Structure Preserving Video Prediction

CVPR 2018, Poster

Jingwei Xu, Bingbing Ni, Zefan Li, Shuo Cheng and Xiaokang Yang
Shanghai Jiao Tong University

Presented by Yuzhang Hu
2018.12.9

Outline

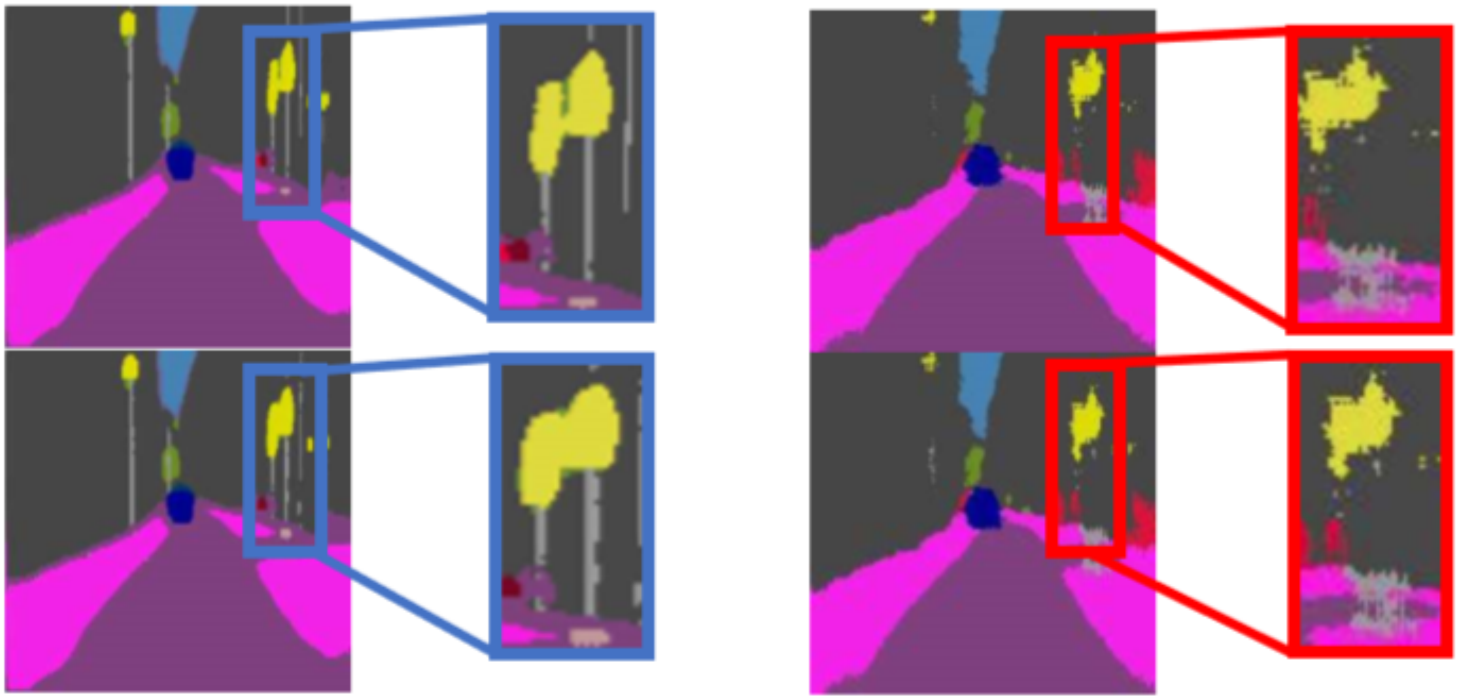
- Authorship
- Background
- Proposed method
- Training
- Experiment
- Conclusion

Background

- Video prediction is a long-standing task but faces two problems
  - Rich structure information like object boundary
  - Detailed motion like body movement

Background

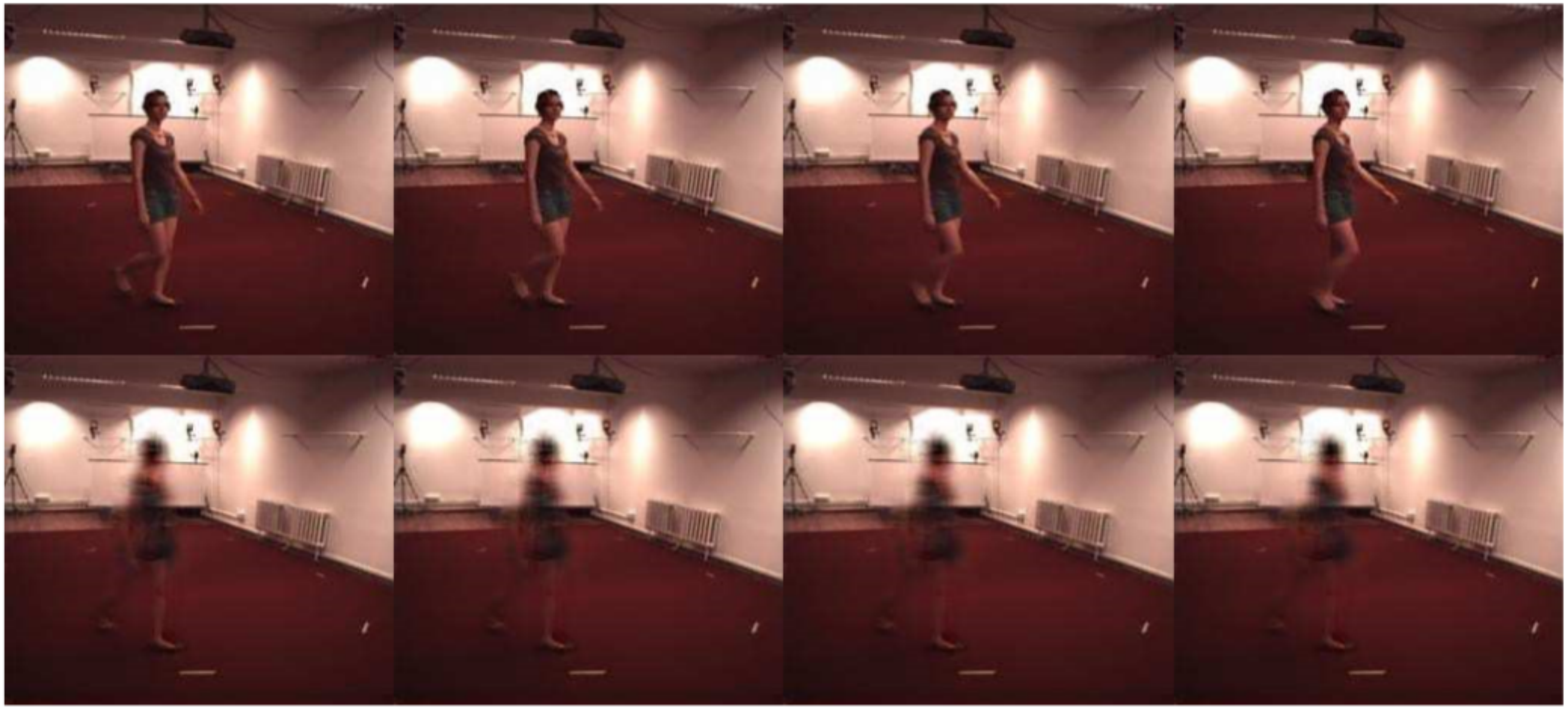
- **Static Structure loss**



result from the motion of camera

Background

- **Dynamic Structure Loss**



moving direction of body parts are different

Proposed Method

$$\mathcal{X} = (\mathcal{X}_1, ..., \mathcal{X}_T)$$

Encoder

$$\mathcal{F}^L = (\mathcal{F}_1^L, ..., \mathcal{F}_T^L) \qquad \mathcal{F}^H = (\mathcal{F}_1^H, ..., \mathcal{F}_T^H)$$

Prediction module

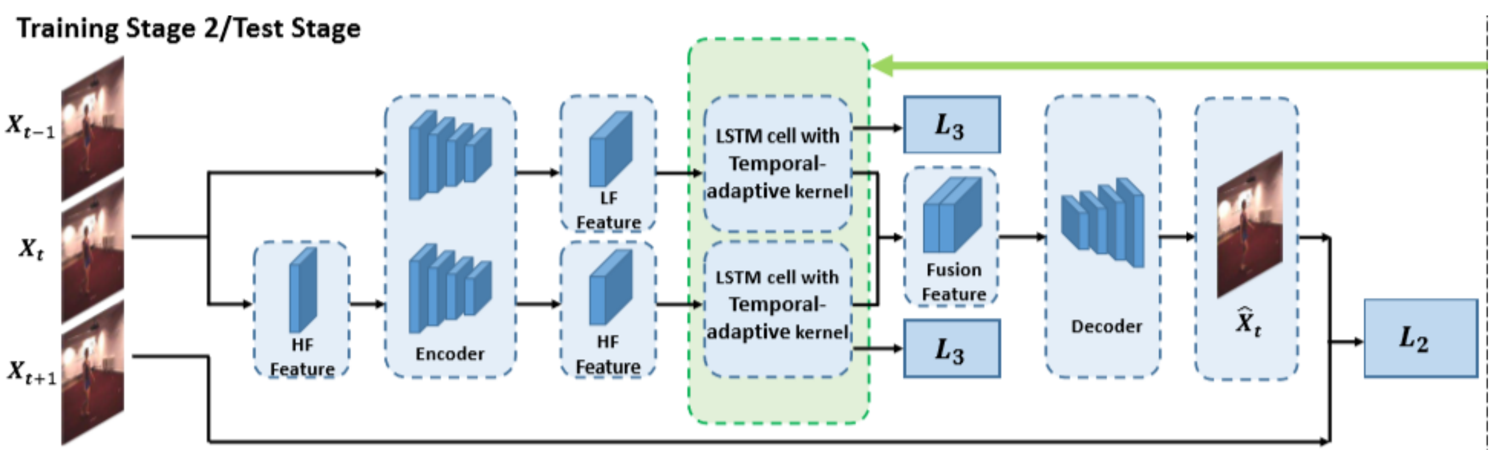
$$(\hat{\mathcal{F}}_{N+1}, ..., \hat{\mathcal{F}}_{N+M}) \qquad (\hat{\mathcal{F}}_{N+1}, ..., \hat{\mathcal{F}}_{N+M})$$

Decoder

$$(\hat{\mathcal{X}}_1, ..., \hat{\mathcal{X}}_N, \hat{\mathcal{X}}_{N+1}, ..., \hat{\mathcal{X}}_{N+M})$$

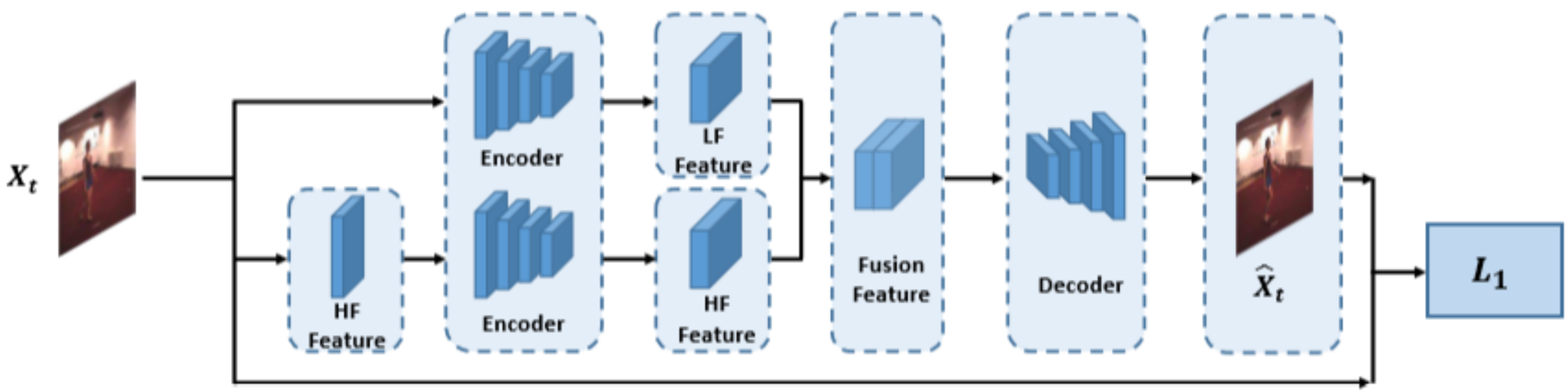
Three parts:

- Encoder
- Prediction module
- Decoder



Proposed Method

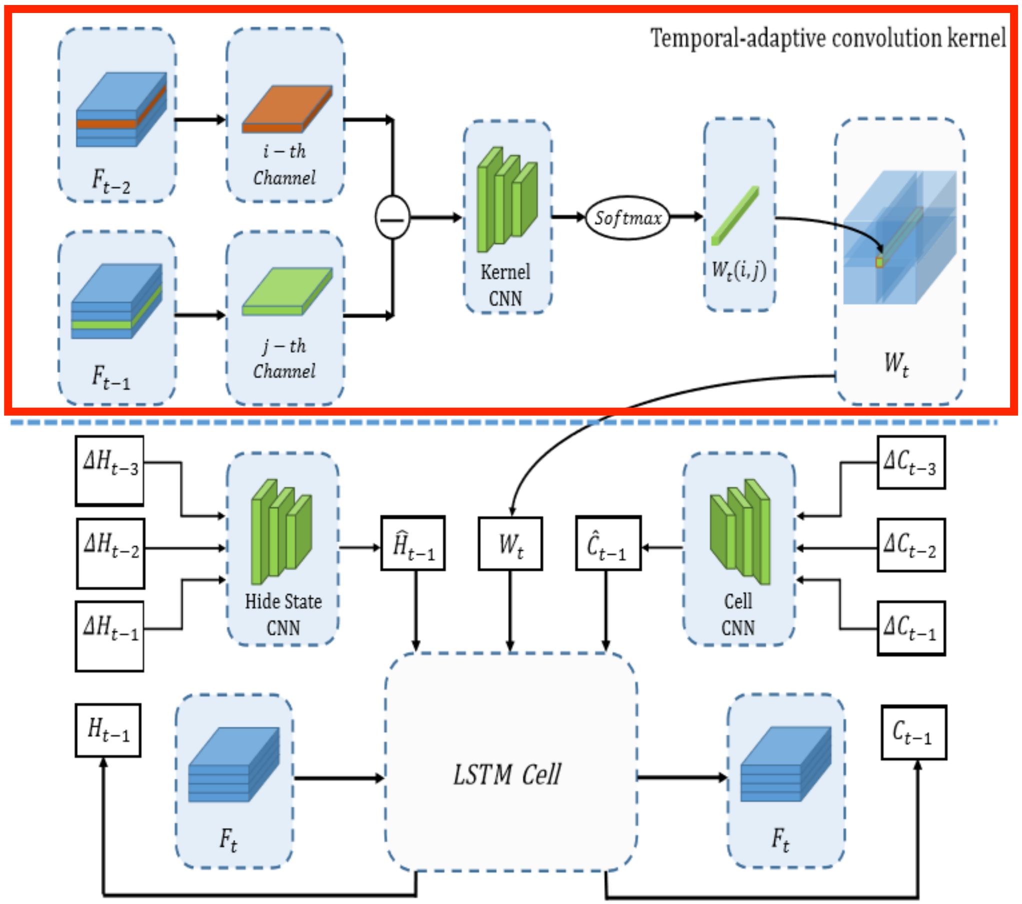
Two-branch video prediction framework



- The two-branch encoders designed for two different frequency domains
- raw pixels directly passed to the first encoder
- process the raw inputs with a high pass filter and then encode

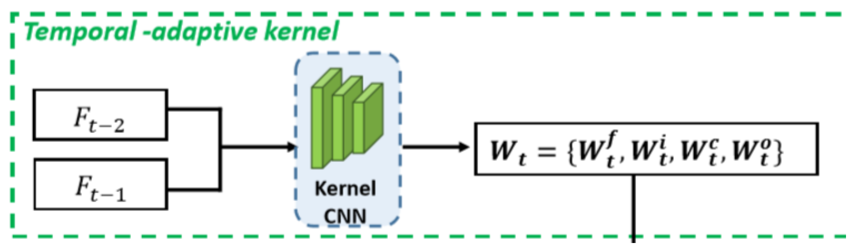
Proposed Method

Temporal adaptive prediction module



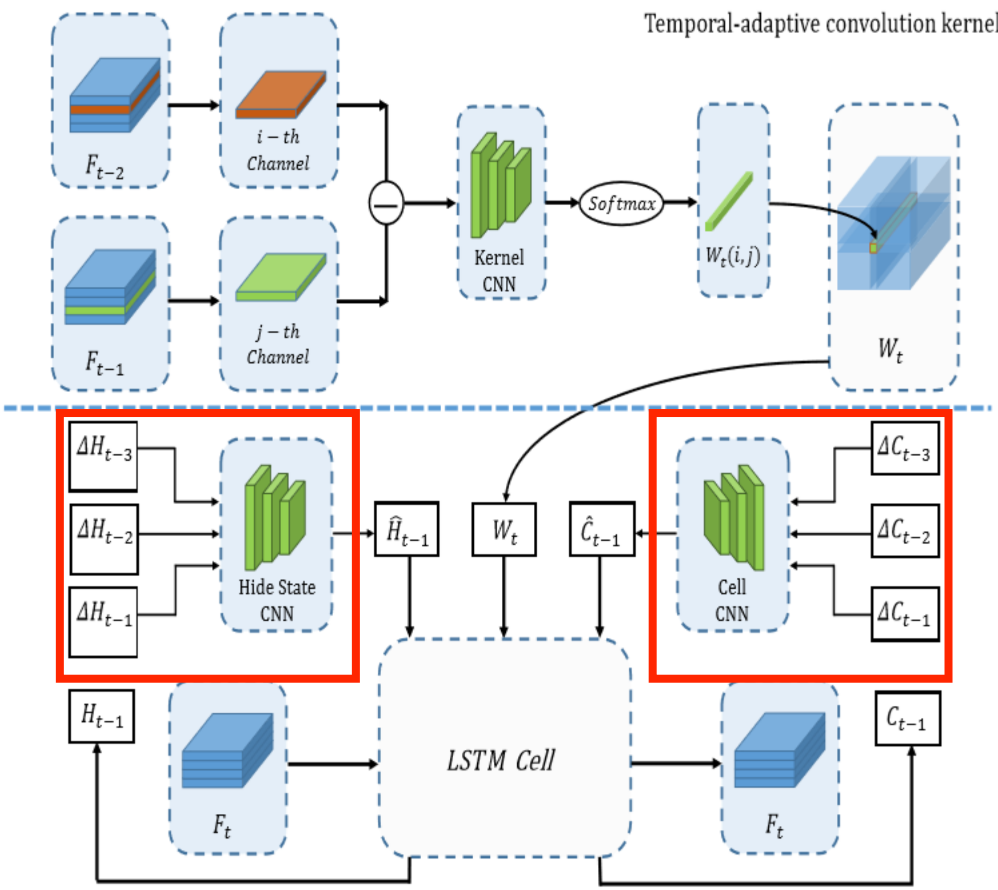
To fully utilize the temporal variation information:

$$\widetilde{\mathcal{W}}_t(i,j) = \tilde{\phi}_{\mathcal{W}}(\mathcal{F}_t(i) - \mathcal{F}_{t-1}(j))$$



Proposed Method

Temporal adaptive prediction module



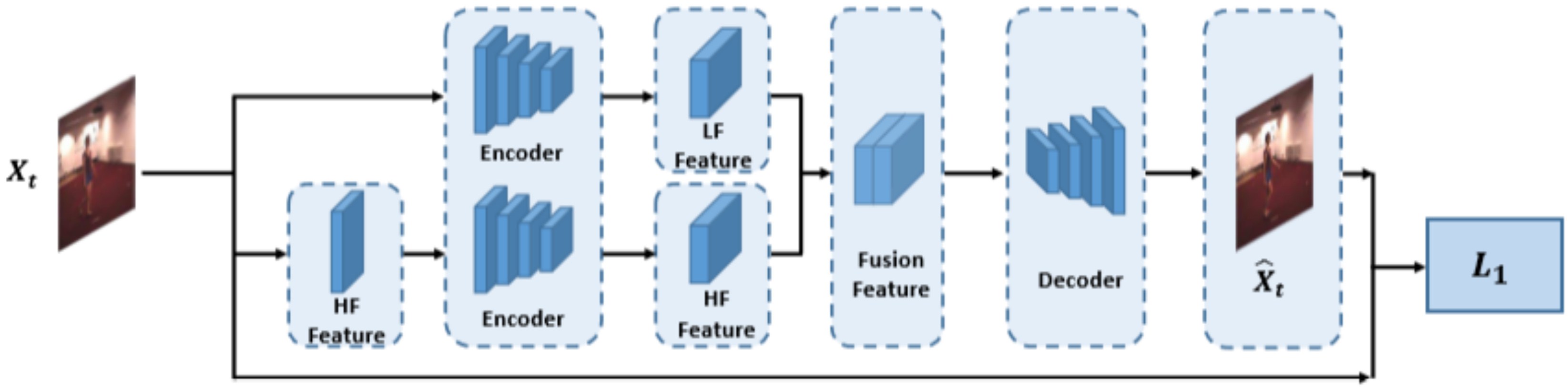
Aim to purse a more efficient temporal information sharing mechanism to facilitate the video prediction task

$$\Delta\mathcal{H}_{t-i} = \mathcal{H}_{t-(i+1)} - \mathcal{H}_{t-(i+2)}, i = 1, 2, 3,$$
$$\Delta\mathcal{C}_{t-i} = \mathcal{C}_{t-(i+1)} - \mathcal{C}_{t-(i+2)}, i = 1, 2, 3.$$

Training

Divide the training process into two process

Stage #1



$$\mathcal{L}_1 = ||\mathcal{X} - \hat{\mathcal{X}}||_1 + ||HF(\mathcal{X}) - HF(\hat{\mathcal{X}})||_1,$$

Training

Divide the training process into two process

Stage #2

$$\mathcal{L}_1 = ||\mathcal{X} - \hat{\mathcal{X}}||_1 + ||HF(\mathcal{X}) - HF(\hat{\mathcal{X}})||_1,$$

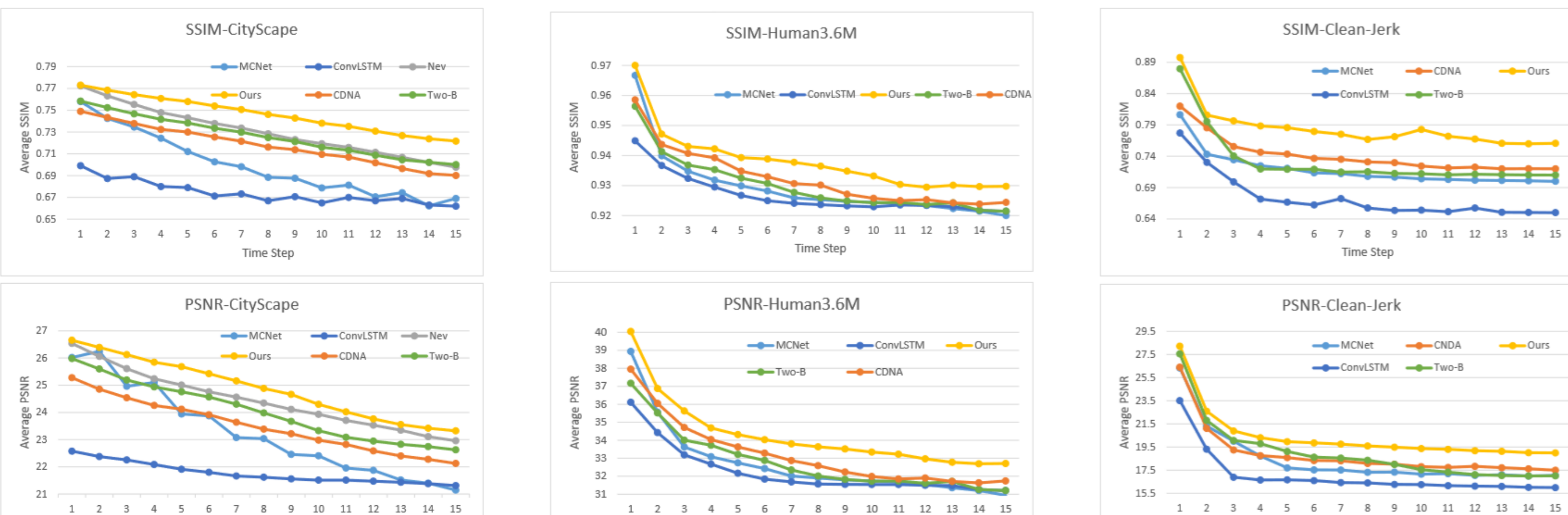
$$\mathcal{L}_2 = \sum_{i=1}^{N+M-1} (||\mathcal{X}_{i+1} - \hat{\mathcal{X}}_i||_1 + ||\mathcal{F}_{i+1} - \hat{\mathcal{F}}_i||_1 + ||HF(\mathcal{X}_{i+1}) - HF(\hat{\mathcal{X}}_i)||_1).$$

$$\mathcal{L}_3 = \frac{1}{N+M} \sum_{t=1}^{N+M} ||(||\mathcal{F}_t - \hat{\mathcal{F}}_t||_1) - \sigma_{ths}||_1$$

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_3 + \lambda_4 \sum ||\Theta||_2^2,$$

Experiment

Quantitative Evaluation



Experiment

Qualitative Evaluation

Model	CityScape/Human3.6M/Clean-Jerk PSNR	SSIM
ConvLSTM	22.8/36.2/23.4	0.70/0.94/0.78
Two-B	25.2/37.2/25.3	0.74/0.96/0.85
Two-B+Fus-4	25.7/37.5/25.7	0.76/0.96/0.85
Two-B+Fus-4+Tem-K	**26.6/39.7/27.5**	**0.77/0.97/0.89**

Two-B: two-branch framework
Fus-4: Use of the hidden state of the last 4 time-steps
Tem-K: kernel generation

Conclusion

Solution
- Two-branch video prediction framework
- Temporal adaptive prediction module

Discussion
- Combine pixel domain with frequency domain
- Apply LSTM in inter prediction
- Transfer from inter to intra

Thanks!