Deep video models

Viorica Pătrăucean, Research scientist

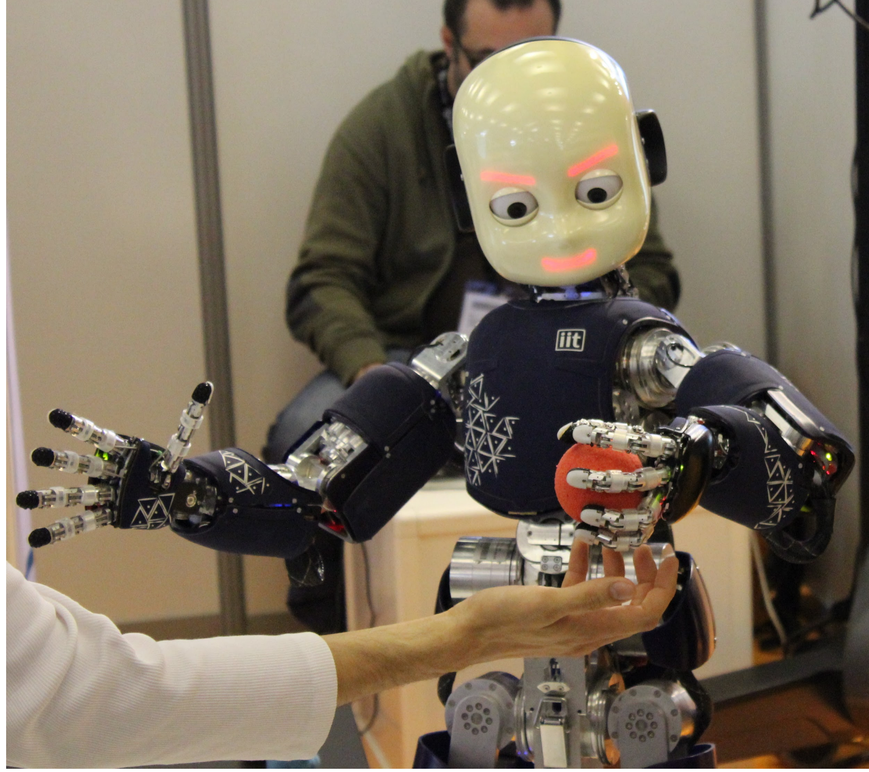
DeepMind

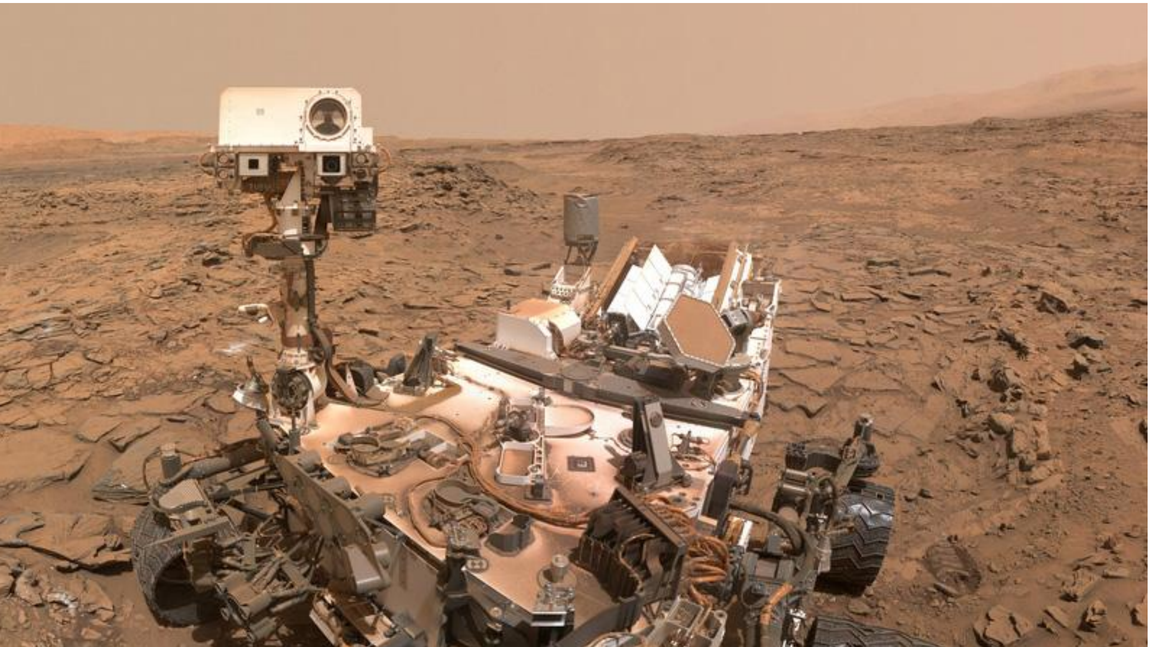
EEML

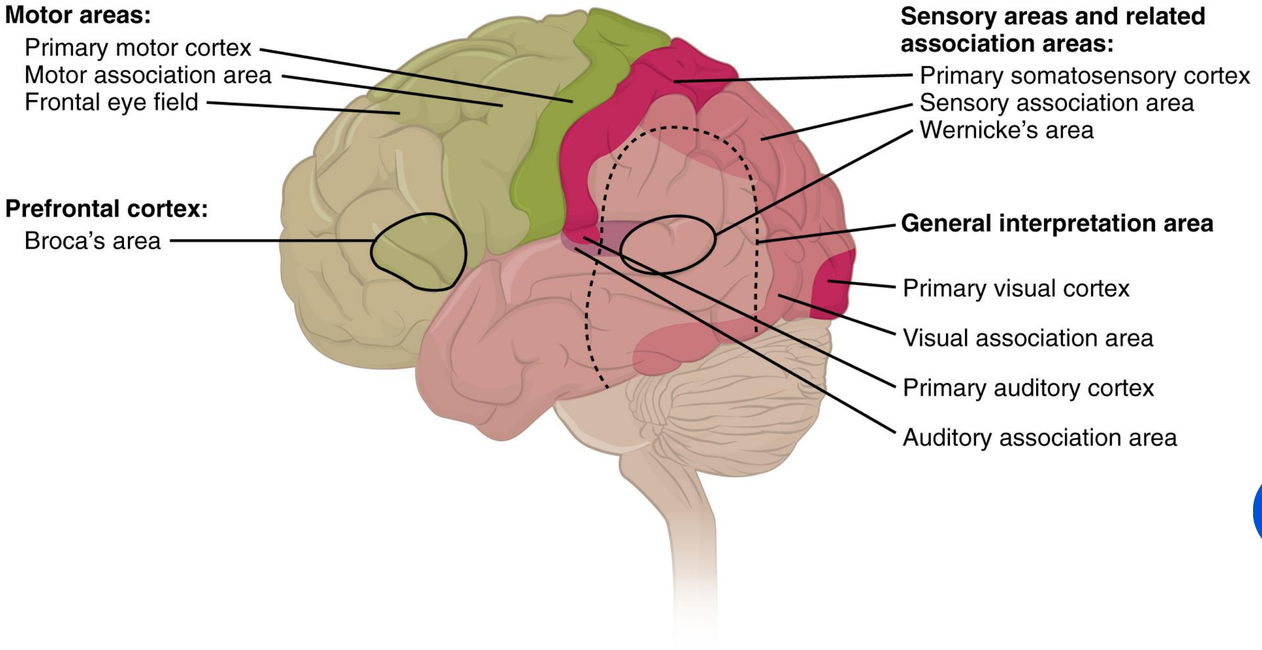
Vilnius Machine Learning Workshop 2021

About myself

- Undergrad in Bucharest Romania — Computer Engineering
- Master’s and PhD in Toulouse France — Applied Maths (2D image processing)
- PostDoc in Paris and Cambridge — Research on 3D shapes and videos
- Research scientist in DeepMind since 2016
- EEML organiser since 2017






Motor areas:
Primary motor cortex
Motor association area
Frontal eye field
Prefrontal cortex:
Broca's area
Sensory areas and related association areas:
Primary somatosensory cortex
Sensory association area
Wernicke's area
General interpretation area
Primary visual cortex
Visual association area
Primary auditory cortex
Auditory association area

Outline

01
Overview
Why videos?

Popular video models

Efficiency challenges

02
Learning paradigms
Multimodal learning

Self-supervised learning

03
Depth parallelism
Inference

Training

Biological plausibility

04
Conclusion
Summary

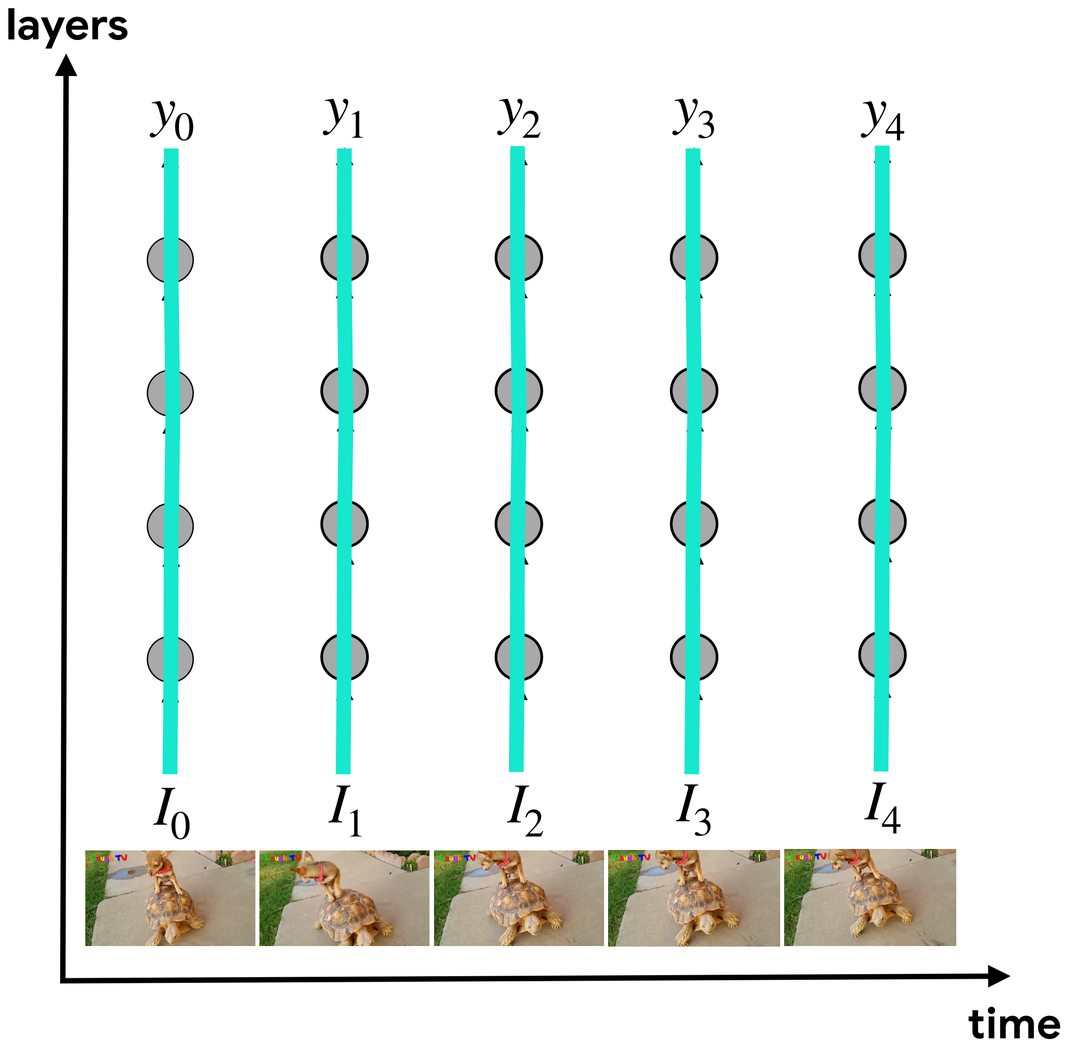
My vision for vision

Disclaimer: There are many concepts and works in this space. I will cover only a subset of them.

01

Overview

Why care about videos?



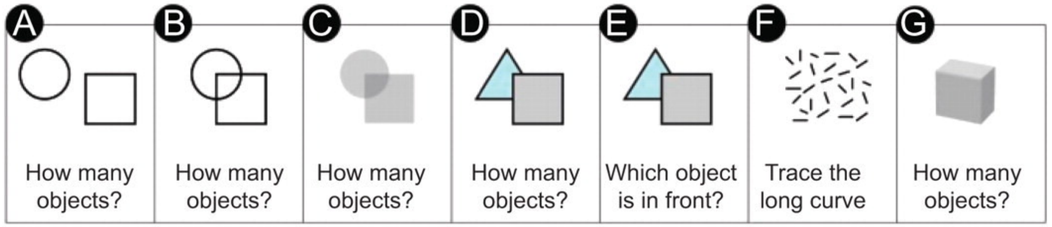
Frame-by-frame video model

- Inference: run image model on consecutive frames
- Image model trained on single video frames
- Cannot integrate temporal information, no motion features
- Cannot exploit temporal redundancy to improve efficiency

1. Frame-by-frame model

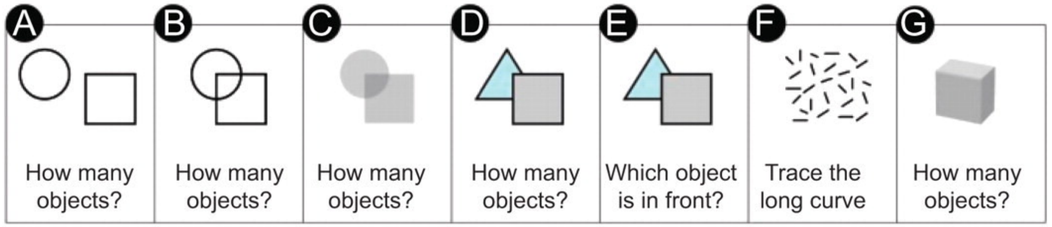
Improving Semantic Segmentation via Video Propagation and Label Relaxation, Zhu et al, 2019

Importance of motion



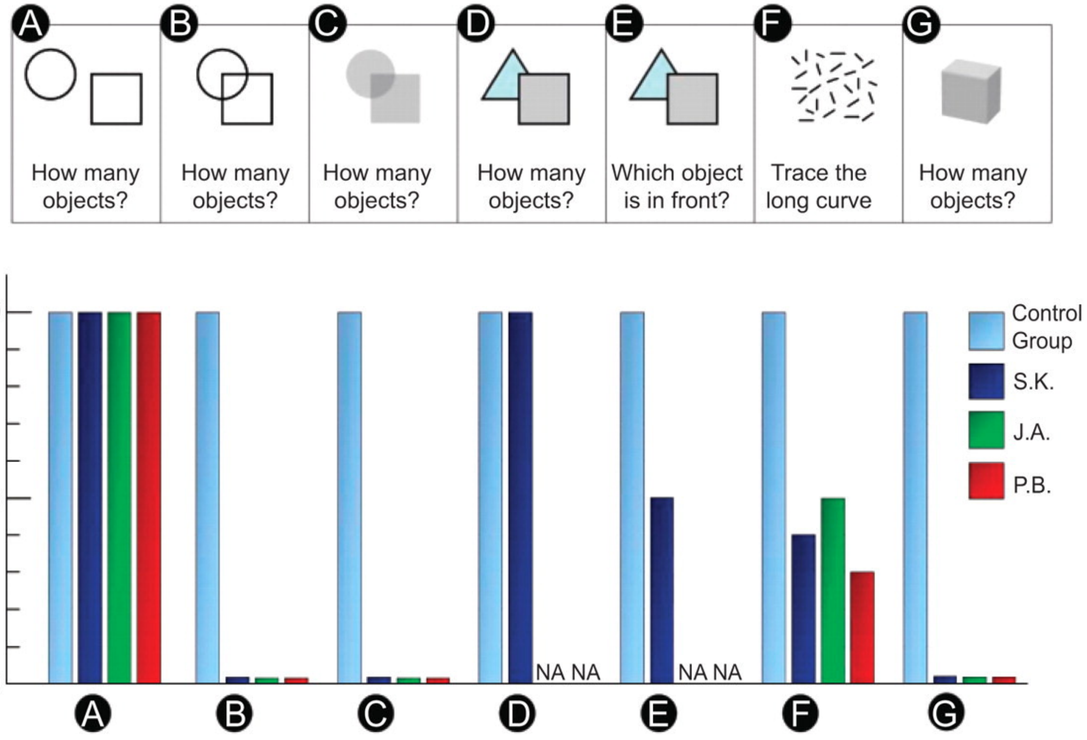
Visual Parsing After Recovery From Blindness, Ostrovsky et al, 2009

Importance of motion



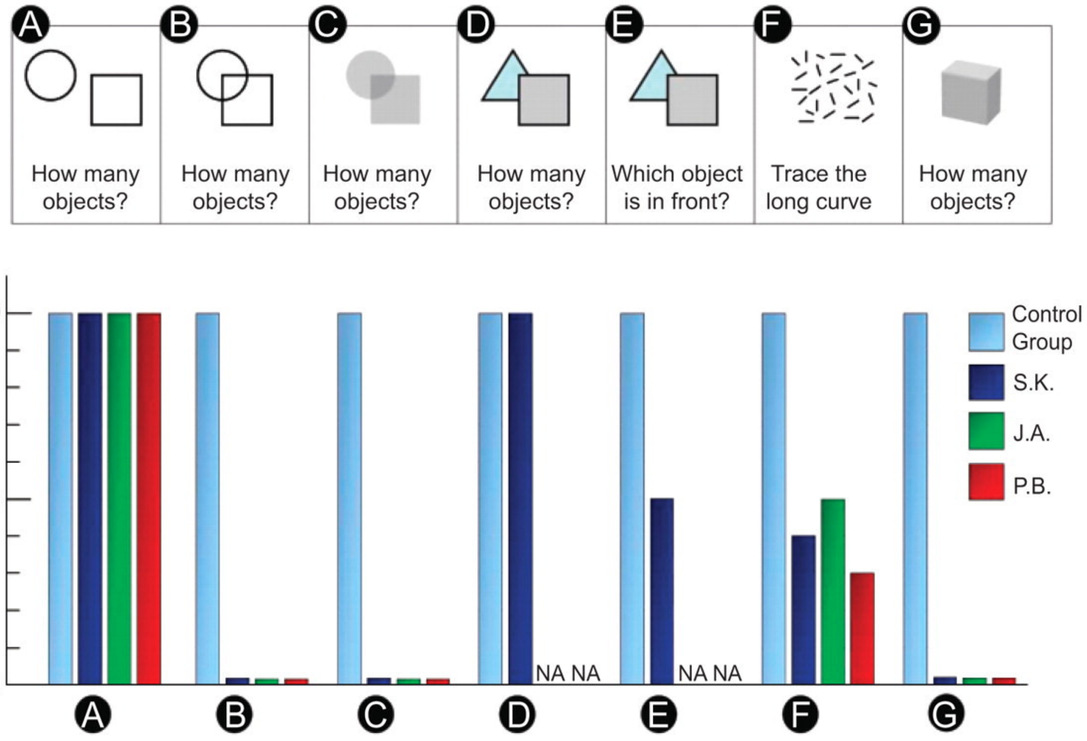
Visual Parsing After Recovery From Blindness, Ostrovsky et al, 2009

Importance of motion



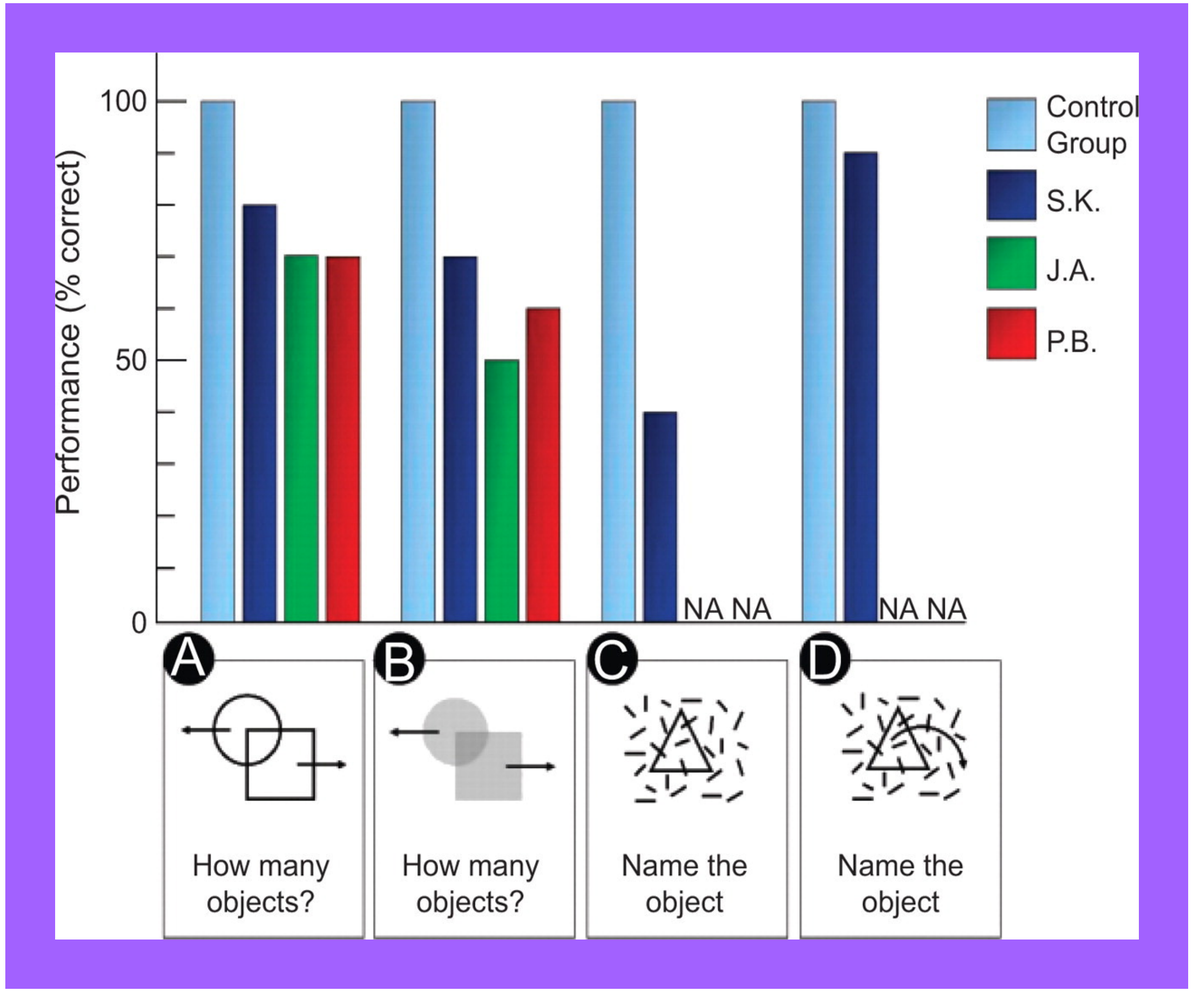
Visual Parsing After Recovery From Blindness, Ostrovsky et al, 2009

Importance of motion



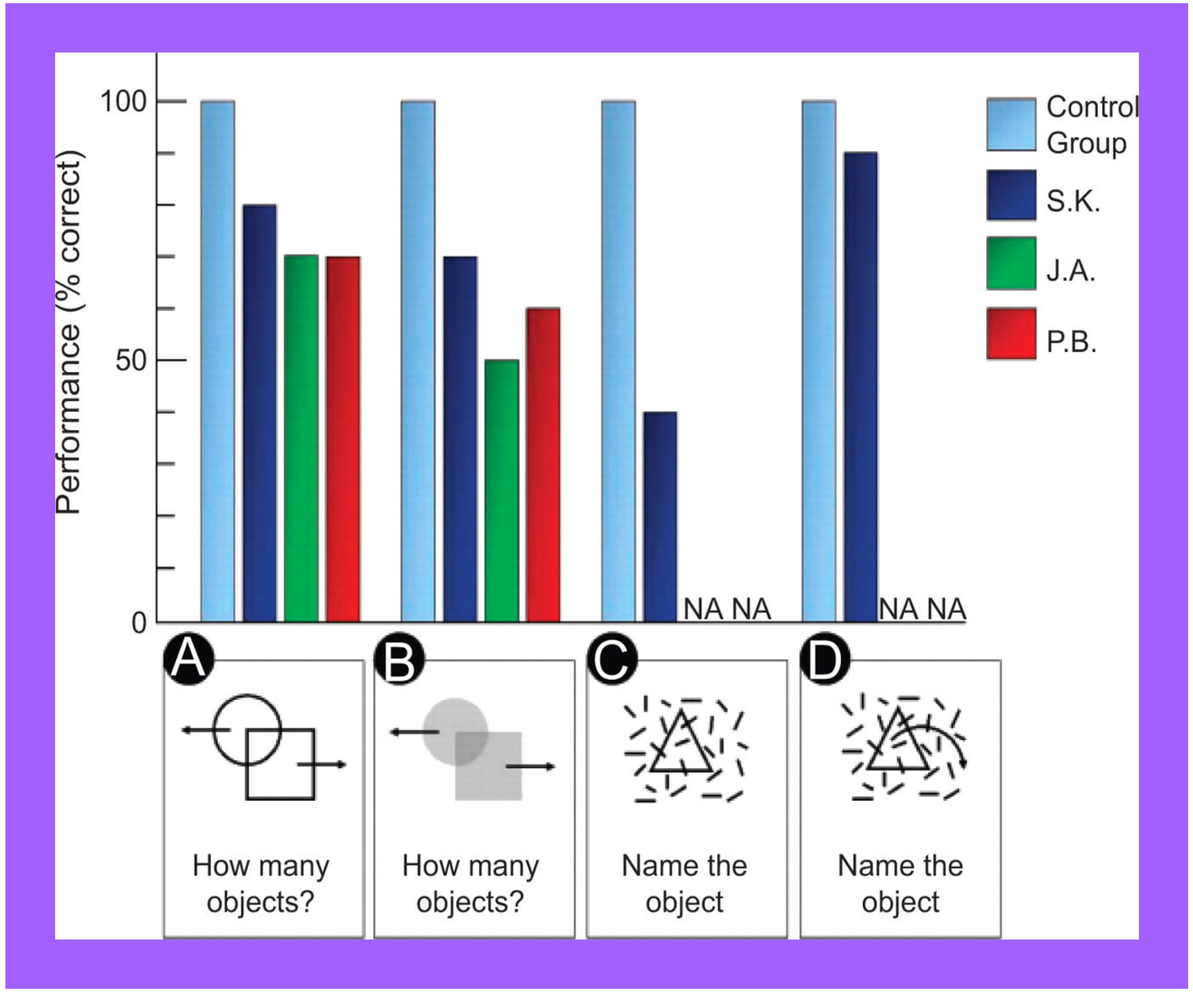
Visual Parsing After Recovery From Blindness, Ostrovsky et al, 2009

Importance of motion



Visual Parsing After Recovery From Blindness, Ostrovsky et al, 2009

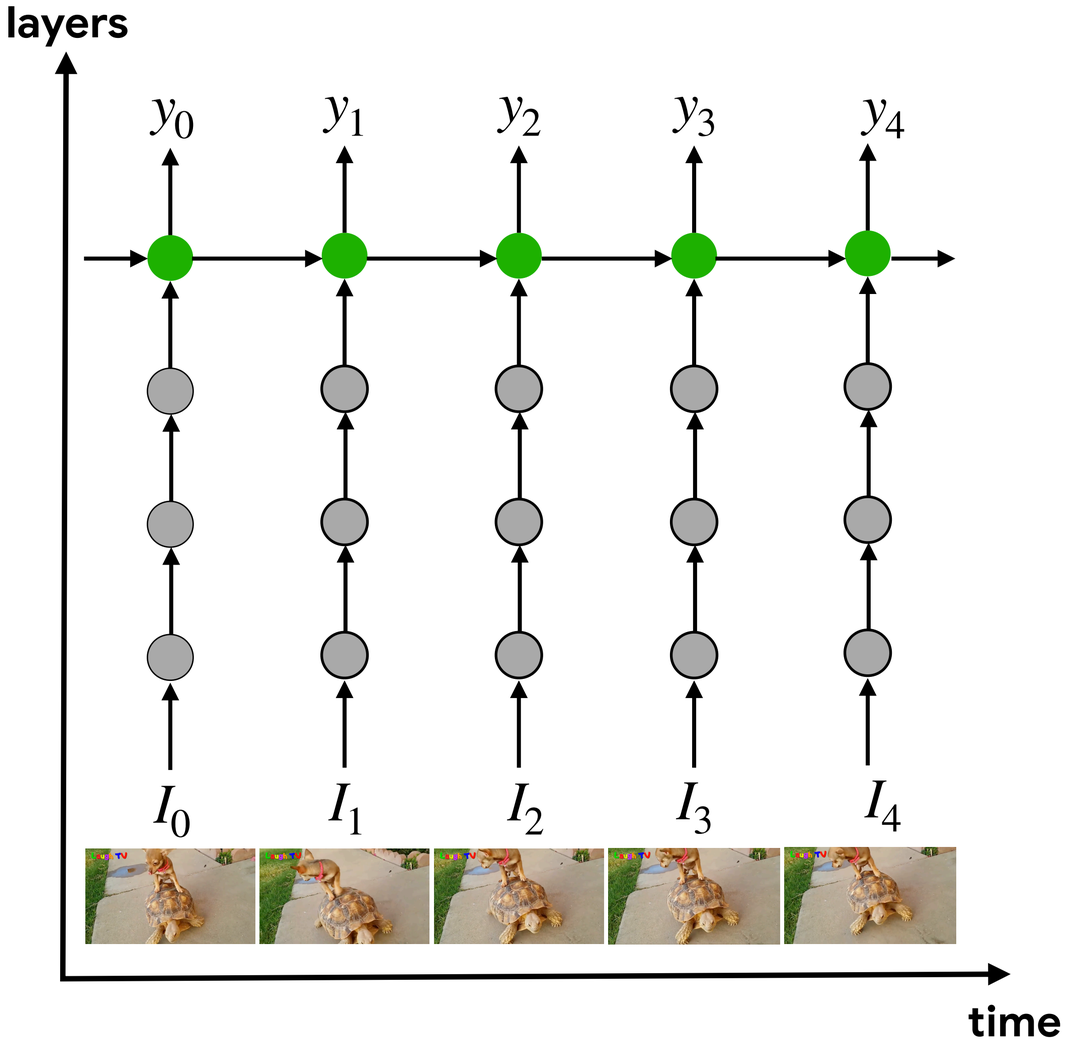
Importance of motion



Motion helps object recognition when learning to see.

Visual Parsing After Recovery From Blindness, Ostrovsky et al, 2009

2. Image encoder + RNN

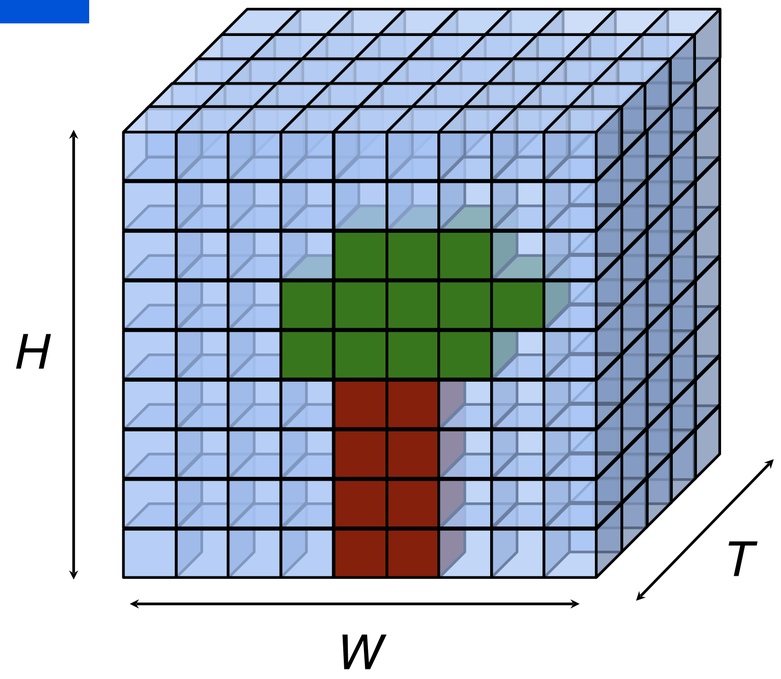


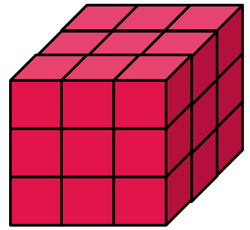
- Image encoder: standard image model (pre-trained on Imagenet)
- RNN: LSTM / GRU fully connected or convolutional
- Need large memory for training with BPTT

3. 3D convolutional models

Video as a volume

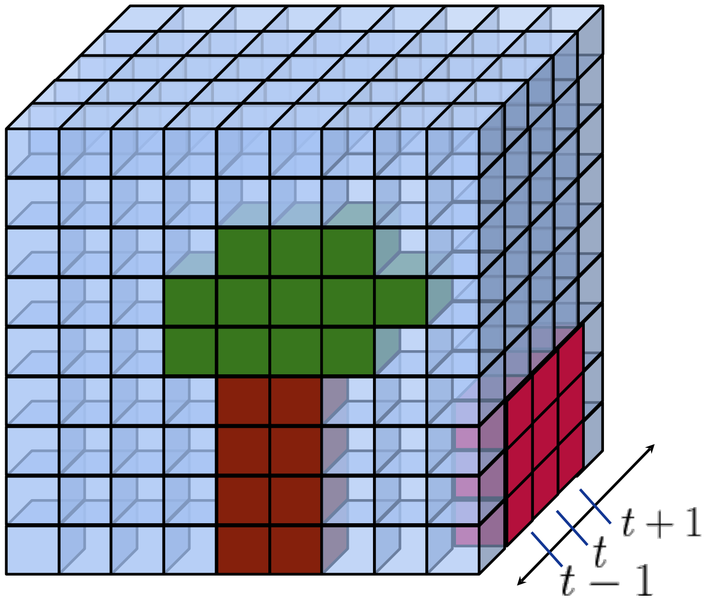
- stack frames *T* x *H* x *W* x *3*
- apply 3D convolutions

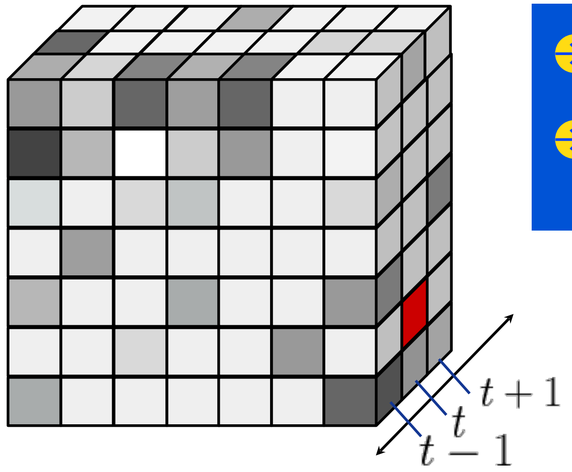




$$y = \sum_{i \in 3 \times 3 \times 3} \mathbf{w}_i \mathbf{x}_i + b$$

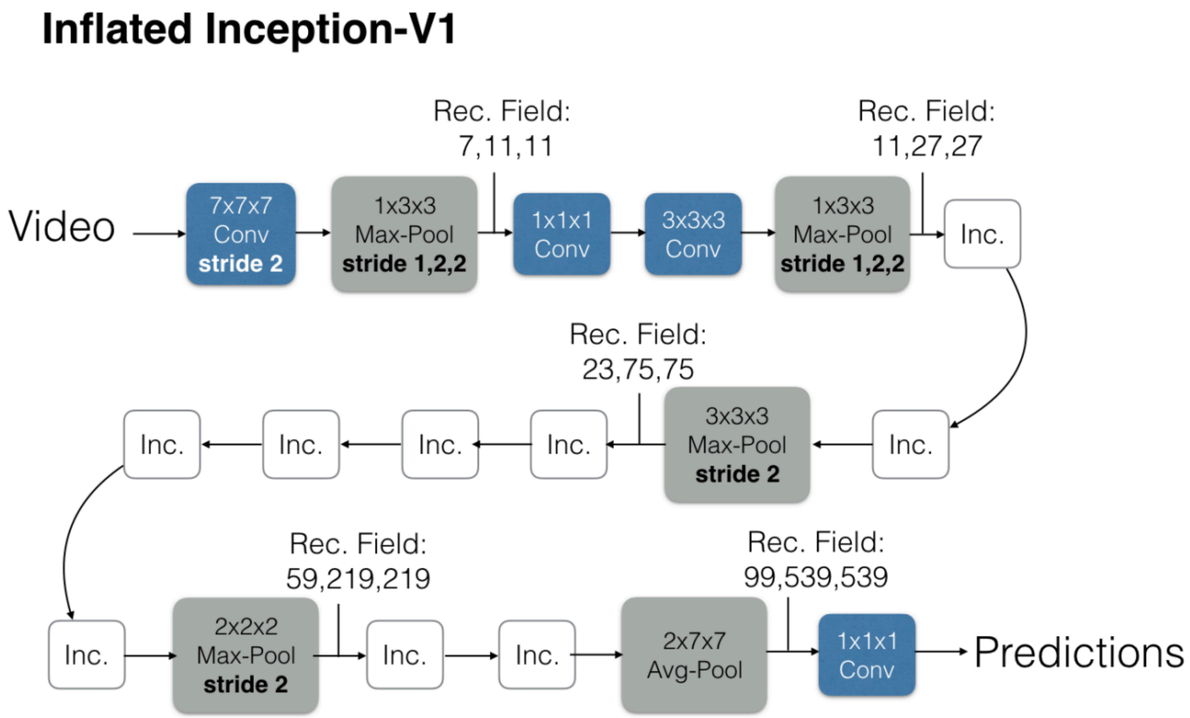
3. 3D convolutional models





- 3D convolutions are non-causal
- masked 3D convolutions are causal

3. 3D convolutional models



- 3D convnet
- Temporal striding
- Possibly inflated 2D model (transfer learning)

Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, Carreira and Zisserman, 2018

Models for action recognition

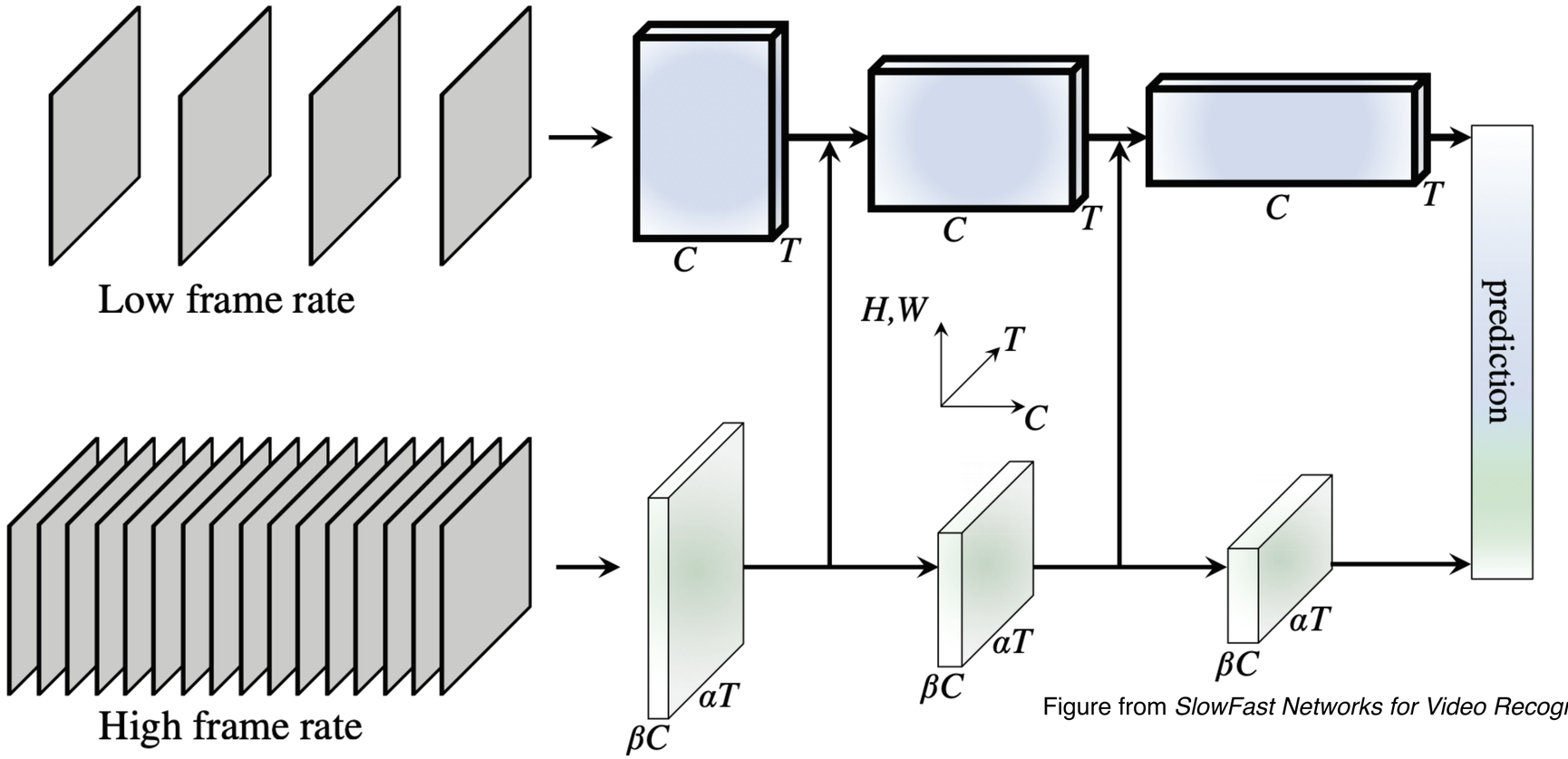
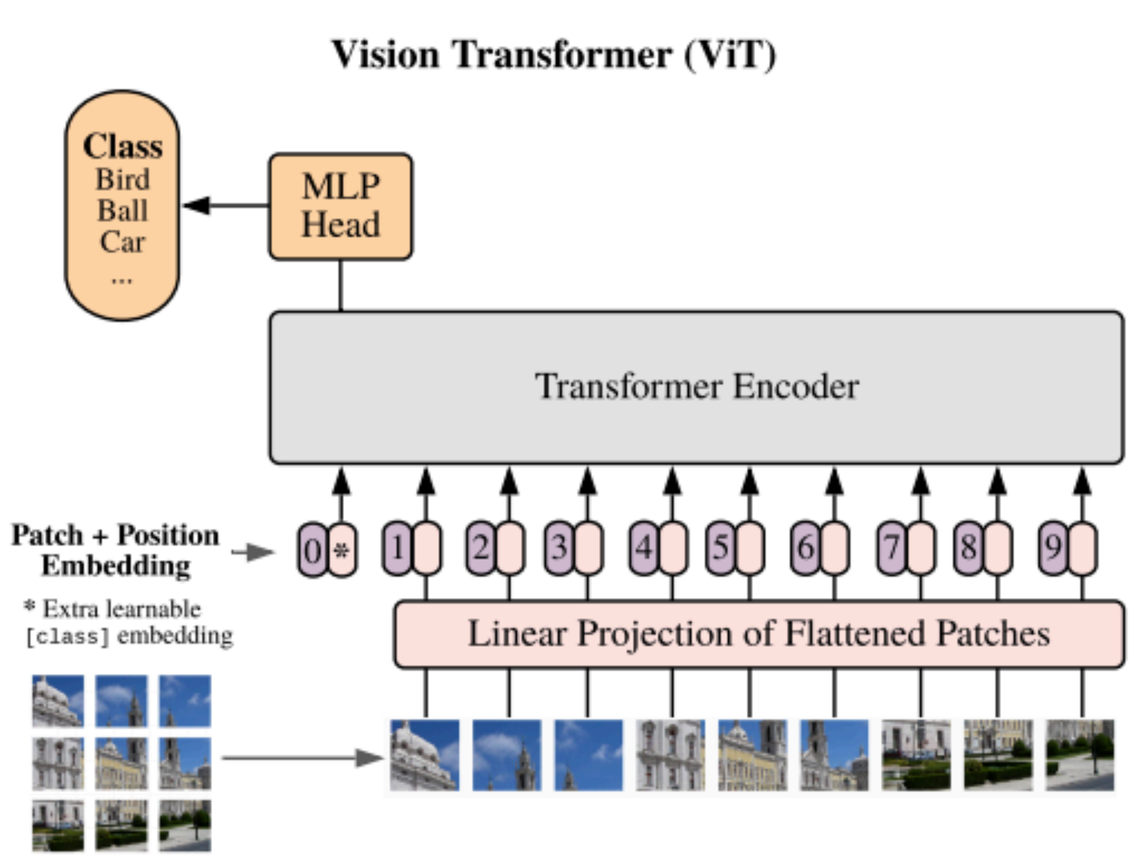
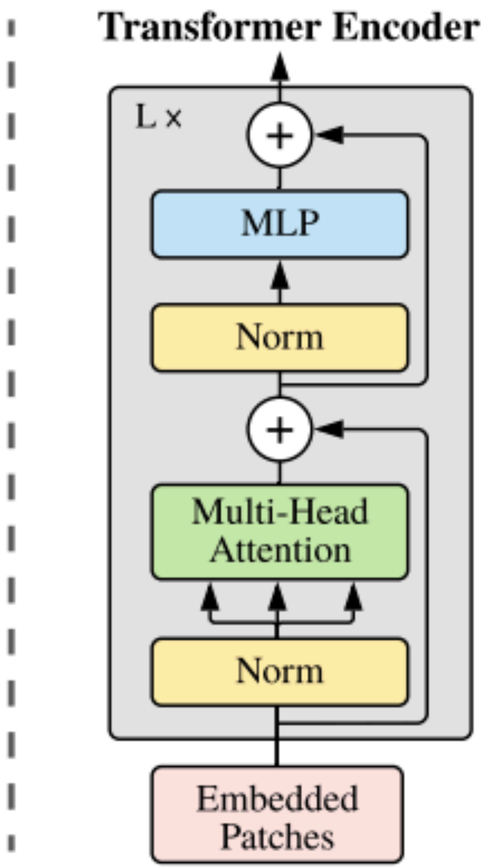


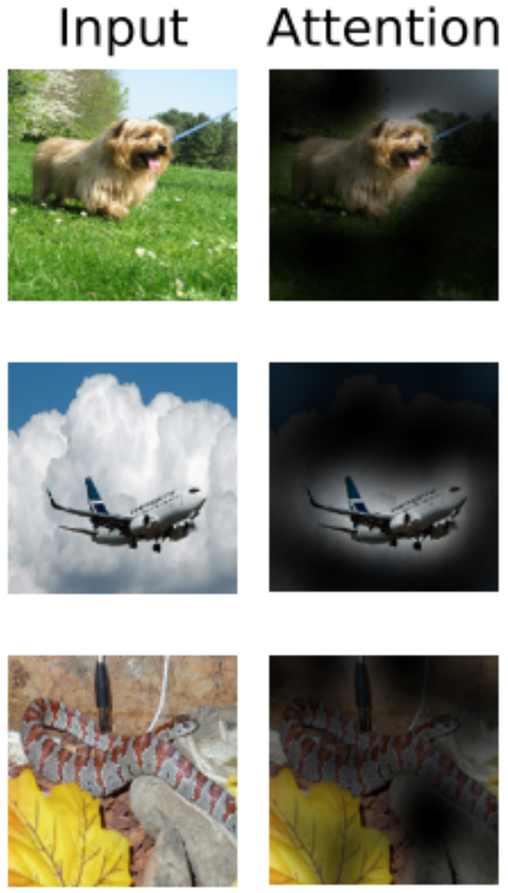
Figure from *SlowFast Networks for Video Recognition*, Feichtenhofer et al, 2019

I3D – Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, Carreira and Zisserman, 2018
SlowFast Networks for Video Recognition, Feichtenhofer et al, 2019
TSM: Temporal Shift Module for Efficient Video Understanding, Lin et al, 2018
.....

4. Self-attention models — Vision Transformers







Figures from *An image is worth 16x16 words*, Dosovitskiy et al, 2021

4. Self-attention models — Video Transformers

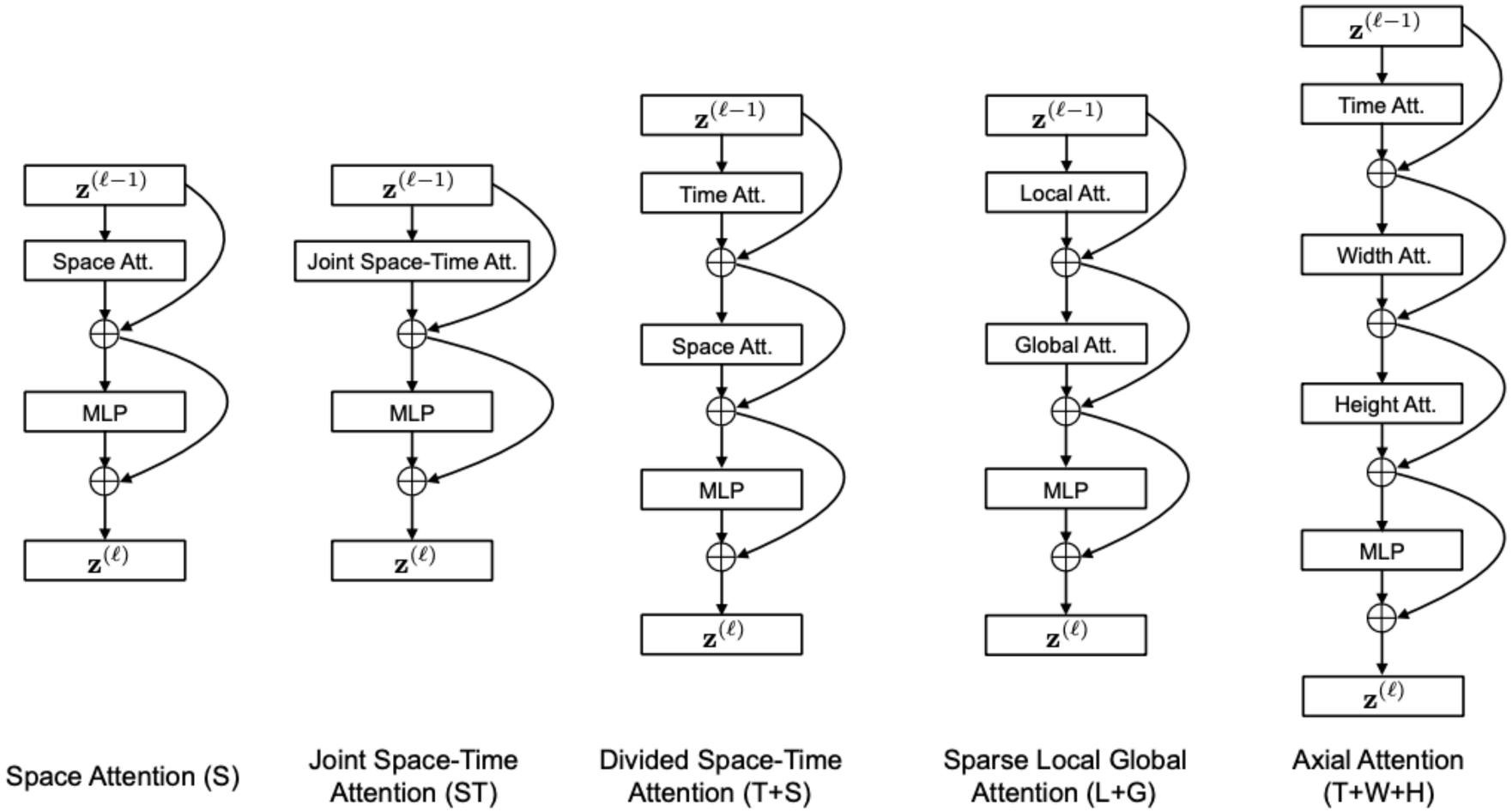
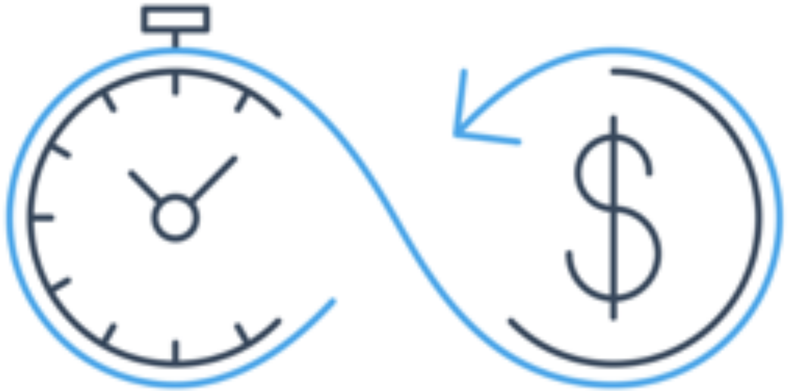
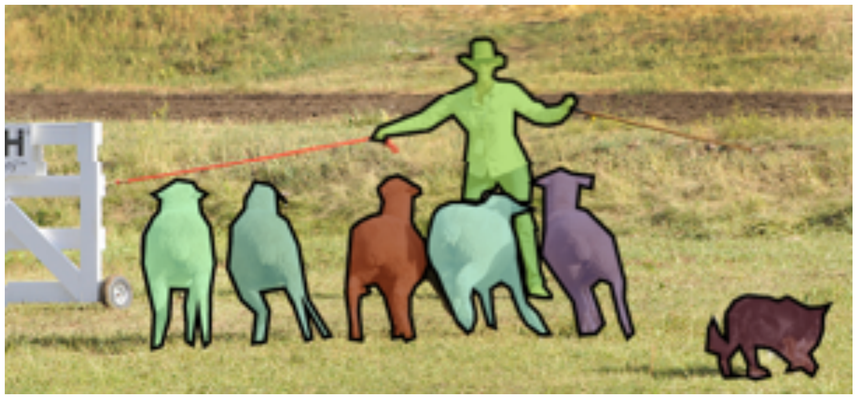


Figure from *Is Space-Time Attention All You Need for Video Understanding?* Bertasius et al, 2021

Challenges in video processing



Labels are expensive



Agreement: definition? granularity?

Challenges in video processing

✓ Ok
✓ Not great, not terrible
✗ Pretty bad

	Frame-by-frame (2D) models	2D-encoder + RNN	3D-conv models	Video transformer
Accuracy	✗	✓	✓	✓
Latency / throughput @ inference	✓	✓	✗	✗
Memory @ train	✓	✗	✗	✗
Memory @ test	✓	✓	✗	✗
Energy efficiency	✗	✗	✗	✗
Causal	✓	✓	✗	✗

Accuracy and latency/throughput dependent on the depth of the model
Video transformer - quadratic complexity in all dimensions

Latency & throughput

Video from Davis2017 dataset @ 25fps

Same video @ 5fps

Attempts to improve them

Boost efficiency of image models

- Model compression / binarisation
  Chen et al., Courbarieux et al.
- Distillation
  Hinton et al.
- Lighter convolutional modules
  MobileNet, Xception
- Budget methods
  Karayev et al., Mathe et al.

Improve efficiency of video models

- Temporal multi-scale models
  I3D, SlowFast, TSN
- 2D *temporal* models
  TSM, R(2d+1)
- Reuse of features (warping)
  Zhu et al.
- Variable update rates
  Shelhamer et al.

Self-supervised methods

02

Learning paradigms

Learning landscape

Want to learn more?

Zador. **A critique of pure learning and what artificial neural networks can learn from animal brains (2019)**

Dataset $\{x_1, x_2, \ldots, x_n\}$

Proxy task $\{(x_1, z_1), (x_2, z_2), \ldots, (x_n, z_n)\}$

Labels generated automatically

Learn mapping g $\hat{z} = g_w(x)$

Transfer learning from g

Self-supervised learning

Dataset $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

Model $p(Y \mid X = x)$

Learn mapping f $\hat{y} = f(x)$

Maximise similarity $(\hat{y}, y)$

Supervised learning

Dataset $\{x_1, x_2, \ldots, x_n\}$

Model $p(X)$

Learn the structure in the data

Unsupervised learning

Evolution

Environment

MDP (S, A, R, T)

Learn policy $\pi_\theta : S \rightarrow A$

Maximise reward

Reinforcement learning

Downstream tasks in videos

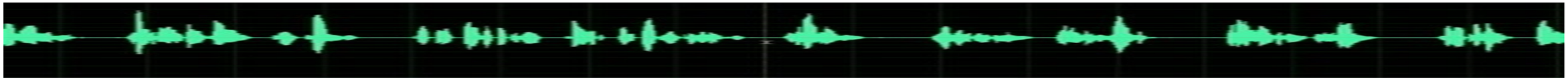
Tracking / segmentation

Video from Davis dataset

Action recognition

Video from Kinetics dataset

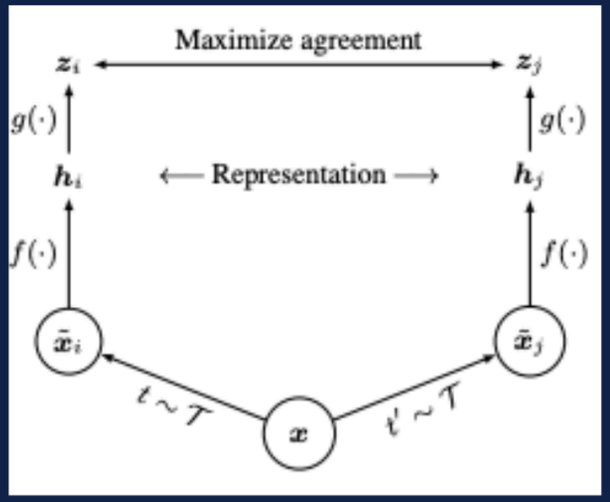
Multimodal sensing of the environment



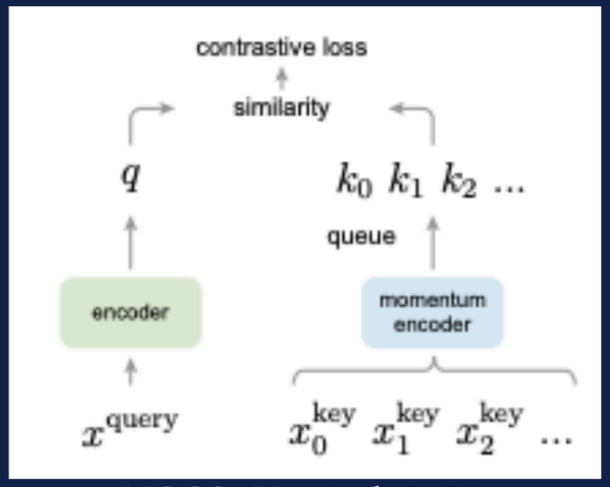
[...] towards the root and try to get as close to the root as possible, nice long strokes [...]

Self-supervised learning for pre-training

Vision

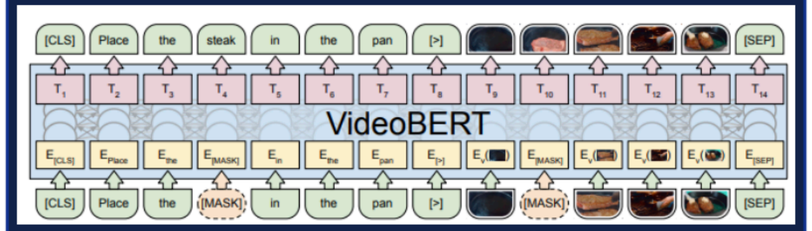


SimCLR: Chen et al, 2020

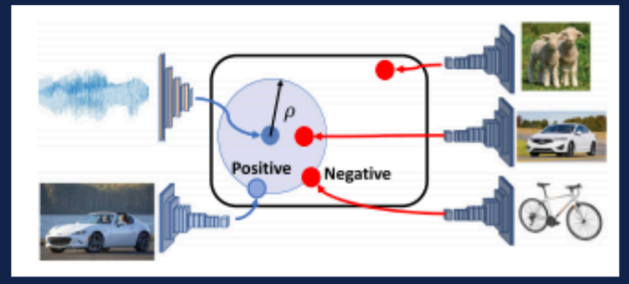


MOCO: He et al, 2020

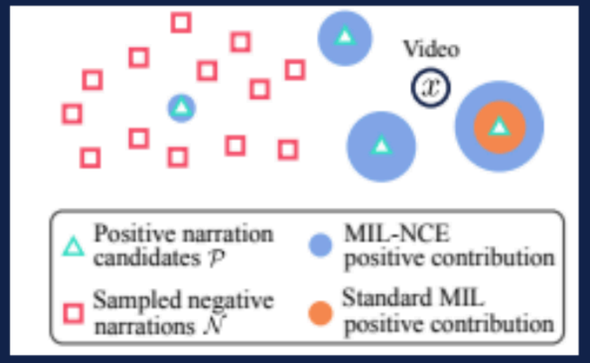
Vision+Language



VideoBERT: Sun et al, 2019

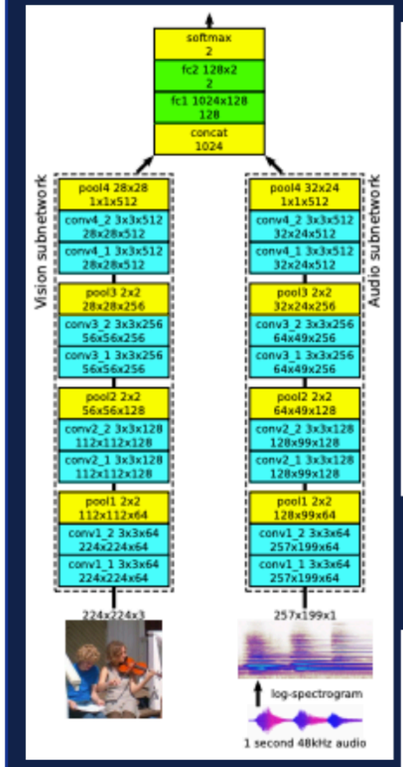


DaveNet: Harwath et al, 2018

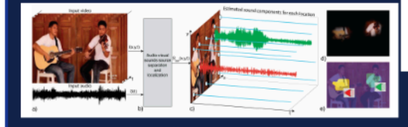


MIL-NCE: Miech, Alayrac et al, 2020

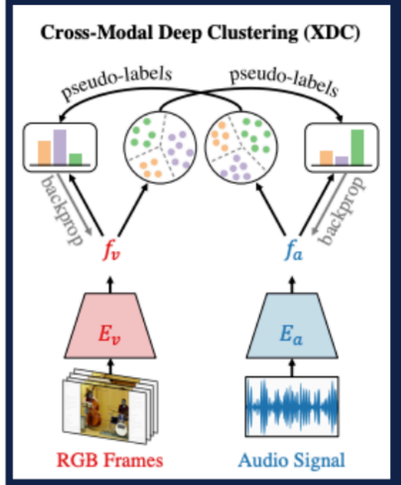
Vision+Audio



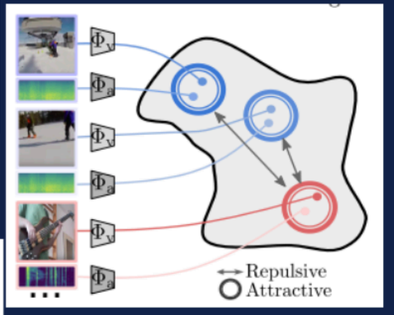
L3: Arandjelovic and Zisserman, 2017



Sound of Pixels: Zhao et al, 2018



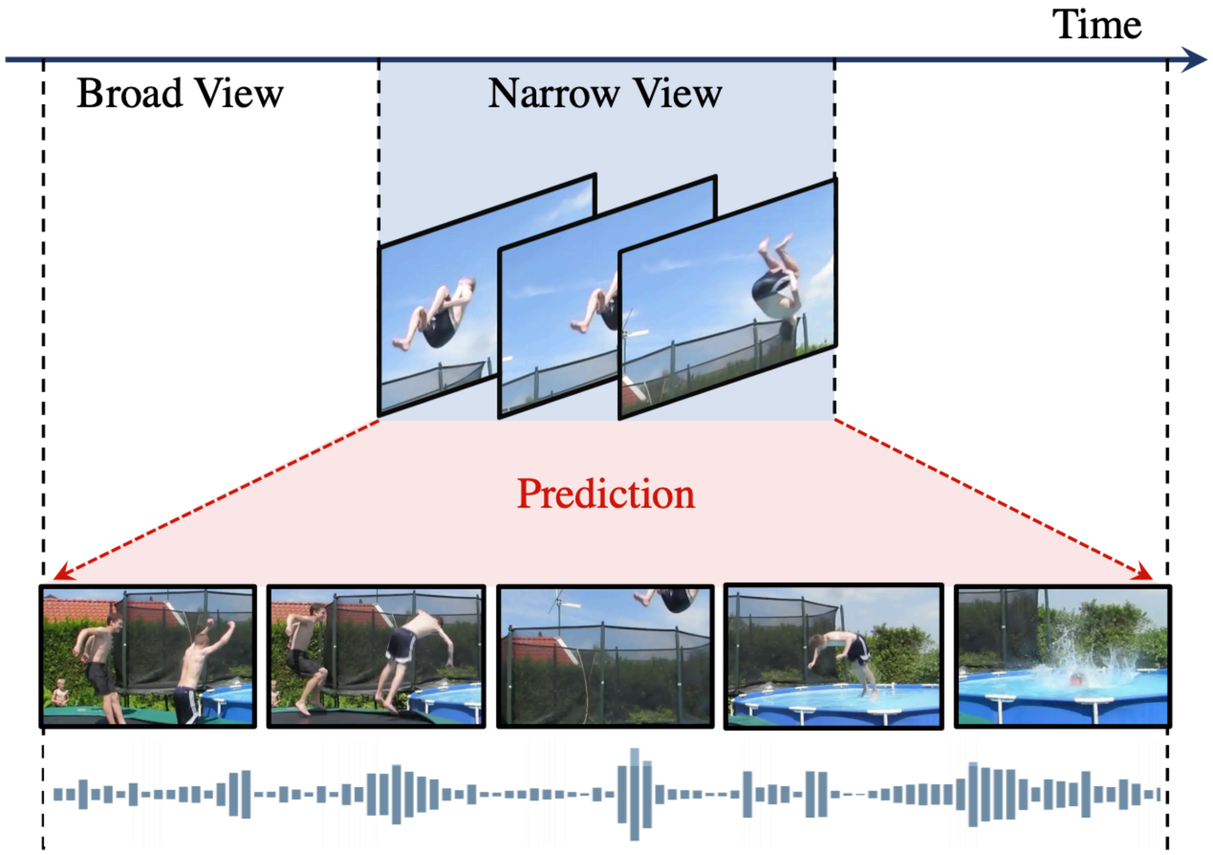
XDC: Alwassel at al, 2020



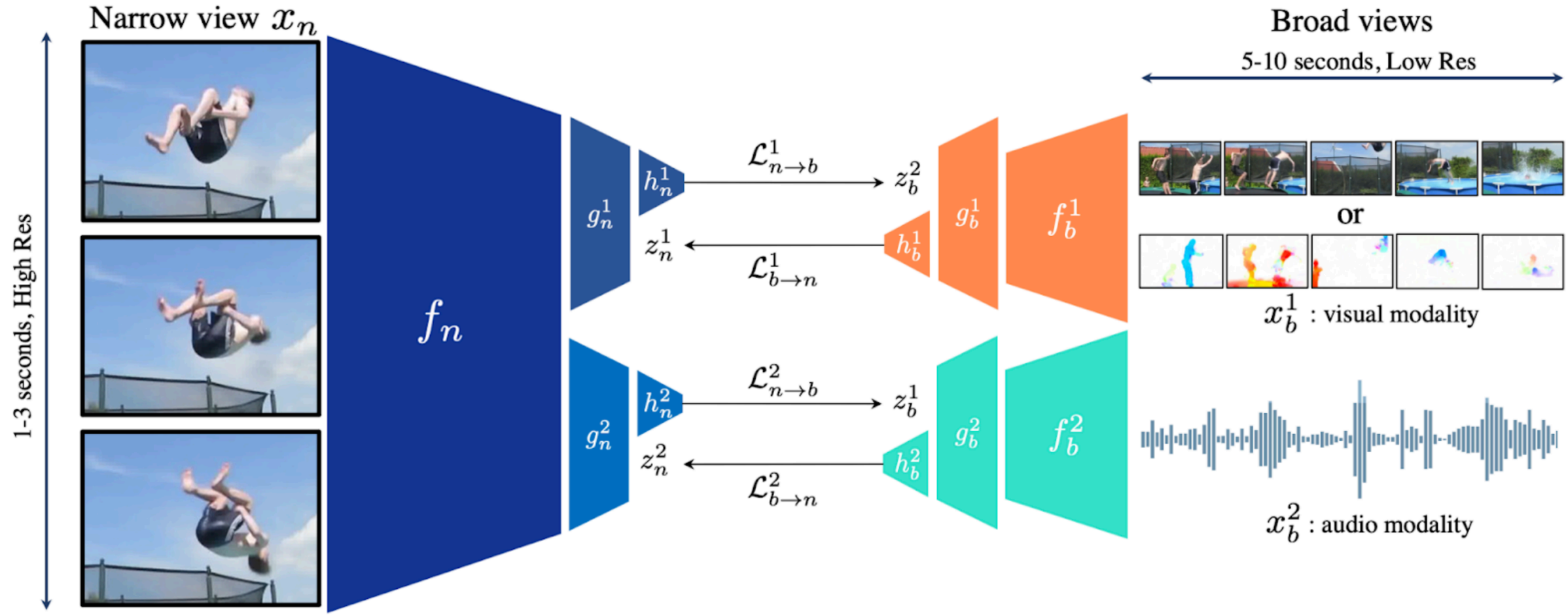
GDT: Patrick at al, 2020

Broaden your views for self-supervised video learning

Recasens et al, ICCV2021



BraVe


Narrow view x_n
1-3 seconds, High Res
f_n
g_n^1
h_n^1
$\mathcal{L}_{n\to b}^1$
z_b^2
z_n^1
$\mathcal{L}_{b\to n}^1$
h_b^1
g_b^1
f_b^1
g_n^2
h_n^2
$\mathcal{L}_{n\to b}^2$
z_b^1
z_n^2
$\mathcal{L}_{b\to n}^2$
h_b^2
g_b^2
f_b^2
Broad views
5-10 seconds, Low Res
or
x_b^1 : visual modality
x_b^2 : audio modality

BraVe training loss

Global loss

$$\mathcal{L}(x) = \underbrace{\mathcal{L}_{n\to b}(x)}_{\text{Narrow}\to\text{Broad}} + \underbrace{\mathcal{L}_{b\to n}(x)}_{\text{Broad}\to\text{Narrow}}$$

Narrow to Broad Loss

$$\mathcal{L}_{n\to b}(x) = \left\| \frac{h_n(z_n)}{\|h_n(z_n)\|_2} - \mathrm{sg}\left[\frac{z_b}{\|z_b\|_2}\right] \right\|_2^2$$

Broad to Narrow Loss

$$\mathcal{L}_{b\to n}(x) = \left\| \frac{h_b(z_b)}{\|h_b(z_b)\|_2} - \mathrm{sg}\left[\frac{z_n}{\|z_n\|_2}\right] \right\|_2^2$$

BraVe results

- When using the same backbone and dataset, our model significantly beats the state-of-the-art.

					UCF101		HMDB51		K600	ESC-50	AS
Method	Backbone (#params)	Dataset	Years	$\mathcal{M}$	Linear	FT	Linear	FT	Linear	Linear	MLP
AVTS [45]	MC3 (11.7M)	AS	1	VA		89.0		61.6		80.6	
ELo [67]	R(2+1)D-50 (46.9M)	YT8M	13	VFA		93.8	64.5	67.4			
AVID [58]	R(2+1)D-50 (46.9M)	AS	1	VA		91.5		64.7		89.2	
GDT [64]	R(2+1)D-18 (33.3M)	AS	1	VA		92.5		66.1		88.5	
MMV [3]	R(2+1)D-18 (33.3M)	AS	1	VA	83.9	91.5	60.0	70.1	55.5	85.6	29.7
XDC [4]	R(2+1)D-18 (33.3M)	AS	1	VA		93.0		63.7		84.8	
XDC [4]	R(2+1)D-18 (33.3M)	IG65M	21	VA		95.5		68.9		85.4	
BraVe:V↔A (ours)	R(2+1)D-18 (33.3M)	AS	1	VA	89.9	94.1	64.8	71.1	63.6	90.4	34.7
BraVe:V↔A (ours)	TSM-50 (23.5M)	AS	1	VA	93.0	94.8	69.4	72.6	70.1	90.5	34.4
BraVe:V↔FA (ours)	TSM-50 (23.5M)	AS	1	VFA	93.1	95.4	70.0	74.6	69.3	90.1	34.5
BraVe:V↔FA (ours)	R(2+1)D-50 (46.9M)	AS	1	VFA	92.5	95.1	68.3	73.6	69.4	**91.6**	34.5
BraVe:V↔FA (ours)	TSM-NF-F0 (71.5M)	AS	1	VFA	**94.1**	**95.8**	**71.4**	73.1	**72.6**	90.2	34.5
BraVe:V↔FA (ours)	TSM-50x2 (93.9M)	AS	1	VFA	93.1	95.7	70.5	**77.8**	71.4	91.1	**34.8**
Supervised [11, 44, 67, 87]						96.8	71.5	75.9	82.4	94.7	43.9

03

Depth parallelism

Attempts to improve efficiency of video models

Boost efficiency of image models

- Model compression / binarisation
  Chen et al., Courbarieux et al.
- Distillation
  Hinton et al.
- Lighter convolutional modules
  MobileNet, Xception
- Budget methods
  Karayev et al., Mathe et al.

Improve efficiency of video models

- Temporal multi-scale models
  I3D, SlowFast, TSN
- 2D *temporal* models
  TSM, R(2d+1)
- Reuse of features (warping)
  Zhu et al.
- Variable update rates
  Shelhamer et al.

All models are executed sequentially over depth

Our direction: Is sequential mode optimal?

NUMBERS | NEUROSCIENCE

Why Is the Human Brain So Efficient?

How massive parallelism lifts the brain's performance above that of AI.

BY LIQUN LUO

APRIL 12, 2018

Our direction: Is sequential mode optimal?

General-purpose processors use **pipelined** instructions enabling **parallel** processing.

Maximise throughput

Efficient use of hw resources

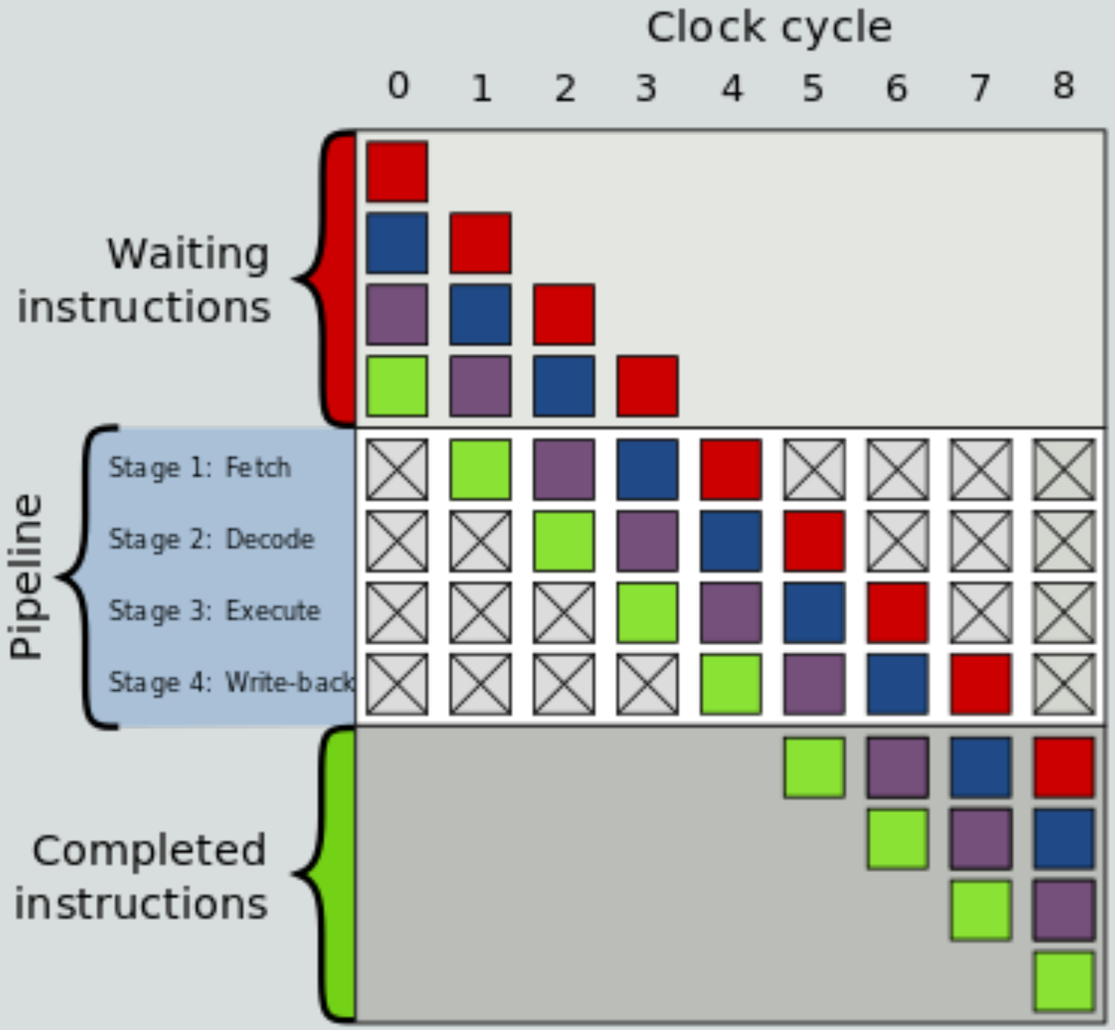


Figure from https://en.wikipedia.org/wiki/Instruction_pipelining

Outline

Depth-parallel video models

A. Depth-parallel inference

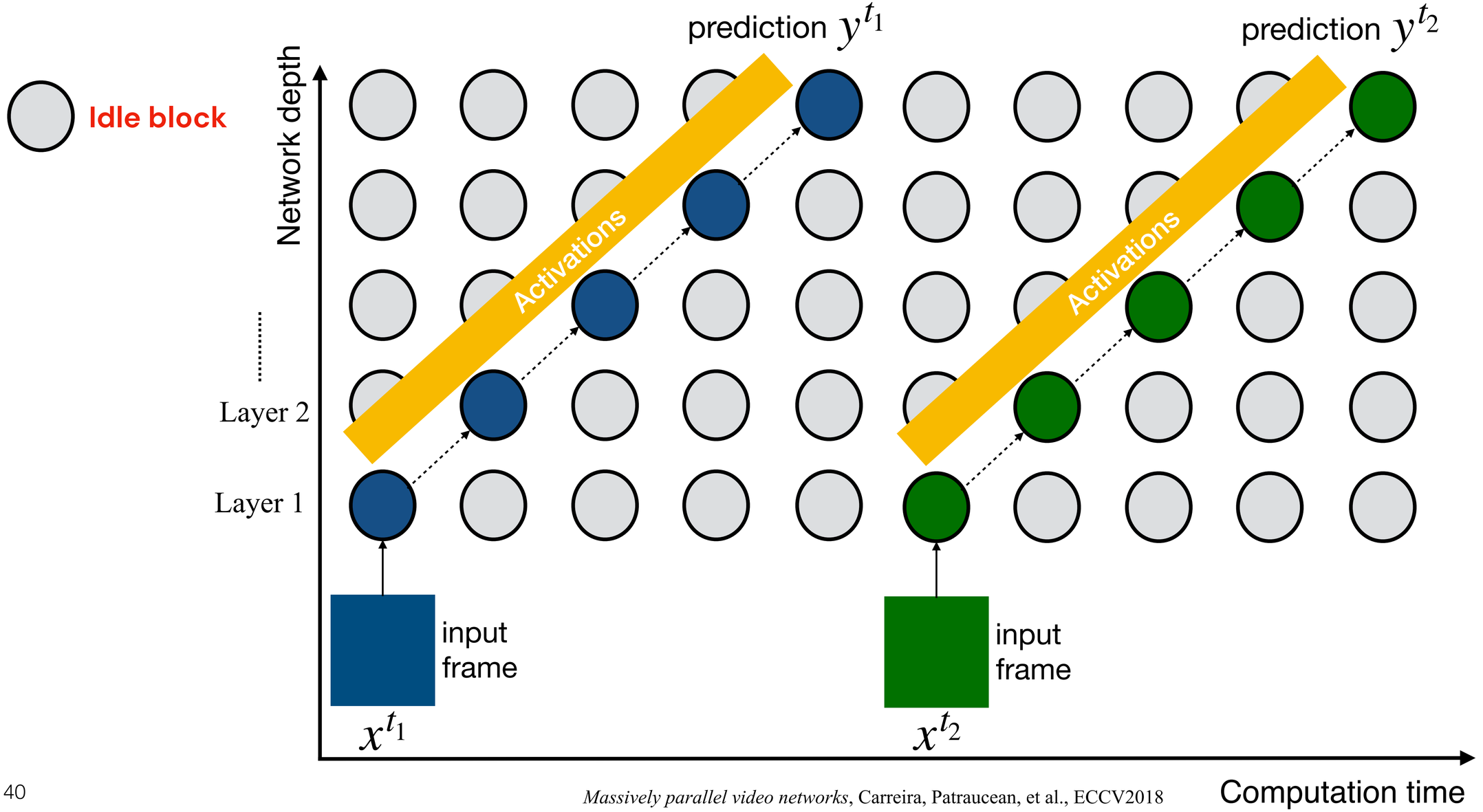
B. Depth-parallel training

Goals

- Increase throughput
- Reduce latency
- Reduce clock cycles (energy)
- Biological plausibility

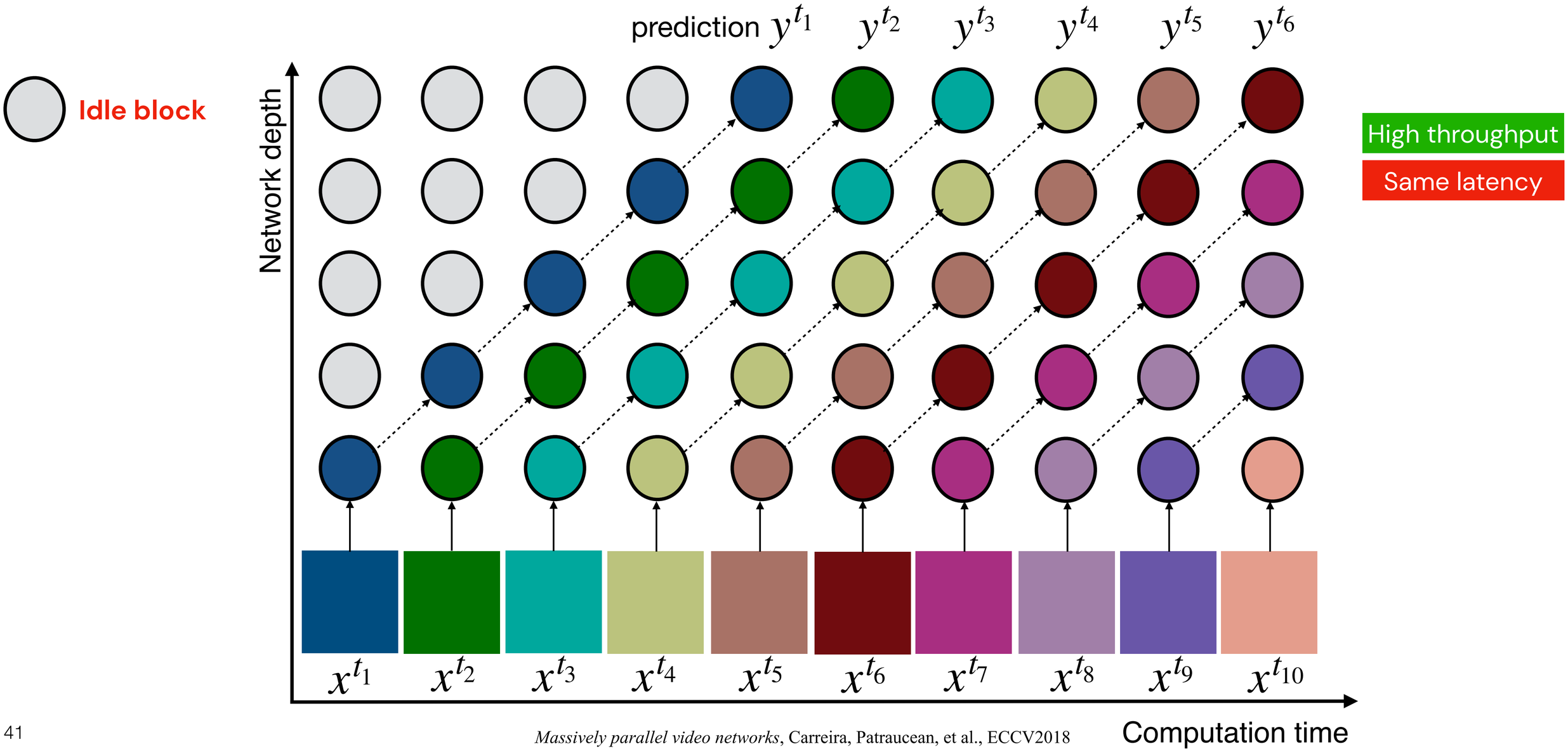
A. Depth-parallel inference

Sequential (frame-by-frame) model



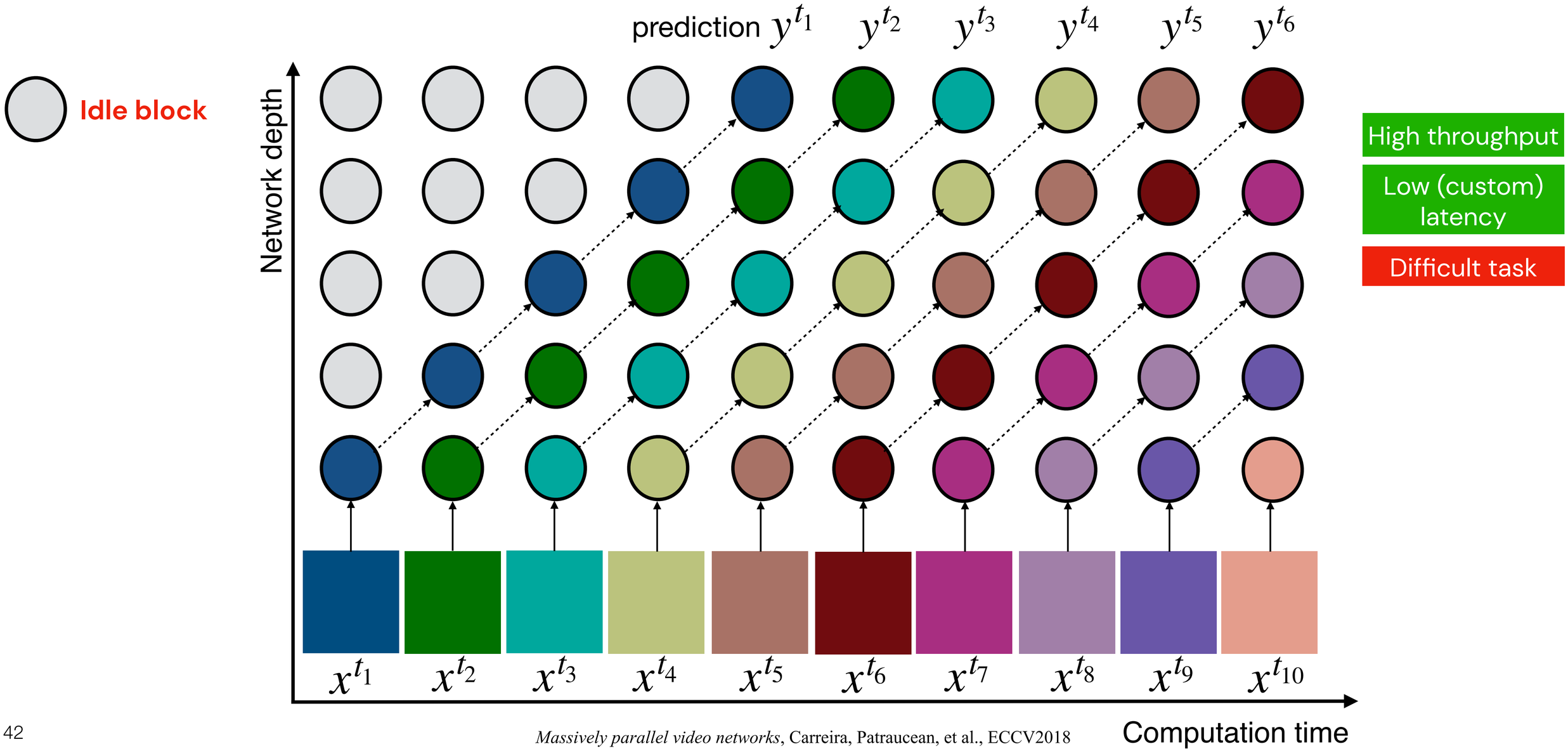
Massively parallel video networks, Carreira, Patraucean, et al., ECCV2018

Pipelined model



Massively parallel video networks, Carreira, Patraucean, et al., ECCV2018

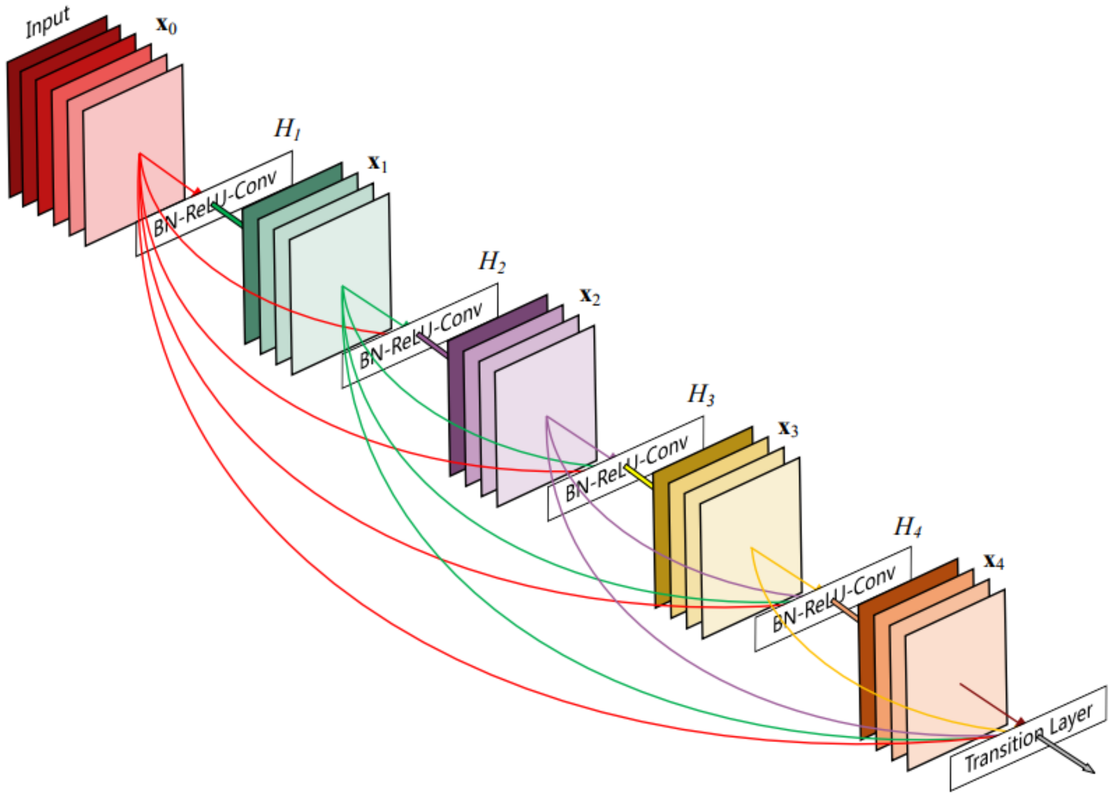
Pipelined model



Massively parallel video networks, Carreira, Patraucean, et al., ECCV2018

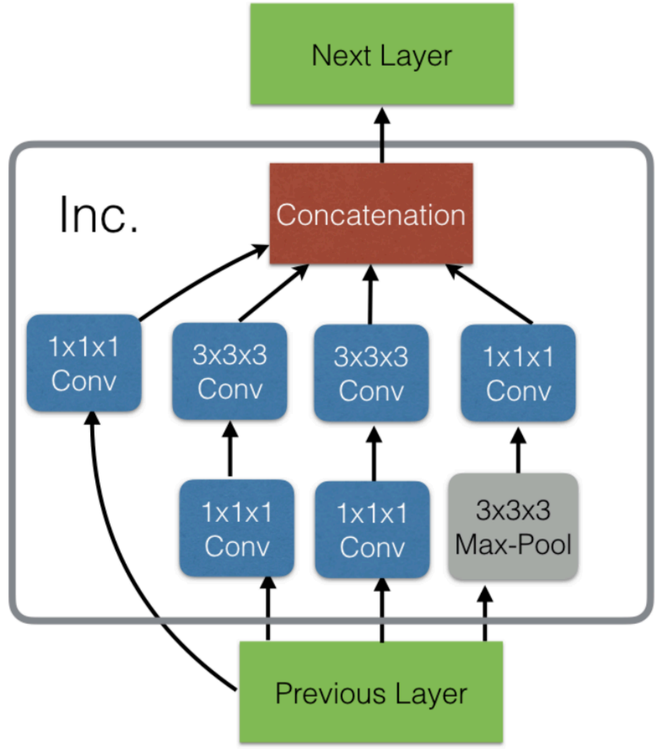
Any existing model can be pipelined

DenseNet



Densely connected convolutional neural networks, Huang et al., CVPR2017

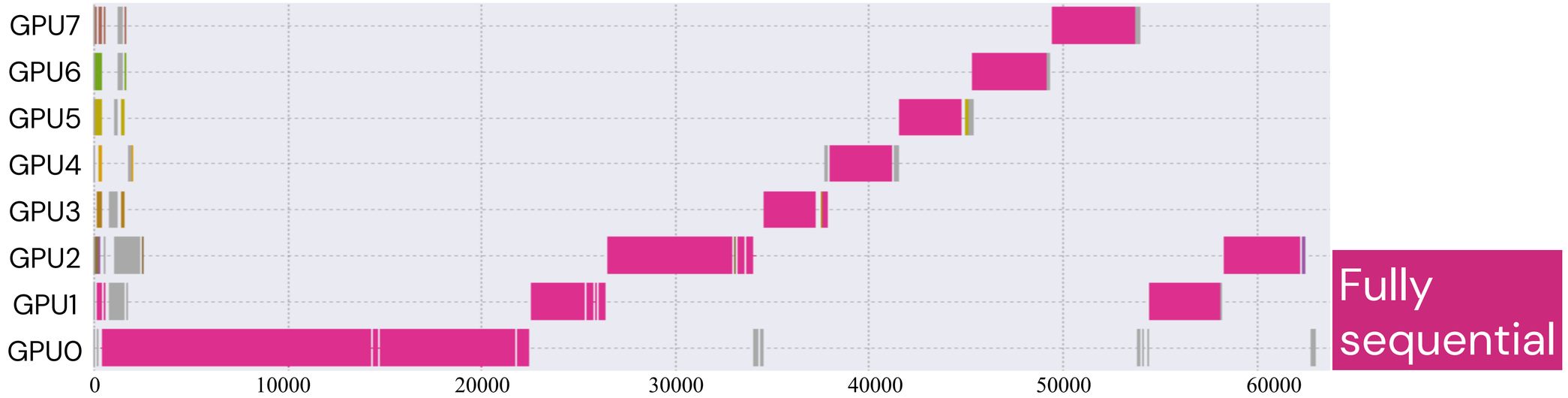
Inception

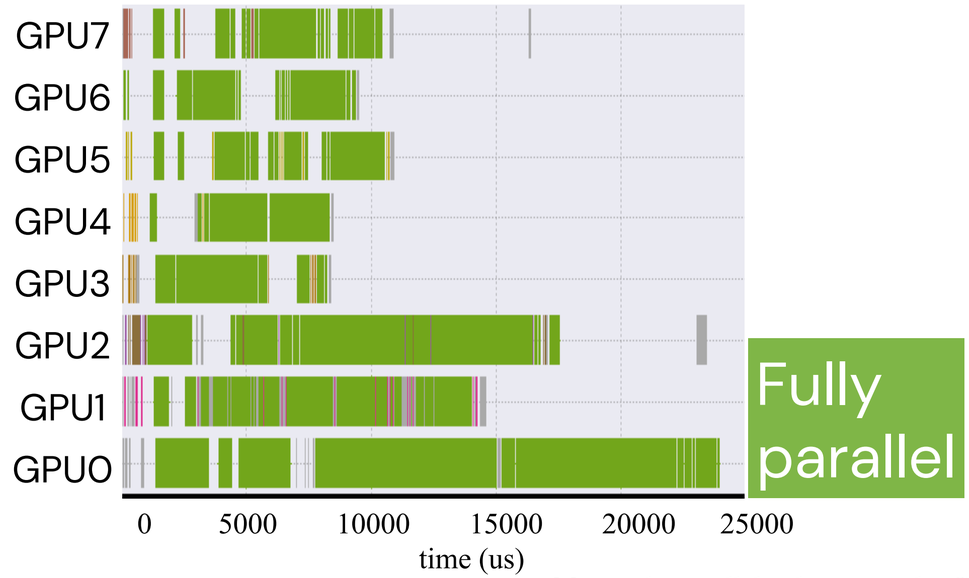


Quo Vadis, action recognition? Carreira and Zisserman, CVPR2017

Massively parallel video networks, Carreira, Patraucean, et al., ECCV2018

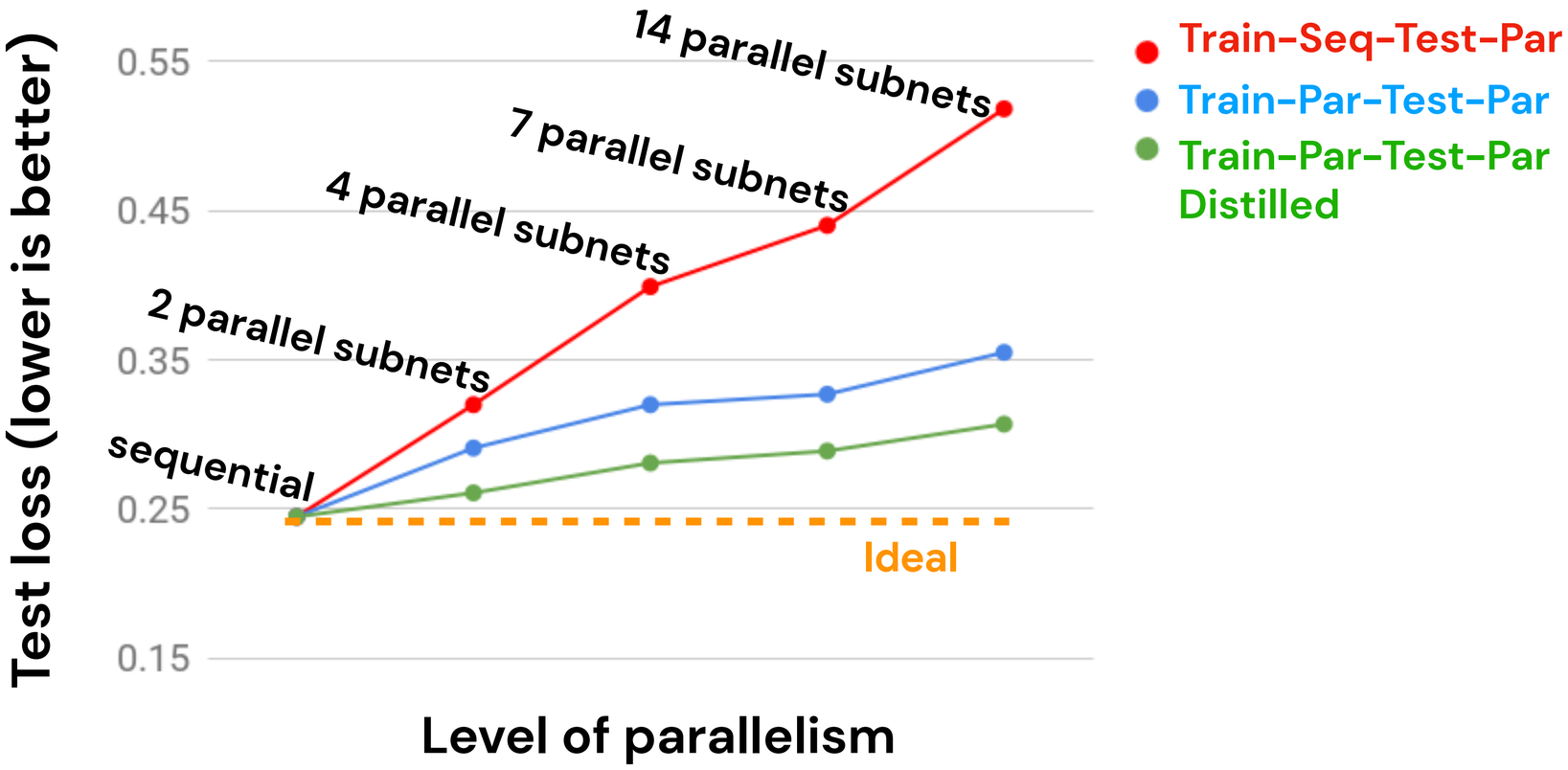
GPU utilisation





Massively parallel video networks, Carreira, Patraucean, et al., ECCV2018

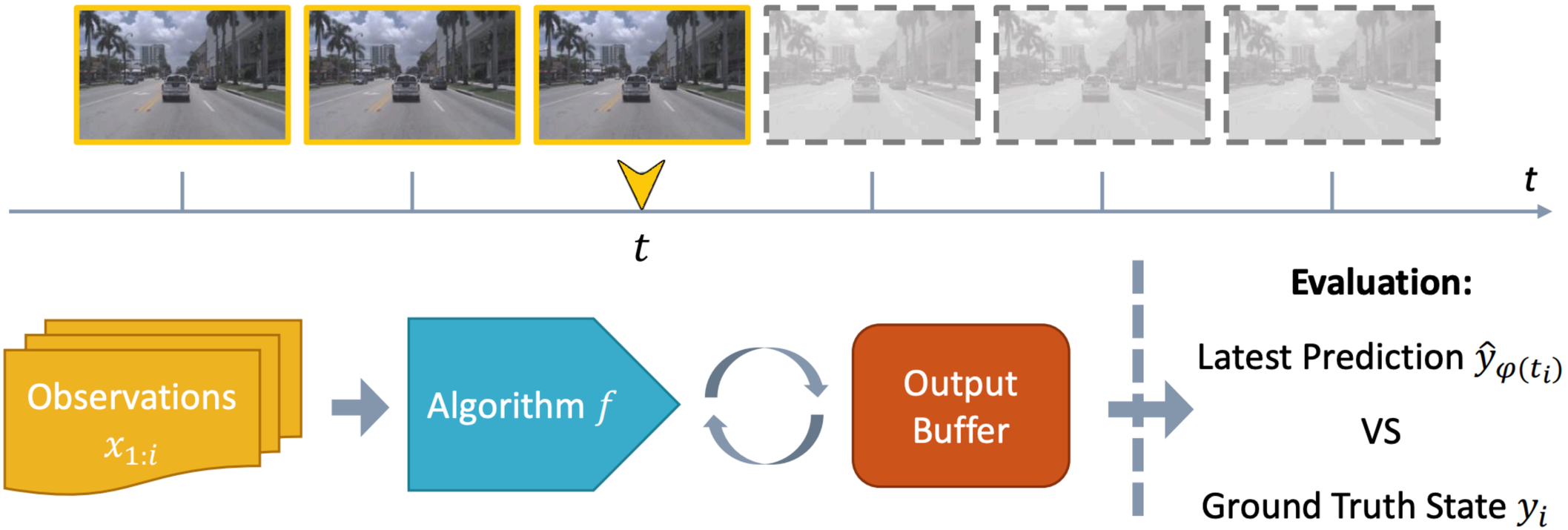
Human keypoint localisation



Massively parallel video networks, Carreira, Patraucean, et al., ECCV2018

Fair(er) asynchronous evaluation

Towards streaming image understanding, Li et al, ECCV2020

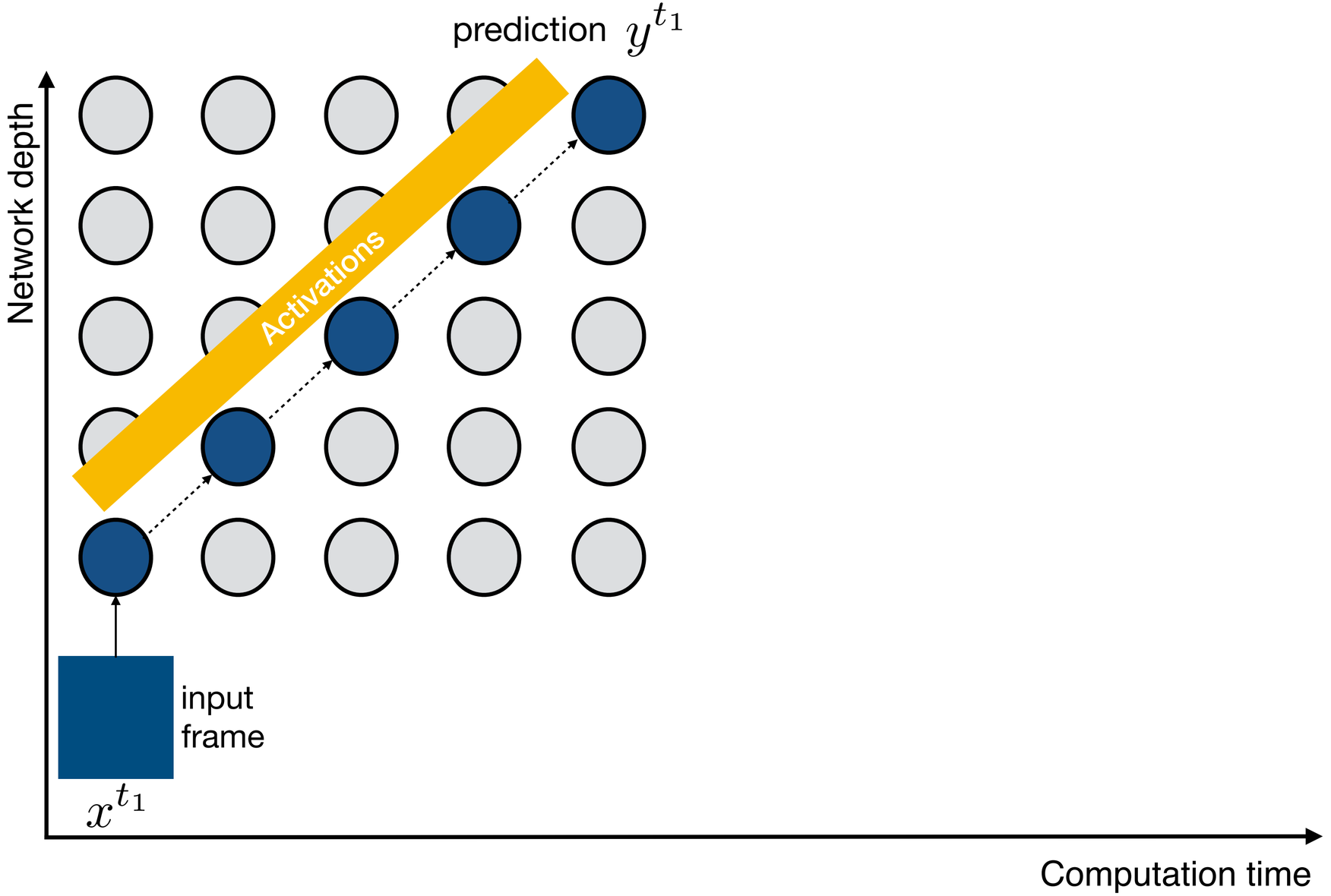


Object detector: offline precision 38.0, streaming average precision 20.3

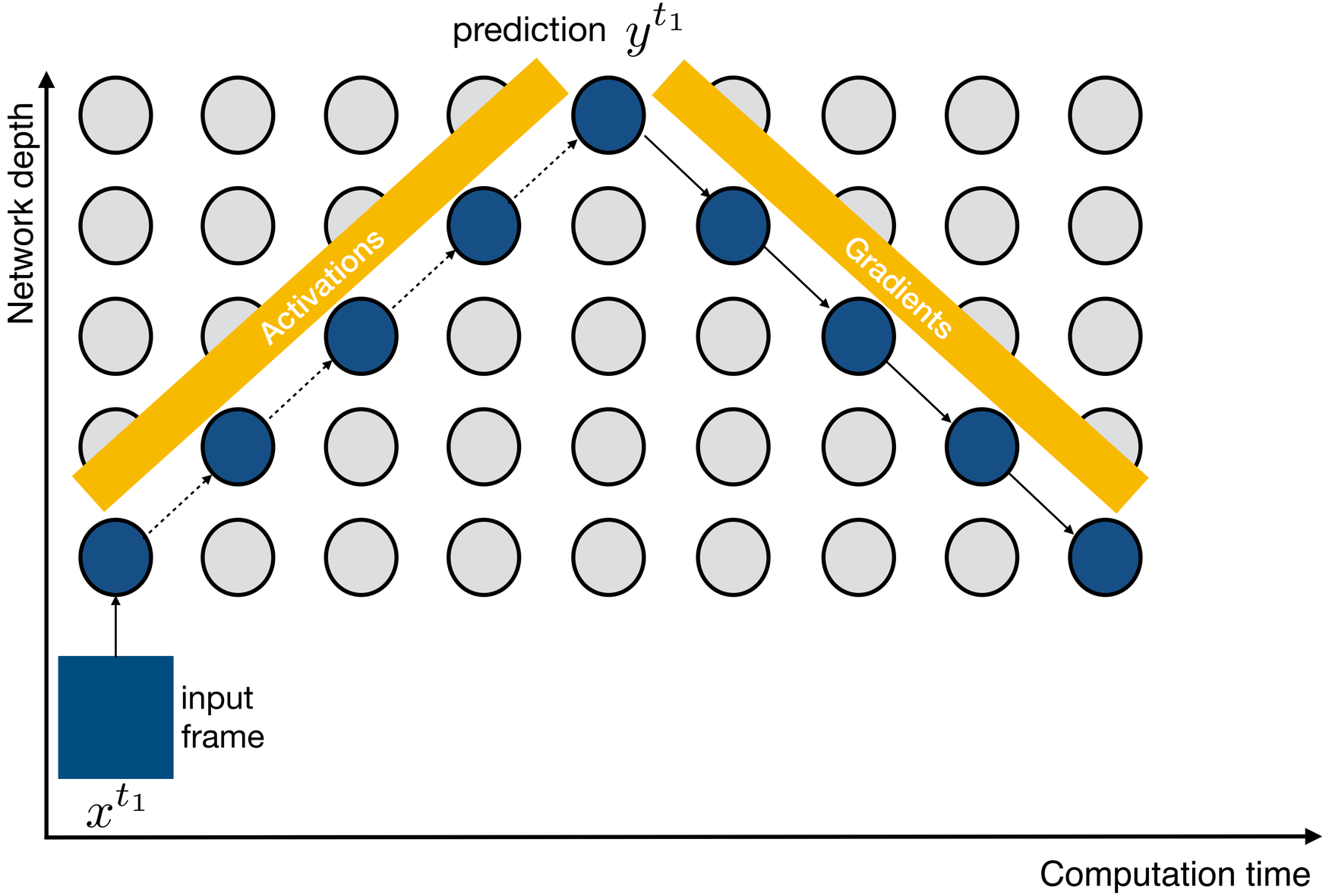
Figure from *Towards streaming image understanding,* Li et al, ECCV2020

B. Depth-parallel training

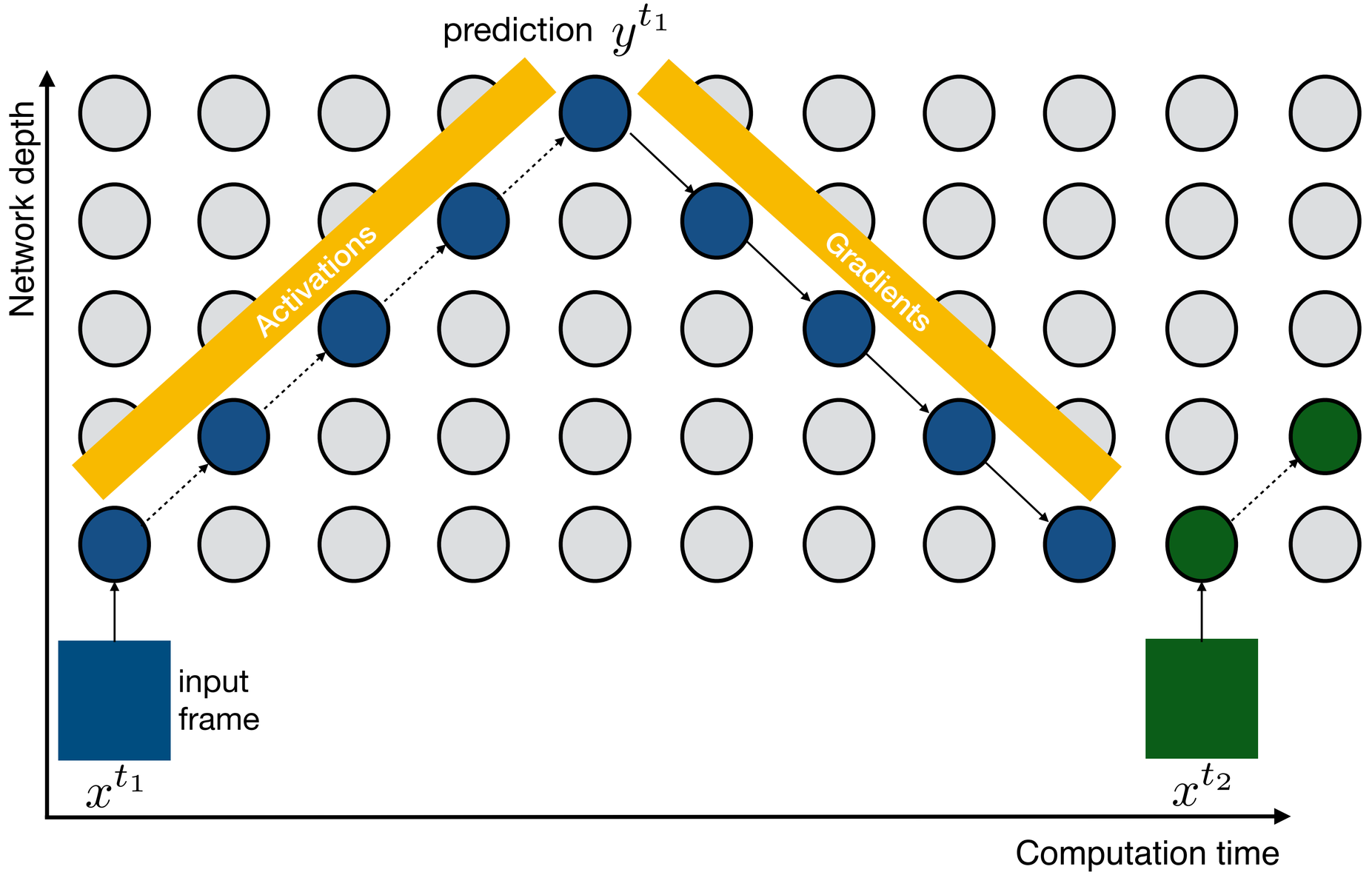
Sequential (frame-by-frame) model



Sequential (frame-by-frame) model

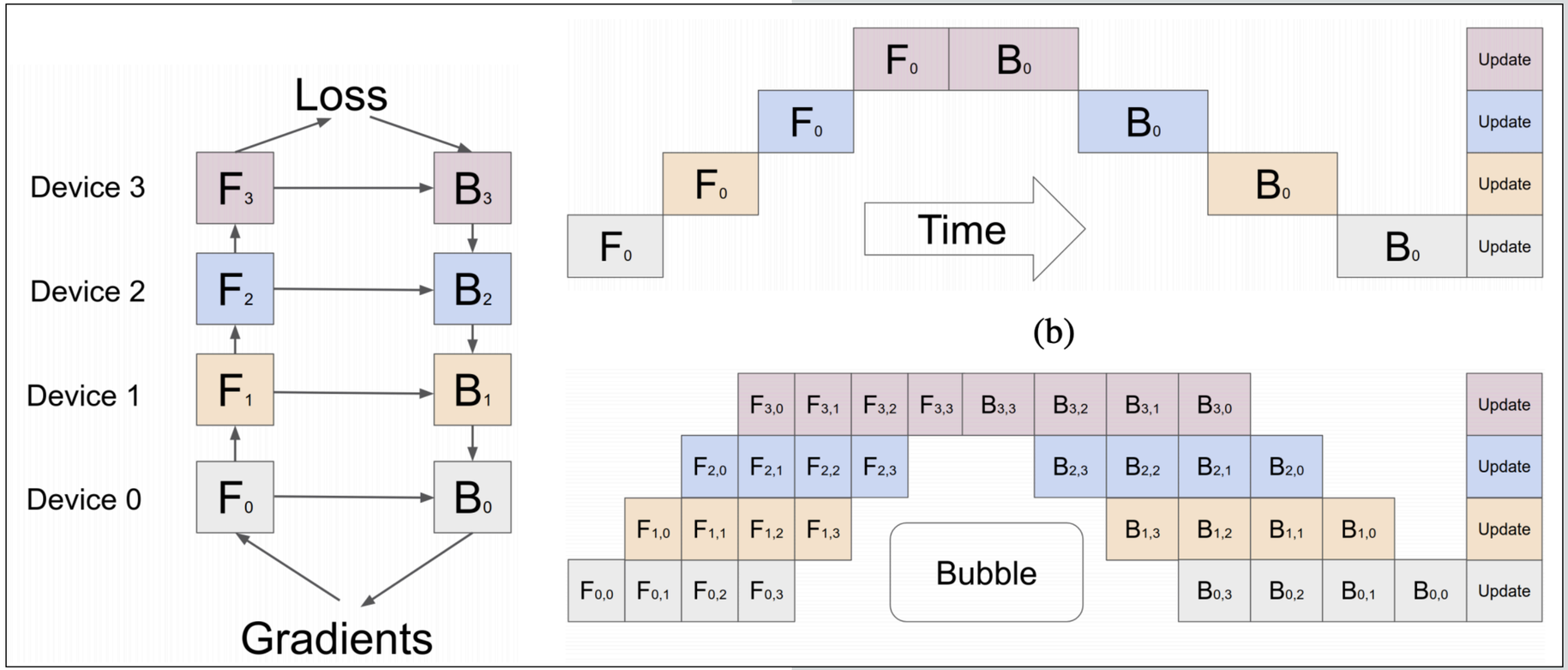


Sequential (frame-by-frame) model



Pipelined backpropagation: no blocking

GPipe: Easy Scaling with Micro-Batch Pipeline Parallelism, Yanping et al, 2019



Pipelined training of image classifiers by buffering activations

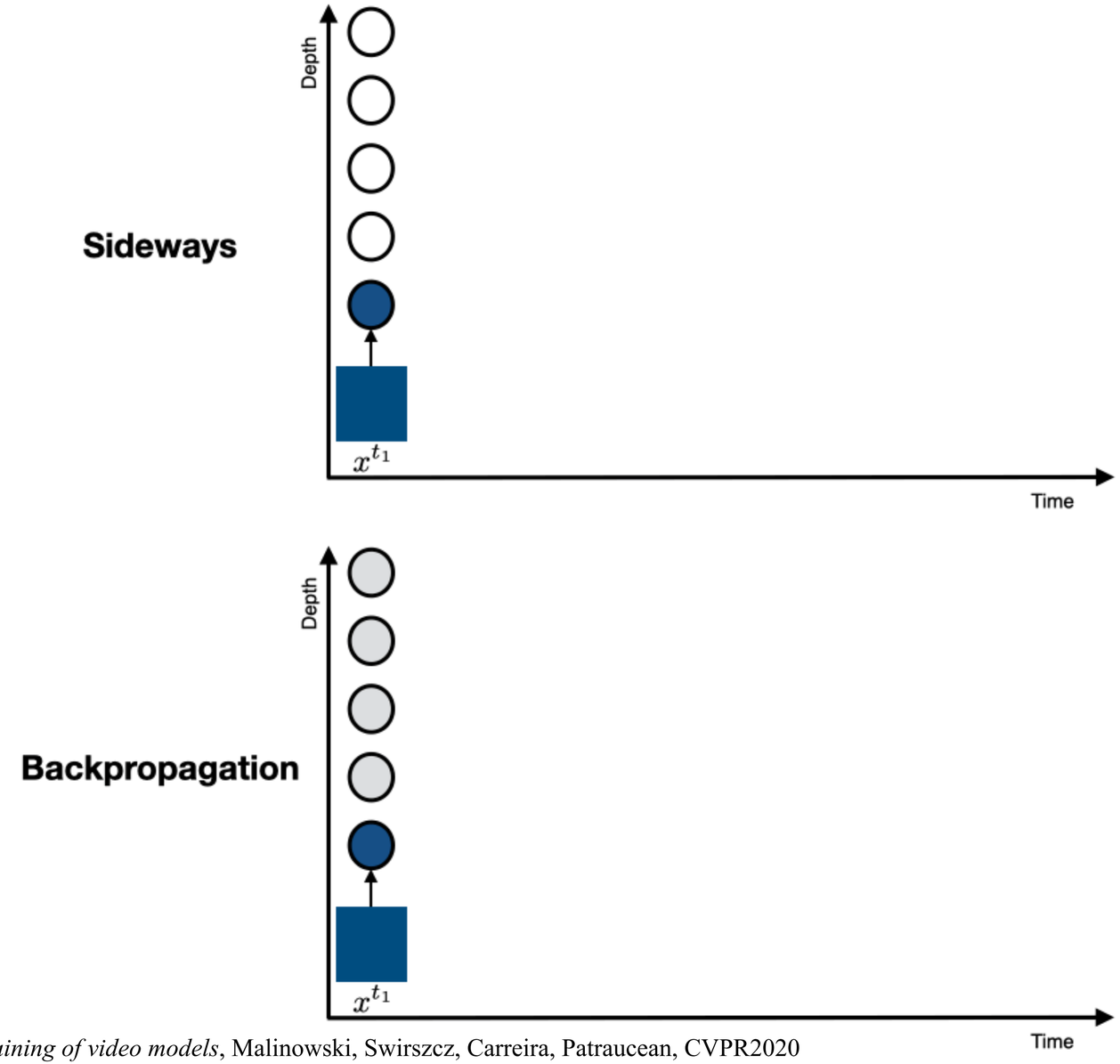
Figure from *GPipe: Easy Scaling with Micro-Batch Pipeline Parallelism*, Yanping et al, 2019

Videos are temporally smooth: do we need buffering?

Video from Kinetics600 dataset

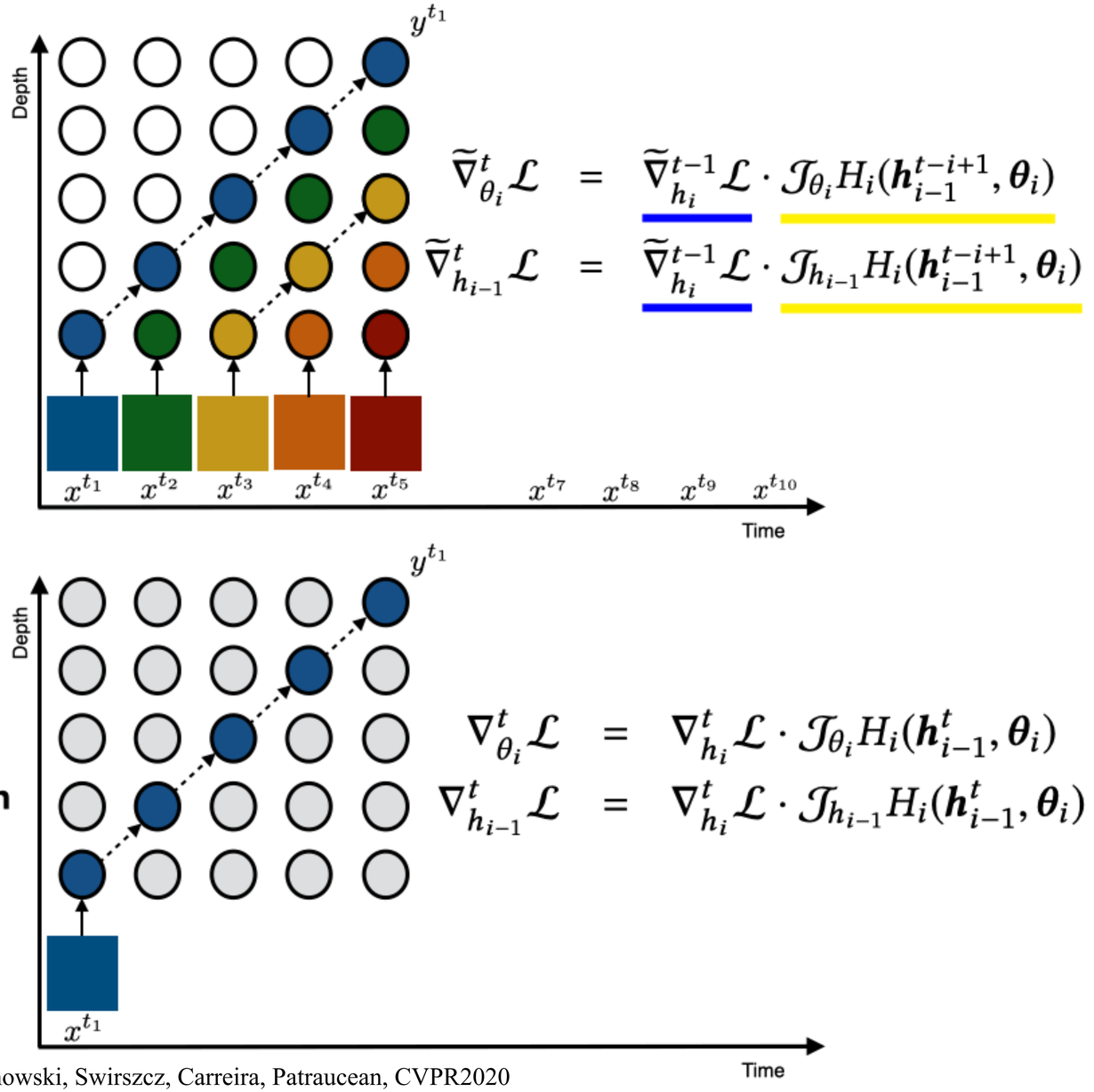
Sample consecutive frames extracted from the video

Sideways vs. Backprop



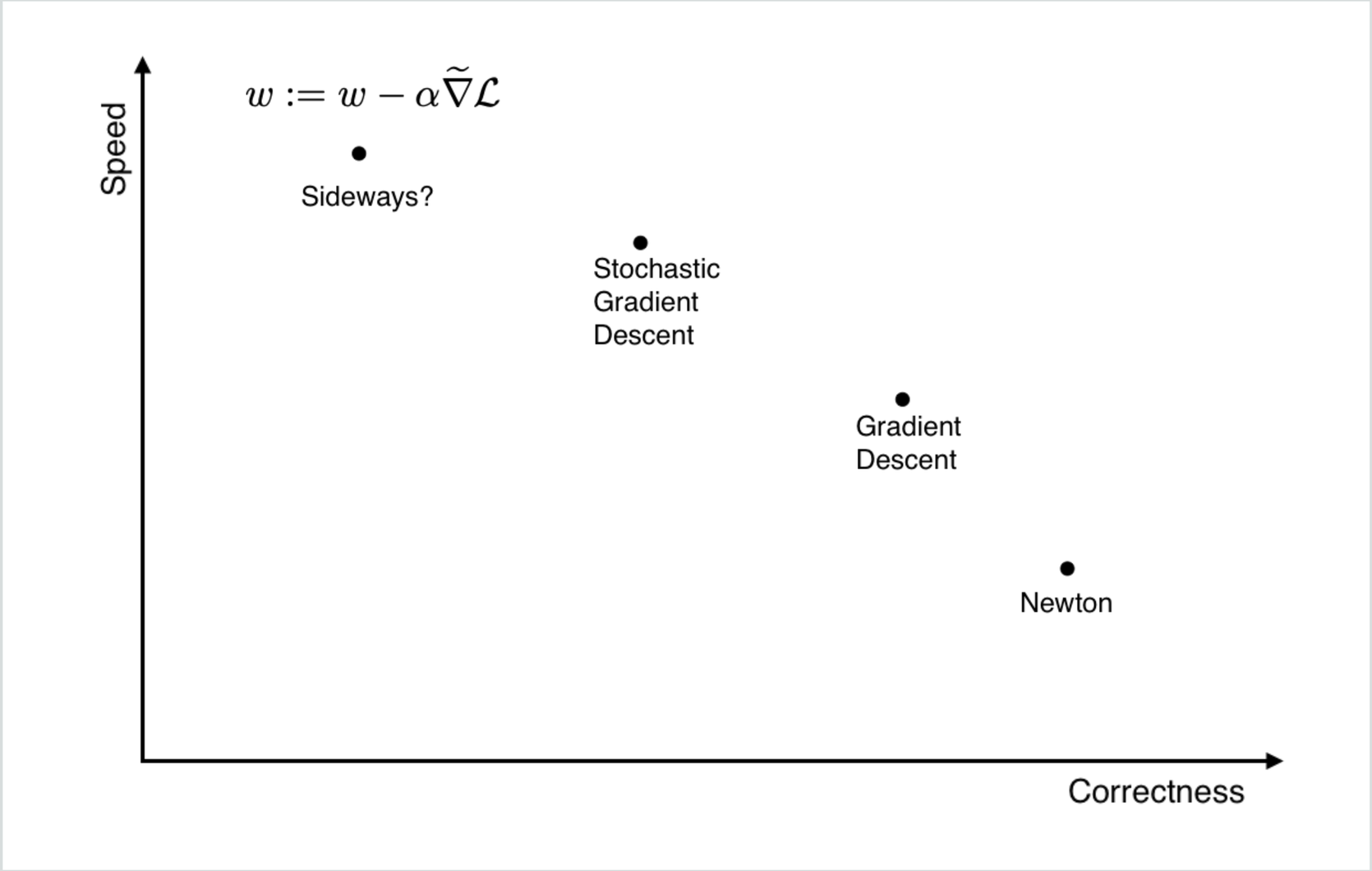
Sideways: Depth-parallel training of video models, Malinowski, Swirszcz, Carreira, Patraucean, CVPR2020

Sideways vs. Backprop



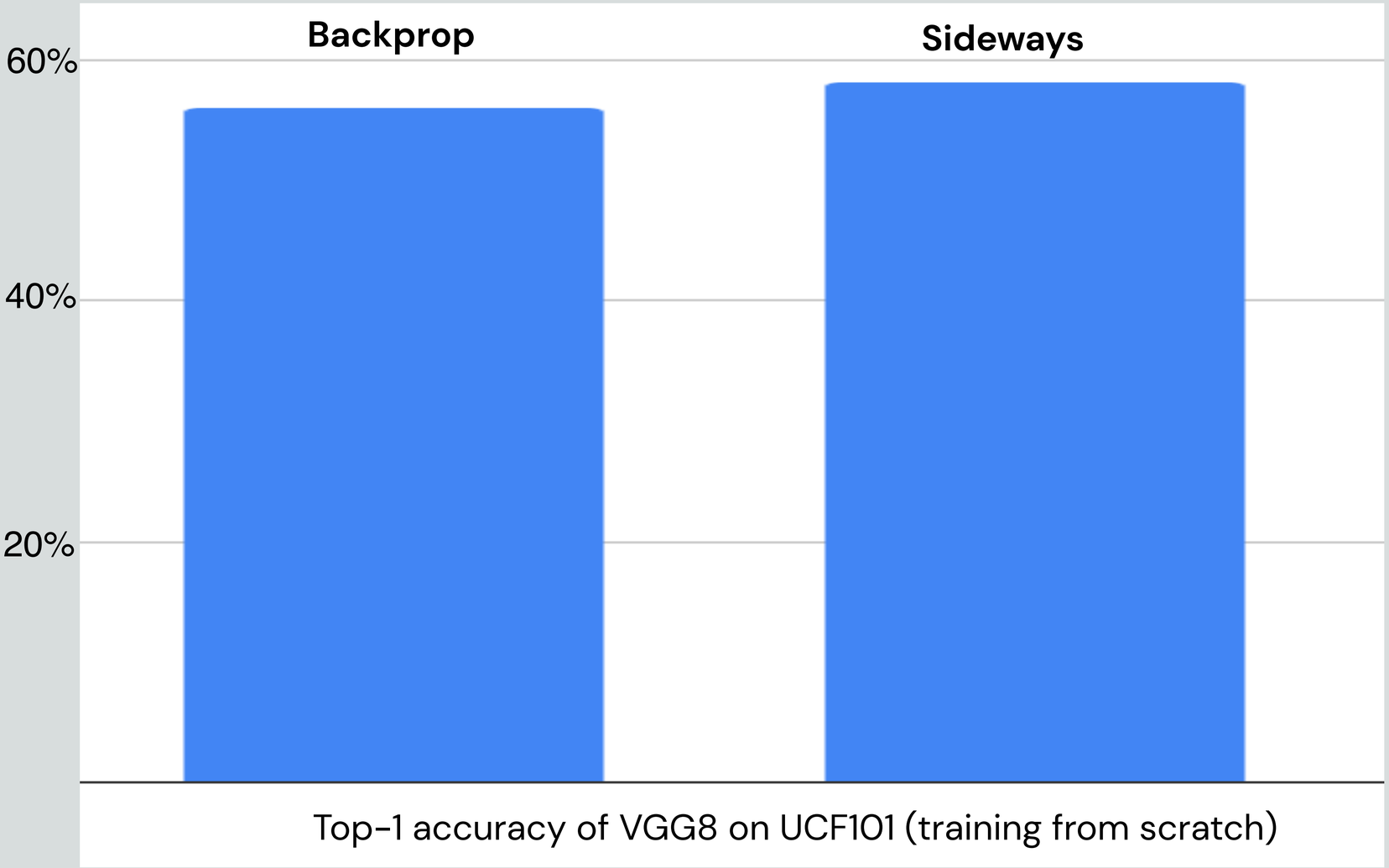
Sideways: Depth-parallel training of video models, Malinowski, Swirszcz, Carreira, Patraucean, CVPR2020

Mathematical correctness of gradients vs. speed



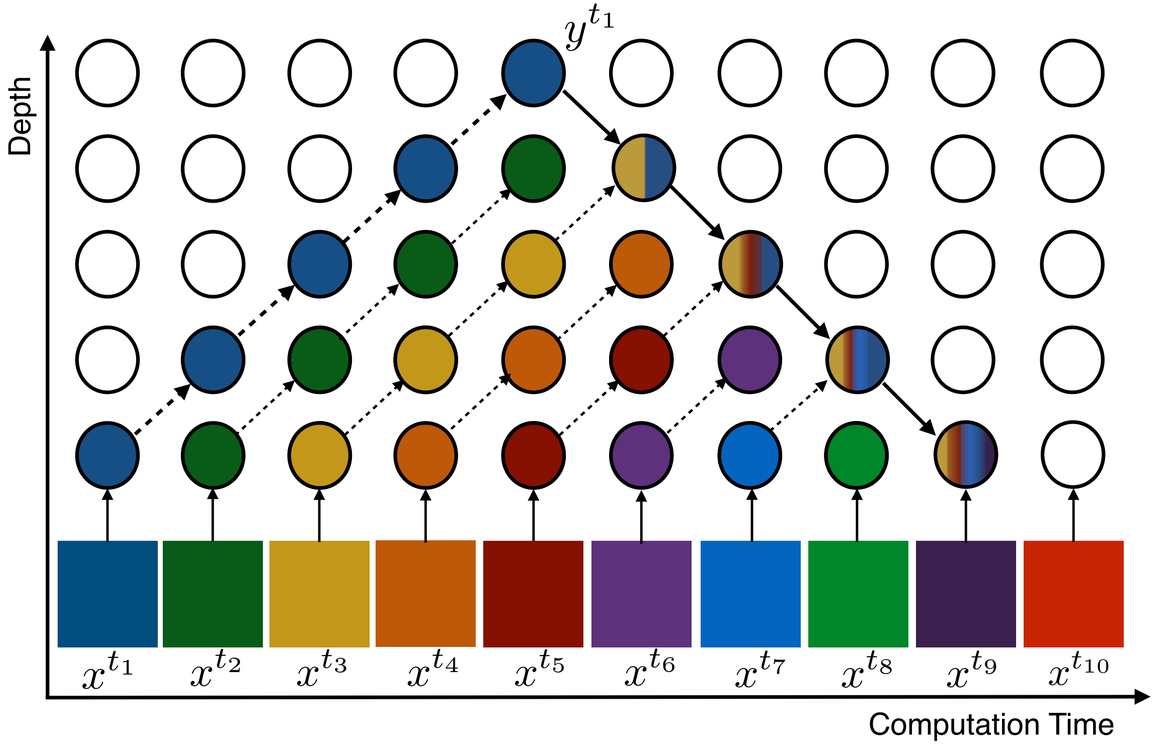
Sideways: Depth-parallel training of video models, Malinowski, Swirszcz, Carreira, Patraucean, CVPR2020

Sideways vs. Backprop

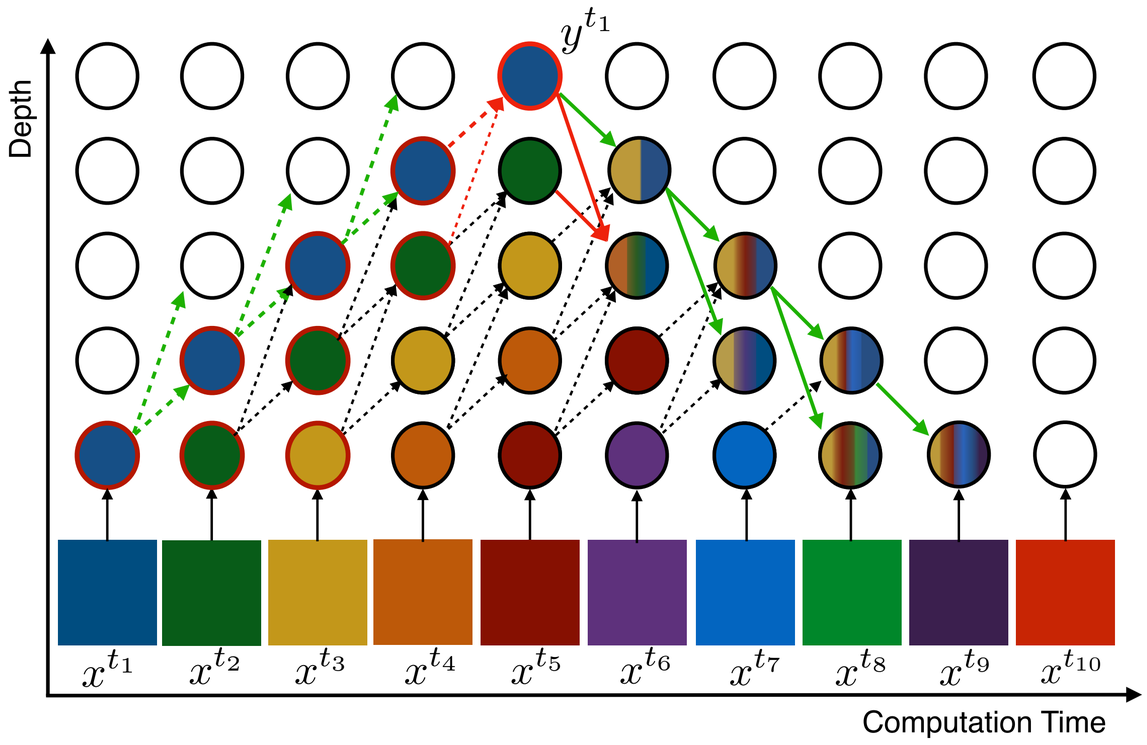


Sideways: Depth-parallel training of video models, Malinowski, Swirszcz, Carreira, Patraucean, CVPR2020

Skip-Sideways: 2D *temporal* model



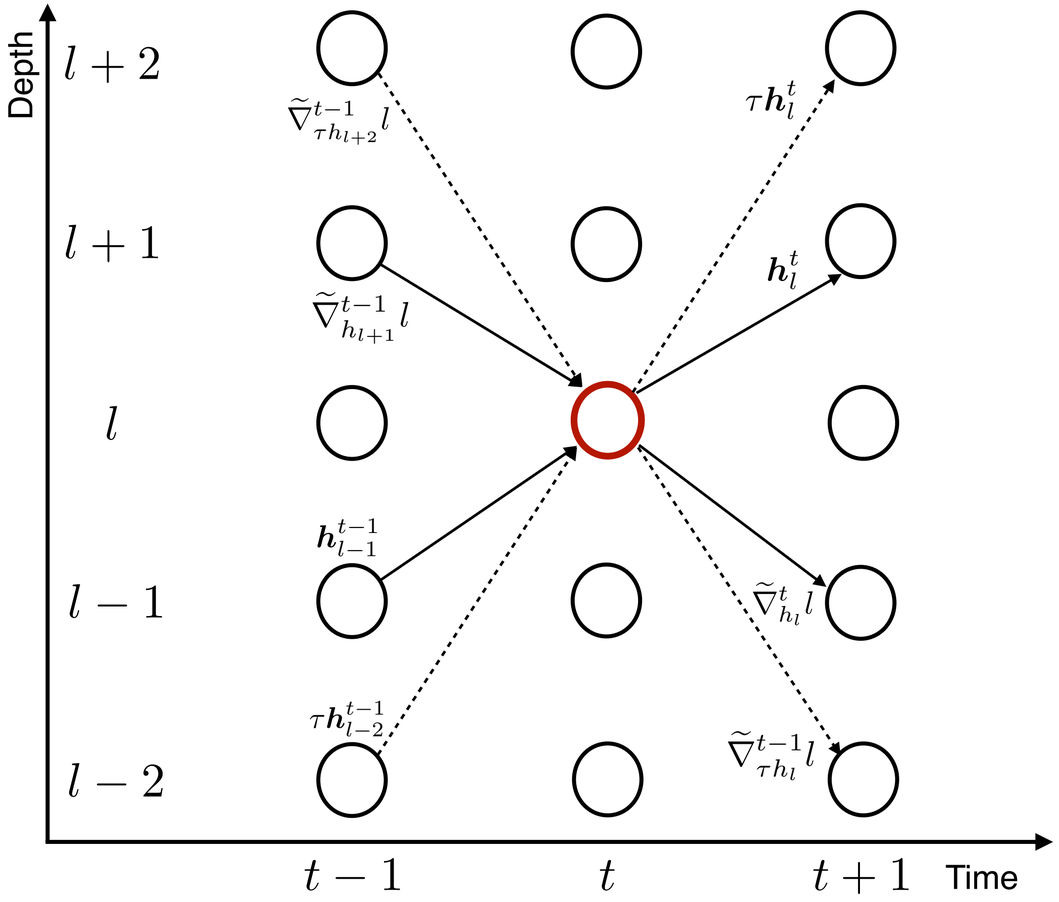
Sideways

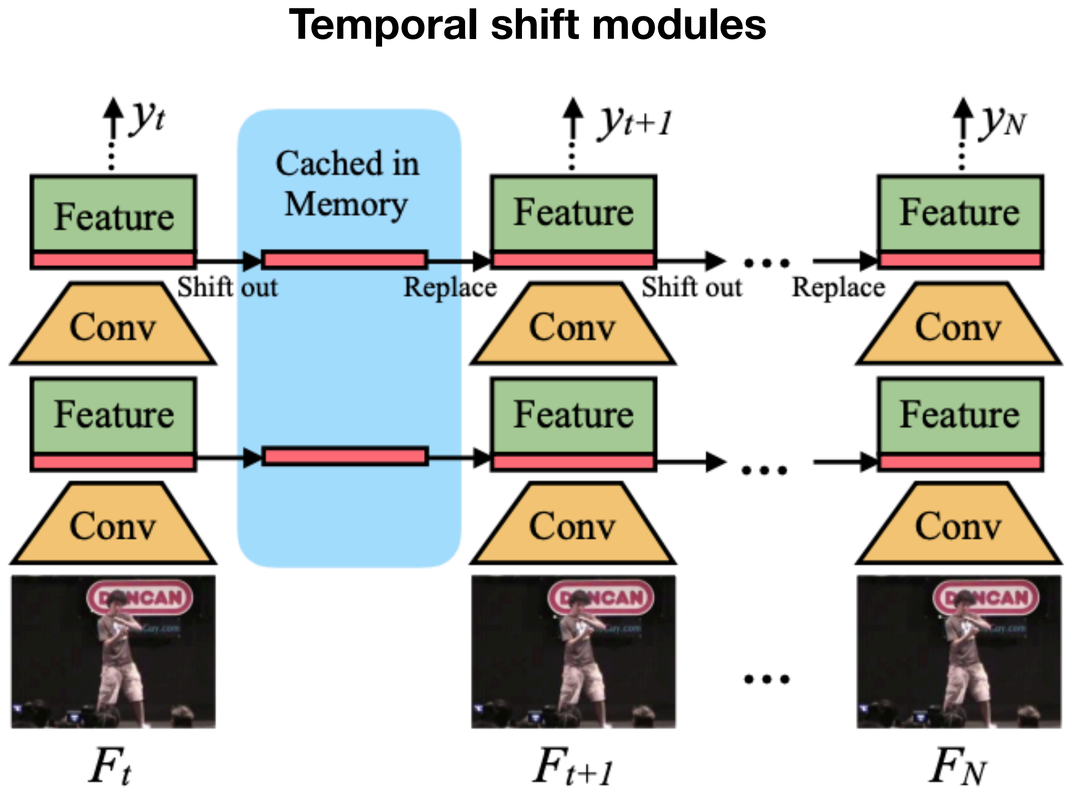


Skip-Sideways

Gradient forward propagation for large-scale temporal video modelling, Malinowski et al, CVPR2021

Skip-Sideways vs. Temporal shift modules (TSM)

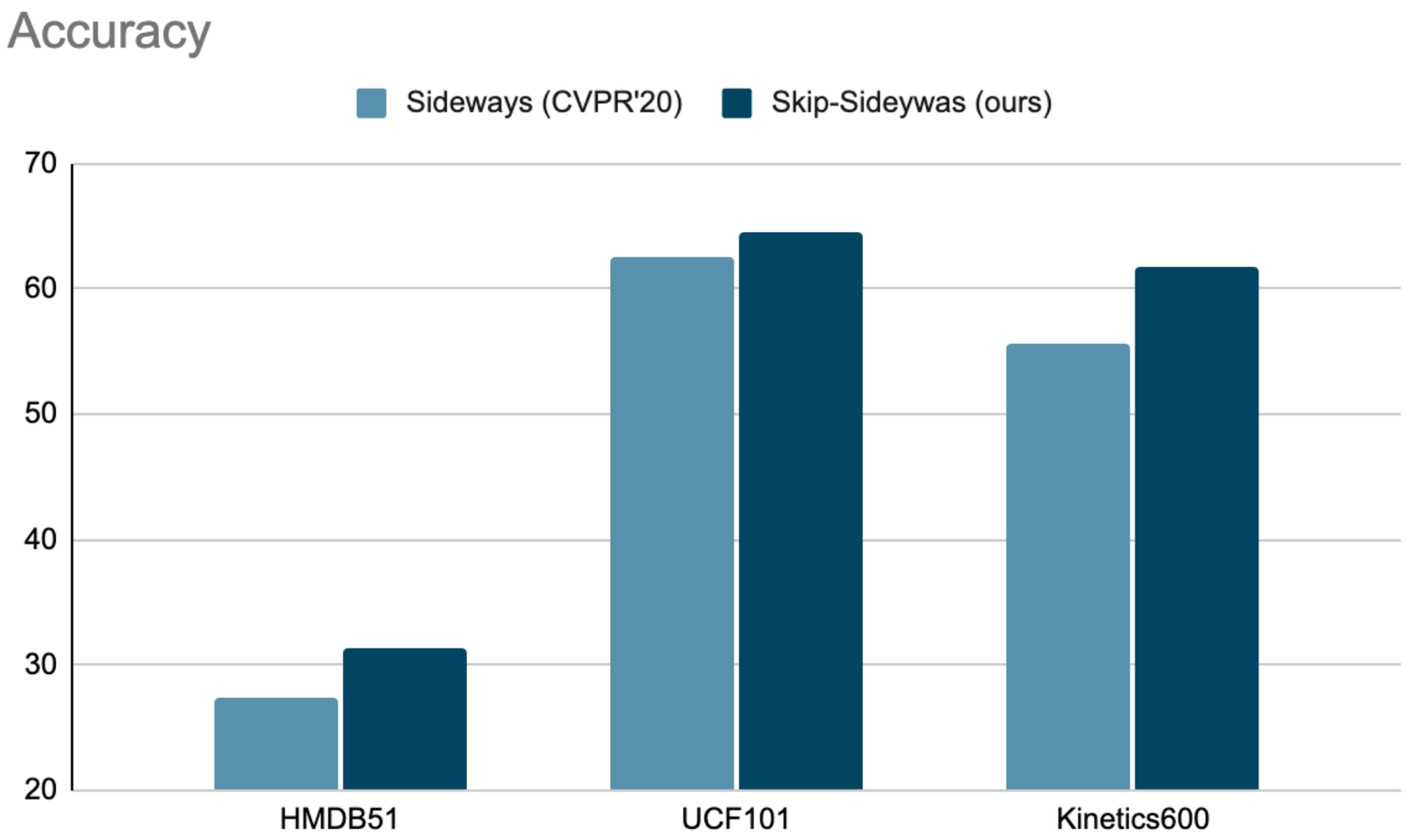




TSM: Temporal Shift Module for Efficient Video Understanding
Lin et al, 2019

Gradient forward propagation for large-scale temporal video modelling, Malinowski et al, CVPR2021

Skip-Sideways: 2D *temporal* model



Gradient forward propagation for large-scale temporal video modelling, Malinowski et al, CVPR2021

(Skip-)Sideways vs. GPipe vs. Backprop

	Inputs	Uses	Blocking	Requires buffering	Correct gradients
BP	any	chain rule	yes	yes	yes
GPipe	any	pipelining of operations	partial	yes	yes
(Skip-)Sideways	**temporally smooth sequences**	pipelining and smoothness of inputs	no	no	approximate

Biological (im)plausibility of backprop

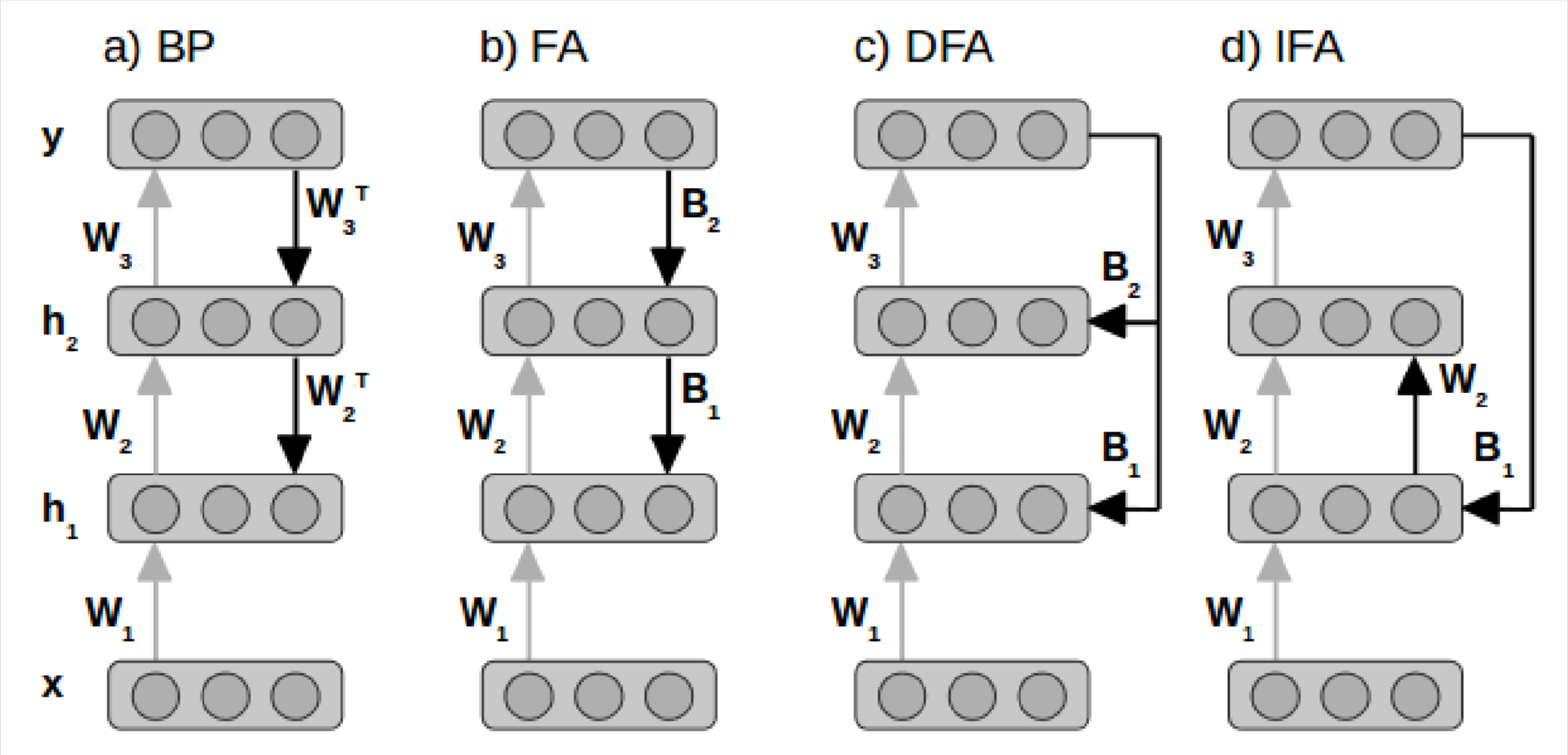


Figure from *Direct Feedback Alignment Provides Learning in Deep Neural Networks,* Nokland, 2016

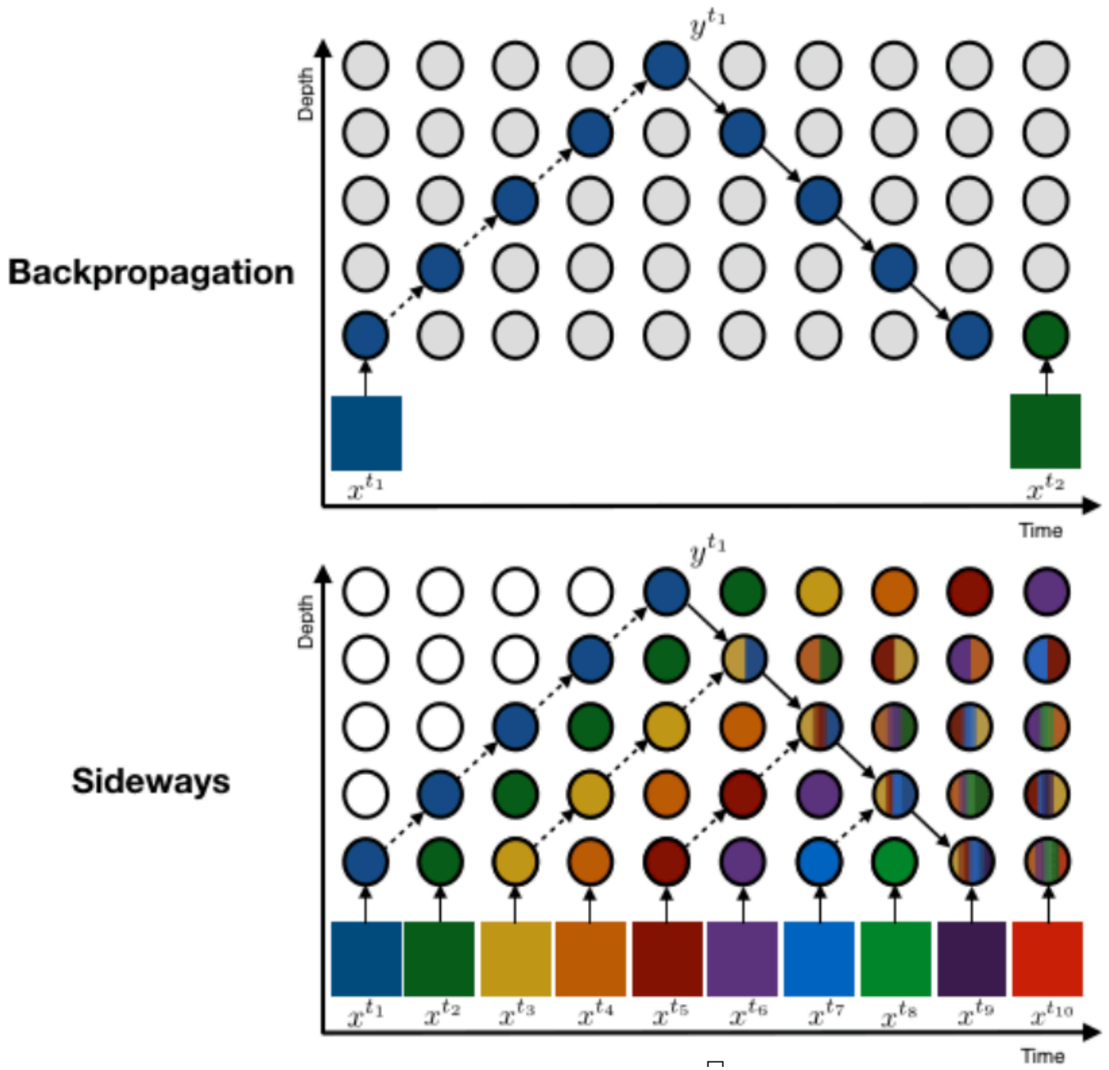
Sideways: more biologically plausible

BP: instantaneous computation (blocking)

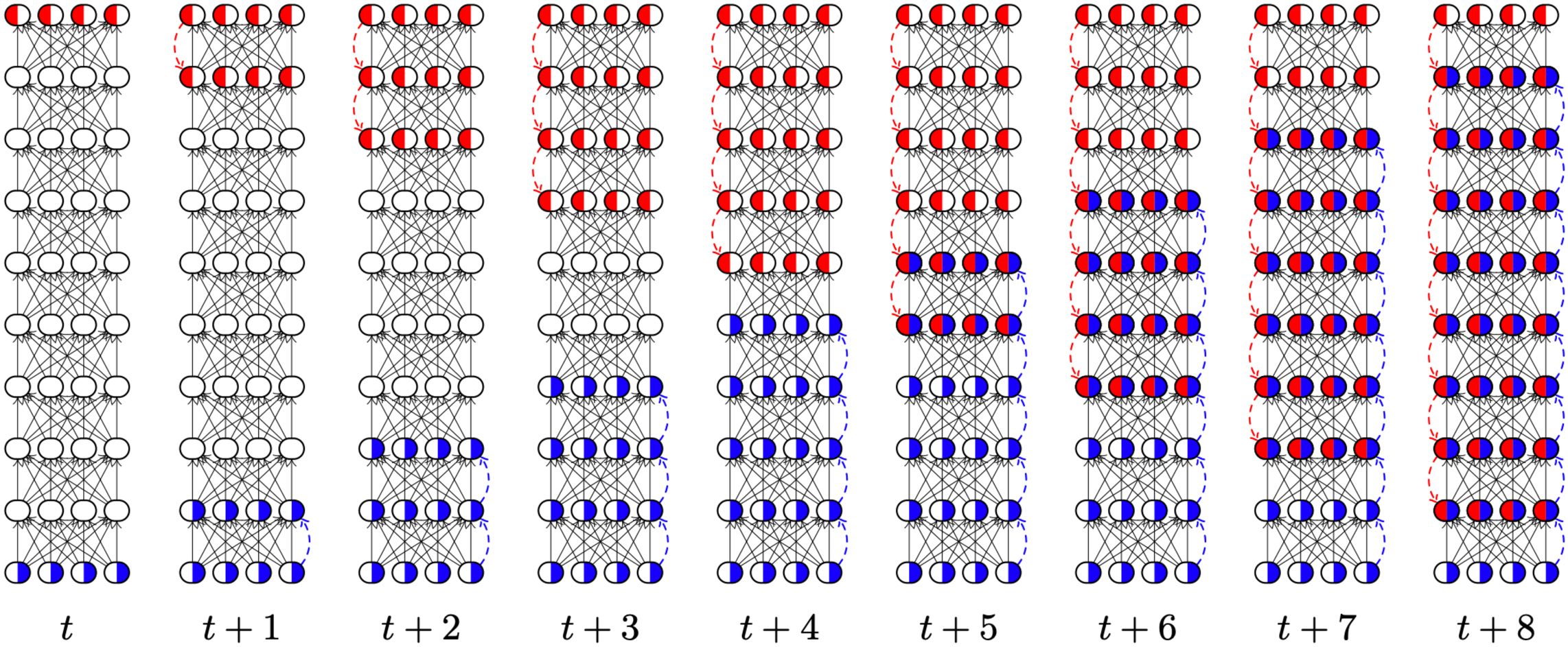
BPTT: sends gradients back in time

RTRL: sends gradients forward, but computationally intractable

Sideways: sends *approximate* gradients forward



Backprop as diffusion



Backprop Diffusion is Biologically Plausible, Betti and Gori, 2020, hal-02878574

04

Conclusion

Summary and Future work

- Design principles for efficient training and inference through
  - self-supervision
  - pipelining
- Exploit the temporal smoothness of videos to speed up training
- Latency and throughput are not dependent on depth: we can increase depth (and accuracy) without affecting the latency and throughput
- **Next steps**: theoretical framework for convergence with approximate gradients
- **Next steps**: Skip-Sideways + self-supervision -> local learning rule

Roadmap/curriculum to solve vision

LeCun's cake

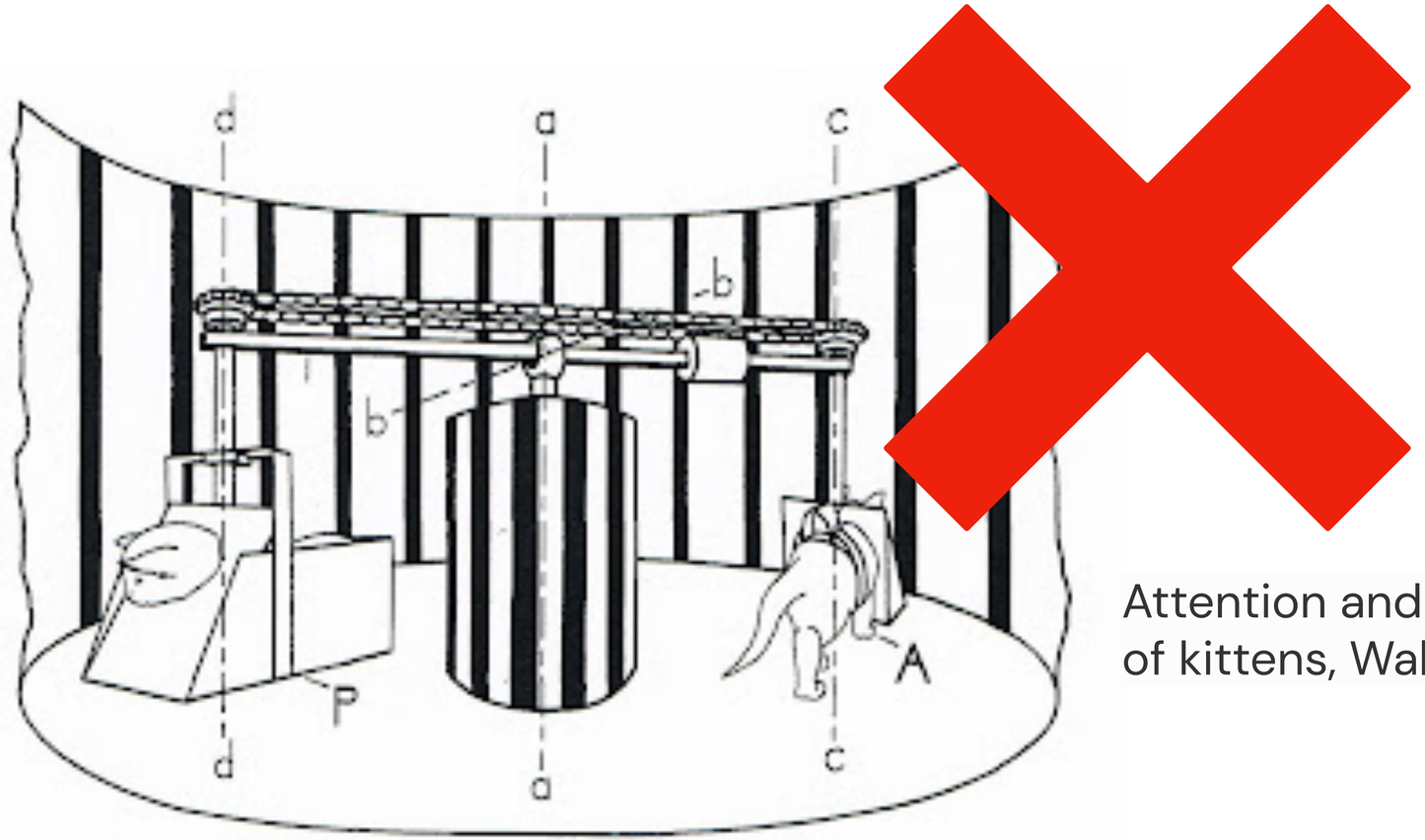
Self-supervised learning

Self-supervised multimodal learning

Active learning Interaction and intervention

Rough model of the visual observations → Grounding across senses (common-sense) → Causal perceptual model of the world

Learning vision requires active locomotion



Attention and the depth perception of kittens, Walk et al, 1988

[Held, R. and Hein A. (1963). Movement-produced stimulation in the development of visually guided behavior. Jouranal of Comparative and Physiological Psychology 56(5): 872-876.]

Final advice

Question everything! Don’t take things for granted!