Wavelet Attention Embedding Networks for Video Super-Resolution

Young-Ju Choi[1], Young-Woon Lee[2], and Byung-Gyu Kim*[1]

[1]Dept. of IT Engineering, Sookmyung Women's University, Seoul, Korea

[2]Dept. of Computer Science and Engineering, Sunmoon University, A-san, Korea

Introduction

MOTIVATION

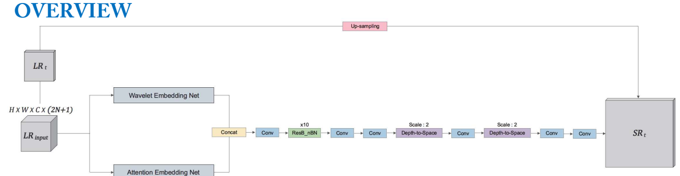
- ✓ In **video super-resolution (VSR)**, the frames contain different type of information including **low and high-frequency components**.
- ✓ However, the previous methods **handled the features identically** or **simply combined the edge map** so the high resolution (HR) output image **lack meaningful 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

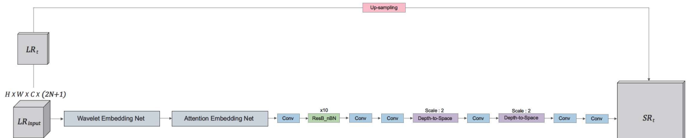
- ✓ In this paper, we propose the **wavelet attention embedding networks (WAEN)** consisting of **two embedding modules** to jointly exploit the spatio-temporal dependencies for VSR.
- ✓ One module is the **wavelet embedding network (WENet)** for **spatial features**, and the other one is the **attention 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.

Methodology

OVERVIEW



(a) Parallel structure.



(b) Serial structure.

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

- ✓ 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.

WAVELET EMBEDDING NET

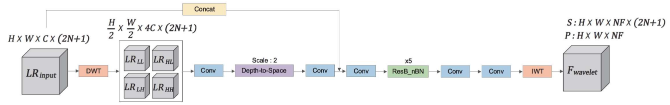


Fig. 2. The structural details of wavelet embedding net.

- ✓ 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.

ATTENTION EMBEDDING NET

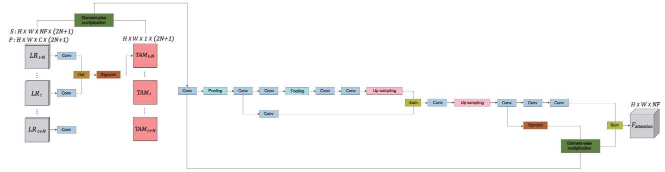


Fig. 3. The structural details of attention embedding net.

- ✓ 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.

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 $\mathbf{4 \times 10^{-4}}$.

Quantitative Comparison Results

Table 1. Quantitative comparison on Vid4 for 4× video SR on **Y (luminance) channel**.
Red and Blue indicates the best and the second best performance, respectively.

Method Params.	Bicubic -	SOF-VSR [21] 1.0M	WDVR [8] 1.2M	FRVSR [3] 5.1M	**WAEN P (Ours)** **9.5M**	**WAEN S (Ours)** **9.6M**
Calendar	20.45	21.56	23.47	23.02	**23.63**	**23.81**
City	25.22	26.24	27.36	27.93	**27.48**	**27.61**
Foliage	23.57	24.65	25.84	26.26	**25.89**	**26.00**
Walk	26.27	28.41	30.11	29.61	**30.16**	**30.37**
Average	23.88	25.21 (26.00)	26.69 (26.62)	26.71 (26.69)	**26.79**	**26.95**

Table 2. Quantitative comparison on Vid4 for 4× video SR on **RGB channel**.
Red and Blue indicates the best and the second best performance, respectively.

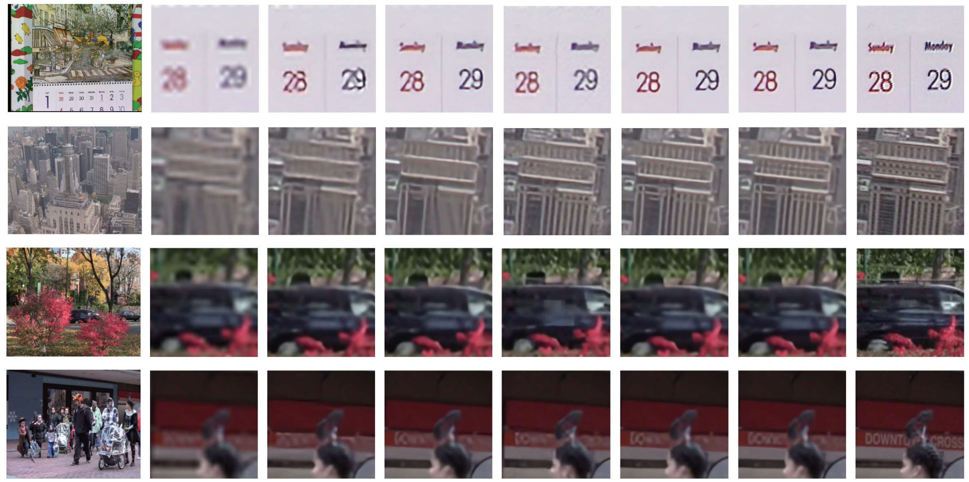
Method Params.	Bicubic -	SOF-VSR [21] 1.0M	WDVR [8] 1.2M	FRVSR [3] 5.1M	**WAEN P (Ours)** **9.5M**	**WAEN S (Ours)** **9.6M**
Calendar	18.96	19.97	21.75	21.37	**21.87**	**22.04**
City	23.75	24.76	25.84	26.39	**25.96**	**26.08**
Foliage	22.21	23.25	24.44	24.84	**24.47**	**24.59**
Walk	24.94	27.07	28.74	28.24	**28.79**	**28.99**
Average	22.47	23.76	25.19	25.21	**25.27**	**25.42**

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

Method Params.	EDVR TSA [6] 5.0M	WENet 8.5M	**WAEN P (Ours)** **9.5M**	**WAEN S (Ours)** **9.6M**
WENet	X	V	V	V
AENet	V	X	X	V
Reconstruction	V	V	V	V
Y	26.75	26.67	**26.79**	**26.95**
RGB	25.24	25.15	**25.27**	**25.42**

- ✓ 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.

Qualitative Comparison Results



(a) The original. (b) Bicubic. (c) SOF-VSR [21]. (d) WDVR [8]. (e) FRVSR [3]. (f) **WAEN P.** (g) **WAEN S.** (h) GT.

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

Conclusions

- ✓ 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**.