SSA-GAN: End-to-End Time-Lapse Generation with Spatial Self-Attention

Daichi Horita, Keiji Yanai

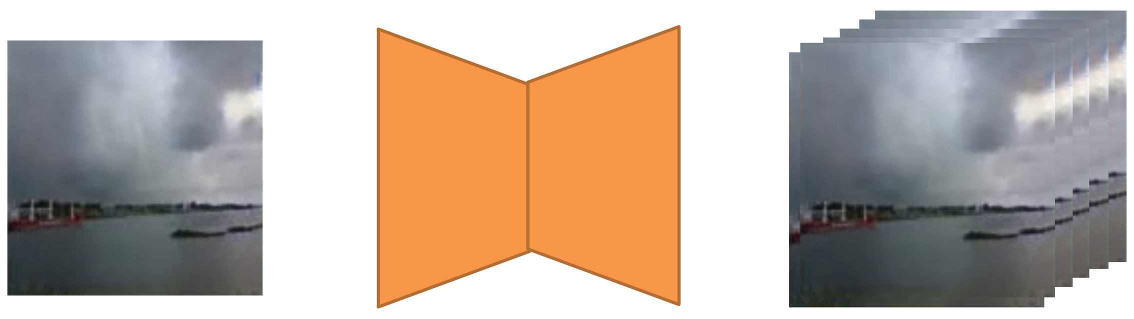
The University of Electro Communications, Tokyo, Japan

ACPR2019

Demo page!

1. Introduction

- We usually predict how objects will move in the near future in our daily lives. *However, how do we predict?* To address this problem, we propose a GAN-based network to predict the near future for fluid object domains.
- Our model takes one frame and is able to predict future frames.
- We propose introducing the spatial self-attention mechanism into the model.



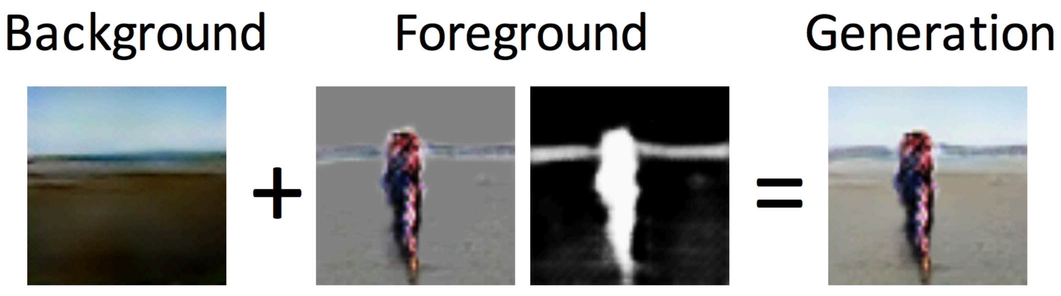
Demo iPad

ground truth | our model | [2] first | [2] second stage

2. Related Works

Video prediction

- Since VGAN[1] generates the background and foreground of the image separately, the background is fixed.



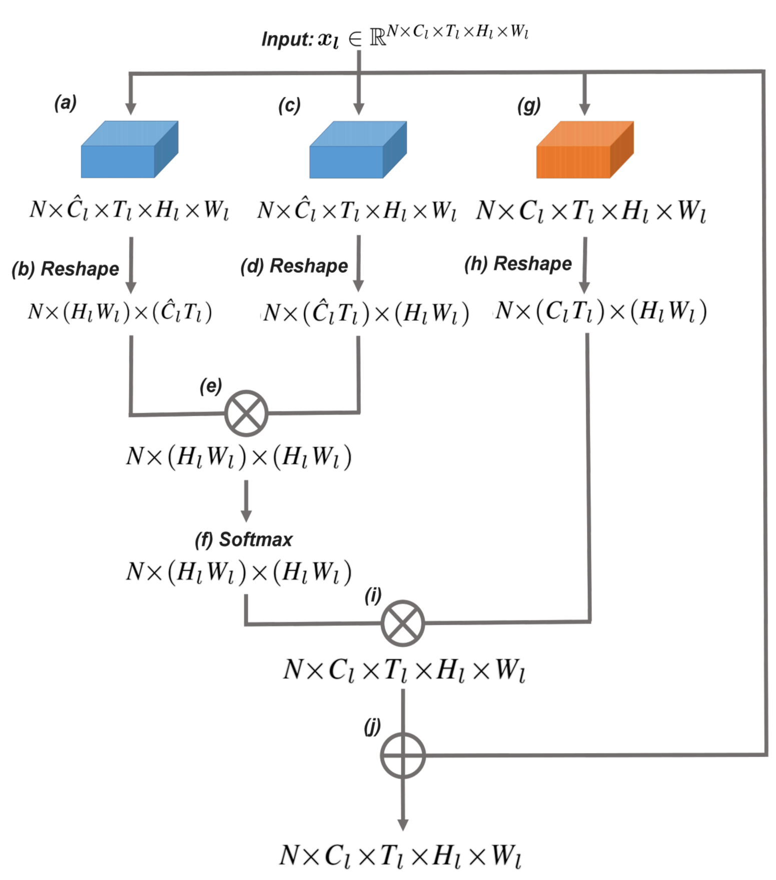
➡ <u>Propose to generate the both content at the same time.</u>

- MDGAN[2] generate rough movements in the first stage and add detailed appearances and motions in the second stage.

➡ <u>Propose the model with only one-stage learning.</u>

3. Method - Spatial Self Attention Module

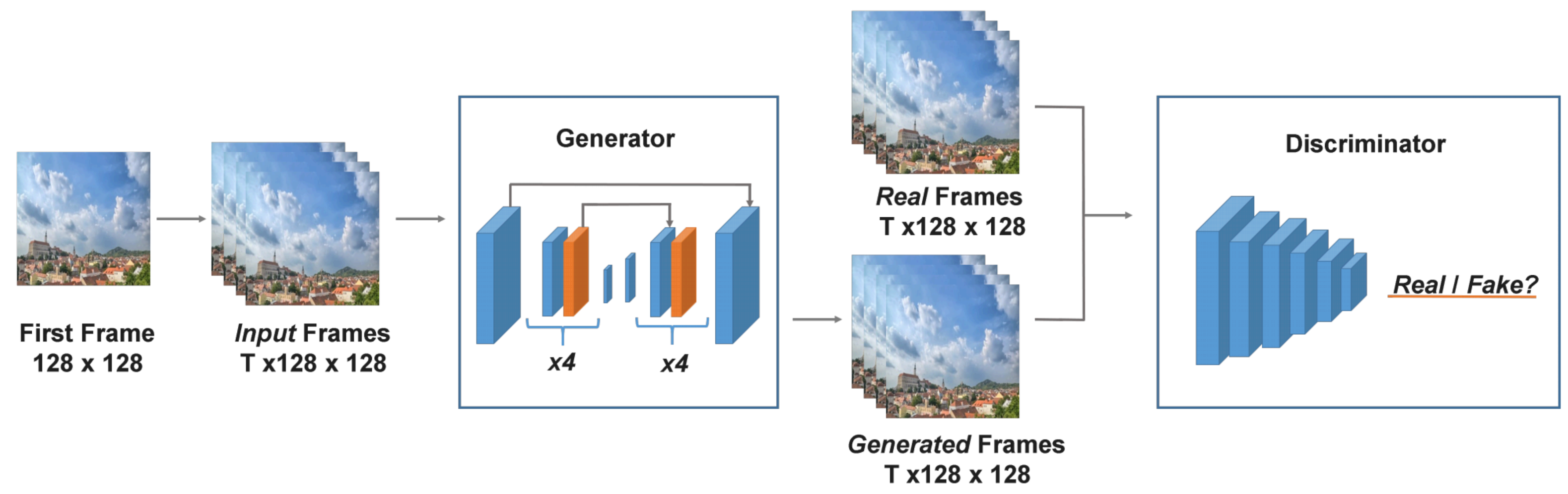
- Propose a spatial self-attention module <u>to learn the long-range dependence within a frame.</u>
- The network assigns more weight to areas outside the neighborhood.
- γ play the important role to avoid the over-weighting(See 5.2).



(i) Output

$$y_l = \gamma o_l + x_l,$$

4. Method - Spatial Self-Attention GAN



- (Blue) 3D convolutional/transposed convolutional layers.
- (Orange) Spatial Self Attention Modules.
- **Objectives**

<u>Adversarial Loss</u>

$$\mathcal{L}_{adv} = \min_{G} \max_{D} \mathop{E}_{Y \sim P_r} [\log D(Y)] + \mathop{E}_{\bar{X} \sim P_g} [\log (1 - D(\bar{X}))],$$

<u>Content Loss</u>

$$\mathcal{L}_{con} = \mathop{E}_{Y \sim P_r, \bar{X} \sim P_g} [\|Y - \bar{X}\|],$$

<u>Full Loss</u>

$$\mathcal{L}_D = -\mathcal{L}_{adv},$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{con}\mathcal{L}_{con},$$

5. Experiments

1. Quantitative evaluation

- Cloud dataset[3]

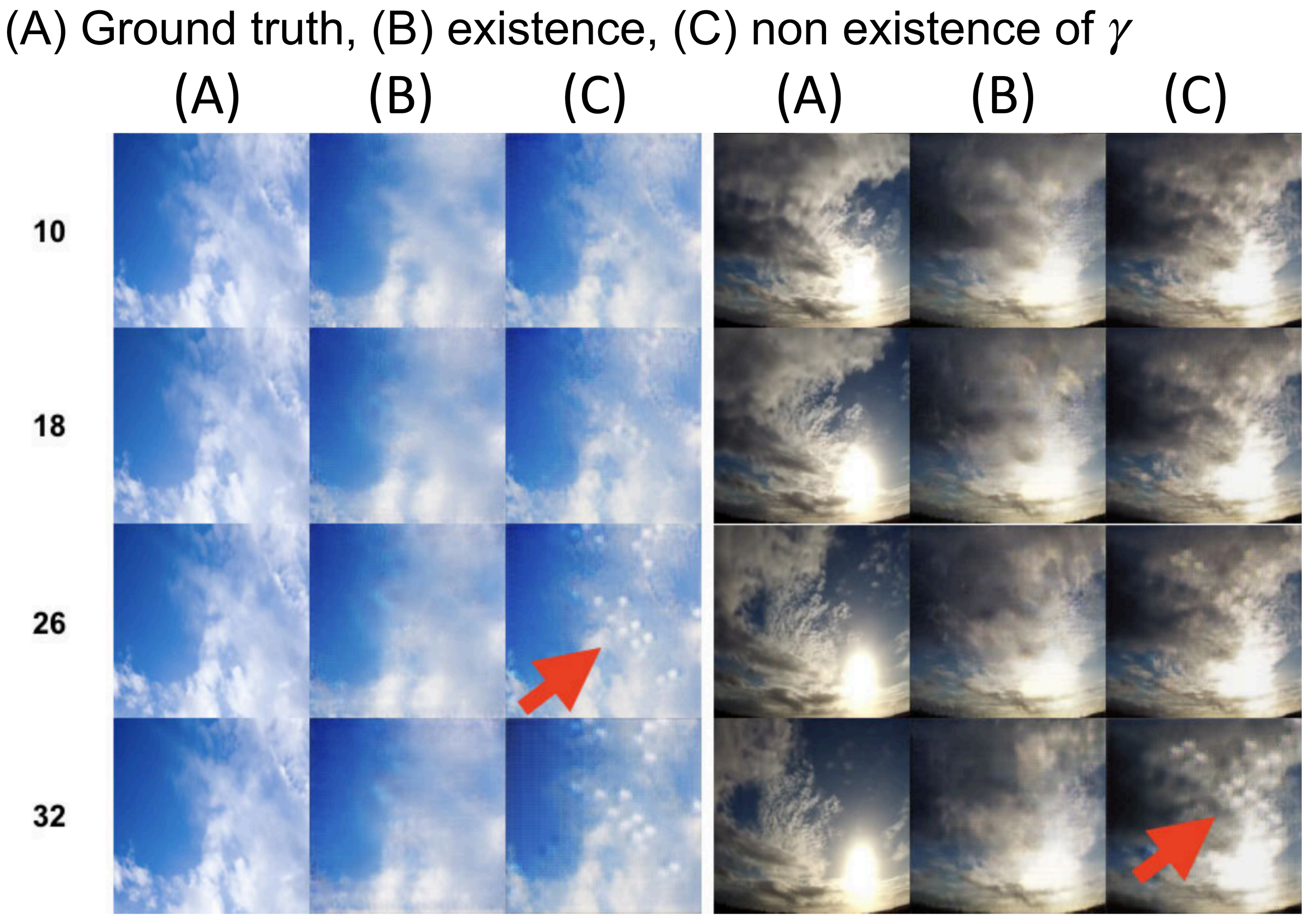
Method	MSE↓	PSNR↑	SSIM↑
MD-GAN Stage *I*	0.0970	16.9019	0.3583
MD-GAN Stage *II*	0.0307	22.7372	0.5920
SSA-GAN (Ours)	**0.0232**	**24.9100**	**0.6805**

Beach dataset[1]

Method	MSE↓	PSNR↑	SSIM↑
RNN-GAN	0.1849	7.7988	0.5143
VGAN	0.0958	11.5586	0.6035
MD-GAN Stage *II*	0.0422	16.1951	**0.8019**
Ours (a)	0.0379	23.6601	0.7320
Ours (b)	**0.0374**	**25.6432**	0.7346

2. Ablation study

- (A) Ground truth, (B) existence, (C) non existence of γ



6. Future Works

- Use more temporal features to generate more realistic frames.
- Experiment with mode frames.

- References

[1] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In Proc.of Neural Information Processing Systems(NIPS), 2016.

[2] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu,and Jiebo Luo. Learning to generate time-lapse videos usingmulti-stage dynamic generative adversarial networks. In Proc. of CVPR, 2018.