MotionRNN: A Flexible Model for Video Prediction with Spacetime-Varying Motions

Haixu Wu*, Zhiyu Yao*, Jianmin Wang, Mingsheng Long (✉)
School of Software, BNRist, Tsinghua University, China
{whx20,yaozy19}@mails.tsinghua.edu.cn, {jimwang,mingsheng}@tsinghua.edu.cn

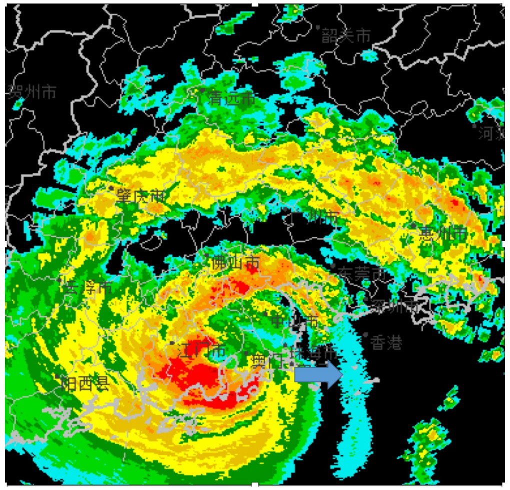

Haixu Wu


Zhiyu Yao

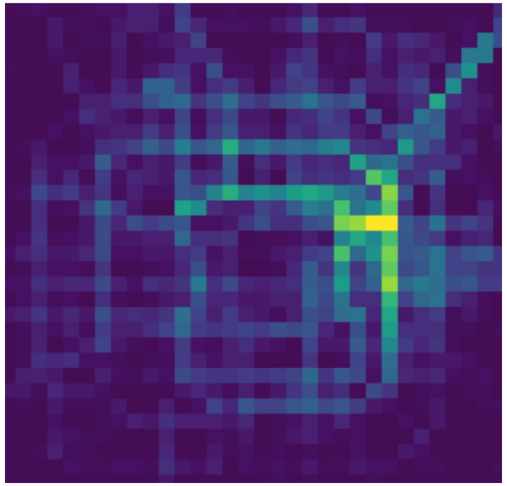

Mingsheng Long


Jianmin Wang

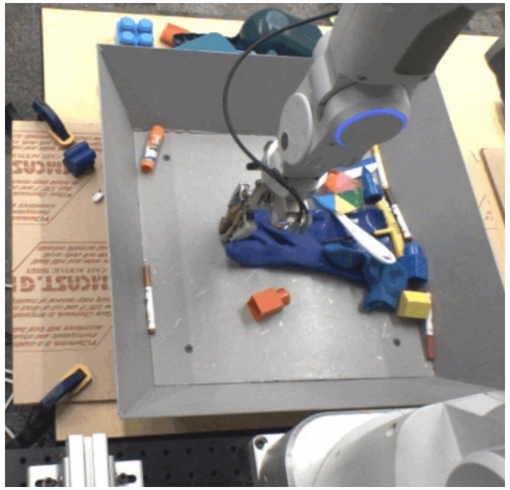
Video Prediction



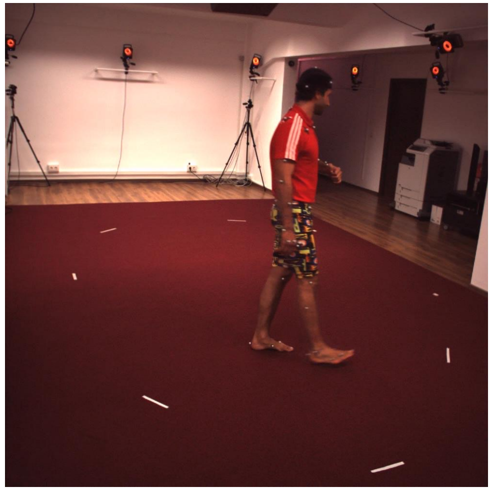
Precipitation nowcasting
Radar



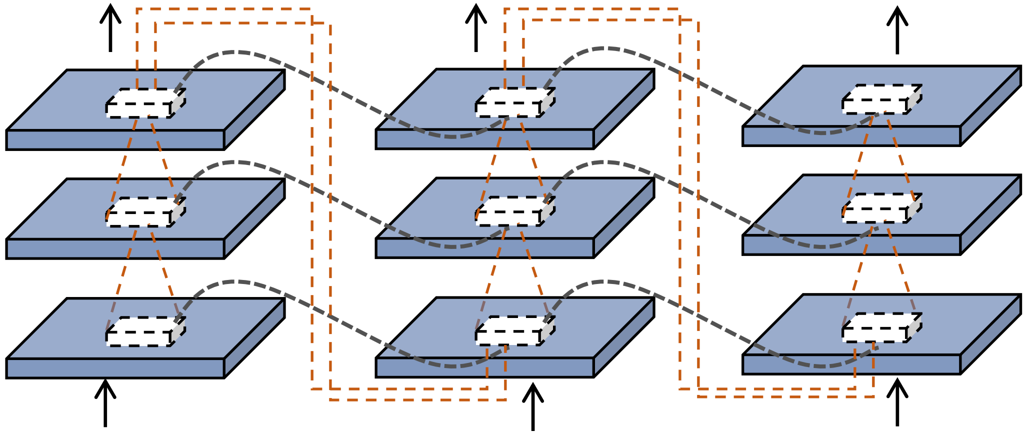
Traffic Planning
TaxiBJ



Visual Foresight
Bair Robot Pushing



Pedestrians Forecasting
Human3.6M



LSTM-based Predictive Models

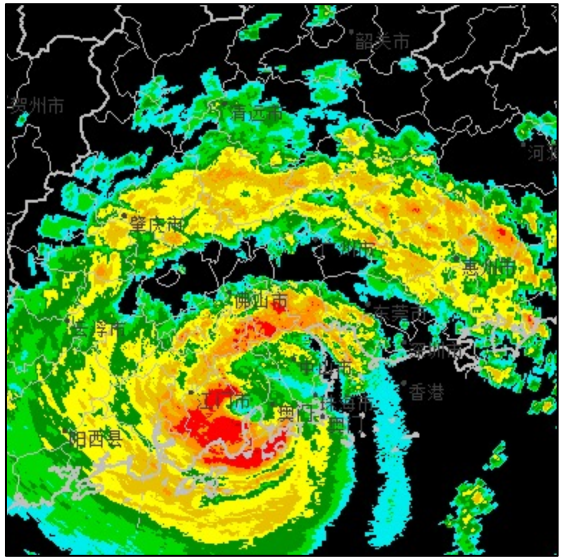
Modeling Spatiotemporal State Transition with gate mechanism.

Gate are easily saturated.

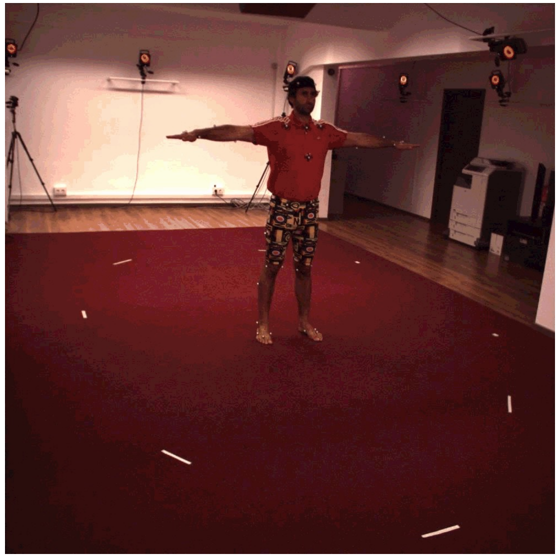
Motion Vanished.

ConvLSTM[Shi et al. NeurIPS15], PredRNN[Wang et al. NeurIPS17], MIM[Wang et al. CVPR19], E3D-LSTM[Wang et al. ICLR19]

Spacetime-Varying Motions



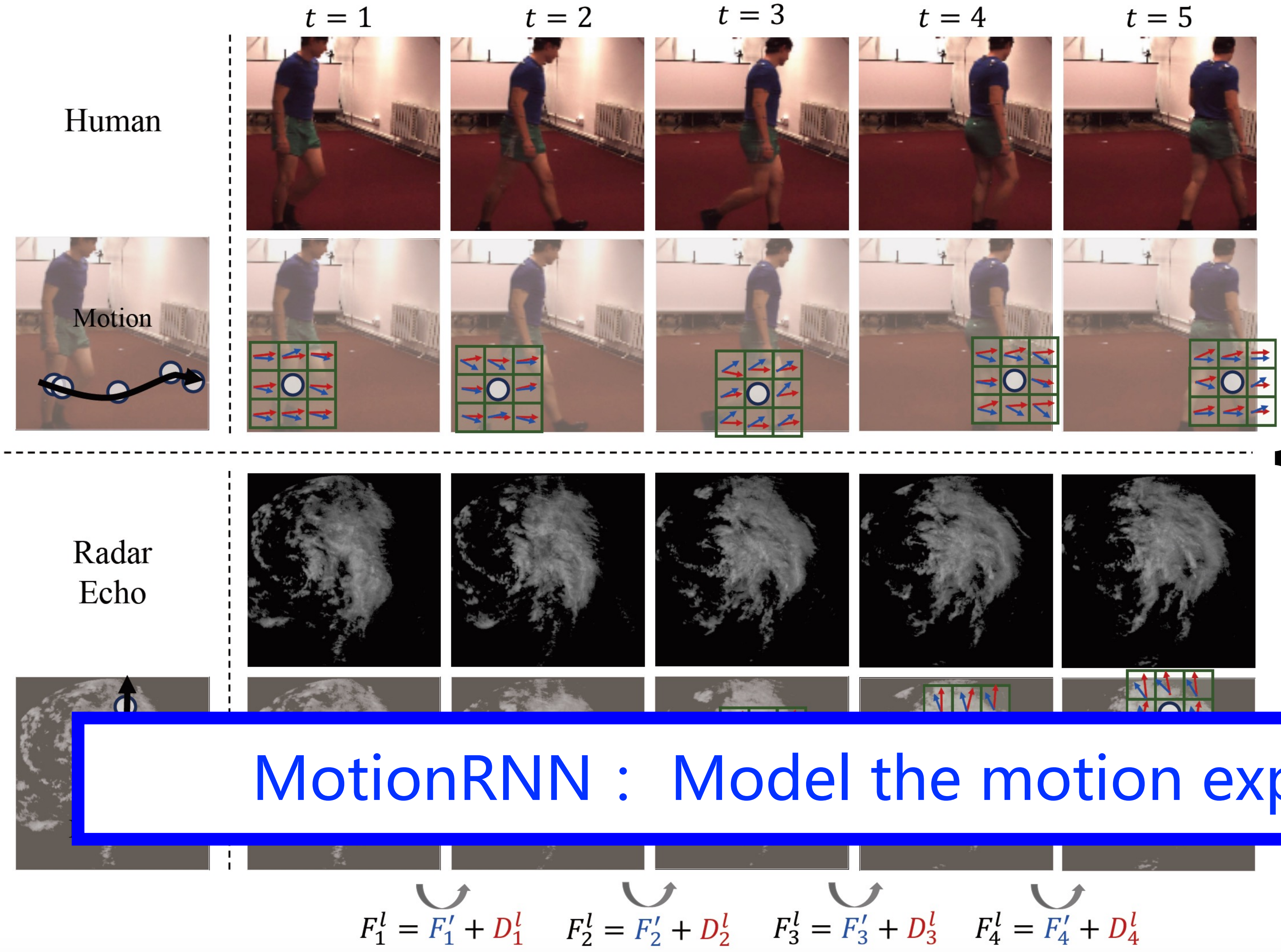
Clouds generation, dissipation, translation, deformation...



Change direction, speed, and body movement...

Challenge: Real-world motions are ever-changing

Motion Decomposition

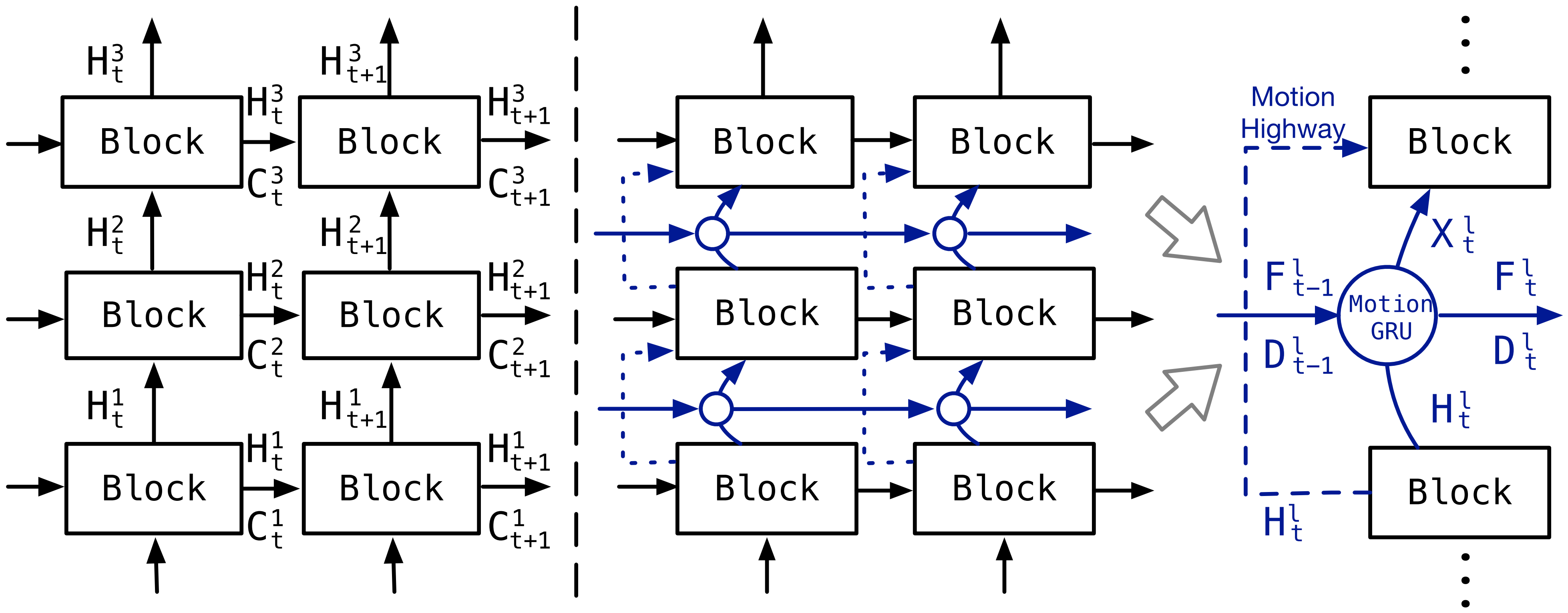


Transient Variation

Motion Trend

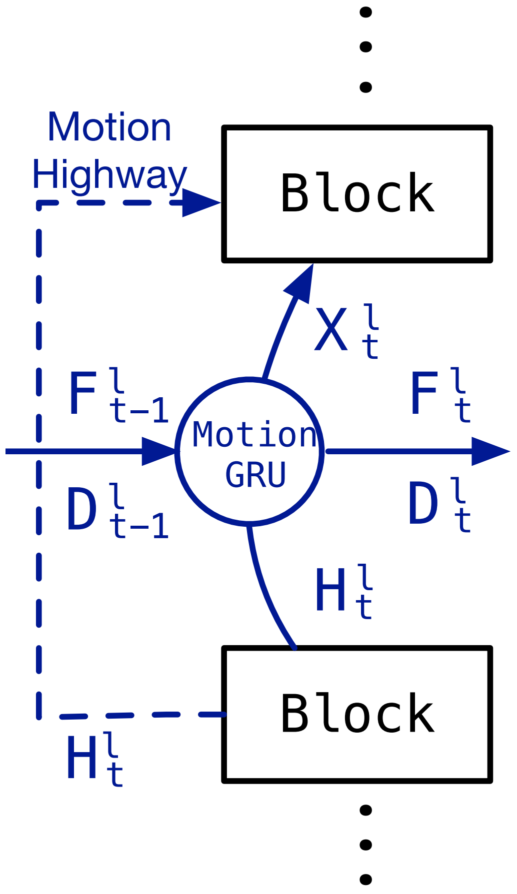
MotionRNN： Model the motion explicitly

MotionRNN



Flexible Model: can be applied to any LSTM-based models

MotionRNN



MotionGRU

Modeling the motion explicitly.

$$\mathcal{X}_t^l, \mathcal{F}_t^l, \mathcal{D}_t^l = \text{MotionGRU}(\mathcal{H}_t^l, \mathcal{F}_{t-1}^l, \mathcal{D}_{t-1}^l)$$

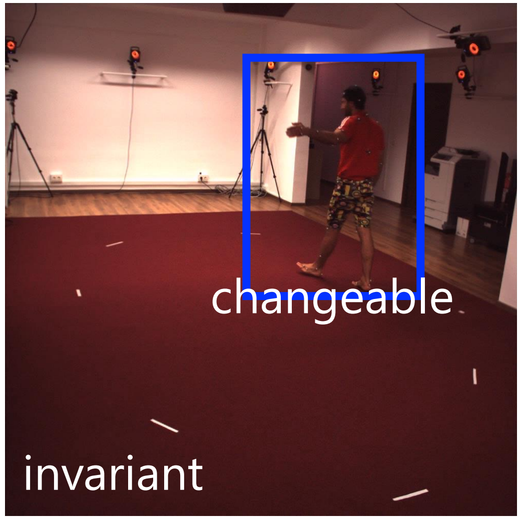
$$\mathcal{H}_t^{l+1}, \mathcal{C}_t^{l+1} = \text{Block}(\mathcal{X}_t^l, \mathcal{H}_{t-1}^{l+1}, \mathcal{C}_{t-1}^{l+1})$$

LSTM-based Predictive Block

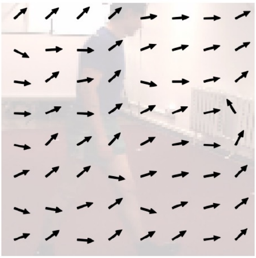
$$\boxed{\mathcal{H}_t^{l+1} = \mathcal{H}_t^{l+1} + (1 - o_t) \odot \mathcal{H}_t^l,}$$

Motion Highway

Balance the invariant part and the changeable motion part.

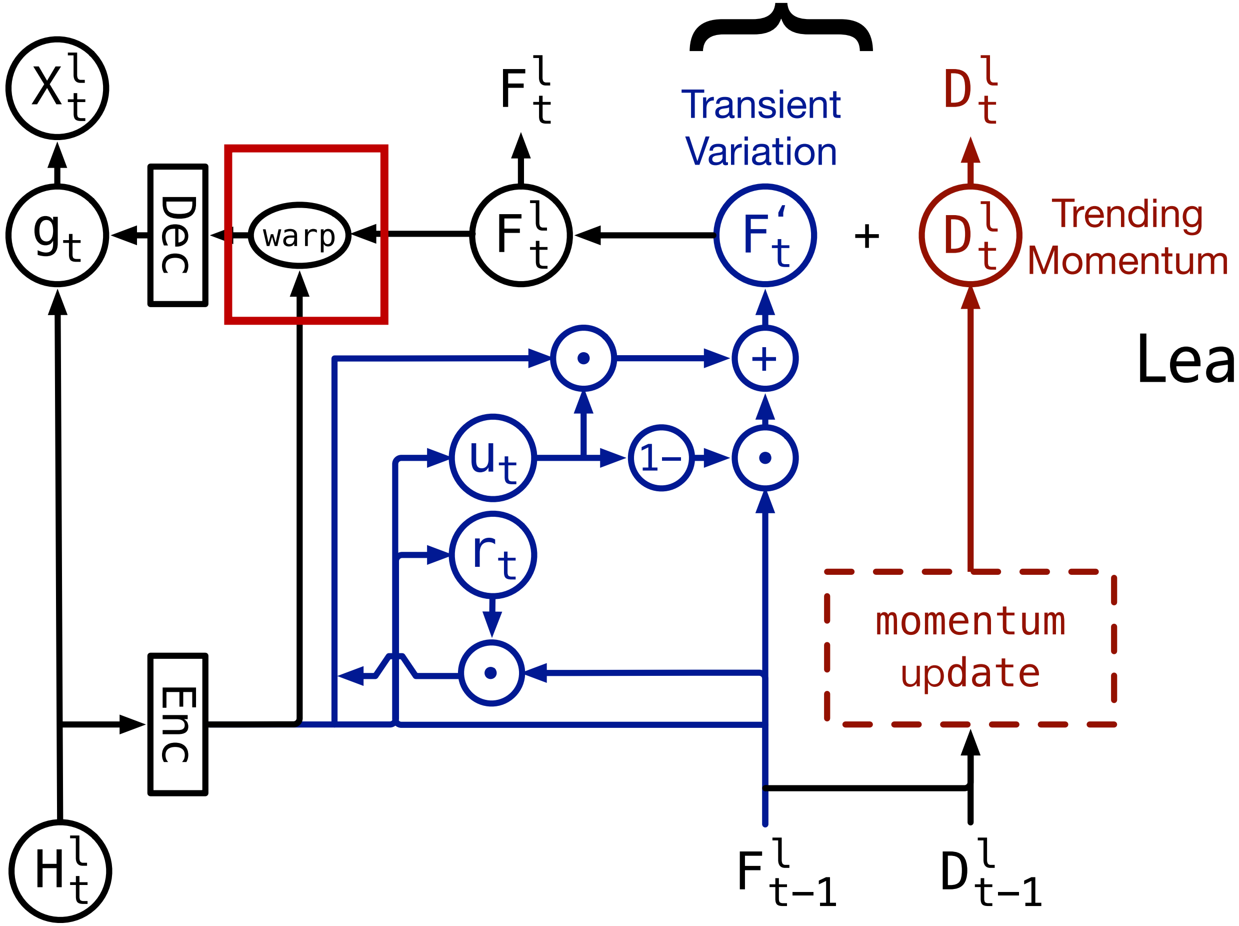


MotionGRU



Offset space

Pixel-wise displacement

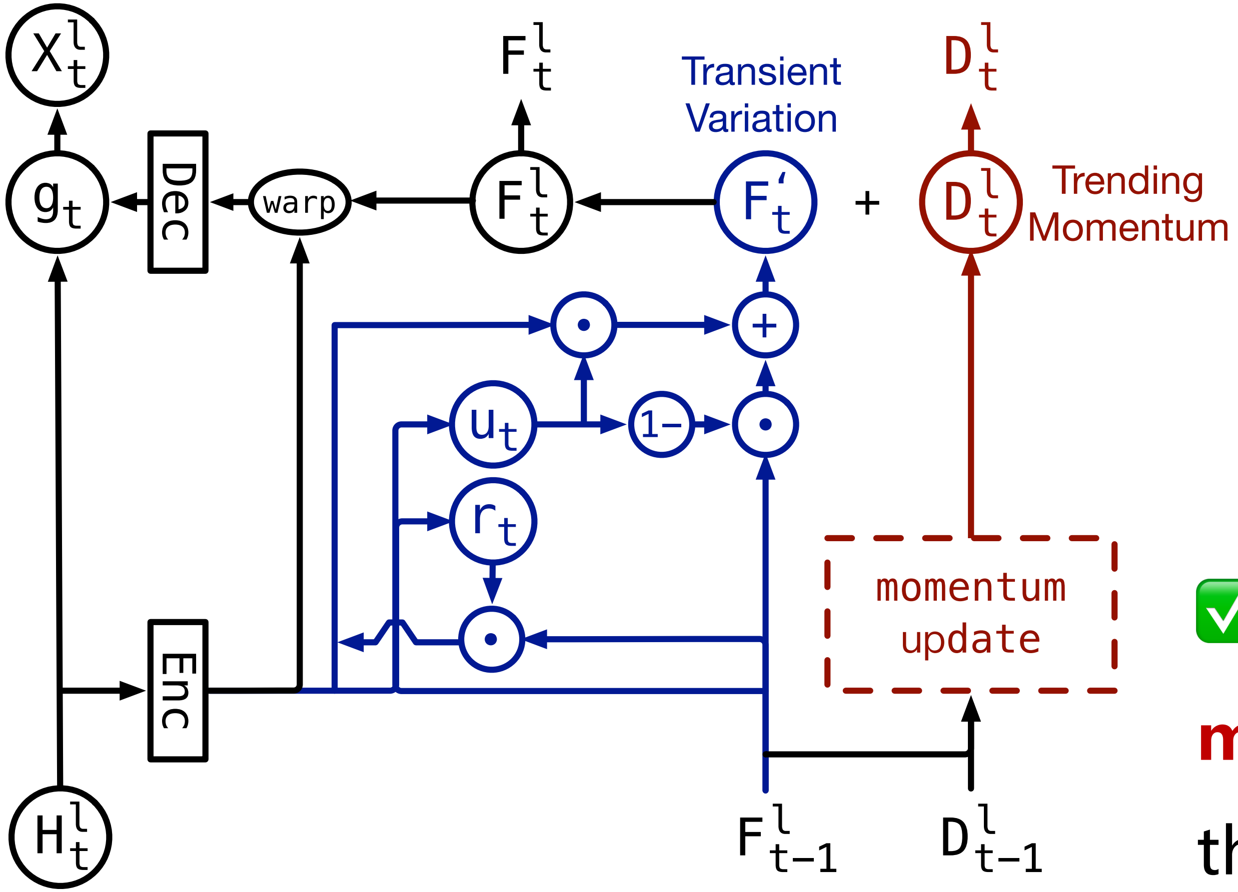


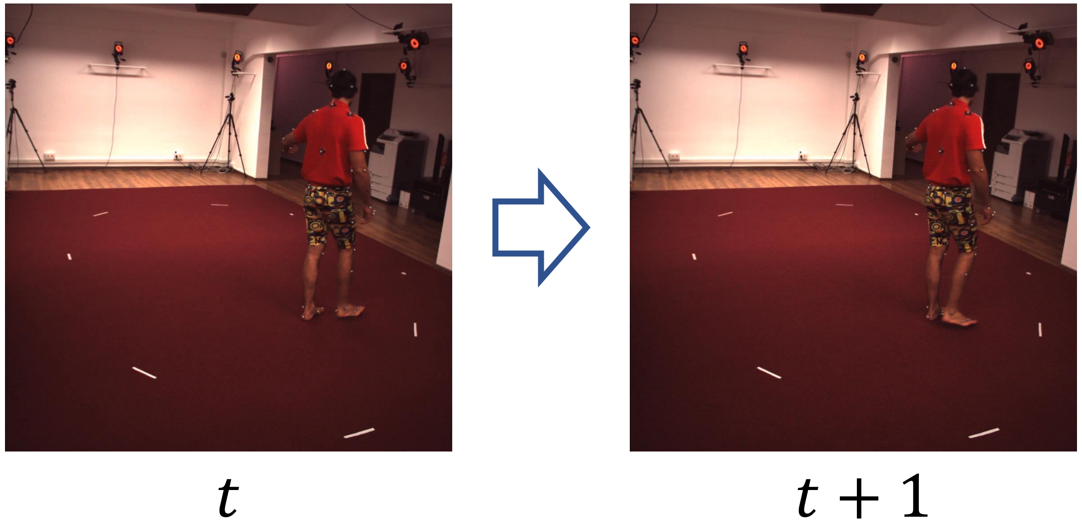
Transient Variation

Learned Motion Filter $\mathcal{F}_t^l = \mathcal{F}_t' + \mathcal{D}_t^l$

Motion Trend

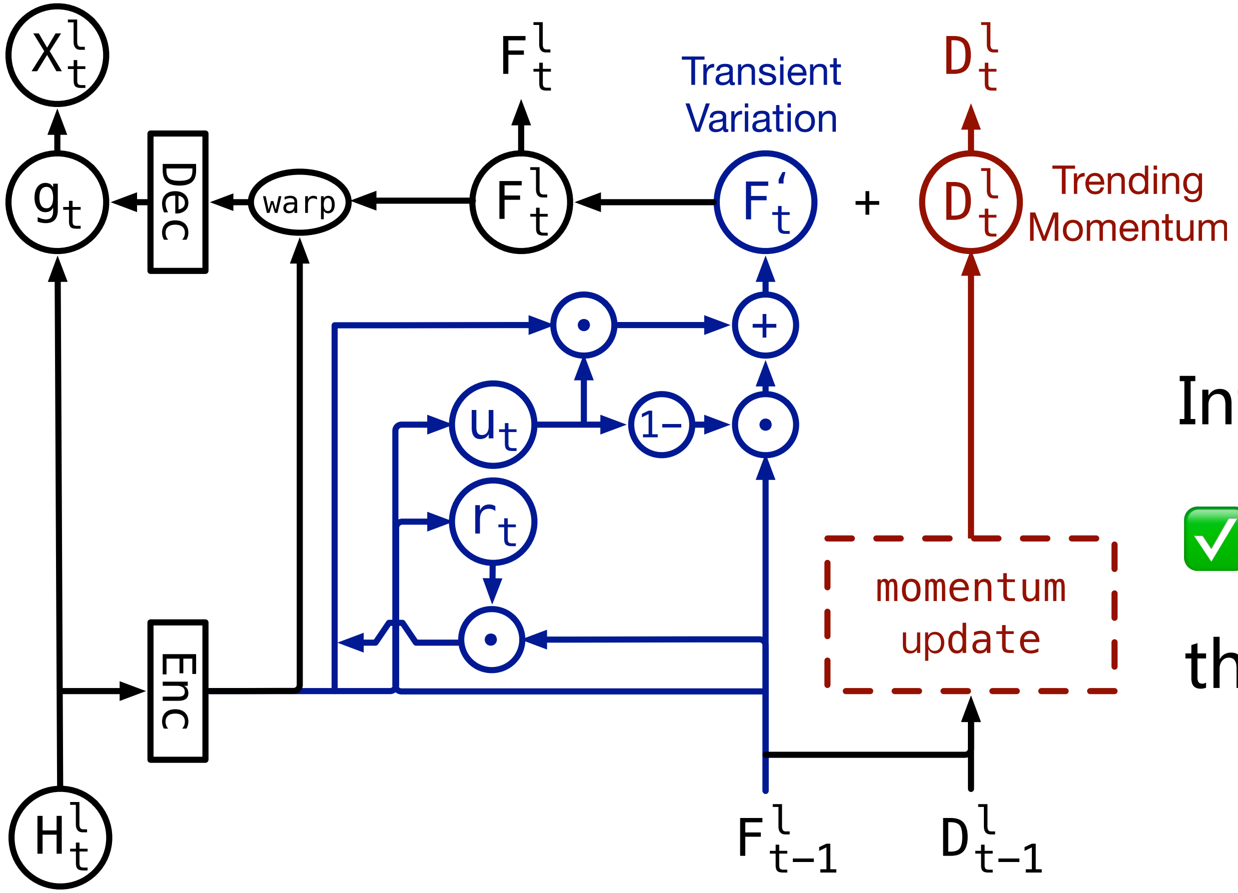
Transient Variation

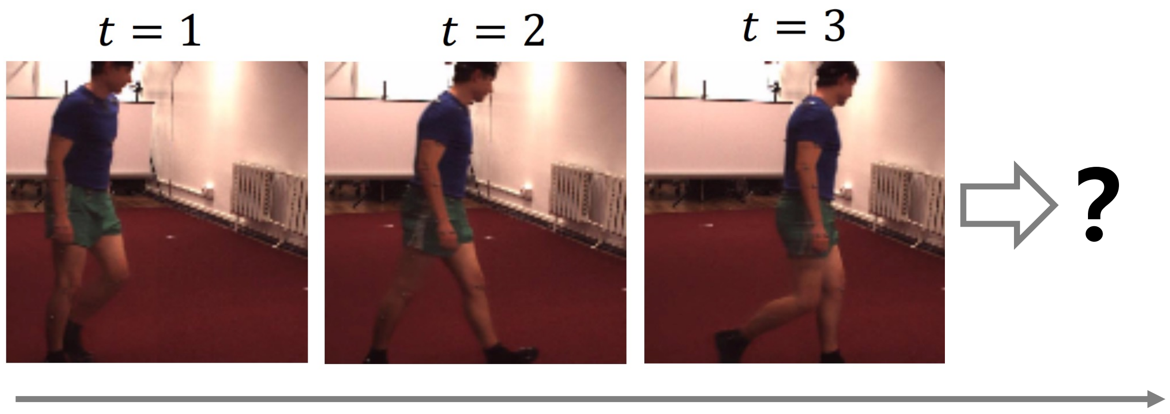




✅Using ConvGRU to model the **motion filter** transition and maintain the spatiotemporal coherence.

Motion Trend





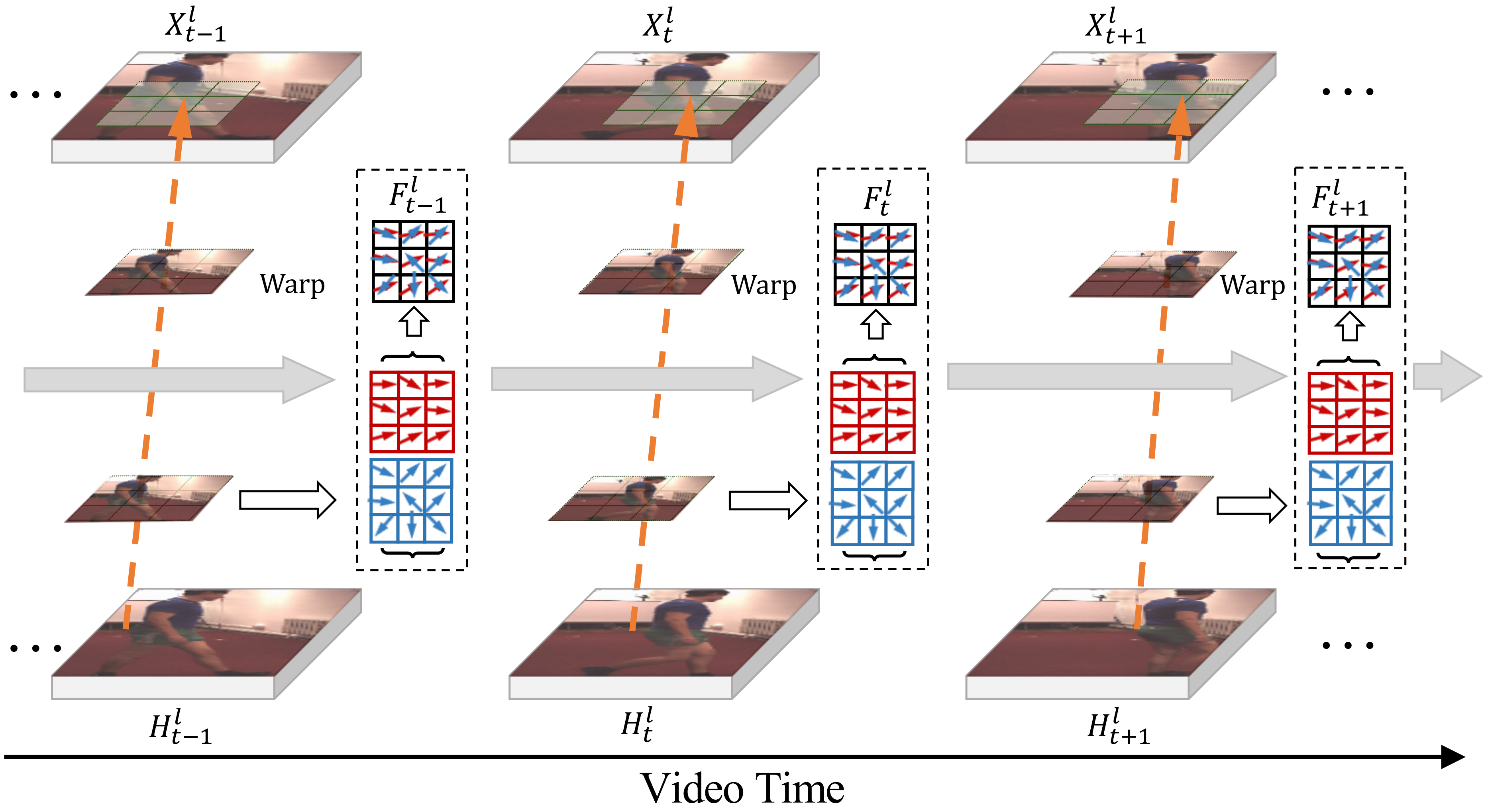
Inference the trend: with future unknown

✅Using Temporal Difference to estimate the motion tendency.

$$\mathcal{D}_t^l = \mathcal{D}_{t-1}^l + \alpha \left(\mathcal{F}_{t-1}^l - \mathcal{D}_{t-1}^l \right)$$

Will converge to the weighted sum of motion filters.

Overall Procedure



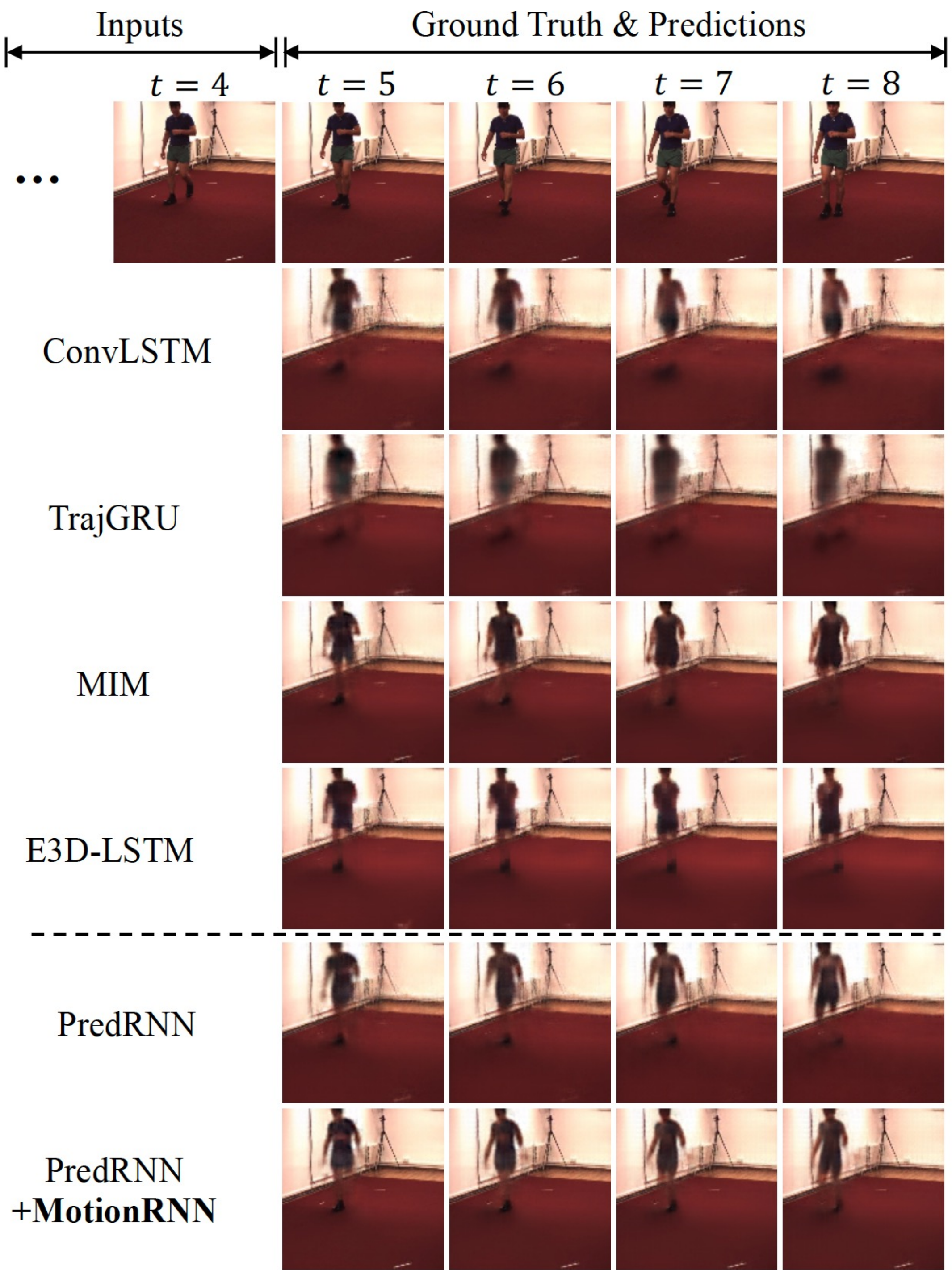
Human Motion（Human3.6M）

Method	SSIM	MSE/10	MAE/100	FVD
TrajGRU [22]	0.801	42.2	18.6	26.9
Conv-TT-LSTM [26]	0.791	47.4	18.9	26.2
ConvLSTM [21]	0.776	50.4	18.9	28.4
+ MotionRNN	0.800	44.3	18.6	26.9
MIM [37]	0.790	42.9	17.8	21.8
+ MotionRNN	0.841	35.1	14.9	18.3
PredRNN [36]	0.781	48.4	18.9	24.7
+ MotionRNN	0.846	**34.2**	**14.8**	**17.6**
E3D-LSTM [35]	0.869	49.4	16.6	23.7
+ MotionRNN	**0.881**	44.5	15.8	21.7

Based on PredRNN，on Human3.6M

MSE promotion: **48.4→34.2**

Achieves **sate-of-the-art** performance



Parameter and Computation Efficiency

Method	Params(MB)	FLOPs(G)	MSEΔ
ConvLSTM	4.41	31.6	-
+ **MotionRNN**	5.21(↑ 18%)	36.6(↑ 16%)	12%
PredRNN	6.41	46.0	-
+ **MotionRNN**	7.01(↑ 9.3%)	49.5(↑ 7.6%)	29%
MIM	9.79	70.2	-
+ **MotionRNN**	10.4(↑ 6.2%)	73.7(↑ 5.0%)	18%
E3D-LSTM	20.4	292	-
+ **MotionRNN**	21.3(↑ 4.4%)	303(↑ 3.8%)	10%

Based on PredRNN , MotionRNN achieves **29%** improvement with little extra Params and FLOPS.

Ablation Study

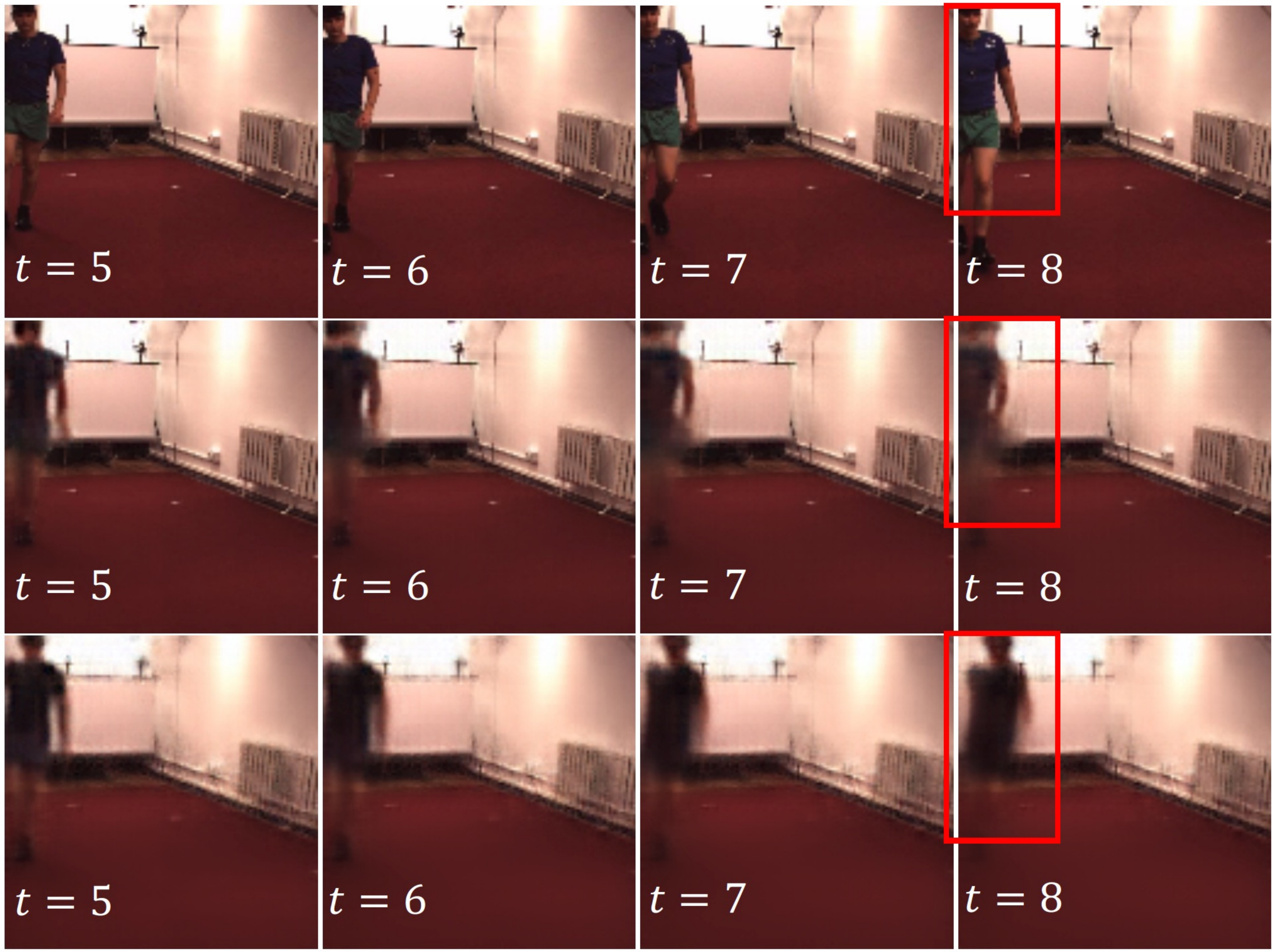
Transient Variation
13% ↑

Motion Highway
12% ↑

Motion Trend
9% ↑

Method	MH	TV	TM	$\frac{\text{MSE}}{10}$	Δ
PredRNN				48.4	-
+ Motion Highway	√			42.5	12%
+ MotionGRU w/o Momentum		√		41.5	14%
+ MotionGRU w/o Transient			√	43.5	10%
+ MotionGRU		√	√	40.3	17%
+ MotionRNN w/o Momentum	√	√		38.9	20%
+ MotionRNN w/o Transient	√		√	40.6	16%
+ MotionRNN	√	√	√	**34.2**	**29%**

Ablation of Motion Highway

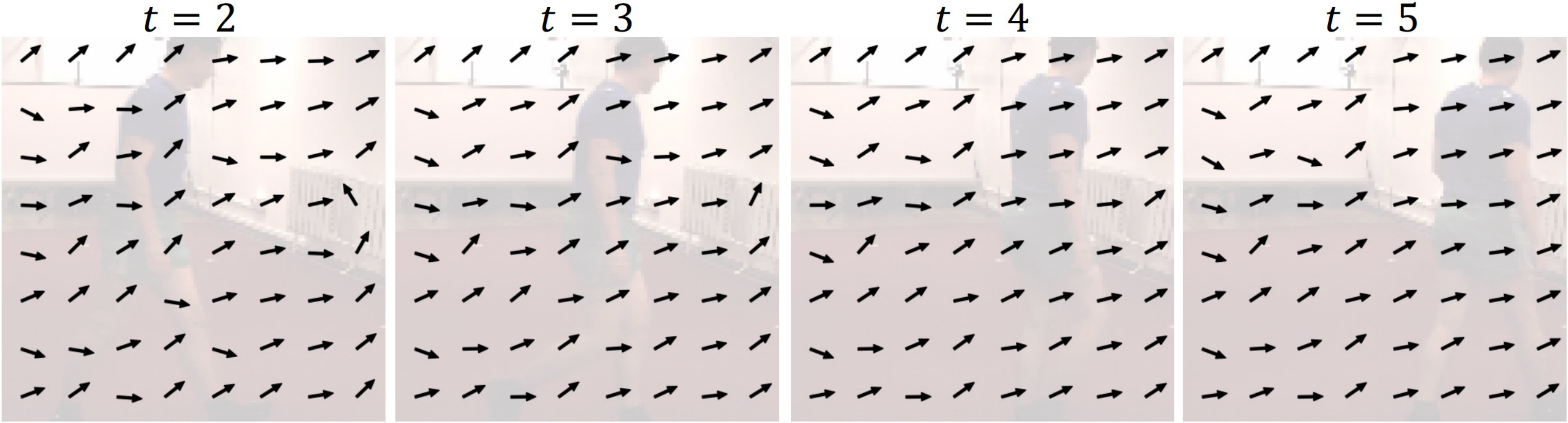


Ground Truth

PredRNN +MotionRNN

PredRNN +MotionRNN **-Motion Highway**

Motion Trend Visualization



Visualization of learned motion trend $\mathcal{D}_t^1$

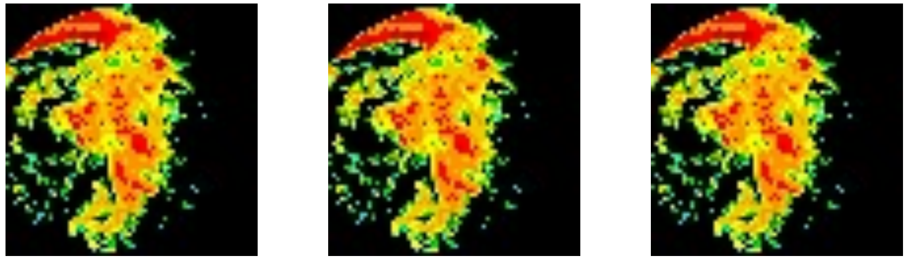
Precipitation Nowcasting (Radar Shanghai)

Method	SSIM	GDL	CSI30	CSI40	CSI50
TrajGRU	0.815	13.9	0.576	0.545	0.484
Conv-TT-LSTM	0.820	13.6	0.571	0.530	0.469
ConvLSTM	0.837	12.3	0.624	0.605	0.560
+ MotionRNN	0.850	11.9	0.646	0.629	0.586
MIM	0.849	11.3	0.654	0.646	0.609
+ MotionRNN	0.863	11.1	0.668	0.654	0.614
PredRNN	0.841	11.9	0.633	0.622	0.581
+ MotionRNN	0.865	10.9	**0.678**	**0.664**	**0.623**
E3D-LSTM	0.842	12.7	0.615	0.615	0.590
+ MotionRNN	**0.880**	**9.67**	0.671	0.659	0.621

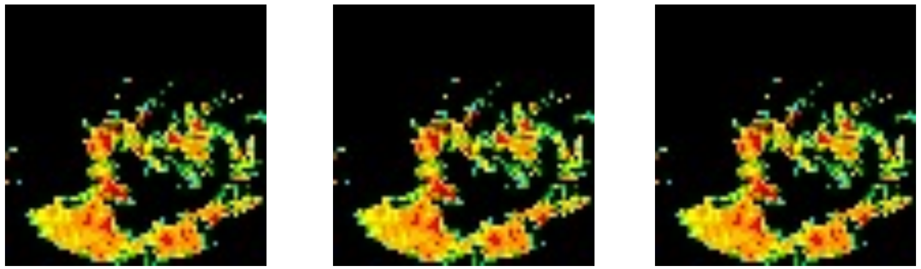
$$\mathrm{CSI} = \frac{\mathrm{Hits}}{\mathrm{Hits}+\mathrm{Misses}+\mathrm{FalseAlarms}}$$

MotionRNN can significantly improve the prediction of cloud with **high density**.

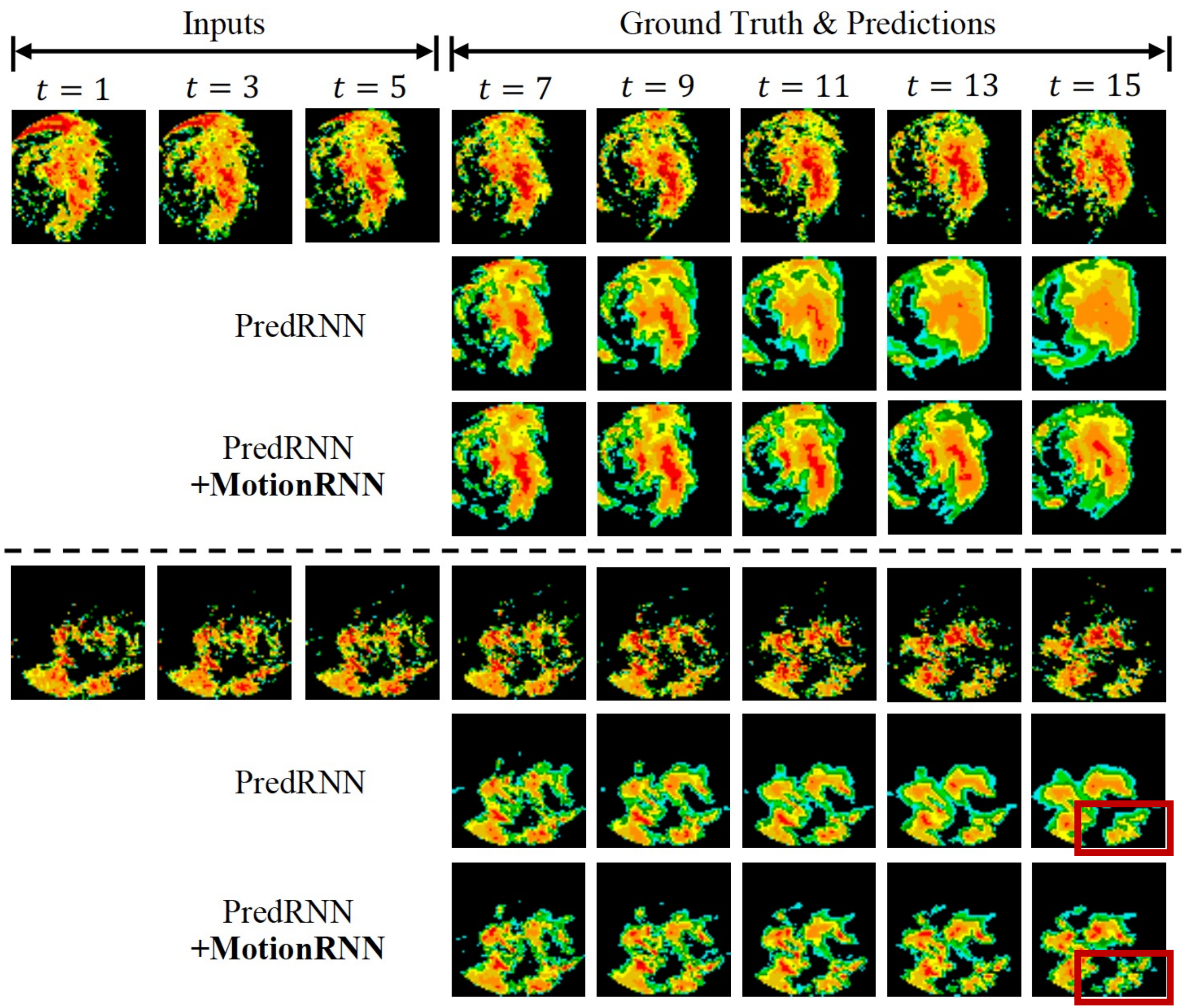
Precipitation Nowcasting（Radar Shanghai）



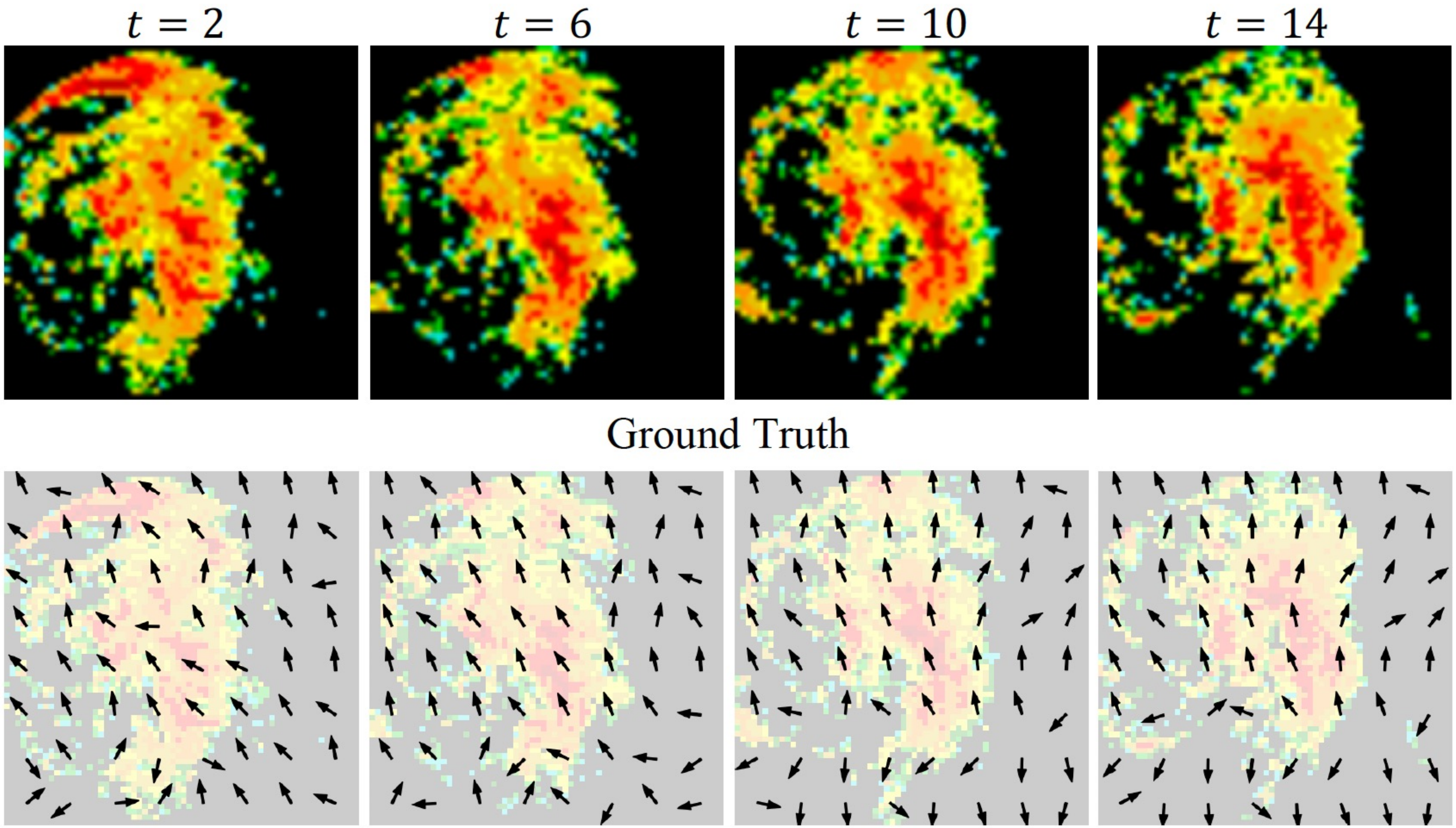
Ground Truth PredRNN MotionRNN



Ground Truth PredRNN MotionRNN



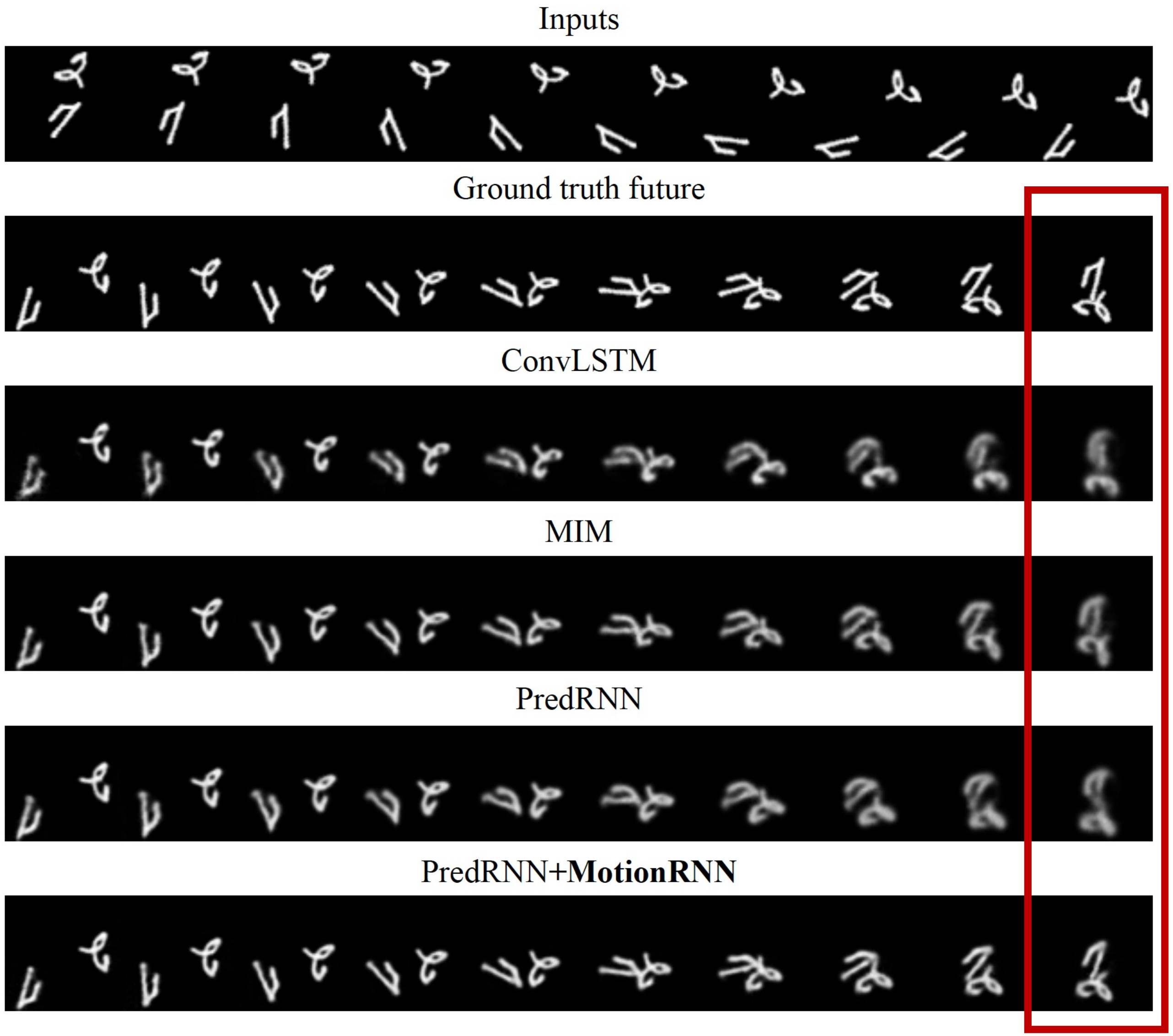
Motion Trend Visualization



MotionRNN Learned Trending Momentum D_t^1

Varied Moving Digits

Method	MSE	SSIM	PSNR	GDL
TrajGRU	109	0.515	15.9	69.3
Conv-TT-LSTM	71.1	0.744	18.4	53.6
E3D-LSTM	57.6	0.852	19.7	44.6
+ MotionRNN	52.8	0.867	20.3	42.4
ConvLSTM	47.0	0.845	20.6	41.8
+ MotionRNN	44.4	0.861	20.9	40.3
MIM	34.6	0.888	22.3	34.6
+ MotionRNN	28.9	0.906	23.1	30.9
PredRNN	35.6	0.891	22.1	34.7
+ MotionRNN	**25.1**	**0.920**	**24.0**	**27.7**



Add **rotation and scaling** to digits.

MotionRNN can greatly improve the **sharpness(GDL)** of prediction results.

Summary

- Based on motion decomposition, we design a new **MotionGRU** unit to obtain the **motion trend** and **transient variation** in a unified way.
- We propose the MotionRNN framework, which unifies the MotionGRU and a new **Motion Highway** structure to mitigate motion vanishing.
- Our MotionRNN is **flexible** to be applied together with a rich family of predictive models to yield consistent improvements and **SOTA** results.

Thank You!

whx20@mails.tsinghua.edu.cn