Learning from one continuous stream

João Carreira

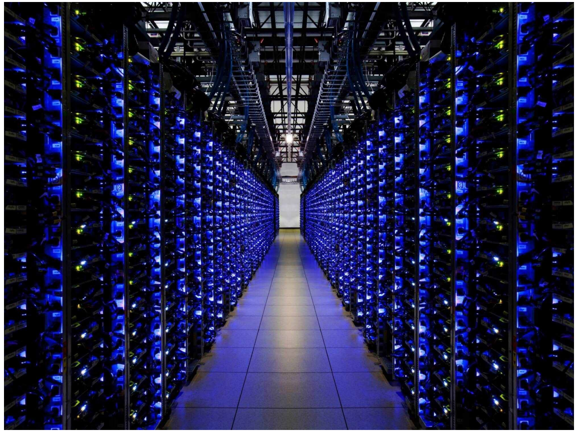
Time is precious: self-supervised learning beyond images tutorial – ECCV 2024 Milan
30th of September 2024

Google DeepMind

Human learning… just happens

Contrast to powerful modern AI systems

GPUs



Data – the Internet

Data cleanup team

GPT-4 Technical Report

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain et al. (181 additional authors not shown)

Big difference between regimes

From: *This Hand Is My Hand: A Probabilistic Approach to Hand Disambiguation in Egocentric Video, Lee et al*

- Data arrives sequentially
- Highly multimodal (2 eyes, ears, proprioception, etc)
- Single very long (boring) stream

- All data available in parallel
- Images / mono-video + text (audio)
- Many diverse images / videos

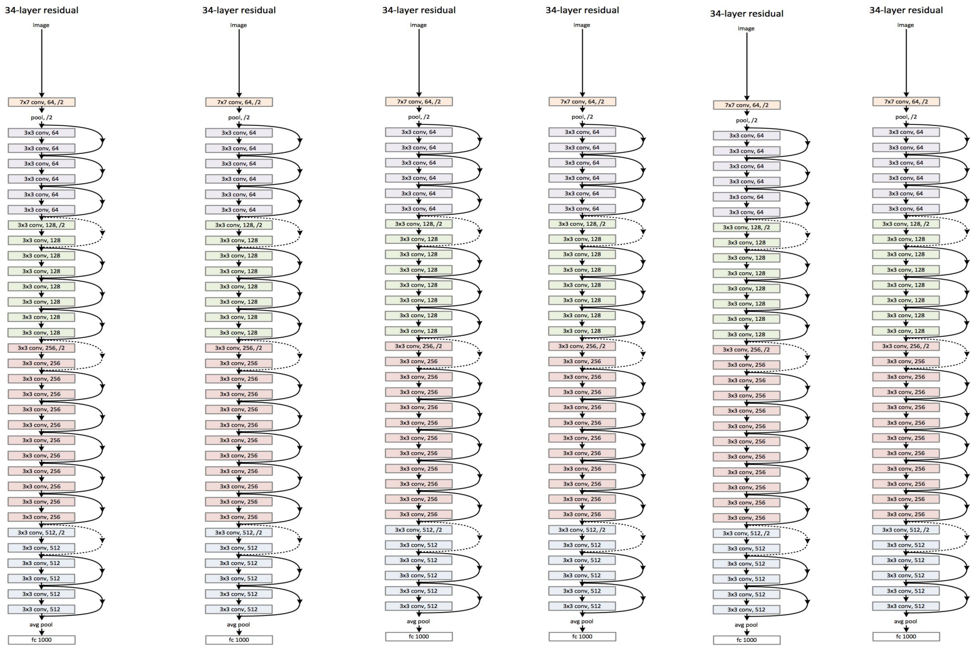
Problem #1 – how to use modern highly parallel hardware ?

- **Data arrives sequentially (batch size 1)**
- Highly multimodal (2 eyes, ears, proprioception, etc)
- Single very long (boring) stream

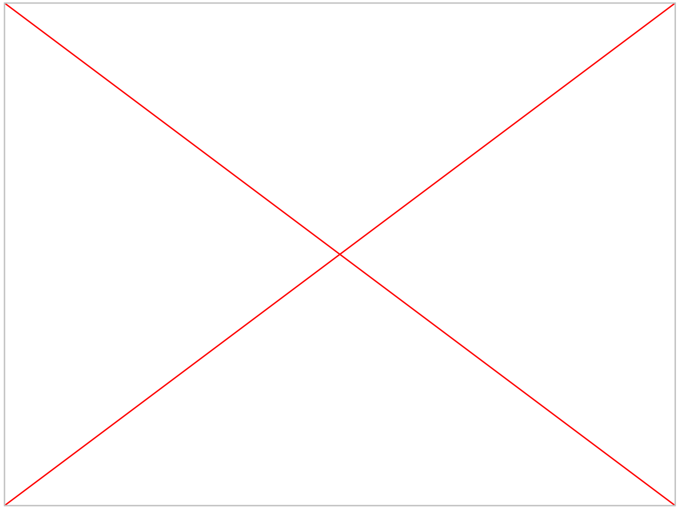
- **All data available in parallel**
- Images / mono-video + text (audio)
- Many diverse images / videos

Efficient use of computational resources + batch size 1 ?

Deep models at 25fps



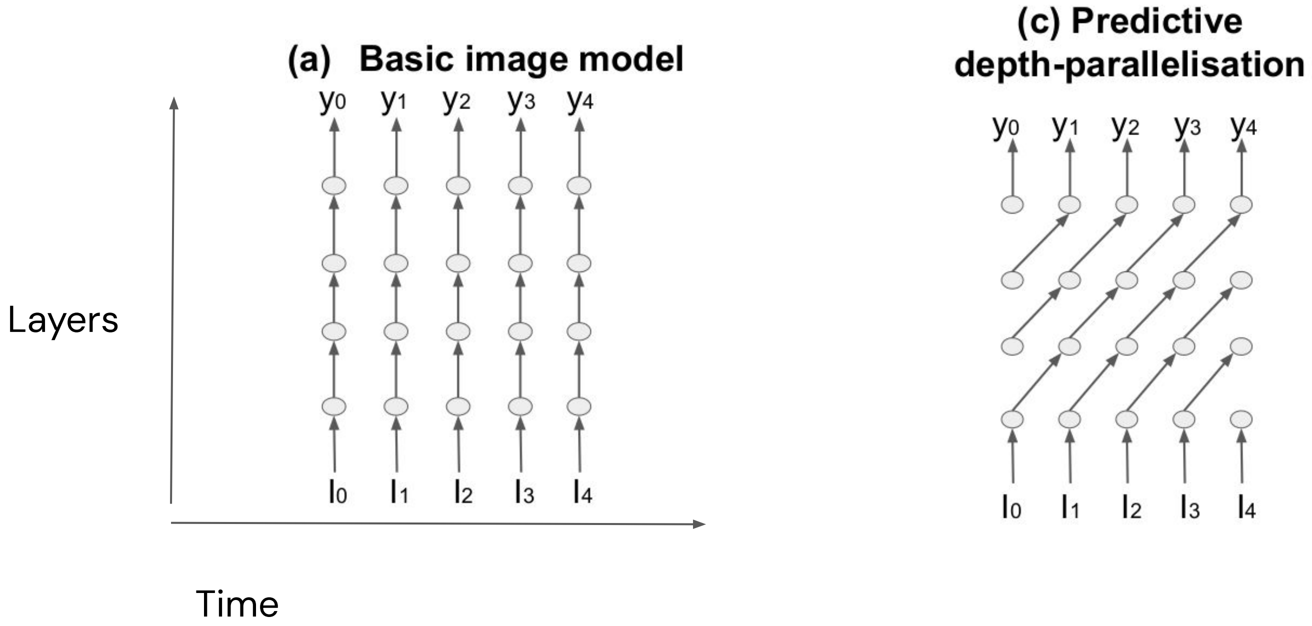
Frame 1 Frame 2 Frame k



Massively Parallel Video Networks

ECCV 2018

João Carreira[†,1], Viorica Pătrăucean[†,1], Laurent Mazare[1],
Andrew Zisserman[1,2], Simon Osindero[1]



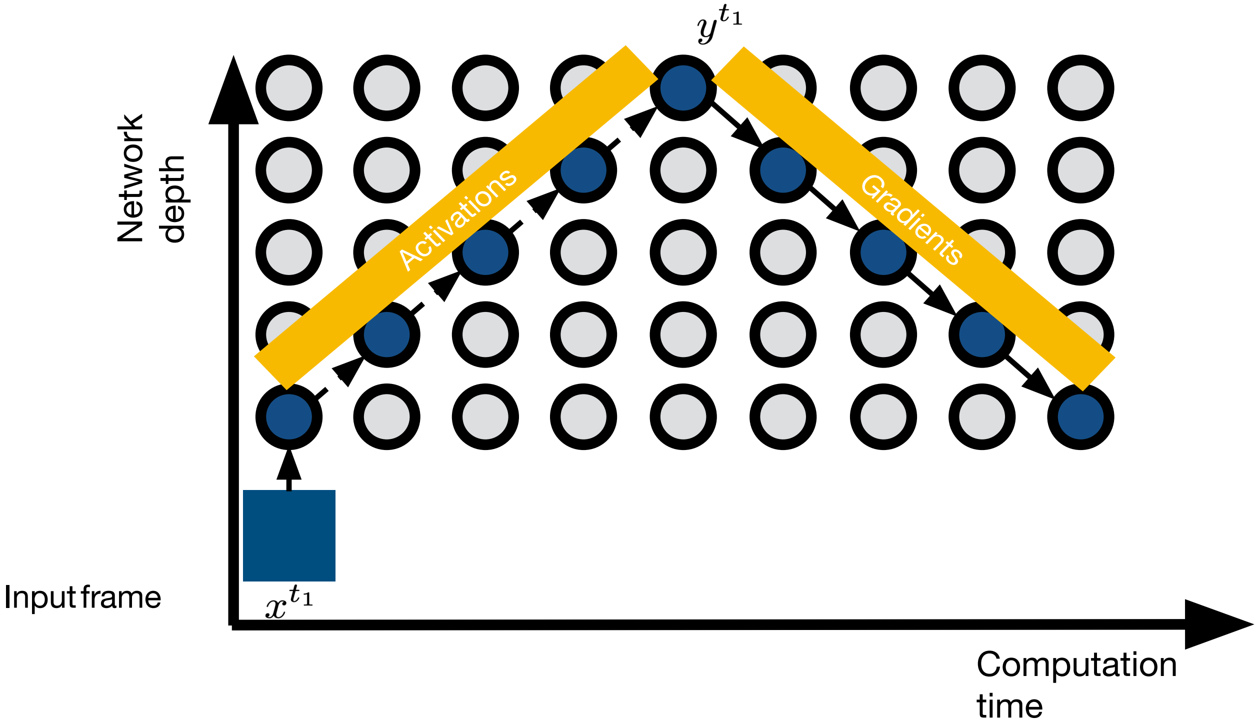

Original video
#Subnets: 14, speedup: 7.4x
#Subnets: 7, speedup: 6.1x
#Subnets: 4, speedup: 5.2x
#Subnets: 2, speedup: 4.0x
#Subnets: 1, speedup: 3.4x

Sideways: Depth-Parallel Training of Video Models

CVPR 2020

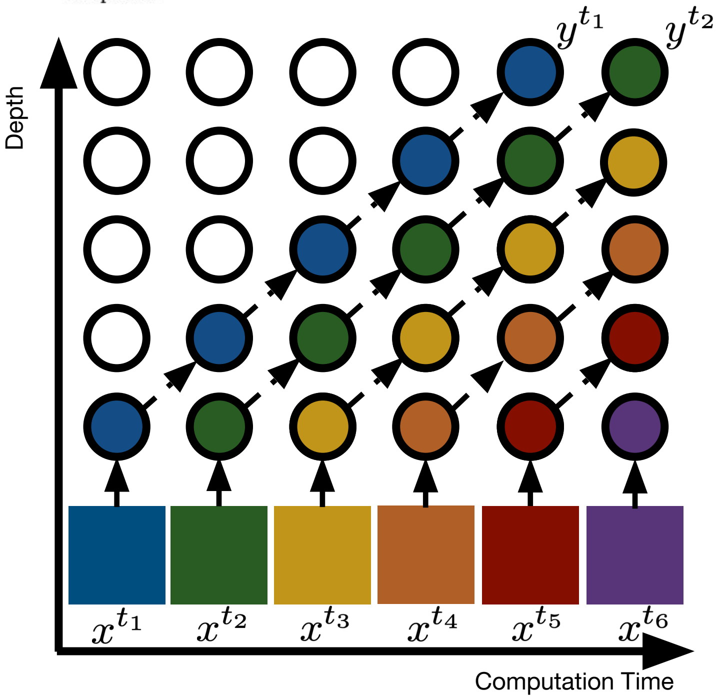
Mateusz Malinowski[1], Grzegorz Świrszcz[1], João Carreira[1] and Viorica Pătrăucean[1]
[1]DeepMind

- Expanding idea from inference to both inference+training

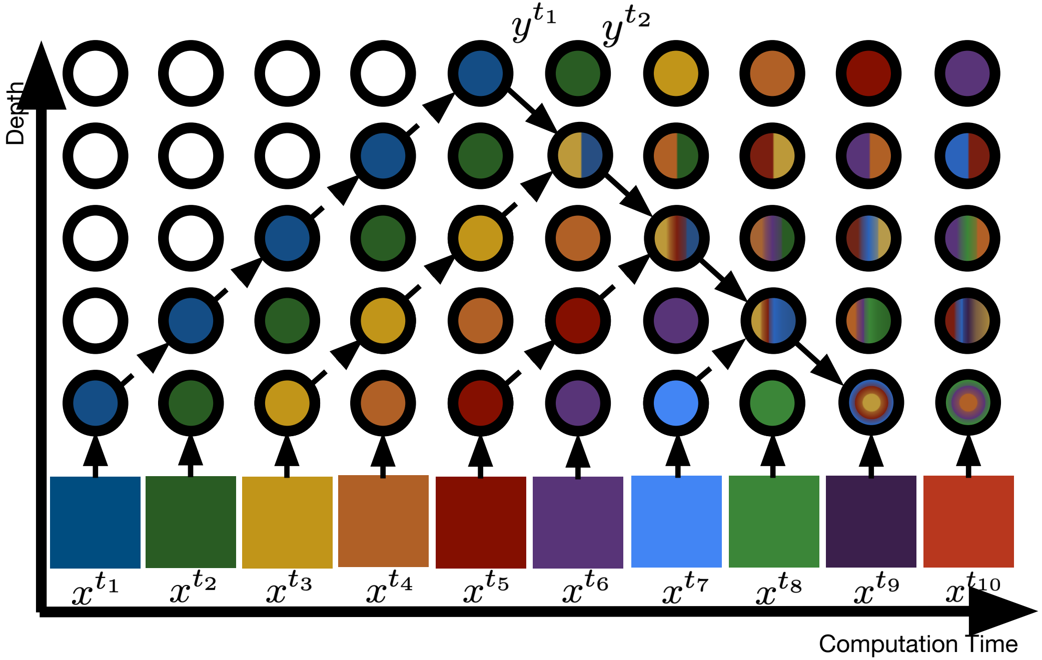


Sideways: Depth-Parallel Training of Video Models

Mateusz Malinowski[1], **Grzegorz Świrszcz**[1], **João Carreira**[1] **and Viorica Pătrăucean**[1]
[1]DeepMind



Massively Parallel Video Models
(ECCV'18)

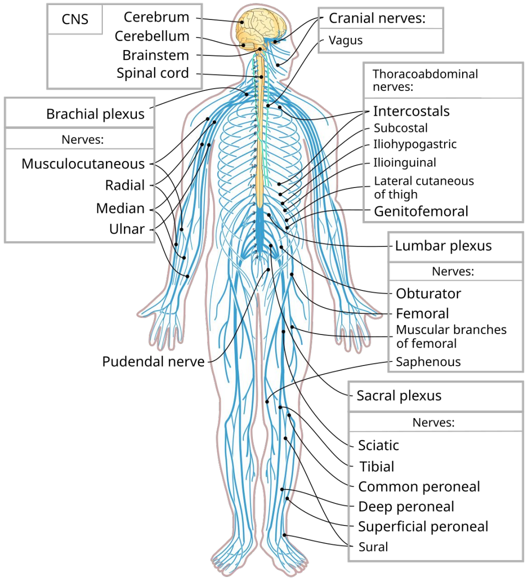


Sideways: Depth-Parallel Training of Video Models
(CVPR'20)

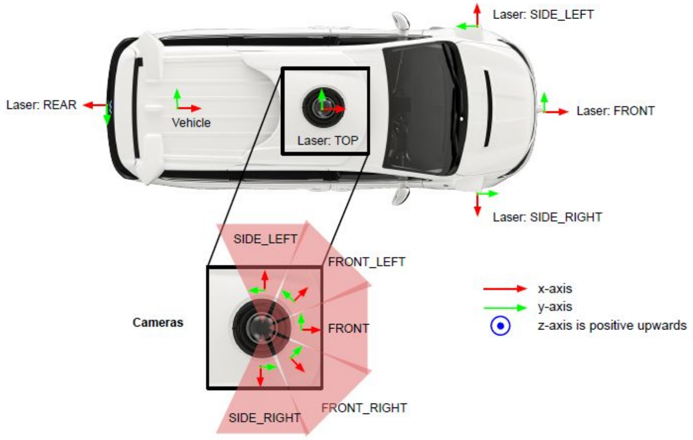
Problem #2 – integrating information from many modalities + time

- Data arrives sequentially (batch size 1)
- **Highly multimodal (2 eyes, ears, proprioception, etc)**
- Single very long (boring) stream

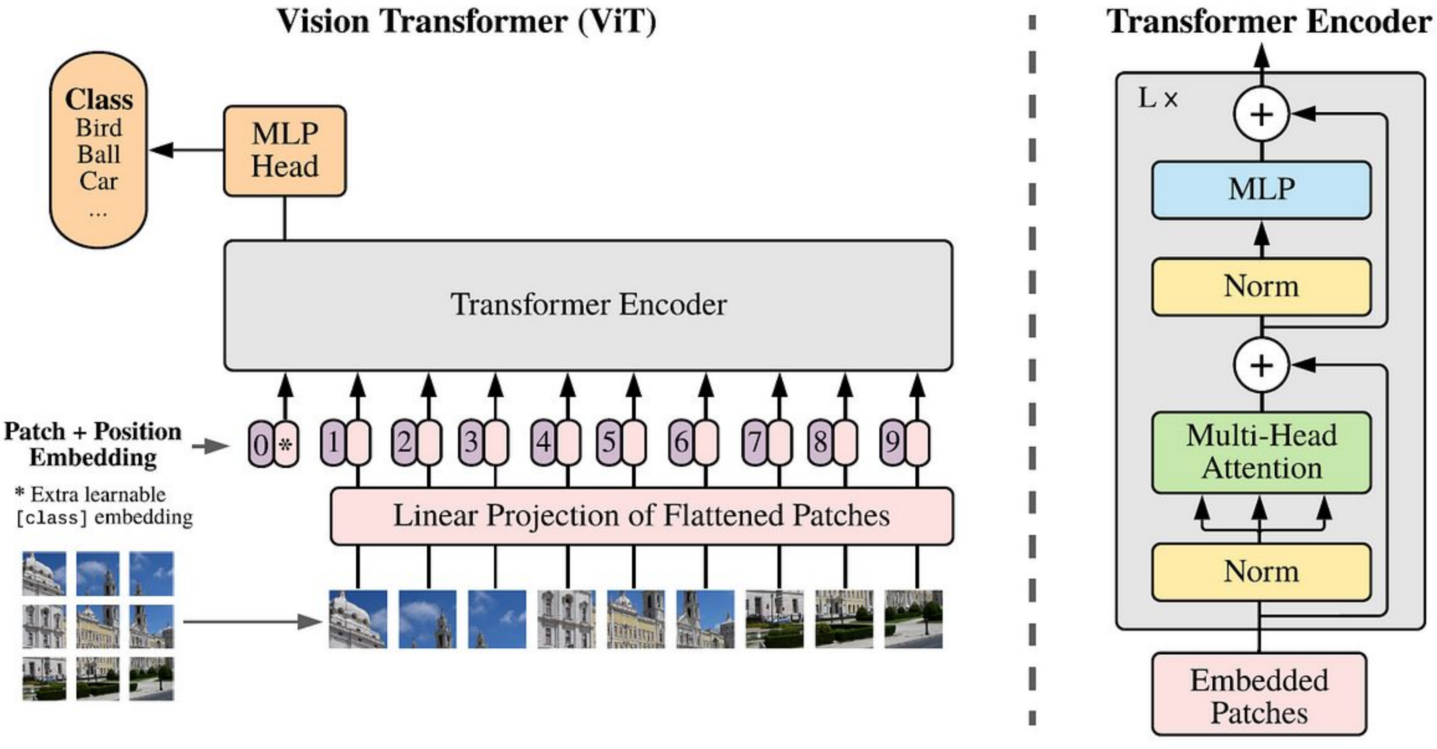
- All data available in parallel
- **Image / video + text (audio)**
- Many diverse images / videos



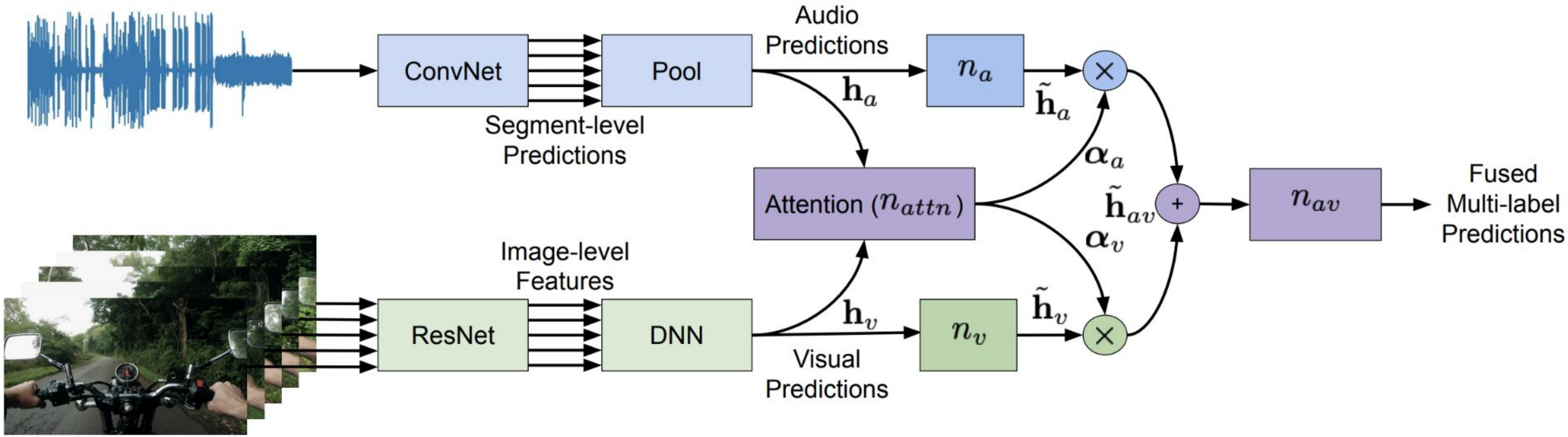
Self-driving cars' streams also have lots of sensors



Modality specific operations (e.g. convolutions)



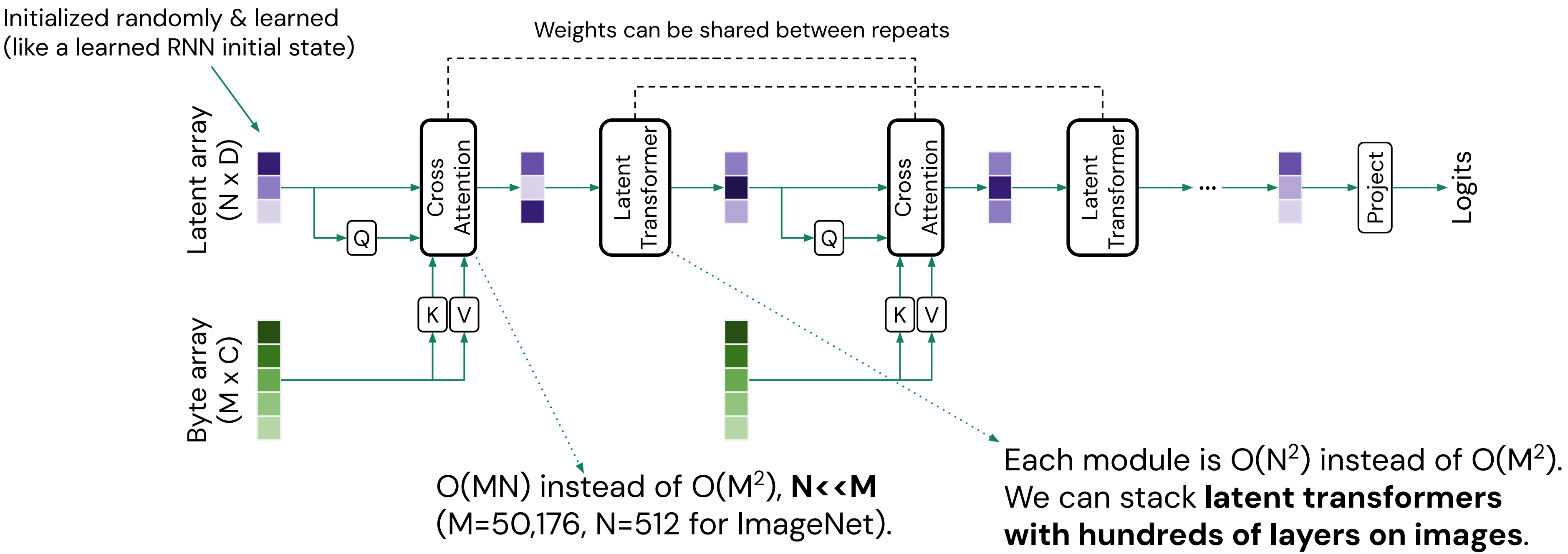
Multimodal fusion



Large Scale Audiovisual Learning of Sounds with Weakly Labeled Data

Haytham M. Fayek, Anurag Kumar 2020

The Perceiver



- Minimal assumptions about spatial structure (no patches/grids): **not just an image model**.
- Byte array features for images: [RGB + Fourier feature position encoding]

Perceiver: General Perception with Iterative Attention

ICML 2021

Andrew Jaegle[1] Felix Gimeno[1] Andrew Brock[1] Andrew Zisserman[1] Oriol Vinyals[1] Joao Carreira[1]

PERCEIVER IO: A GENERAL ARCHITECTURE FOR STRUCTURED INPUTS & OUTPUTS

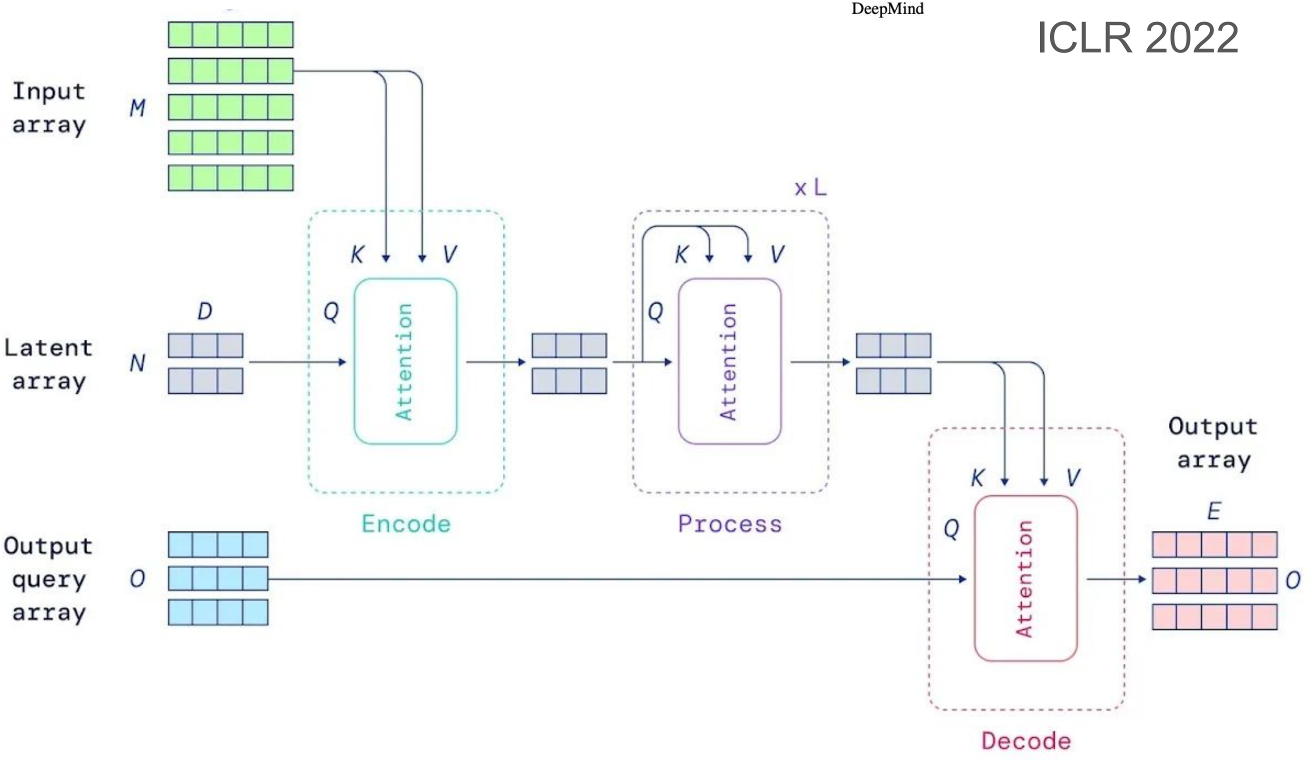
Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu,

David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff,

Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, João Carreira

DeepMind

ICLR 2022



Optical flow

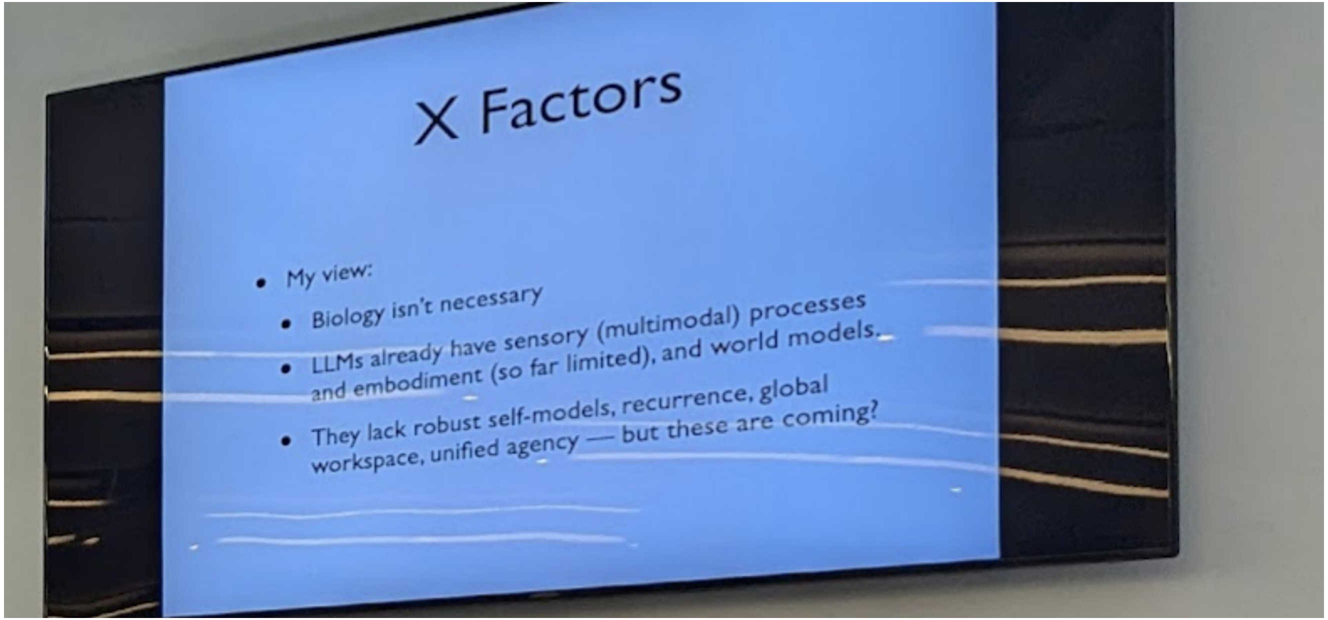
Multimodal auto-encoding

Original audio+video

Reconstruction

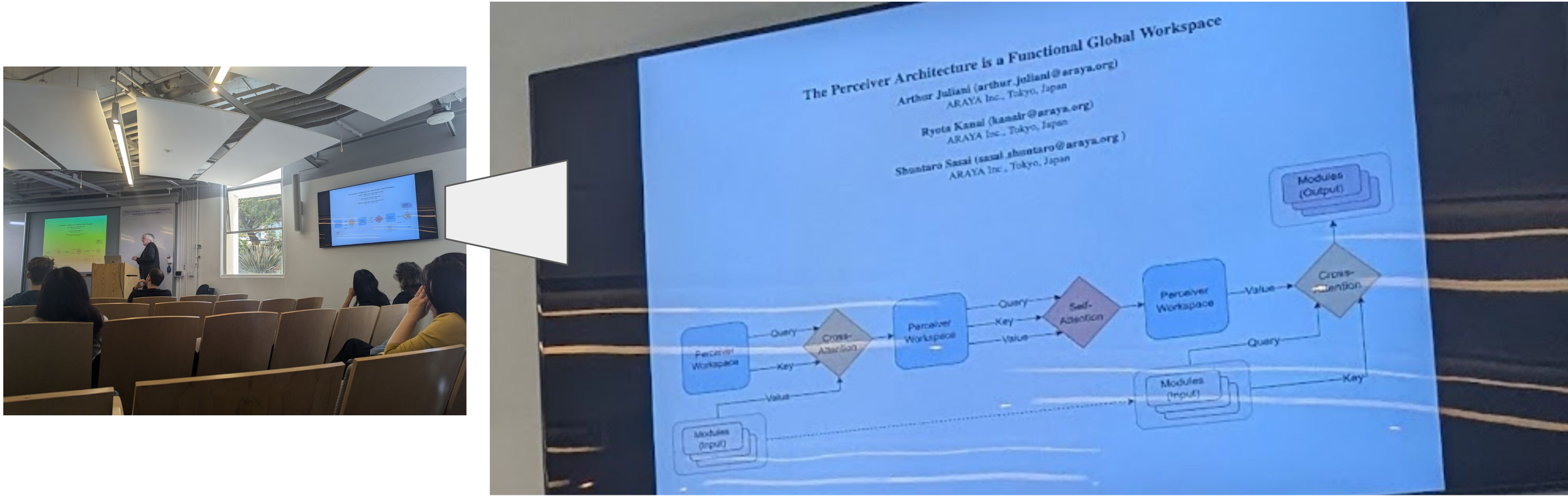
Consider Perceiver if you want your model to ever be conscious

- David Chalmers recent talk on whether LLMs are or will ever be conscious



Consider Perceiver if you want your model to ever be conscious

- David Chalmers recent talk on whether LLMs are or will ever be conscious



MooG - Moving Off the Grid – New!

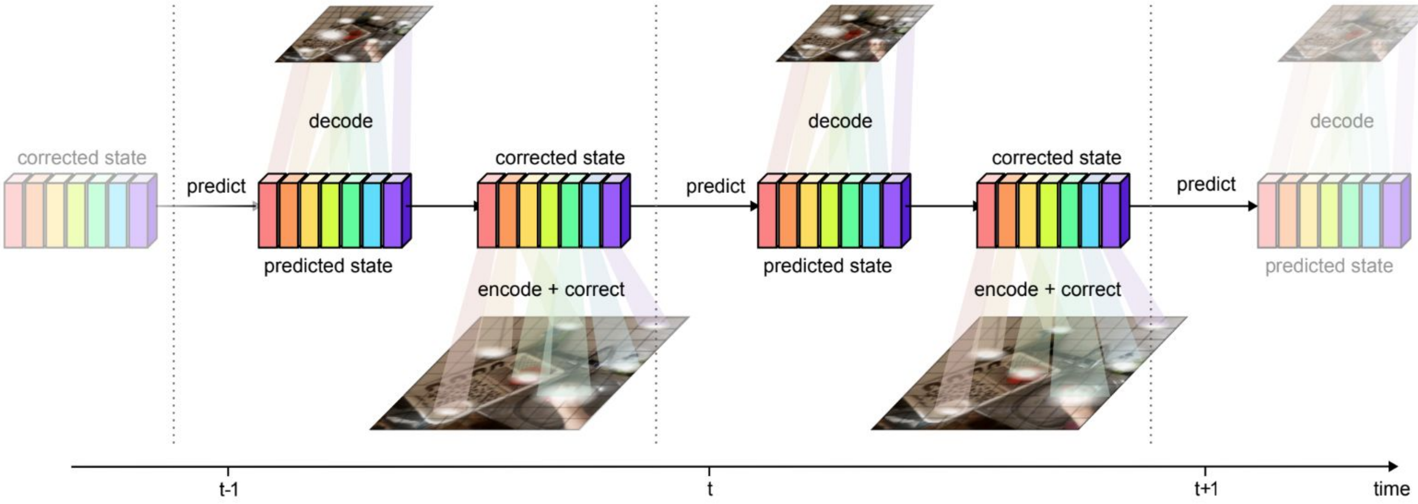
A self-supervised, recurrent, perceiver-like representation learning model which allows the representation to “bind” to scene elements and track them as they move.

16239 Moving Off-the-Grid: Scene-Grounded Video Representations

Download PDF

Sjoerd van Steenkiste, Daniel Zoran, Yi Yang, Yulia Rubanova, Rishabh Kabra, Carl Doersch, Dilara Gokay, Joseph Heyward, Etienne Pot, Klaus Greff, Drew A. Hudson, Thomas Albert Keck, Joao Carreira, Alexey Dosovitskiy, Mehdi S. M. Sajjadi, Thomas Kipf Hide authors

NeurIPS 2024 spotlight



[MooG](#) - **Mo**ving **O**ff the **G**rid

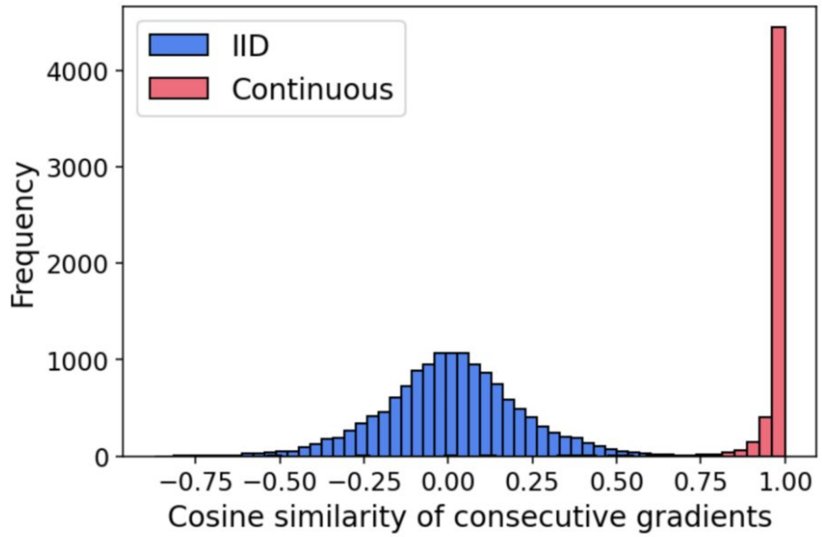
A self-supervised, recurrent, perceiver-like representation learning model which allows the representation to “bind” to scene elements and track them as they move.

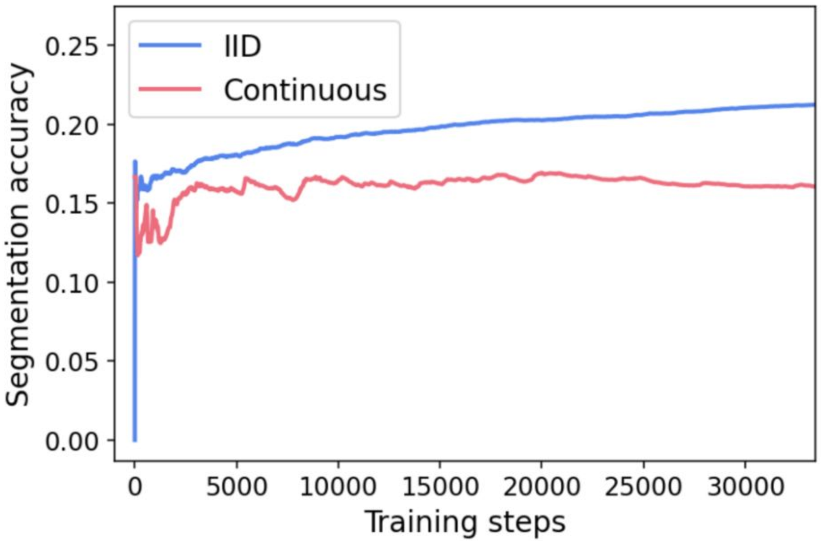
Argmax token | Blended Argmax | Input | Predictions

Problem #3 – how to make the model learn

- Data arrives sequentially
- Highly multimodal (2 eyes, ears, proprioception, etc)
- **Single very long (boring) stream**

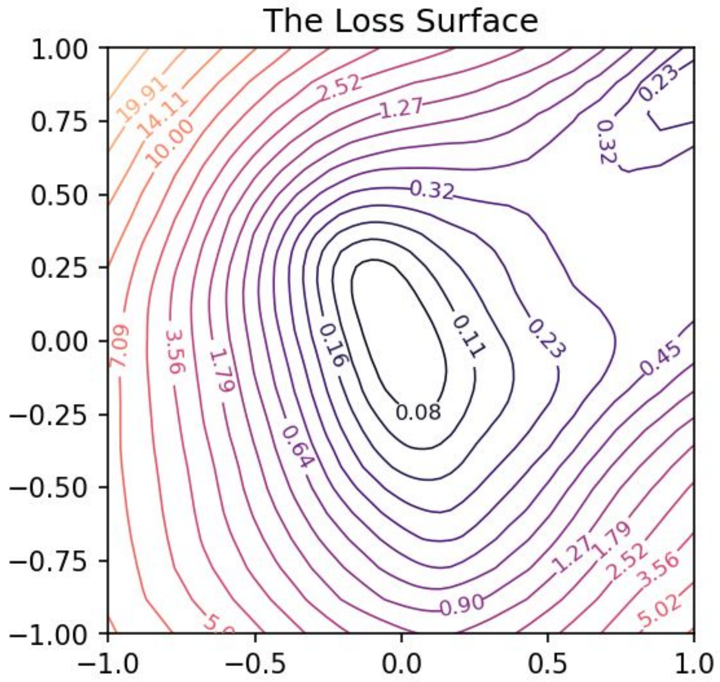
- All data available in parallel
- Images / mono-video + text (audio)
- **Many diverse images / videos**





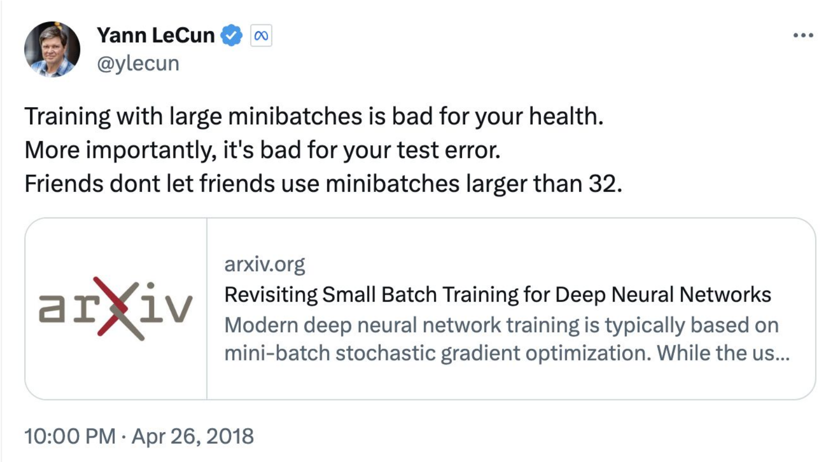
SGD + IID (independent and identically distributed data)

- Engine of deep learning
- Main empirical result: can find good enough local minima of loss function by approximating global gradient (over full dataset) by sequence of **random** local gradients (based on single example)



SGD + IID (Independent and identically distributed data)

- Original SGD = batch size 1 — this works fine


Yann LeCun
@ylecun

Training with large minibatches is bad for your health.
More importantly, it's bad for your test error.
Friends dont let friends use minibatches larger than 32.

arxiv.org
Revisiting Small Batch Training for Deep Neural Networks
Modern deep neural network training is typically based on mini-batch stochastic gradient optimization. While the us...

10:00 PM · Apr 26, 2018

Large-scale machine learning with stochastic gradient descent. Bottou et al

Revisiting Small Batch Training for Deep Neural Networks, Masters et al

From: *This Hand Is My Hand: A Probabilistic Approach to Hand Disambiguation in Egocentric Video, Lee et al*

Learning from One Continuous Video Stream

João Carreira, Michael King, Viorica Patraucean, Dilara Gokay, Catalin Ionescu, Yi Yang, Daniel Zoran, Joseph Heyward, Carl Doersch, Yusuf Aytar, Dima Damen, Andrew Zisserman

CVPR 2024

https://sites.google.com/view/one-stream-video

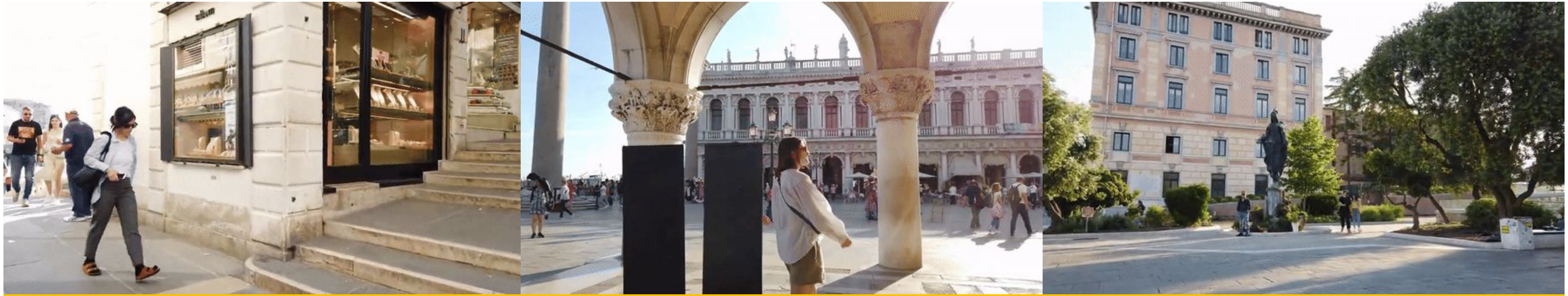
Continuous streams

Stream name	# videos train	# frames train	# videos val	# frames val	Max. length	Median length
Ego4D-stream	21,704	294M (3,265h)	2302	31M (348h)	1.95h	8.8 minutes
ScanNet-stream	1,199	1.8M (20h)	312	0.5M (5.7h)	5.5 minutes	1 minute

ScanNet

Ego4D

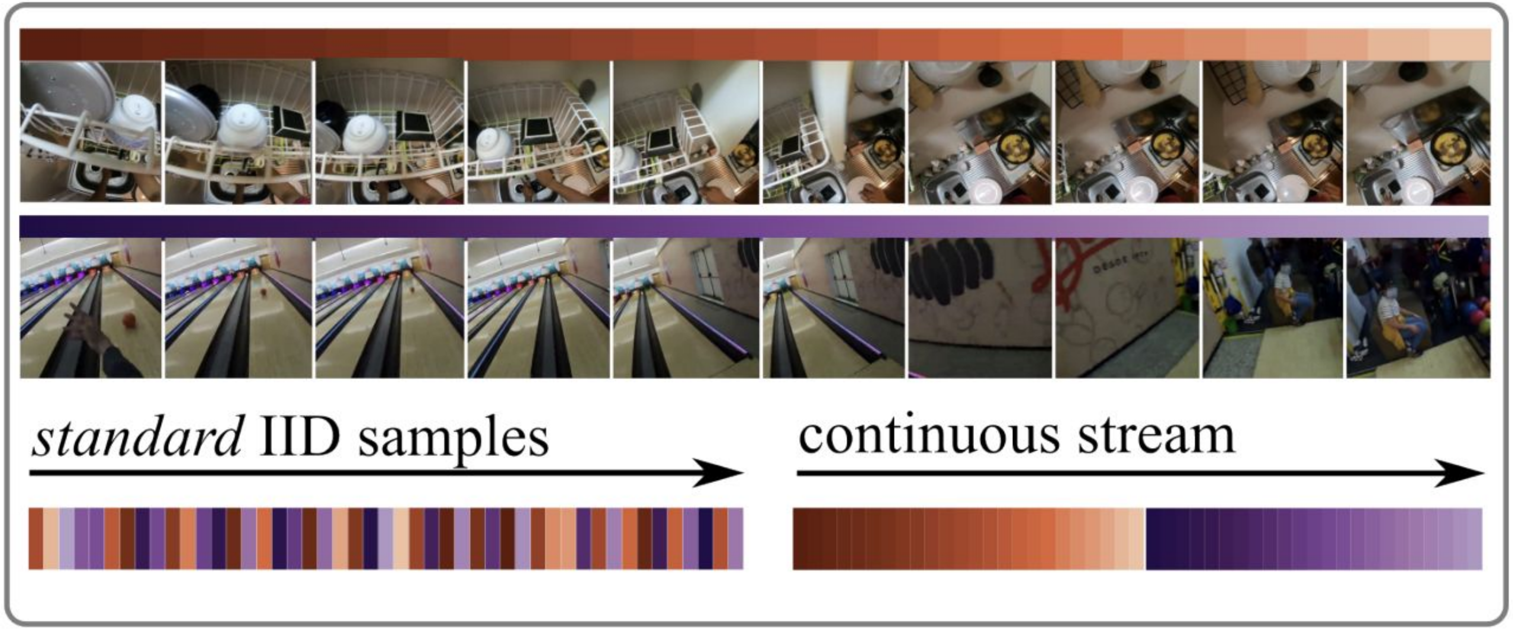
Walking Tour Dataset – also a good dataset for this



10 x 4K videos from different cities, Avg duration – 1hr 38min, ~700 classes, License - CC-BY

Wiles *et al.*, Compressed vision for efficient video understanding. In ACCV, 2022.
Venkataramanan *et al*., Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video. ICLR 2024

Continuous streams



Tasks

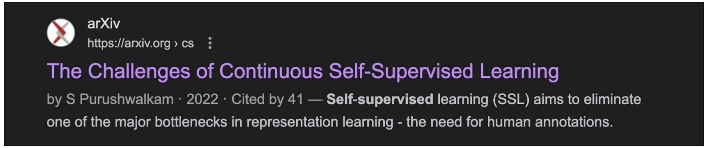
Future frame prediction:

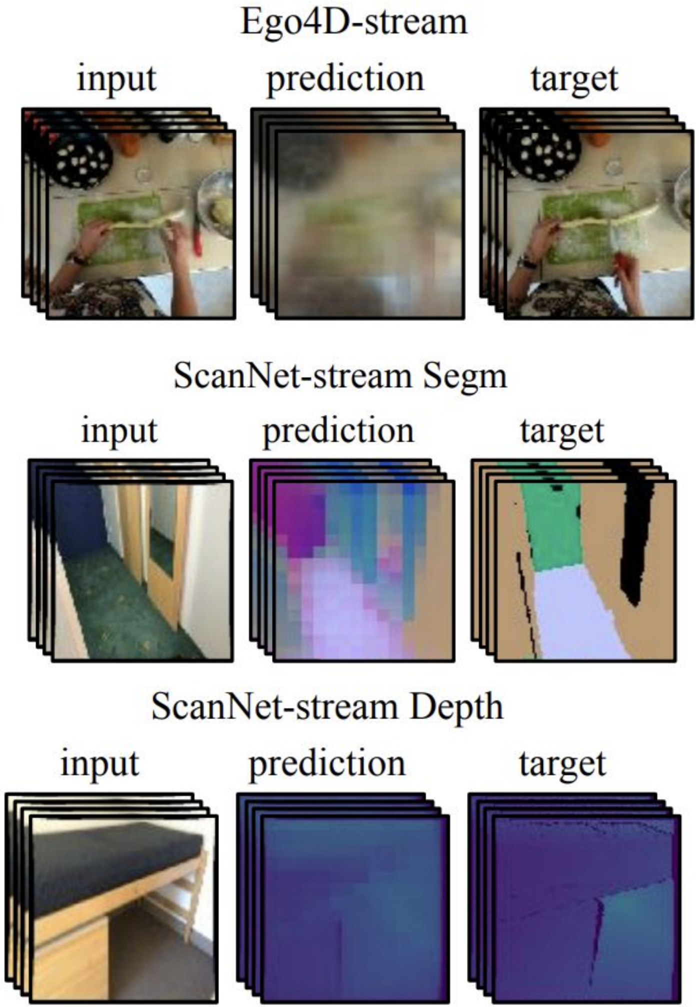
- Pixel space
- Semantic segmentation
- Depth

Displacement 0 (present), 4 and 16 frames

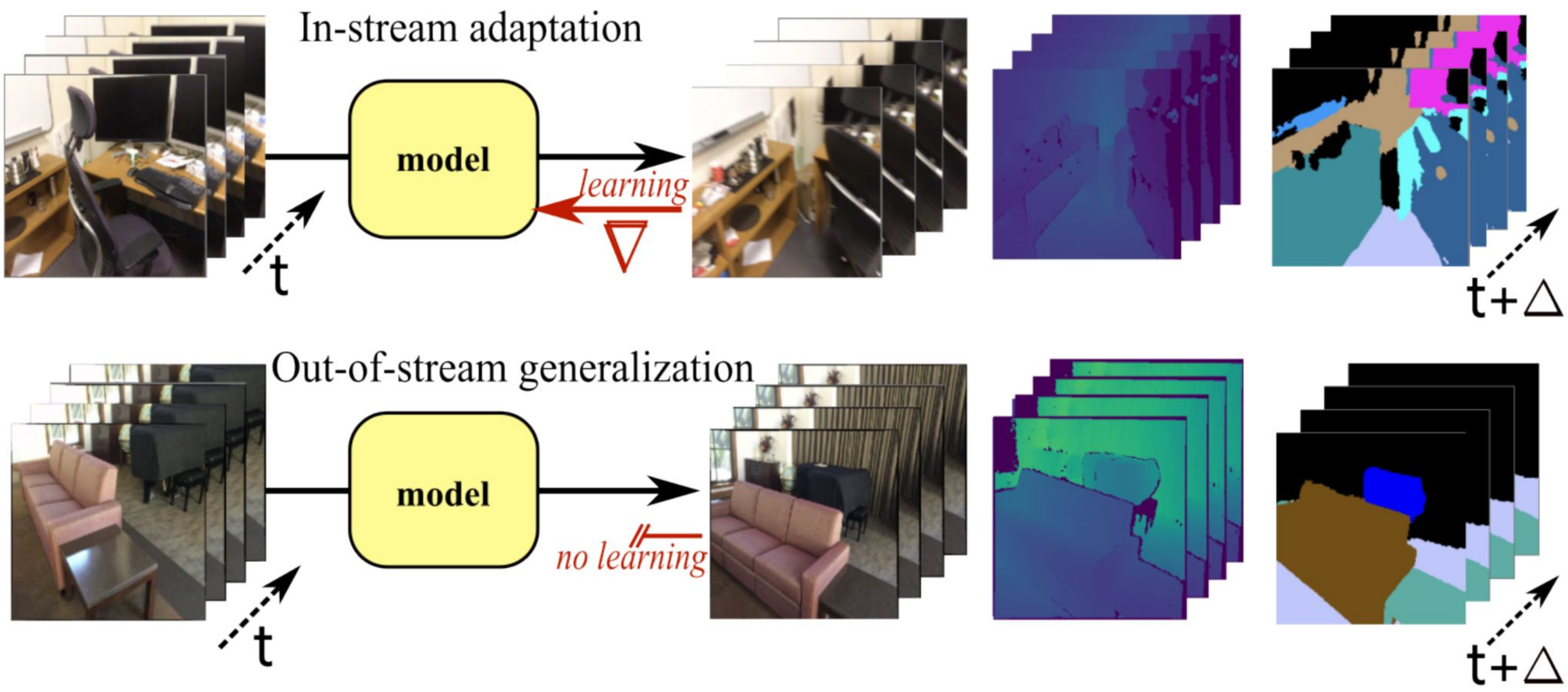
Models see a 24h-long stream

Note: could also evaluate via representation learning evals as in:


arXiv
https://arxiv.org › cs
The Challenges of Continuous Self-Supervised Learning
by S Purushwalkam · 2022 · Cited by 41 — **Self-supervised** learning (SSL) aims to eliminate one of the major bottlenecks in representation learning - the need for human annotations.

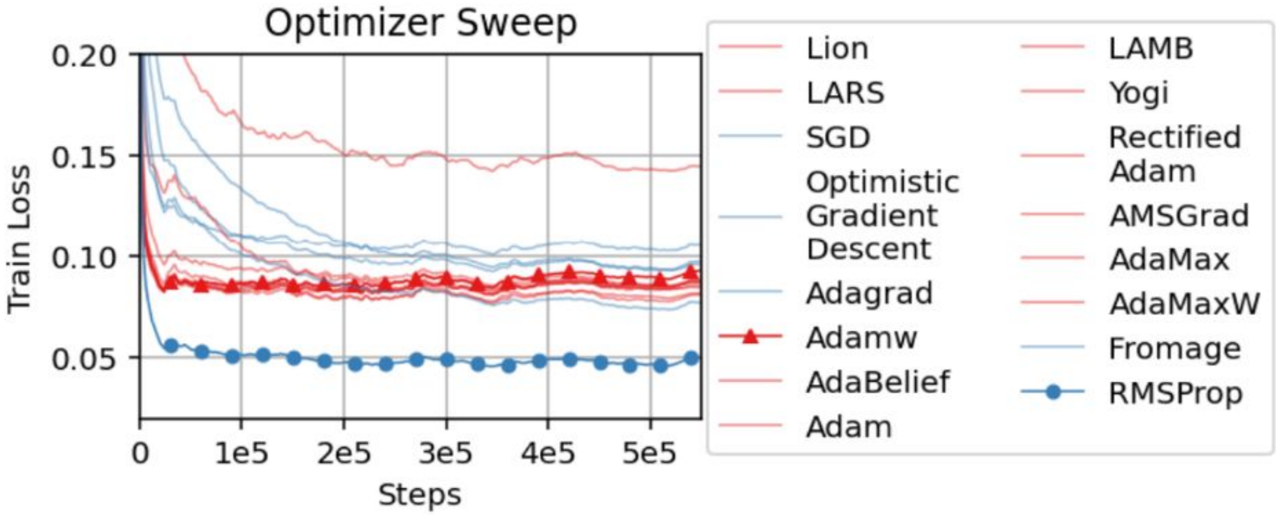


Evaluation


In-stream adaptation
model
learning
t
t+△
Out-of-stream generalization
model
no learning
t
t+△

What did we learn with the framework ?

- **Non-standard optimization settings help**
- Pretraining helps



What did we learn with the framework ?

- **Non-standard optimization settings help**
- Pretraining helps

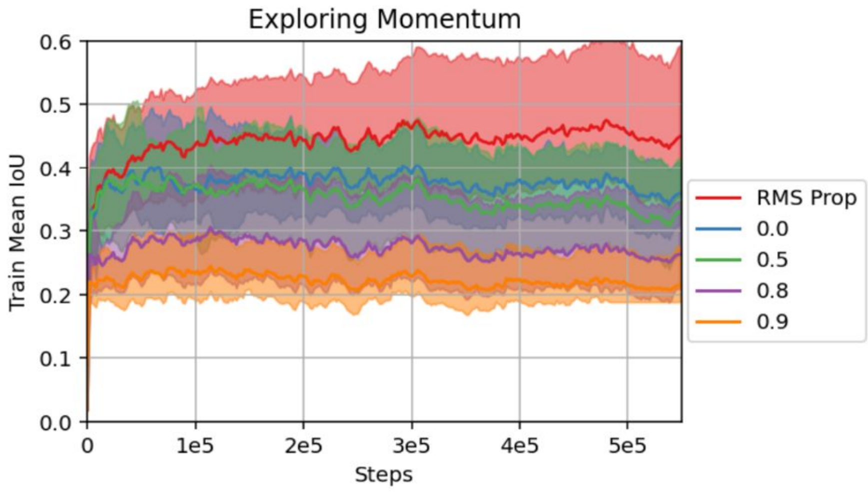


Figure 5. Reducing momentum with the AdamW optimizer helps to recover some of the performance of RMS Prop.

What did we learn with the framework ?

- **Non-standard optimization settings help**
- Pretraining helps

→ updating weights less frequently helps generalization, hurts adaptation

			n steps per update			
Stream	dataset	model	1	4	16	64
In	Ego4D (↓)	UNet	**.036**	.038	.037	.039
		ViT	**.035**	.037	.039	.047
	Segm (↑)	UNet	**.420**	.292	.211	.195
		ViT	**.457**	.395	.302	.232

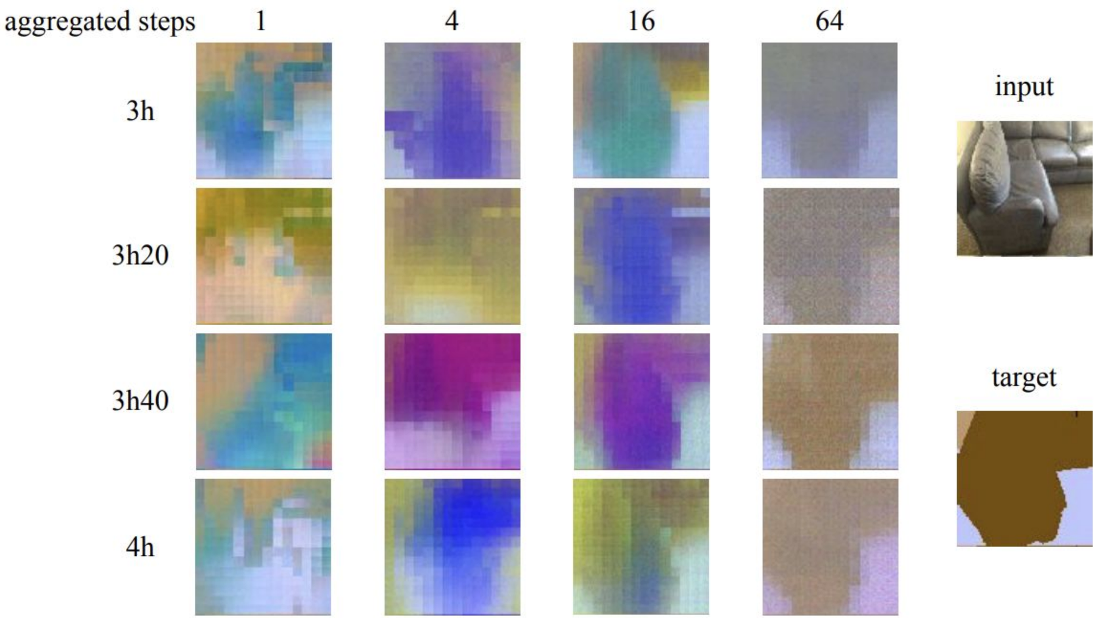
What did we learn with the framework ?

- **Non-standard optimization settings help**
- Pretraining helps

→ updating weights less frequently helps generalization, hurts adaptation

			n steps per update			
Stream	dataset	model	1	4	16	64
Off	Ego4D (↓)	UNet	.095	.051	.047	**.042**
		ViT	.076	.062	.046	**.044**
	Segm (↑)	UNet	.176	.179	**.205**	.183
		ViT	.251	.272	**.280**	.274

Out-of-stream outputs – visualization

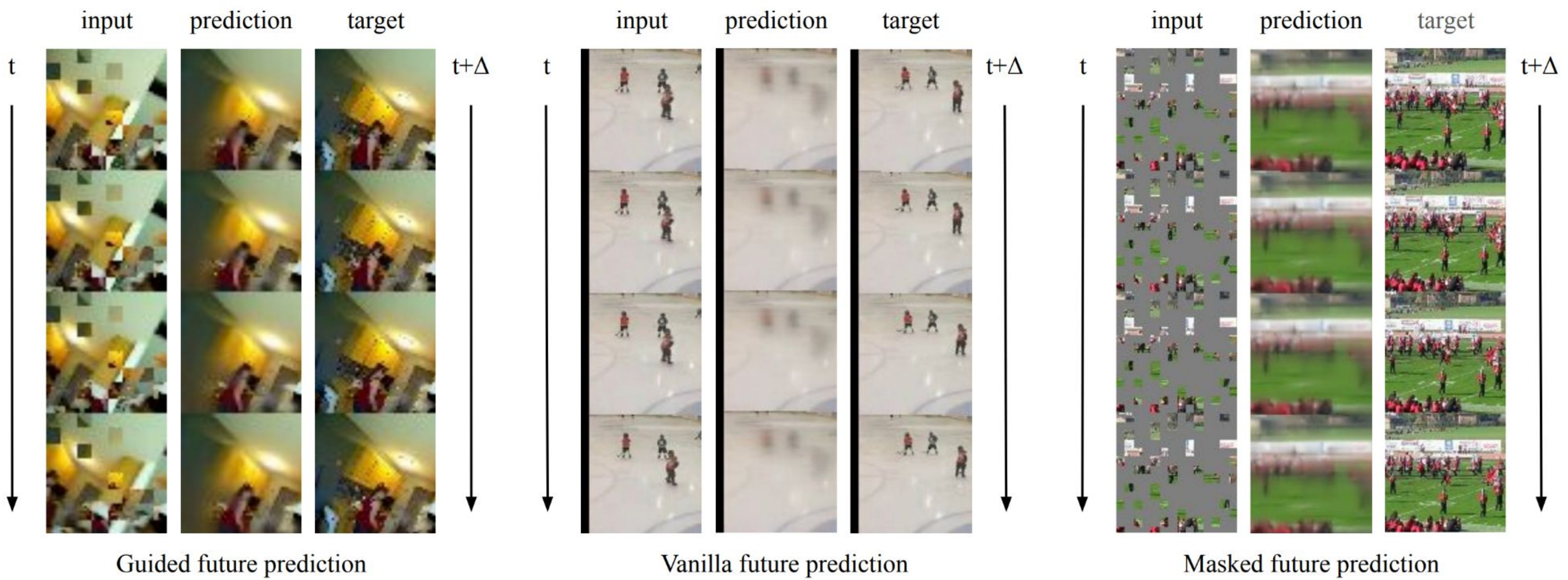


What did we learn with the framework ?

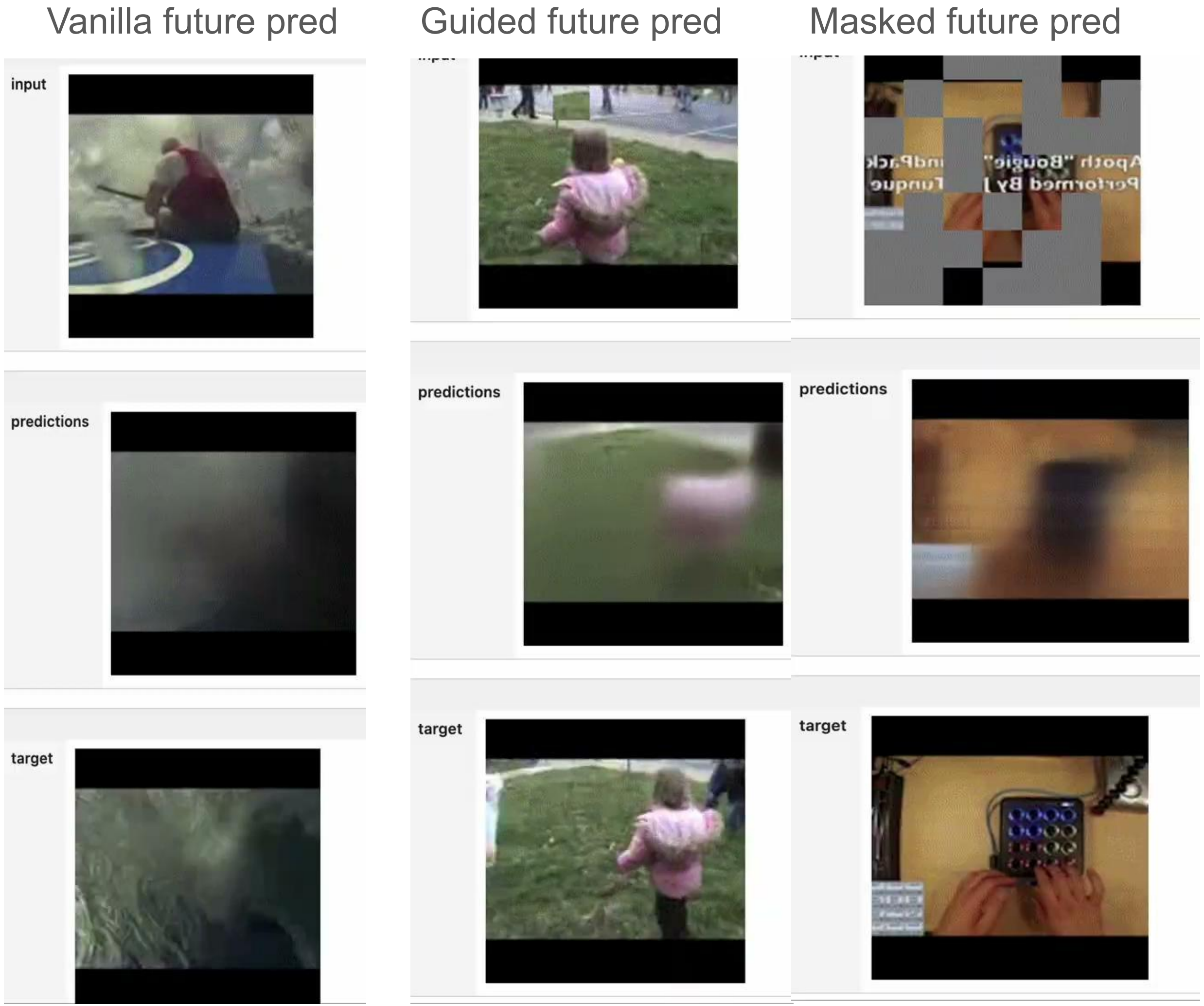
- Non-standard optimization settings help
- **Pretraining helps**

Pretraining – Kinetics

- No pretraining
- Imagenet-based pretraining (e.g. classification, MAE)
- Video-based pretraining



Visualization: 3.84s displacement



What did we learn with the framework ?

- Non-standard optimization settings help
- **Pretraining helps**

Pretraining Checkpoint	Ego4D (↓)	ScanNet Depth (↓)	ScanNet Segm (↑)
None	.074 / .105	1.969 / 2.163	.177 / .188
ViT-L-I1K-CLS	.043 / .048	1.821 / 2.040	.288 / .234
ViT-L-I21K-CLS	.042 / .048	1.735 / 2.013	.244 / .192
ViT-L-I1K-MAE	.040 / .044	1.806 / 2.045	.360 / **.320**
Guided Future Prediction	**.036** / **.043**	**1.622** / **1.990**	**.390** / .313

Comparison to IID training – VIT-L

STDL (standard deep learning, baseline):

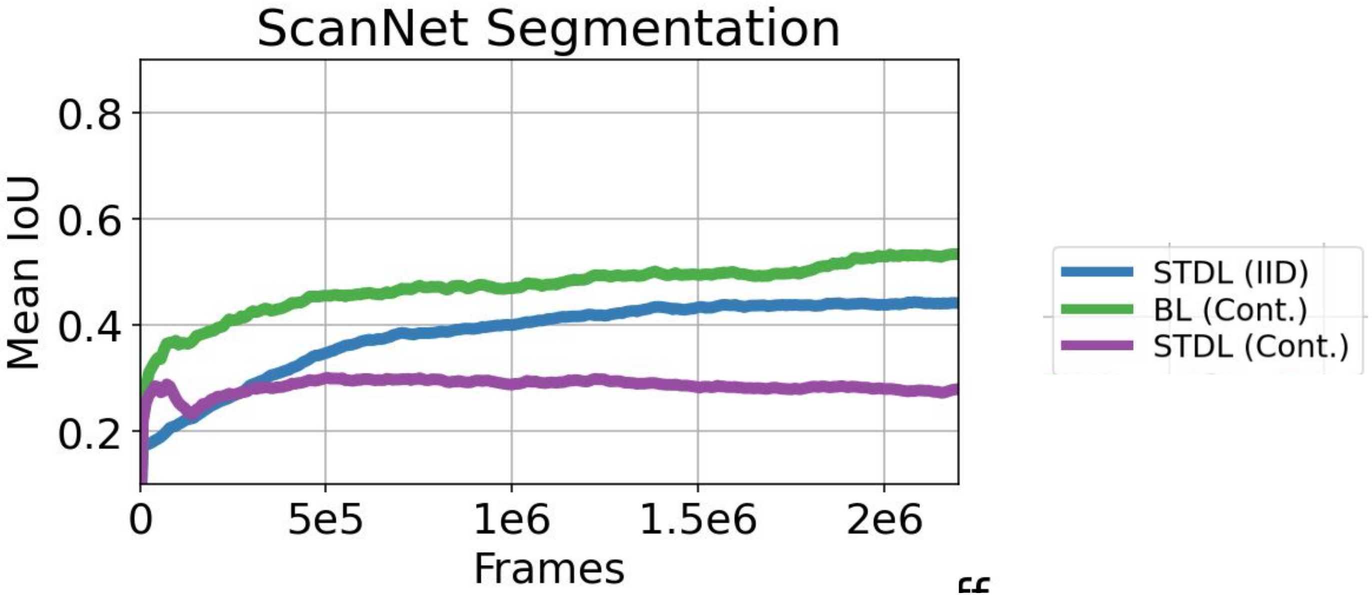
- ADAM
- Update weights after every video chunk
- ImageNet-MAE
- Batch size 1

BL (baby learning):

- RMSProp
- Update weights after 16 video chunks
- Kinetics guided future prediction
- Batch size 1

Comparison to IID training – VIT-L

In-stream

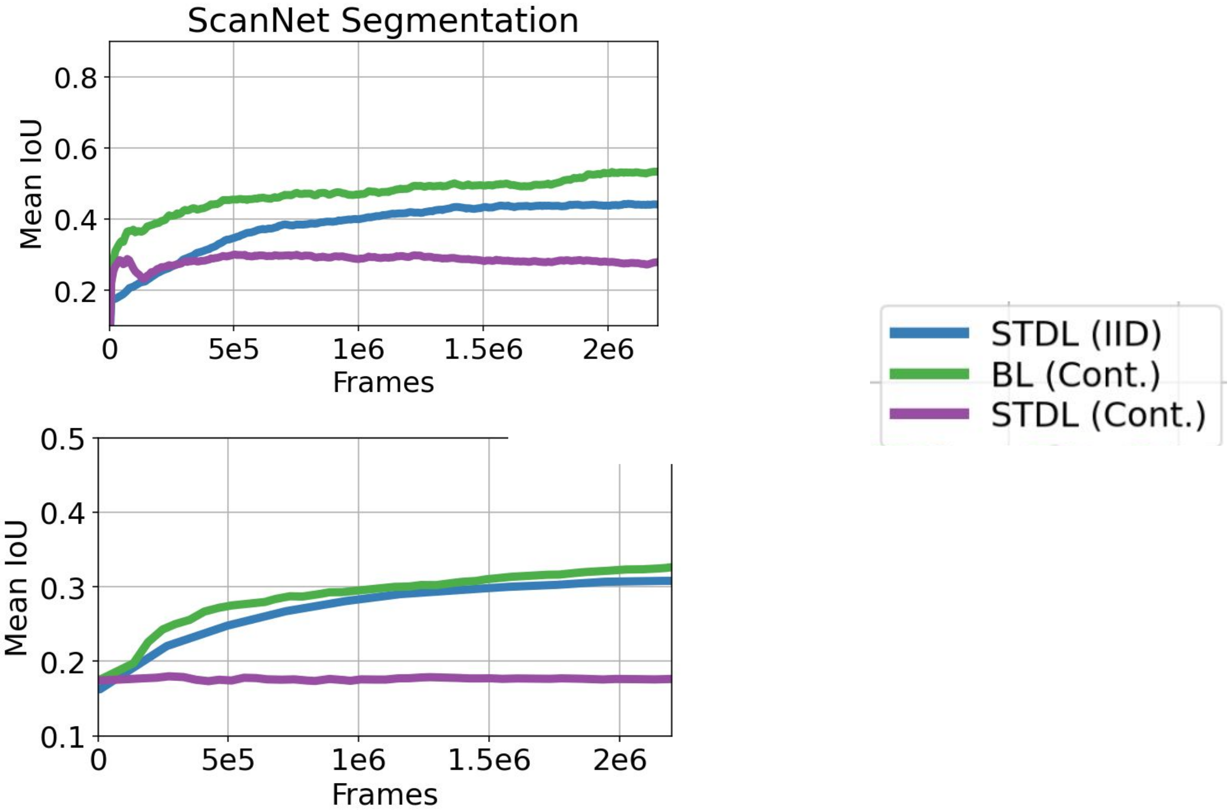


Temporal displacement 0 (no future prediction here)

Comparison to IID training – VIT-L

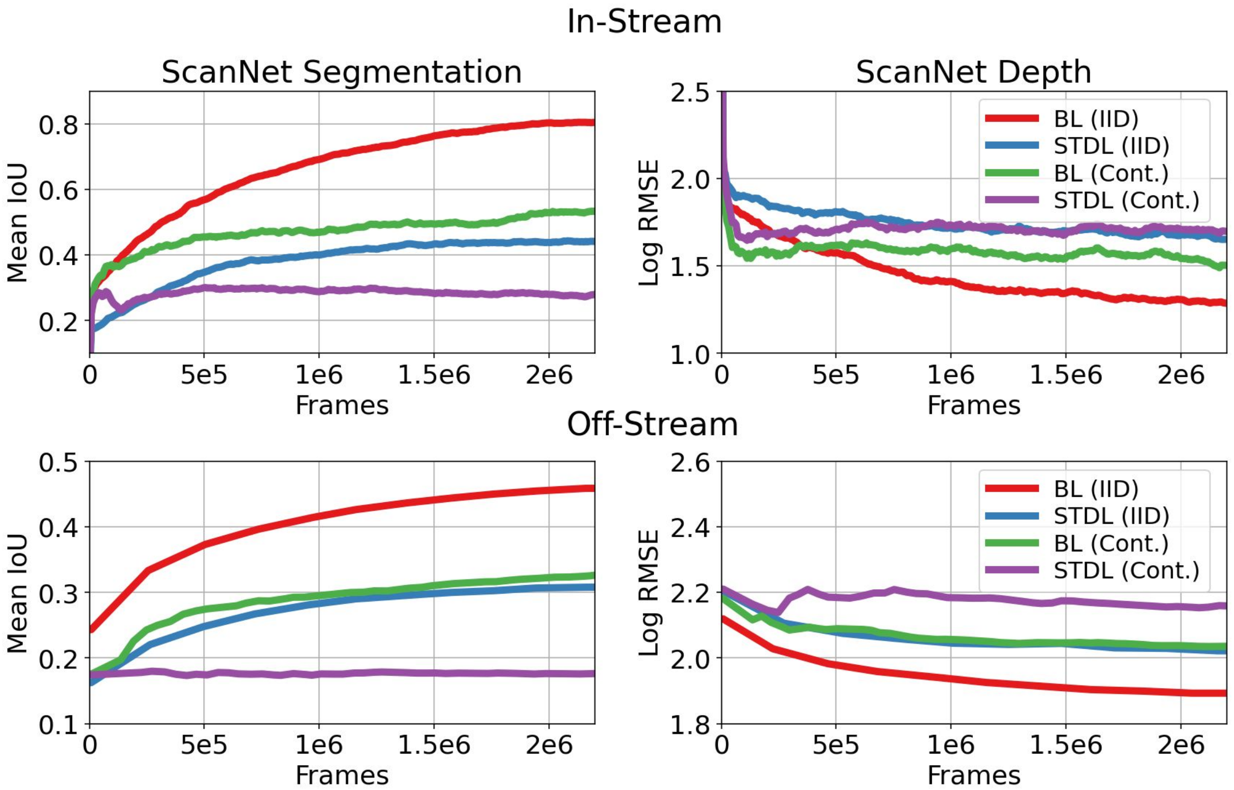
In-stream

Out-of-stream



Comparison to IID training – VIT-L

- Improved pretraining making a lot of difference!
- Much progress to be done bridging IID and continual



Conclusion

- Just scratching the surface of single-stream learning, much (all) research to do
  - Cheap (no need for many GPUs)
  - Needs great ideas (shuffled learning should not work best!)
  - It’s the future
  - Risky!
    - Great for academia

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain.

Turing