JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

A Dynamic Multi-Scale Voxel Flow Network for Video Prediction

Xiaotao Hu Zhewei Huang Ailin Huang Jun Xu Shuchang Zhou

Poster: TUE-PM-191

Project page: https://huxiaotaostasy.github.io/DMVFN/

Paper: https://arxiv.org/abs/2303.09875

VCIP, CS, Nankai University， http://mmcheng.net/

Overview

Key challenge:

- The motions of different objects between two adjacent frames are usually of different scales.

Large scale

Small scale

Contributions:

- We design a light-weight **D**ynamic **M**ulti-scale **V**oxel **F**low **N**etwork (DMVFN) with only RGB frames as inputs.
- We propose an effective **Routing Module** to dynamically select a suitable sub-network according to the input.
- DMVFN achieves SoTA results while being an order of magnitude faster than previous methods.

More Details About DMVFN

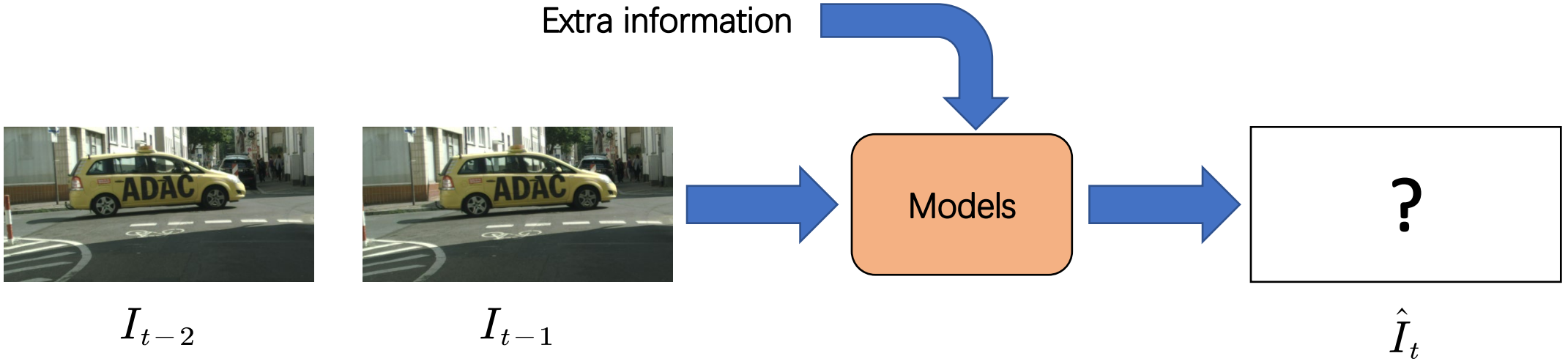
Video Prediction

Task:

Video prediction aims to predict future video frames from the current ones.

Prior approaches:

Extra information is exploited by video prediction methods in pursuit of better performance (e.g. FVS). However, the extra information may not always be available in practical scenarios, which limits the application scope of these methods.
Most of prior video prediction methods are computationally expensive.

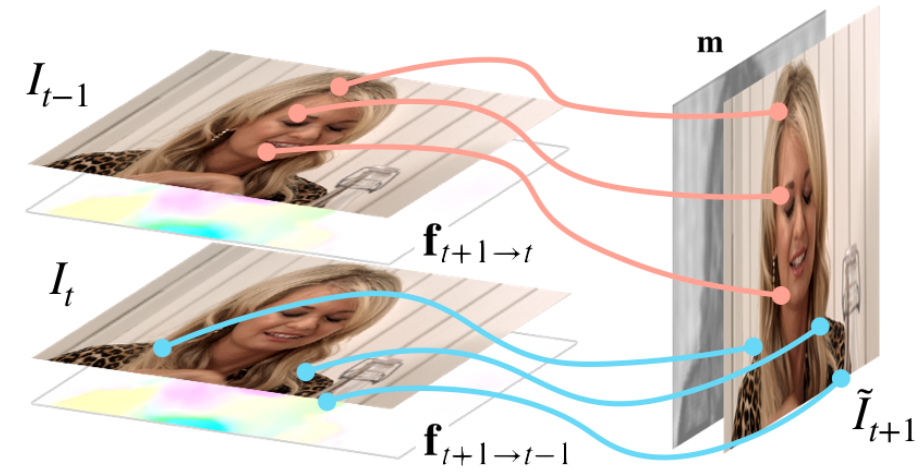


[FVS] Future Video Synthesis With Object Motion Prediction, CVPR 2020.

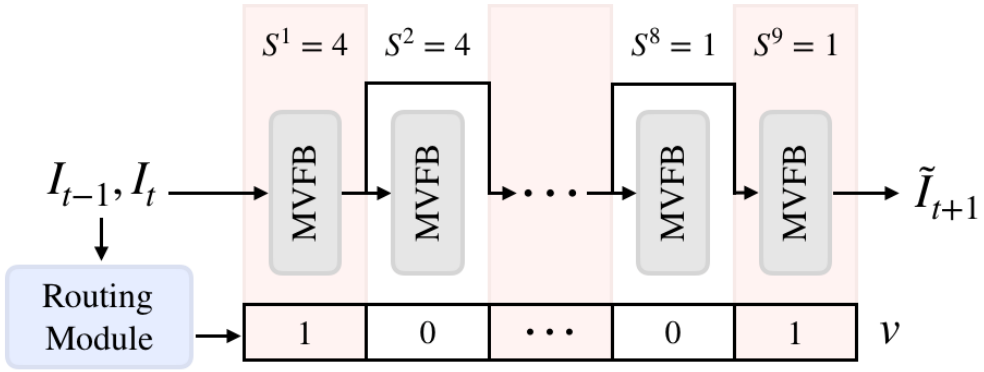
Ways to Improve

- Design a new block that can model different motion scales.
- Propose a new network that dynamically selects sub-network according to the input reducing computational costs.

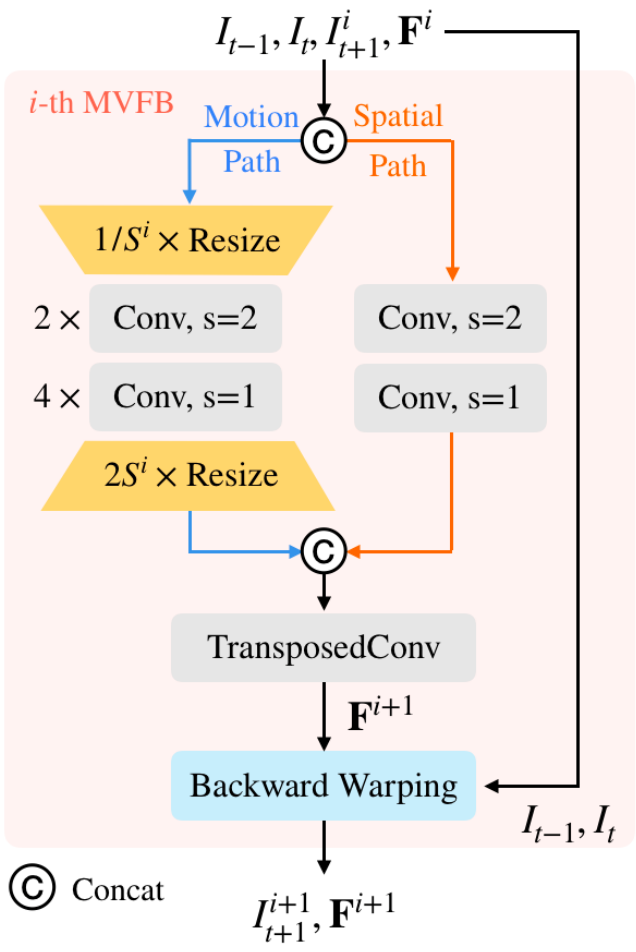
Dynamic Multi-scale Voxel Flow Network



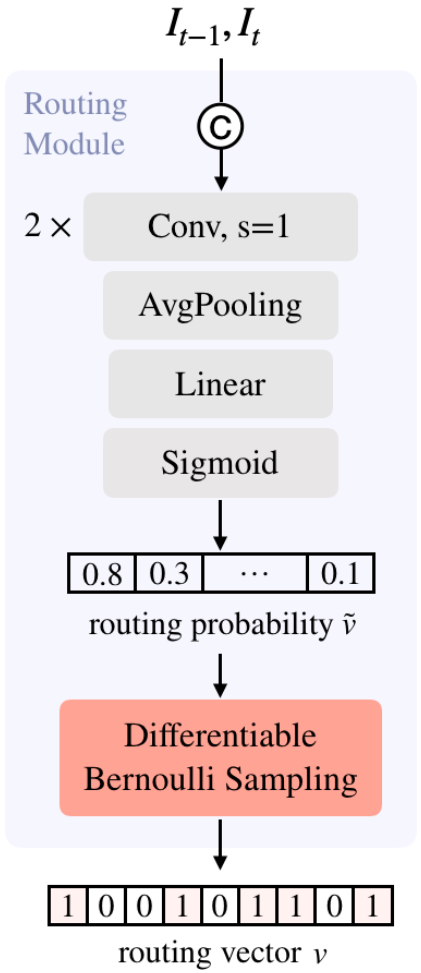
(a) Voxel Flow-based Image Fusion



(b) DMVFN



(c) MVFB



(d) Routing Module

Quantity comparison

Quantitative results of different methods on the Cityscapes datasets. “RGB”, “F”, “S” and “I” denote the video frames, optical flow, semantic map, and instance map, respectively. “N/A” means not available.

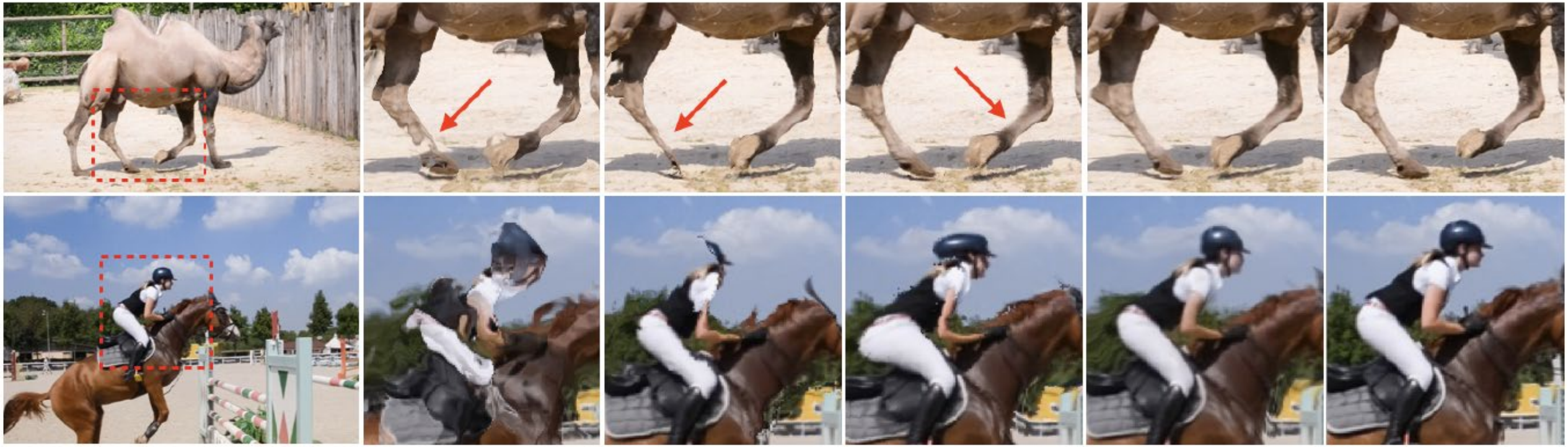
Method	Inputs	GFLOPs	MS-SSIM ($x10^{-2}$)			LPIPS ($x10^{-2}$)		
			t+1	t+3	t+5	t+1	t+3	t+5
Vid2vid	RGB+S	603.79	88.16	80.55	75.13	10.58	15.92	20.14
Seg2vid	RGB+S	455.84	88.32	N/A	61.63	9.69	N/A	25.99
FVS	RGB+S+I	1891.65	89.10	**81.13**	75.68	8.50	**12.98**	16.50
SADM	RGB+S+F	N/A	**95.99**	N/A	**83.51**	**7.67**	N/A	**14.93**
DVF	RGB	409.78	83.85	76.23	71.11	17.37	24.05	28.79
CorrWise	RGB	944.29	92.80	N/A	**83.90**	8.50	N/A	15.00
OPT	RGB	313482.15	94.54	86.89	80.40	6.46	12.50	17.83
DMVFN	RGB	12.71	**95.73**	**89.24**	83.45	**5.58**	**10.47**	**14.82**

[Vid2vid] Video-to-video synthesis, NIPS 2018.
[Seg2vid] Video generation from single semantic label map, CVPR 2019.
[FVS] Future video synthesis with object motion prediction, CVPR 2020.
[SADM] Learning semantic-aware dynamics for video prediction , CVPR 2021.
[DVF] Video frame synthesis using deep voxel flow, ICCV 2017.
[CorrWise] Comparing correspondences: Video prediction with correspondence-wise losses, CVPR 2022.
[OPT] Optimizing video prediction via video frame interpolation, CVPR 2022.

* More results can be found in the main paper.

Quality comparison

Visual results of different methods on the DAVIS-17 datasets.



[DVF] Video frame synthesis using deep voxel flow, ICCV 2017.
[DYAN] Dyan: A dynamical atoms-based network for video prediction, ECCV 2018.
[OPT] Optimizing video prediction via video frame interpolation, CVPR 2022.

Conclusion

- We developed an efficient Dynamic Multi-scale Voxel Flow Network (DMVFN) that excels previous video prediction methods on dealing with complex motions of different scales.
- With the proposed routing module, our DMVFN adaptively activates different sub-networks based on the input frames, improving the prediction performance while reducing the computation costs.
- Experiments on diverse benchmark datasets demonstrated that our DMVFN achieves state-of-the-art performance with greatly reduced computation burden.

Thank you!

Code & demo are available at: https://huxiaotaostasy.github.io/DMVFN/ **(*or scan the QRCode*)**