



Space-Time Video Super-Resolution Using Deformable Attention Network

Hai Wang¹, Xiaoyu Xiang^{2*}, Yapeng Tian³, Wenming Yang¹, Qingmin Liao¹

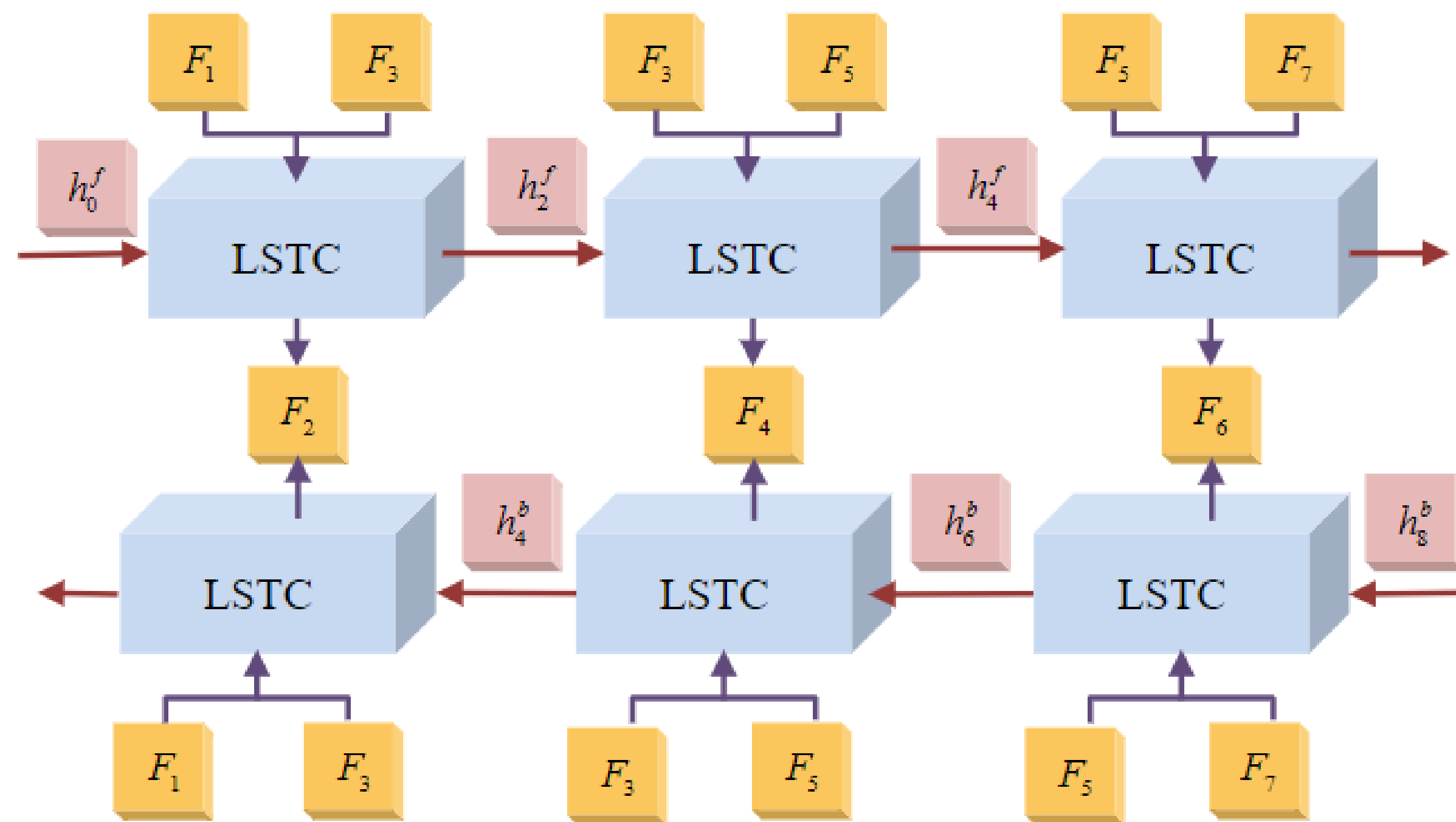
¹Tsinghua University, ²Meta Reality Lab, ³University of Rochester



T4V: Transformers for Vision

Introduction

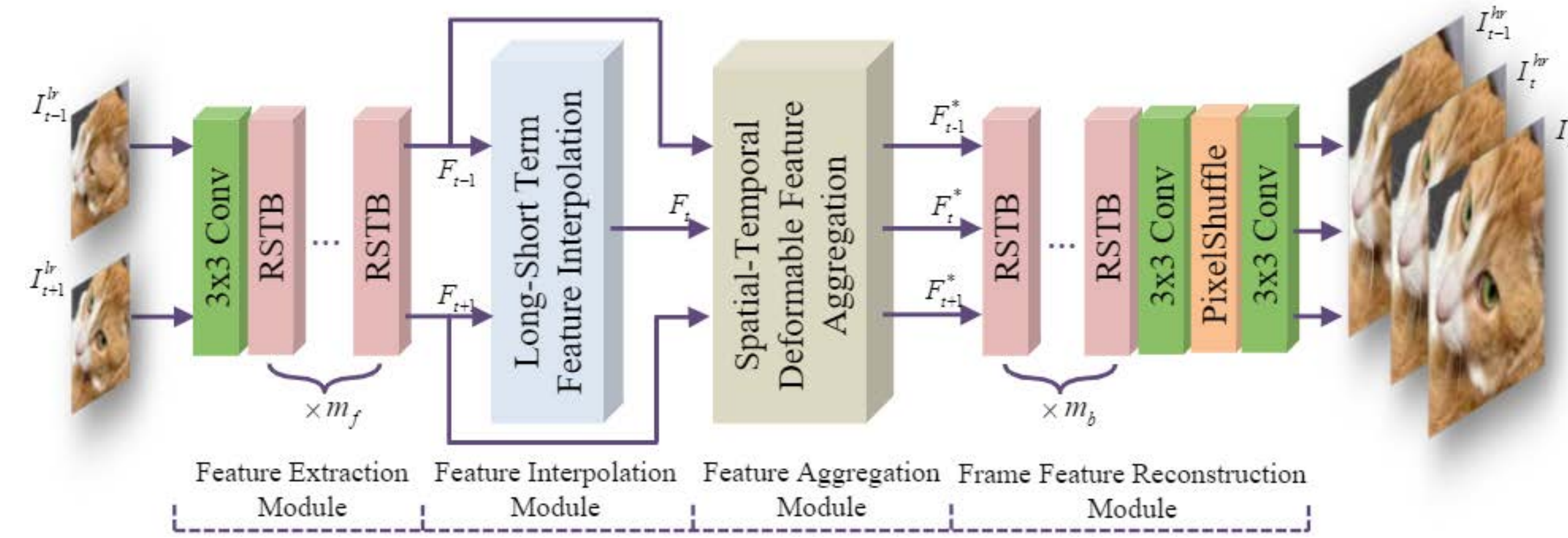
- **Motivation:** Most space-time video super-resolution (STVSR) models only use two adjacent frames, that is, short-term features, to synthesize the missing frame embedding, which cannot fully explore the information flow of consecutive input low-resolution frames. In addition, existing STVSR methods 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.



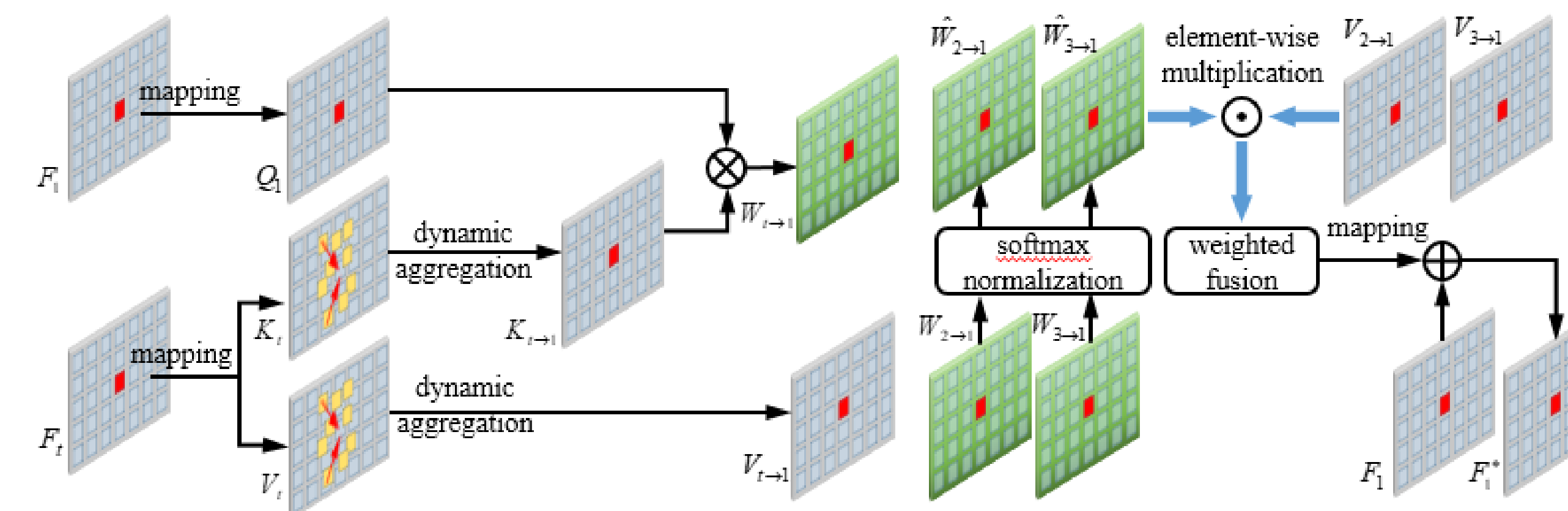
LSTFI module consists of long-short term cells (LSTCs) with bidirectional RNN, which can fully exploit the whole input video frame features during the interpolation process.

*Corresponding author

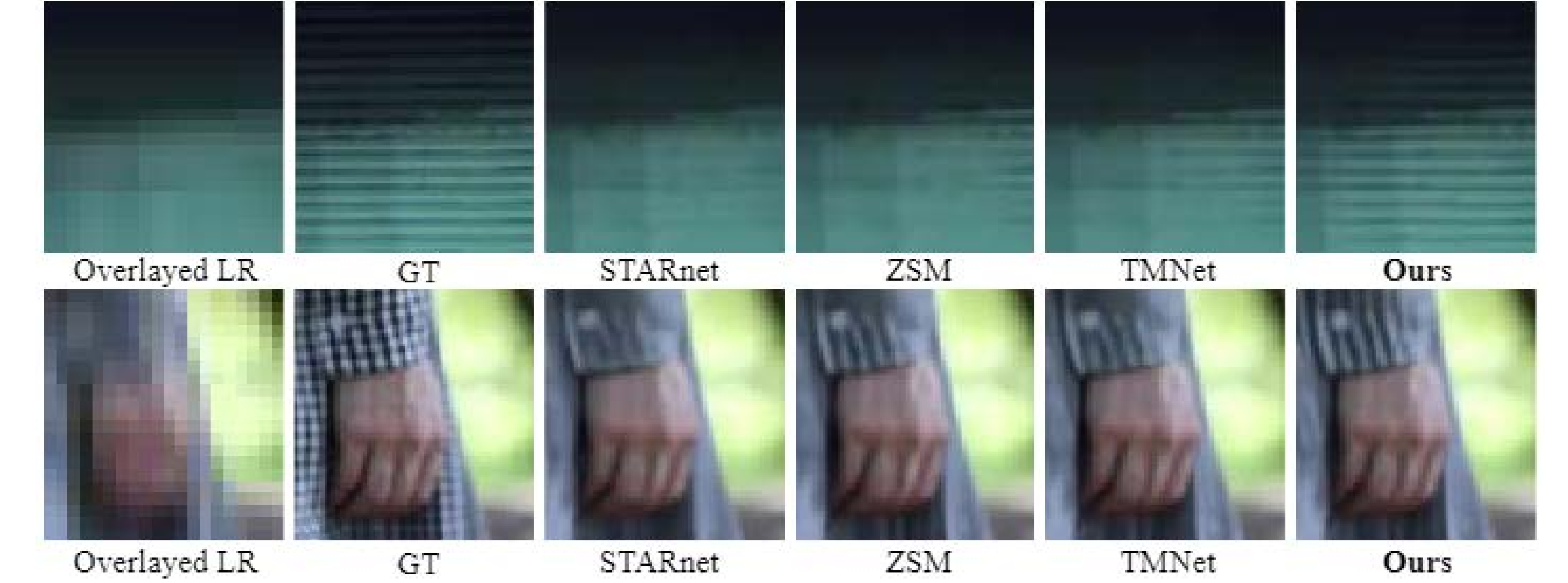
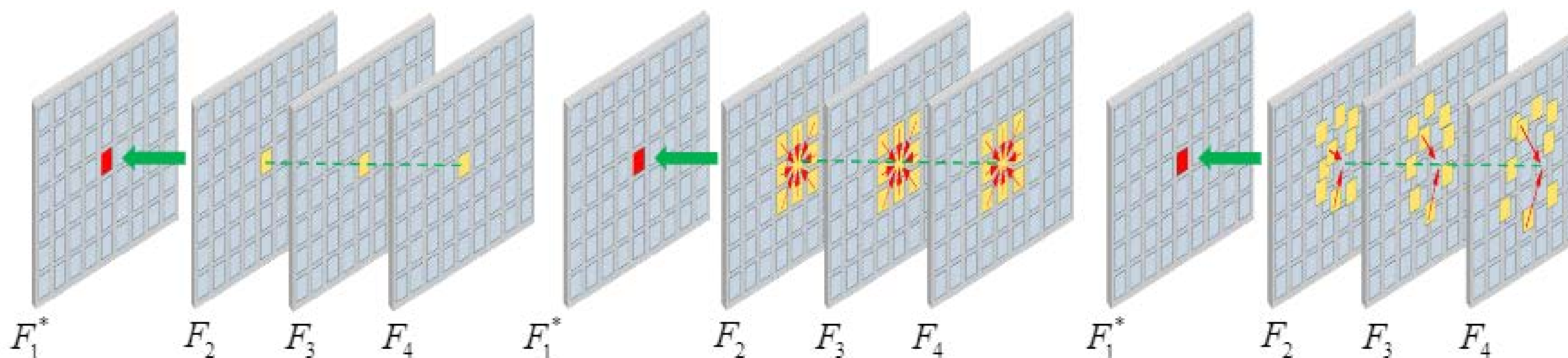
Framework



- **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.



Experiments

Quantitative Comparisons

STVSR Method	Vid4 PSNR	Vid4 SSIM	Vimeo-Slow PSNR	Vimeo-Slow SSIM	Vimeo-Medium PSNR	Vimeo-Medium SSIM	Vimeo-Fast PSNR	Vimeo-Fast SSIM	Params (M)
STARnet [1]	25.99	0.7819	33.10	0.9164	34.86	0.9356	36.19	0.9368	111.61
ZSM [7]	26.14	0.7974	33.36	0.9138	35.41	0.9361	36.81	0.9415	11.10
TMNet [8]	26.23	0.8011	33.51	0.9159	35.60	0.9380	37.04	0.9435	12.26
STDAN (Ours)	26.28	0.8041	33.66	0.9176	35.70	0.9387	37.10	0.9437	8.29

- ✓ We can see that our STDAN with the least parameters obtains state-of-the-art performance on both Vid4 and Vimeo.

Ablation Study

Method	Ω ₁	Ω ₂	Ω ₃	Ω ₄	Ω ₅
Params (M)	5.44	5.54	5.54	5.82	8.29
Feature Interpolation	✓	✓	✓	✓	✓
Feature Aggregation	1×1 fixed window	✓			
	3×3 fixed window		✓		
	deformable window			✓	✓
Vid4 (slow motion)	25.27	25.69	25.85	25.97	26.28
Vimeo-Fast (fast motion)	35.88	36.22	36.41	36.63	37.10

- ✓ Feature aggregation module can improve the reconstruction results.
- ✓ The larger the spatial range of feature aggregation, the more useful information can be captured to enhance recovery quality of HR frames.
- ✓ Combining long-term and short-term information can achieve better feature interpolation results.