Efficient Reference-based Video Super-Resolution (ERVSR): Single Reference Image Is All You Need

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

Abstract

Reference-based video super-resolution (RefVSR) is a promising domain of super-resolution that recovers high-frequency textures of a video using reference video. The multiple cameras with different focal lengths in mobile devices aid recent works in RefVSR, which aim to super-resolve a low-resolution ultra-wide video by utilizing wide-angle videos. Previous works in RefVSR used all reference frames of a Ref video at each time step for the super-resolution of low-resolution videos. However, computation on higher-resolution images increases the runtime and memory consumption, hence hinders the practical application of RefVSR. To solve this problem, we propose an Efficient Reference-based Video Super-Resolution (ERVSR) that exploits a single reference frame to super-resolve whole low-resolution video frames. We introduce an attention-based feature align module and an aggregation upsampling module that attends LR features using the correlation between the reference and LR frames. The proposed ERVSR achieves 12× faster speed, 1/4 memory consumption than previous state-of-the-art RefVSR networks, and competitive performance on the RealMCVSR dataset while using a single reference image.

1. Introduction

Super-resolution (SR) aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) image. To recover the high-frequency details, reference-based super-resolution (RefSR) approaches [2, 20, 22, 25, 27, 30, 32, 33] utilize additional reference images, *e.g.* web-crawled high-resolution images [20] and image taken from slightly different viewpoints [16]. The super-resolved image, incorporating high-frequency detail from the reference images, is precise and visually pleasing compared to synthesized textures [33]. In addition, since many recent mobile phones are equipped with dual or triple cameras, the demand for RefSR is increasing.

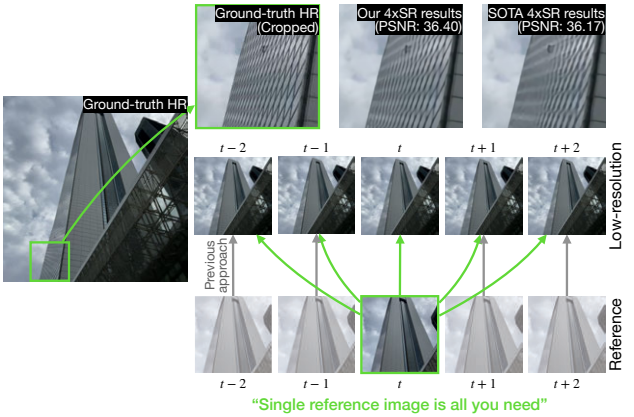

Figure 1. Illustration of input configuration for ERVSR. Unlike previous RefVSR, we use only a single Ref frame (a center frame within the time window) for alignment. Note that there is a lot of overlap between Ref frames.

Reference-based video super-resolution (RefVSR) inherits the advantages of both RefSR and Video Super-Resolution (VSR), which leverages the temporal information in videos. Lee *et al*. [16] presented the dataset for RefVSR, called RealMCVSR dataset, consisting of ultra-wide and wide-angle videos captured with asymmetric multi-cameras in smartphones. As these cameras capture the scene with different field of view (FoV), the wide-angle image of smaller FoV can be used as a Ref image to super-resolve the ultra-wide angle image of larger FoV. To recover the HR video, Lee *et al*. [16] computes the correlation between every Ref and LR frame. Guided by the additional information, their approach outperforms VSR and RefSR models. However, their best-performing model consumes 19GB of memory and takes up to 16× longer inference time than the other VSR approaches [4, 5]. This large memory consumption and computation time of RefVSR mainly

*Equal contribution. †Co-corresponding authors.
Code: https://github.com/haewonc/ERVSR

comes from the computation of correlation between Ref and LR frames at every time step. Here the question arises: Can we achieve competitive super-resolved video results with less memory consumption while utilizing Ref data? In this paper, we show that competitive performance can be obtained by designing efficient and effective modules for the alignment between Ref and LR data. Unlike the existing work, we do not use Ref frames at every time step, but only use a single Ref frame in the center of the window. As shown in Fig. 1, there is a large overlap between consecutive wide-angle Ref frames. From this point of view, we observe that a single Ref image is sufficient as long as such high-frequency information can be used effectively.

In this paper, we introduce an Efficient Reference-based Video Super-Resolution (ERVSR) framework that super-resolves HR videos from LR videos using a single Ref image. We propose an attention-based feature align (AFA) module which aligns the center Ref frame with the LR frame, thus propagating the Ref information without explicit alignment. Furthermore, high-frequency features of the Ref frame are transferred to every low-resolution feature by attention-based aggregation (AA) upsampling. The AFA module and AA upsampling both exploit widely used attention [21] mechanism where reference features are query and LR features are key and value. Benefiting from attention modules that allow the network to fully exploit a single Ref image, ERVSR achieves competitive performance in the RealMCVSR dataset [16] compared to models that use reference images for every time step [11, 30, 17, 16] with faster speed and less GPU memory.

Our contributions are summarized as follows:

- We propose the ERVSR, which is the first work to tackle the large computation of RefVSR. We successfully optimized the accuracy-efficiency trade-off of RefVSR, opening the possibility of using RefVSR on a mobile device.
- We exploit the attention-based similarity computation and fusion in RefVSR. We also show that a single reference image is enough to recover high-frequency details of the entire video.

2. Related Works

2.1. Reference-based Image Super-resolution

Reference-based image super-resolution methods can be categorized into two: texture transfer methods [32, 30, 27] and the methods that exploit alignment [20, 23]. Texture transfer methods usually unfold images to patches and compute the similarity of reference patches for each LR patch. To measure the similarity, RefSR-Net [32], and TTSR [27] use the inner product between the raw pixels of patches, whereas SRNTT [30] use the inner product between the features of patches. On the other hand, some works [32, 30] concatenate most similar reference patches for each LR patch and fuse the patches using convolution layers.

Contrastingly, CrossNet [33] estimate the optical flow and warp the reference image to the LR image. SSEN [20] align reference features with LR features using stacked deformable convolution. DCSR [23] exploits both texture transfer and alignment; It first finds a reference patch that maximizes cosine similarity for each low-resolution patch and warp reference patch using estimated spatial transformation [11].

2.2. Video Super-Resolution

Many VSR methods perform the alignment by estimating the optical flow field between the target frame and neighboring frames [3, 13, 26], then use convolutions or recurrent networks to fuse features from aligned frames [3, 7]. EDVR [24] aligns the features of each frame using deformable convolution, then fuses them via attention mechanism in both spatial and temporal manner. BasicVSR [4] established the usage of the bidirectional propagation scheme in VSR by showing it maximizes the information gathering. IconVSR [4] extends BasicVSR by adding the additional feature extractor and the coupled propagation mechanism, which interconnect the propagation modules to exploit further information in the sequences. BasicVSR++ [5] extends BasicVSR with second-order grid propagation and flow-guided deformable alignment.

2.3. Reference-based Video Super-Resolution

EFENet [31] utilizes the first frame of a high-resolution ground-truth video as a reference to super-resolve a low-resolution (LR) video. ERVSR is different from EFENet in two folds. First, ERVSR only computes the correlation between the single Ref frame and the single LR frame, whereas EFENet computes flow maps between the Ref and every LR frame using a shared flow estimator. Second, EFENet requires the guidance of an HR video, which is impractical in a real-world scenario.

Lee *et al*. [16] propose a practical setup that captures the Ref video with an asymmetric multi-camera in a smartphone. They follow a bidirectional propagation scheme, with reference alignment [20] and propagation module in each recurrent cell. Then the aligned reference features are fused with temporally aggregated features using convolutional layers.

Our proposed framework ERVSR differs from previous works in two folds. First, ERVSR does not use the Ref frame every time step, but only one frame in the center time step. There is a lot of overlap between the Ref video frames, and we believe it is not necessary to utilize every Ref frames. Second, existing methods use the reference alignment module proposed in [22] to obtain Ref features aligned to LR frames. Instead of patch-wise calculation of

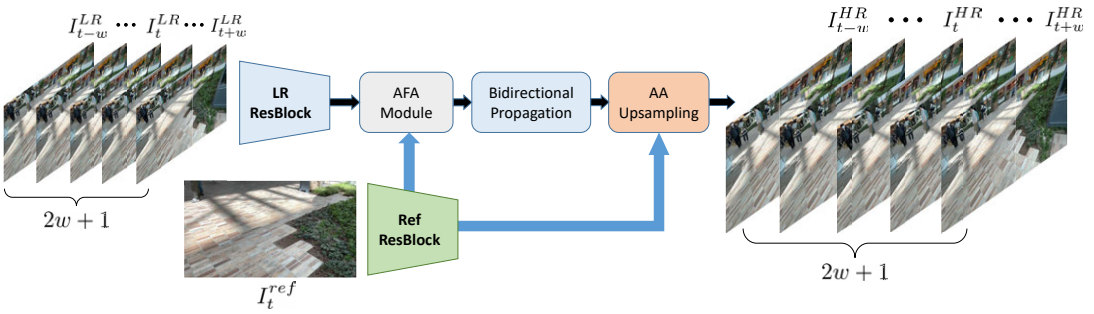

Figure 2. Overall framework of our proposed method. $2w+1$ denotes window size, which means the number of frames, and t denotes the center time step.

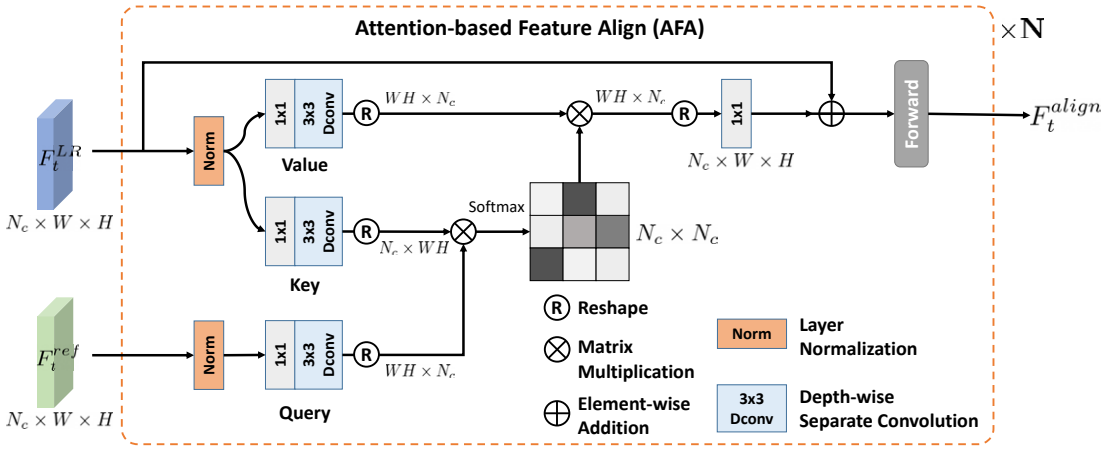

Figure 3. The proposed Attention-based Feature Align (AFA) module.

similarity and alignment, which is computationally heavy, ERVSR computes the similarity between LR and Ref feature using learnable projection [21]. Note that ERVSR is the first approach to exploit attention-based similarity computation and fusion in RefVSR.

3. Proposed Method

In this section, we present ERVSR, an end-to-end neural network for efficient reference-based video super-resolution. As shown in Fig. 2, our proposed network consists of three components: Attention-based Feature Align (AFA), Bidirectional Propagation, and Attention-based Aggregation (AA) upsampling module. The AFA module aligns the LR frame at center time t, the Bidirectional Propagation propagates the temporal information to the other time steps, and the AA module transfers the features from the Ref image to upsample the LR feature.

Let LR frames $\{I_{t-w}^{LR}, \cdots, I_t^{LR}, \cdots, I_{t+w}^{LR}\} \in \mathbb{R}^{3\times W\times H}$ and a reference frame at center time t, $I_t^{ref} \in \mathbb{R}^{3\times W\times H}$, where $2w+1$ denotes size of the window and W, H denote the width and height in spatial dimension of LR frame, respectively. AFA module aligns the LR frame at center time t with the reference frame I_t^{ref}. Fig. 3 illustrates the overview of the AFA module. I_t^{LR} and I_t^{ref} are first deeply represented by residual layers [8], resulting $F_t^{LR}, F_t^{ref} \in \mathbb{R}^{N_c\times W\times H}$ respectively, where N_c denotes the number of channel dimensions. Since the properties of the features to be extracted from the LR frame and Ref frame are different, the two residual blocks (ResBlock) do not share weights.

3.1. Attention-based Feature Align (AFA) Module

Extracted features are normalized through normalization layer [1] and projected to query **Q**, key **K**, and value **V** [21]. We calculate the **Q**, **K**, **V** as follows:

$$\begin{aligned} \mathbf{Q} &= P_Q(F_t^{ref}) \in \mathbb{R}^{WH\times N_c}, \\ \mathbf{K} &= P_K(F_t^{LR}) \in \mathbb{R}^{N_c\times WH}, \\ \mathbf{V} &= P_V(F_t^{LR}) \in \mathbb{R}^{WH\times N_c}, \end{aligned} \tag{1}$$

where P_Q, P_K, P_V denote the projection consisting of 1×1 convolutional layer, efficient depth-wise separate convolutional layer [9], and reshaping operation. AFA module then computes the correlation between projected reference fea-

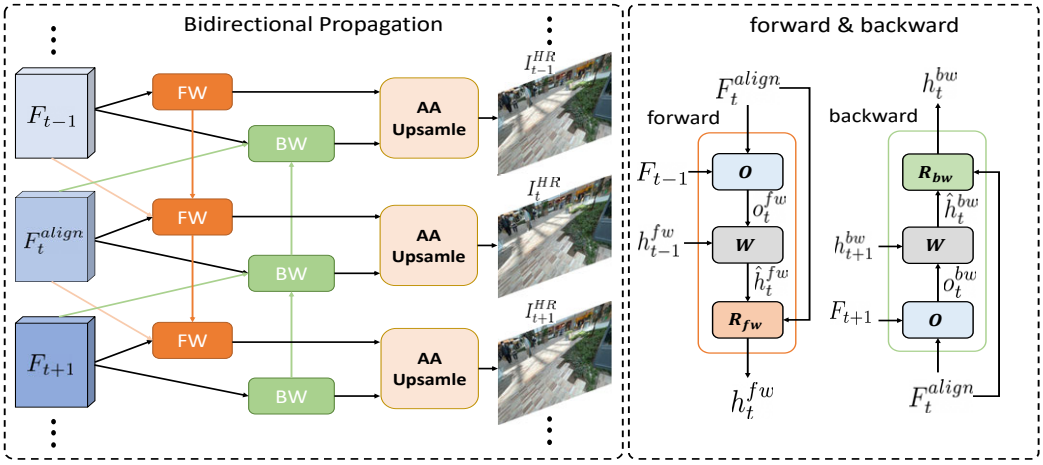

Figure 4. The overview of bidirectional propagation. The bidirectional propagation consists of cascaded forward (FW) and backward (BW) modules.

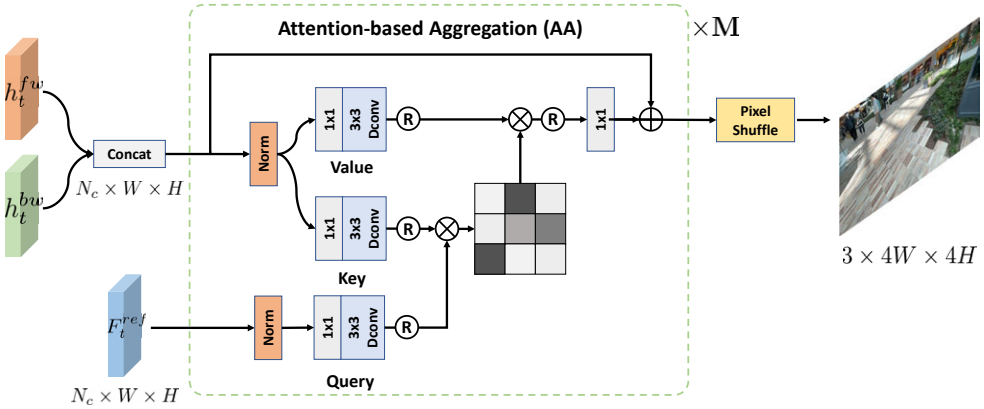

Figure 5. The proposed Attention-based Aggregation (AA) upsampling module.

ture **Q** and projected LR feature **K**, resulting in an attention map $\mathcal{A}$:

$$\mathcal{A} = \text{Softmax}(\mathbf{KQ}) \in \mathbb{R}^{N_c \times N_c}. \quad (2)$$

Projected LR feature **V** is attended using the information of reference, $\mathcal{A}$, resulting an aligned LR feature F_t^{align} where out projection P_O includes 1×1 convolutional layer and reshaping operation:

$$F_t^{align} = P_O(\mathbf{V}\mathcal{A}) + F_t^{LR}. \quad (3)$$

In attention calculation, we also utilize depth-wise separate convolutions for efficient computation as demonstrated in Fig. 3. For the forward network, we adopt the Gated-Dconv Feed-forward Network (GDFN) in [28]. The AFA module is repeated **N** times. Using attention mechanism, ERVSR can fully exploit Ref features, not only the most similar features, and benefit from repetitive textures.

By extracting and propagating F_t^{align} instead of the low-resolution feature, we can exploit information from reference efficiently. As proposed in [16], aligning all LR frames with reference frames would guide more information to the network, but it leads to a huge computational cost. Therefore, LR frames other than I_t^{LR}, are extracted to features by residual block, not the AFA module, resulting F_i where $i \in \{t-w, \cdots, t+w\}$ and $i \notin \{t\}$.

3.2. Bidirectional Propagation

The generally used approach to propagate the temporal information is the unidirectional propagation method [10, 6], where the information is sequentially propagated from the first frame to the last frame. However, the problem with this method is that the first frame receives no information from the other frames. In other words, there can be an information imbalance. Therefore, instead of a unidirectional propagation approach, we adopted the bidirectional method [4] to propagate temporal information in each frame.

As shown in Fig. 4, given consecutive LR features $\{F_{t-1}, F_t^{align}, F_{t+1}\}$, we can obtain the forward feature h_t^{fw} and backward feature h_t^{bw} where fw and bw denote

	Model	Reference usage	PSNR (dB)	SSIM	Inference time (ms/frame)	GPU memory usage (GB)
Image SR	Bicubic	No	26.65	0.8	N/A	N/A
	SRGAN [15]	No	29.38	0.877	-	-
	RCAN-ℓ_1 [29]	No	31.07	0.915	-	-
Reference Image SR	TTSR [27]	All time steps	30.31	0.905	-	-
	TTSR-ℓ_1 [27]	All time steps	30.83	0.911	-	-
	C^2-Matching-rec* [12]	All time steps	30.58	0.887	-	-
	DCSR [23]	All time steps	30.63	0.895	-	-
	DCSR-ℓ_1 [23]	All time steps	32.43	0.933	-	-
Video SR	EDVR [24]	No	33.26	0.946	512.82	6.474
	BasicVSR [4]	No	33.66	**0.951**	**77.72**	**4.792**
	IconVSR [4]	No	**33.80**	**0.951**	**102.03**	**5.097**
	BasicVSR++* [5]	No	32.80	0.942	**100.19**	6.452
Reference Video SR	Lee *et al*.-IR [16]	All time steps	31.73	0.916	1204.61	19.089
	Lee *et al*.-IR-ℓ_1 [16]	All time steps	**34.86**	**0.959**	1204.61	19.089
	ERVSR (Ours)	1 per window	**34.44**	**0.957**	107.02	**5.073**

Table 1. Quantitative evaluation for the 13 frames per window on the RealMCVSR dataset. - indicates that the information is not provided in that paper. The **best** and **top-3** results are highlighted. * denotes the trained by ours.

the forward and backward module respectively:

$$\begin{aligned} h_t^{fw} &= fw(F_t^{align}, F_{t-1}, h_{t-1}^{fw}), \\ h_t^{bw} &= bw(F_t^{align}, F_{t+1}, h_{t+1}^{bw}). \end{aligned} \tag{4}$$

Each module exploits the flow-based methods for spatial alignment as:

$$\begin{aligned} s_t^{\{fw,bw\}} &= O(F_t^{align}, F_{t\pm1}), \\ \hat{h}_t^{\{fw,bw\}} &= W(h_{i\pm1}^{\{fw,bw\}}, s_i^{\{fw,bw\}}), \\ h_t^{\{fw,bw\}} &= R_{\{fw,bw\}}(F_t^{align}, \hat{h}_i^{\{fw,bw\}}), \end{aligned} \tag{5}$$

where O and W denote the flow estimation and explicit feature-level warping, respectively. Here, $R_{\{fw,bw\}}$ denotes a stack of residuals for each forward and backward warping.

3.3. Attention-based Aggregation (AA) Upsampling

We propose the AA upsampling module to transfer the high-frequency feature of reference while upsampling the LR feature. As illustrated in Fig. 5, given the forward aggregated feature and backward aggregated feature from bidirectional propagation, the final high-resolution (HR) frames are obtained by:

$$I_i^{HR} = U(h_i^{fw}, h_i^{bw}), \quad i \in \{t-w, \cdots, t+w\}, \tag{6}$$

where U denotes the upsampling module consisting of pixel-shuffle [19], attention mechanism using AA attention map obtained by computing correlation of LR feature and reference feature. In the same way as the AFA Module, the attention map is calculated using depth-wise separate convolution layer and then scaled up using pixel shuffle. In contrast to the AFA module, AA upsampling module transfers the information from the Ref feature to LR feature for every time step. The AA module is repeated **M** times.

4. Experiments

4.1. Dataset

Our model is trained and tested on the RealMCVSR dataset [16]. RealMCVSR dataset provides real-world HD video triplets concurrently recorded by Apple iPhone 12 Pro Max equipped with triple cameras having fixed focal lengths: ultra-wide (30mm), wide-angle (59mm), and telephoto (147mm). The video triplets are split into training, validation, and testing sets, each of which has 137, 8, and 16 triplets of 19,426, 1,141, and 2,540 frames, respectively. Following the [16], we set the ultra-wide and wide-angle as LR frames and Ref frames, respectively.

4.2. Implementation Details

The network and experiments are implemented using the Pytorch framework. We use NVIDIA A100-40GB when measuring the inference time and GPU memory usage. For the training, we use ℓ_1 loss as a loss function and Adam [14] optimizer. For each iteration, we randomly sample batches of frame triplets from the RealMCVSR training set while setting the batch size 32. We used the pretrained optical flow network [18] in bidirectional propagation. We trained our model to super-resolve a $4\times$ bicubic downsampled LR

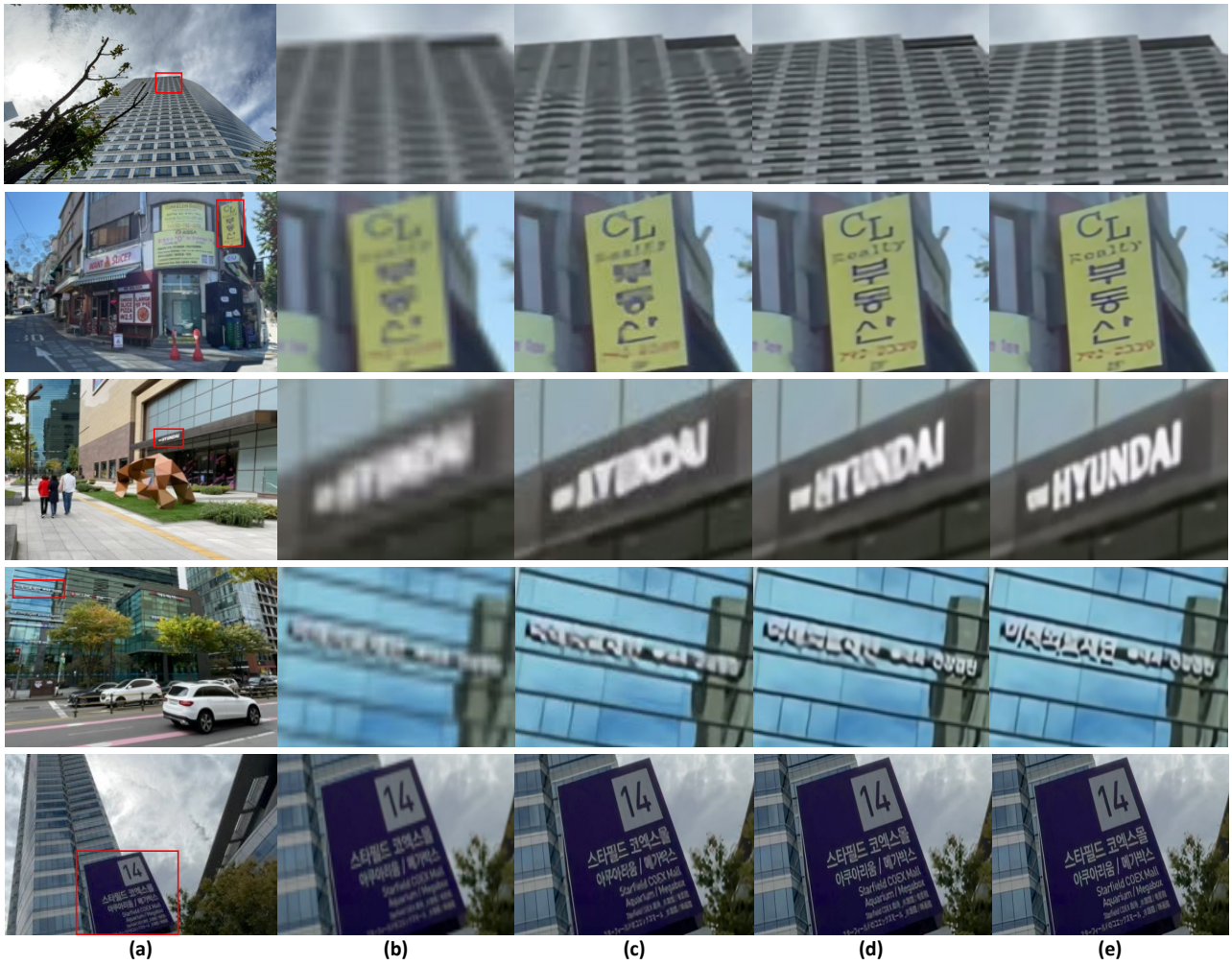

Figure 6. Qualitative comparison of our methods with previous works. For clarity, the magnified parts of the resultant images are zoomed. From left to right, (a): LR input, (b): Bicubic interpolation, (c): BasicVSR++ [5], (d):Lee *et al.*-IR-$\ell 1$ [16], and (e): Ours, respectively.

ultra-wide video using a wide-angle video frame as a Ref image. In the training phase, the ultra-wide LR frames and wide-angle Ref frames are cropped to 64 $\times$ 64. We set the number of layers to $\mathbf{N} = \mathbf{M} = 4$ (in sec 3.1, 3.3) equally in our network and the window size to 13 as a default size. More details about the experiment can be found in our supplementary materials.

4.3. Quantitative Comparison

Table 1 shows a quantitative comparison on RealMCVSR dataset. We used 4$\times$ bicubic downsampled low-resolution ultra-wide video and wide-angle reference video as input. We compared ERVSR with previous works: SRGAN [15], RCAN [29], TTSR [27], DCSR [23], EDVR [24], BasicVSR [4], BasicVSR++ [5], Lee *et al.* [16]. We demonstrated the comparisons on PSNR, SSIM, inference time per frame, and GPU memory usage. Models trained only on ℓ_1 loss function are indicated with ℓ_1.

Our ERVSR outperforms the previous single image SR (SISR), RefSR, and VSR methods and achieves the second-best in both PSNR and SSIM among several models. The performance difference between Lee *et al.*-IR-$\ell 1$ [16] (best performing model among several models in [16]) and our model is acceptable considering that our model is 12 times faster, consumes 3.8 times smaller GPU memory usage, and uses only one Ref image at inference time. Even when only one Ref image is given, we show that our method efficiently and effectively exploits and transfers the high-frequency information from the Ref image by using our novel attention-based feature align (AFA) module (in sec 3.1) and attention-based aggregation (AA) upsampling module (in sec 3.3). The efficiency of ERVSR is also competitive with the methods in VSR in inference time per frame and GPU memory usage, while PSNR and SSIM values are much higher. This indicates that the efficiency of our method is comparable with the VSR models, even incorporating the Ref frame in

Model	PSNR (dB)	SSIM
Baseline	31.88	0.931
Baseline + AFA	32.11	0.932
Baseline + AA	32.17	0.933
ERVSR (Baseline + AFA + AA)	**34.44**	**0.957**

Table 2. Quantitative ablation study of our proposed components.

the network. Therefore, our proposed ERVSR successfully reduces the accuracy-efficiency of the trade-off of RefVSR.

4.4. Qualitative Comparison

In Fig. 6, we show the super-resolved results of Bicubic, BasicVSR++ [5], and Lee *et al*.-IR-$\ell 1$ [16] trained on RealMCVSR dataset. Non-reference-based SR methods such as Bicubic and BasicVSR++ tend to produce blurred textures and letters. Contrastingly, RefVSR methods benefit from the Ref information and better super-resolve the details such as letters. Also, our proposed ERVSR shows competitive visual quality with Lee *et al*.-IR-$\ell 1$ [16], even though our method uses only one Ref image per window.

Lee *et al*.-IR-$\ell 1$ [16] often produces the blurry artifact, especially in small-sized features and non-overlapping FoV area. It is due to the misalignment between the LR image and the Ref image since it explicitly calculates the similarity map. Benefiting from the attention-based mechanism, our ERVSR can provide better visual results on small-sized features by utilizing the feature-level alignment. Also, since our AA module transfers the high-resolution detail to LR from the Ref feature for overlapping as well as non-overlapping regions, ERVSR can make better results in the non-overlapping FOV area.

4.5. Ablation Studies

1. Contribution of each component of ERVSR to the performance. We conducted quantitative ablation studies to analyze the effect of each of our proposed modules, AFA and AA. We set the baseline network as a model solely consisting of bidirectional propagation without AFA and AA modules. As demonstrated in Table 2, the models with our each proposed module show better PSNR and SSIM than the baseline network. AA module has a more remarkable performance improvement than the AFA module since it is effective in directly transferring the LR feature to temporally aligned features by bidirectional propagation. We find that the performance gain is significant when the network uses both AA and AFA modules. We hypothesize that the network needs to learn the encoded and decoded feature simultaneously. Using the solely consisted module becomes extremely difficult to transfer the high-resolution information from the Ref feature to the LR feature. On the other hand, using both AA and AFA module at encode- and decode-level achieve to align the Ref to LR feature without explicit alignment.

As shown in Fig. 7, we also conducted qualitative ablation studies to show the visual effect of each component on SR. The Baseline cannot restore the details such as letters or repetitive patterns. The Baseline + AFA shows reduced distortion than the Baseline but still produces blurry artifacts. Our ERVSR model with both AFA and AA modules achieves the best visual quality with reduced blurry artifacts and distortion and accurate restoration of edges.

2. Can reference at the center frame guide the LR frames with a large time difference? This question is crucial since we use a single reference frame per window. To answer this question, we compared the PSNR of each super-resolved video frame of ERVSR and BasicVSR++ with respect to its time difference from the center frame (See Fig. 8). Performance degradation decreases as the time difference increases, but such an amount of degradation is also observed in BasicVSR++. If reference at the center frame cannot effectively guide the LR frames with a large time difference, the performance degradation will be greater than that of the VSR. However, the performance gap between ERVSR and BasicVSR++ is constant, which implies that the details of the reference center frame manage to guide the feature extraction and upsampling of an LR frame even with a large time interval.

3. Effect of frame number in performance. VSR models can benefit from the increased number of input LR frames since the feature can be better refined from the bidirectional flow. The advantageous effect of increased frame number escalates in RefVSR models that exploit reference frames for every time step. In this section, we show the comparison of PSNR and SSIM for the number of frames in the window in Table 3. As the frame number decrease, the performance gap between our model and Lee *et al*.-IR-ℓ_1 [16] decreases, and finally, our model outperforms it when the frame number is 5. Knowing that Lee *et al*.-IR-ℓ_1 [16] still exploits four more reference frames than ERVSR when the frame number is 5, such results imply that our attention-based modules are effective design for the utilization of the reference details.

5. Conclusion

In this work, we propose efficient reference-based video super-resolution (ERVSR) using only a single reference image. To this end, we first propose the attention-based approach to compute similarity using learnable projections in RefVSR. To evaluate our method, we train and evaluate our network in a large-scale benchmark in [16]. Our model outperforms the state-of-the-art image SR, reference-based image SR, and VSR methods in both qualitatively and quantitatively. We accelerate the inference time for $\times 12$ faster while using GPU memory to $\times 3.8$ efficiently

Figure 7. Qualitative results of ablation studies on each component. For clarity, the magnified parts of the resultant images are zoomed in. From left to right, (a): LR input, (b): Baseline, (c): Baseline + AFA, (d): Baseline + AA, and (e): ERVSR (Baseline + AFA + AA), respectively.

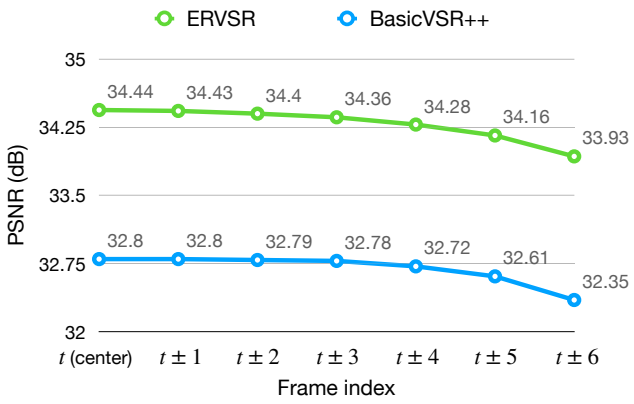

Figure 8. Comparison on PSNR of each video frame.

Model	The number of frame for SR 5	7	13
BasicVSR++ [5]	32.56 0.9381	32.74 0.9404	32.80 0.9416
Lee *et al.*-IR-ℓ_1 [16]	34.02 0.9516	**34.36** **0.9548**	**34.86** **0.959**
ERVSR (Ours)	**34.03** **0.9534**	34.15 0.9541	34.44 0.9567

Table 3. Quantitative evaluation of various networks on various the number of frames in window. Best results are **highlighted** and second-best results are underlined. 1st and 2nd row mean PSNR (db) and SSIM scores, respectively.

where only one reference frame in a video is used, compared with Lee *et al.*-IR-ℓ_1 [16], which is the state-of-the-art approach in RefVSR domain. In short, we reduce the accuracy-efficiency trade-off of RefVSR, opening the possibility of using RefVSR in real-time problems.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01126, Self-learning based Autonomic IoT Edge Computing). (50%)
This research was supported and funded by the Korean National Police Agency. [Project Name: XR Counter-Terrorism Education and Training Test Bed Establishment / Project Number: PR08-04-000-21] (50%)

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Vivek Boominathan, Kaushik Mitra, and Ashok Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In *2014 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2014.

[3] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Con-*

ference on Computer Vision and Pattern Recognition, pages 4778–4787, 2017.

[4] 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.

[5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5972–5981, 2022.

[6] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019.

[7] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[10] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *European Conference on Computer Vision*, pages 645–660. Springer, 2020.

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[12] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2103–2112, 2021.

[13] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 106–122, 2018.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[16] 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.

[17] 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 (CVPR)*, pages 6368–6377, June 2021.

[18] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017.

[19] 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.

[20] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8434, 2020.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[22] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2001–2010, 2021.

[23] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2001–2010, October 2021.

[24] 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/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[25] Yanchun Xie, Jimin Xiao, Mingjie Sun, Chao Yao, and Kaizhu Huang. Feature representation matters: End-to-end learning for reference-based image super-resolution. In *European Conference on Computer Vision*, pages 230–245. Springer, 2020.

[26] 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.

[27] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020.

[28] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang.

Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.

[29] 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 European conference on computer vision (ECCV)*, pages 286–301, 2018.

[30] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7982–7991, 2019.

[31] Yaping Zhao, Mengqi Ji, Ruqi Huang, Bin Wang, and Shengjin Wang. Efenet: Reference-based video super-resolution with enhanced flow estimation. In *CAAI International Conference on Artificial Intelligence*, pages 371–383. Springer, 2021.

[32] Haitian Zheng, Mengqi Ji, Lei Han, Ziwei Xu, Haoqian Wang, Yebin Liu, and Lu Fang. Learning cross-scale correspondence and patch-based synthesis for reference-based super-resolution. In *BMVC*, volume 1, page 2, 2017.

[33] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European conference on computer vision (ECCV)*, pages 88–104, 2018.