Space-Time-Aware Multi-Resolution Video Enhancement

Muhammad Haris\textsuperscript{1}, Greg Shakhnarovich\textsuperscript{2}, and Norimichi Ukita\textsuperscript{1}
\textsuperscript{1}Toyota Technological Institute, Japan \textsuperscript{2}Toyota Technological Institute at Chicago
muhammad.haris@bukalapak.com,greg@ttic.edu,ukita@toyota-ti.ac.jp

Abstract

We consider the problem of space-time super-resolution (ST-SR): increasing spatial resolution of video frames and simultaneously interpolating frames to increase the frame rate. Modern approaches handle these axes one at a time. In contrast, our proposed model called STARnet super-resolves jointly in space and time. This allows us to leverage mutually informative relationships between time and space: higher resolution can provide more detailed information about motion, and higher frame-rate can provide better pixel alignment. The components of our model that generate latent low- and high-resolution representations during ST-SR can be used to finetune a specialized mechanism for just spatial or just temporal SR. Experimental results demonstrate that STARnet improves the performances of space-time, spatial, and temporal video SR by substantial margins on publicly available datasets.

1. Introduction

The goal of Space-Time Super-Resolution (ST-SR), originally proposed by [49], is to transform a low spatial resolution video with a low frame-rate to a video with higher spatial and temporal resolutions. However, existing SR methods treat spatial and temporal upsampling independently. Space SR (S-SR) with multiple input frames, (i.e., multi-image SR [11, 12] and video SR [22, 33, 7, 46, 17]), aims to super-resolve spatial low-resolution (S-LR) frames to spatial high-resolution (S-HR) frames by spatially aligning similar frames (Fig. 1 (a)). Time SR (T-SR) aims to increase the frame-rate of input frames from temporal low-resolution (T-LR) frames to temporal high-resolution (T-HR) frames by temporally interpolating in-between frames [45, 36, 35, 42, 3, 41] (Fig. 1 (b)).

While few ST-SR methods are presented [49, 50, 47, 40, 32], these methods are not learning-based method and require each input video to be long enough to extract meaningful space-time patterns. [48] proposed ST-SR based on a deep network. However, this method fails to fully exploit the advantages of ST-SR schema because it relies only on LR for interpolation.

On the other hand, one can perform ST-SR by using any learning-based S-SR and T-SR alternately and independently. For example, in-between frames are constructed on S-LR, and then their SR frames are produced by S-SR; Fig. 1 (c). The other way around is to spatially upsample input frames by S-SR, and then to perform T-SR to construct their in-between frames; Fig. 1 (d).

However, space and time are obviously related. This relation allows us to jointly employ spatial and temporal representations for solving vision tasks on both human [20, 21, 8] and machine perceptions [39, 62, 6, 55, 17, 57, 31]. Intuitively, more accurate motions can be represented on a higher spatial representation and, the other way around, a higher temporal representation (i.e., more frames all of which are similar in appearance) can be used to accurately extract more spatial contexts captured in the temporal frames as done in multi-image SR and video SR. This intuition is also supported by various joint learning problems [18, 15, 61, 1, 60, 56, 29], which are proven to improve learning efficiency and prediction accuracy.

In order to utilize the complementary nature of space and time, we propose the Space-Time-Aware multi-Resolution Network, called STARnet. STARnet explicitly incorporates spatial and temporal representations for augmenting S-SR and T-SR mutually in LR and HR spaces by presenting direct connections from LR to HR for ST-SR, indicated as purple arrows in Fig. 1 (e). This network also provides the extensibility where the same network can be further finetuned for either of ST-SR, S-SR, or T-SR. As shown in Fig. 2, STAR-based finetuned models perform better than state-of-the-arts [58, 14, 3, 17].

The main contributions of this paper are as follows:

1) The novel learning-based ST-SR method, which trains a deep network end-to-end to jointly learn spatial and temporal contexts, leading to what we call Space-Time-Aware multi-Resolution Networks (STARnet). This approach outperforms the combinations of S-SR and T-SR methods.

2) Joint learning on multiple resolutions to estimate both
3) A novel view of S-SR and T-SR that are superior to direct S-SR and T-SR. In contrast to the direct S-SR and T-SR approaches, our S-SR and T-SR models are acquired by finetuning ST AR. This finetuning from ST AR allows the S-SR and T-SR models to be augmented by ST-SR learning; (1) S-SR is augmented by interpolated frames as well as by input frames and (2) T-SR is augmented by subtle motions observed in S-HR as well as large motion observed in S-LR.

2. Related Work

**Space SR.** Deep SR [9] is extended by better up-sampling layers [51], residual learning [26, 54], back-projection [14, 16], recursive layers [27], and progressive upsampling [30]. In video SR, temporal information is retained by frame concatenation [7, 24] and recurrent networks [22, 46, 17].

**Time SR.** T-SR, or video interpolation, aims to synthesize in-between frames [36, 45, 23, 35, 42, 41, 3, 43, 37, 59]. The previous methods use a flow image as a motion representation [23, 41, 3, 58, 59]. However, the flow image suffers from blur and large motions. DAIN [3] employed monocular depth estimation in order to support robust flow estimation. As another approach, by spatially downscaling input S-HR frames, large and subtle motions can be extracted in downscaled S-LR and input S-HR frames, respectively [37, 43]. While these methods [37, 43] *downscale* input S-HR frames for T-SR with joint training of multiple spatial resolutions, STARNet *upscales* input S-LR frames both in input and interpolated frames for ST-SR with joint training of multiple spatial and temporal resolutions.

**Space-Time SR.** The first work of ST-SR [49, 50] solved huge linear equations, then created a vector containing all the space-time measurement from all LR frames. Later, [47] presented ST-SR from a single video recording under the assumption of spatial and temporal recurrences. These
previous work [49, 50, 47, 32, 40] have several drawbacks, such as dependencies between the equations, its sensitivity to some parameters, and required longer videos to extract meaningful space-time patterns. [48] proposed STSR method to learn LR-HR non-linear mapping. However, it did not investigate the effectiveness of multiple spatial resolutions to improve the ST-SR results. Furthermore, it is also evaluated on a limited test set.

Another approach is to combine S-SR and T-SR, as shown in Fig. 1 (c) and (d). However, this approach treats each context, spatial and temporal, independently. ST-SR has not been investigated thoroughly using joint learning.

3. Space-Time-Aware multi-Resolution

3.1. Formulation

Given two LR frames (I^l_t and I^l_{t+1}) with size of (M^l \times N^l), ST-SR obtains space-time SR frames (I^H_t, I^H_{t+1}) with size of (M^H \times N^H) where n \in [0, 1] and M^l < M^H and N^l < N^H. The goal of ST-SR is to produce \{I^H_t\}_{t=0}^{T+1} from \{I^l_t\}_{t=0}^{T+1}, where T+1 indicates the higher number of frames than T. In addition, STARnet computes an in-between S-LR frame (I^l_{t+n}) from (I^l_t and I^l_{t+1}) for joint learning on LR and HR in space and time. Bidirectional dense motion flow maps, F_{t\rightarrow t-1} and F_{t\rightarrow t+1} (describing a 2D vector per pixel), between I^l_t and I^l_{t+1} are precomputed. Let L_t \in \mathbb{R}^{M^l \times N^l \times c^l} and H_t \in \mathbb{R}^{M^H \times N^H \times c^h} represent the S-LR and S-HR feature-maps on time t, respectively, where c^l and c^h are the number of channels.

STARnet’s operation is divided into three stages: initialization (stage 1), refinement (stage 2), and reconstruction (stage 3); Fig. 3. We train the entire network end-to-end.

Initializations (Stage 1) achieves joint learning of S-SR, T-SR, and ST-SR on LR and HR where T-SR and ST-SR are performed in the same subnet indicated by “ST-SR.” This stage takes four inputs: two RGB frames (I^l_{t-1} and I^l_{t+1}) and their bidirectional flow images (F_{t\rightarrow t-1} and F_{t\rightarrow t+1}). Stage 1 is defined as follows:

S-SR: \[ H_t = \text{Net}_S(I^l_t, I^l_{t+1}, F_{t\rightarrow t-1}; \theta_s) \]
\[ H_{t+1} = \text{Net}_S(I^l_t, I^l_{t+1}, F_{t\rightarrow t+1}; \theta_s) \]
\[ L_t = \text{Net}_D(H_t; \theta_d) \]
\[ L_{t+1} = \text{Net}_D(H_{t+1}; \theta_d) \]  

Motion: \[ M_{t+1} = \text{Net}_M(F_{t\rightarrow t+1}, F_{t\rightarrow t-1}; \theta_m) \]  

ST-SR: \[ H_{t+n}, L_{t+n} = \text{Net}_ST(H_t, H_{t+1}, L_t, L_{t+1}, M_t; \theta_m) \]

In S-SR, S-HR feature-maps (H_t and H_{t+1}) are produced by \text{Net}_S, as expressed in Eq. (1). As with other video SR methods, this S-SR is performed with sequential frames (I^l_t and I^l_{t+1}) and their flow image (F_{t\rightarrow t-1} or F_{t\rightarrow t+1}). \theta denotes a set of weights in each network. Following up- and down-samplings for enhancing features for SR [14, 17], H_t and H_{t+1} are downsampled by \text{Net}_D for updating L_t and L_{t+1}, respectively, as expressed in Eq. (2). \text{Net}_M produces a motion representation (M) which is calculated from the bidirectional optical flows; Eq. (3). The output of \text{Net}_M is flow feature maps, learned by a CNN. While it is hard to interpret these features directly, they are intended to help spatial alignment between F_{t\rightarrow t+1} and F_{t\rightarrow t-1}.

Finally, with the concatenation of all these features, ST-SR in the feature space is performed by \text{Net}_ST; Eq. (4). \text{Net}_ST achieves T-SR as well as ST-SR which are incorporated on LR and HR, shown as blue and purple arrows in Fig. 1 (e). The outputs of stage 1 are HR and LR feature-maps (H_{t+n} and L_{t+n}) for an in-between frame.

In this stage, STARnet maintains cycle consistencies (1) between S-HR and S-LR and (2) between t and t+1, while such a cycle consistency is demonstrated for general purposes [64, 13, 63].

Refinement (Stage 2) further maintains the cycle consistencies for refining the feature-maps again. While raw optical flows (F_{t\rightarrow t-n} and F_{t\rightarrow t+n}) are used in Eq. (1) of Stage 1, the motion feature (M) is used in the first equations of Eqs (5), (7), (9), and (10) in Stage 2. This difference allows us to produce more reliable feature-maps. For further refinement, residual features are extracted in Eqs. (6), (8), and (11), as proposed in RBPN [17] for precise spatial alignment of temporal features.

Finally, Stage 2 is defined as follows:

\[ t: H_t^b = \text{Net}_B(L_{t+n}, L_t, M_t; \theta_b) \]
\[ L_t^b = \text{Net}_D(H_t^b; \theta_d) \]
\[ \hat{H}_t = H_t + \text{ReLU}(H_t - H_t^b) \]
\[ \hat{L}_t = L_t + \text{ReLU}(L_t - L_t^b) \]  

\[ t+1: H_{t+1}^f = \text{Net}_F(L_{t+n}, L_{t+1}, M_t; \theta_f) \]
\[ L_{t+1}^f = \text{Net}_D(H_{t+1}; \theta_d) \]
\[ \hat{H}_{t+1} = H_{t+1} + \text{ReLU}(H_{t+1} - H_{t+1}^f) \]
\[ \hat{L}_{t+1} = L_{t+1} + \text{ReLU}(L_{t+1} - L_{t+1}^f) \]  

\[ t+n: H_{t+n}^f = \text{Net}_F(L_t, L_{t+n}, M_t; \theta_f) \]
\[ L_{t+n}^f = \text{Net}_D(H_{t+n}; \theta_d) \]
\[ \hat{H}_{t+n} = H_{t+n} + \text{ReLU}(H_{t+n} - H_{t+n}^f) + \text{ReLU}(H_{t+n} - H_{t+n}^b) \]
\[ \hat{L}_{t+n} = L_{t+n} + \text{ReLU}(L_{t+n} - L_{t+n}^f) + \text{ReLU}(L_{t+n} - L_{t+n}^b) \]  

Reconstruction (Stage 3) transforms four feature-maps (H_t, \hat{H}_{t+n}, \hat{H}_{t+1}, and L_{t+n}) to their corresponding images (I^H_t, I^H_{t+n}, I^H_{t+1}, and I^l_{t+n}) by using only one conv layer \text{Net}_{rec}; for example, I^H_t = \text{Net}_{rec}(\hat{H}_t; \theta_{rec}).
Space-Time loss is evaluated only on \( I_{t+1} \) and \( I_{t+1}^{sr} \).

**Space loss** is evaluated on \( I_{t+1}^{sr} \).

**Time loss** is evaluated only on \( I_{t+1}^{sr} \).

**Space-Time loss** is evaluated only on \( I_{t+1}^{sr} \).

Our framework provides the following four variants, which are trained with different training objectives.

**STAR** is trained using all of the aforementioned three losses on LR and HR in space and time. STAR produces \( \{ I_t^{sr} \}_{t=0}^T \) and \( \{ I_t^h \}_{t=0}^T \) simultaneously as in Fig. 4 (a). **STAR-ST** is a fine-tuned model from STAR using Space and Space-Time losses on HR in space and time. The network is optimized on the space-time super-resolved frames \( \{ I_t^{sr} \}_{t=0}^T \) as in Fig. 4 (b). **STAR-S** is a fine-tuned model from STAR using Space loss on S-HR, optimizing only \( \{ I_t^{sr} \}_{t=0}^T \) as in Fig. 4 (c). **STAR-T** is a fine-tuned model from STAR using Time loss on T-HR as in Fig. 4 (d). STAR-T can be trained on two different regimes, S-LR and S-HR. While STAR-T uses the original frames (S-HR) as input frames, STAR-T uses the downsampled frames (S-LR) as input frames.

### 3.3. Loss Functions

Each of Space, Time, and Space-Time losses consists of two types of loss functions, \( L_1 \) and \( L_{vgg} \). \( L_1 \) is the loss per-pixel between a predicted super-resolved frame \( (I_t^{sr}) \) and its ground-truth HR frame \( (I_t^h) \) where \( t \in [T] \).

\[
L_1 = \sum_{t=0}^{T} ||I_t^h - I_t^{sr}||_1 \tag{12}
\]

\( L_{vgg} \) is calculated in the feature space using a pretrained VGG19 network [52]. For computing \( L_{vgg} \), both \( I^h \) and \( I^{sr} \) are mapped into the feature space by differentiable functions \( f_m \) from the VGG multiple max-pool layer \( (m = 5) \).

\[
L_{vgg} = \sum_{t=0}^{T} \| f_m(I_t^h) - f_m(I_t^{sr}) \|_2^2 \tag{13}
\]

\( L_1 \) is for fulfilling standard image quality assessment metrics such as PSNR and validated for SR [42, 5], while
$L_{vgg}$ improves visual perception [25, 10]. Based on this fact, only $L_1$ or a weighted sum of $L_1$ and $L_{vgg}$ is utilized for training STARnet depending on the purpose.

3.4. Flow Refinement

As mentioned in Section 3.1, we use flow images precomputed by [34]. As revealed in many video interpolation papers [36, 45, 23, 35, 42, 41, 3, 43, 37, 59], large motions between $t$ and $t+1$ make video interpolation difficult. Flow noise due to such large motions has a bad effect on the interpolation results. While STARnet suppresses this bad effect by T-SR not only in S-HR but also in S-LR, it is difficult to fully resolve this problem. For further improvement, we propose a simple solution to refine or denoise the flow images, called a Flow Refinement (FR) module.

Let $F_{t \rightarrow t+1}$ and $F_{t+1 \rightarrow t}$ are flow images between frames $I_t$ and $I_{t+1}$ on forward and backward motions, respectively. During training, $F_{t \rightarrow t+n}$ can be calculated from an input frame at $t$ to the ground truth (i.e., from $I_t$ to $I_{t+n}$). $Net_{flow}$ is a U-Net which defines as follows.

$$FR: \hat{F}_{t \rightarrow t+1} = Net_{flow}(F_{t \rightarrow t+1}, I_t, I_{t+1}; \theta_{flow})$$
$$F_{t+1 \rightarrow t} = Net_{flow}(F_{t+1 \rightarrow t}, I_t, I_{t+1}; \theta_{flow})$$

(14)

To reduce the noise, we propose the following flow refinement loss.

$$L_{flow} = ||\hat{F}_{t \rightarrow t+1} - (F_{t \rightarrow t+1} + F_{t+1 \rightarrow t+1})||_2^2 + ||\hat{F}_{t+1 \rightarrow t} - (F_{t+1 \rightarrow t} + F_{t \rightarrow t+1})||_2^2$$

(15)

With $L_{flow}$, the loss functions for training STARnet are defined as follows:

$$L_r = w_1 * L_1 + w_2 * L_{flow}$$
$$L_f = L_r + w_3 * L_{vgg}$$

(16) (17)

4. Experimental Results

In all experiments, we focus on 4x SR factor and $n = 0.5$. $I_{t+1}^{rc}$ and $I_{t+1}^{cr}$ denote the SR frames of input frames and in-between frames, respectively.

4.1. Implementation Details

Stage 1. For $Net_S$ and $Net_P$, we use DBPN [14] or RBPNN [17] that have up- and down-sampling layers to simultaneously produce a pair of S-LR and S-HR features with $c^d=64$ and $c^r=128$. $Net_M$ is constructed with two residual blocks where each block consists of two conv layers with $3 \times 3$ with stride = 1 and pad by 1. $Net_{ST}$ has five residual blocks followed by deconv layers for upsampling.

Stage 2. Both $Net_F$ and $Net_B$ are constructed using five residual blocks and deconv layers.

Train Dataset. We use the triplet training set in Vimeo90K [58] for training. This dataset has 51,313 triplets from 14,777 video clips with a fixed resolution, $448 \times 256$. During training, we apply augmentation, such as rotation, flipping, and random cropping. The original images are regarded as S-HR and downsampled to $112 \times 64$ S-LR frames ($4 \times$ smaller than the originals) with Bicubic interpolation.

Test Dataset and Metrics. We evaluate our method on several test sets. The test set of Vimeo90K [58] consists of 3,782 triplets with the original resolution of $448 \times 256$ pixels. While UCF101 [53] is developed for action recognition, it is also used for evaluating T-SR methods. This test set consists of 379 triplets with the original resolution of $256 \times 256$ pixels. Middlebury [2] has the original resolution of $640 \times 480$ pixels. We evaluate PSNR, SSIM, and interpolation error (IE) on the test sets.

Training Strategy. The batch size is 10 with $112 \times 64$ pixels (S-LR scale). The learning rate is initialized to $1e-4$ for all layers and decreased by a factor of 10 on every 30 epochs for total 70 epochs. For each finetuned model, we use another 20 epochs with learning rate $1e-4$ and decreased by a factor of 10 on every 10 epochs. We initialize the weights based on [19]. For optimization, we used AdaMax [28] with momentum to 0.9. All experiments were conducted using Python 3.5.2 and PyTorch 1.0 on NVIDIA Tesla V100 GPUs. For the loss setting, we use $w_1: 1$, $w_2: 0.1$, and $w_3: 0.1$.

4.2. Ablation Studies

Here, we evaluate STARnet without T-SR paths (blue arrows in Fig. 1(e)) in order to clarify the effectiveness of our core contribution (i.e., joint learning in time and space on multiple resolutions) with a simplified network using direct ST-SR paths (purple arrows). The test set of Vimeo90K [58] is used.

Basic components. We evaluate the basic components on STARnet. In the first experiment, we remove the refinement part (i.e., Stage 2), leaving only the initialization part. Second, we omit input flow images and $Net_M$, so no motion context is used (STAR w/o Flow). Third, the FR module is removed. Finally, the full model is evaluated. The results of these four models are shown in “STAR w/o Stage 2,” “STAR w/o Flow,” “STAR w/o FR,” and “STAR” in Table 1. Compared with the full model, the PSNR of STAR w/o Stage 2 decreases to 0.36dB and 1.0dB on $I_{t}^{rc}$ and $I_{t}^{cr}$, respectively. The flow information can also improve the PSNR 0.28dB and 0.43dB on $I_{t}^{rc}$ and $I_{t}^{cr}$, respectively.

While FR is also useful, the quantitative improvement by FR is not substantial compared with those of the other two components. The examples of $I_{t}^{rc}$ are shown in Fig. 5 where flow images are computed only by $I_{t}$ and $I_{t+1}$, only by $I_{t}$ and $I_{t+1}$, and refined by FR, and by $I_{t}$, (i.e., GT in-between frame) in addition to $I_{t}$ and $I_{t+1}$ in (a), (b), and (c), respectively. In Fig. 5, the visual improvement by FR is substantial. This result reveals that (1) erroneous flows are
Table 1. Baseline comparison of STAR with DBPN [16] and $L_f$. Red in all tables indicates the best performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>$I^{sr}_t$ PSNR</th>
<th>$I^{sr}_t$ SSIM</th>
<th>$I^{sr}_t$ PSNR</th>
<th>$I^{sr}_t$ SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAR w/o Stage 2</td>
<td>30.920</td>
<td>0.921</td>
<td>30.002</td>
<td>0.917</td>
</tr>
<tr>
<td>STAR w/o Flow</td>
<td>31.489</td>
<td>0.928</td>
<td>30.086</td>
<td>0.918</td>
</tr>
<tr>
<td>STAR w/o FR</td>
<td>31.601</td>
<td>0.929</td>
<td>30.229</td>
<td>0.920</td>
</tr>
<tr>
<td>STAR</td>
<td>31.920</td>
<td>0.933</td>
<td>30.365</td>
<td>0.923</td>
</tr>
</tbody>
</table>

Figure 5. Visual analysis of $I^{sr}_t$ with and w/o FR. (a) Flows are computed by $I^t$ and $I_{t+1}$. (b) Flows are computed by $I^t$ and $I_{t+1}$ and refined by FR. (c) Flows are computed by $I^{sr}_t$, (i.e., GT inter-frame) in addition to $I^t$ and $I_{t+1}$.

Table 2. Analysis on different training objectives using STARnet with DBPN [16] and $L_f$.

<table>
<thead>
<tr>
<th>Method</th>
<th>$I^{sr}_t$ PSNR</th>
<th>$I^{sr}_t$ SSIM</th>
<th>$I^{sr}_t$ PSNR</th>
<th>$I^{sr}_t$ SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAR</td>
<td>31.601</td>
<td>0.929</td>
<td>30.229</td>
<td>0.920</td>
</tr>
<tr>
<td>STAR-ST</td>
<td>31.883</td>
<td>0.933</td>
<td>30.350</td>
<td>0.928</td>
</tr>
<tr>
<td>STAR-S</td>
<td>32.026</td>
<td>0.935</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>STAR-T</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>30.028</td>
</tr>
</tbody>
</table>

Table 3. Analysis on two loss functions using STAR-ST with RBPN. Higher PSNR and SSIM indicate better results, while a lower NIQE indicates a better perceptual index.

<table>
<thead>
<tr>
<th>Loss</th>
<th>$I^{sr}_t$ PSNR</th>
<th>$I^{sr}_t$ SSIM</th>
<th>$I^{sr}_t$ PSNR</th>
<th>$I^{sr}_t$ SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_f$</td>
<td>32.153</td>
<td>0.936</td>
<td>6.280</td>
<td>0.925</td>
</tr>
<tr>
<td>$L_r$</td>
<td>32.349</td>
<td>0.938</td>
<td>6.905</td>
<td>30.704</td>
</tr>
</tbody>
</table>

Larger scale T-SR. The performance on a larger scale T-SR is investigated. While the S-SR factor is the same, the PSNR and SSIM in other experiments (i.e., $4 \times$), the frame-rate is upsampled to $4 \times$. We compare two upsampling paths: (1) STAR-ST ($2 \times$ S-SR and $2 \times$ T-SR) → STAR-ST ($2 \times$ S-SR and $2 \times$ T-SR) (2) STAR-ST ($4 \times$ S-SR and $2 \times$ T-SR) → STAR-SR ($2 \times$ T-SR). For training $4 \times$ T-SR, the training set of the Vimeo90K setuplet test set contains 7 frames, used. Then, the 1st and 5th frames in the Vimeo90K set are used as input frames for evaluation. As shown in Table 5, the second path is better. This result may suggest that a higher spatial resolution provides better results on T-SR.

Table 4. Analysis on the S-SR module using STAR-ST and $L_r$.

<table>
<thead>
<tr>
<th>Method</th>
<th>$I^{sr}_t$ PSNR</th>
<th>$I^{sr}_t$ SSIM</th>
<th>$I^{sr}_t$ PSNR</th>
<th>$I^{sr}_t$ SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAR with DBPN [16]</td>
<td>32.150</td>
<td>0.936</td>
<td>30.540</td>
<td>0.925</td>
</tr>
<tr>
<td>STAR with RBPN [17]</td>
<td>32.349</td>
<td>0.938</td>
<td>30.704</td>
<td>0.928</td>
</tr>
</tbody>
</table>

Table 5. Analysis on larger scale T-SR ($4 \times$) on the Vimeo90K settest set with $L_r$.

<table>
<thead>
<tr>
<th>Method</th>
<th>$I^{sr}_t$ PSNR</th>
<th>$I^{sr}_t$ SSIM</th>
<th>$I^{sr}_t$ PSNR</th>
<th>$I^{sr}_t$ SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Only ST-SR</td>
<td>33.007</td>
<td>0.941</td>
<td>27.186</td>
<td>0.893</td>
</tr>
<tr>
<td>(2) STAR-ST → STAR-SR</td>
<td>34.146</td>
<td>0.950</td>
<td>27.640</td>
<td>0.901</td>
</tr>
</tbody>
</table>

Table 6. Analysis on ST-SR jointly trained with T-SR with RBPN [17] and $L_r$. Models are optimized for STAR-ST w/ FR.
Table 7. Comparison on ST-SR ($I_r^{sr}$) using $L_r$. $\alpha \rightarrow \beta$ indicates the output of $\alpha$ is the input of $\beta$. Red indicates the best and blue indicates the second best performance in all tables in Section 4.3. * indicates a joint learning of RBPN and DAIN methods to perform ST-SR.

ST-SR. As discussed in Section 2, older ST-SR methods [49, 50, 47, 32, 40] cannot be applied to videos in the Vimeo90K dataset. We can combine more modern S-SR and T-SR methods to perform ST-SR. We use DBPN [16] and RBPN [17] as S-SR. For T-SR, we choose TOFlow [58] and DAIN [3]. In Table 7, we present the results of ST-SR obtained by six combinations of these methods.

It is found that S-SR→T-SR performs better than T-SR→S-SR. The margin is up to 1dB on Vimeo90K, showing that the performance of previous T-SRs significantly drops on LR images. Even STAR is better than the combination of state-of-the-arts (RBPN [17]→DAIN [3]), while the best result is achieved by STAR-ST, which is the fine-tuned model from STAR. STAR-ST has a better performance around 0.38dB than RBPN [17]→DAIN [3] on Vimeo90K test set.

We can also present ST-SR as a joint learning of RBPN [17] and DAIN [3], indicated as (*). It shows that
joint learning is effective to improve this combination as well as STAR. However, STAR, which leverages direct connections for ST-SR (i.e., purple arrows in Fig. 1 (e)) and joint learning in space and time, shows the best performance. Visual results shown in Fig. 6 demonstrate that STAR-ST produces sharper images than others.

**S-SR.** The results on S-SR are shown in Table 8. Our methods are compared with DBPN [16], DBPN-MI, and RBPN [17]. DBPN is a single image SR method. A Multi-Image extension of DBPN (DBPN-MI) uses DBPN with a temporal concatenation of RGB and optical flow images. DBPN-MI and RBPN have the same input regimes using sequential frames and optical flow images.

It shows that multiple frames are able to improve the performance of DBPN for around 0.3dB on Vimeo90K. RBPN successfully leverages temporal connections of sequential frames for performance improvement compared with DBPN and DBPN-MI. As expected, STAR-S is the best, which is also better than STAR-ST. It can improve the PSNR by 1.19dB dB, 0.87dB, and 0.55dB compared with DBPN [16], DBPN-MI, and RBPN [17], respectively, on Vimeo90K test set.

**T-SR.** Our method is compared with eight state-of-the-art T-SR methods: SPyNet [44], EpicFlow [45], MIND [36], DVF [35], TOFlow [58], SepConv [42], MEMC-Net [4], and DAIN [3]. Input frames are the original size of the test set without downscaling. As shown in Table 9, STAR-T\textsubscript{HR} is comparable with the state-of-the-art T-SR methods.

The visual results are shown in Fig. 7. We can see that STAR produces better interpolation on subtle and large motions, and also sharper textures. DAIN [3] and TOFlow [58] tend to produce blur images on subtle and large motion areas as shown by the red arrows.

We also investigate the performance on S-LR. There are different motion magnitudes between S-HR and S-LR. Naturally, when the frames are downscaled, the magnitude of pixel displacements is reduced as well. Therefore, each spatial resolution has a different access to the motion variance. The evaluation on S-LR images focuses on subtle motions, while S-HR images focus on large motions. Table 9 shows that STAR-T\textsubscript{HR} is superior to STAR-T\textsubscript{LR} and other methods on S-HR (original size). Likewise, STAR-T\textsubscript{LR} is superior to STAR-T\textsubscript{HR} on S-LR (original size) ↓ with Bicubic) as shown in Table 10. It shows that if we finetune the network on the same domain, it can increase the performance. Furthermore, we can see that STAR-T\textsubscript{LR} is much superior than TOFlow and DAIN.

5. Conclusion

We proposed a novel approach to space-time super-resolution (ST-SR) using a deep network called Space-Time-Aware multi-Resolution Network (STARnet). The network super-resolves jointly in space and time. We show that a higher resolution presents detailed motions, while a higher frame-rate provides better pixel alignment. Furthermore, we demonstrate a special mechanism to improve the performance for just S-SR and T-SR. We conclude that the integration of spatial and temporal contexts is able to improve the performance of S-SR, T-SR, and ST-SR by substantial margin on publicly available datasets.

This work was supported by JSPS KAKENHI Grant Number 19K12129.
References


