Wavelet Attention Embedding Networks for Video Super-Resolution

Young-Ju Choi¹, Young-Woon Lee², and Byung-Gyu Kim*¹

¹Dept. of IT Engineering, Sookmyung Women’s University, Seoul, Korea
²Dept. of Computer Science and Engineering, Sunmoon University, A-san, Korea
1. Introduction

MOTIVATION

✓ In video super-resolution (VSR), the frames contain different type of information including low and high-frequency components.

✓ However, the previous methods handled the features identically or simply combined the edge map so the high resolution (HR) output image lack meaningful information.

✓ The compensated feature generated by pixel-based frame alignment can involve some discontinuous regions.

✓ This inaccurate motion alignment and compensation cause the HR output containing visual artifacts.

APPROACH

✓ In this paper, we propose the wavelet attention embedding networks (WAEN) consisting of two embedding modules to jointly exploit the spatio-temporal dependencies for VSR.

✓ One module is the wavelet embedding network (WENet) for spatial features, and the other one is the attention embedding network (AENet) for temporal features.

✓ Our WAEN can enhance low-frequency features and recover high-frequency details by utilizing appropriate spatial and temporal information.
Given $2N + 1$ consecutive low resolution (LR) input frames $LR_{input}$, our WAEN has a purpose of estimating a HR center frame $SR_t$.

We designed two types of pipeline structure (parallel and serial).

In the parallel structure, input frames are fed to both WENet and AENet.

In the serial structure, input frames are fed to only WENet, and the output features of WENet become the input of AENet.

The output features after the embedding network pass through a reconstruction (with residual blocks) and up-sampling (with depth-to-space transformation) module.

Fig. 1. The network architecture of the proposed WAEN.
2. Methodology

The WENet is operated as a spatial feature extractor of individual low and high-frequency information based on 2-D Haar discrete wavelet transform (DWT).

Through separating each given feature to four sub-band wavelet feature by DWT, more precise and sharp features can be extracted.
2. Methodology

Our AENet is based on the **temporal and spatial attention (TSA) module** in [6].

In **neighboring frames** with different degrees of motion information, there is a **high probability that necessary information** for the reference frame exists.

By utilizing **the relationship between frames**, **discontinuities** in output feature **can be reduced** rather than extracting explicit or implicit motion feature.
3. Experimental Results

DATASETS AND IMPLEMENTATION DETAILS

- We use Vimeo-90K dataset for training and Vid4 dataset for testing.
- For evaluation, we use peak signal-to-noise ratio (PSNR).
- The network takes 7 frames (3 channel patches of 64×64 for training).
- The scale of SR was set to 4.
- We used Charbonnier penalty function for loss function.
- We trained with setting the size of mini-batch to 20.
- We used Adam optimizer.
- We initially set learning rate to $4 \times 10^{-4}$.

QUANTITATIVE COMPARISON RESULTS

Table 1. Quantitative comparison on Vid4 for 4× video SR on Y (luminance) channel.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Calendar</td>
<td>20.45</td>
<td>21.56</td>
<td>23.47</td>
<td>23.02</td>
<td>23.63</td>
<td>23.81</td>
</tr>
<tr>
<td>City</td>
<td>25.22</td>
<td>26.24</td>
<td>27.36</td>
<td>27.93</td>
<td>27.48</td>
<td>27.61</td>
</tr>
<tr>
<td>Foliage</td>
<td>23.57</td>
<td>24.65</td>
<td>25.84</td>
<td>26.26</td>
<td>25.89</td>
<td>26.00</td>
</tr>
<tr>
<td>Walk</td>
<td>26.27</td>
<td>28.41</td>
<td>30.11</td>
<td>29.61</td>
<td>30.16</td>
<td>30.37</td>
</tr>
</tbody>
</table>

Table 2. Quantitative comparison on Vid4 for 4× video SR on RGB channel.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>23.75</td>
<td>24.76</td>
<td>25.84</td>
<td>26.39</td>
<td>25.96</td>
<td>26.08</td>
</tr>
<tr>
<td>Foliage</td>
<td>22.21</td>
<td>23.25</td>
<td>24.44</td>
<td>24.84</td>
<td>24.47</td>
<td>24.59</td>
</tr>
<tr>
<td>Average</td>
<td>22.47</td>
<td>23.76</td>
<td>25.19</td>
<td>25.21</td>
<td>25.27</td>
<td>25.42</td>
</tr>
</tbody>
</table>

Table 3. Adopted modules in our WAEN on Vid4 for 4× video SR.

<table>
<thead>
<tr>
<th>Method Params.</th>
<th>EDVR TSA [6] 5.0M</th>
<th>WENet 8.5M</th>
<th>WAEN P (Ours) 9.5M</th>
<th>WAEN S (Ours) 9.6M</th>
</tr>
</thead>
<tbody>
<tr>
<td>WENet</td>
<td>X</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>AENet</td>
<td>V</td>
<td>X</td>
<td>X</td>
<td>V</td>
</tr>
<tr>
<td>Reconstruction</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>Y</td>
<td>26.75</td>
<td>26.67</td>
<td>26.79</td>
<td>26.95</td>
</tr>
<tr>
<td>RGB</td>
<td>25.24</td>
<td>25.15</td>
<td>25.27</td>
<td>25.42</td>
</tr>
</tbody>
</table>

- Our WAEN S shows the best performance and WAEN P is in second place on average in both Y and RGB channels.
- From the results about adopted modules, we can explain that combination of two feature extractors produces better performance than using a single module.
3. Experimental Results

QUALITATIVE COMPARISON RESULTS

Fig. 4. Visual results on Vid4 for 4× video SR. Zoom in to see better visualization.
4. Conclusions

- In this paper, we have proposed a wavelet attention embedding network for VSR.

- The proposed model extracts the enhanced spatial features by handling four different components individually in wavelet embedding network.

- The effective temporal features can be extracted by generating attention map with neighboring frames in attention embedding network.

- The WAEN can derive the meaningful feature for more accurate HR reconstruction by applying a powerful spatio-temporal structure.

- We compared the proposed models with other recent state-of-the-art VSR approaches and the results demonstrated that our proposed method could obtain better quality of SR.
Thank You