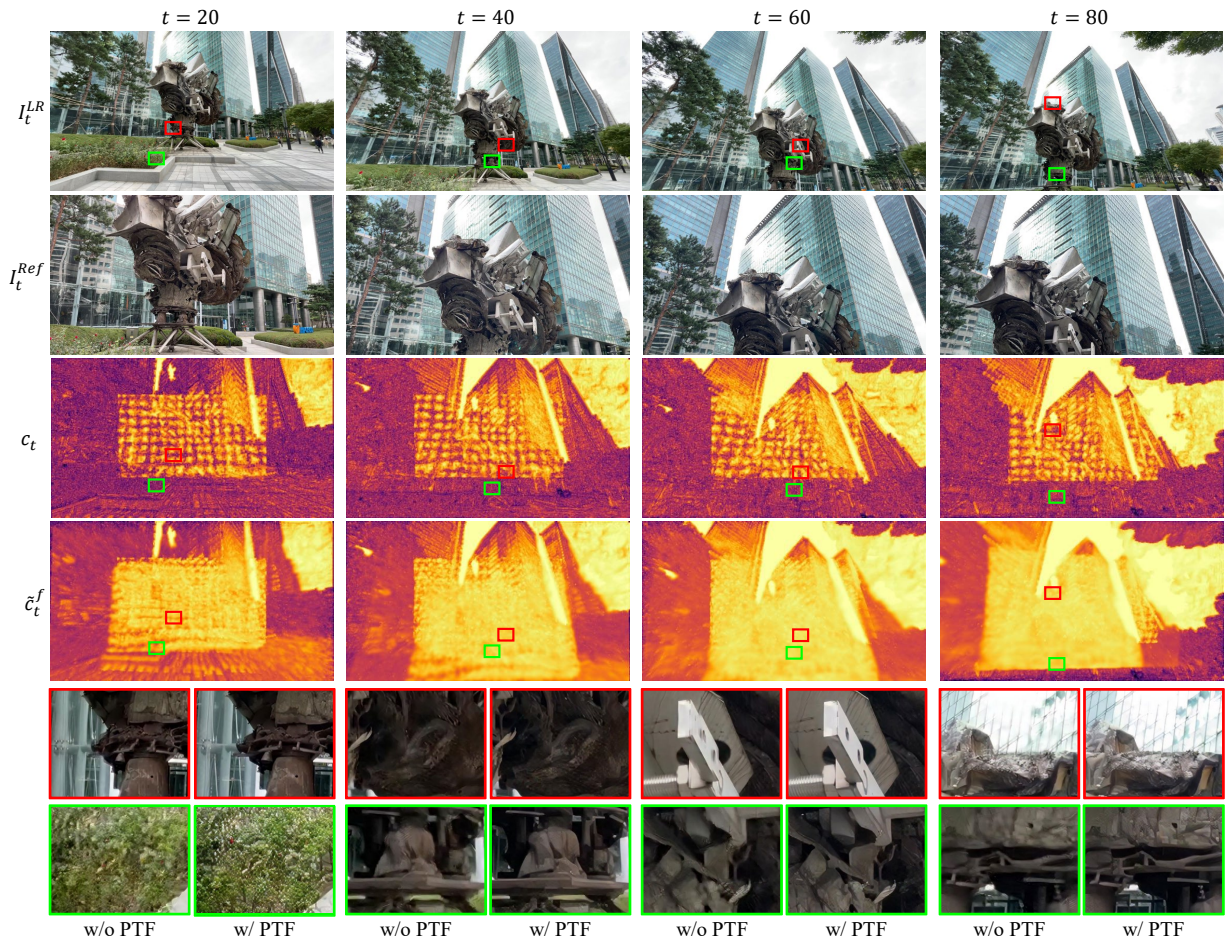

Figure 3. Effect of the proposed Propagative Temporal Fusion (PTF) module. c_t is the confidence map computed when the input LR frame I_t^{LR} is matched with the Ref frame I_t^{Ref} at the current time step. $\widetilde{c}_t^f$ is the accumulated matching confidence of the forward propagation branch. As can be seen in the figure, confidence values in $\widetilde{c}_t^f$ are accumulated following the motion in the video. The red and green boxes show zoomed-in cropped patches from the region inside and outside the overlapped FoV between I_t^{LR} and I_t^{Ref}, respectively. Note that the matching confidence maps are noisy due to HEVC/H.265 compression artifacts contained in video frames.

ral fusion module, we use a modified fusion module that does not utilize accumulated matching confidence $c_{t\pm1}^{\{f,b\}}$ during the fusion. Specifically, we modify Eq. 6 in the main paper for the modified fusion module:

$$h_t^{\{f,b\}} = \{\text{conv}(c_t) \otimes \text{conv}([\widetilde{h}_t^{Ref}, \widehat{h}_t^{\{f,b\}}])\} + \widehat{h}_t^{\{f,b\}}.$$

In Fig. 3, matching confidence c_t at the current time step shows a high matching score mainly concentrated in the region inside the overlapped FoV between an LR frame I_t^{LR} and a Ref frame I_t^{Ref} (the third row of the figure). However, in the accumulated matching confidence $\widetilde{c}_t^f$, the matching scores spread out following the motion of the video (the fourth row). As we can observe from the figure, compared to the model with the modified fusion module that does not utilize $\widetilde{c}_t^f$, the model with the propagative temporal fusion module restores more accurate structures and details for the region inside the overlapped FoV (red boxes in the figure), due to $\widetilde{c}_t^f$ providing better matched temporal Ref feature during the fusion. Moreover, the model with the propagative temporal fusion module shows finer details in reconstructed SR frames for the region outside the overlapped FoV (green boxes), as $\widetilde{c}_t^f$ guides temporal Ref features outside the overlapped region to be utilized during the fusion.

3. Effect of Bidirectional Scheme

The effect of the bidirectional scheme has been widely explored in previous VSR works [1, 2, 3]. To validate the effect on our method, we conduct the ablation study on the model with a unidirectional forward branch (Table 2), in addition to the ablation study reported in Sec. 5.1 of the main paper. In the table, models with bidirectional branches show better VSR quality than models with only a forward branch. Moreover, the proposed components (ℓ_{Mfid} and PTF in the table) improve the VSR performance for both unidirectional and bidirectional schemes.

$F_{\{f,b\}}$	ℓ_{Mfid}	PTF	PSNR↑	SSIM↑	Params (M)
			30.07	0.890	4.2653
	✓	✓	31.02	0.906	4.2656
✓			30.71	0.894	4.2768
✓	✓	✓	**31.68**	**0.914**	4.2772

Table 2. Ablation study including bidirectional branches. $F_{\{f,b\}}$ indicates the model with bidirectional branches. ℓ_{Mfid} is the model trained with the multi-Ref fidelity loss. PTF is the model with the propagative temporal fusion module.

Inter-frame		RA		PSNR↑	SSIM↑	Params (M)	MACs (T)
OF	PM	OF	PM				
	✓	✓		29.33	0.872	4.408	8.303
	✓		✓	31.68	0.920	3.059	9.364
✓		✓		29.35	0.878	5.627	1.975
✓			✓	**31.68**	**0.914**	4.277	**2.737**

Table 3. Effect of alignment modules for inter-frame alignment (Inter-frame) and reference alignment (RA). OF and PM indicate the optical flow [6] and patch-match-based alignment [7] methods, respectively. Our model adopts the combination in the last row. MACs are computed on 256×256 frames.

4. Effect of Alignment Methods

For the proposed RefVSR network, different alignment methods can be used for inter-frame and reference alignments. For inter-frame alignment, resolving local disparity is important [4], and we use flow-based alignment [6] that shows similar VSR quality to patch-match-based alignment, but with much smaller computational cost (2nd vs. 4th rows in Table 3). For reference alignment, establishing global correspondence is important [5], and we adopt patch-match-based alignment [7] that shows significant performance gain with slight computational overhead compared to flow-based alignment (3rd vs. 4th rows in the table).

5. Failure Case

Our network may fail to accurately utilize Ref frames for a resulting SR frame when matching between LR and Ref frames is inaccurate. Fig. 4 qualitatively shows the failure cases. In the figure, we show an LR patch (the second column), and its corresponding patches from a Ref frame (the third column) and a resulting SR frame (the last column). We can observe from the figure that when an LR frame does not contain enough cue needed for accurate matching with a Ref frame (*e.g.*, texture patterns), our model fails to accurately utilize Ref patches for recovering an SR frame.

6. Additional Qualitative Results

Figs. 5 and 6 show additional qualitative comparisons on 8K 4×SR results from real-world HD videos in the proposed RealMCVSR test set[2]. Note that all the compared models are trained with the proposed training strategy (Sec. 4 in the main paper).

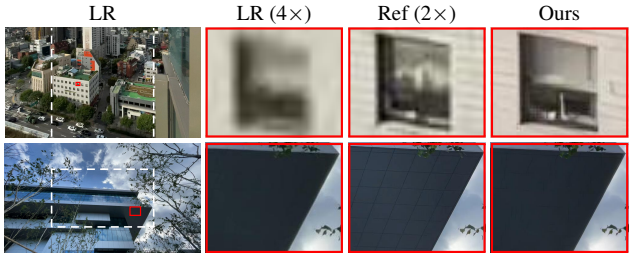

Figure 4. Failure cases. The first column shows LR real-world HD input frames. The other columns show zoomed-in cropped patches of bicubic upsampled LR and Ref input frames and SR frames resulting from our method, corresponding to the red box in an LR frame. In these examples, LR and Ref input frames do not contain enough cues needed for accurate matching, and the structure and detail in the results are not restored as those in Ref frames.

References

[1] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 5, 6

[2] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 3

[3] Yan Huang, Wei Wang, and Liang Wang. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):1015–1028, 2018. 3

[4] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[5] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *Proceedings of the Springer European Conference on Computer Vision (ECCV)*, 2020. 4

[6] Anurag Ranjan and Michael Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[7] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 4, 5, 6

[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 *Proceedings of the Springer European Conference on Computer Vision (ECCV)*, 2018. 5, 6

[2] More video results can be found at the following link: https://junyonglee.me/projects/RefVSR

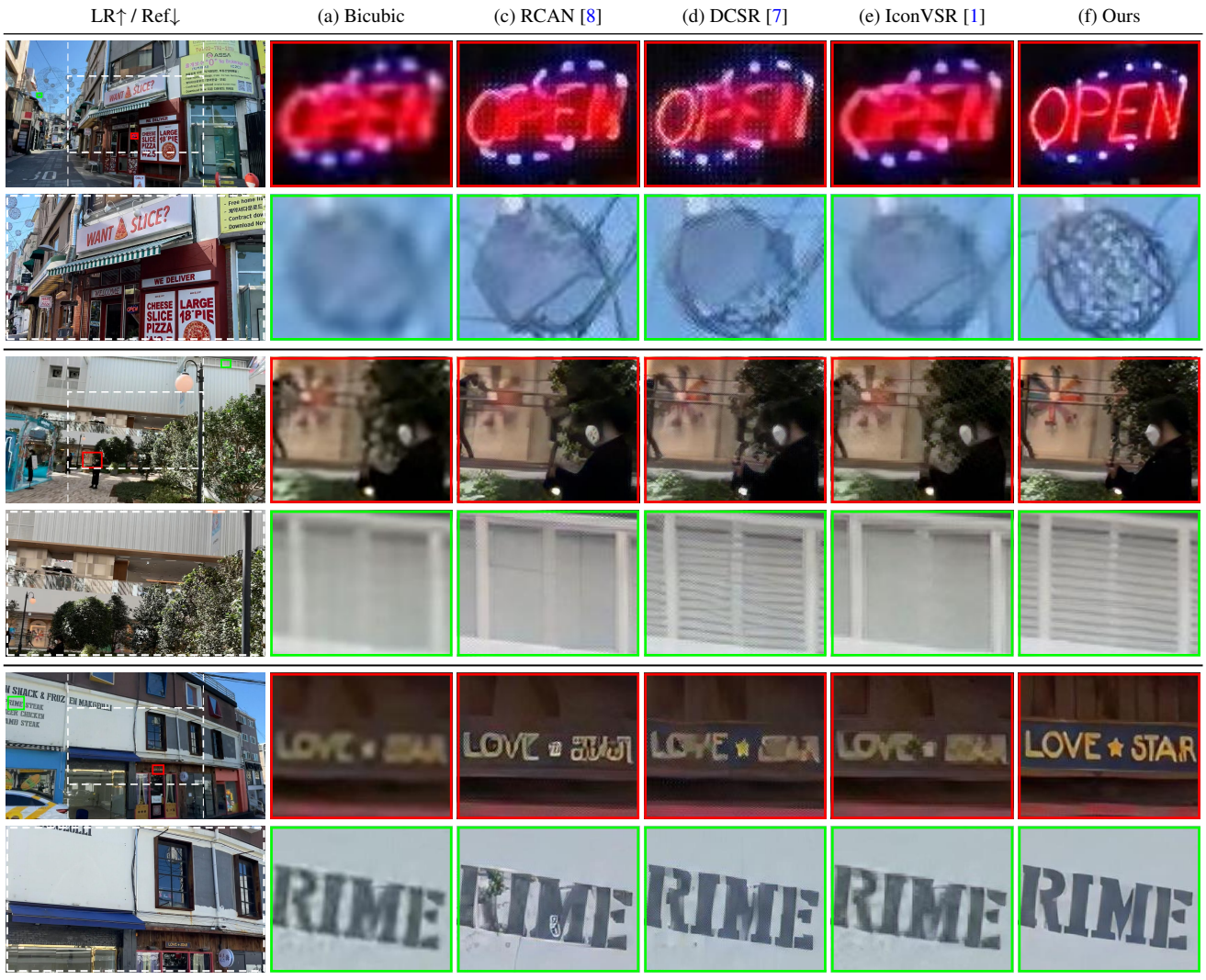

Figure 5. Qualitative comparison on 8K 4×SR video results from real-world HD videos in the proposed RealMCVSR dataset.

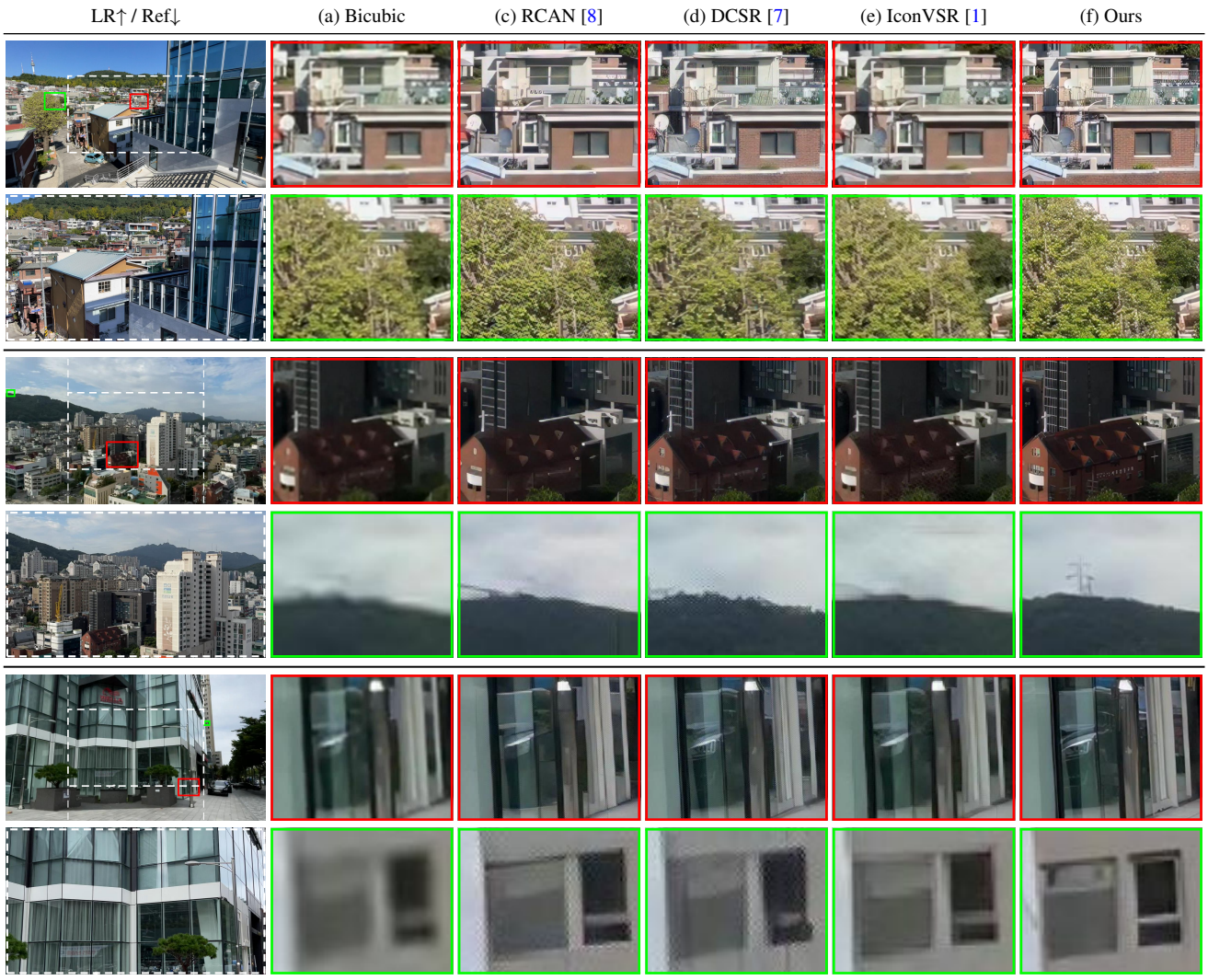

Figure 6. Qualitative comparison on 8K 4×SR video results from real-world HD videos in the proposed RealMCVSR dataset.