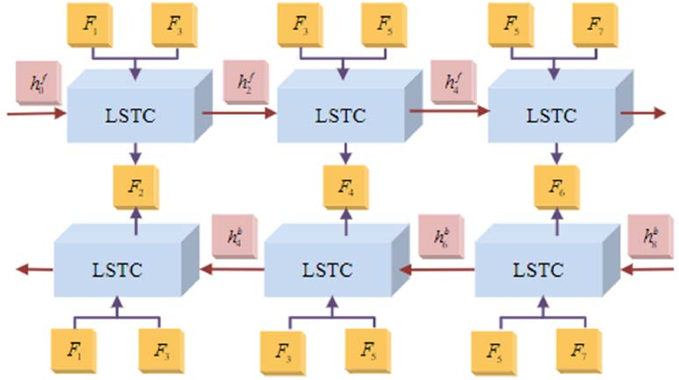
Space-Time Video Super-Resolution Using Deformable Attention Network

Hai Wang[1], Xiaoyu Xiang[2]*, Yapeng Tian[3], Wenming Yang[1], Qingmin Liao[1]

[1]Tsinghua University, [2]Meta Reality Lab, [3]University of Rochester

Introduction

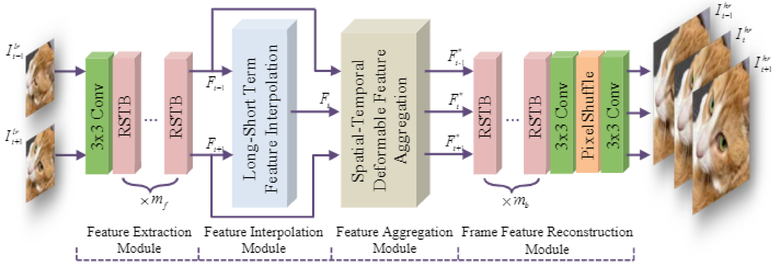
- **Motivation**: Most of space-time video super-resolution (STVSR) only use two adjacent frames, that is, short-term features, to synthesize the missing frame embedding, which cannot fully explore the information flow of consecutive input low-resolution frames. In addition, existing STVSR methods hardly exploit the temporal contexts explicitly to assist high-resolution frame reconstruction.
- **Key Ideas:** we propose a deformable attention network called STDAN for STVSR.
  - First, we devise a long-short term feature interpolation (LSTFI) module, which is capable of excavating abundant content from more neighboring input frames for the interpolation process through a bidirectional RNN structure.
  - Second, we put forward a spatial-temporal deformable feature aggregation (STDFA) module, in which spatial and temporal contexts in dynamic video frames are adaptively captured and aggregated to enhance super-resolution reconstruction.



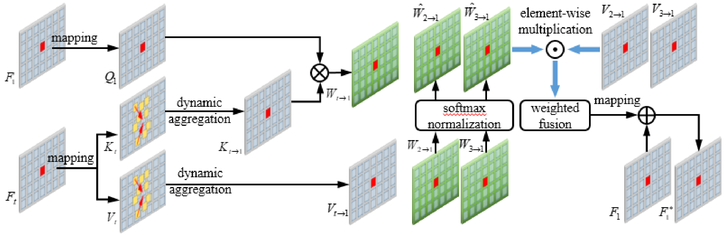
LSTFI module consists of long-short term cells (LSTCs) with bidirectional RNN, which can fully exploit the whole input video frame features during the interpolation process.

*Corresponding author

Framework



- **Spatial-Temporal Deformable Feature Aggregation:** through deformable attention, the cross-frame spatial aggregation phase dynamically fuses useful content from different frames. The adaptive temporal aggregation phase mixes the temporal contexts among these fused frame features further to acquire enhanced features.



- **Three different aggregation methods:** the feature vector (red point) attends the valuable spatial content (yellow points) in a (a) 1x1 window, (b) 3x3 window, and (c) deformable window.