V2V: Scaling Event-Based Vision through Efficient Video-to-Voxel Simulation

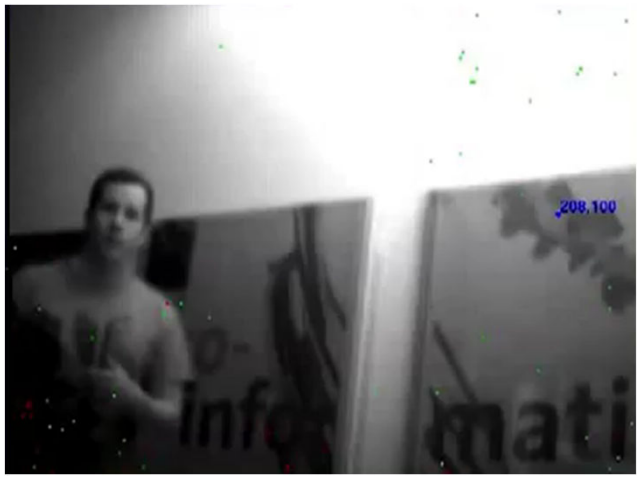
Hanyue Lou[1,2] Jinxiu Liang[3] Minggui Teng[1] Yi Wang[2,4] Boxin Shi[1]

[1] Peking University [2] Shanghai Innovation Institute

[3] National Institute of Informatics [2] Shanghai AI Laboratory

Event-based vision

- Event cameras are novel visual sensors that report intensity changes, providing high dynamic range and high temporal resolution.
- Potential applications: imaging, robotics, wildlife monitoring……



Intensity images & Event signals

*Video courtesy of Elias Mueggler

Data shortage in event-based vision

Massive image resources on the Internet
Everyone with a smartphone takes new images

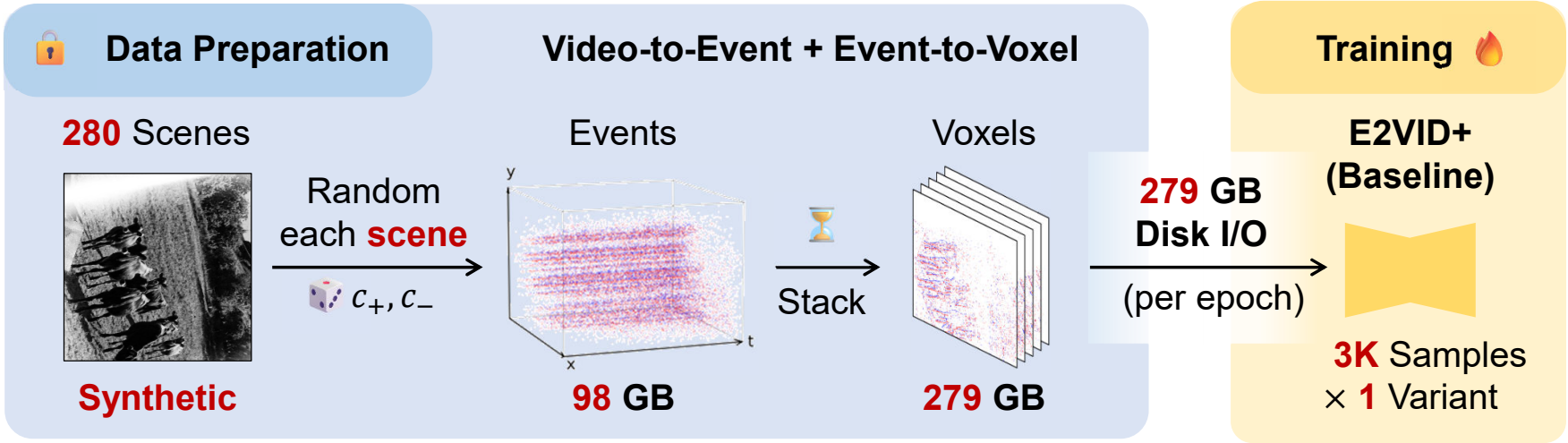
- Image-based CV: **Big data** + Big Models + GPUs → High Performance

- Event-based CV: **Big data** + Big Models + GPUs → **?**

Event cameras are expensive, only labs buy them
Data amount is small, growth is slow

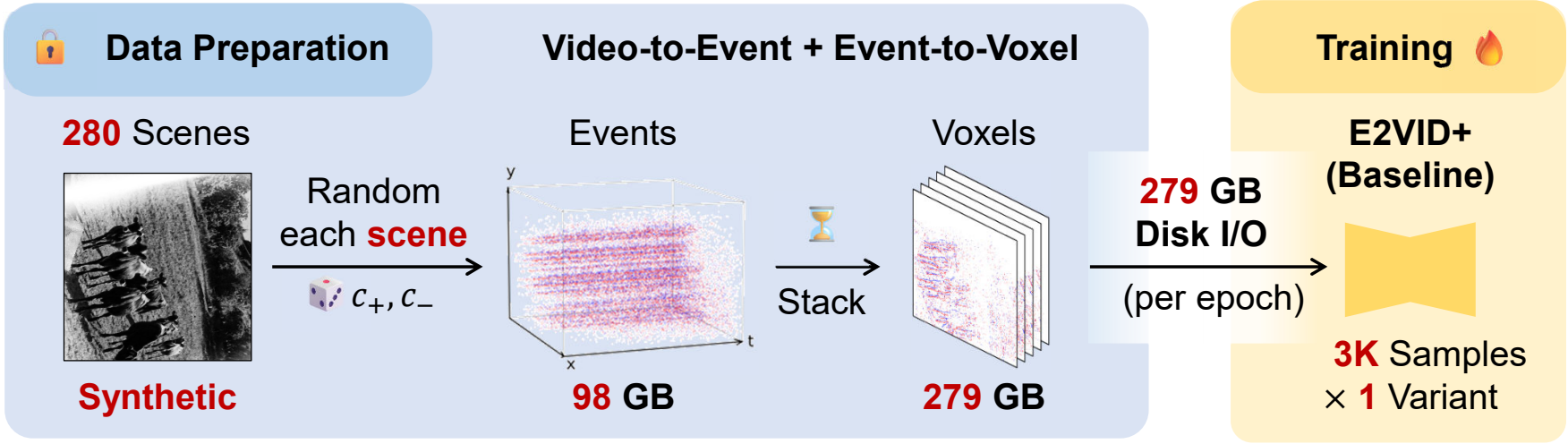
Current synthetic data pipeline

- Assign event camera parameters (e.g. thresholds) to scenes
- Simulate (N, 4) event streams, save to disk
- Pre-compute voxel representations, save to disk
- Read all voxels out for training



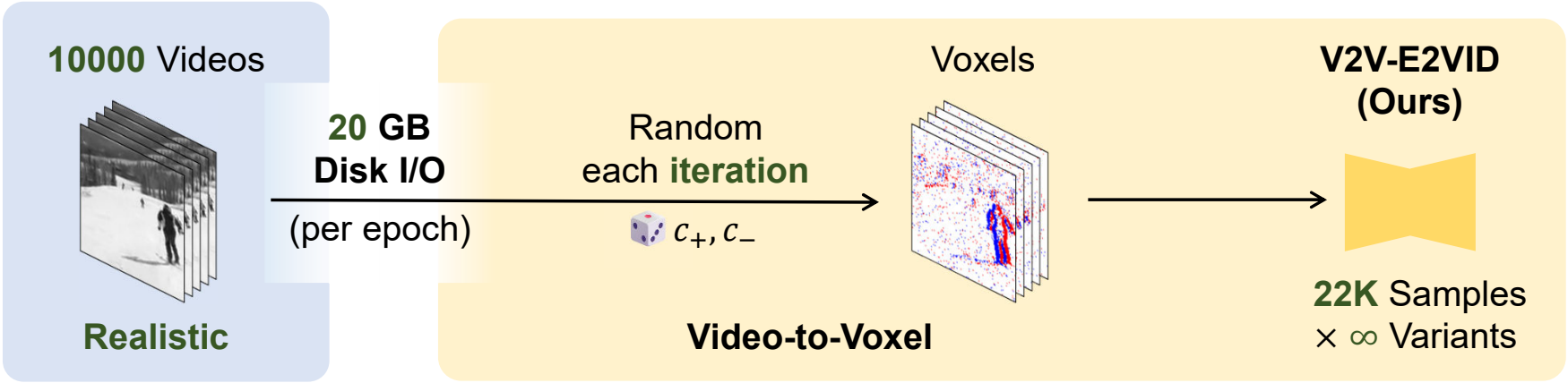
Weaknesses of current pipeline

- Storage and I/O bandwidth requirement is high
- N scenes × M camera parameters → N×M requirement
- Impossible to efficiently scale up



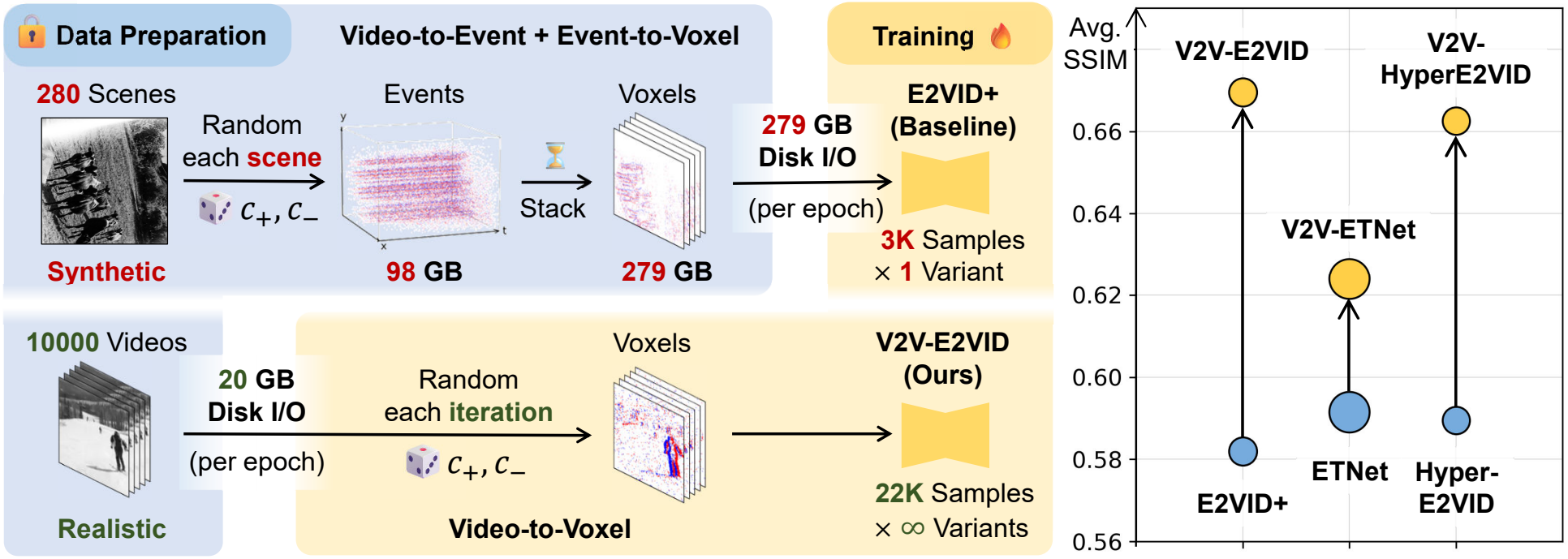
Our observation

- When the model uses discrete voxel representations, videos can be directed converted into voxels!
- Skipping event stream generation & storage
- Enabling parameter randomization in each iteration

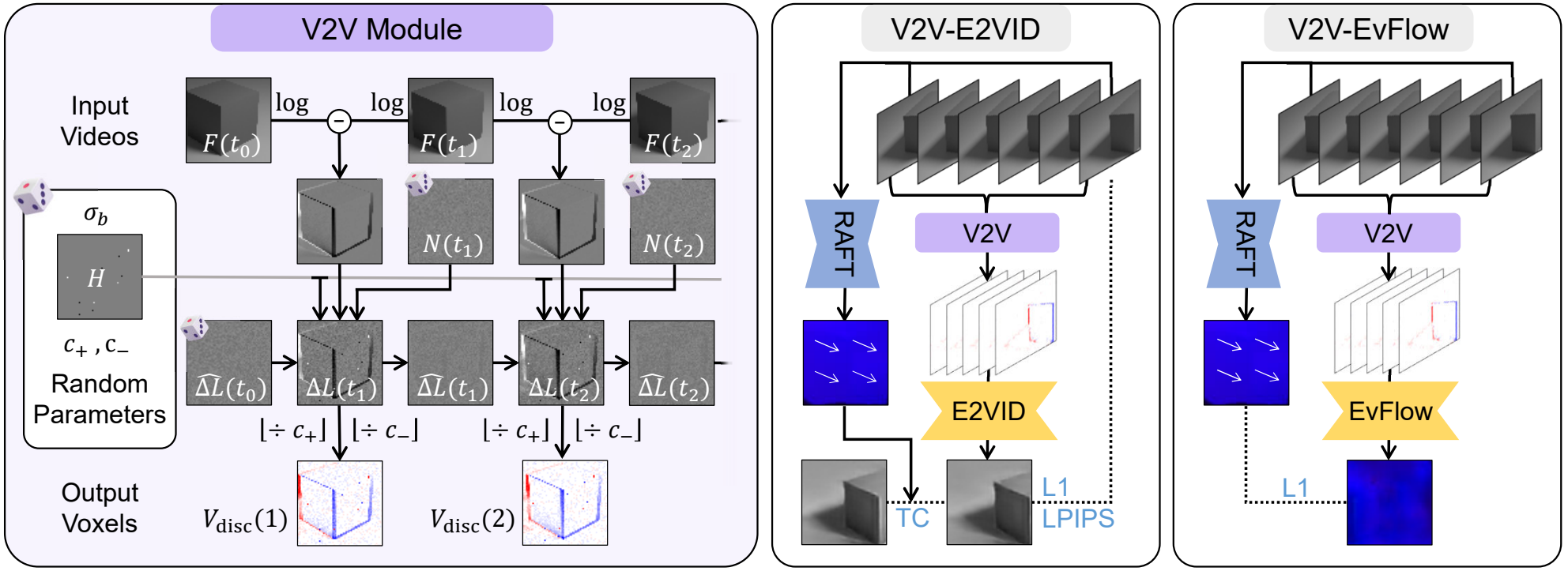


ESIM-280 vs V2V-WebVid10K

- Ours: 7.1% Storage & bandwidth usage; 33× Scene diversity; On-the-fly parameter randomization augmentation



V2V Module & Applications

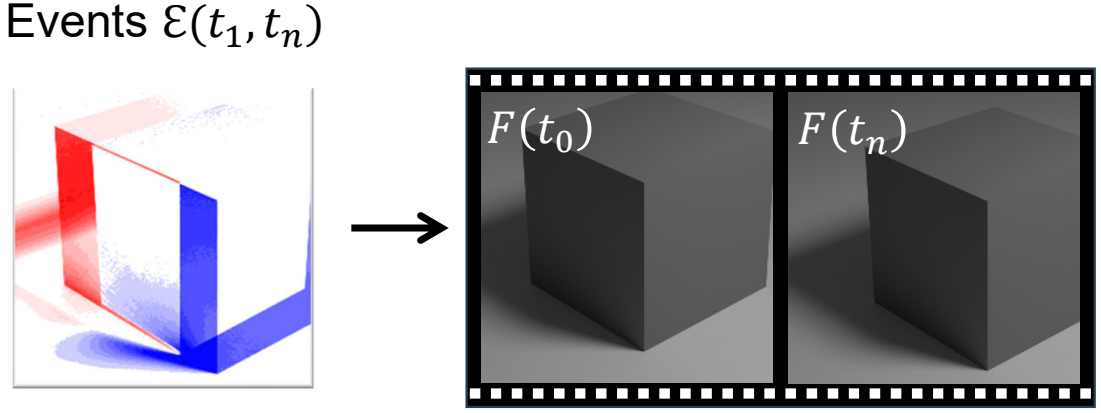


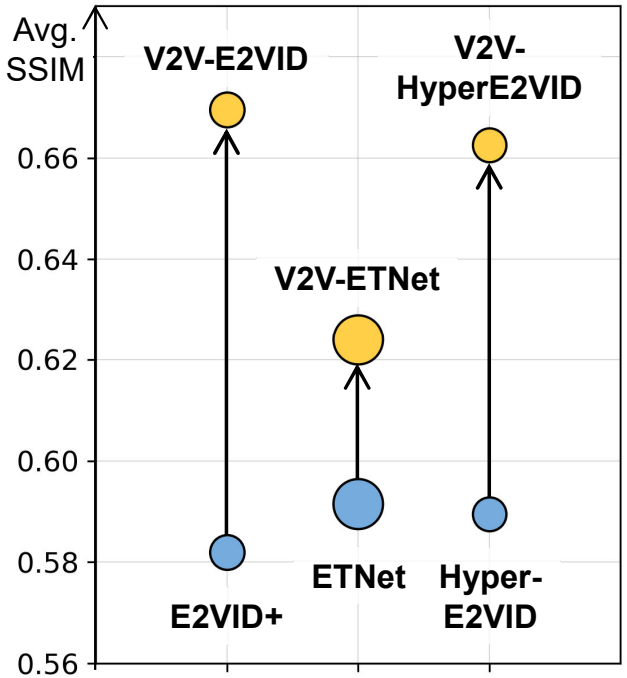
: Data Flow. ····· : Loss Constraints.

L1 : L1 Loss. LPIPS : LPIPS Loss. TC : Temporal Consistency Loss.

Application 1: Video reconstruction

- E2VID, ETNet and HyperE2VID: SOTA models trained with ESIM-280
- Retrain with V2V-WebVid10K
- Significant performance gain





Dataset quality of V2V-WebVid10K

- Scaling up with much more diversity.

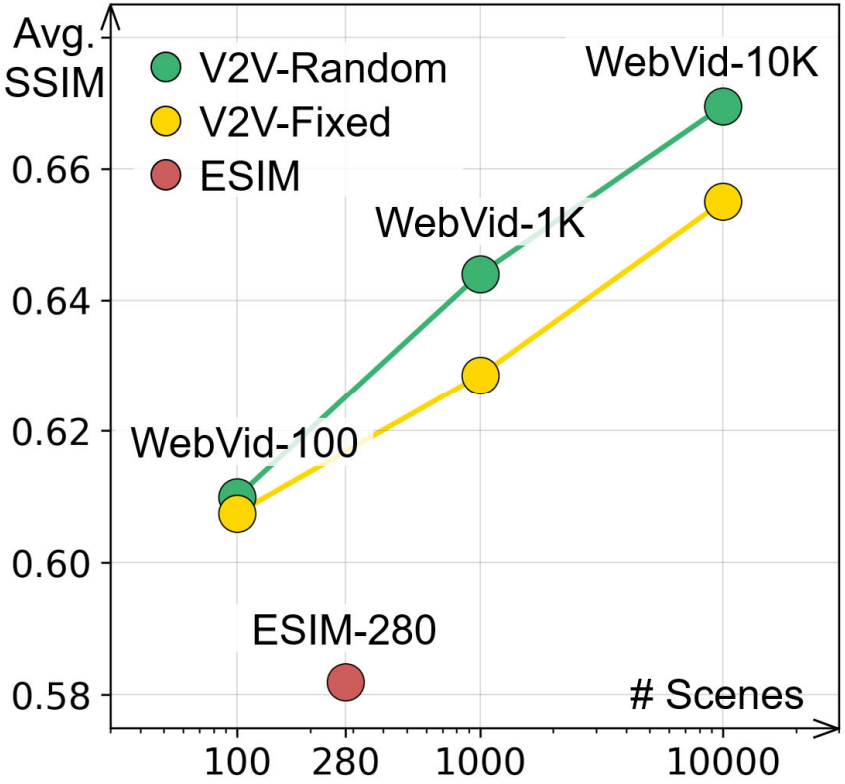
(a) WebVid scenes

(b) ESIM scenes

Where is the performance gain from?

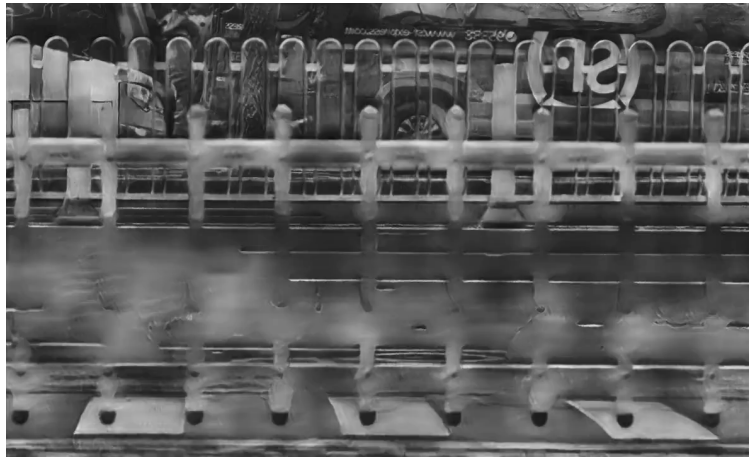
- More data diversity → Better priors
- Flexible parameter selection → More data augmentation

- Controlled variable experiment:
- #Videos 100→1000→10000
- Fixed thresholds→Random thresholds



Qualitative results of reconstruction

E2VID+ (Baseline)

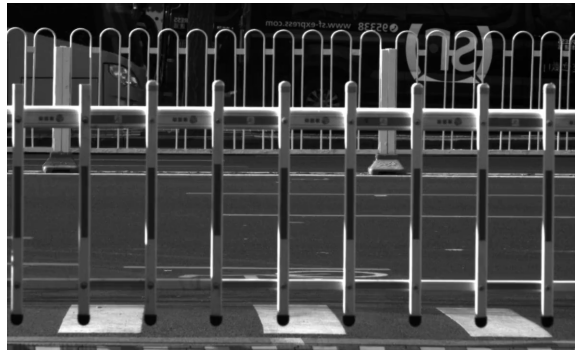


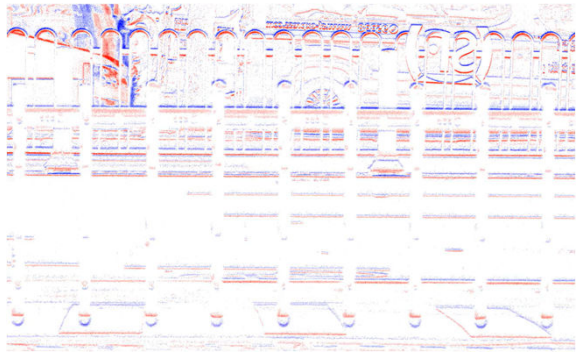
Model Architecture: E2VID
Training Data: ESIM-280 (Pretrained)
Representation: Interpolated Voxel
Loss: Alex + Temporal Consistency

V2V-E2VID (Ours)

Model Architecture: E2VID
Training Data: WebVid10K
Representation: Discrete Voxel
Loss: VGG + L1 + Half TC

Ground Truth





Events

Application 2: Optical flow estimation

- EvFlow+ : Originally trained on ESIM-280
- V2V-EvFlow: Retrain the model with WebVid-10K

Sequence	indoor_flying1		indoor_flying2		indoor_flying3		outdoor_day1		outdoor_day2	
AEE ↓	D	S	D	S	D	S	D	S	D	S
Zeros	1.772	2.041	2.705	3.463	2.412	2.772	3.985	3.585	2.793	2.562
EvFlow+	0.967	0.752	1.154	0.990	1.184	0.833	2.903	1.445	1.983	1.580
V2V-EvFlow	0.732	0.745	0.870	0.990	0.741	0.758	2.114	1.191	1.627	1.365
3PE (%) ↓	D	S	D	S	D	S	D	S	D	S
Zeros	14.5	20.2	36.1	52.5	30.2	38.6	57.1	54.5	32.9	31.7
EvFlowt+	2.7	0.8	5.4	2.4	7.0	1.6	36.1	12.1	19.8	14.4
V2V-EvFlow	0.5	0.5	1.3	2.3	0.4	0.6	23.6	8.4	15.0	11.2

Camera Intelligence
A Computational Photography Lab @ PKU

北京大学
PEKING UNIVERSITY

上海创智学院
Shanghai Innovation Institute

NEURAL INFORMATION PROCESSING SYSTEMS

Thank You!

Lab page

https://camera.pku.edu.cn

Code available

https://github.com/HYLZ-2019/V2V