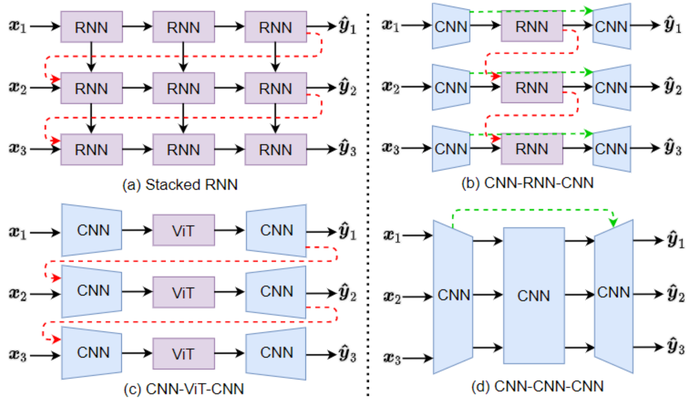
# SimVP: Simpler yet Better Video Prediction

Zhangyang Gao, Cheng Tan, Lirong Wu, Siyuan Li, Stan Z. Li
AI Lab, School of Engineering, Westlake University.
Institute of Advanced Technology, Westlake Institute for Advanced Study

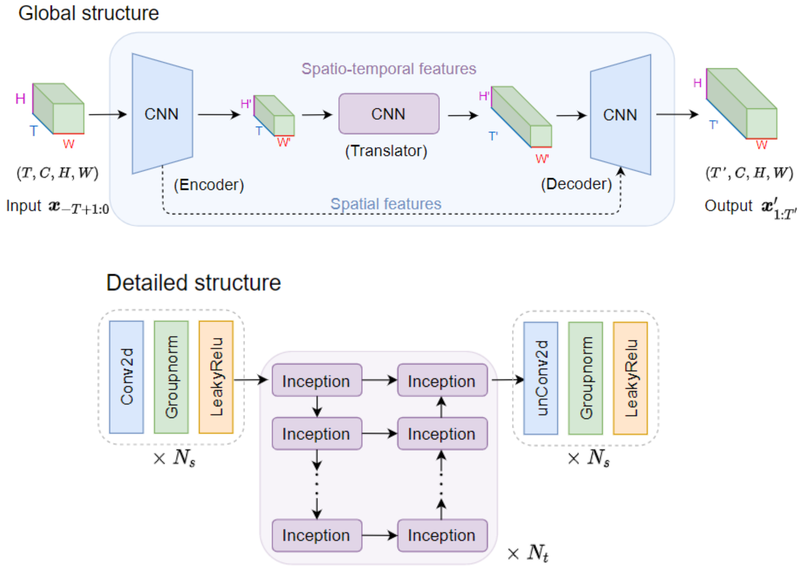
## Section 1 Previous methods




- We have witnessed many terrific methods that have achieved outstanding performance.
- However, as the models become more complex, understanding their performance gain is an inevitable challenge, and scaling them into large datasets is intractable.
- Can we develop a simpler model to provide better understanding and performance?

## Section 2 SimVP

- SimVP consists of an encoder, a translator and a decoder built on CNN.
- SimVP can achieve state-of-the-art results without introducing any complex modules, strategies and tricks.
- SimVP has better computational time efficiency than baselines.




The **encoder** and **decoder** stack $N_s$ ConvNormReLU blocks to extract spatial features, i.e., convoluting $C$ channels on $(H, W)$. The hidden feature is:

$$z_i = \sigma\left(\mathrm{LayerNorm}\left(\mathrm{Conv2d}(z_{i-1})\right)\right), 1 \leq i \leq N_s$$

The **translator** employs $N_t$ Inception modules to learn temporal evolution, i.e., convoluting $T{\times}C$ channels on $(H, W)$. The Inception module consists of a bottleneck Conv2d with $1{\times}1$ kernel followed by parallel GroupConv2d operators. The hidden feature is:

$$z_j = \mathrm{Inception}\left(z_{j-1}\right), N_s \leq j \leq N_s + N_t$$

## Section 3 Experimental results

- SimVP achieves state-of-the-art MSE and SSIM on Moving MNIST, TrafficBJ, Human3.6, Caltech Pedestrian, and KTH.
- The simplicity leads to good computational efficiency..
- SimVP extends well to the case of flexible predictive length.

| | | Moving MNIST | | | TrafficBJ | | | Human3.6 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE↓ | MAE↓ | SSIM↑ | MSE × 100 ↓ | MAE↓ | SSIM↑ | MSE / 10 ↓ | MAE / 100 ↓ | SSIM↑ |
| | ConvLSTM | 103.3 | 182.9 | 0.707 | 48.5 | 17.7 | 0.978 | 50.4 | 18.9 | 0.776 |
| | PredRNN | 56.8 | 126.1 | 0.867 | 46.4 | 17.1 | 0.971 | 48.4 | 18.9 | 0.781 |
| | CausalLSTM | 46.5 | 106.8 | 0.898 | 44.8 | 16.9 | 0.977 | 45.8 | 17.2 | 0.851 |
| | MIM | 44.2 | 101.1 | 0.910 | 42.9 | 16.6 | 0.971 | 42.9 | 17.8 | 0.790 |
| | E3D-LSTM | 41.3 | 86.4 | 0.910 | 43.2 | 16.9 | 0.979 | 46.4 | 16.6 | 0.869 |
| Train for 10 day | PhyDNet | 24.4 | 70.3 | 0.947 | 41.9 | **16.2** | **0.982** | 36.9 | 16.2 | 0.901 |
| Train for 2 day | SimVP | **23.8** | **68.9** | **0.948** | **41.4** | **16.2** | **0.982** | **31.6** | **15.1** | **0.904** |

| Method | Caltech Pedestrian (10 → 1) | | |
|---|---|---|---|
| | MSE↓ | SSIM↑ | PSNR↑ |
| BeyondMSE [40] | 3.42 | 0.847 | - |
| MCnet [62] | 2.50 | 0.879 | - |
| DVF [35] | - | 0.897 | 26.2 |
| Dual-GAN [32] | 2.41 | 0.899 | - |
| CtrlGen [19] | - | 0.900 | 26.5 |
| PredNet [37] | 2.42 | 0.905 | 27.6 |
| ContextVP [5] | 1.94 | 0.921 | 28.7 |
| GAN-VGG [54] | - | 0.916 | - |
| G-VGG [54] | - | 0.917 | - |
| SDC-Net [50] | 1.62 | 0.918 | - |
| rCycleGan [29] | 1.61 | 0.919 | 29.2 |
| DPG [16] | - | 0.923 | 28.2 |
| G-MAE [54] | - | 0.923 | - |
| GAN-MAE [54] | - | 0.923 | - |
| CrevNet [73] | - | 0.925 | 29.3 |
| STMFANet [26] | - | 0.927 | 29.1 |
| SimVP (ours) | **1.56** | **0.940** | **33.1** |

| Method | KTH (10 → 20) | | KTH (10 → 40) | |
|---|---|---|---|---|
| | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ |
| MCnet [62] | 0.804 | 25.95 | 0.73 | 23.89 |
| ConvLSTM [71] | 0.712 | 23.58 | 0.639 | 22.85 |
| SAVP [30] | 0.746 | 25.38 | 0.701 | 23.97 |
| VPN [28] | 0.746 | 23.76 | – | – |
| DFN [25] | 0.794 | 27.26 | 0.652 | 23.01 |
| fRNN [43] | 0.771 | 26.12 | 0.678 | 23.77 |
| Znet [74] | 0.817 | 27.58 | – | – |
| SV2Pi [3] | 0.826 | 27.56 | 0.778 | 25.92 |
| SV2Pv [3] | 0.838 | 27.79 | 0.789 | 26.12 |
| PredRNN [67] | 0.839 | 27.55 | 0.703 | 24.16 |
| VarNet [27] | 0.843 | 28.48 | 0.739 | 25.37 |
| SVAP-VAE [30] | 0.852 | 27.77 | 0.811 | 26.18 |
| PredRNN++ [65] | 0.865 | 28.47 | 0.741 | 25.21 |
| MSNET [31] | 0.876 | 27.08 | – | – |
| E3d-LSTM [66] | 0.879 | 29.31 | 0.810 | 27.24 |
| STMFANet [26] | 0.893 | 29.85 | 0.851 | 27.56 |
| SimVP (ours) | **0.905** | **33.72** | **0.886** | **32.93** |

- More experiments can be found in our paper.

## Summary

- We propose SimVP, a simpler yet effective CNN model for video prediction.
- SimVP can achieve state-of-the-art results with better computational time efficiency.
- We believe simpler is better, and SimVP may serve as a strong baseline and provide inspiration for future researches.