Deep video models

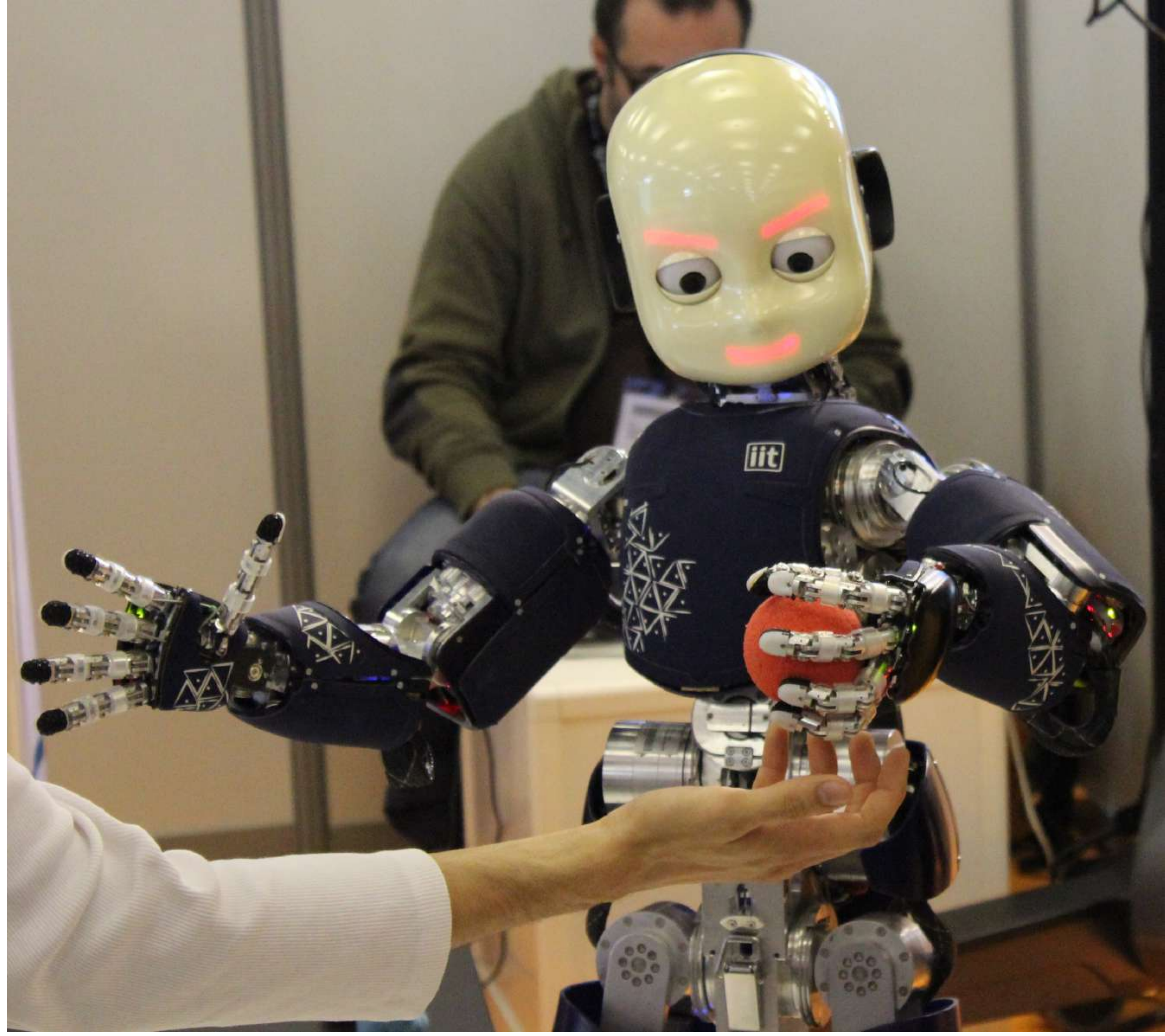
Viorica Pătrăucean, Research scientist

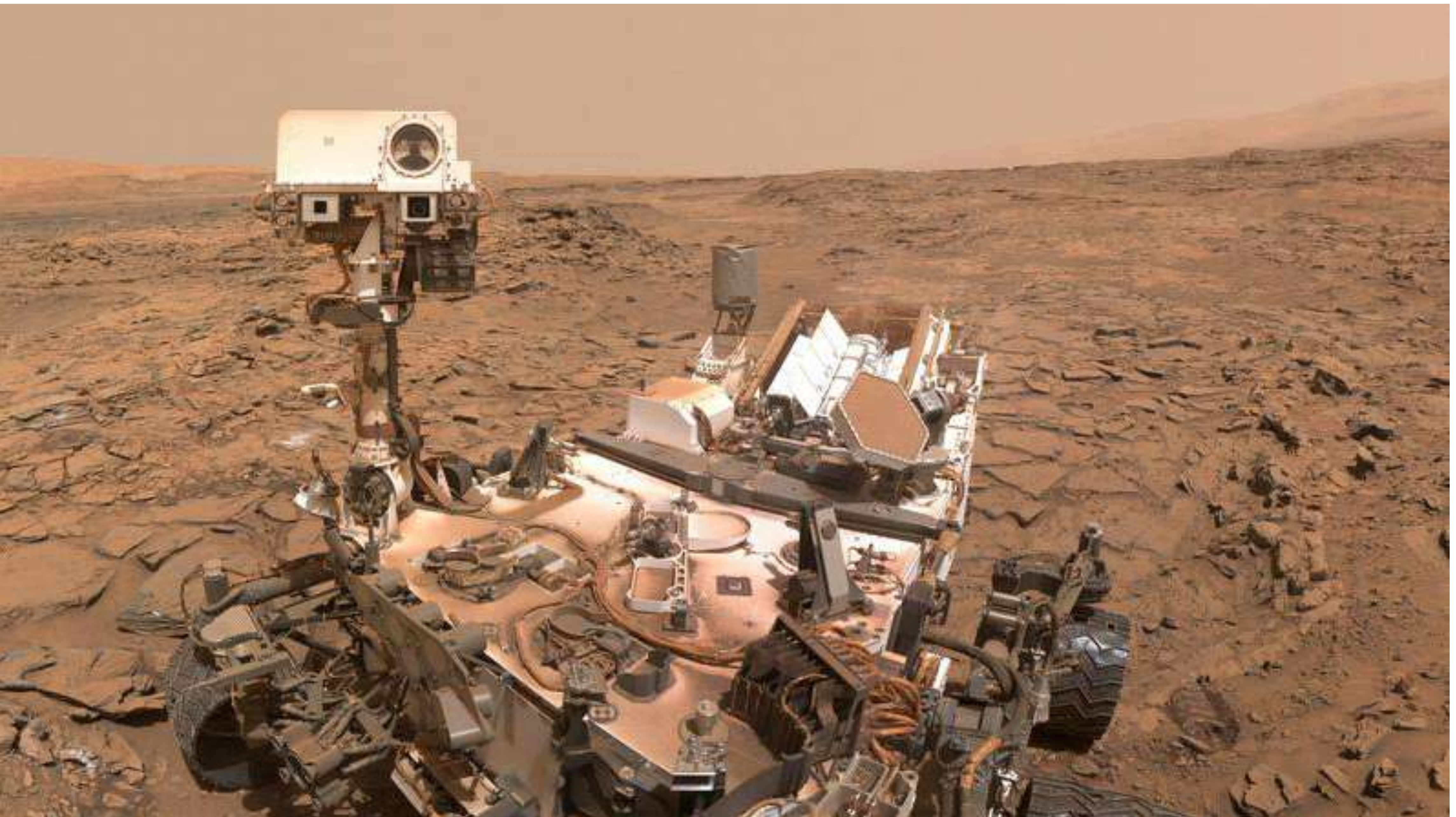
DeepMind

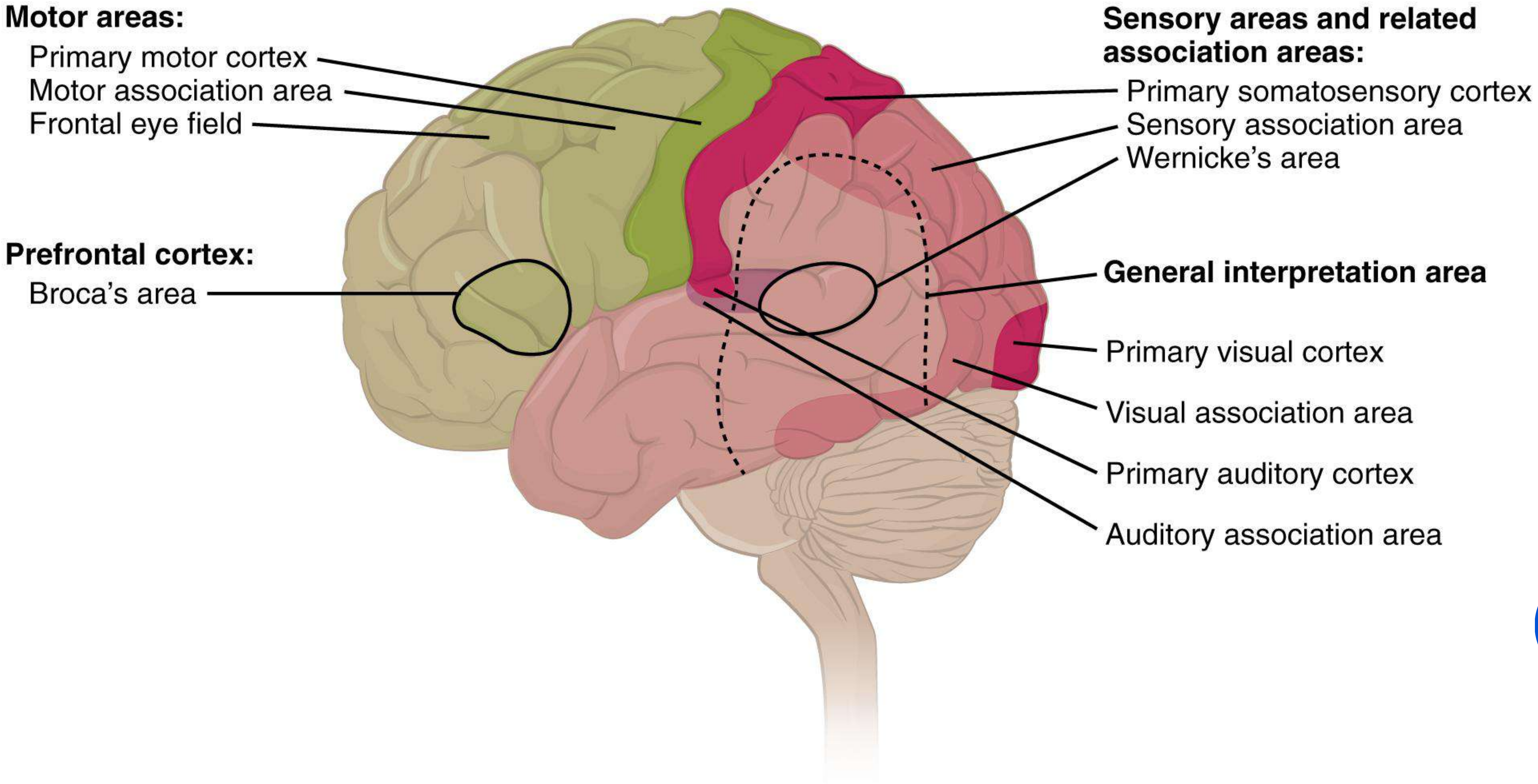
ORASIS 2021

About myself

- Undergrad in Bucharest Romania — Computer Engineering
- Master and PhD in Toulouse France — Multimedia & Image processing
- PostDoc in Paris and Cambridge — Research on 3D shapes and videos
- Research scientist in DeepMind since 2016


iit




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

Deep video models

- Videos: high dimensional data
- Video model:
  - Difficult to generalise
  - Data-hungry

Deep video models design

Generalisation

Data efficiency

Expert labelling

Real-time constraints

Outline

01

Inductive biases

Video architectures

02

Task design

Multimodal learning

Self-supervised learning

03

Operation mode

Inference and Training

Biological plausibility

04

Conclusion

Summary

Future work

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

01 Architecture inductive biases

Inductive biases for video modelling

Assumptions used by a learner to improve generalisation (Battaglia et al. 2018)

Design choices that constrain the solution space:

- **architecture** side for data efficiency

- **loss function** to compensate for lack of labels

- **training procedure** for real-time operation

Want to learn more?

Battaglia et al. **Relational inductive biases, deep learning, and graph networks (2018)**

Inductive biases on the architecture side

Assumptions used by a learner to improve generalisation (Battaglia et al. 2018)

A video is a sequence of images - use spatial inductive biases

Locality and spatial invariance

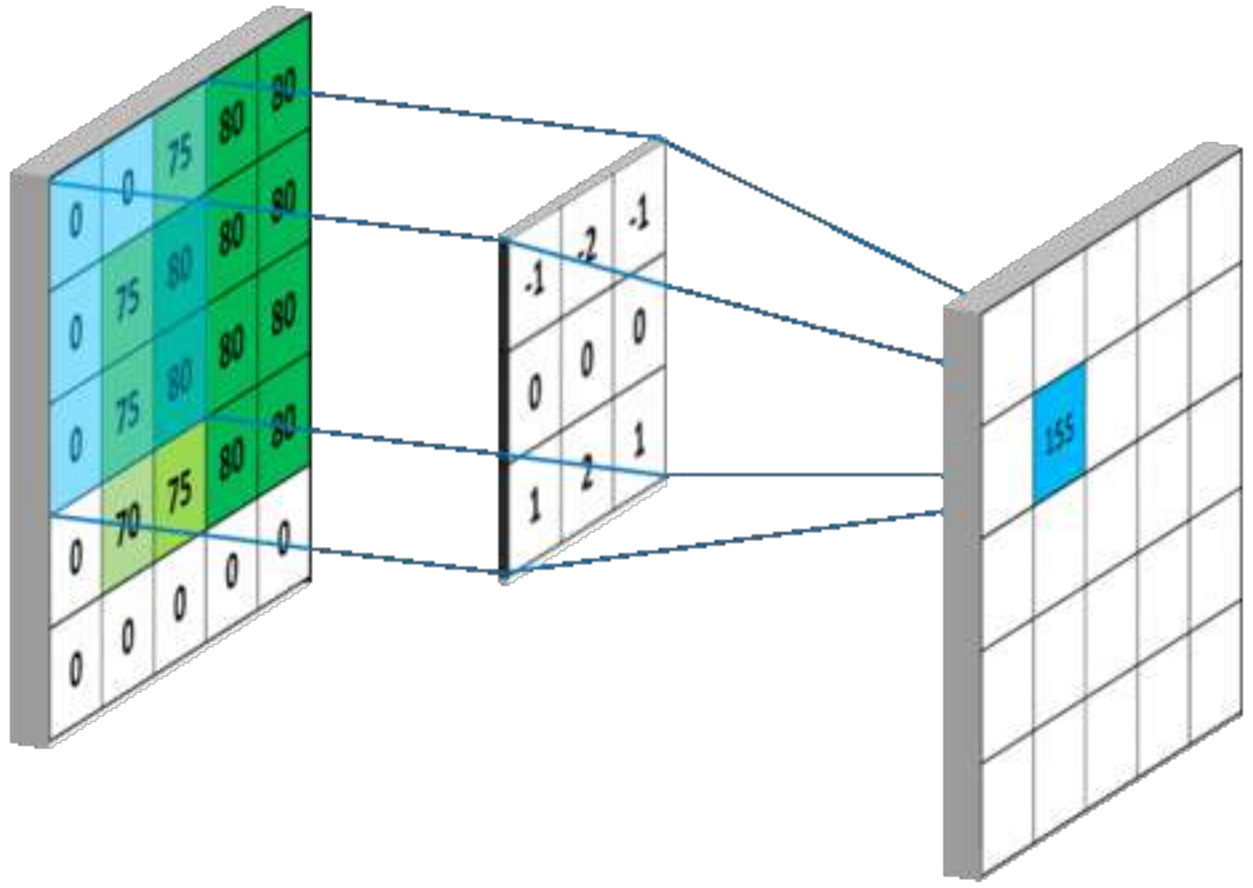
Hierarchical structure

Inductive biases on the architecture side

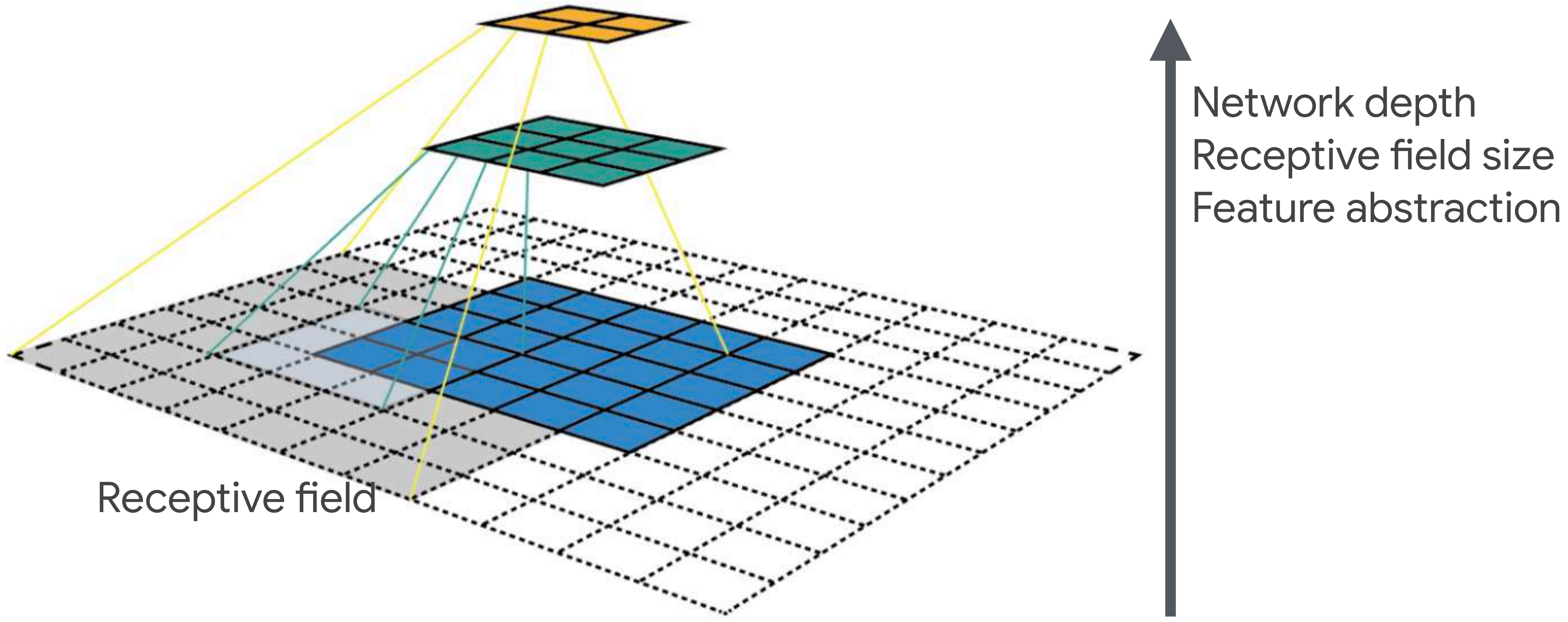
Assumptions used by a learner to improve generalisation (Battaglia et al. 2018)

A video is a sequence of images - use spatial inductive biases

ConvNets incorporate well such biases

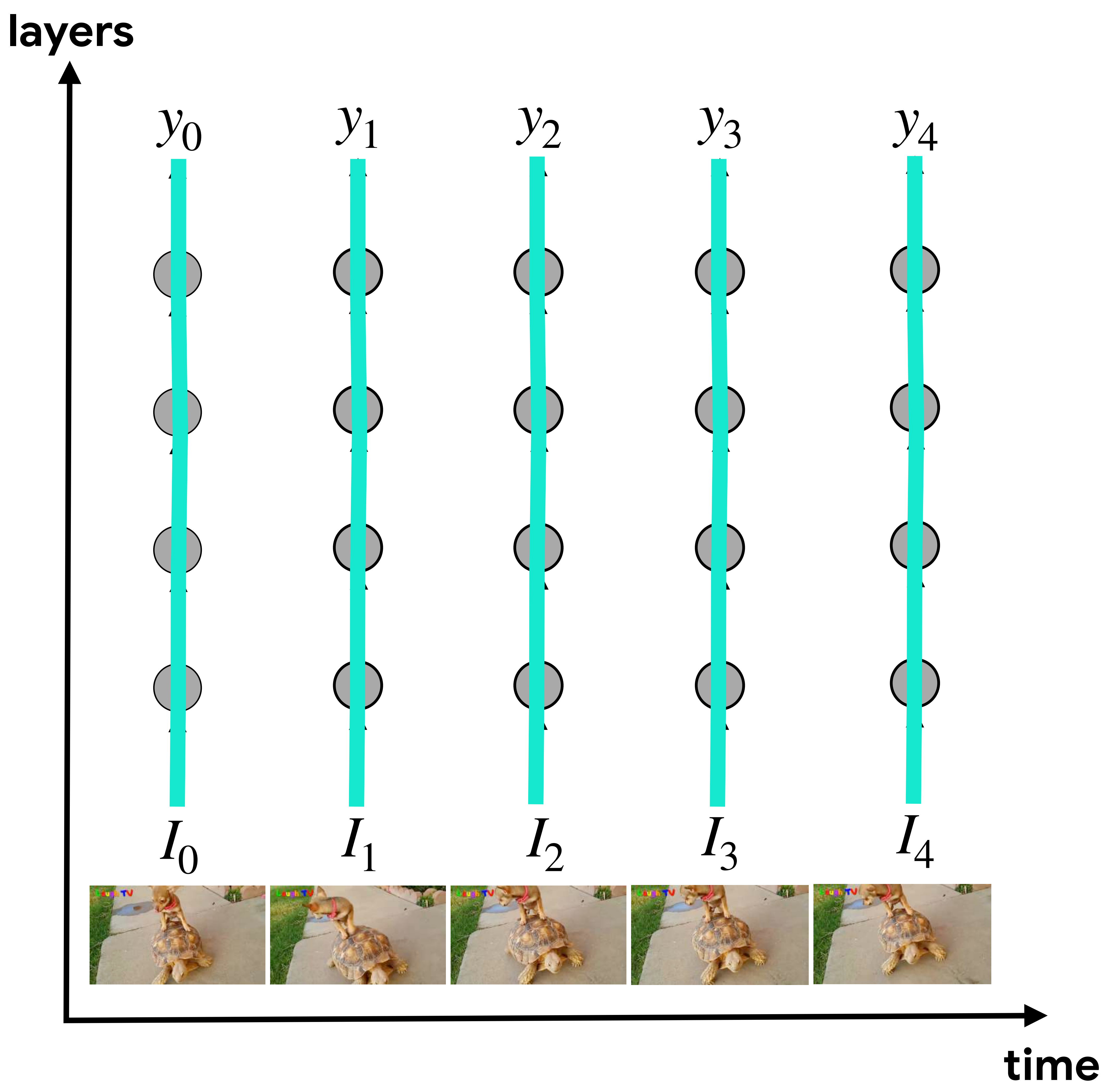


Local filters and weights sharing



Hierarchical processing

Frame-by-frame video model

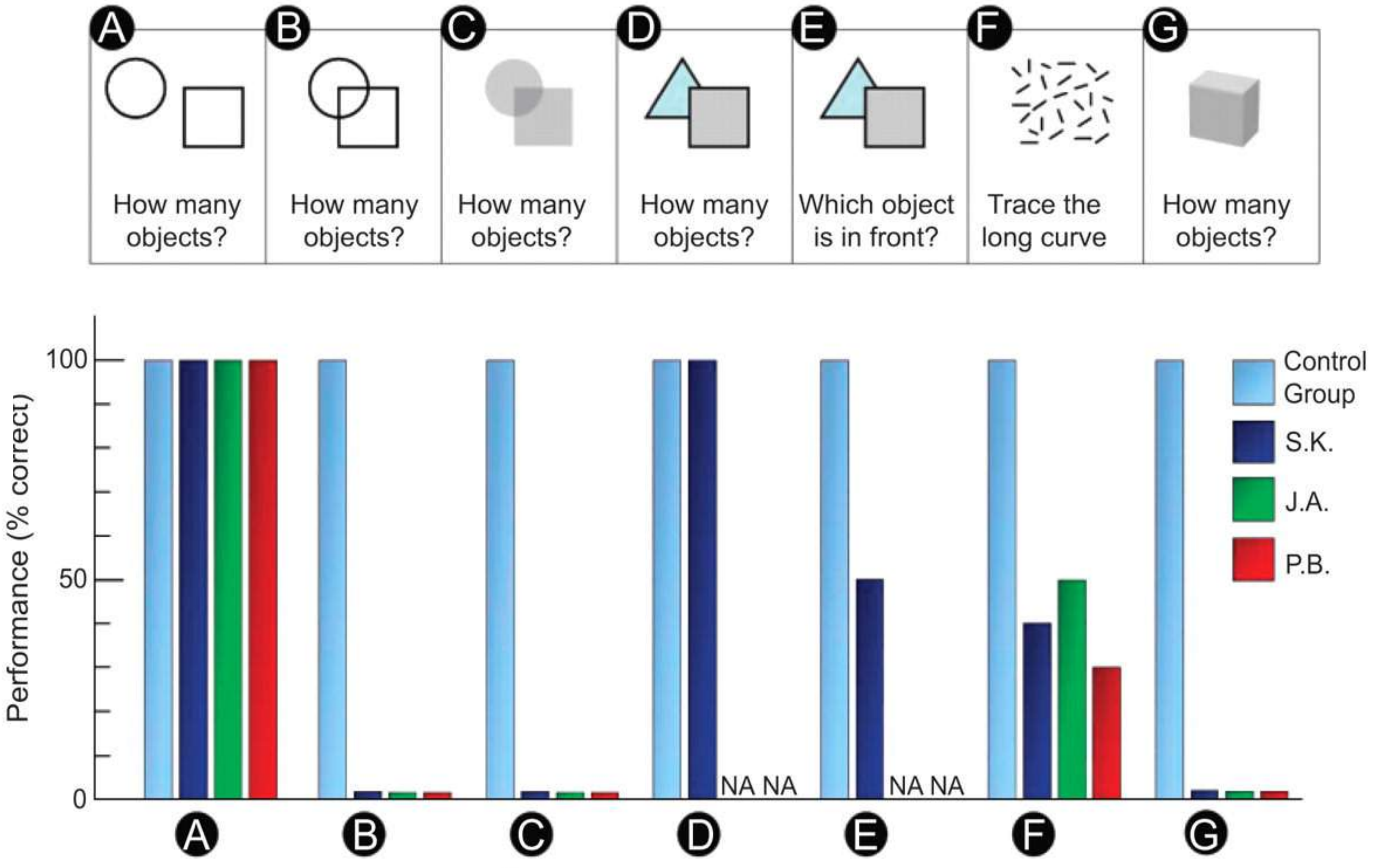


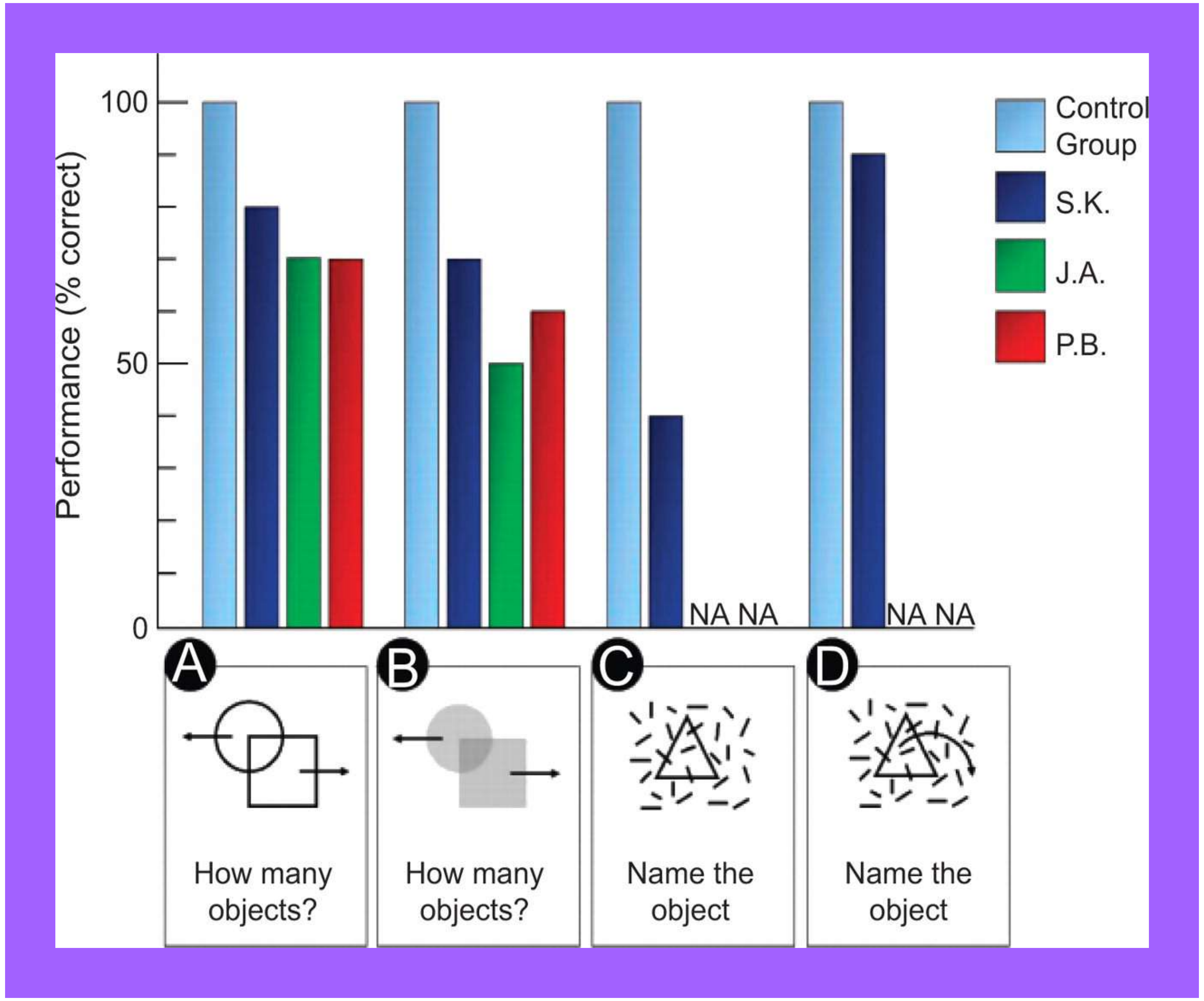
- Image model trained on isolated video frames
- Test time: run image model on consecutive frames
- Cannot integrate temporal information, no motion features
- Cannot exploit temporal redundancy to improve efficiency

Frame-by-frame video model

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

Importance of motion

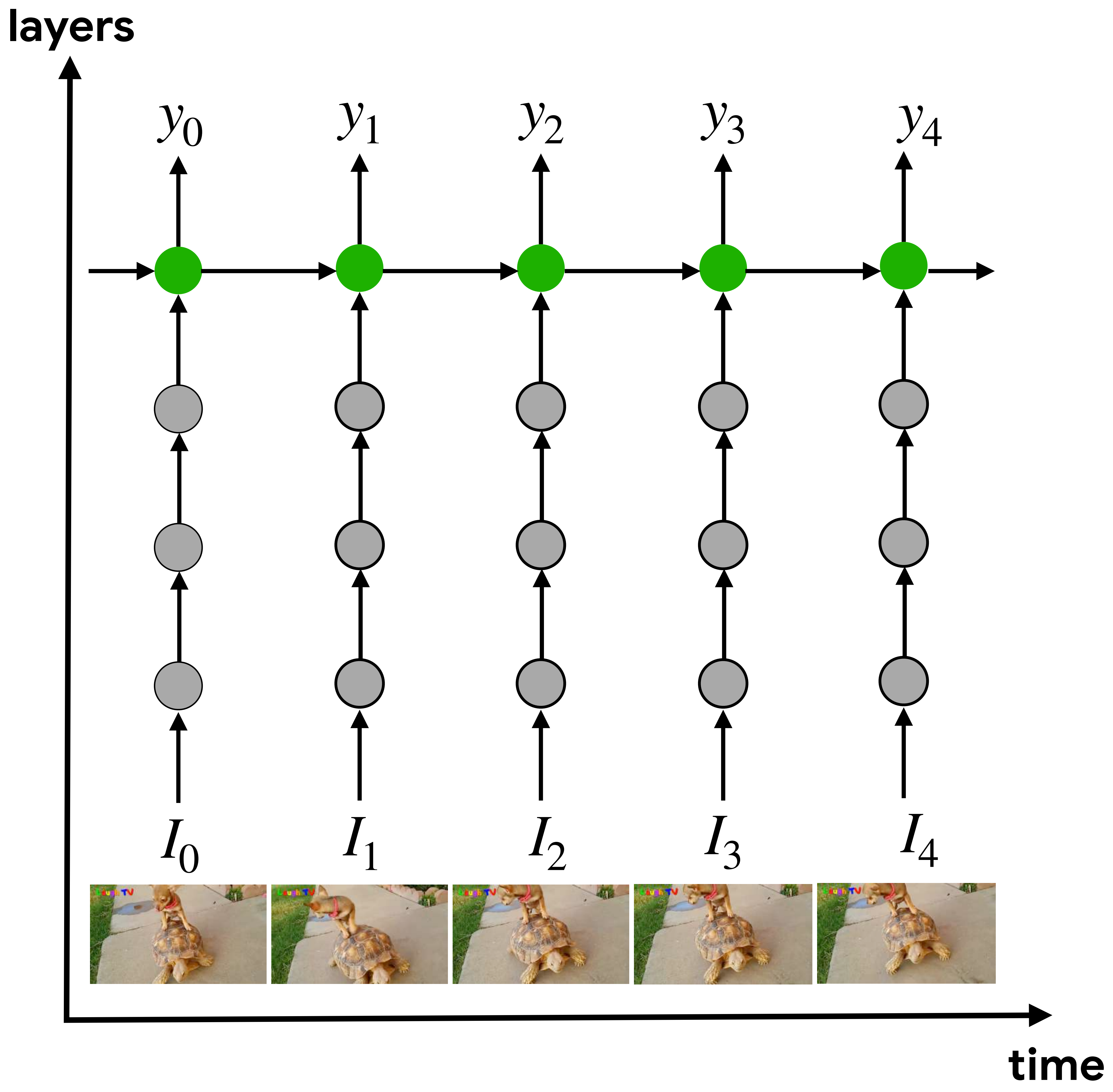




Motion helps object recognition when learning to see.

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

Temporal inductive biases

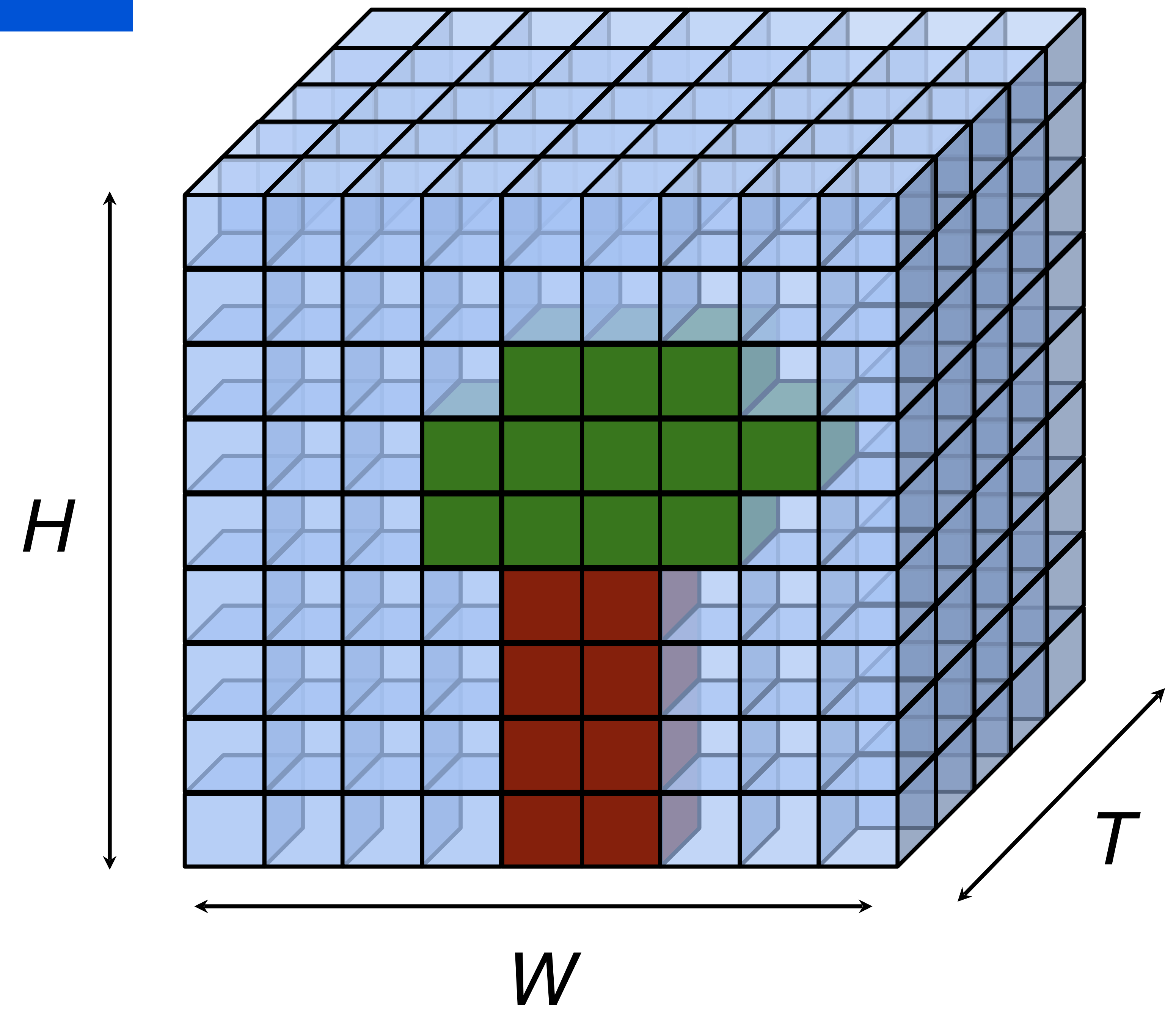


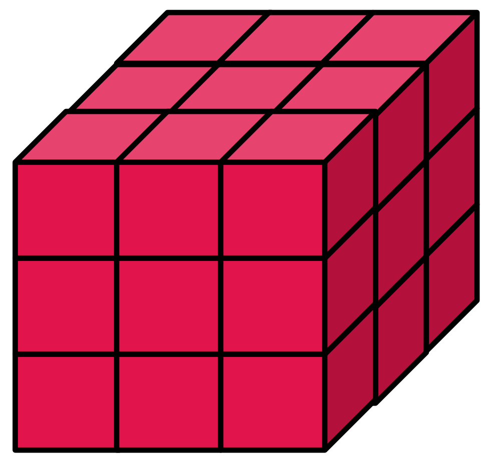
- Temporal locality & temporal invariance
- Use RNNs: Markov assumption and parameters shared over time
- On top of a conv image encoder with spatial inductive biases
- No hierarchical processing in time

3D convolutional models: better temporal inductive biases

Video as a volume

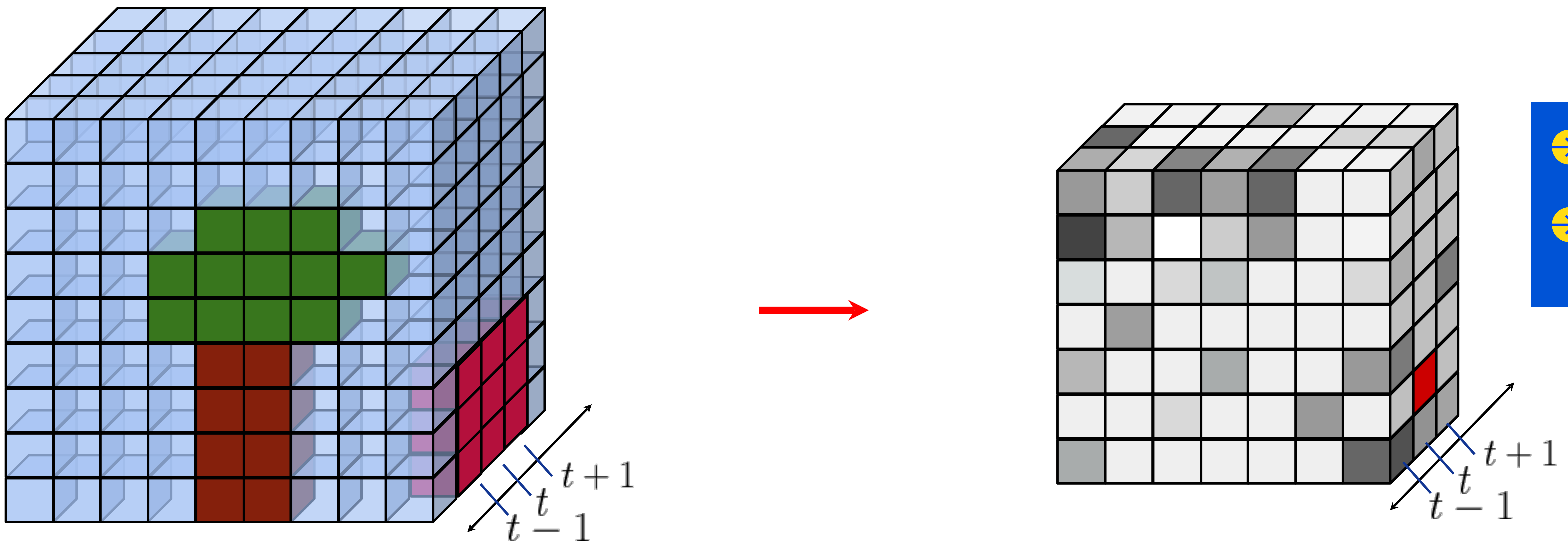
- stack frames *T x H x W x 3*
- apply 3D convolutions





$$y = \sum_{i \in 3\times3\times3} \mathbf{w}_i \mathbf{x}_i + b$$

3D convolutional models: better temporal inductive biases



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

Multi-resolution models

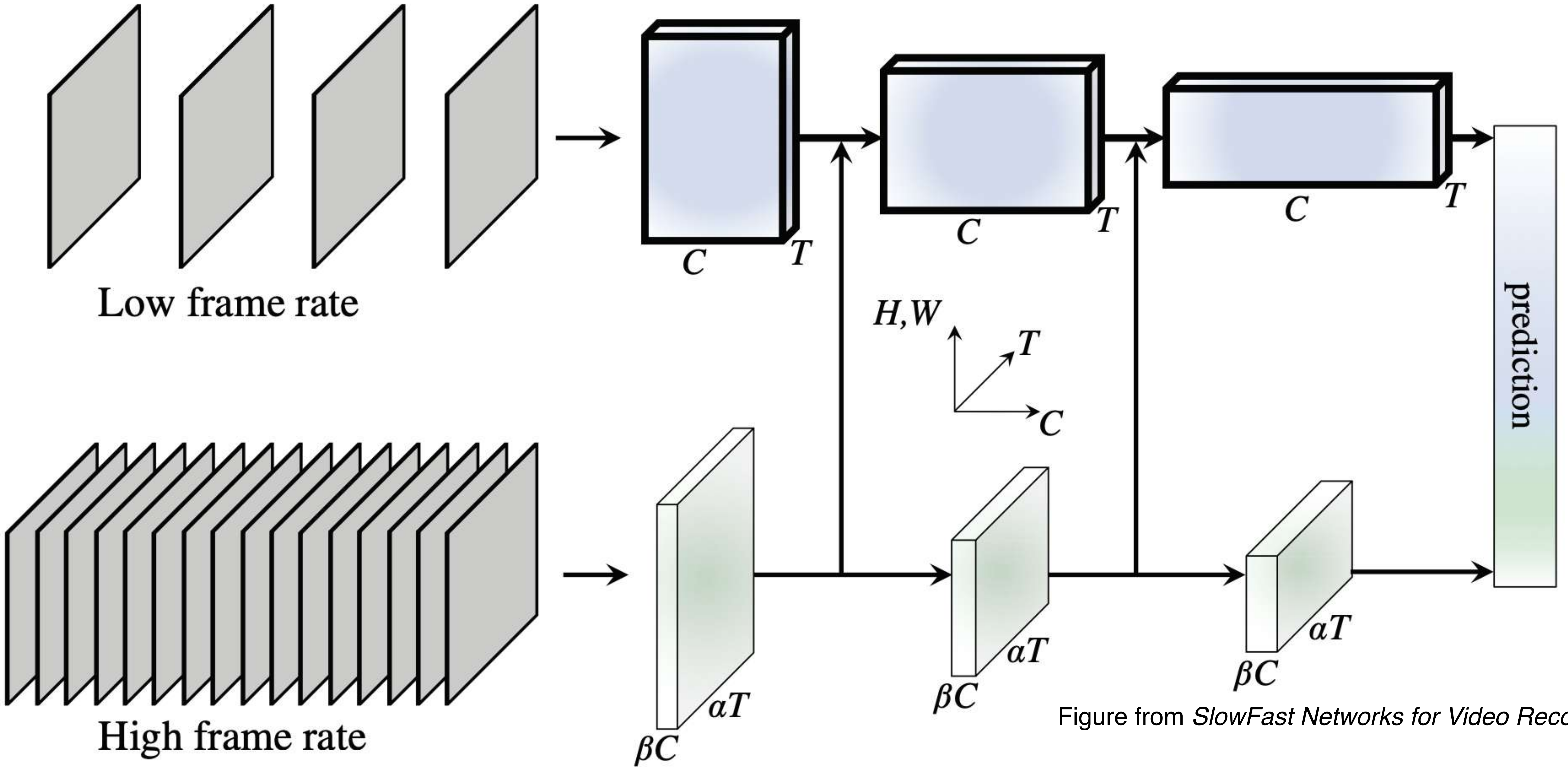


Figure from *SlowFast Networks for Video Recognition*, Feichtenhofer et al, 2019

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
.....

Towards bias-free models

Inductive biases help generalisation and data efficiency.

But if we don’t care about data efficiency, biases may hurt performance.

Inductive biases reduce the generality of the models.

Vision Transformer (ViT) vs. ResNet

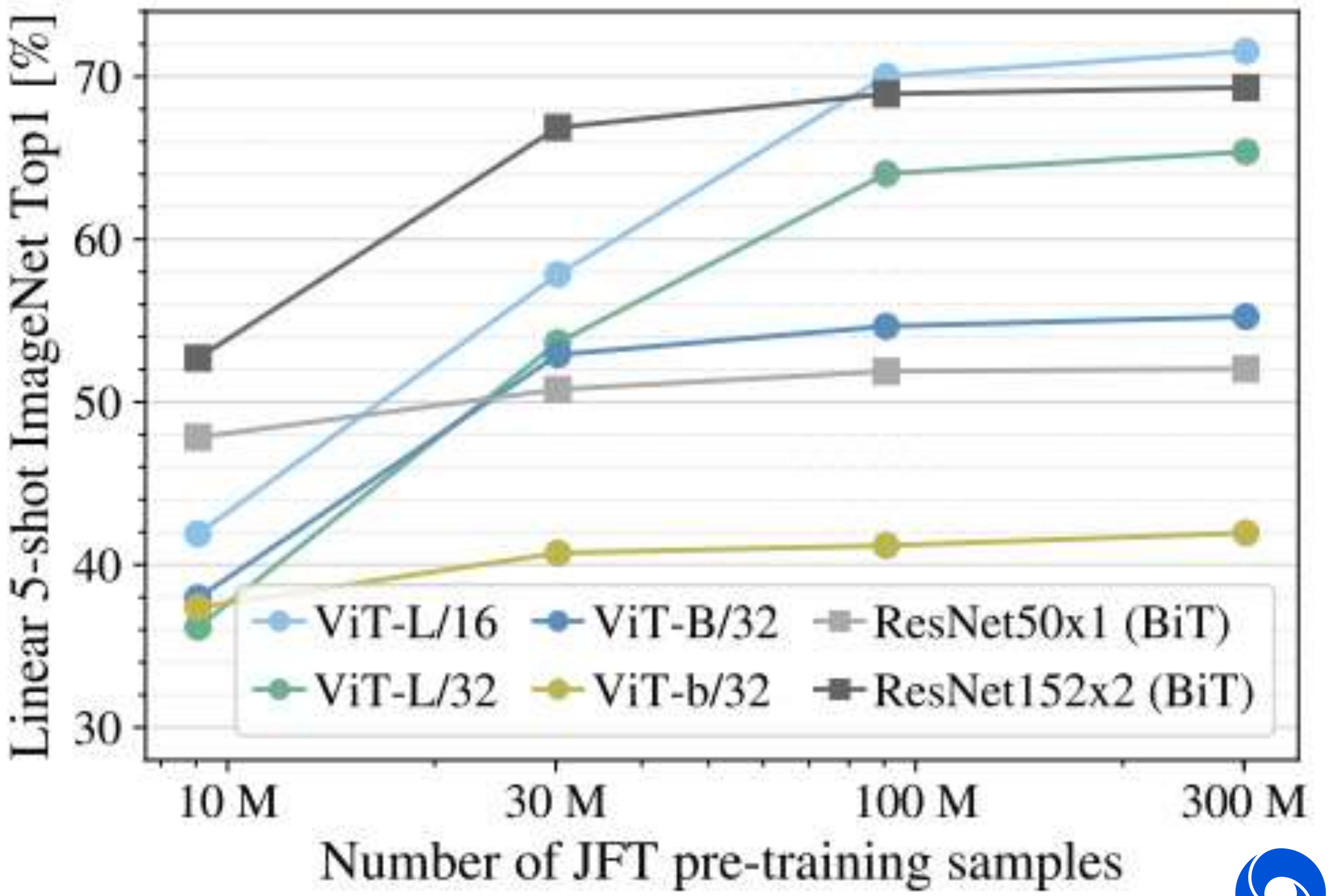
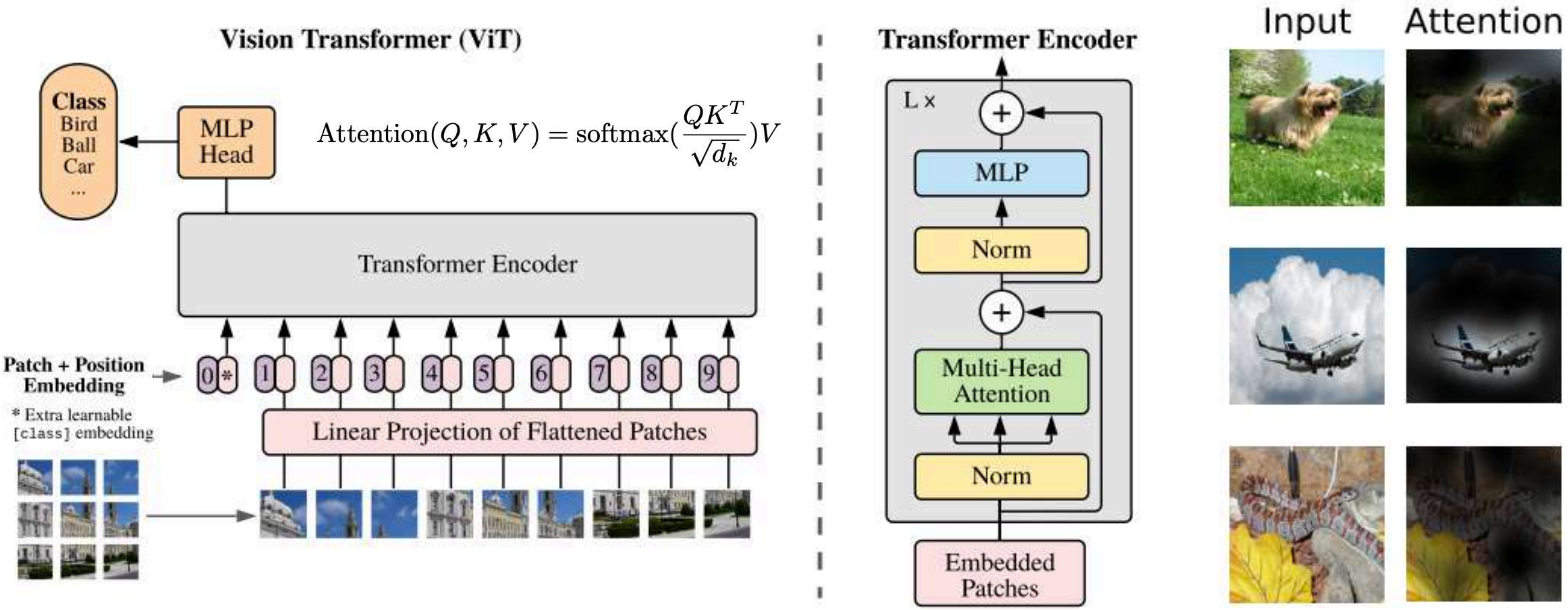


Figure from *An image is worth 16x16 words,* Dosovitskiy et al. (2021)

Vision Transformers — Self-attention models



No hierarchical processing, (almost) no locality, no translation invariance

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

Video Transformers — ViT extended in time

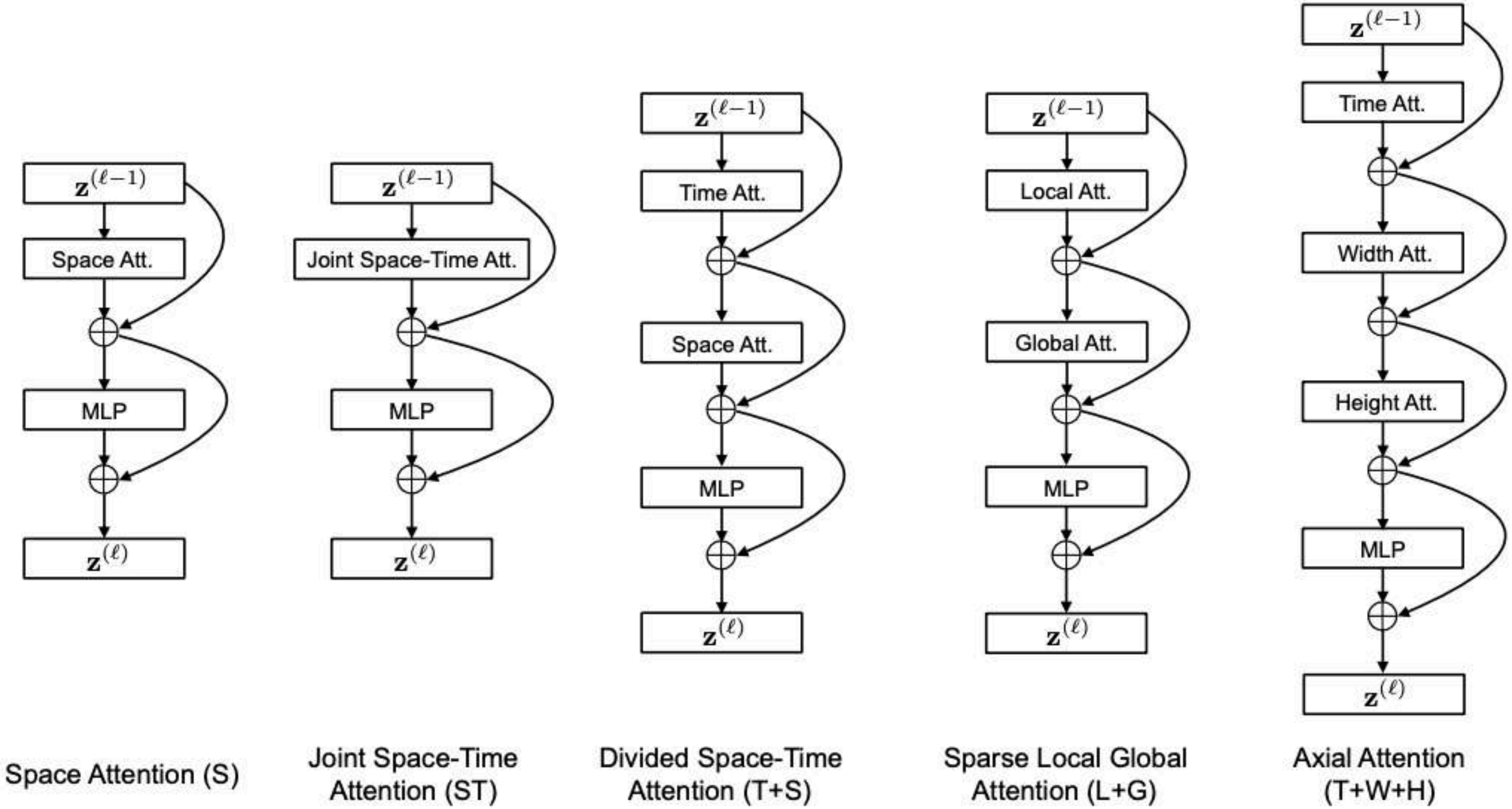


Figure from *Is Space-Time Attention All You Need for Video Understanding?* Bertasius et al, 2021

PerceiverIO — A general architecture for structured inputs & outputs

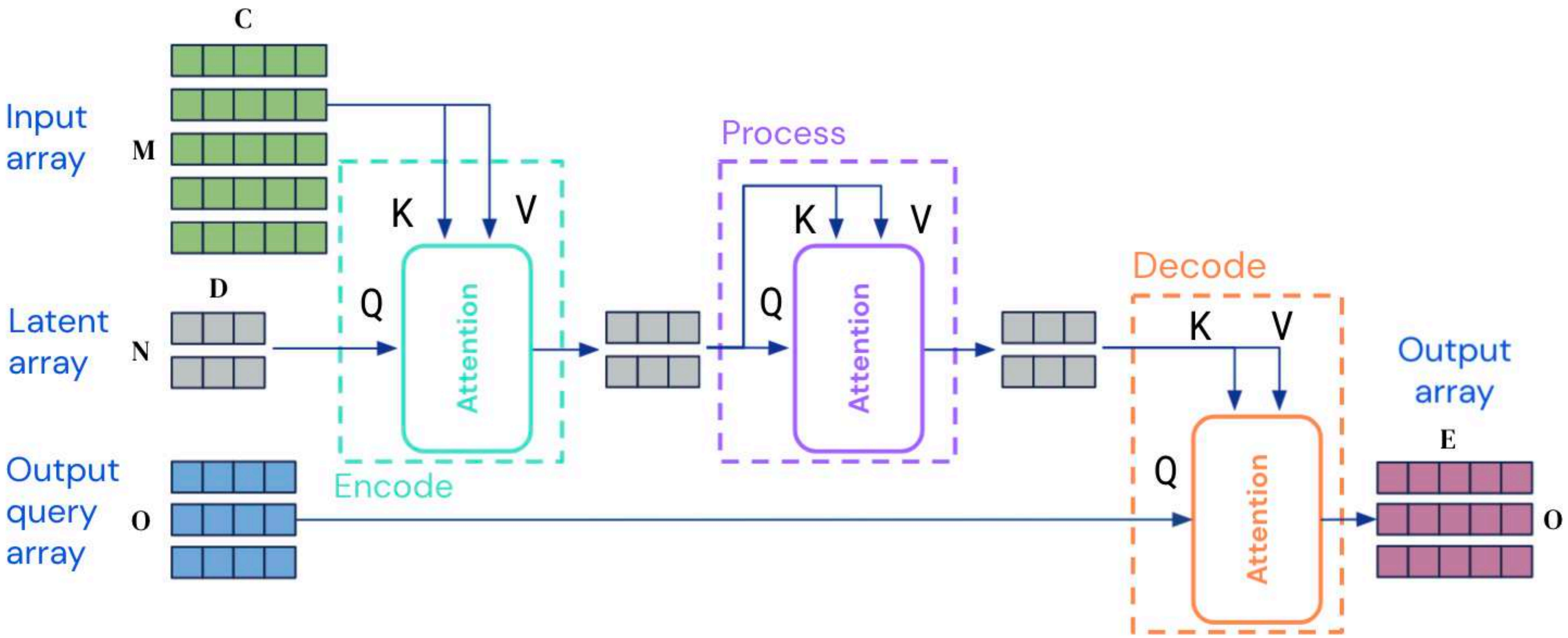


Figure from *PerceiverIO - A General Architecture for Structured Inputs and Outputs,* Jaegle et al (2021)

Summary

- Spatial and temporal inductive biases in convolutional feedforward and recurrent models help generalisation in lower data regime, but may hurt performance in large data regime

- Towards models free of inductive biases:

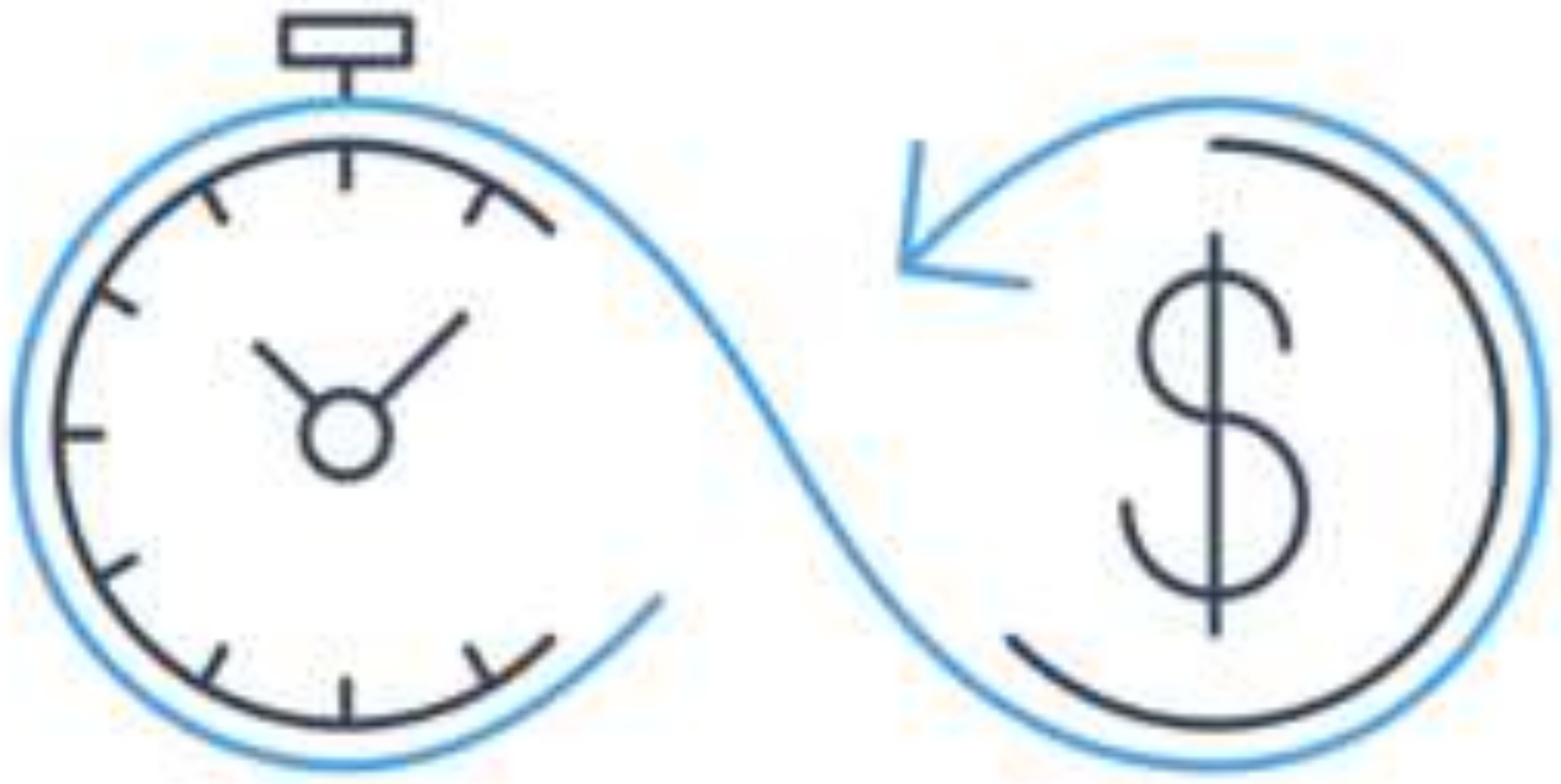
  - improved performance in large data regime

  - great for multimodal processing

  - promising to tackle data in domains where we don’t have good inductive biases

02 Task design

Challenges in video supervised learning



Labels are expensive

Agreement: definition? granularity?

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* $\hat{z} = g_w(x)$

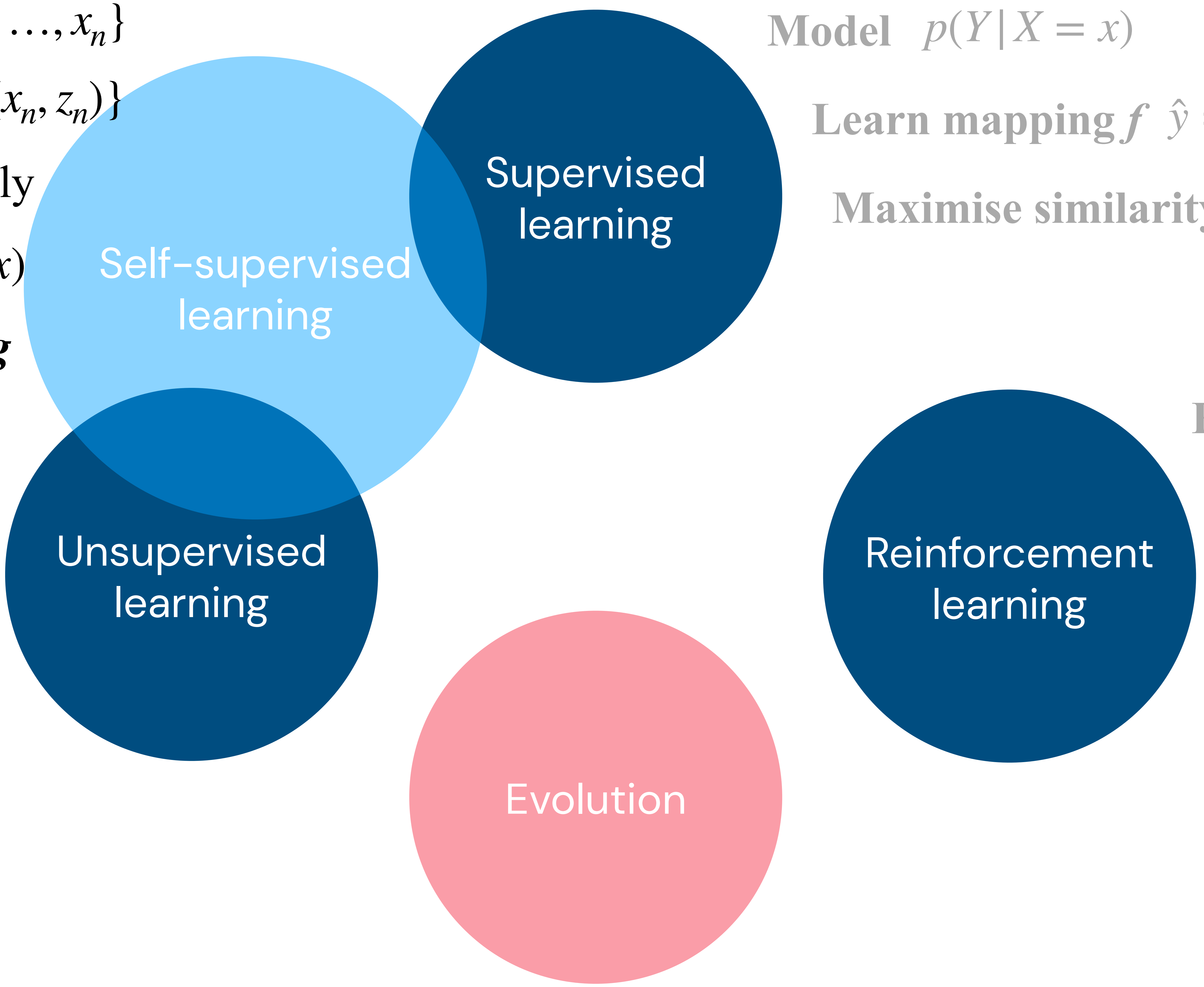
Transfer learning from *g*

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 $(\hat{y}, y)$



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

Model $p(X)$

Learn the structure in the data

Environment

MDP (S, A, R, T)

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

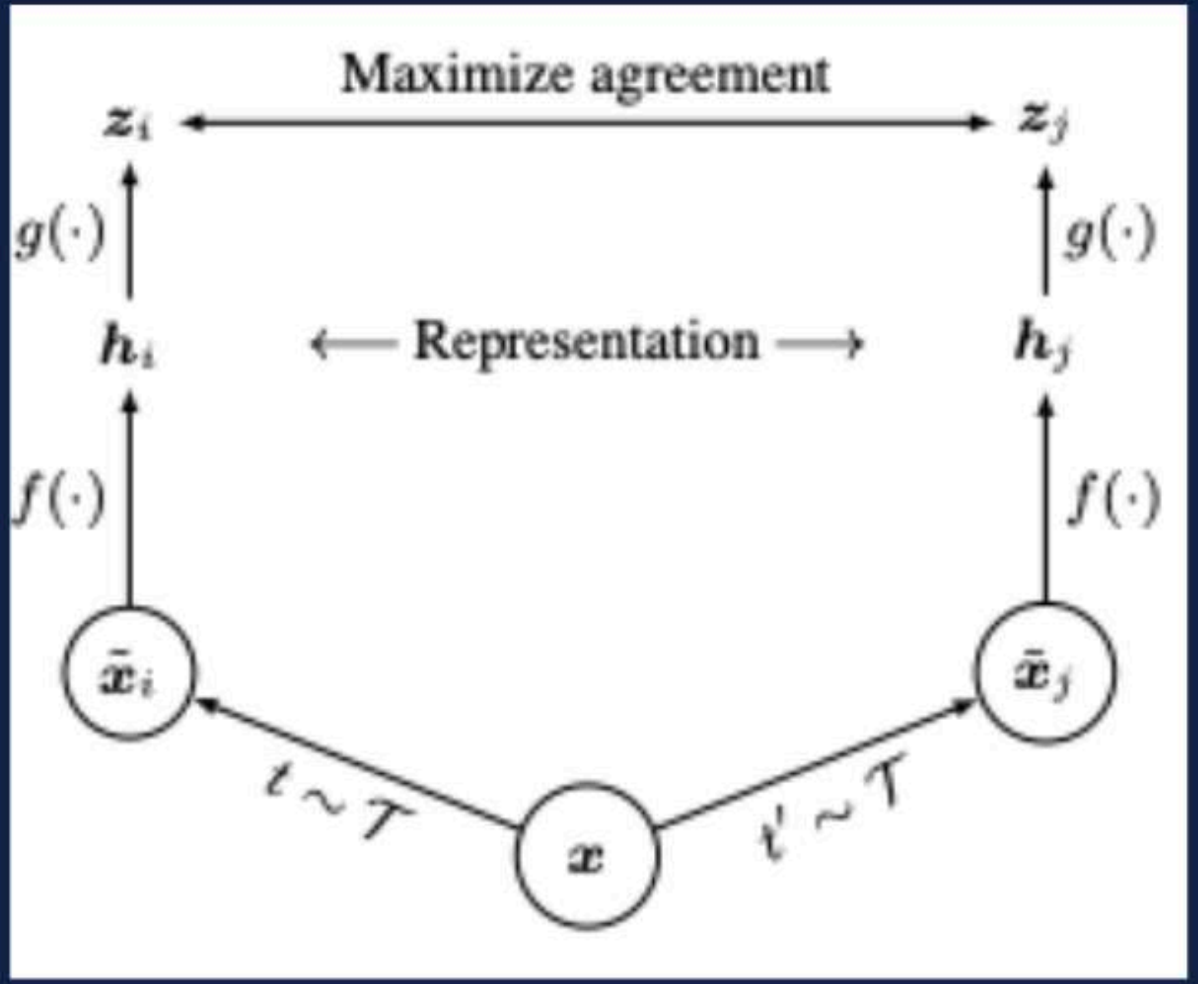
Maximise reward

Multimodal sensing of the environment

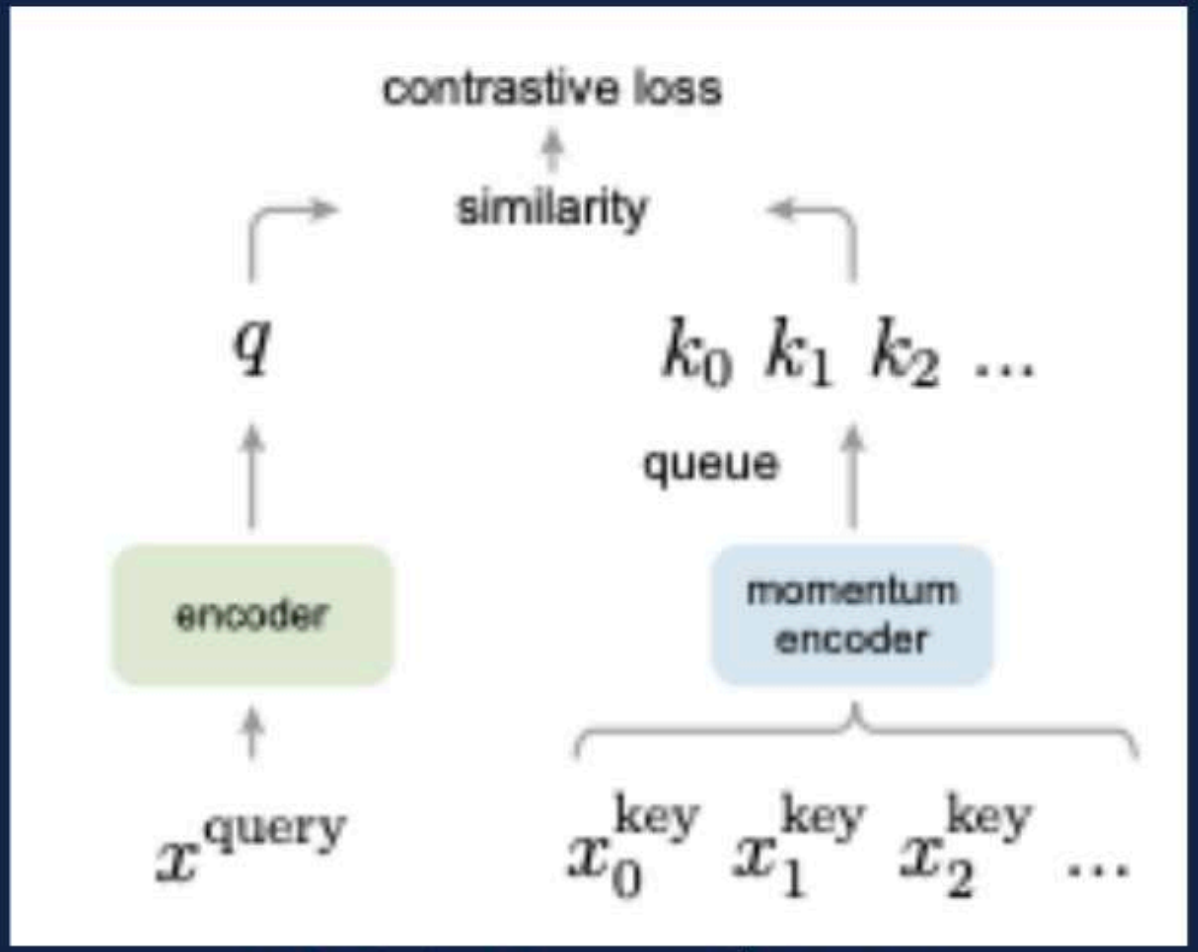
[...] towards the root and try to get as close to the root as possible, nice long strokes [...]

Self-supervised learning for pre-training

Vision

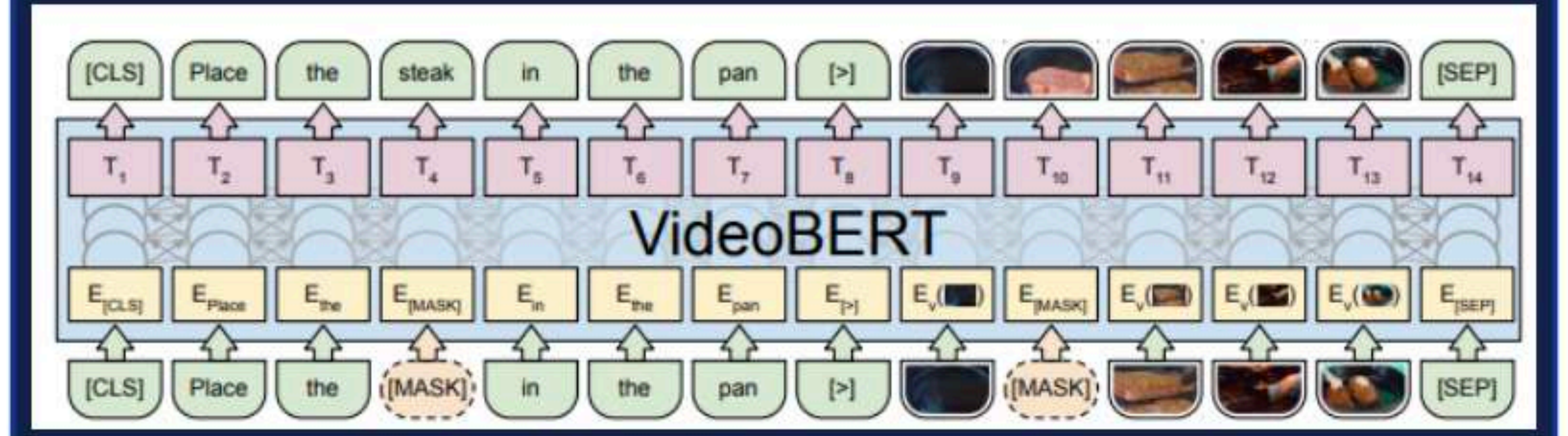


SimCLR: Chen et al, 2020

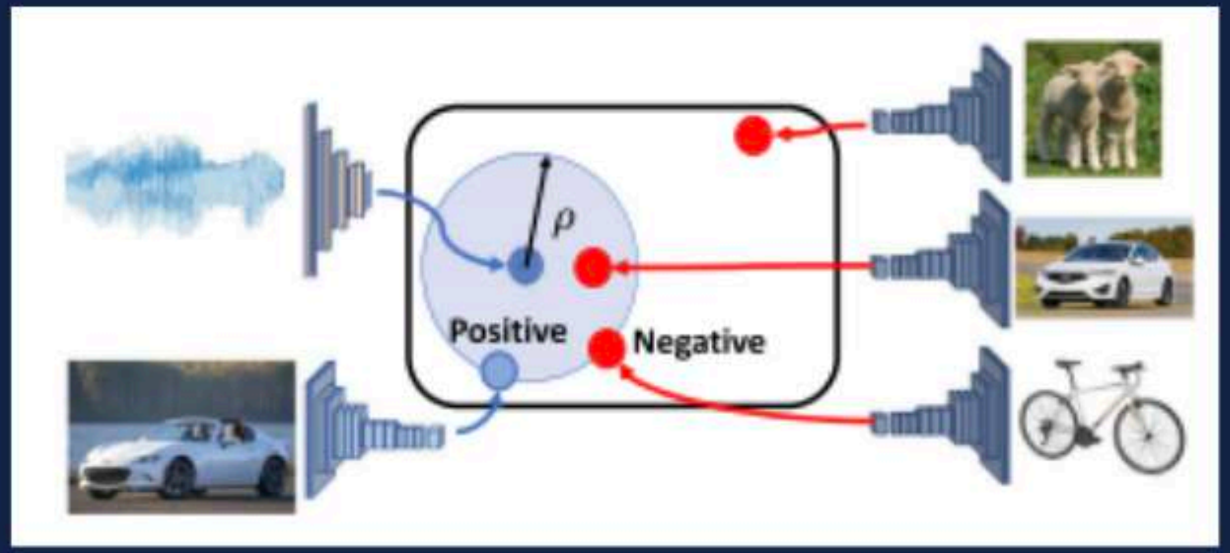


MOCO: He et al, 2020

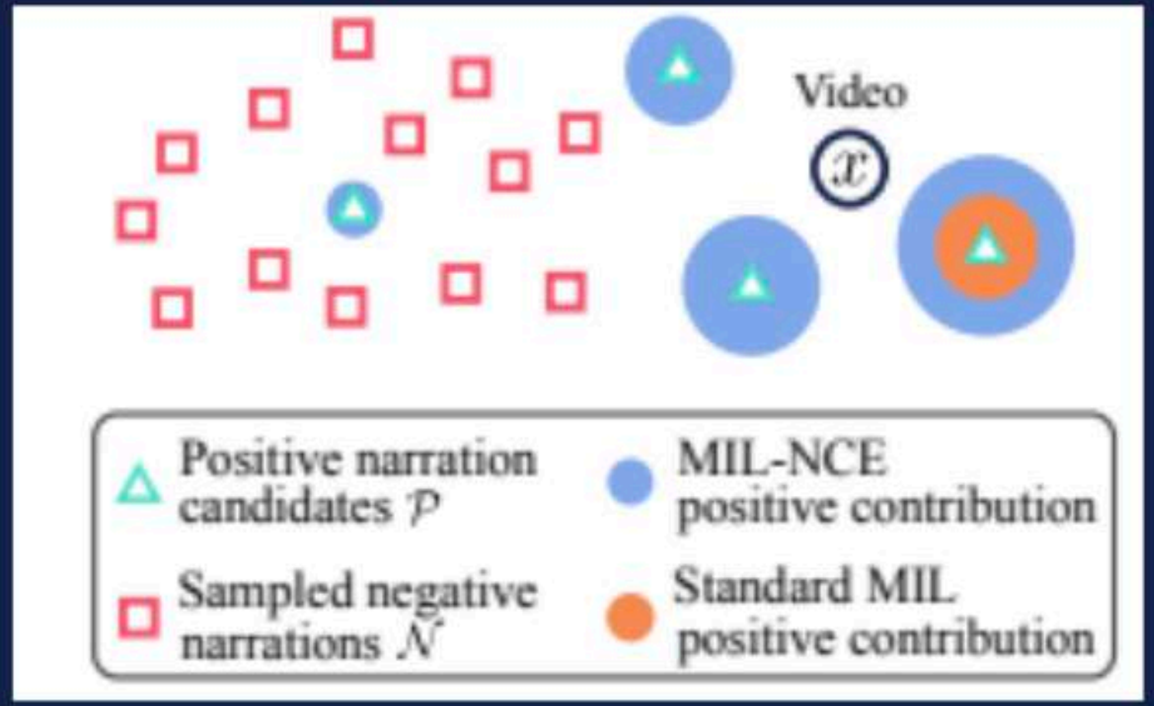
Vision+Language



VideoBERT: Sun et al, 2019

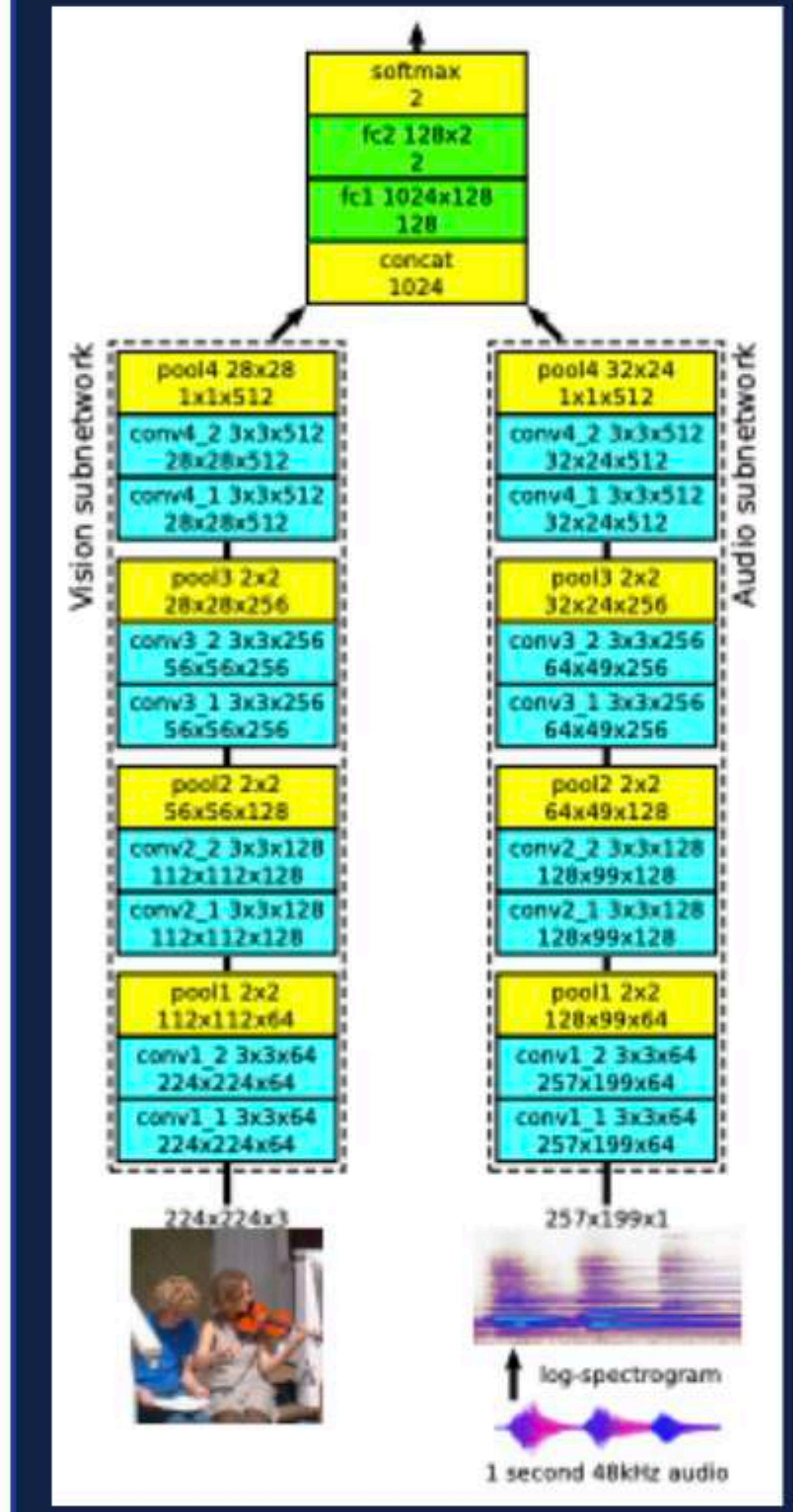


DaveNet: Harwath et al, 2018

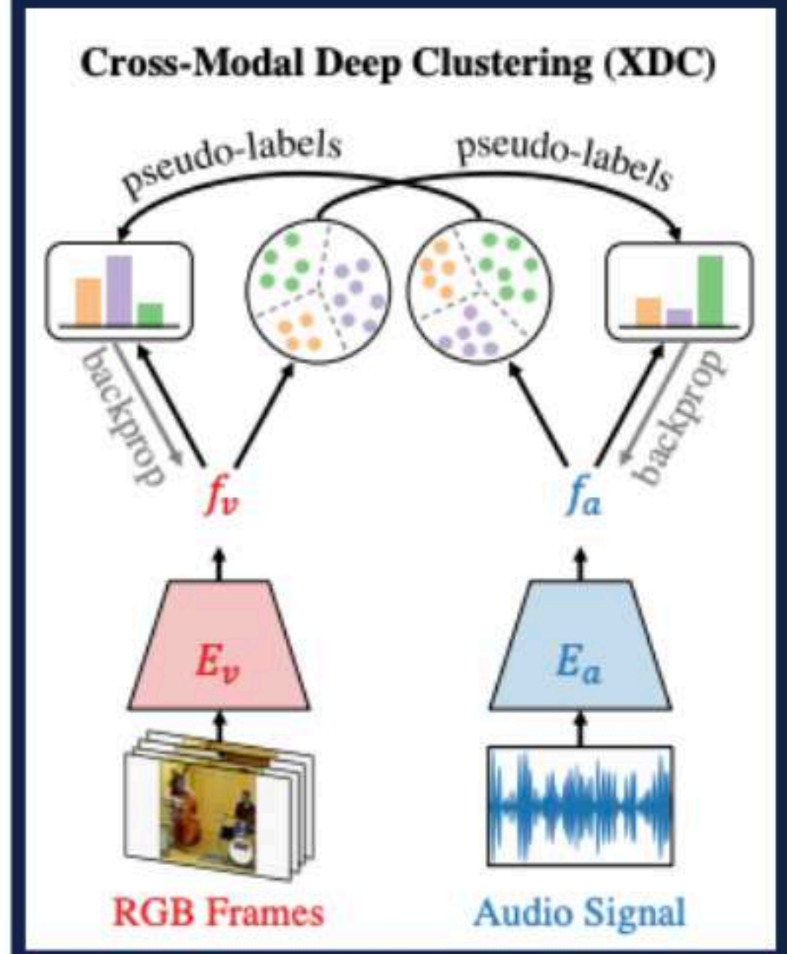


MIL-NCE: Miech, Alayrac et al, 2020

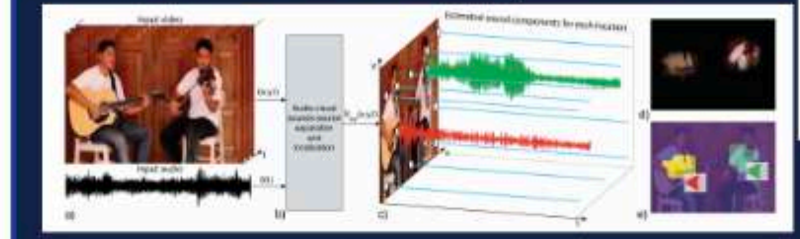
Vision+Audio



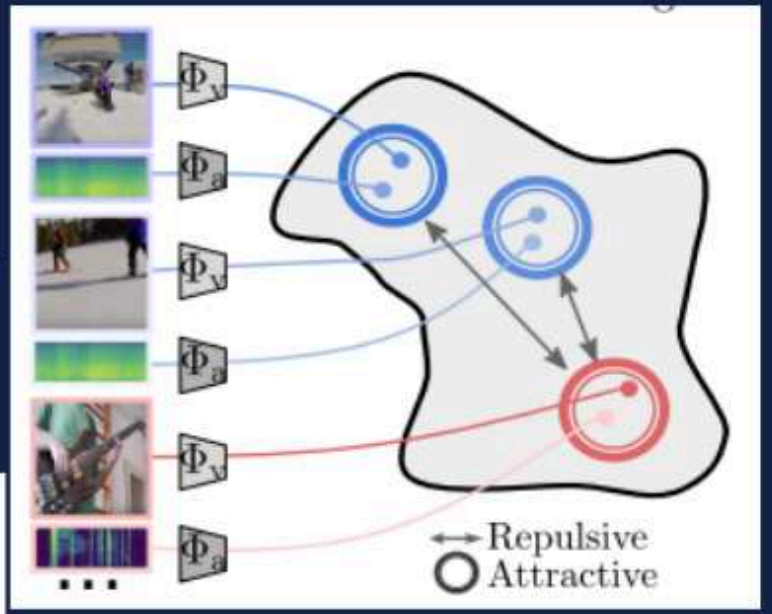
L3: Arandjelovic and Zisserman, 2017



XDC: Alwassel at al, 2020



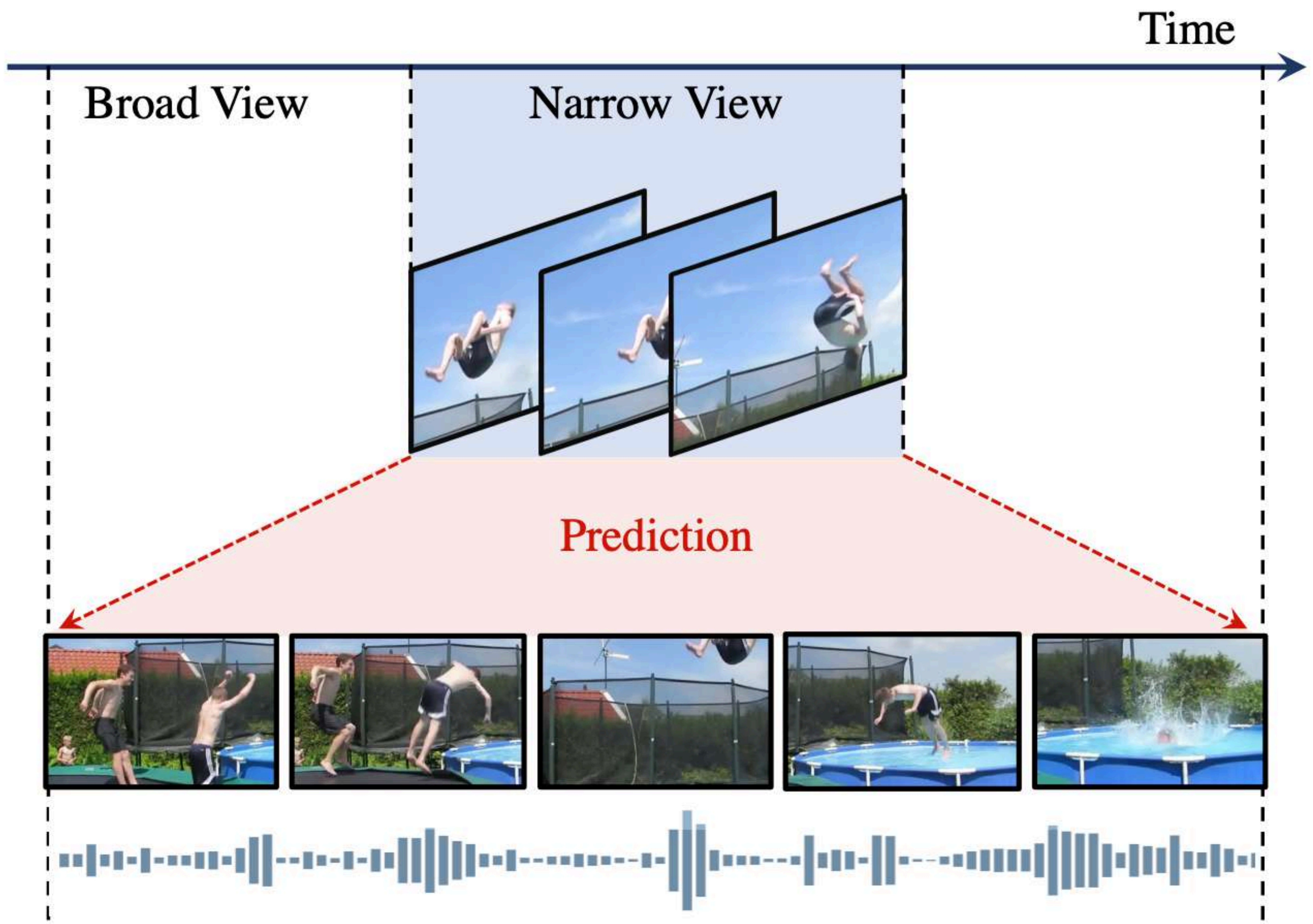
Sound of Pixels: Zhao et al, 2018



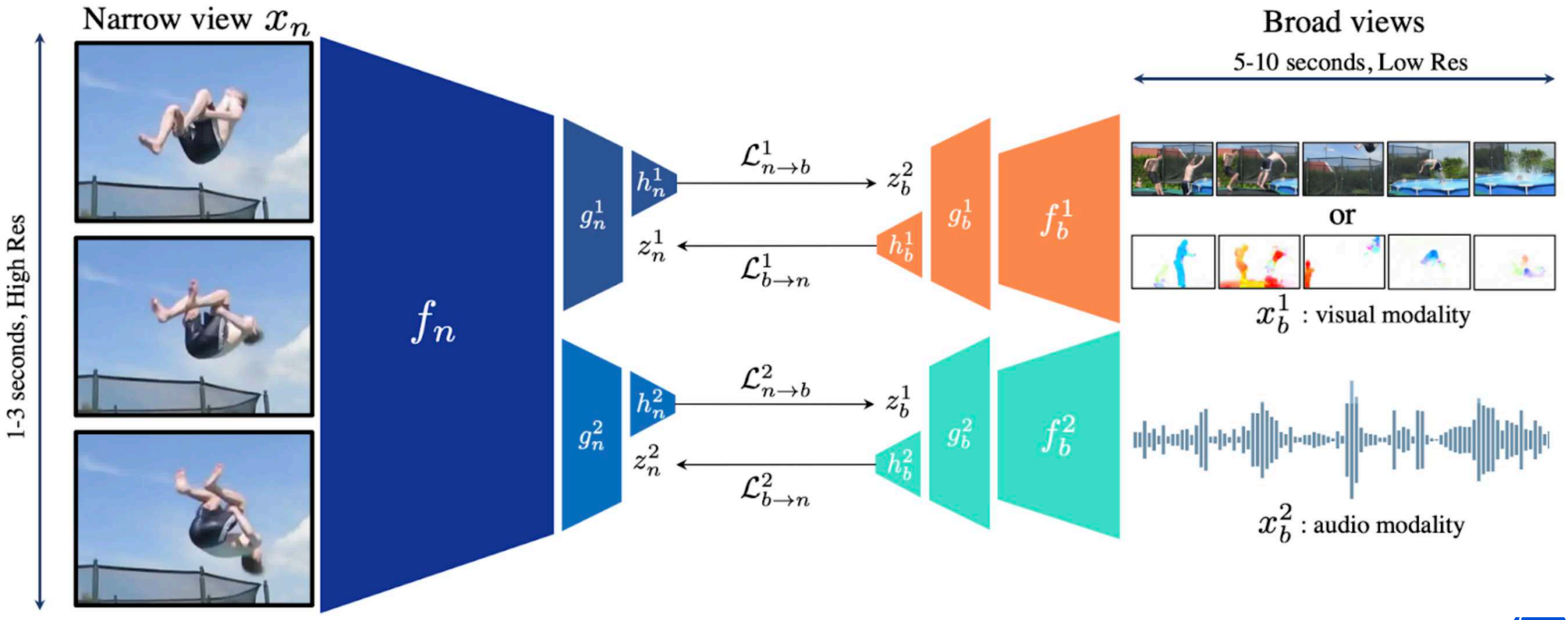
GDT: Patrick at al, 2020

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
z_n^1
g_n^2
h_n^2
z_n^2
$\mathcal{L}^1_{n\to b}$
$\mathcal{L}^1_{b\to n}$
$\mathcal{L}^2_{n\to b}$
$\mathcal{L}^2_{b\to n}$
z_b^2
z_b^1
h_b^1
g_b^1
f_b^1
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 Operation mode

Efficiency challenges in video processing

Ideal model

- Accurate
- Low latency, high throughput* (during training and inference)
- Memory efficient
- Energy efficient

Existing models (deep nets)

- Accurate (to some extent)
- High latency, low throughput (at inference)
- Large memory requirements
- Energy hungry

*During training, throughput can be improved by using distributed training with many parallel resources and large batches.

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.

All models are executed sequentially over depth

Inductive biases for the operation mode

- Training: backpropagation – general purpose algorithm
- Sequential processing along depth and time (for causal models)
- Using domain knowledge about videos, what inductive biases can be used?
  - Videos are temporal sequences, not isolated frames
  - Local smoothness of videos over time

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

Is sequential mode optimal?

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

Maximise throughput

Efficient use of hw resources

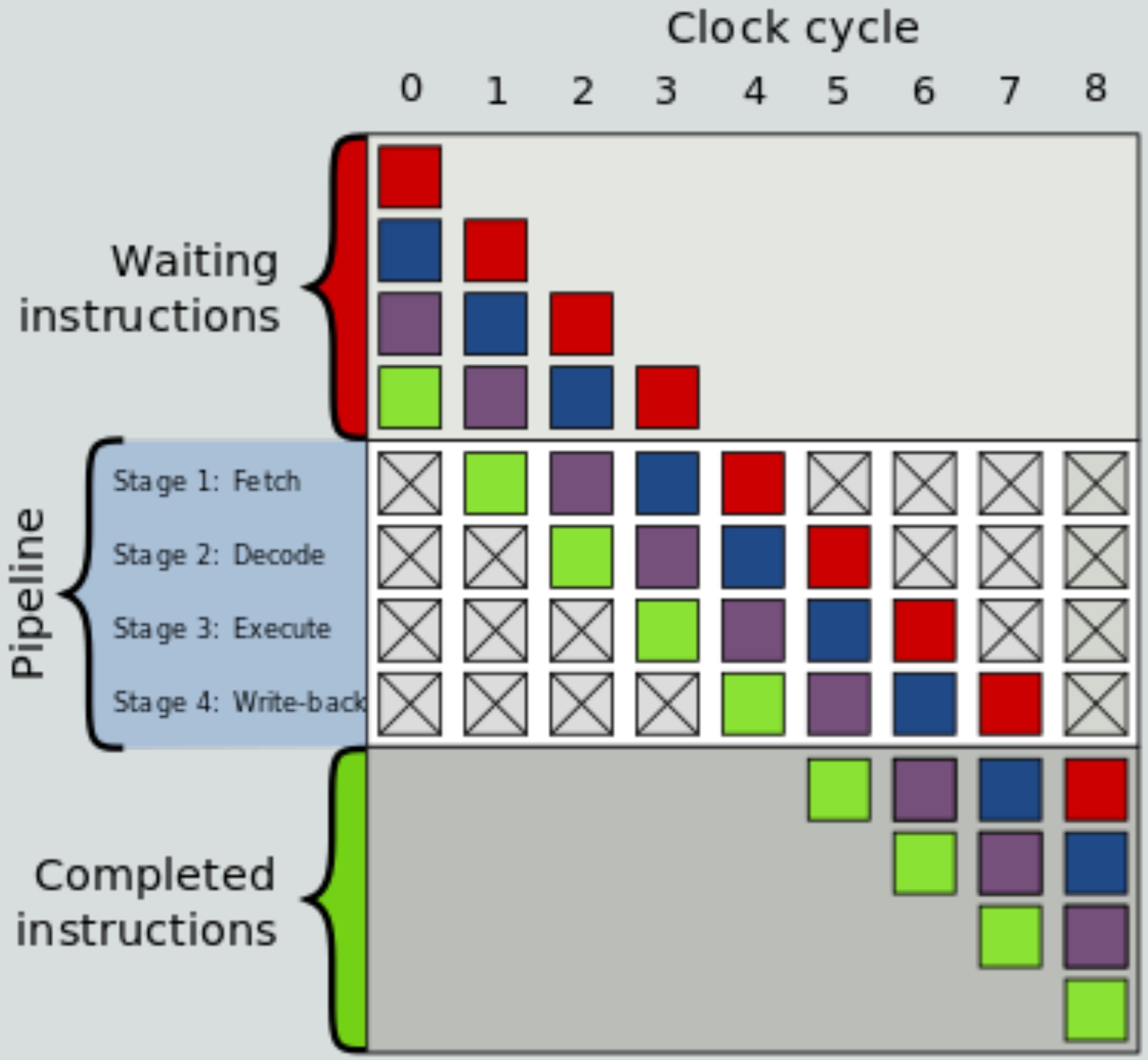


Figure from https://en.wikipedia.org/wiki/Instruction_pipelining

Our proposed algorithm: Sideways

ZDNet

CENTRAL EUROPE MIDDLE EAST SCANDINAVIA AFRICA UK ITALY SPAIN MORE NEWSLETTERS

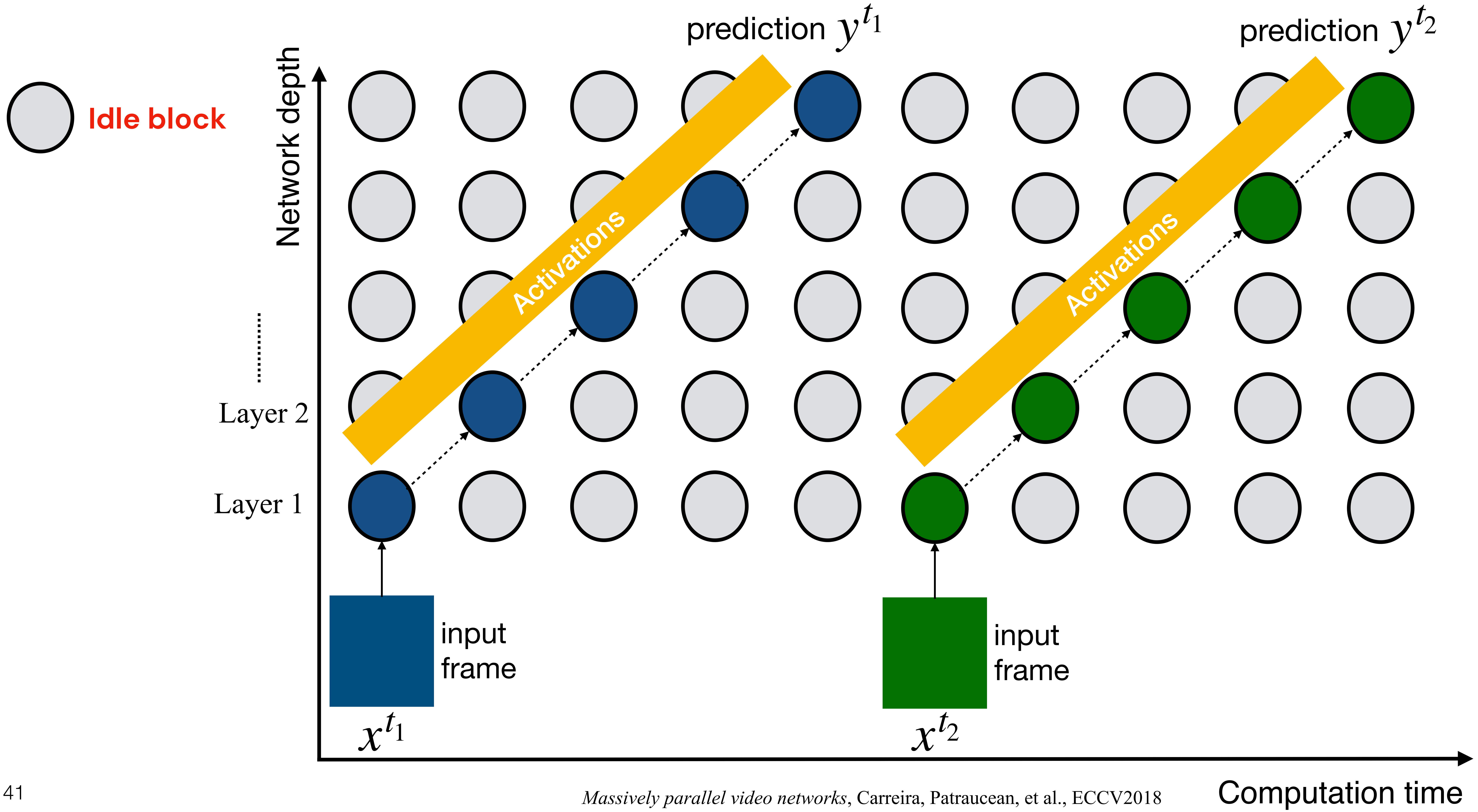
MUST READ: Tech workers are frustrated and thinking about quitting. Here's how to persuade them to stay

Google DeepMind's 'Sideways' takes a page from computer architecture

To get greater efficiency, Google DeepMind's researchers did what chip designers have long done, built a pipeline so that the learning rule for machine learning -- backpropagation -- is more efficient.

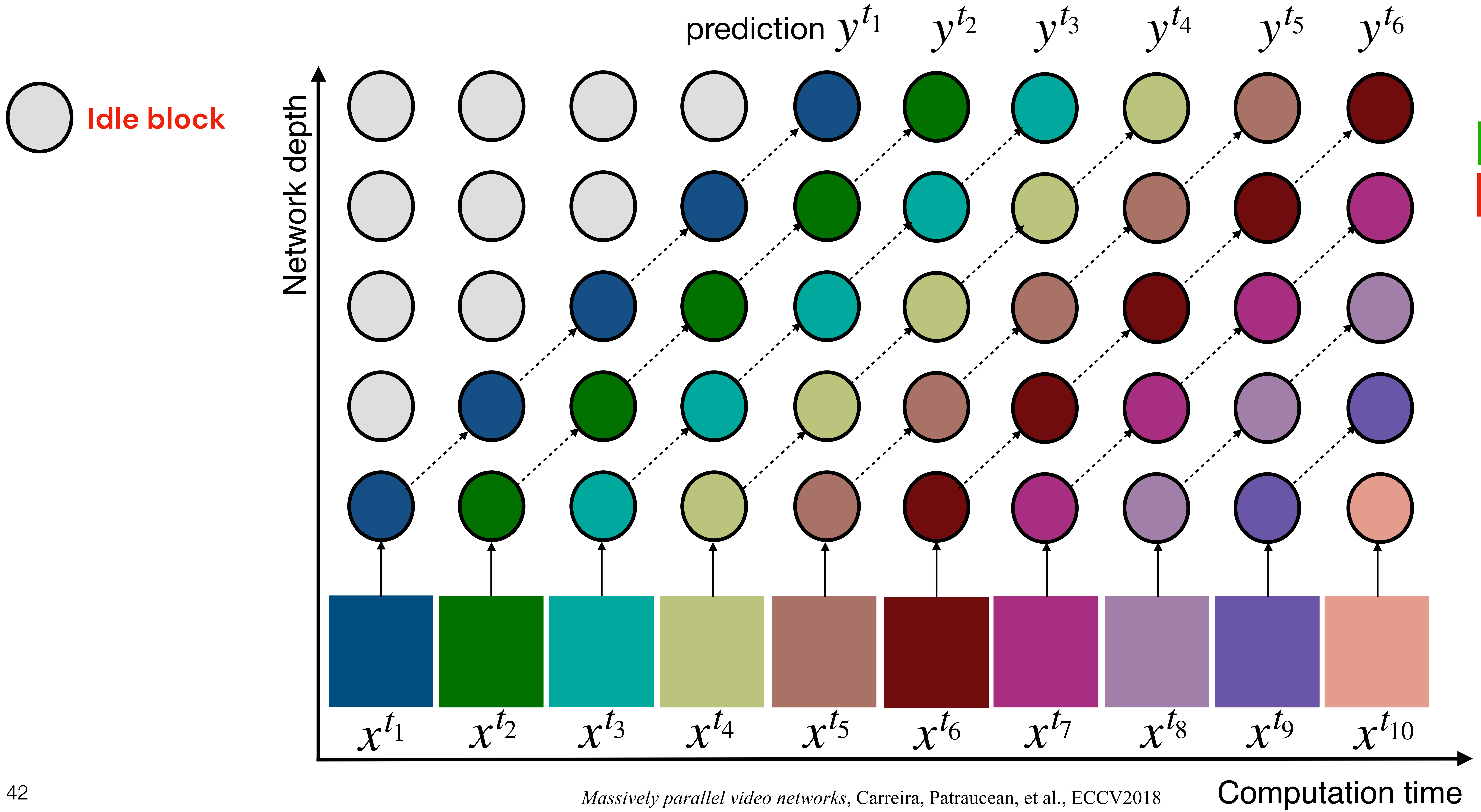
A. Depth-parallel inference

Sequential (frame-by-frame) model



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

Pipelined model

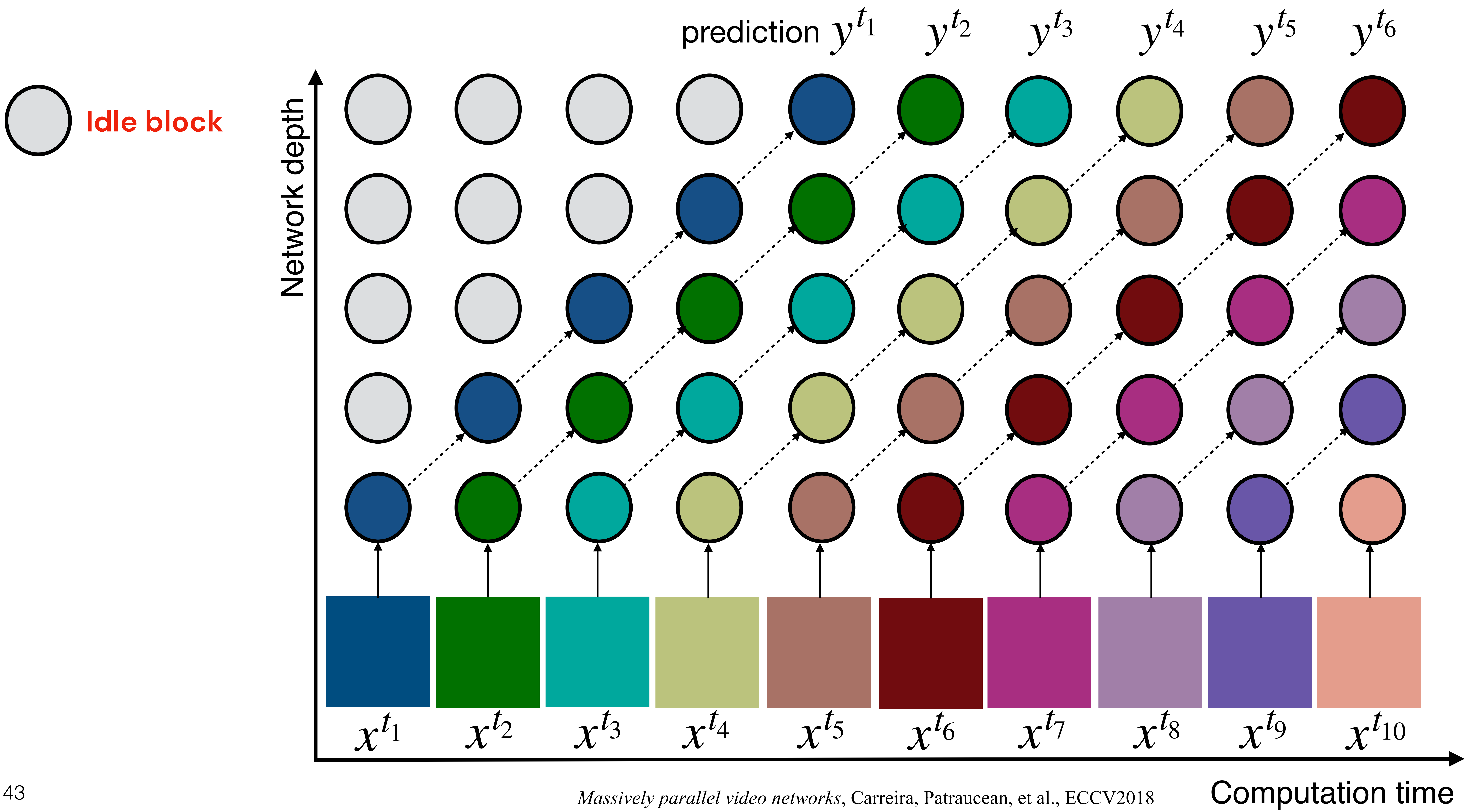


High throughput

Same latency

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

Pipelined model



High throughput

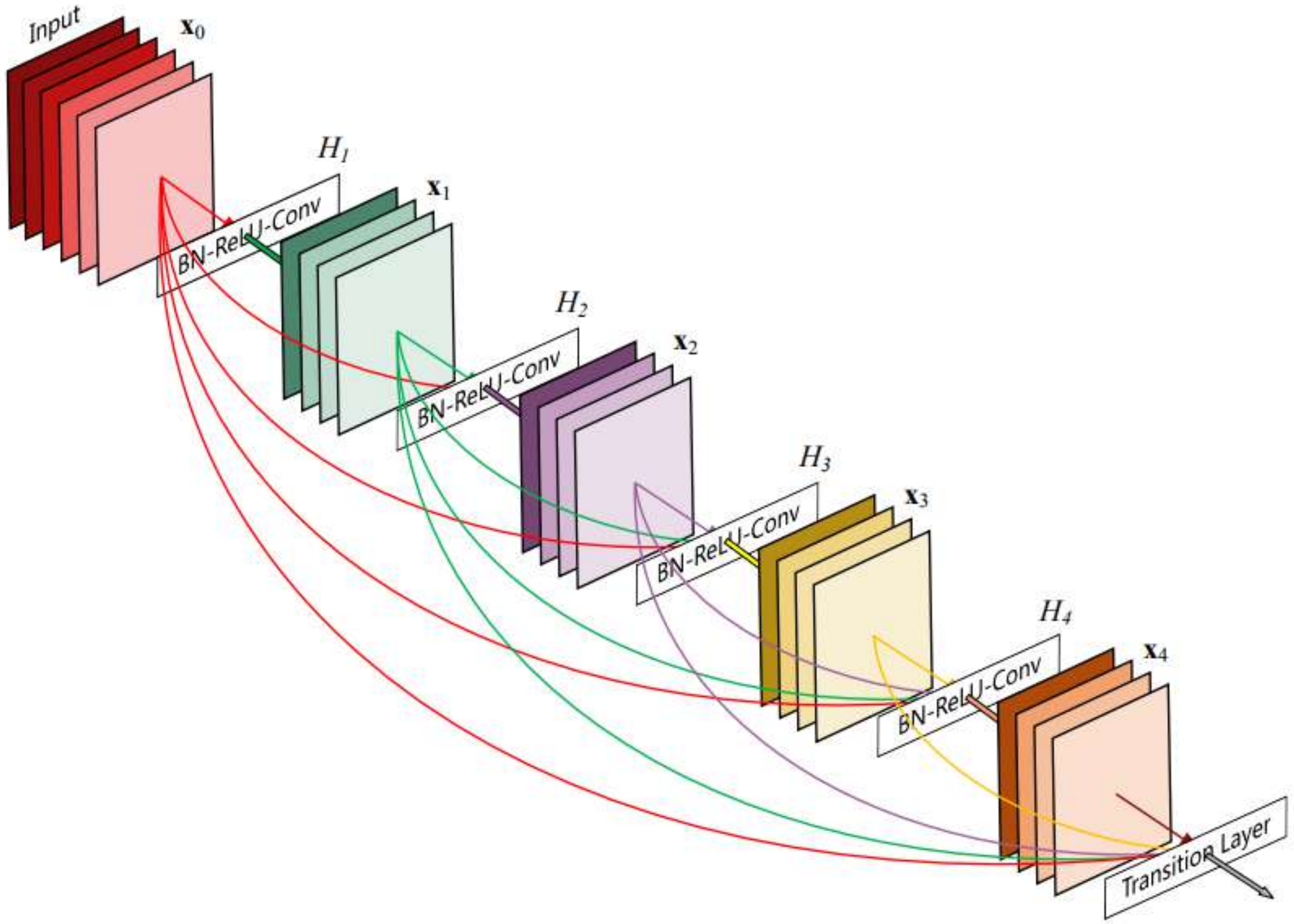
Low (custom) latency

Difficult task

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

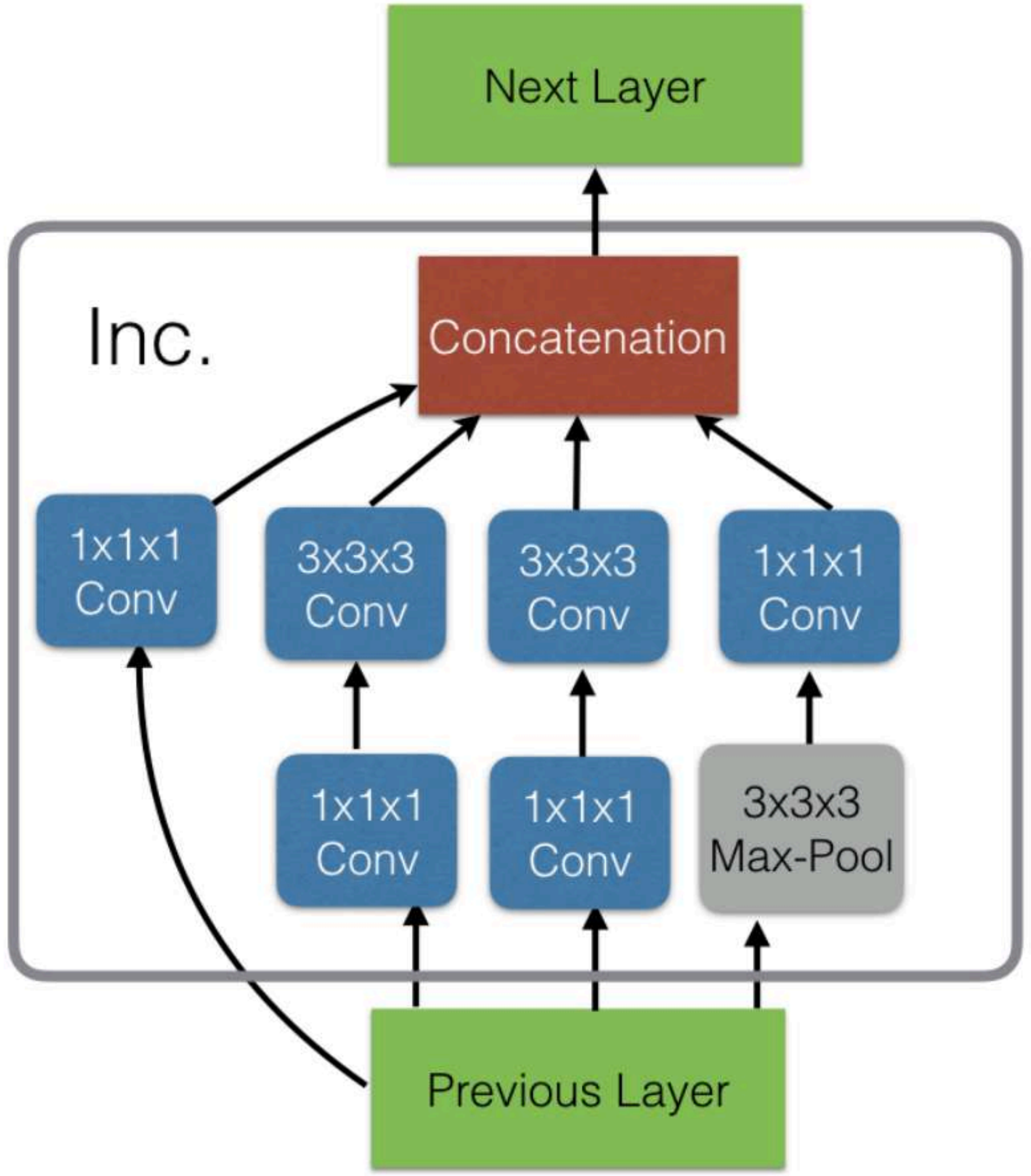
Any existing model can be pipelined

DenseNet



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

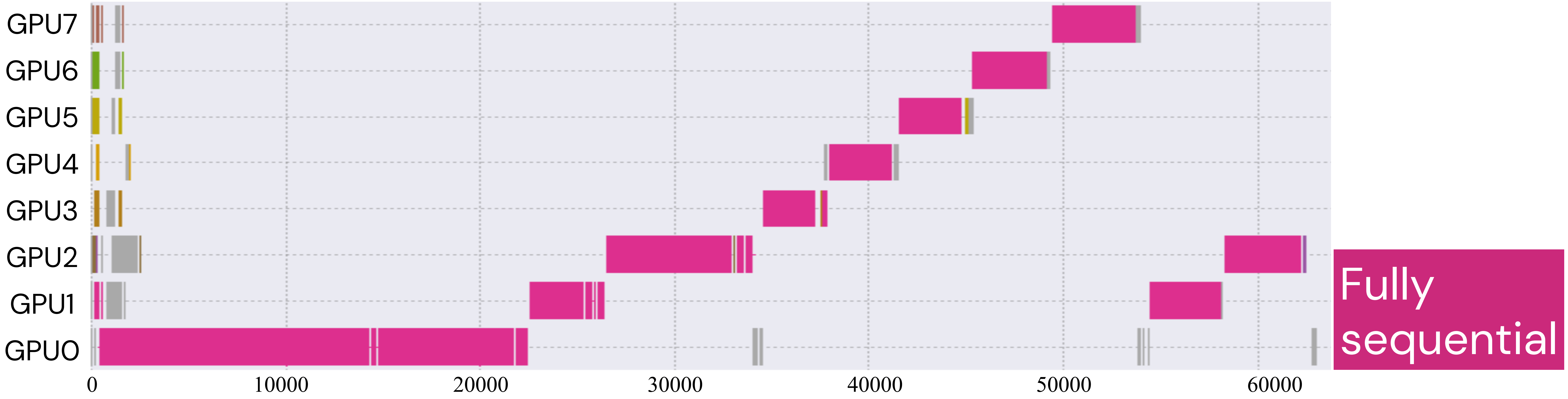
Inception

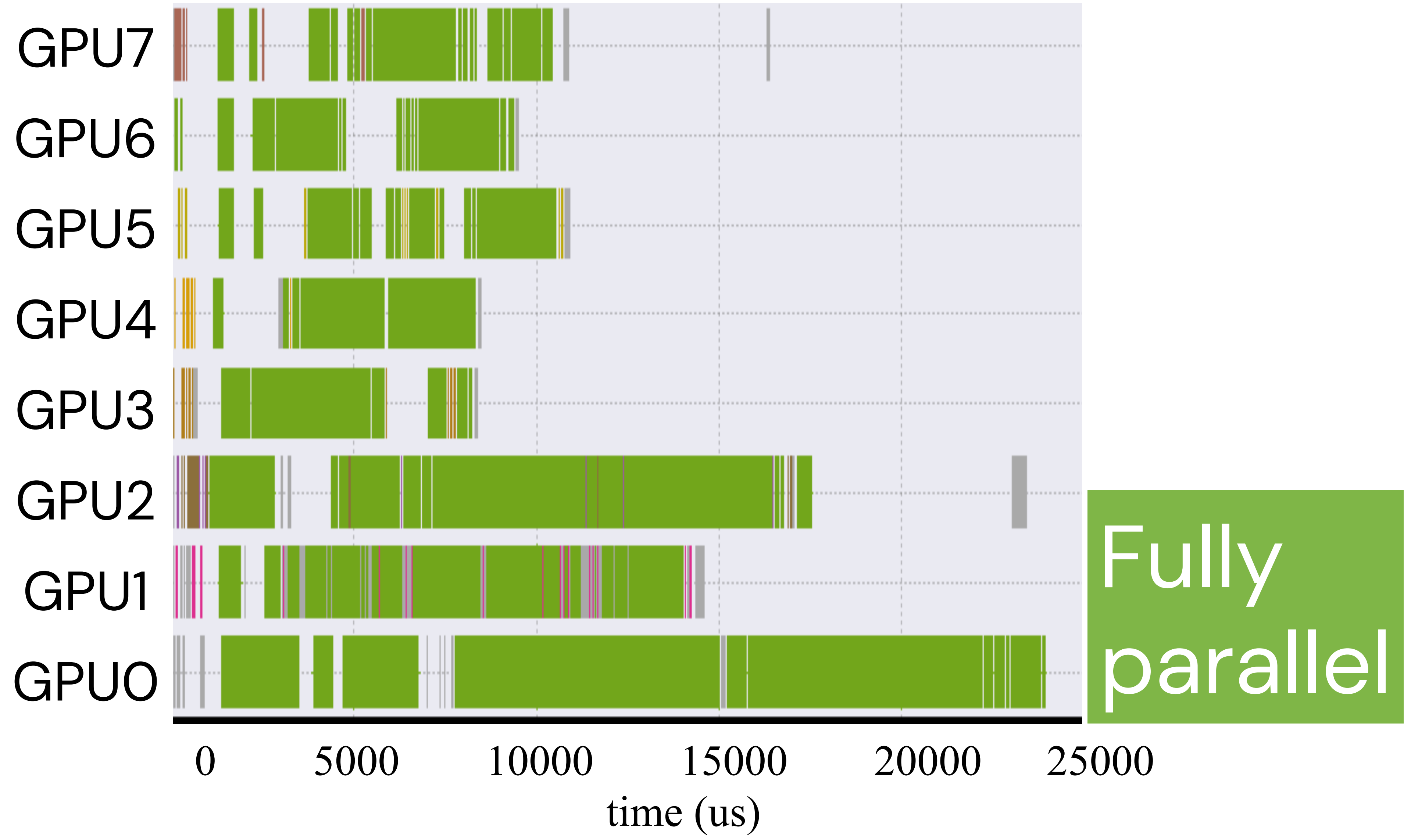


Quo Vadis, action recognition? Carreira and Zisserman, CVPR2017

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

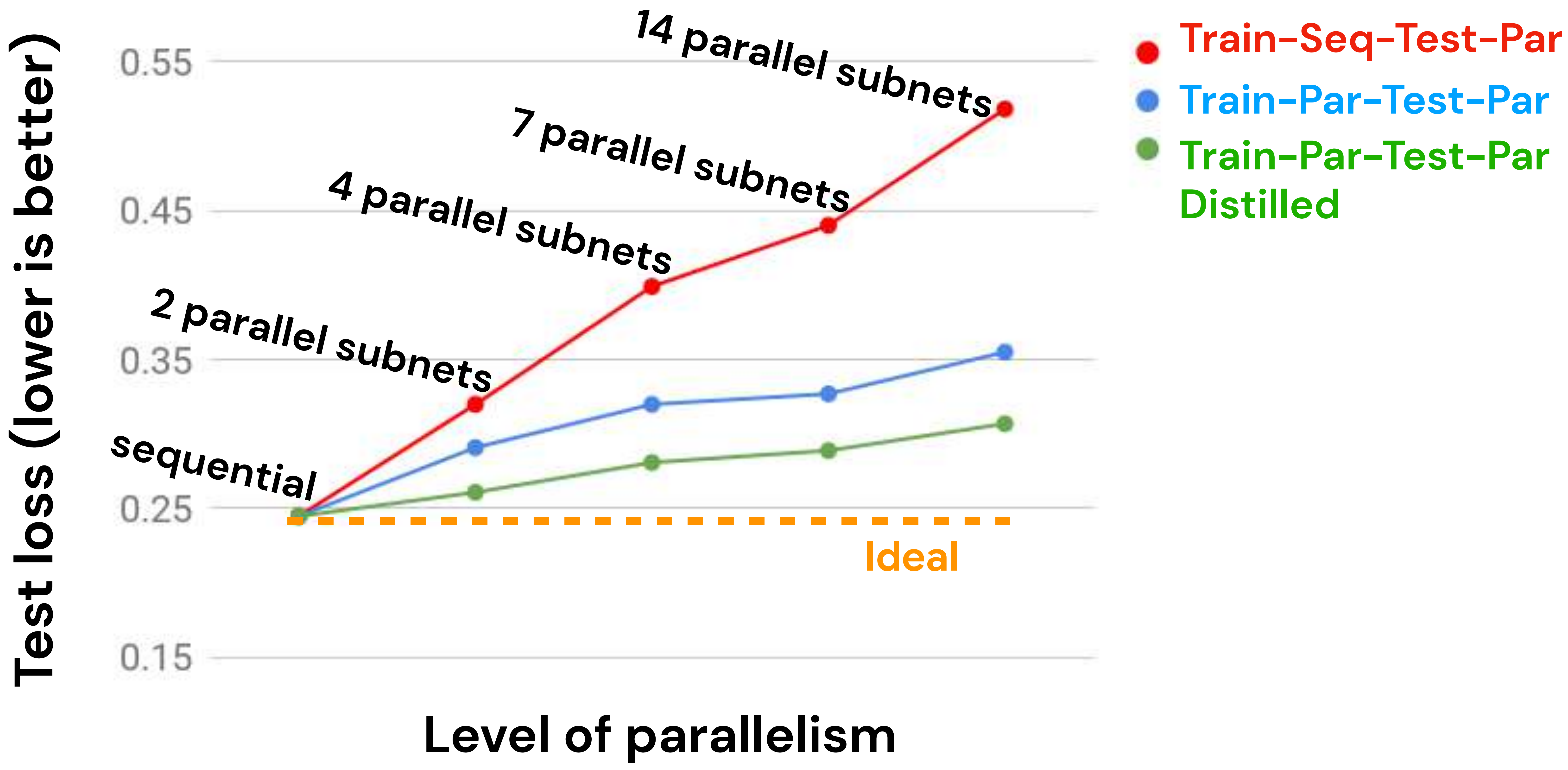
GPU utilisation





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

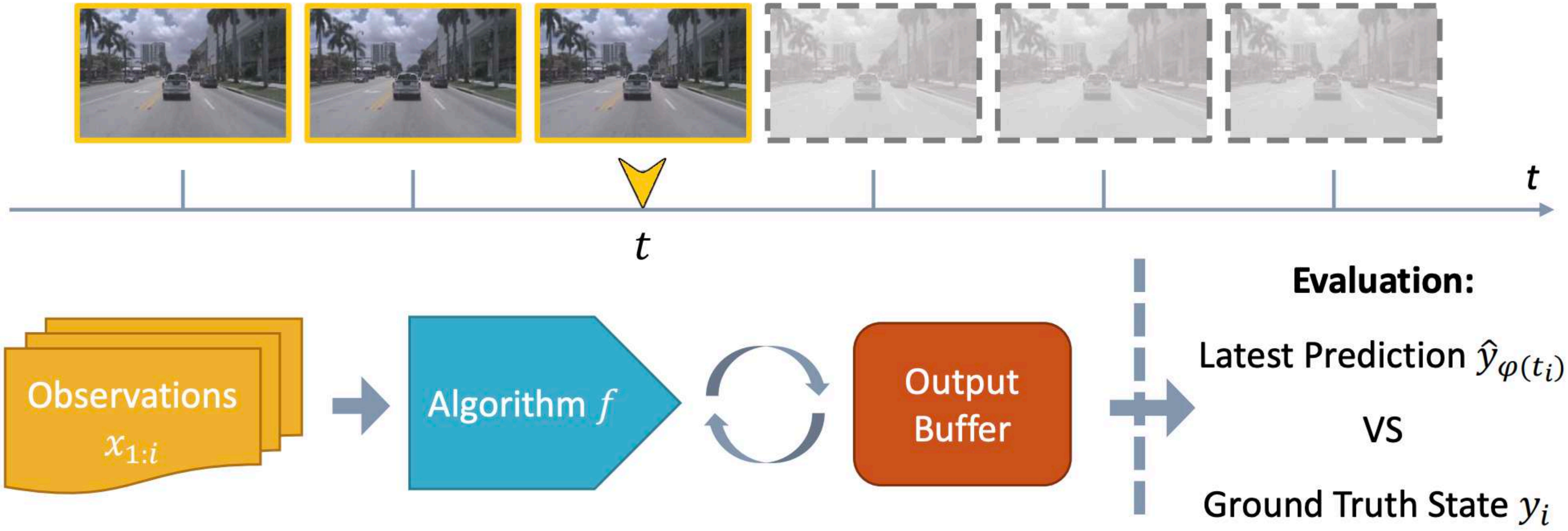
Human keypoint localisation



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

Fair(er) asynchronous evaluation

Towards streaming image understanding, Li et al, ECCV2020

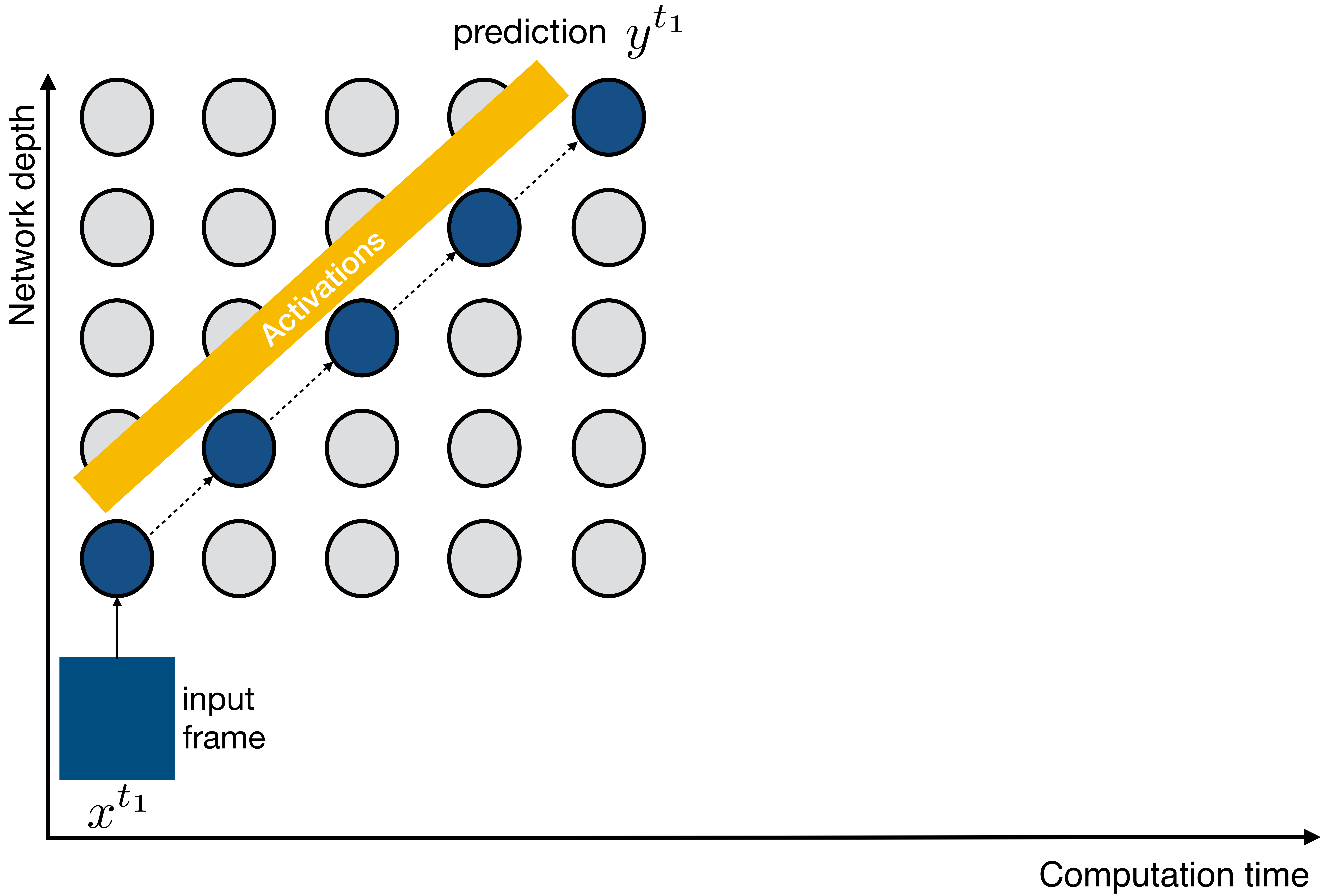


Object detector: offline precision 38.0, streaming average precision 20.3

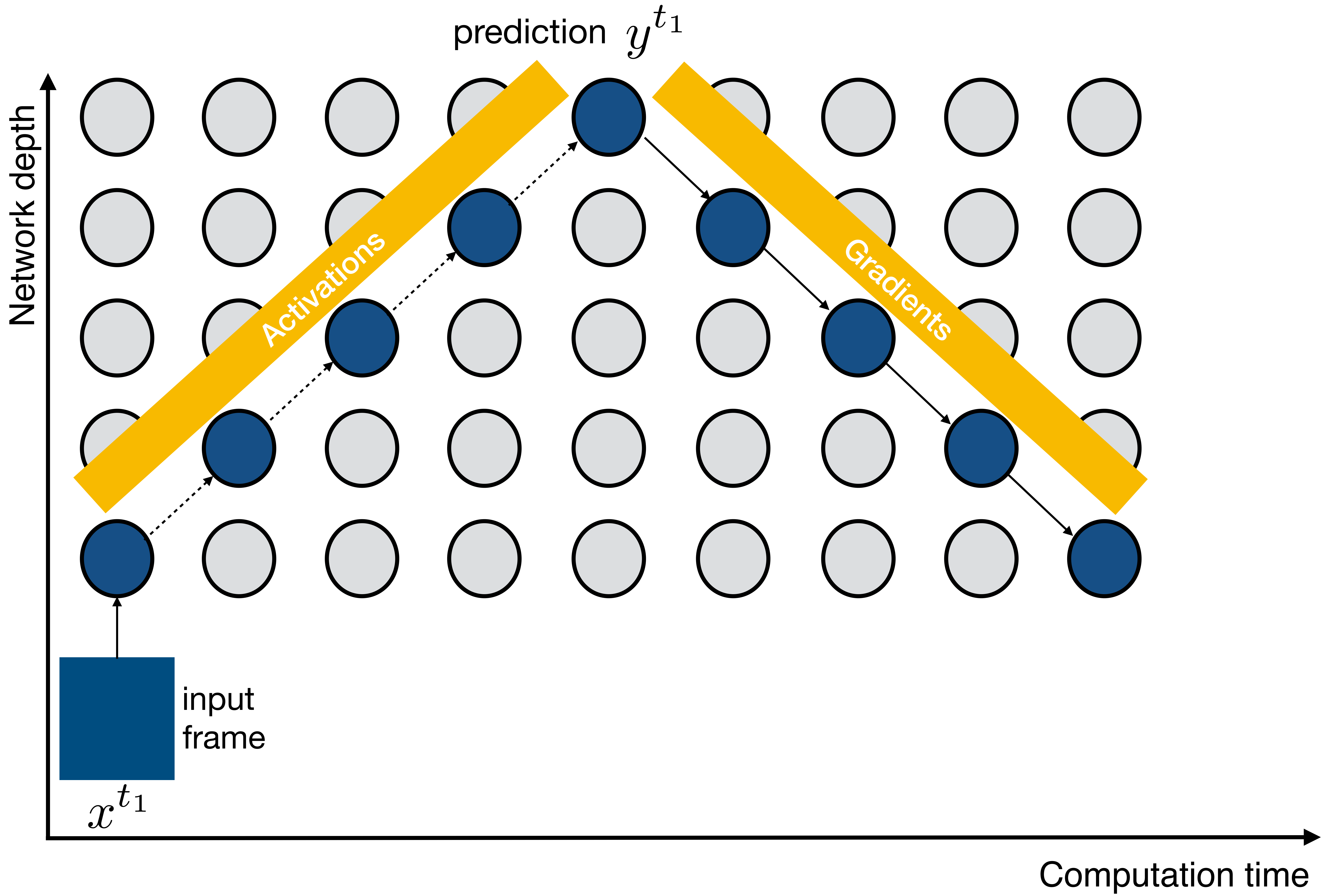
Figure from *Towards streaming image understanding,* Li et al, ECCV2020

B. Depth-parallel training

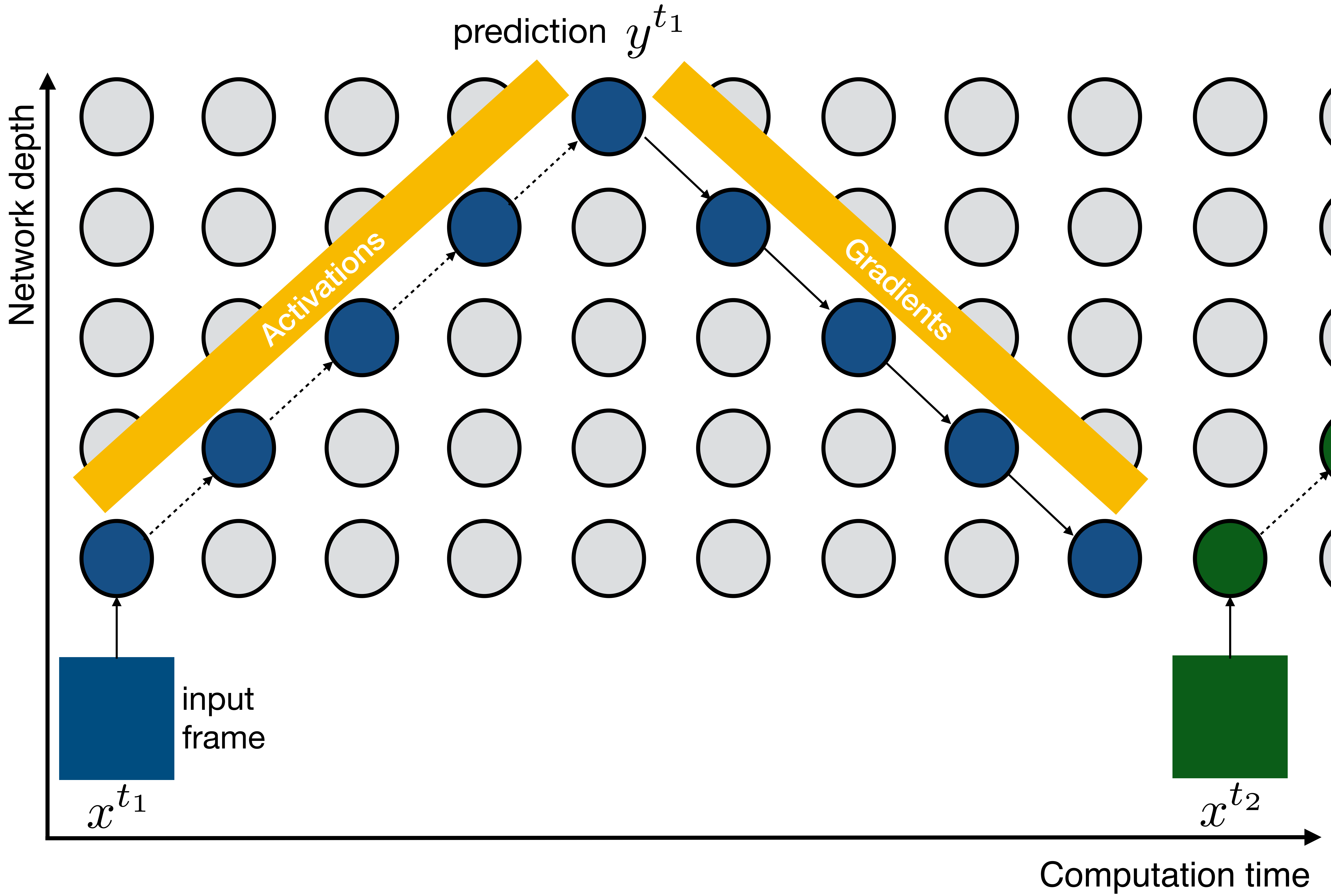
Sequential (frame-by-frame) model



Sequential (frame-by-frame) model

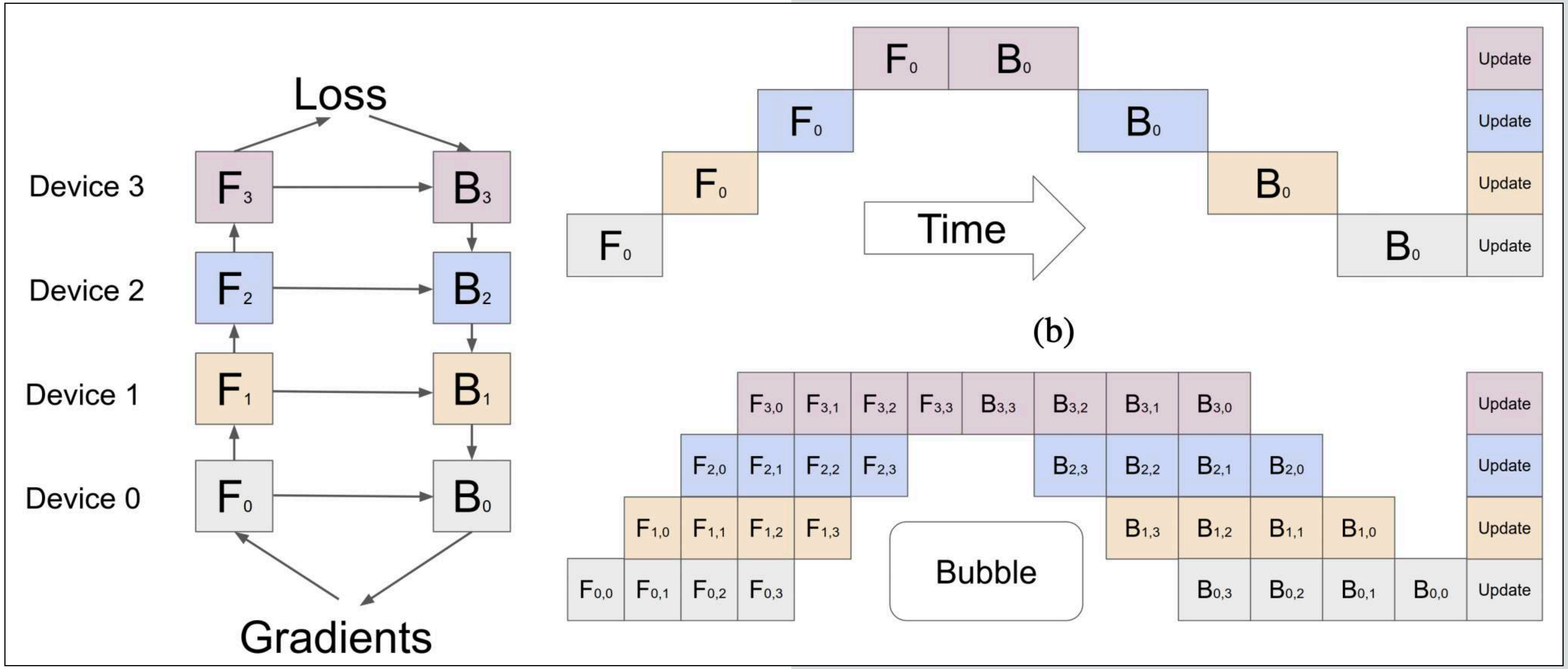


Sequential (frame-by-frame) model



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?

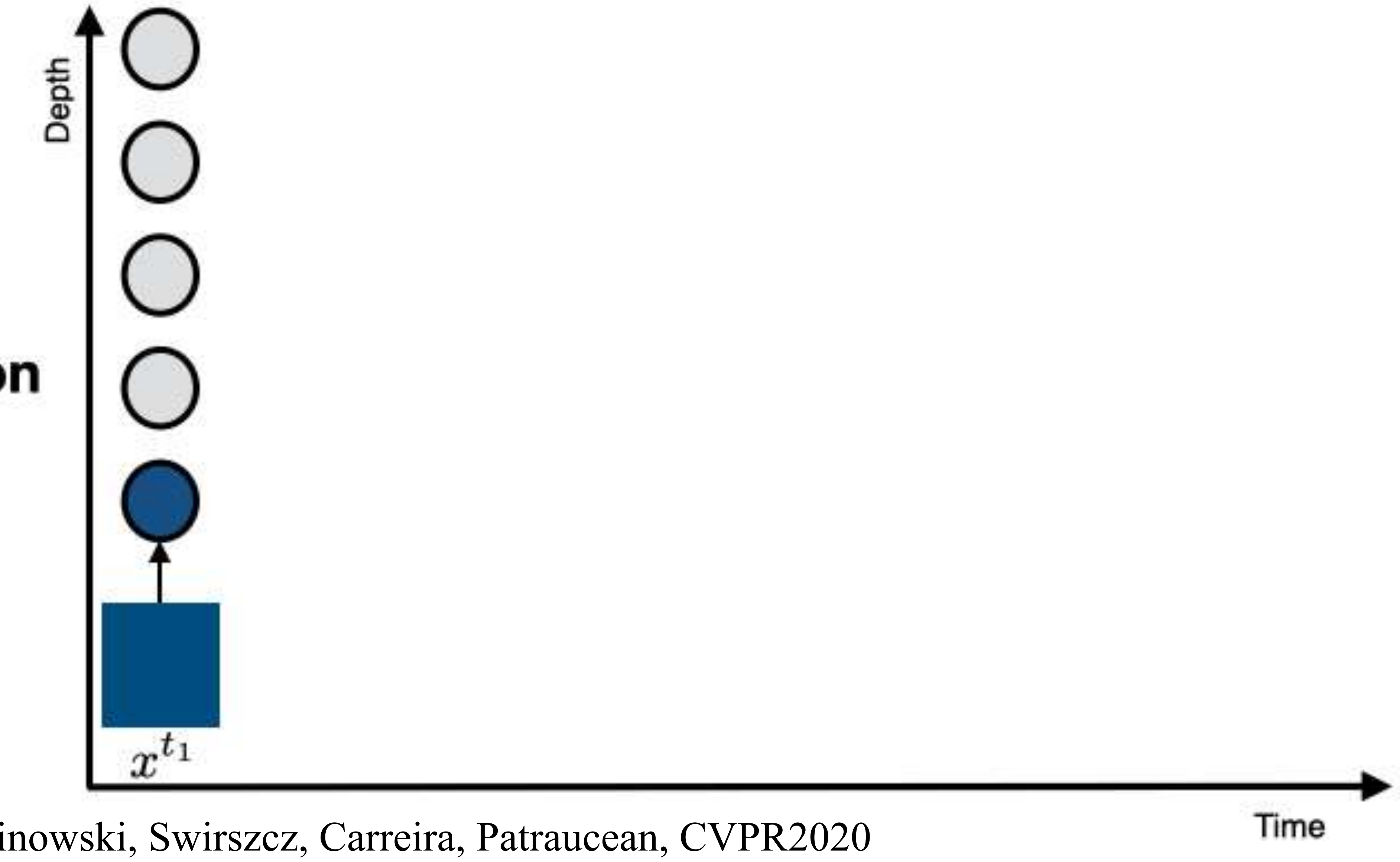
Video from Kinetics600 dataset

Sample consecutive frames extracted from the video

Sideways vs. Backprop

Sideways

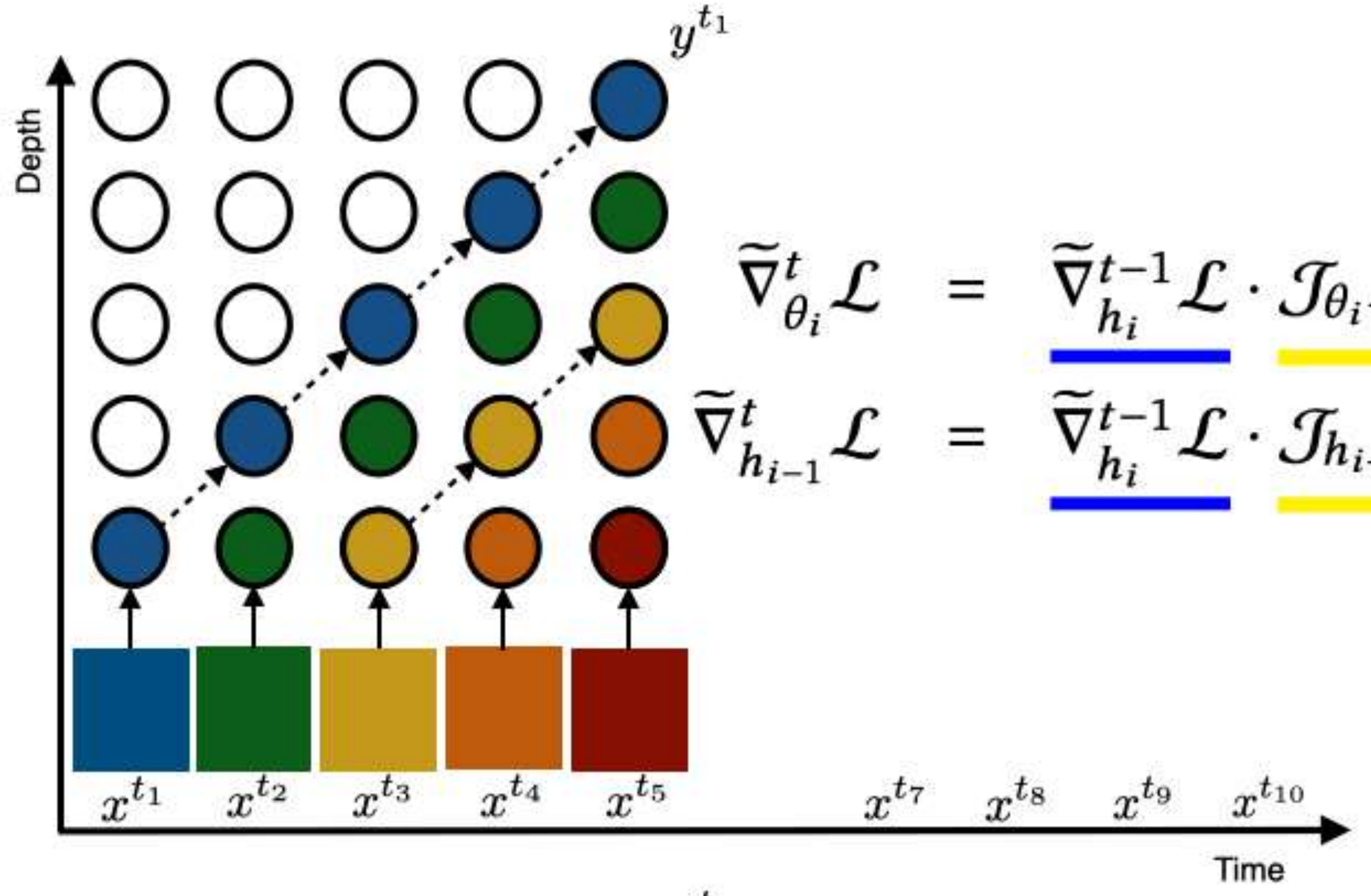
Backpropagation



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

Sideways vs. Backprop

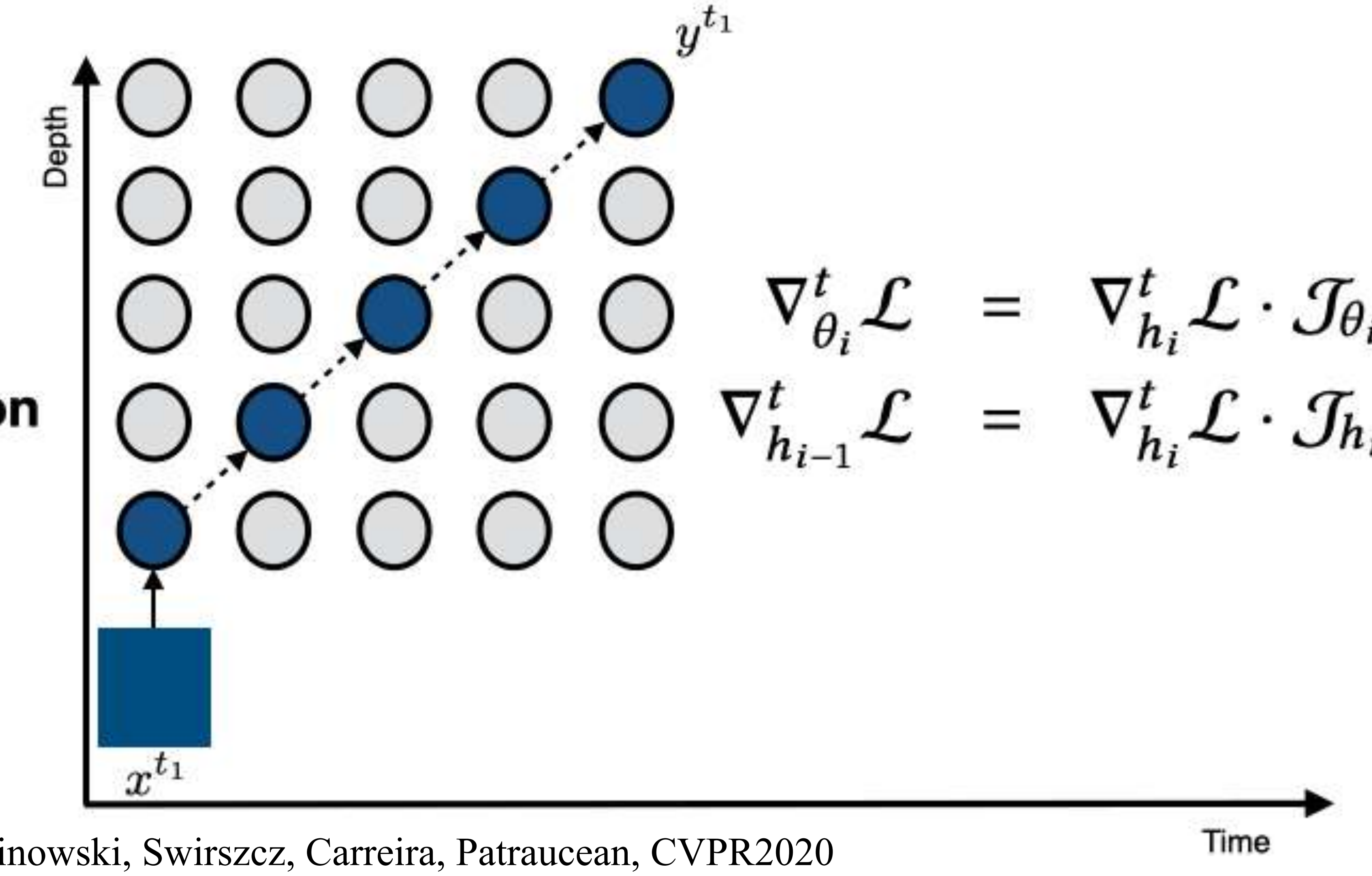
Sideways



$$\widetilde{\nabla}^{t}_{\theta_i}\mathcal{L} = \widetilde{\nabla}^{t-1}_{h_i}\mathcal{L} \cdot \mathcal{J}_{\theta_i} H_i(\boldsymbol{h}^{t-i+1}_{i-1}, \boldsymbol{\theta}_i)$$

$$\widetilde{\nabla}^{t}_{h_{i-1}}\mathcal{L} = \widetilde{\nabla}^{t-1}_{h_i}\mathcal{L} \cdot \mathcal{J}_{h_{i-1}} H_i(\boldsymbol{h}^{t-i+1}_{i-1}, \boldsymbol{\theta}_i)$$

Backpropagation

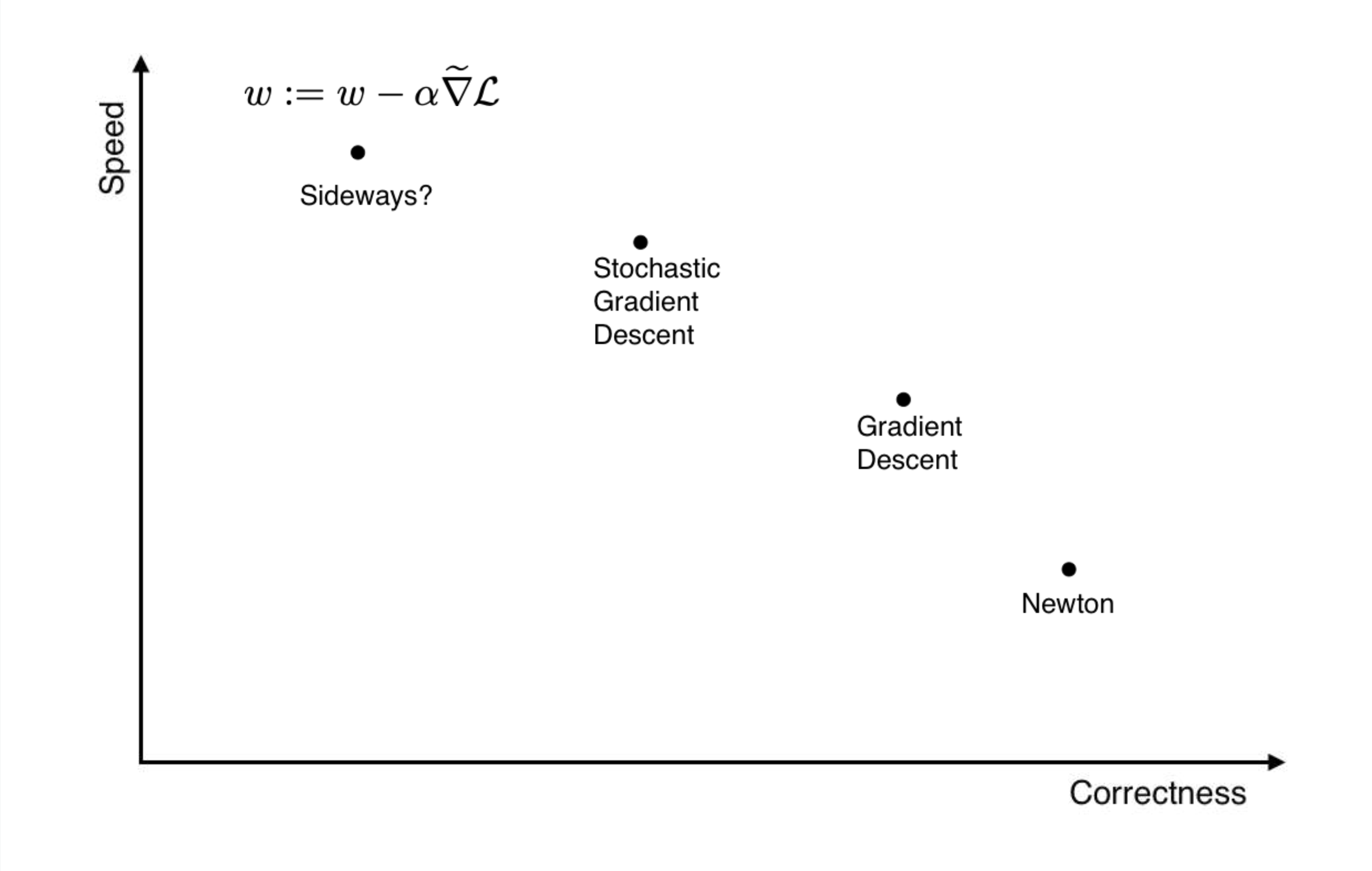


$$\nabla^{t}_{\theta_i}\mathcal{L} = \nabla^{t}_{h_i}\mathcal{L} \cdot \mathcal{J}_{\theta_i} H_i(\boldsymbol{h}^{t}_{i-1}, \boldsymbol{\theta}_i)$$

$$\nabla^{t}_{h_{i-1}}\mathcal{L} = \nabla^{t}_{h_i}\mathcal{L} \cdot \mathcal{J}_{h_{i-1}} H_i(\boldsymbol{h}^{t}_{i-1}, \boldsymbol{\theta}_i)$$

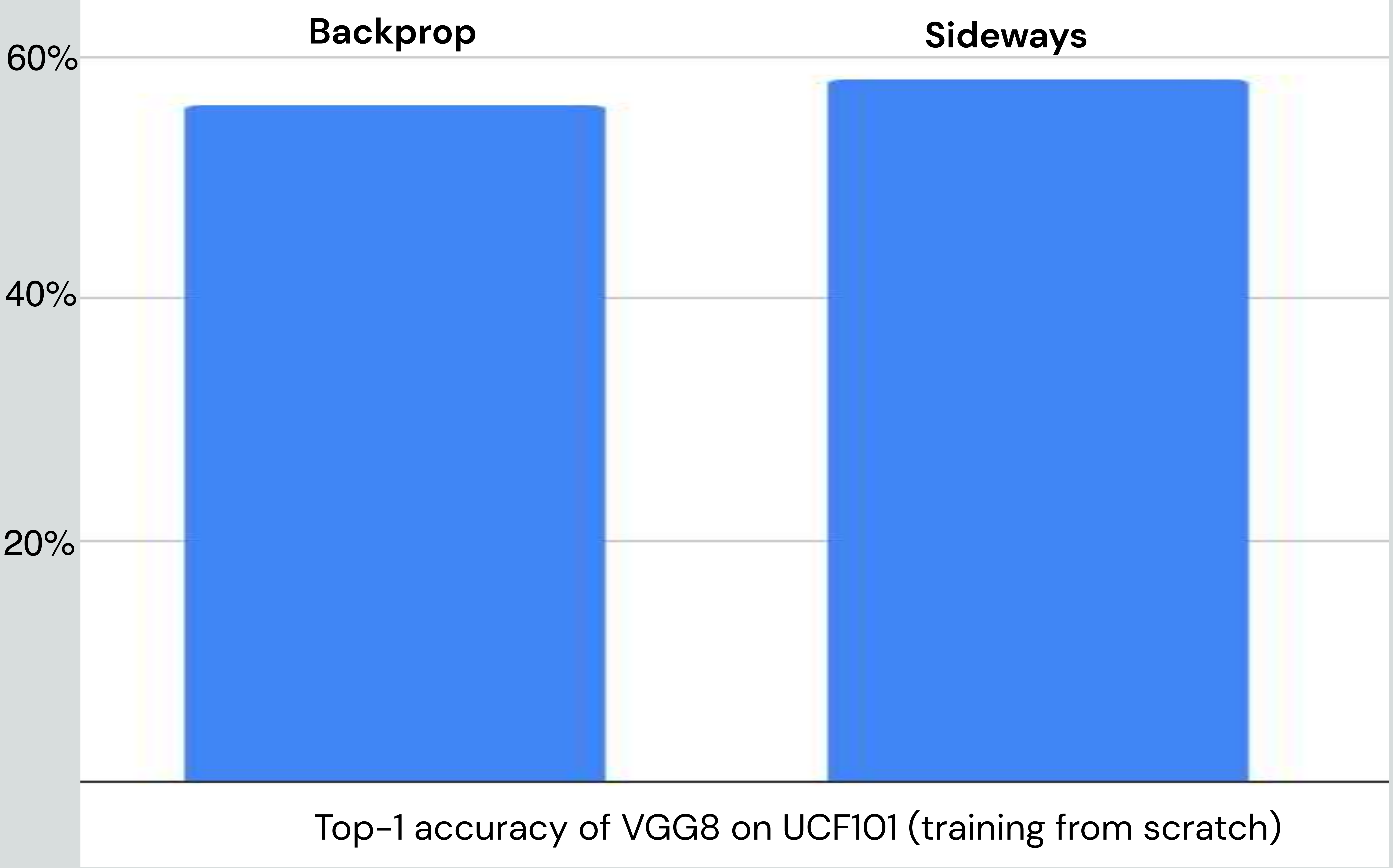
Sideways: Depth-parallel training of video models, Malinowski, Swirszcz, Carreira, Patraucean, CVPR2020

Mathematical correctness of gradients vs. speed



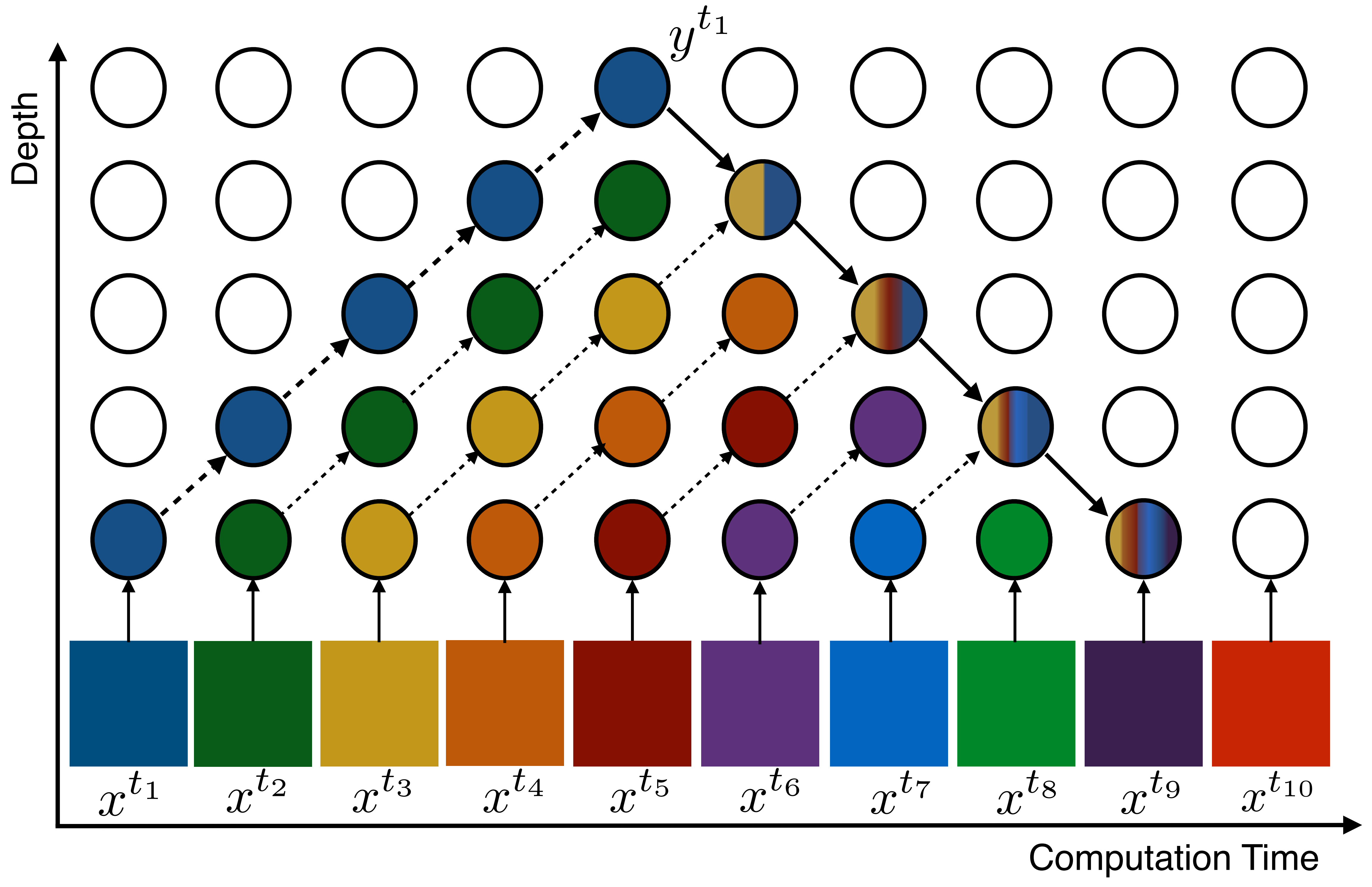
Sideways: Depth-parallel training of video models, Malinowski, Swirszcz, Carreira, Patraucean, CVPR2020

Sideways vs. Backprop

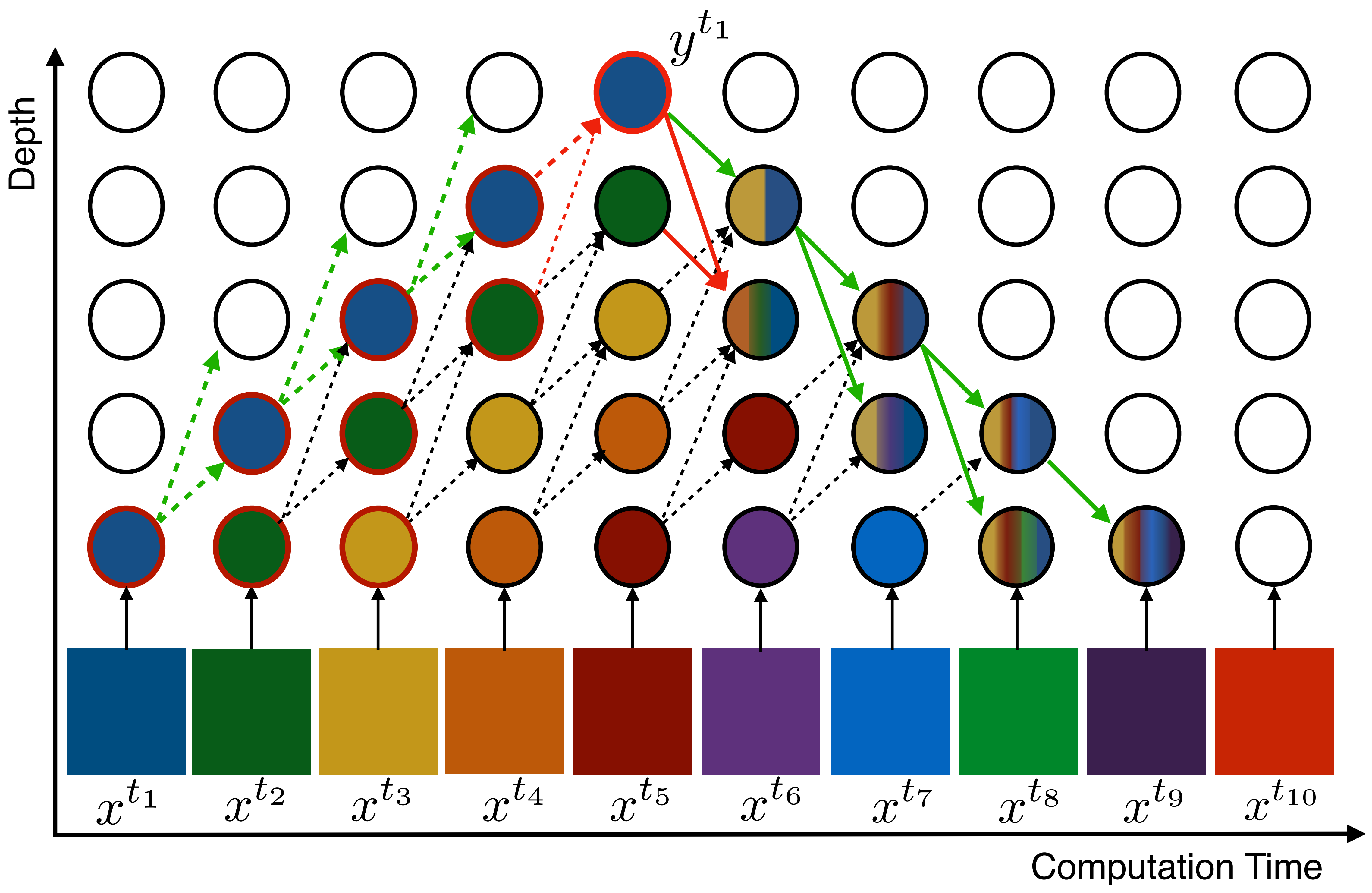


Sideways: Depth-parallel training of video models, Malinowski, Swirszcz, Carreira, Patraucean, CVPR2020

Skip-Sideways: 2D *temporal* model



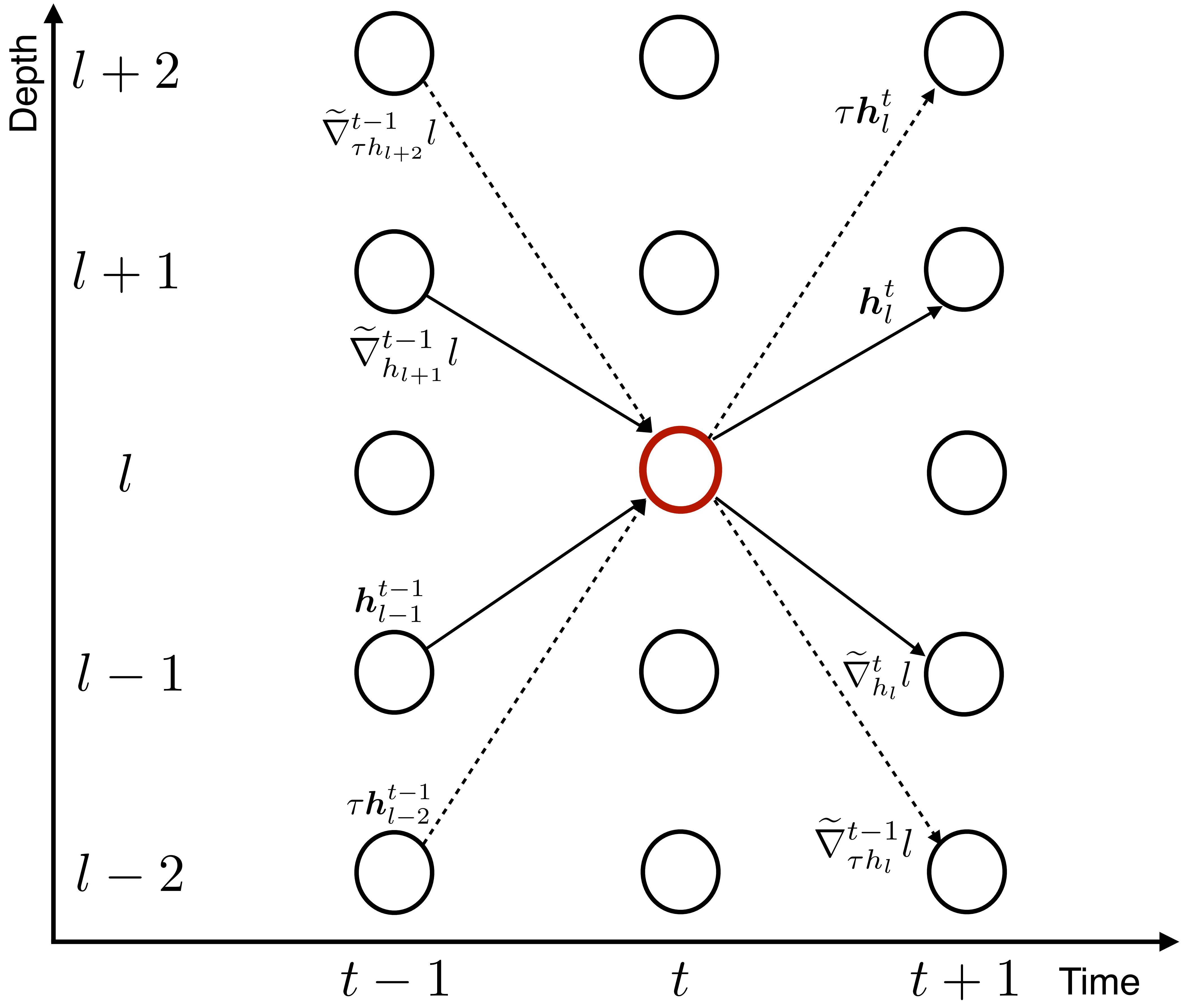
Sideways



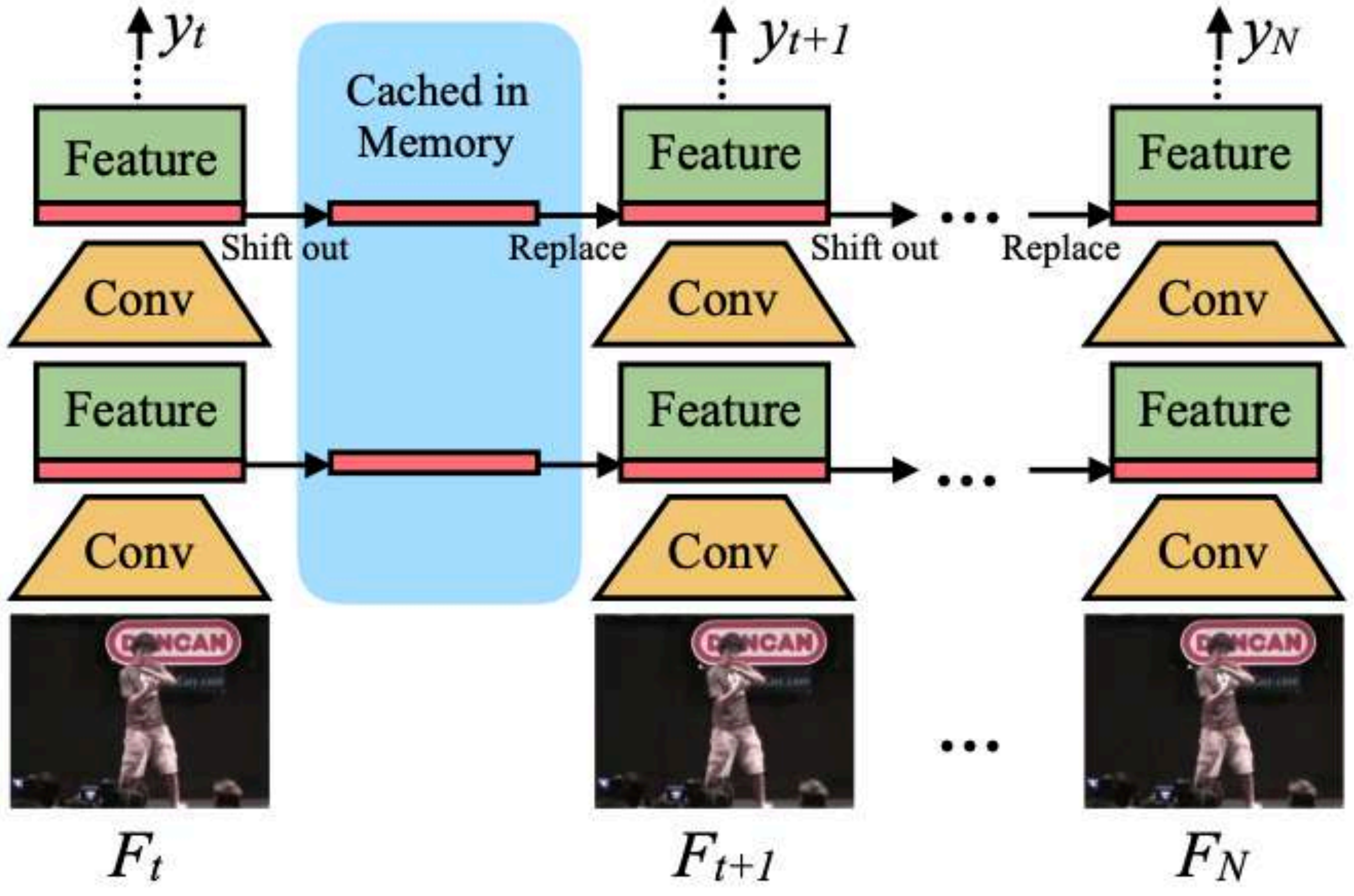
Skip-Sideways

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

Skip-Sideways vs. Temporal shift modules (TSM)



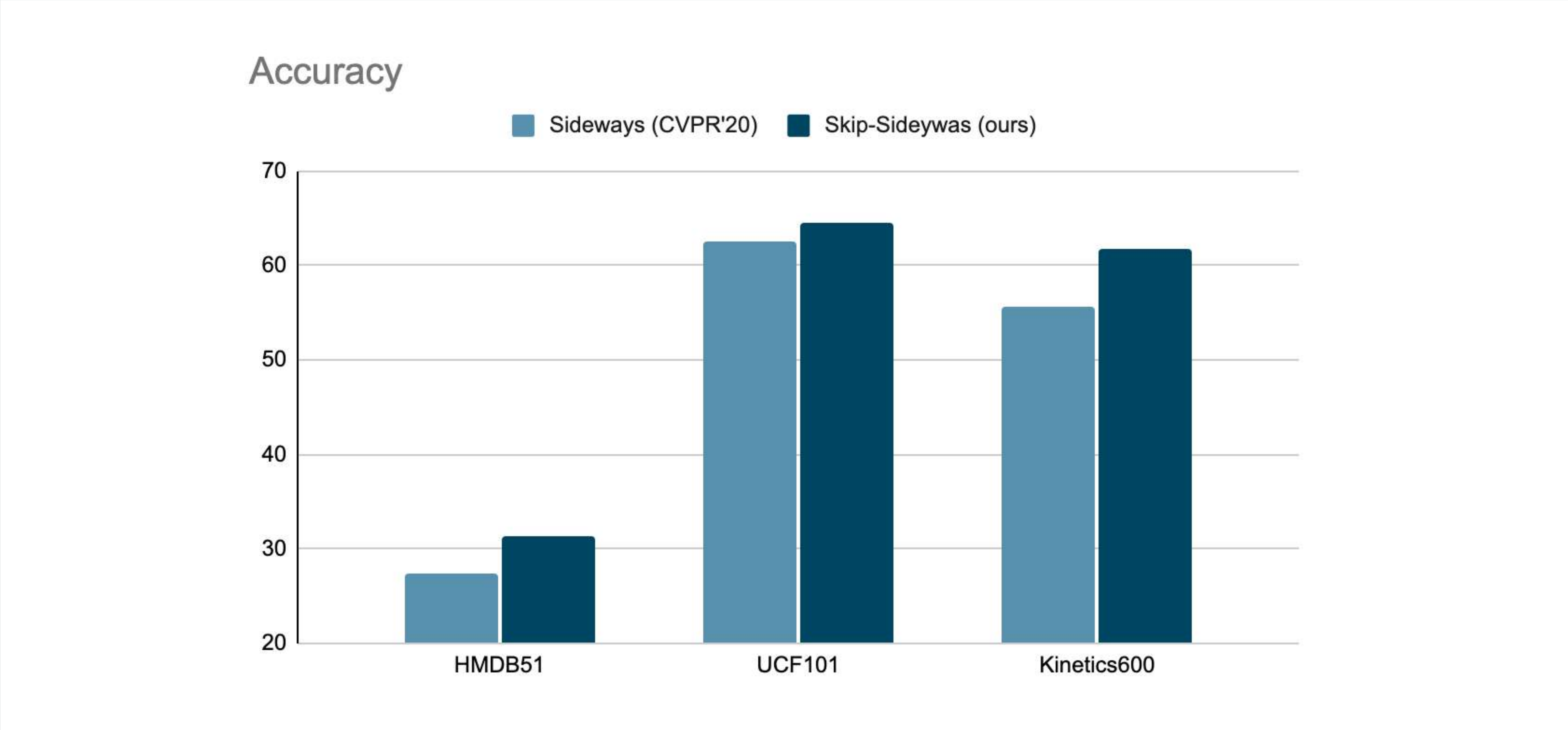
Temporal shift modules



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

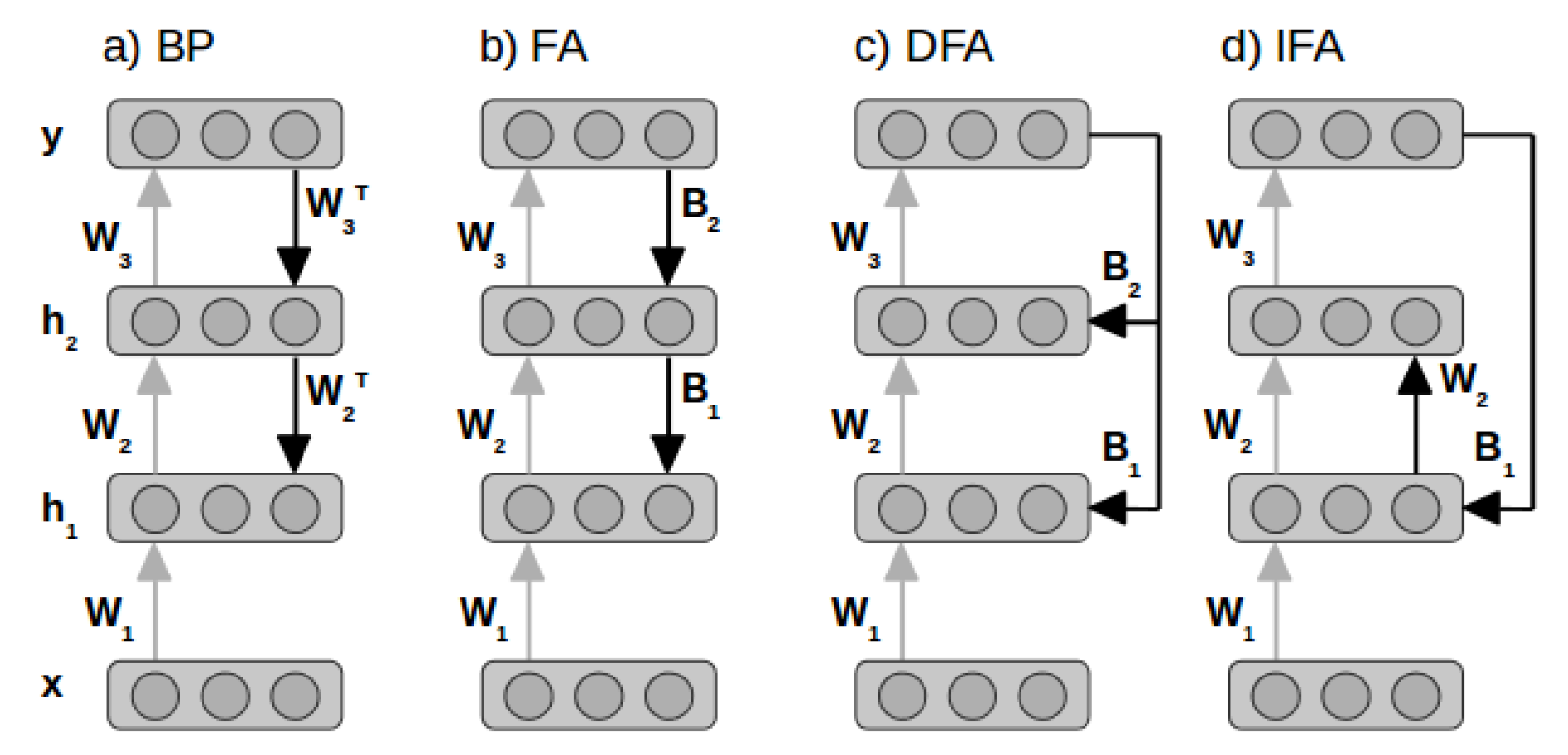


Figure from *Direct Feedback Alignment Provides Learning in Deep Neural Networks,* Nokland, 2016

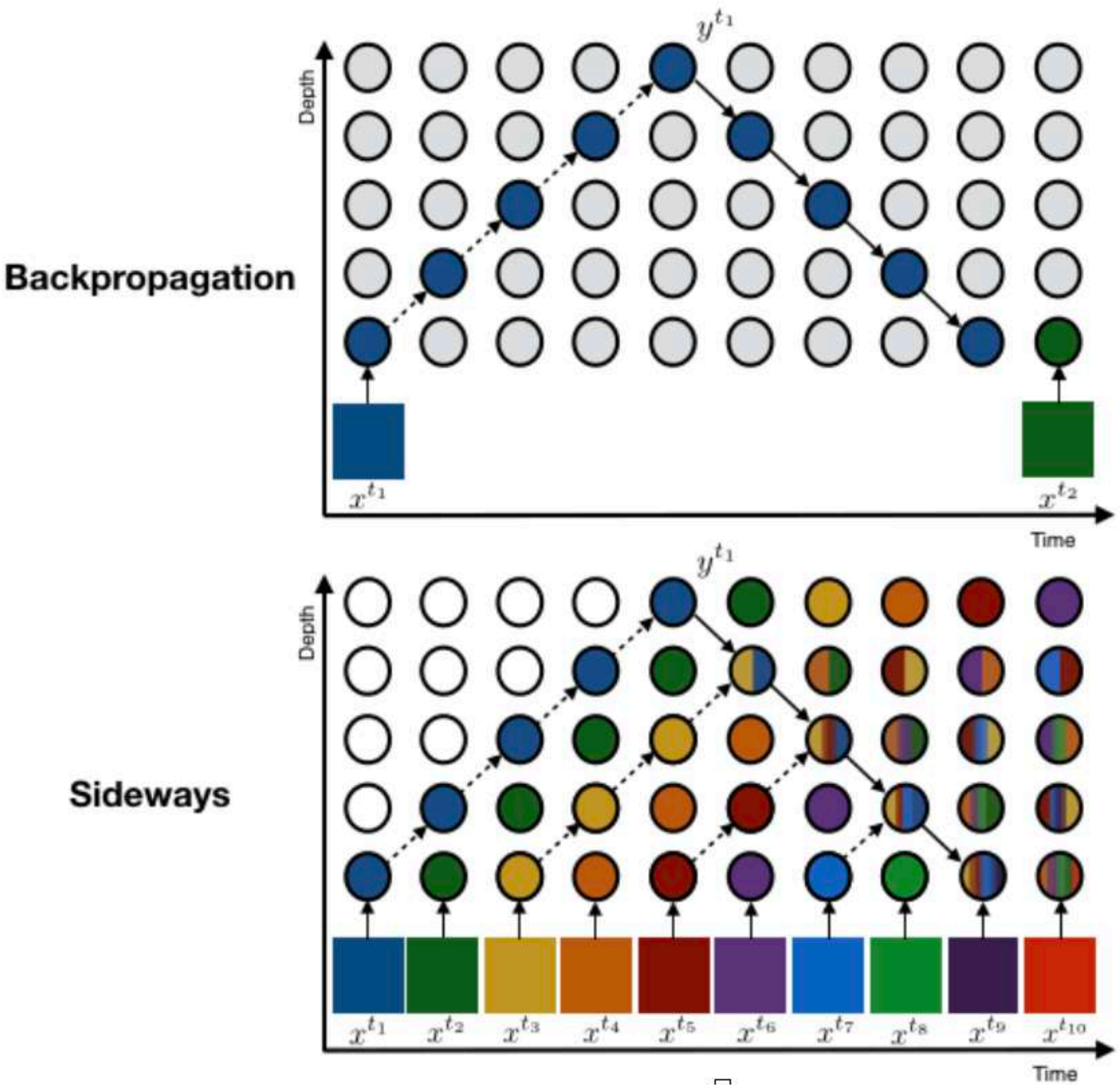
Sideways: more biologically plausible

BP: instantaneous computation (blocking)

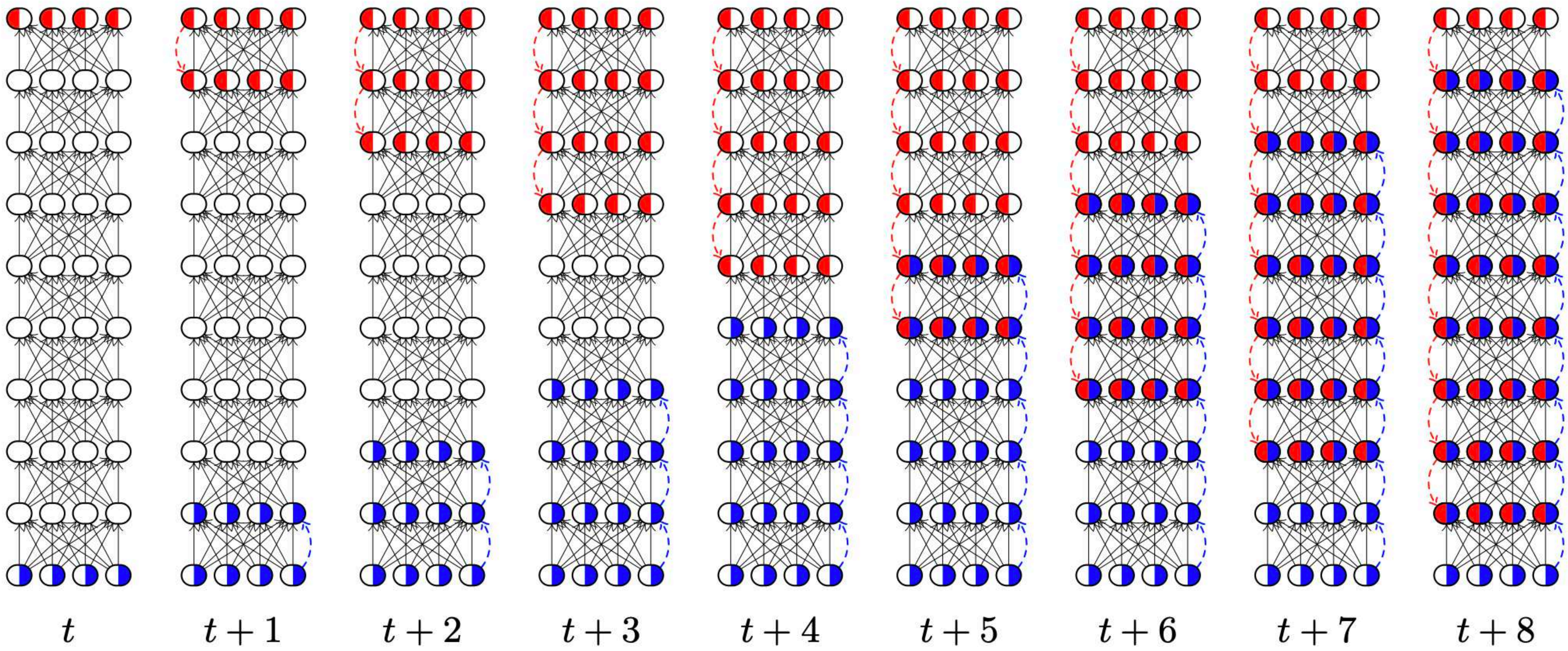
BPTT: sends gradients back in time

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

- Inductive biases in video modelling
  - Architecture side
  - Training objective
  - Training algorithm
- Improve generalisation, data efficiency, and latency for real-time operation

Future work

Towards embodied perception

Self-supervised learning

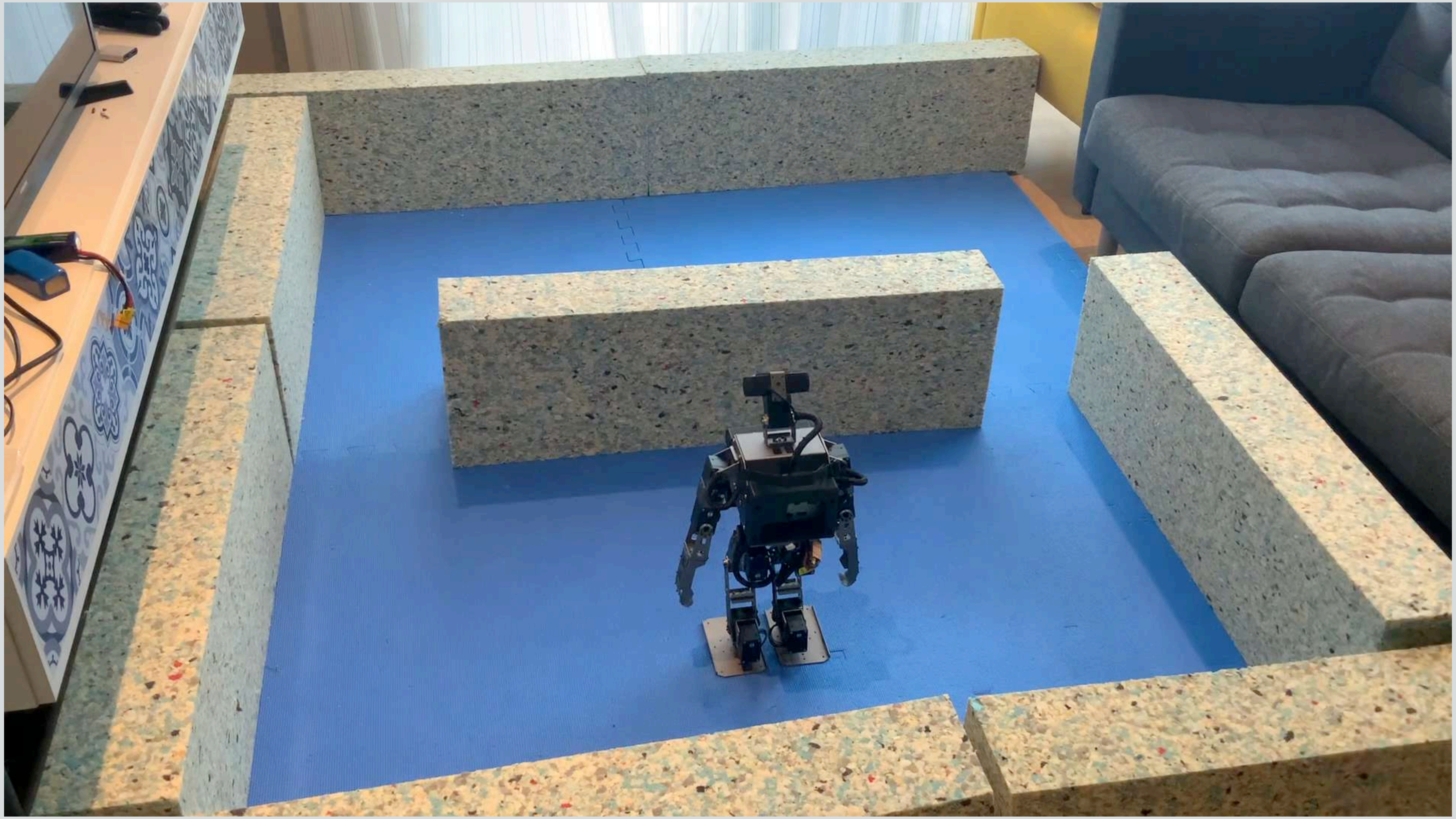
Self-supervised multimodal learning

Active perception Interaction and intervention

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

Future work

Embodied perception - Towards learning in the wild (CORL2021)

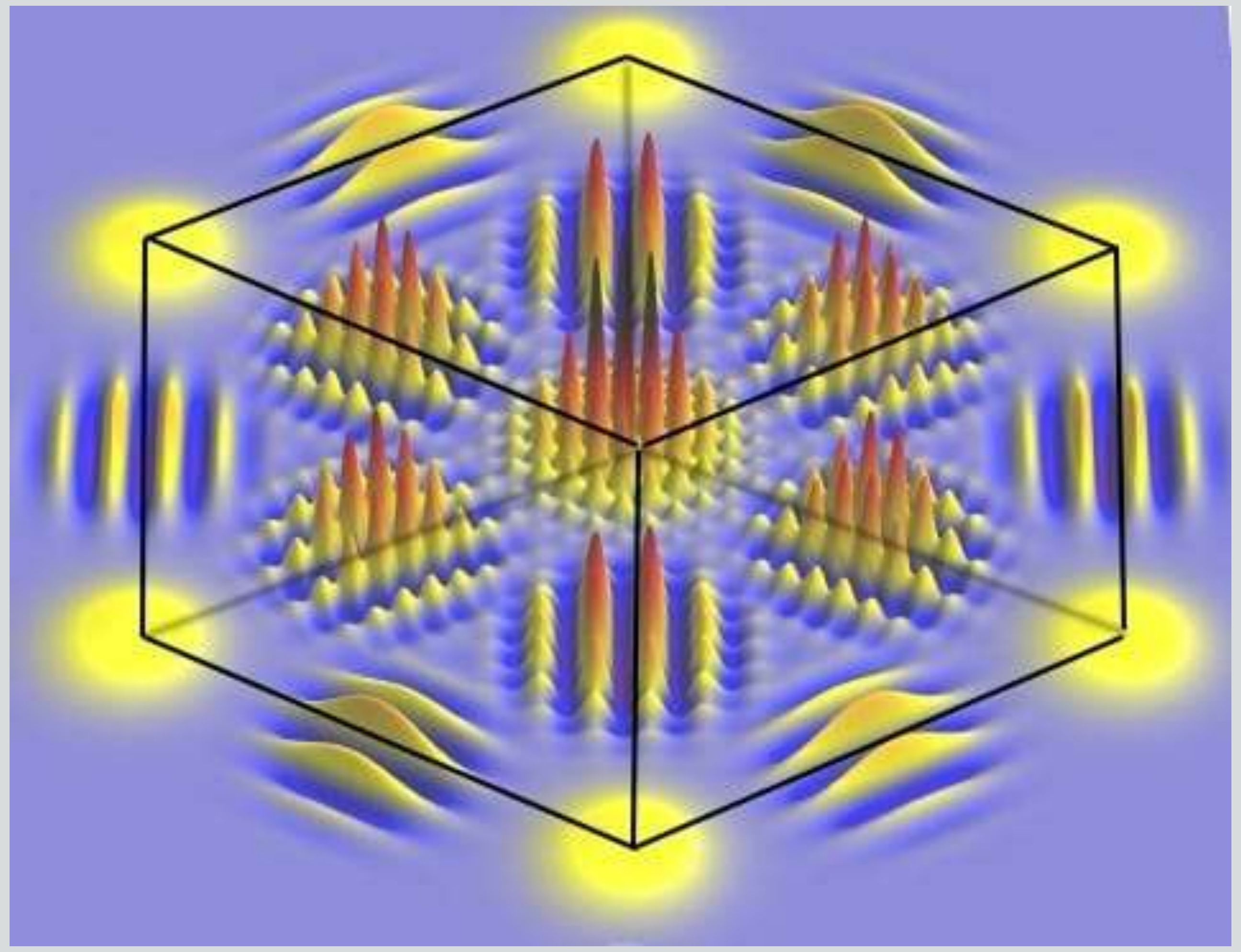


HAGI

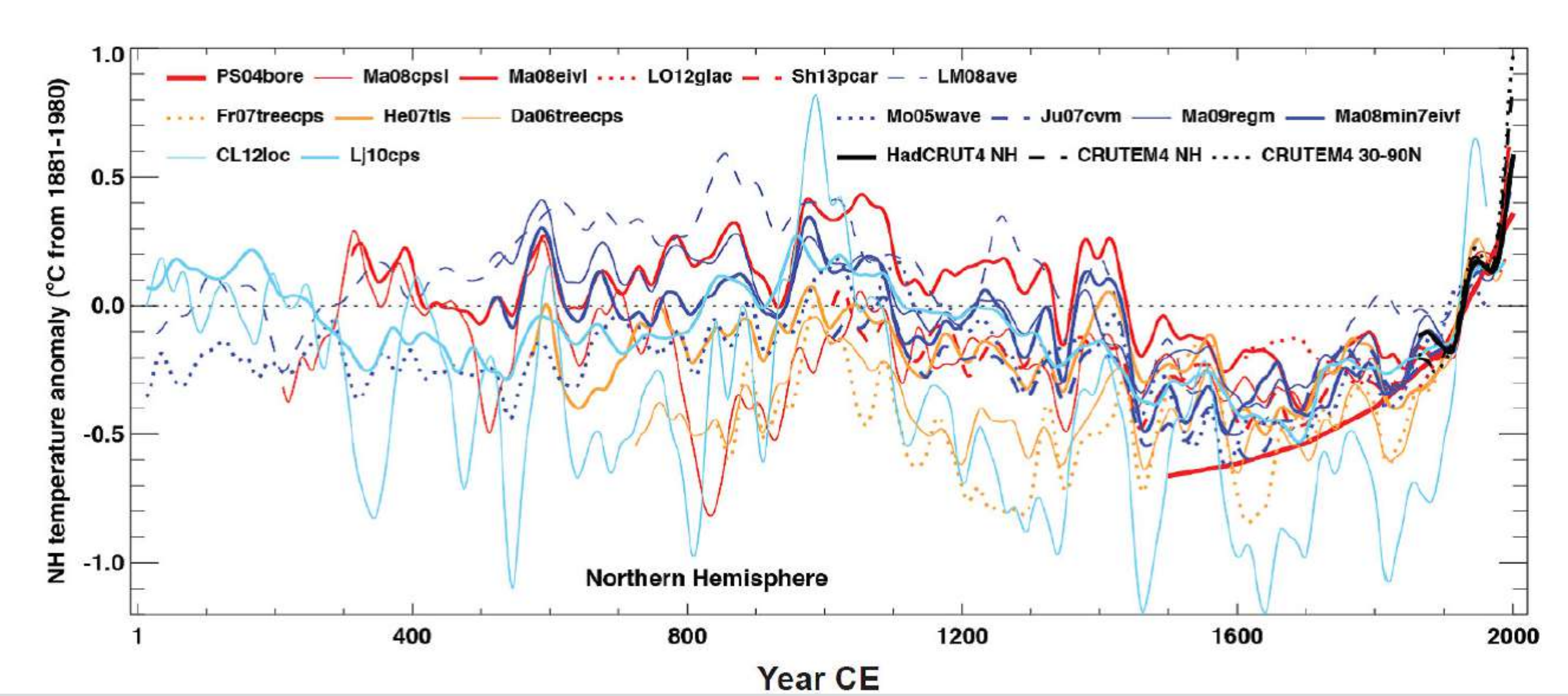
Future work

Generalised perception — going beyond human senses

Genomics



Quantum physics



Geology

Thank you!
Questions?

Viorica Pătrăucean, viorica@google.com