University of Chinese Academy of Sciences

INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

NEURAL INFORMATION PROCESSING SYSTEMS

MAU: A Motion-Aware Unit for Video Prediction and Beyond

Speaker：Zheng Chang

CONTENTS

CONTENTS

University of Chinese Academy of Sciences 中国科学院大学

INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES 中科院计算所

MOTIVATIONS

- Accurately predicting inter-frame motion information is important for video prediction.
- The temporal receptive in current predictive methods are usually narrow.
- LSTM-based methods may be not efficient for video prediction.

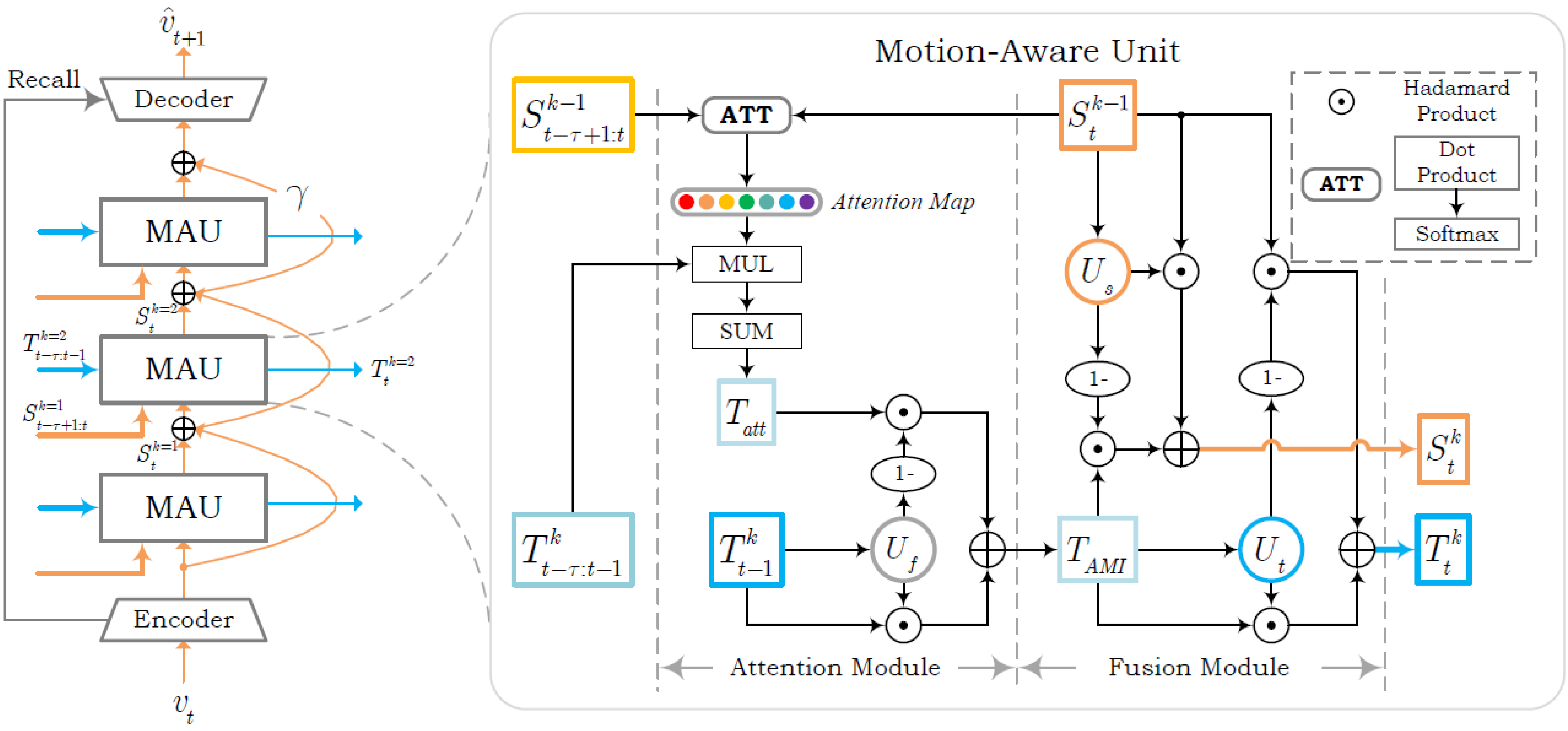
CONTENTS

CONTRIBUTIONS

- The Motion-Aware Unit (MAU) is proposed to improve the model expressivity in capturing motion information.
- For each MAU, the attention module is designed for efficient attention and the fusion module is designed for efficient fusion.
- An information recalling scheme is applied to further preserve the visual details.
- Best performance in video prediction and early action recognition tasks.

CONTENTS

METHODS



$\hat{v}_{t+1}$
Recall
Decoder
γ
MAU
$S_t^{k=2}$
$T_{t-\tau:t-1}^{k=2}$
MAU
$T_t^{k=2}$
$S_{t-\tau+1:t}^{k=1}$
$S_t^{k=1}$
MAU
Encoder
v_t
Motion-Aware Unit
$S_{t-\tau+1:t}^{k-1}$
ATT
S_t^{k-1}
Hadamard Product
Dot Product
ATT
Softmax
Attention Map
MUL
SUM
T_{att}
U_s
1-
1-
1-
S_t^k
$T_{t-\tau:t-1}^k$
T_{t-1}^k
U_f
T_{AMI}
U_t
T_t^k
Attention Module
Fusion Module

CONTENTS

University of Chinese Academy of Sciences 中国科学院大学

INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES 中科院计算所

EXPERIMENTS: Moving MNIST (64×64)

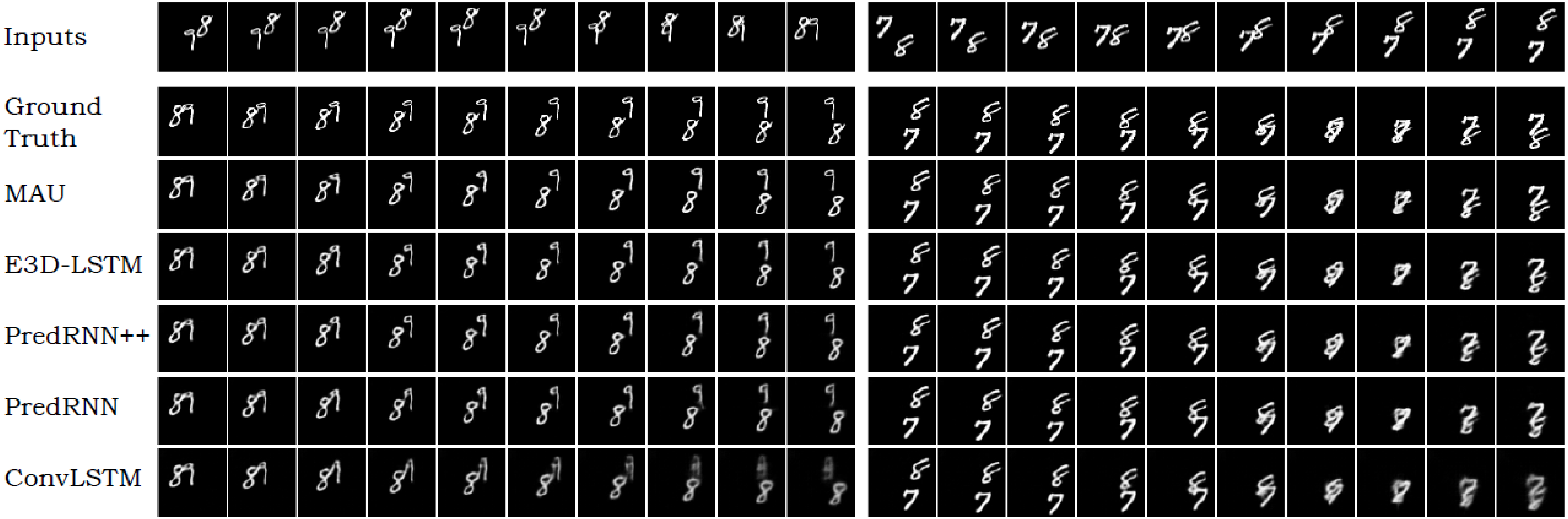


Table 4: Model performance of MAU with different temporal receptive field τ. In particular, $\gamma = 0$, $\lambda = 0$. The percentage values are calculated based the MAU with $\tau = 1$.

	$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 10$
MSE/frame	10.5	10.2 (↓2.9%)	9.7 (↓7.6%)	9.6 (↓8.6%)
Inference time	14.90s	15.85s (↑6.4%)	17.36s (↑16.5%)	20.23s (↑35.8%)

EXPERIMENTS: Moving MNIST

Table 2: Quantitative results of different methods on the Moving MNIST dataset (10 frames → 10 frames). Lower MSE, FVD scores and higher SSIM score indicate better visual quality. The results of the compared methods are reported in [36].

Method	Moving MNIST SSIM/frame↑	MSE/frame↓	FVD/10 frames↓
ConvLSTM (NeurIPS2015) [9]	0.707	103.3	153.1
FRNN (ECCV2018) [12]	0.819	68.4	-
VPN (ICML2017) [29]	0.870	70.0	-
PredRNN (NeurIPS2017) [13]	0.869	56.8	77.0
PredRNN++ (ICML2018) [14]	0.898	46.5	91.5
MIM (CVPR2019) [15]	0.910	44.2	-
E3D-LSTM (ICLR2019) [16]	0.910	41.3	88.7
CrevNet (ICLR2020) [17]	0.949	22.3	63.6
MAU (w/o recalling)	0.977	9.7	39.8
MAU	**0.978**	**8.9**	**37.0**

EXPERIMENTS: Moving MNIST

Table 3: Ablation study on the Moving MNIST dataset. For fair comparison, the encoders and decoders are with the same structure for all models and All models are trained using Adam optimizer based on the MSE loss.

Method	Backbone	MSE↓	SSIM↑	Parameters	Inference time
ConvLSTM (NeurIPS2015) [9]	4×ConvLSTMs	102.1	0.747	0.98M	**16.47s**
ST-LSTM (NeurIPS2017) [13]	4×ST-LSTMs	54.5	0.839	1.57M	17.74s
Casual-LSTM (ICML2018) [14]	4×Casual-LSTMs	46.3	0.899	1.80M	21.25s
MIM (CVPR2019) [15]	4×MIMs	44.1	0.910	3.03M	45.13s
E3D-LSTM (ICLR2019) [16]	4×E3D-LSTMs	40.1	0.912	4.70M	57.21s
RPM (ICLR2020) [17]	4×RPMs	23.7	0.934	1.77M	18.01s
MotionGRU (CVPR2021) [28]	4×MotionGRUs	25.3	0.919	1.16M	17.58s
MAU	4×MAUs	**9.7**	**0.977**	**0.78M**	**17.34s**

EXPERIMENTS: Caltech Pedestrian (128×160)

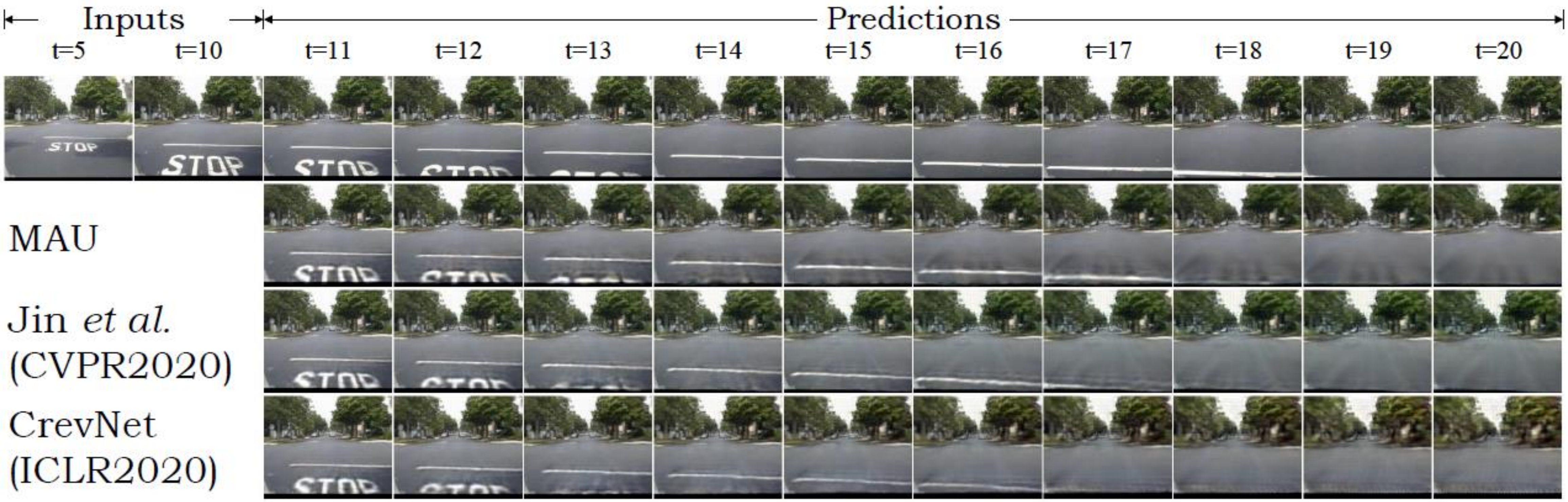


Figure 3: The qualitative results from different methods on the Caltech Pedestrian dataset.

University of Chinese Academy of Sciences 中国科学院大学

INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES 中科院计算所

EXPERIMENTS: Caltech Pedestrain

Table 5: Quantitative results of different methods on the Caltech Pedestrian dataset (10 frames → 1 frame). Lower MSE, LPIPS, FVD scores and higher SSIM, PSNR scores indicate better visual quality. The results of the compared methods are reported in [36].

	Caltech Pedestrian				
Method	MSE(10^{-3})↓	SSIM↑	PSNR↑	LPIPS(10^{-2})↓	FVD/10 frames↓
BeyondMSE (ICLR2016) [26]	3.42	0.847	-	-	-
MCnet (ICLR2017) [38]	2.50	0.879	-	-	-
CtrlGen (CVPR2018) [39]	-	0.900	26.5	-	-
PredNet (ICLR2017) [37]	2.42	0.905	27.6	7.47	2860.8
ContextVP (ECCV2018) [40]	1.94	0.921	28.7	6.03	2451.6
E3D-LSTM (ICLR2019) [16]	2.12	0.914	28.1	6.31	2311.2
Kwon *et al.* (CVPR2019) [24]	1.61	0.919	29.2	4.91	1663.2
CrevNet (ICLR2020) [17]	1.55	0.925	29.3	5.94	1709.6
Jin *et al.* (CVPR2020)[27]	1.59	0.927	29.1	5.89	1441.1
MAU (w/o recalling)	1.34	0.939	29.4	4.90	1269.9
MAU	**1.24**	**0.943**	**30.1**	**4.85**	**1204.0**

EXPERIMENTS: TownCentreXVID(1088×1920)

Figure 4: Qualitative results from different methods on the TownCentreXVID dataset (4 frames → 1 frame).

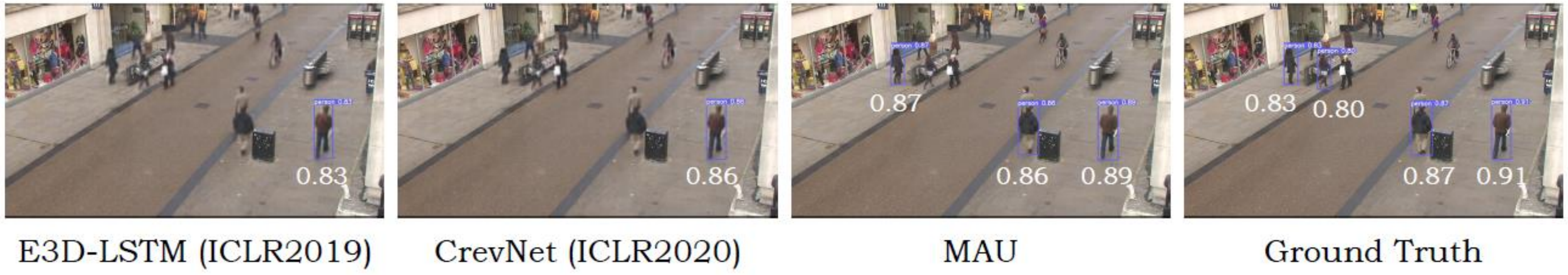


Figure 5: Object detection experiments on the predictions (4 frames → 1 frame) from different methods using Yolov5s pre-trained model [41]. Confidence threshold is set to 0.8.

University of Chinese Academy of Sciences
中国科学院大学
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
中科院计算所

EXPERIMENTS: TownCentreXVID

University of Chinese Academy of Sciences 中国科学院大学

INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES 中科院计算所

Table 6: Quantitative results of different methods on the TownCentreXVID dataset (4 frames → 4 frame). Higher SSIM and PSNR scores indicate better objective quality. Lower LPIPS score indicates better perceptual quality.

Method	TownCentreXVID $t=5$			$t=8$		
	PSNR↑	SSIM↑	LPIPS(10^{-2})↓	PSNR↑	SSIM↑	LPIPS(10^{-2})↓
ConvLSTM (NeurIPS2015) [9]	27.22	0.894	39.90	23.29	0.876	46.12
PredRNN (NeurIPS2017) [13]	28.95	0.921	32.48	23.82	0.885	37.85
PredRNN++ (ICML2018) [14]	29.50	0.926	30.59	24.37	0.894	39.54
E3D-LSTM (ICLR2019) [16]	29.70	0.929	29.47	24.34	0.901	36.82
CrevNet (ICLR2020) [17]	30.12	0.933	27.87	24.62	0.910	33.70
MAU (w/o recalling)	30.84	0.939	24.07	25.52	0.914	30.87
MAU	**31.87**	**0.969**	**8.28**	**27.14**	**0.942**	**12.89**

EXPERIMENTS: Something-Somethingv2

University of Chinese Academy of Sciences 中国科学院大学

ICT 中科院计算所 INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Table 7: The results of the early action recognition experiment of different methods on the Something-Something V2 dataset.

Method	Something-SomethingV2 Front 25%			Front 50%		
	PSNR↑	top-1↑	top-5↑	PSNR↑	top-1↑	top-5↑
ST-LSTM (NeurIPS2017) [13]	14.98	5.77	13.98	15.77	9.23	19.33
Casual-LSTM (ICML2018) [14]	15.44	6.54	17.11	16.44	10.33	22.64
E3D-LSTM (ICLR2019) [16]	16.32	6.98	18.33	17.01	10.45	24.34
RPM (ICLR2020) [17]	16.67	8.01	18.87	17.68	12.50	24.67
MotionGRU (CVPR2021) [28]	17.03	8.44	20.19	17.98	14.31	27.79
MAU	**17.57**	**8.93**	**25.60**	**18.59**	**16.07**	**30.36**

Thanks

Q&A