LDIP: Long Distance Information Propagation for Video Super-Resolution

Michael Bernasconi[1,2] Abdelaziz Djelouah[2] Yang Zhang[2] Markus Gross[1,2] Christopher Schroers[2]

[1]ETH Zürich [2]DisneyResearch|Studios

michael.bernasconi@inf.ethz.ch abdelaziz.djelouah@disney.com

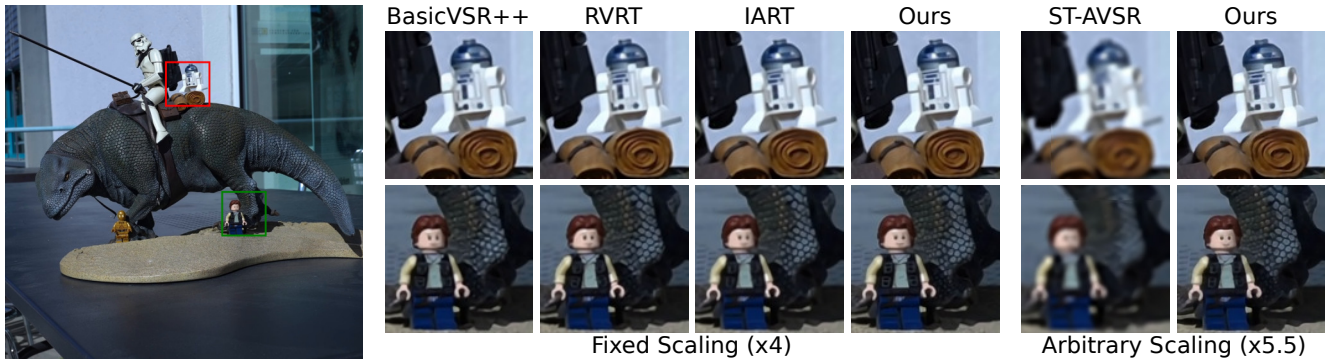

Figure 1. Visual comparison of our method against SOTA fixed scaling and arbitrary scaling Video Super-Resolution (VSR) methods. Thanks to the long range temporal propagation, our model extracts more information from the input video and achieves significantly higher quality output frames.

Abstract

Video super-resolution (VSR) methods typically exploit information across multiple frames to achieve high quality upscaling, with recent approaches demonstrating impressive performance. Nevertheless, challenges remain, particularly in effectively leveraging information over long distances. To address this limitation in VSR, we propose a strategy for long distance information propagation with a flexible fusion module that can optionally also assimilate information from additional high resolution reference images. We design our overall approach such that it can leverage existing pre-trained VSR backbones and adapt the feature upscaling module to support arbitrary scaling factors. Our experiments demonstrate that we can achieve state-of-the-art results on perceptual metrics and deliver more visually pleasing results compared to existing solutions.

1. Introduction

Upscaling low-resolution video content is a fundamental task in computer vision that has been studied for several decades. On a high level, there are two different strategies in which recent deep learning based methods were able to increase the quality: On the one hand by designing better and larger model architectures, and on the other hand by utilizing more of the available information. Regarding the second point, consider *single image super-resolution* (SISR) methods [12, 14, 28, 31, 33] . While they can be used to upscale a video in a frame-by-frame manner, they will obviously ignore any temporal information in the sequence. Therefore, methods specifically designed for *video super-resolution* (VSR) [3, 4, 25, 26] that take temporal information into account achieve results with much higher quality. Similarly, *reference-based super-resolution* (RefSR) methods [30] are able to take high quality reference images into account to produce results well beyond the capability of SISR methods.

In this work we propose a novel VSR method with one main strategy in mind – utilizing all available information to produce higher quality results in a wide variety of scenarios. We identify three main areas where current state-of-the-art VSR methods fall short and individually address them.

First, is the propagation of information over long temporal distances. While techniques like second-order propagation [4] have been shown to greatly improve information

propagation, they still fall short on long video sequences. This is because alignment errors gradually add up and dilute the information. Furthermore, this strategy does not properly handle occlusions and objects going out of frame. Our method solves this problem by allowing information propagation from any frame in the video to any other frame in the video. We show that our long range information propagation scheme results in output frames with significantly improved perceptual quality.

Second, current VSR methods are not able to utilize all available information when high-resolution reference images are available. Reference VSR methods like RefVSR [11] made significant advances in this area. However their considered use case with triple cameras having different fixed focal lengths is too narrow to cover most real-world applications. In contrast, our method is able to inject any number of reference images at any point in the sequence by leveraging the same mechanism that we proposed for the long range information propagation. At the same time, our method continues to support scenarios where no references are available. These are crucial differences compared to RefVSR which ultimately make our method applicable to a wider variety of real-world scenarios. *Third*, the best performing VSR methods [3, 4, 26] are limited to a single scaling factor (typically $\times 4$). Therefore, we propose a simple adaptation to enable arbitrary scaling factors. It only requires minimal training time and is fully compatible with our long range information propagation scheme. We show that our approach significantly outperforms existing arbitrary scaling methods for both in-distribution and out-of-distribution scaling factors.

Our key contributions can be summarized as follows:

- We enable temporal propagation over arbitrarily long distances for VSR.
- We design the first VSR method capable of optionally accepting high quality reference images.
- We obtain state-of-the-art performance for arbitrary scaling in VSR.

2. Related Work

Increasing image resolution and quality is a fundamental task in computer vision. Hence, there exists a large body of work on the topic. On a high level existing SR methods can be distinguished along three main directions. The first is whether they operate on single images or leverage the temporal information available in video inputs. The second is whether they are fixed to a single scaling factors (typically $\times 4$) or support arbitrary (fractional) scaling factors. Finally, some method are able to use high quality reference images to help upscale a given low-resolution image.

Single Image Super-Resolution (SISR) Dong *et al*. [6] were among the first to apply deep learning to the task of image super-resolution. Since then, improvements in network architecture [12, 14, 28, 31, 33] have significantly increased the resulting image quality. The introduction of GANs to SISR by Ledig *et al*. [10] has improved the perceptual quality of images and was leveraged [24, 29] to produce high quality results even for degraded input images. A variety of model designs were proposed to support arbitrary scaling factors [5, 7] and geometric transforms [1, 21].

Video Super-Resolution (VSR) While SISR method can be applied directly to video sequences [20] using the temporal information available in the sequence leads to better results [25]. Chan *et al*. [3] performed extensive experiments to determine the best way to handle the available temporal information, leading to bidirectional recurrent architecture explicitly aligning frames using optical flow. While this works well over short temporal distances, it struggles to propagate high quality information over long temporal distances. A number of methods have been proposed to address this shortcoming. BasicVSR++ [4] introduced second-order propagation where information is propagated from two previous frames. RVRT [13] further addresses the problem dividing video sequences into shorter disjoint clips, and each clip is processed using information from neighboring ones. Recently, IART [26] introduces implicit alignment, which avoids interpolation errors when aligning feature maps and results in a noticeable increase in quality over previous methods. Finally, ST-AVSR [19] is the first method supporting arbitrary scaling factors for VSR. The ability to handle arbitrary scaling factors does, however, come at the cost of noticeable reduction in performance for $\times 4$ scaling when compared to fixed scaling SOTA VSR methods.

Reference-based Super-Resolution (RefSR) Similar to VSR methods, RefSR methods push beyond the limits of SISR. Instead of using multiple low-resolution images from the same video they use high-resolution reference image(s) to help in the upscaling task. Early methods such as [2, 8, 34] were limited to using a single reference image only. Zhang *et al*. [30] addressed this problem and proposed a RefSR method capable of utilizing multiple references. Recently, Lee *et al*. [11] proposed a dataset that consists of video clips captured with multiple FOVs simultaneously, with the narrower FOV frames serving as references for the wider FOV images. Their proposed method is well adapted for this constrained setting but does not generalize to other, more realistic, scenarios. We can also note that a more efficient strategy was introduced [9], to lower its computational cost.

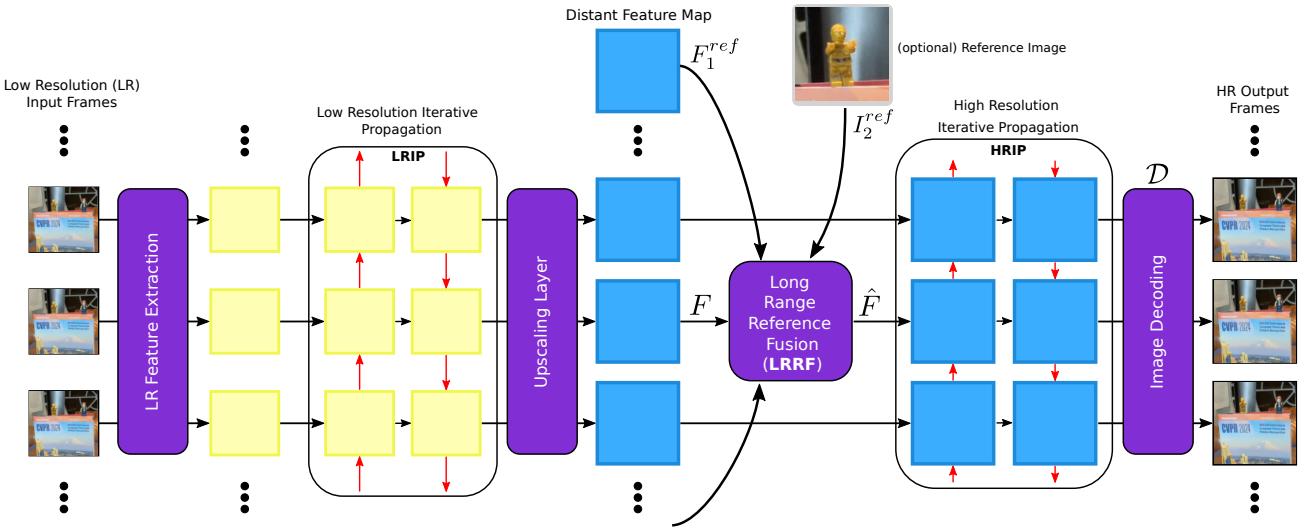

Figure 2. Method overview. Starting from a sequence of low-resolution (LR) frames, LR features are extracted, then refined through the low-resolution iterative propagation (LRIP) module. The upscaling layer upsamples the LR features to the output resolution, obtaining the high-resolution (HR) features F for each frames. After that, the long range reference fusion (LRRF) module injects information from temporally distant HR feature maps (F_1^{ref}) or additional reference image I_2^{ref} when available. The output $\hat{F}$ of the LRRF module is an enhanced version of the initial HR features. The information is further propagated locally with a high resolution iterative propagation (HRIP) module, before final decoding of the HR output frame.

3. Method

We propose a VSR method that tackles existing limitations with long range temporal information propagation. We also extend it to arbitrary scaling, achieving state-of-the-art results. We first describe the overall approach, before detailing the core contributions.

An overview of our method is illustrated in Figure 2. The initial stages of the pipeline are similar to existing VSR methods [3, 4, 26]. Namely, features are first extracted from the low-resolution (LR) frames then refined through a low-resolution iterative propagation (LRIP) module, propagating information frame-by-frame, back and forth through the sequence. After this LR processing, the upscaling layer upsamples the LR features to the output resolution, obtaining the high-resolution (HR) features. In most VSR methods this upscaling layer is a pixel-shuffle, limiting the possible scaling factor, and we modify this to achieve arbitrary scaling.

Contrary, to existing works we do not directly decode the HR features into the HR images. Instead, we introduce a long range reference fusion (LRRF) module. The objective is to inject information from temporally distance HR feature maps. Indeed, depending on the sequence, further away frames may contain richer information that is not easily propagated with the LRIP module. The output of the long range LRRF module is an enhanced version of the initial HR features. We note that this module is flexible and can also fuse information from reference images when available.

Despite its clear benefits, using the LRRF module for all the frames is infeasible due to compute and memory limitations. As a result it is only used for a sparse set of frames, and the information is further propagated locally with a high-resolution iterative propagation (HRIP) module. Finally each HR feature map is converted to a HR output frame using the image decoding module.

We can note that many SOTA VSR methods use a similar pipeline but without the LRRF and HRIP modules. This means our pipeline is fully compatible with existing VSR methods and can make use of powerful pre-trained LR feature extraction, upscaling layer, and image decoding. In our experiments section we use the pre-trained modules from both BasicVSR++ [4] and IART [26]. We will not discuss details of LR feature extraction, upscaling layer, and image decoding in this paper. We instead refer to the respective method for more details.

3.1. Long Range Reference Fusion (LRRF)

Our long range reference fusion (LRRF) is designed to allow information propagation over arbitrarily long distances, by injecting information from any frame in the video sequence into any other frame. As shown in Figure 2 the LRRF operates on HR feature maps produced by the upscaling layer. Given an HR feature map F, it uses an arbitrary number of distant HR feature maps $F_1^{ref}, \cdots, F_N^{ref}$ as input and returns a refined HR feature map $\hat{F}$.

On a high level the LRRF module operates in two stages. First, from every reference we obtain a feature map that is aligned to F. Second, our multi reference fusion (Ref-Fusion) module uses the information contained in the aligned reference feature maps to refine the original feature map F

$$\begin{aligned} \tilde{F}_i^{ref} &= \text{Ref-Align}(F, F_i^{ref}) \\ \hat{F} &= \text{Ref-Fusion}(F, \tilde{F}_1^{ref}, \cdots, \tilde{F}_N^{ref}) \end{aligned} \tag{1}$$

Reference Alignment. Given the *anchor* HR feature map F and a reference HR feature map F^{ref} (we drop the index for simplicity), the alignment can be described as

$$\begin{aligned} I &= \mathcal{D}(F) \\ I^{ref} &= \mathcal{D}(F^{ref}) \\ W &= \text{DenseMatching}(I, I^{ref}) \\ \tilde{F}^{ref} &= \text{RefFeatureExtraction}(I^{ref}, W) \end{aligned} \tag{2}$$

To compute the mapping (or warp grid) W, we decode the HR feature maps into images using the decoder $\mathcal{D}$. These are not the final images, but they allow using any state-of-the-art optical flow or image matching method. The aligned reference feature map $\tilde{F}^{ref}$ is extracted from I^{ref} and W using our RefFeatureExtraction module. The RefFeatureExtraction module first passes I^{ref} to a neural network and then warps the resulting feature map according to W. The warping is performed using nearest neighbor sampling. Further details are provided in the supplementary material.

Our module is flexible and can easily include additional reference images (not from the video sequence). In this case the image I^{ref} is the input, instead of F^{ref}.

Note that the design of our **Ref-Align** module allows it to extract an aligned feature map $\tilde{F}^{ref}$ of the same spatial dimensions as F irrespective of the reference images resolution. Further details regarding the exact architecture used are provided in supplemental material.

Multi-reference Fusion. Our **Ref-Fusion** module utilizes the information contained in an arbitrary number of aligned reference feature maps to refine the anchor feature map F. We achieve this via a pixel-wise attention mechanism that uses the feature map F to extract queries and the warped HR reference feature maps $\tilde{F}_1^{ref}, \cdots, \tilde{F}_N^{ref}$ to extract key-value pairs. Further details regarding the exact architecture used are provided in supplemental material.

3.2. High Resolution Iterative Propagation (HRIP)

Although the LRRF module offers a great degree of flexibility it is relatively expensive to run, especially when many reference images are provided. To avoid excessive computational cost we only run the LRRF module for a subset of high-resolution feature maps. The information contained in these refined HR feature maps is propagated via our HR iterative propagation (HRIP) module. This module operates similarly to the LRIP module but in a less computationally expensive manner. Specifically, this means we use first-order propagation and perform one forward and one backward pass. For comparison, LRIP typically uses second-order propagation and performs two forward and two backward passes.

3.3. Arbitrary Scaling

We note that the only reason most SOTA VSR methods are limited to a fixed scaling factor is the implementation of the upscaling layer. Typically, pixel-shuffle is used to resample the LR feature maps to HR feature maps. We propose replacing the pixel-shuffle warp with our arbitrary scaling module (ASM). The ASM takes an LR feature map F^{LR} and a warp grid W as input and returns a high resolution feature map F^{HR}. Similar to arbitrary scaling SISR methods like LIIF [5] the warp grid W is a mapping from HR pixel coordinates to the corresponding LR coordinates.

$$\begin{aligned} x, y &= W[X, Y] \\ i, j &= \lfloor y \rceil, \lfloor x \rceil \\ dx, dy &= j - x, i - y \\ \phi^{HR}[X, Y] &= [F^{LR}[i, j], dx, dy] \\ F^{HR} &= \mathcal{N}(\phi^{HR}) \end{aligned} \tag{3}$$

Here, we first warp F^{LR} by performing nearest neighbor sampling according to W. Then, we append the sampling offset (dx, dy) along the channel dimension which results in the intermediate HR feature map ϕ^{HR}. The final HR feature map F^{HR} is obtained by passing ϕ^{HR} through the neural network $\mathcal{N}$. Further details regarding $\mathcal{N}$ are provided in the supplemental material.

Enabling arbitrary scaling for a fixed scale VSR method by replacing the upscaling layer with our proposed ASM has the main advantage of faster training time. For instance, ST-AVSR [19] requires 5 days training time, while our ASM can be trained in less than 24hours by starting from a pre-trained VSR backbone (like BasicVSR [4] or IART [26]). Moreover, replacing the fixed scale upscaling layer with our ASM is fully compatible with our long range information propagation scheme.

4. Experiments & Results

We provide details about the models, the number of parameters and exact training setting in supplementary material. In this section we focus on the evaluation of the method with qualitative and quantitative comparisons and

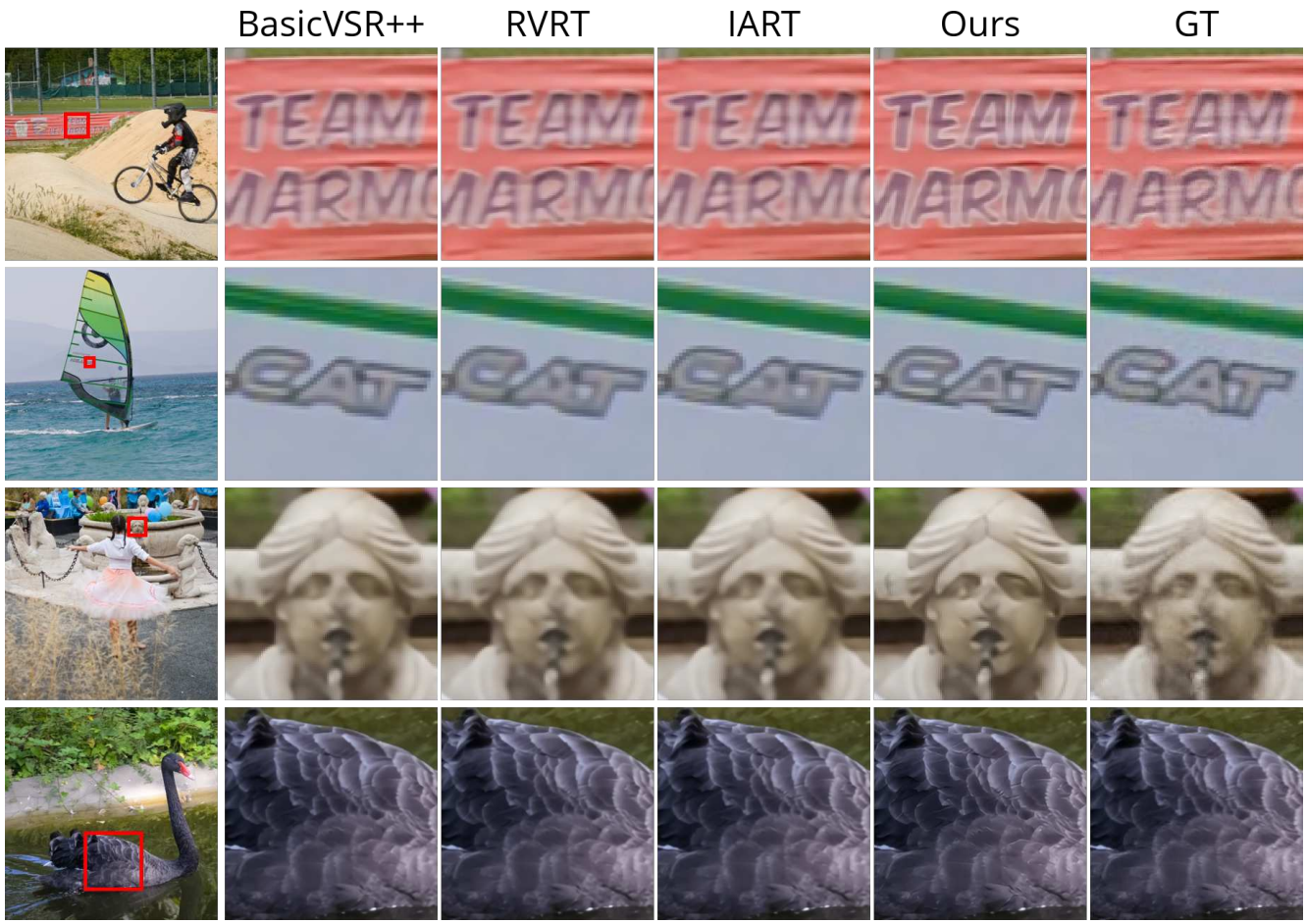

Figure 3. Visual results for $\times 4$ VSR. We compare our method against BasicVSR++ [4], RVRT [13], and IART [26]. Our method produces noticeably sharper results, containing more details and finer lines.

ablations. We used datasets commonly used for VSR evaluation: REDS [16], Vimeo [27] and Vid4 [15]. For qualitative evaluation we additionally include sequences from DAVIS [17] and our own dataset.

For our method there are several ways to set up the long-range propagation. For simplicity we pick a subset of frames as key-frames. The **Ref-Fusion** module only refines the HR feature map for the key-frames. Unless specifically mentioned we use the following: On REDS we pick every 10th frame as a key-frame starting with the 5th; On Vimeo we pick frames 1, 4, and 7 as key-frames; On Vid4 we subdivide the input video into chunks of 5 frames, and pick the center frames as a key-frame.

4.1. Qualitative Results

For this, we show results with our method trained using IART as a backbone. Starting with fixed $\times 4$ scaling, we compare against state-of-the-art methods: BasicVSR++ [4], RVRT [13] and IART [26]. The results are presented in Figure 3. We can clearly see the benefits of long range information propagation, as our method produces results of higher visual quality, with crisper details and finer lines.

In Figure 4 we show visual results for arbitrary scaling factors. The first, second and, third rows correspond to scaling factors $\times 3.25$, $\times 4$, and $\times 5.5$. In all cases our method clearly outperforms existing arbitrary scaling methods. This demonstrates the benefits of leveraging a strong pre-trained VSR backbone.

4.2. Quantitative Evaluation

As the quality of the results in VSR continuously improve and reach higher quality levels, it becomes more challenging to perform quantitative evaluation, as simple metrics such as PSNR and SSIM are limited. LPIPS [32] seems to better correlate with the perceived quality, for our task. This is illustrated in Figure 5. For this $\times 4$ upscale result on REDS, we provide visual comparison to the ground truth along PSNR and LPIPS evaluation. Our results are clearly

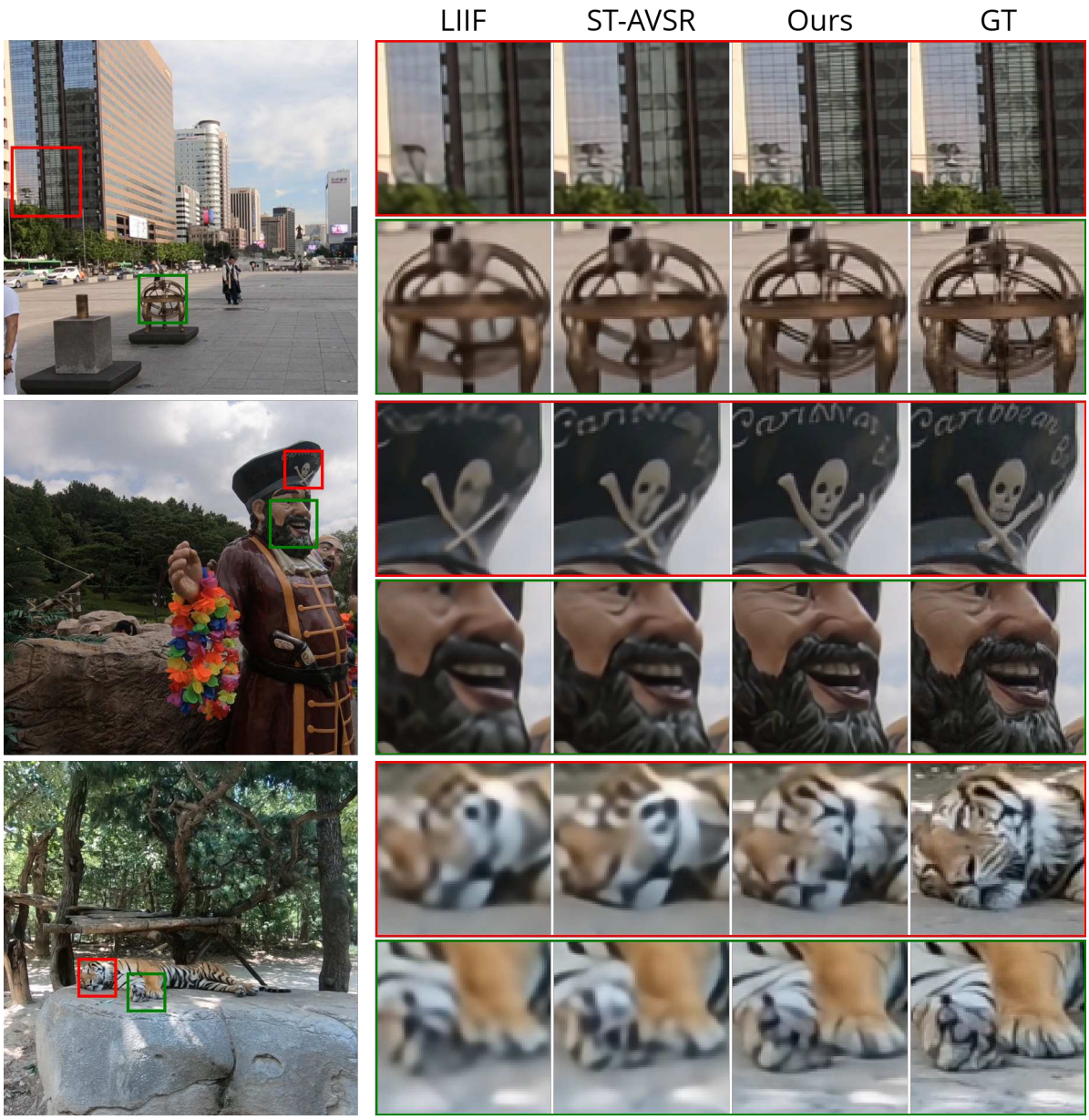

Figure 4. Visual results for arbitrary scaling VSR. The first, second, and third row show results for scaling factors $\times 3.25$, $\times 4$, and $\times 5.5$ respectively. Our method clearly produces sharper, more visually appealing results compared to both LIIF [5] and ST-AVSR [19].

sharper and better match the ground truth images and the LPIPS metric is lower. However this is not reflected in PSNR values.

In our quantitative evaluation we show both metrics. The evaluation is done for both fixed $\times 4$ VSR and arbitrary scaling VSR. The results for fixed $\times 4$ VSR are presented in Table 1. Here, our method is evaluated using both BasicVSR++ [4] and IART [26] as a backbone. The results show that even with the computationally more efficient BasicVSR++ backbone our method outperforms existing methods on both Vimeo and Vid4 on the LPIPS metric. Using IART backbone, our method outperforms existing methods on all datasets on the LPIPS metric.

The evaluation for arbitrary scaling is performed on the REDS dataset with a variety of in-distribution and out-of-distribution scaling factors (our model was trained with scaling factor between $\times 1.5$ and $\times 4.5$). The results are shown in Table 2, comparing against the AS-VSR method ST-AVSR [19] the SISR method LIIF [5]. The publicly available checkpoints for ST-AVSR were trained with-

	REDS		Vimeo		Vid4	
	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
BasicVSR [CVPR 21]	31.52	0.152	36.16	0.077	27.30	0.184
IconVSR [CVPR21]	31.91	0.143	36.26	0.073	27.48	0.172
BasicVSR++ [CVPR 22]	32.41	0.119	36.57	0.072	27.81	0.162
RVRT [NeurIPS 22]	<u>33.03</u>	0.116	<u>36.83</u>	0.068	<u>27.91</u>	0.161
IART [CVPR 24]	**33.20**	<u>0.105</u>	**37.05**	0.067	**28.01**	0.158
Ours (BasicVSR++)	32.02	0.107	35.91	<u>0.065</u>	27.40	<u>0.136</u>
Ours (IART)	32.83	**0.097**	36.45	**0.059**	27.64	**0.134**

Table 1. Numeric evaluation for $\times 4$ VSR on a variety of datasets. For each dataset we report PSNR and LPIPS. The **best** and <u>second best</u> result are highlighted bold and underlined respectively. Our method using IART as a backbone produces significantly better LPIPS than existing methods. Even with the much computationally more efficient BasicVSR++ backbone our method outperforms existing methods in terms of LPIPS on Vimeo and Vid4.

	$\times 2.5$		$\times 3.25$		$\times 4$		$\times 5.5$		$\times 8$	
	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
LIIF [CVPR 21]	32.50	0.134	30.27	0.209	28.91	0.261	27.07	0.346	25.51	0.439
ST-AVSR (no aa)	31.12	0.215	28.31	0.298	27.80	0.368	26.09	0.483	24.77	0.592
ST-AVSR [ECCV 24]	35.07	0.069	29.59	0.139	30.55	0.189	27.89	0.270	26.38	0.377
Ours (BasicVSR++)	<u>36.29</u>	<u>0.033</u>	<u>33.54</u>	<u>0.075</u>	<u>31.67</u>	<u>0.113</u>	<u>29.41</u>	<u>0.192</u>	<u>26.64</u>	<u>0.299</u>
Ours (IART)	**36.63**	**0.030**	**33.82**	**0.069**	**32.36**	**0.103**	**29.72**	**0.181**	**26.66**	**0.279**

Table 2. Numeric evaluation for a variety of scaling factors on the REDS dataset. For each scaling factor we report PSNR and LPIPS. The **best** and <u>second best</u> result are highlighted bold and underlined respectively. Independent of the backbone, our method significantly outperforms existing arbitrarily scaling methods across the board.

Figure 5. Visual illustration of the differences between PSNR and LPIPS. Our method produces sharper, more visually appealing results than IART. LPIPS [32] seems to better correlate with this perceived quality.

out applying anti-aliasing to images before downsampling. This deviates from standard practice used by other methods. For a fair comparison we retrain ST-AVSR using the publicly available code with anti-aliasing applied. We always report the results for both the original version *ST-AVSR (original)* and our re-trained version *ST-AVSR*. The results show that our re-trained version of ST-AVSR performs significantly better than the publicly available version. Irrespective of the backbone and scaling factors, our method performs best.

4.3. Ablation Study

Training Procedure for the LRRF module. Training a model for long range temporal propagation is challenging: training on clips containing hundreds of frames is not computationally feasible. To tackle this challenge we investigate different options to generate distant *reference* images.

The first option is referred to as "self referential" (or *"Self"*). Here, the reference are created by the model: the input clip is first upscaled using the backbone without our additional modules. Randomly selected frames from this initial upscaled sequence are then injected at randomly chosen key-frames. The second is referred to as HQ-references or *"HQ"*. Here, reference images are generated by downscaling frames from the ground truth clip. Further details regarding the generation of reference images are provided in supplemental material.

We train four different version of our method. All variants are trained on the REDS dataset using clips of length 12. For this experiment we use BasicVSR++ [4] as a backbone, with all the backbone parameters kept frozen during training. The first variant (No Ref) only contains a HRIP module and no LRRF module. The second variant (Self) contains both a HRIP and LRRF module and is trained using self-referential references. For the third variant (Self +

(a) Training setup for LRRF.

	PSNR ↑	LPIPS ↓
Baseline	32.41	0.119
No Ref	32.28	0.129
Self	32.23	0.126
Self + HQ	32.15	0.126
HQ	32.04	0.106

(b) Optical flow methods in HRIP.

	PSNR ↑	LPIPS ↓
Baseline	32.41	0.119
SpyNet	32.12	0.110
RAFT	32.02	0.107
PDCNetPlus	32.04	0.106

(c) Effect of arbitrary scaling layer.

	PSNR ↑	LPIPS ↓
BasicVSR++	32.41	0.119
+ Arbitrary Scaling (AS)	32.13	0.129
+ Long Range Propagation + AS	31.67	0.113

Table 3. Ablation Study. Evaluation of the different components of the method. See text for details.

HQ) we using both self-referential and HQ references. Finally, the fourth variant is trained using only HQ references. The results for all four variants are shown in Table 3a, including the unmodified BasicVSR++ model as a baseline. As the (HQ) setup leads to the best LPIPS evaluation, it is the option we use everywhere.

Optical Flow Method. Our HRIP module requires optical flow to align neighboring frames. We investigate three methods for this task. The first option is SpyNet [18] which is used by the BasicVSR++ [4] backbone. This option is computationally the cheapest. The second is RAFT [22], adding more compute and extra parameters (the added parameters are frozen). The third option is, PDCNetPlus [23]. This is the most computationally expensive option.

The results are shown in Table 3b. All variants are trained on REDS with clips of length 12 and using HQ-references. Our conclusion is to select RAFT as the best choice. It performs almost as well as PDCNetPlus while being significantly less expensive to run.

Arbitrary Scaling Module. We add the capability to support arbitrary scaling factors to existing fixed $\times 4$ VSR methods by replacing the upscaling layer. Naturally, we evaluate how this change affects performance for $\times 4$ scaling. We use BasicVSR++ as a starting point and replace its upscaling layer with our arbitrary scaling module. Next, we use AS-BasicVSR++ as a backbone and train our LRRF module on top. We evaluate all three methods on REDS for $\times 4$ scaling (see Table 3c). We can see that the generalization to arbitrary scaling comes at the cost of reduced performance on the specialized $\times 4$ task, however the addition of long distance propagation largely compensate for this, resulting in our arbitrary scaling method largely outperforming its original backbone.

Computational Cost & Key-Frame Density Our method achieves long distance information propagation by injecting information from multiple distance reference images at given key-frames. Throughout our evaluation we have chosen the simple strategy of selecting key and reference frames at regular intervals. In Table 4 we evaluate how the density of selected key and reference frames affects our methods quality and computational cost. For large numbers of references the LRRF module becomes the main computational bottleneck and we provide a more detailed computational cost evaluation of this module in isolation in the supplementary material. We can see that in all configurations our method clearly outperforms its backbone in perceptual quality. Note that injecting a sparse set of reference images at a dense set of key-frames (5 ref, every 5) is a viable option in terms of quality and uses the least amount of GPU memory which is the main limiting factor in practice. This means that for short and medium length sequences we are abel to employ our simplistic scheme of selecting reference and key frames at regular intervals. For long sequences, where previous method were not able to propagate any information, our methods performance can be optimized by intelligently selecting suitable reference images for each key-frame and we look forward to future methods proposing such selection schemes.

	References (Number)	LRRF (frequency)	HRIP	PSNR↑	LPIPS↓	Time [s/f]	Mem [GB]
Ours (IART)	5 ref	every 1		32.74	0.096	2.35	31.1
	5 ref	every 5	✓	32.69	0.096	1.93	22.0
	10 ref	every 10	✓	32.83	0.097	1.91	25.4
	20 ref	every 20	✓	32.88	0.095	1.90	34.4
	20 ref	every 5	✓	32.72	0.094	2.39	36.2
IART	N/A	N/A	N/A	33.20	0.105	1.18	12.5
Ours (BVSR++)	5 ref	every 1		32.10	0.107	1.22	31.0
	5 ref	every 5	✓	31.92	0.105	0.80	22.0
	20 ref	every 5	✓	31.95	0.102	1.26	36.2
BVSR++	N/A	N/A	N/A	32.41	0.119	0.05	12.5

Table 4. Performance and computational cost on the REDS dataset (100 frames per clip). The option with the smallest memory (highlighted) still outperforms IART and BasicVSR++.

5. Conclusion

In this work we have proposed a novel VSR method capable of propagating temporal information over long distances through a Long Range Reference Fusion module. By design, this module naturally allows to also leverage high resolution reference images when available. By further designing our approach such that it can use pre-trained VSR backbones and extending it with arbitrary scaling functionality, we obtain a very flexible method that is applicable to a wide range of real world scenarios while also being efficient to train and reaching state-of-the-art results.

References

[1] Michael Bernasconi, Abdelaziz Djelouah, Farnood Salehi, Markus Gross, and Christopher Schroers. Kernel aware resampler. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22347–22355, 2023. 2

[2] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *European conference on computer vision*, 2022. 2

[3] 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*, 2021. 1, 2, 3

[4] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5962–5971, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[5] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8628–8638, 2021. 2, 4, 6

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks, 2014. 2

[7] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution. 2019. 2

[8] 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. 2

[9] Youngrae Kim, Jinsu Lim, Hoonhee Cho, Minji Lee, Dongman Lee, Kuk-Jin Yoon, and Ho-Jin Choi. Efficient reference-based video super-resolution (ervsr): Single reference image is all you need. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1828–1837, 2023. 2

[10] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. 2

[11] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[12] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021. 1, 2

[13] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. In *Advances in Neural Information Processing Systems*, pages 378–393. Curran Associates, Inc., 2022. 2, 5

[14] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 1, 2

[15] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):346–360, 2013. 5

[16] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*, 2019. 5

[17] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5

[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*, 2017. 8

[19] Wei Shang, Dongwei Ren, Wanying Zhang, Yuming Fang, Wangmeng Zuo, and Kede Ma. Arbitrary-scale video super-resolution with structural and textural priors. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LVII*, page 73–90, Berlin, Heidelberg, 2024. Springer-Verlag. 2, 4, 6

[20] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. 2016. 2

[21] Sanghyun Son and Kyoung Mu Lee. SRWarp: Generalized image super-resolution under arbitrary transformation. In *CVPR*, 2021. 2

[22] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020. 8

[23] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. 2023. 8

[24] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*. 2

[25] Xintao Wang, Kelvin Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. pages 1954–1963, 2019. 1, 2

[26] Kai Xu, Ziwei Yu, Xin Wang, Michael Bi Mi, and Angela Yao. Enhancing video super-resolution via implicit resampling-based alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2546–2555, 2024. 1, 2, 3, 4, 5, 6

[27] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127 (8):1106–1125, 2019. 5

[28] W. Yifan, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O Sorkine-Hornung, and C. Schroers. A fully progressive approach to single-image super-resolution. In *CVPR Workshops*, 2018. 1, 2

[29] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision*, pages 4791–4800, 2021. 2

[30] Lin Zhang, Xin Li, Dongliang He, Fu Li, Errui Ding, and Zhaoxiang Zhang. Lmr: A large-scale multi-reference dataset for reference-based super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13118–13127, 2023. 1, 2

[31] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2998–3008, 2021. 1, 2

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 7

[33] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 1, 2

[34] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *arXiv:1903.00834v1*, 2019. 2