Supplementary Material : Efficient Reference-based Video Super-Resolution (ERVSR)

Youngrae Kim*[1], Jinsu Lim*[1], Hoonhee Cho*[2], Minji Lee*[1],
Dongman Lee†[1], Kuk-Jin Yoon†[2], and Ho-Jin Choi†[1]
[1]School of Computing, KAIST, [2]Mechanical Engineering, KAIST, Daejeon, South Korea
{youngrae.kim, j1n2u, gnsgnsgml, haewon_lee, kjyoon, hojinc}@kaist.ac.kr, dlee@cs.kaist.ac.kr

Due to the limitation of space in the main paper, we provide more quantitative and qualitative results and details of the architecture in the supplementary material.

1. Detailed implementation.

We provide the detailed diagram of our proposed model in Fig. A. The residual block is composed of a convolutional layer, ReLU activation, and a convolutional layer. The Attention-based Feature Align (AFA) module is composed of 4 attention modules. The number of heads and the latent dimension of the transformer module to 2 and 64, respectively. The bidirectional propagation module consists of a warping module and a residual block fusion module. And we use a pre-trained flow net to estimate the optical flow. The Attention-based Aggregation(AA) Upsampling is composed of 2 convolutional layers, $2\times$ Pixel Shuffle [4] upsampling layer, LeakyReLU activation, and one decoder attention. All convolutional layers are 2-dimensional and have kernel size of 3×3.

2. Extended experiments on the effect of the frame number.

We augmented the ablation study on the effect of frame number with additional results on a frame number of 11 in Table A. Lee *et al.*-IR-ℓ_1 [2] shows a greater decrease in PSNR (from 34.86 when using 13 frames to 34.02 when using 5 frames). On the other hand, our model shows relatively less performance degradation (from 34.44 when using 13 frames to 34.03 when using 5 frames). We can conclude from Table A that our model can perform better when the number of input frames and reference images are reduced for practical applications.

3. Extended qualitative results.

We provide extended qualitative results in Fig. B and Fig. C. As described in the main paper, our proposed ERVSR reduces blurry artifacts and restores high-frequency details such as letters.

Model	The number of frame for SR			
	5	7	11	13
BasicVSR++ [1]	32.56 0.9381	32.74 0.9404	32.80 0.9416	32.80 0.9416
Lee *et al.*-IR-ℓ_1 [2]	<u>34.02</u> <u>0.9516</u>	**34.36** **0.9548**	**34.79** **0.9584**	**34.86** **0.959**
ERVSR (Ours)	**34.03** **0.9534**	<u>34.15</u> <u>0.9541</u>	<u>34.38</u> <u>0.9562</u>	<u>34.44</u> <u>0.9567</u>

Table A. Extended quantitative evaluation of various networks on various the number of frames in window. Best results are **highlighted** and second-best results are <u>underlined</u>. 1st and 2nd row mean PSNR (dB) and SSIM scores, respectively.

4. Reproduction of MASA-SR [3]

We measured the PSNR and SSIM values of MASA-SR on RealMCVSR as 27.94 and 0.805, respectively. We guess that the reason for the low performance of the model is due to the implementation issue, so we did not report it in Table 1 of the main paper.

5. Video material.

We provide the video material to show the performance of our method here. We mainly compared the super-resolved video of ours and the bicubic interpolation.

6. Measurement of GPU memory usage.

All of the results regarding GPU memory usage on the main paper are measured on an NVIDIA A100-40GB. The memory usage for each window of video is measured as the peak GPU memory usage during inference.

References

[1] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced

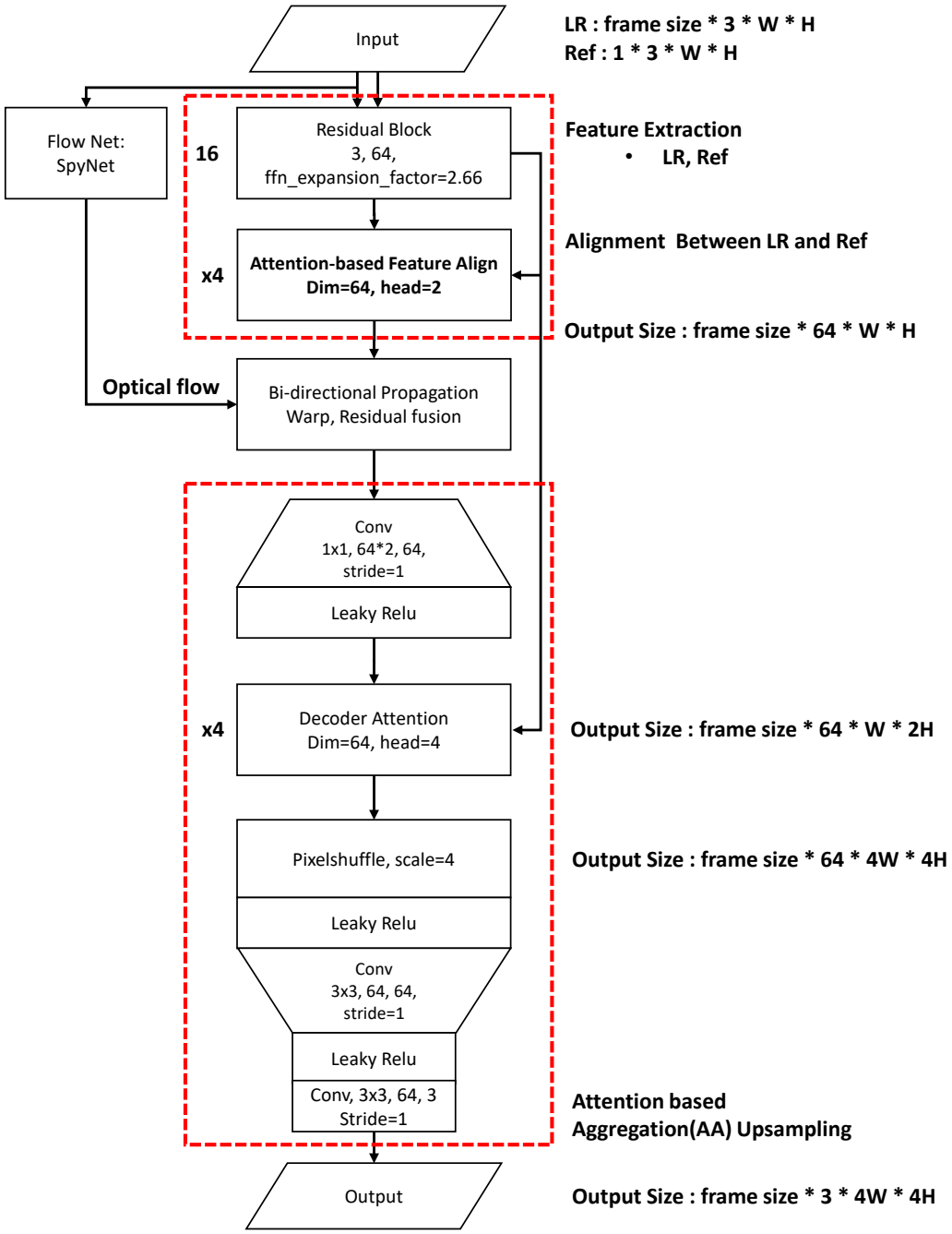

Figure A. Flowchart of the proposed ERVSR framework.

propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5972–5981, 2022.

[2] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. *arXiv preprint arXiv:2203.14537*, 2022.

[3] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6368–6377, 2021.

[4] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

Figure B. Qualitative comparison of our methods with previous works. . For better comparison, we zoomed some area of the images in. (a): LR input, (b): Bicubic interpolation, (c): BasicVSR++ [1], (d): Lee *et al*. [2], and (e): Ours, respectively.

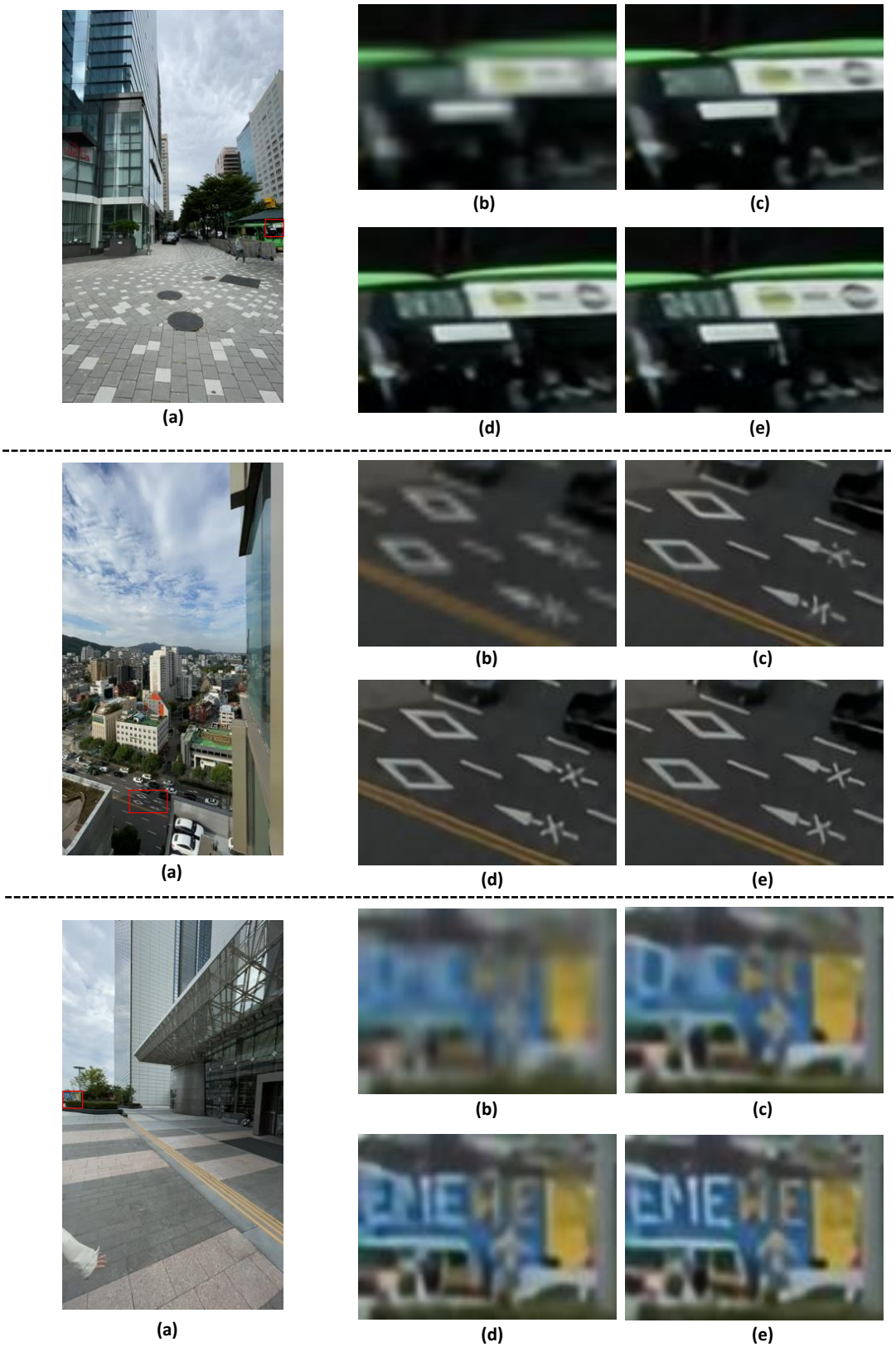

Figure C. Qualitative comparison of our methods with previous works. For better comparison, we zoomed some area of the images in. (a): LR input, (b): Bicubic interpolation, (c): BasicVSR++ [1], (d): Lee *et al*. [2], and (e): Ours, respectively.