AIDA
ARTIFICIAL INTELLIGENCE
DOCTORAL ACADEMY

The Machine Learning of Time: Past & Future

Efstratios Gavves, University of Amsterdam, egavves@uva.nl

About Me

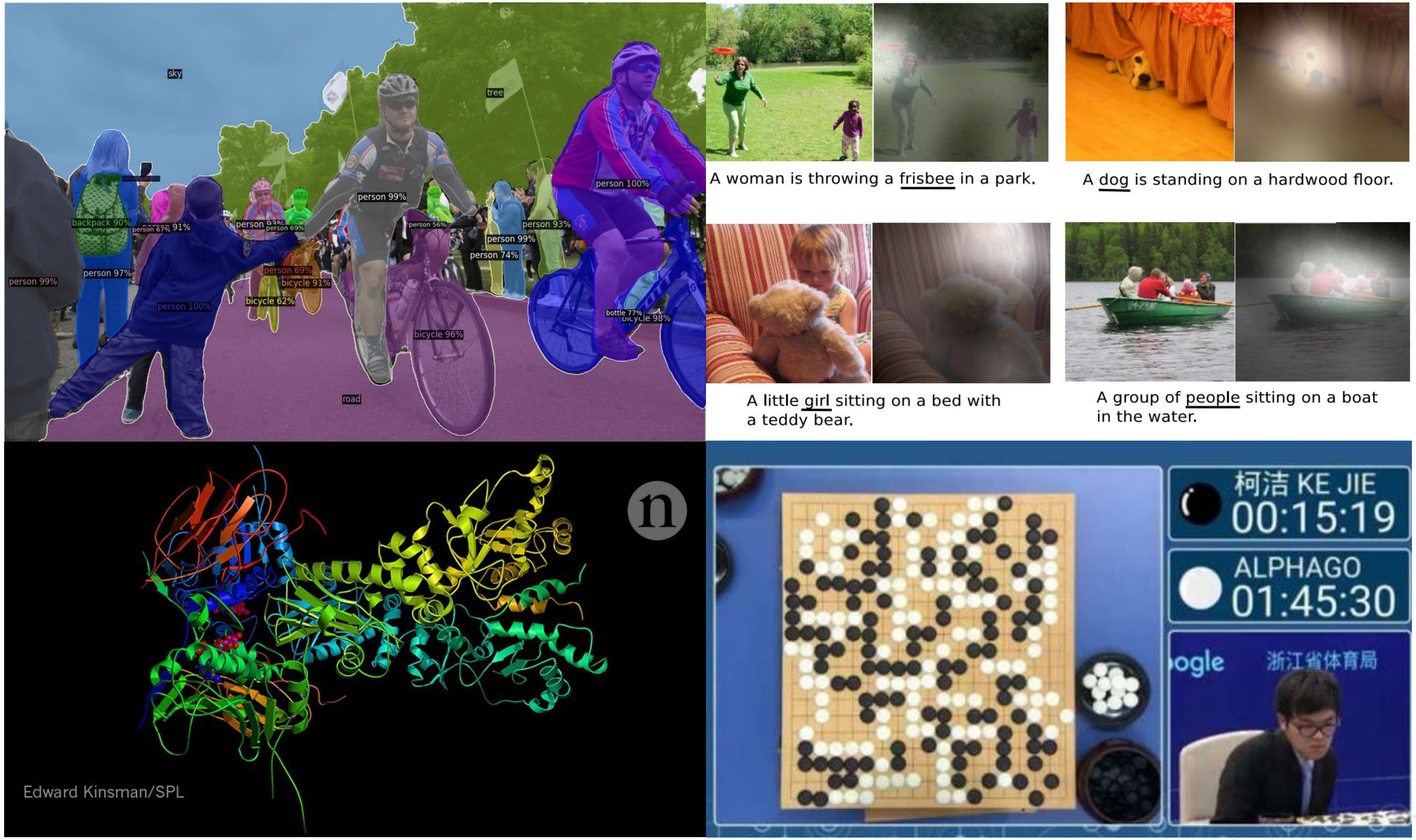
- Associate Professor at the University of Amsterdam
  - Director of QUVA and POP-AART Lab (we will be hiring!)
- Co-founder of Ellogon.AI
  - AI to personalize to immunotherapy in oncology
- ELLIS scholar

Temporal Learning and Dynamics

Efficient Vision and Learning

Machine Learning for Oncology

The Golden Age of Learning Algorithms



[1] Ghodrati, Gavves, Snoek, Video Time: Encoders, Properties, Evaluation, BMVC, 2018
[2] Zhou, Andonian, Oliva, Torralba, Temporal Relational Reasoning in Videos, ECCV, 2018

An Urgent Paradox

Apple falling: videos reversed, shuffled or normal ⇒ no difference[1,2]

Normal video: 83.1%

Reversed frames: 82.9%

State-of-the-art spatiotemporal models ignore time

[1] Ghodrati, Gavves, Snoek, Video Time: Encoders, Properties, Evaluation, BMVC, 2018
[2] Zhou, Andonian, Oliva, Torralba, Temporal Relational Reasoning in Videos, ECCV, 2018

An Urgent Paradox

Apple falling: videos reversed, shuffled or normal ⇒ no difference[1,2]

Normal video: 83.1%

Reversed frames: 82.9%

Urgent for forecasting

State-of-the-art spatiotemporal models ignore time

An Urgent Paradox

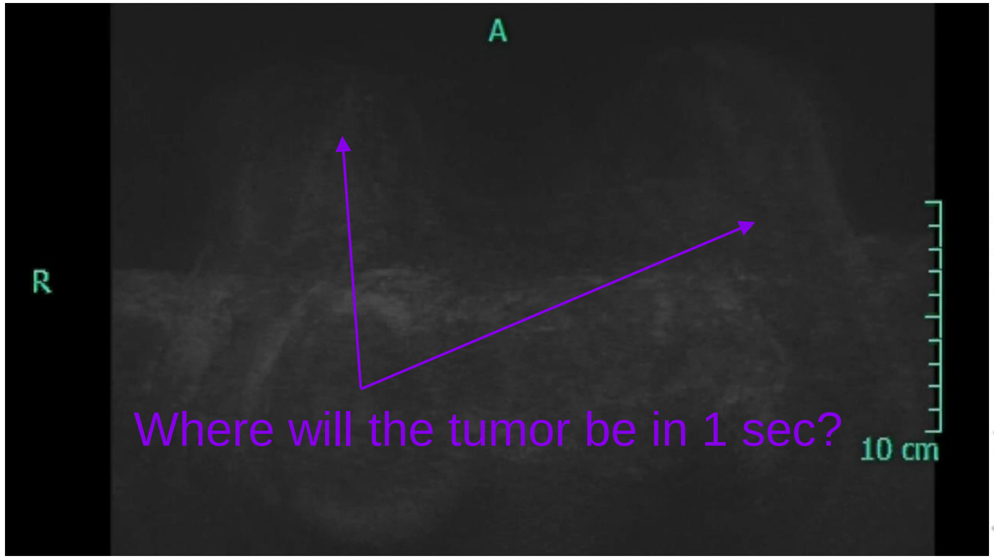
[1] Ghodrati, Gavves, Snoek, Video Time: Encoders, Properties, Evaluation, BMVC, 2018
[2] Zhou, Andonian, Oliva, Torralba, Temporal Relational Reasoning in Videos, ECCV, 2018

Apple falling: videos reversed, shuffled or normal ⇒ no difference[1,2]

Normal video: 83.1%

Reversed frames: 82.9%

Urgent for future planning



State-of-the-art spatiotemporal models ignore time

[1] Ghodrati, Gavves, Snoek, Video Time: Encoders, Properties, Evaluation, BMVC, 2018
[2] Zhou, Andonian, Oliva, Torralba, Temporal Relational Reasoning in Videos, ECCV, 2018

An Urgent Paradox

Apple falling: videos reversed, shuffled or normal ⇒ no difference[1,2]

Normal video: 83.1%

Reversed frames: 82.9%

Urgent for autonomous driving

State-of-the-art spatiotemporal models ignore time

Central Question

What is the role of time in visual recognition?

AIDA
ARTIFICIAL INTELLIGENCE
DOCTORAL ACADEMY

The Vision

Models that learn temporality in entangled spatiotemporal sequences

ARTIFICIAL INTELLIGENCE DOCTORAL ACADEMY

The Vision

Models that learn temporality in entangled spatiotemporal sequences

+[illegible]

The Vision

Models that learn temporality in entangled spatiotemporal sequences

+1.1

-[illegible]

The Vision

Models that learn temporality in entangled spatiotemporal sequences

+1.1

-5.1

[illegible]

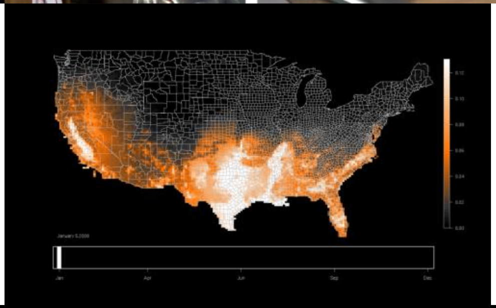
ARTIFICIAL INTELLIGENCE
DOCTORAL ACADEMY

Entangled spatiotemporal data

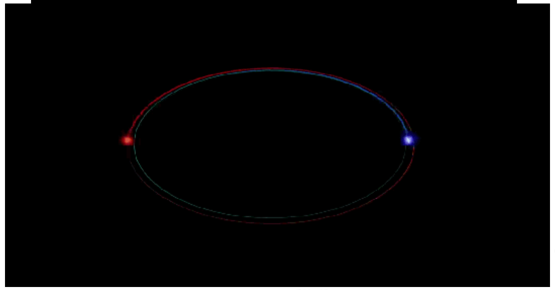
- Data in thousands of dimensions confounded space and time

- Example 1: Long & complex videos

- Example 2: Migration patterns

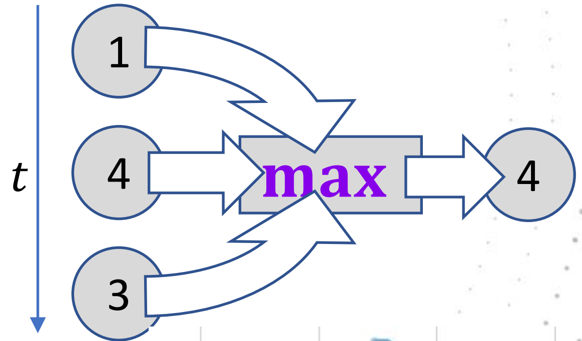


- Example 3: Particles through time



What’s the challenge?

- Thousands of frames → A lot of correlations and dynamics

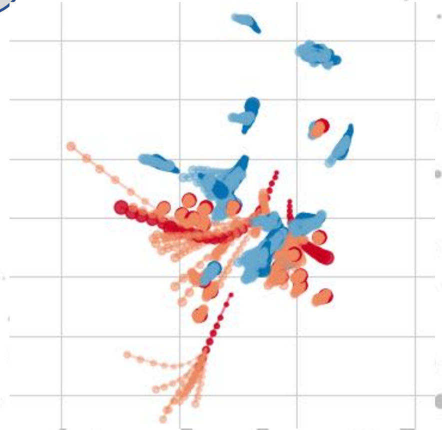


Challenge #1: State-of-the-art discards time by aggregating with set operations

Challenge #2: Hard to annotate manually → supervised learning is debatable

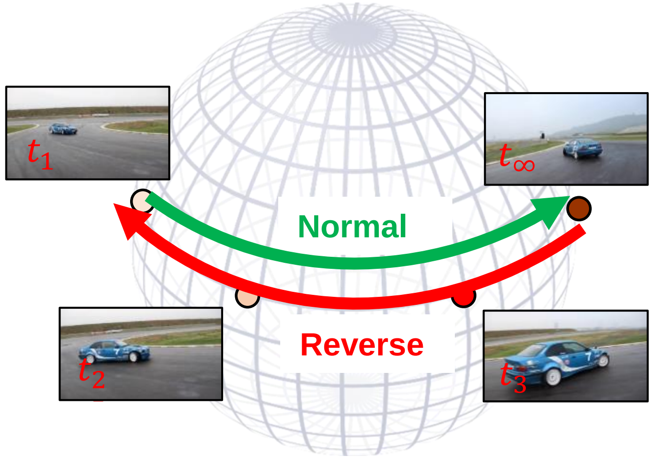
Challenge #3: A sequence is one of myriad possibilities → generative modelling critical

Challenge #4: Lack of standardization ← data is huge, algorithms very complex



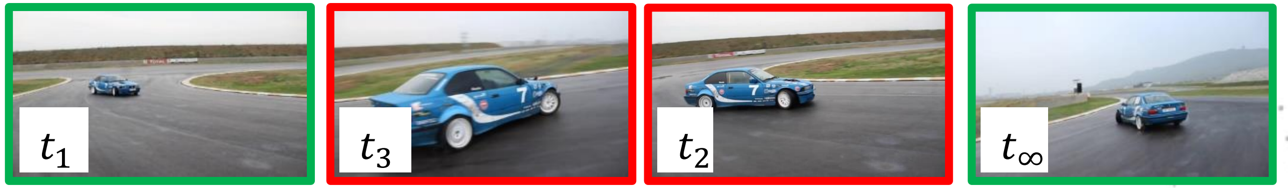
Addressing the challenges

Time Geometry
- Learn spatiotemporal geometric manifolds
- Aggregate over a geodesic time path on manifold

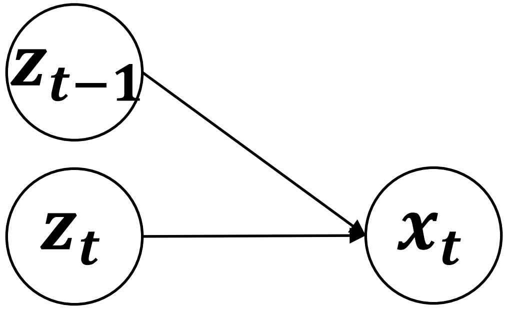


Time Supervision
- Replace manual annotation with time properties
- In particular, combine with time-sensitive models

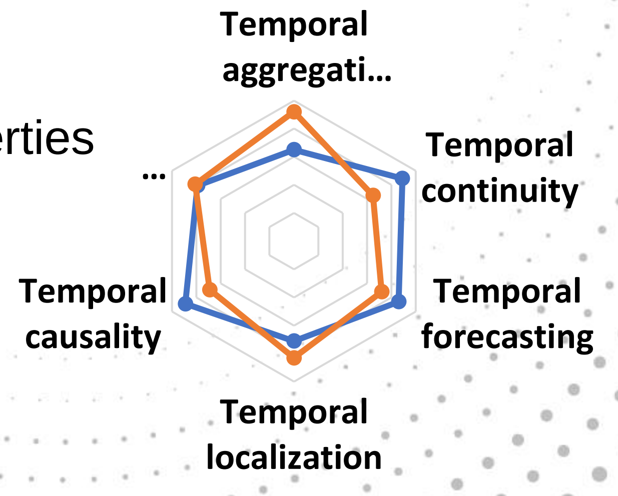
Are frames correctly ordered?



Time Generation
- Models that imagine all possible futures
- Spatiotemporal generative/bayesian models



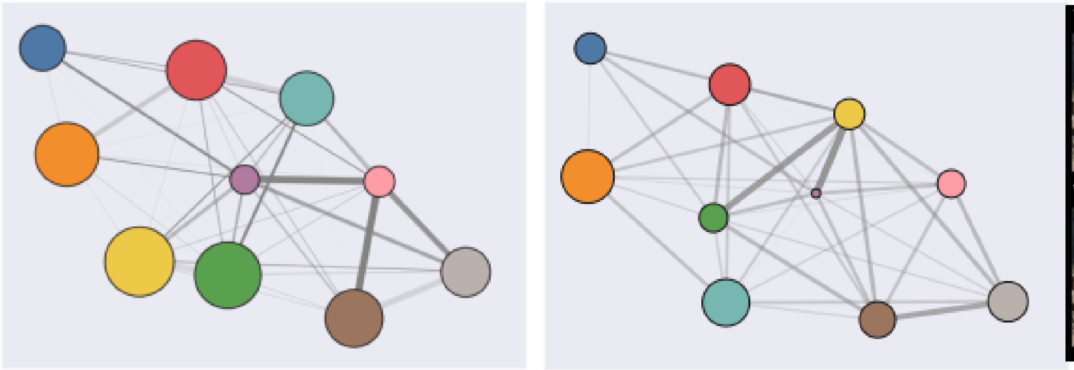
Time Evaluation
- Standardize data
- Evaluate on temporal properties



Time Geometry

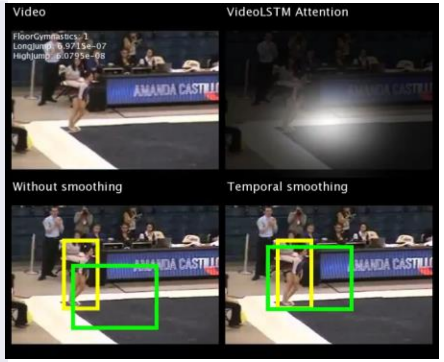
Types of spatiotemporal geometric learning

Random walks on space-time graphs



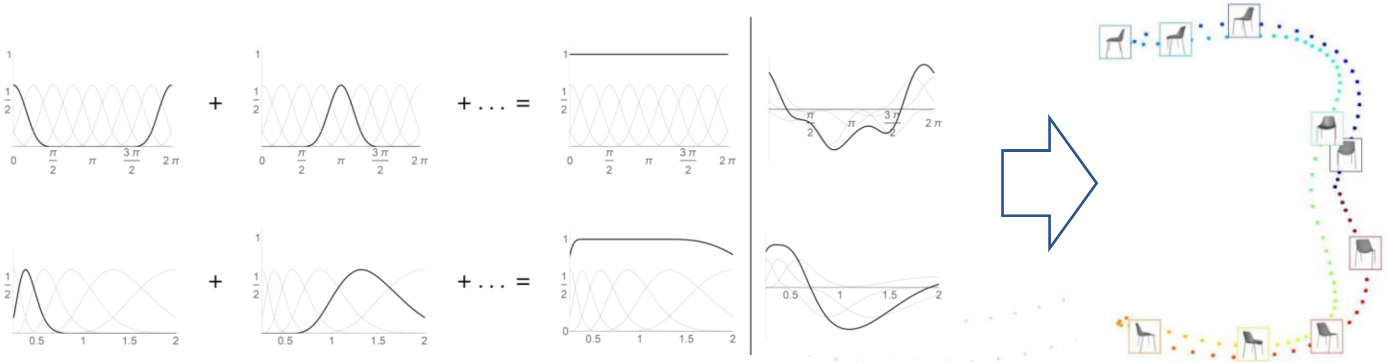
(a) Balance beam (b) Floor exercise

swing flip flares handstands twist balance



[1]

Space-time manifolds



[1] Li, Gavrilyuk, Gavves, Mihir, Snoek, VideoLSTM convolves, attends and flows for action recognition, CVIU 2018

VideoGraph: Recognizing Minutes-Long Human Activities in Videos

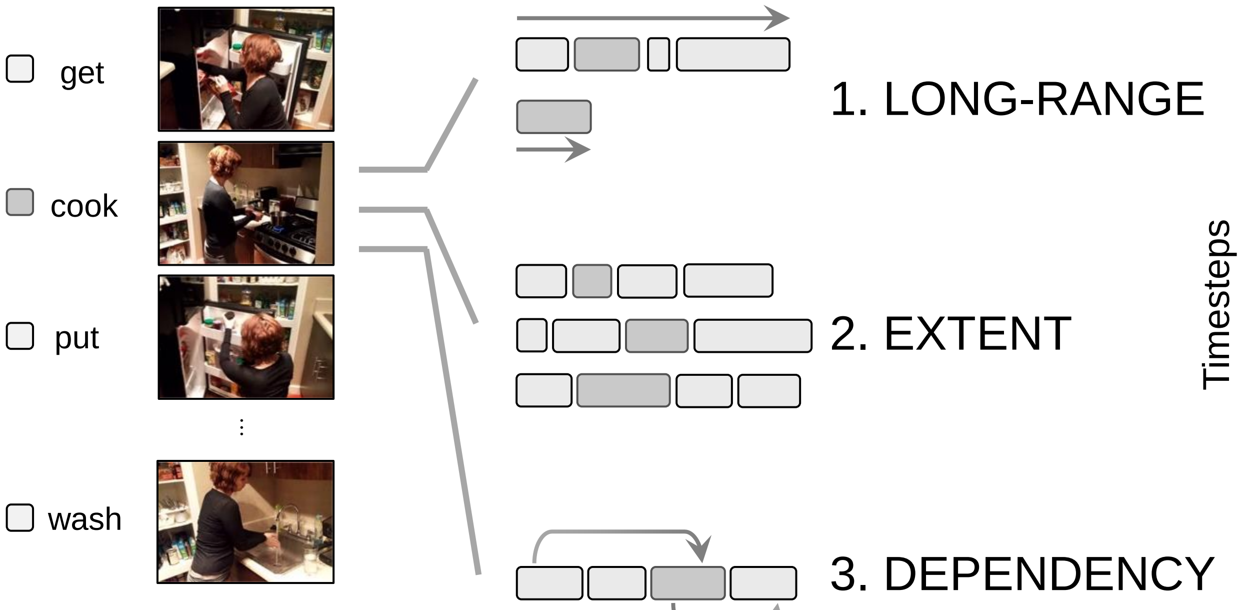

N. Hussein


E. Gavves

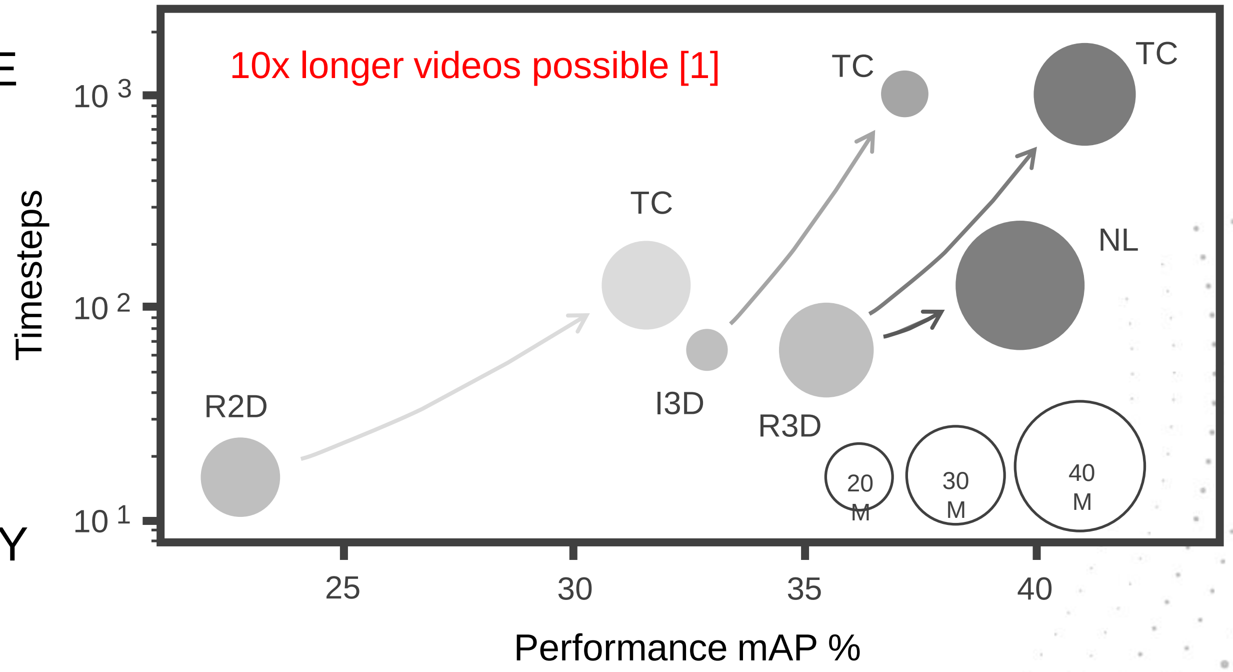

A. Smeulders

Hussein, Gavves, Smeulders, VideoGraph: Recognizing Minutes-Long Human Activities in Videos, 2019

Temporal length is important with complex videos



Cooking a Meal

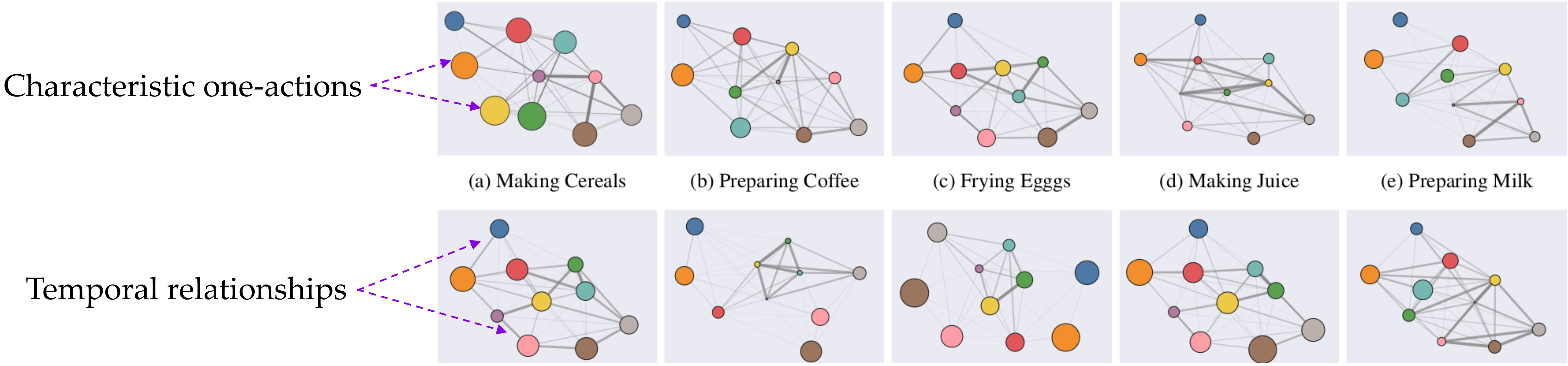


Code available

[1] Hussein, Gavves Smeulders, Timeception for Complex Action Recognition, 2019

How to stretch time further → (Time) Graphs!

- Sublinear temporal representation
- Compositionality
- Interpretability

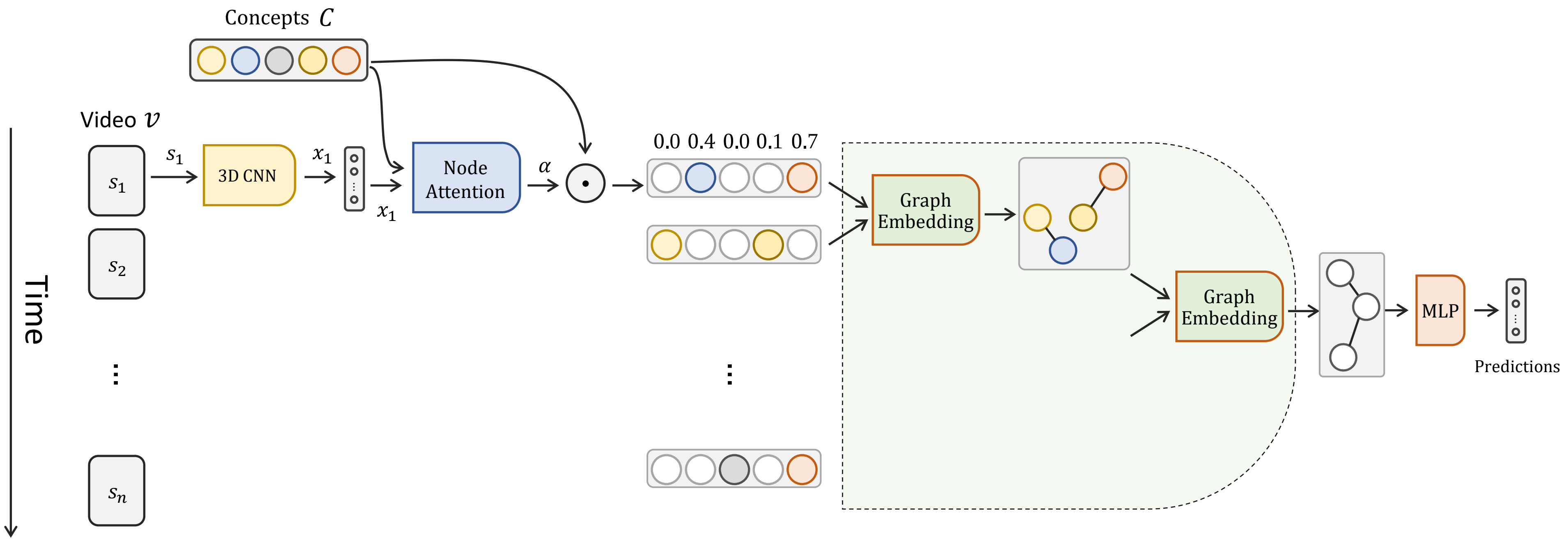


(a) Making Cereals (b) Preparing Coffee (c) Frying Egggs (d) Making Juice (e) Preparing Milk

(f) Making Pancake (g) Making Salat (h) Making Sandwich (i) Making Scrambled Egg (j) Preparing Tea

● cereal, ● pan, ● eggs, ● sandwitch, ● kettle, and ● foodbox

VideoGraph



Concepts C
Video $\mathcal{V}$
Time
s_1
s_2
s_n
s_1
3D CNN
x_1
x_1
Node Attention
α
0.0 0.4 0.0 0.1 0.7
Graph Embedding
Graph Embedding
MLP
Predictions

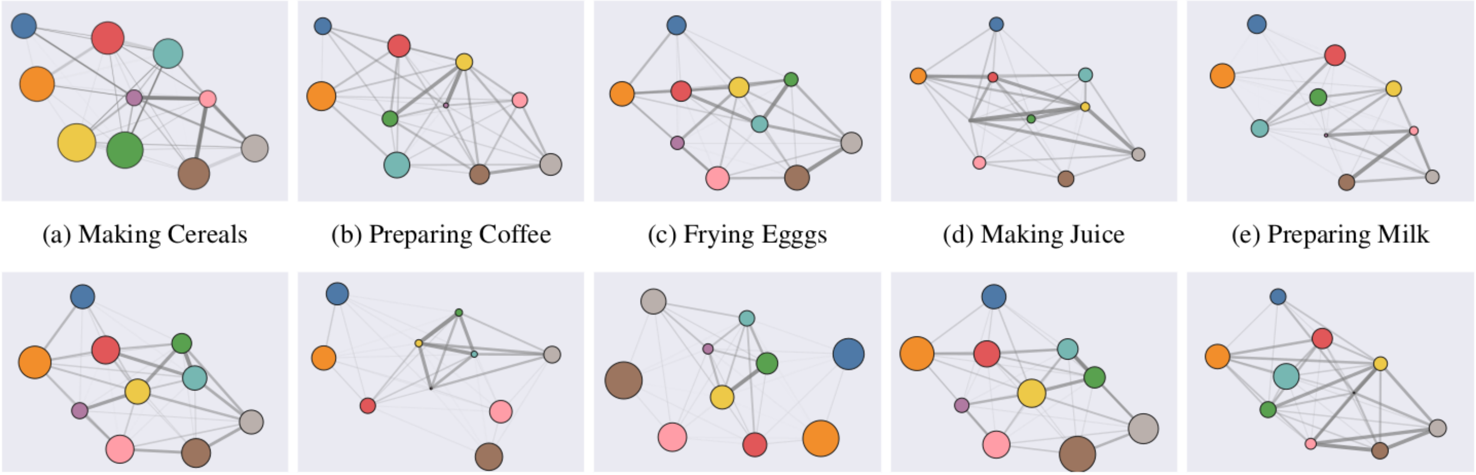
Some results

Charades

Method	Modality	mAP (%)
Two-stream	RGB + Flow	18.6
Two-stream + LSTM	RGB + Flow	17.8
ActionVLAD	RGB + iDT	21.0
Temporal Fields	RGB + Flow	22.4
Temporal Relations	RGB	25.2
ResNet-152	RGB	22.8
ResNet-152 + Timeception	RGB	31.6
I3D	RGB	32.9
I3D + ActionVLAD	RGB	35.4
I3D + Timeception	RGB	37.2
I3D + VideoGraph	RGB	**37.8**

Breakfast

Method	Breakfast Acc. (%)	Breakfast mAP (%)
ResNet-152	41.13	32.65
ResNet-152 + ActionVLAD	55.49	47.12
ResNet-152 + Timeception	57.75	48.47
ResNet-152 + VideoGraph	**59.12**	**49.38**
I3D	47.05	58.61
I3D + ActionVLAD	60.20	65.48
I3D + Timeception	61.82	67.07
I3D + VideoGraph	**63.14**	**69.45**



(a) Making Cereals (b) Preparing Coffee (c) Frying Egggs (d) Making Juice (e) Preparing Milk

(f) Making Pancake (g) Making Salat (h) Making Sandwich (i) Making Scrambled Egg (j) Preparing Tea

● cereal, ● pan, ● eggs, ● sandwitch, ● kettle, and ● foodbox

Code available

AIDA ARTIFICIAL INTELLIGENCE DOCTORAL ACADEMY

www.i-aida.org

Categorical Normalizing Flows via Continuous Transformations

P. Lippe

E. Gavves

Lippe, Gavves, Categorical Normalizing Flows via Continuous Transformations, ICLR 2021

From classifying to generating graphs

- What if we would like to create new, plausible graphs?
- Normalizing Flows is State-of-the-art in generative modelling

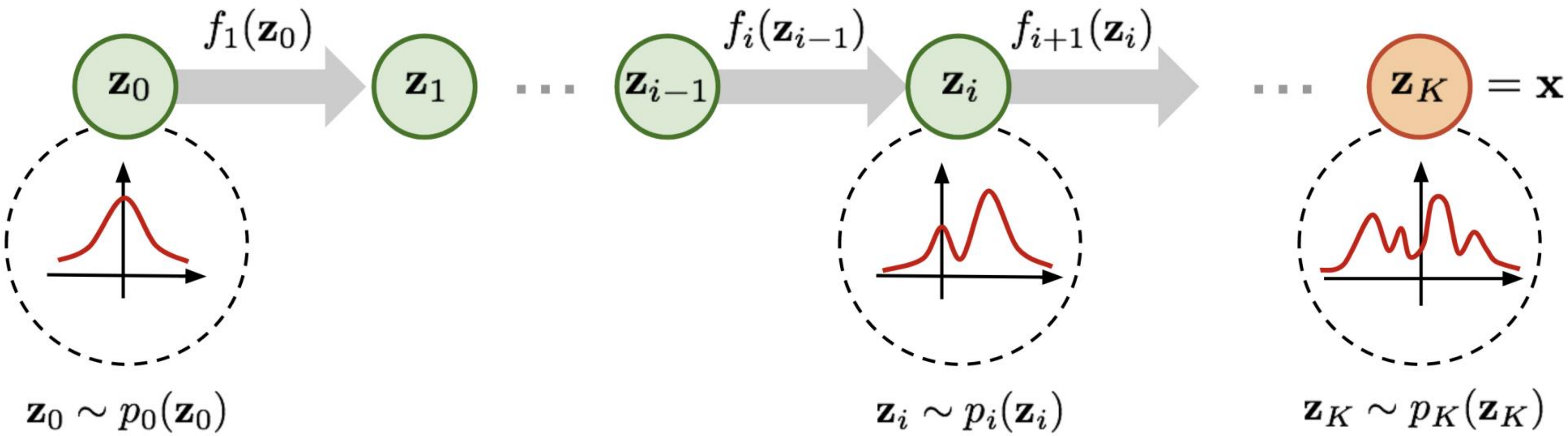
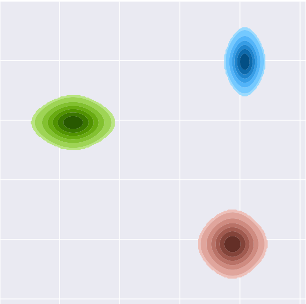


Figure credit: Weng, Lilian. “Flow-based Deep Generative Models”, 2018.

+ Universality
+ Exact likelihood estimate
+ Efficient density evaluation and (parallel) sampling
- Does not work on categorical data seamlessly

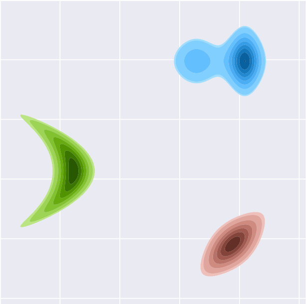
Step 1: Categorical Normalizing Flows

- Learn encoder to represent categorical data in continuous space
  - Must not lose information in the representation
  - Must be smooth and have support for higher dimensions



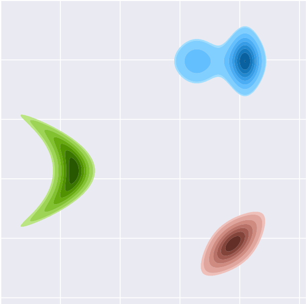
Mixture model

$$q(\boldsymbol{z}|\boldsymbol{x}) = \prod_{i=1}^{N} g(\boldsymbol{z}_i|\mu(x_i), \sigma(x_i))$$



Linear flows

$$q(\boldsymbol{z}|\boldsymbol{x}) = \prod_{i=1}^{N} q(\boldsymbol{z}_i|x_i)$$



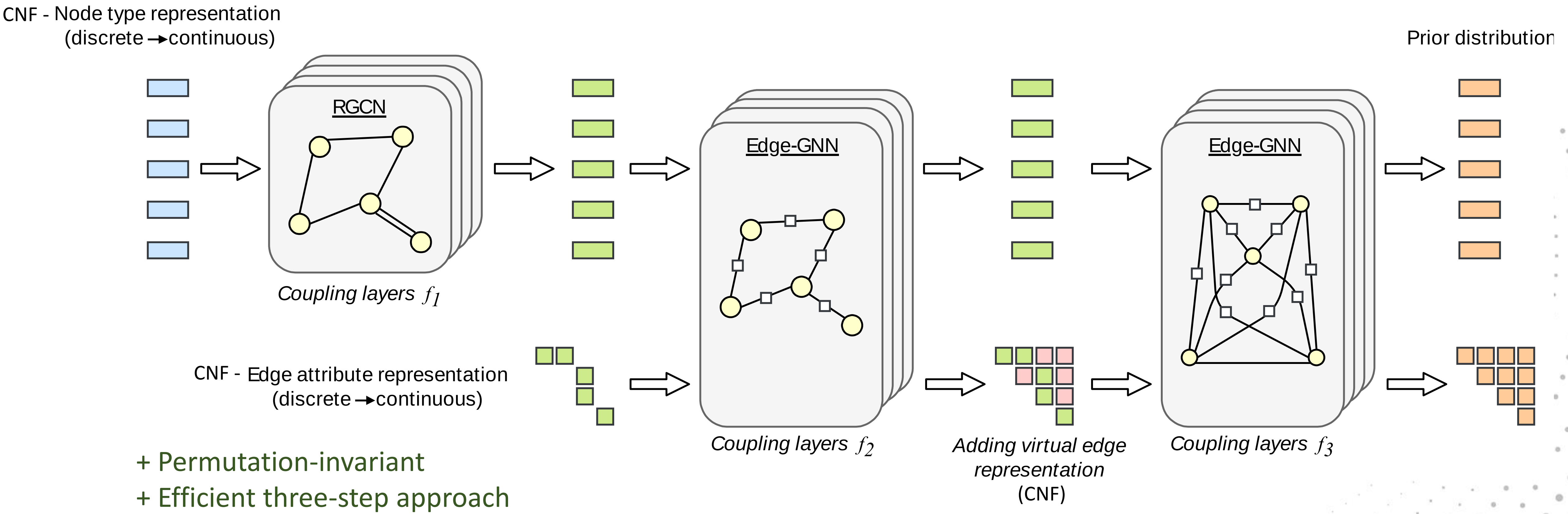
Variational encoding

$$q(\boldsymbol{z}|\boldsymbol{x})$$

- Learn decoder s.t. continuous z contains what is in x exactly
  - Variational inference with factorized decoder

$$p(\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{z}\sim q(\cdot|\boldsymbol{x})}\left[\frac{\prod_i p(x_i|\boldsymbol{z}_i)}{q(\boldsymbol{z}|\boldsymbol{x})} p(\boldsymbol{z})\right]$$

Step 2: Graph generation with CNF



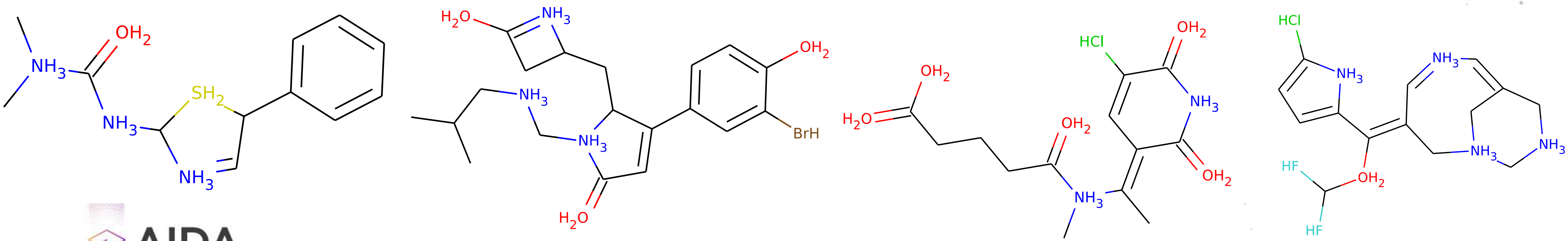
+ Permutation-invariant
+ Efficient three-step approach

Some results

Code available

Molecule generation with Zinc250k dataset (224k examples)

Method	Validity	Uniqueness	Novelty	Reconstruction	Parallel	General
JT-VAE	100%	100%	100%	71%	✗	✗
GraphAF	68%	99.10%	100%	100%	✗	✓
R-VAE	34.9%	100%	–	54.7%	✓	✓
GraphNVP	42.60%	94.80%	100%	100%	✓	✓
GraphCNF	83.41% ($\pm$2.88)	99.99% ($\pm$0.01)	100% ($\pm$0.00)	100% ($\pm$0.00)	✓	✓
+ Sub-graphs	96.35% ($\pm$2.21)	99.98% ($\pm$0.01)	99.98% ($\pm$0.02)	100% ($\pm$0.00)	✓	✓



Rotation Equivariant Siamese Networks for Tracking


D. Gupta


D. Arya


E. Gavves

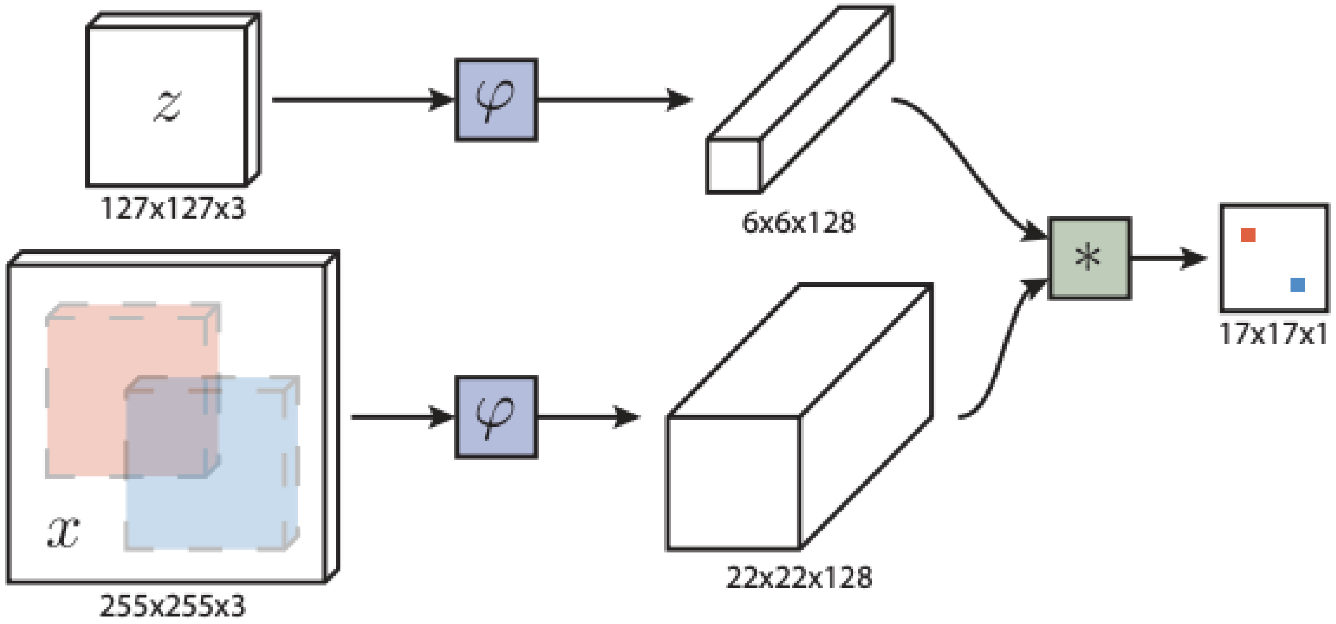
Gupta, Arya, Gavves, Rotation Equivariant Siamese Networks for Tracking, ongoing

In-plane rotations in tracking

- Drone, surveillance, ego-motion recordings, etc.

SoTA: Siamese Trackers [1, 2]

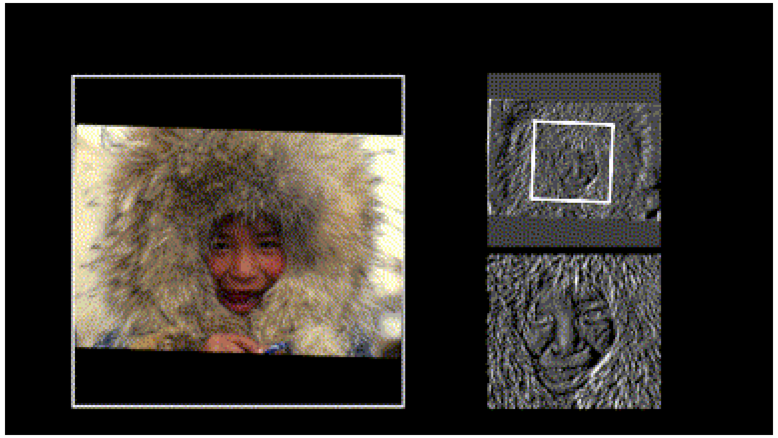
- Tracking as matching the target query to per frame instances
- Sensitive to rotations ← Convolutions are sensitive to rotations



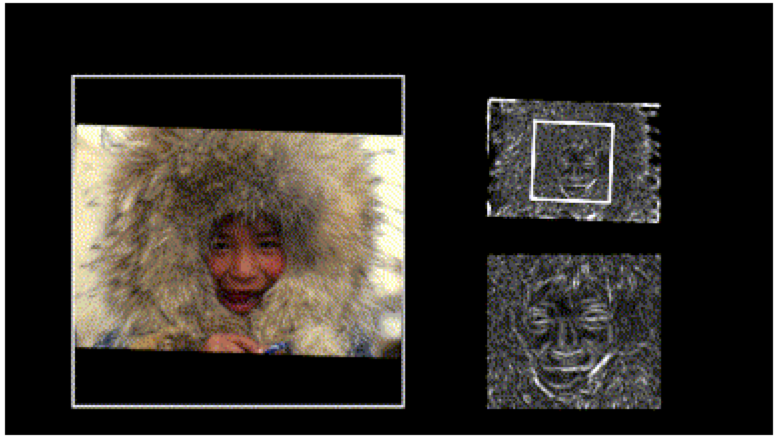
[1] Tao, Gavves, Smeulders, Siamese Instance Search for Tracking, CVPR 2016
[2] Bertinetto et al., Fully-Convolutional Siamese Networks for Object Tracking, ECCV16 2017

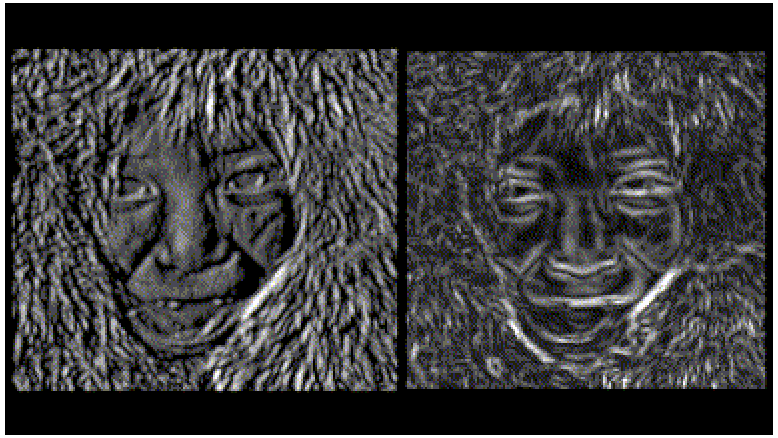
Rotation equivariance in CNNs

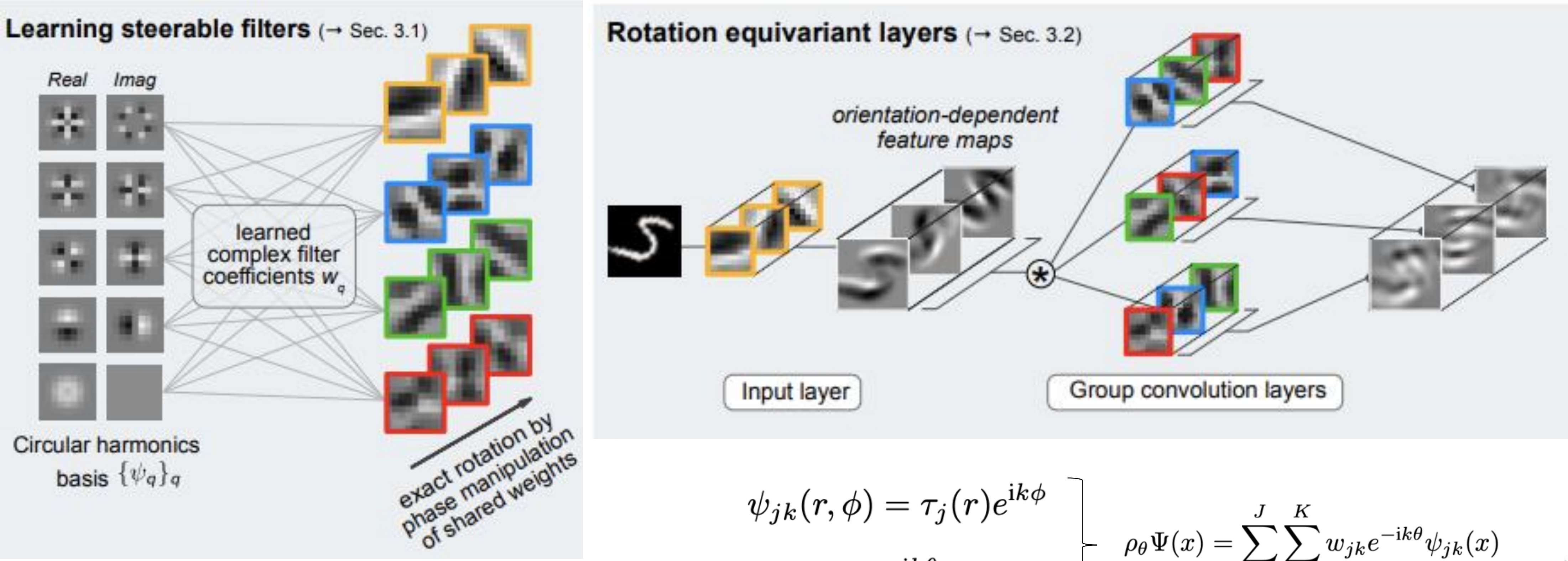
Regular CNN



Rotation Equivariant Net
(H-Net; Worrall et al, CVPR 2017)

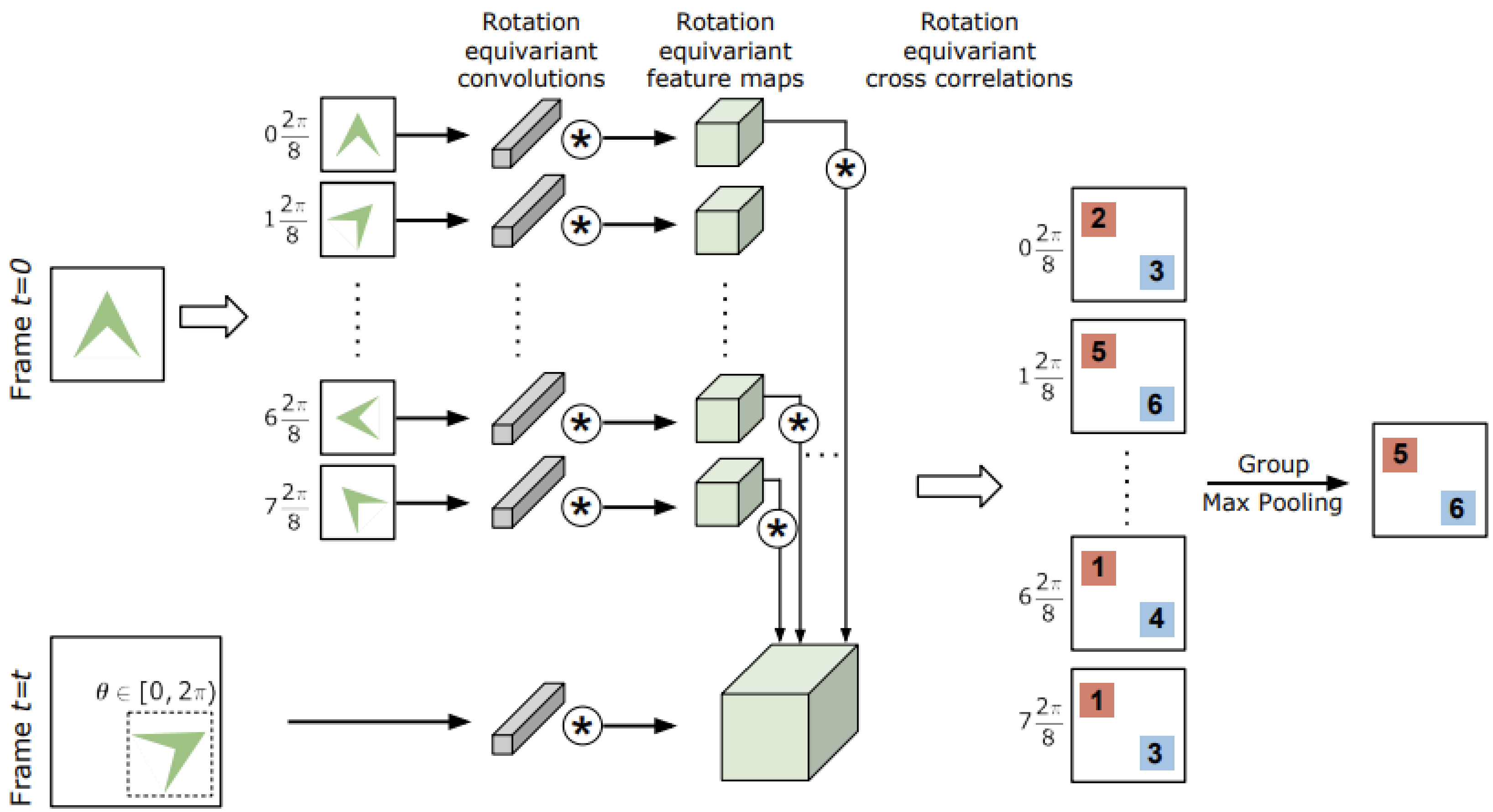






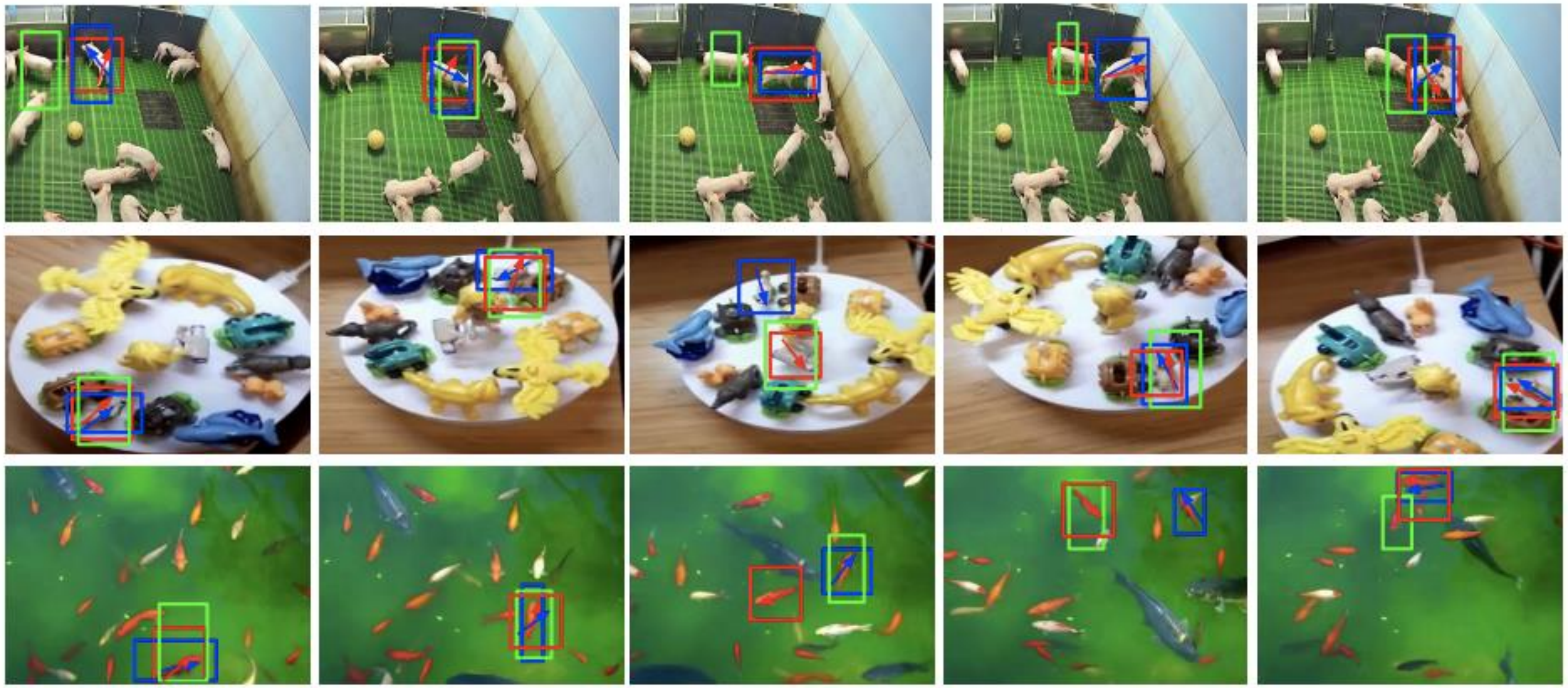
$$\left.\begin{aligned}\psi_{jk}(r,\phi) &= \tau_j(r)e^{\mathrm{i}k\phi}\\ \rho_\theta\psi_{jk}(x) &= e^{-\mathrm{i}k\theta}\psi_{jk}(x)\end{aligned}\right\}\quad \rho_\theta\Psi(x) = \sum_{j=1}^{J}\sum_{k=0}^{K} w_{jk}e^{-\mathrm{i}k\theta}\psi_{jk}(x)$$

Rotation Equivariant Siamese Trackers



Some results

Code soon available



Ground-truth, SiamFC, RE-SiamFC

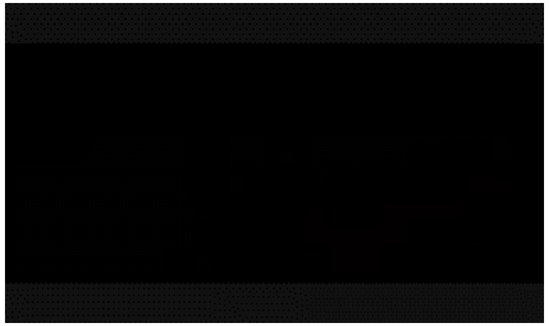
Model	Type	Succ	Pr
SiamFC [1]	-	0.315	0.523
	R4	0.360	0.629
	R8	0.423	0.676
SiamFCv2	-	0.288	0.473
	R4	0.348	0.622
	R8	0.425	0.678
	R16	0.423	0.688
SiamFCv2	aug	0.317	0.541
SiamRPN++ [18]	-	0.461	0.634
SiamRPN++	R4	0.485	0.679
DiMP18 [2]	-	0.429	0.643
DiMP50 [2]	-	0.447	0.668

AIDA ARTIFICIAL INTELLIGENCE DOCTORAL ACADEMY

Time Supervision

(Some) properties of time

Temporal Asymmetry

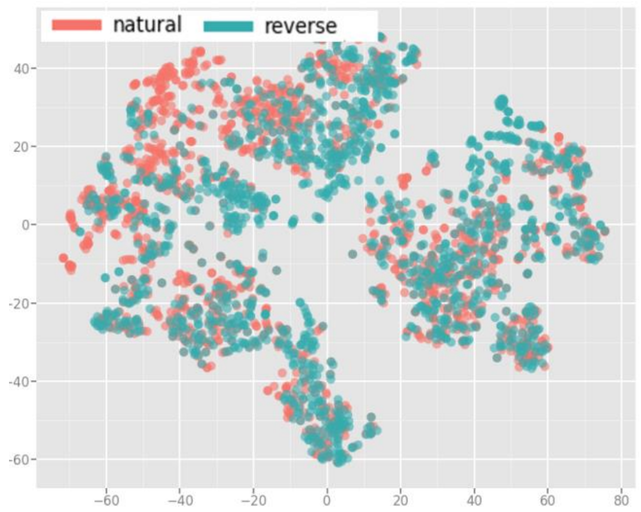


Temporal Causality

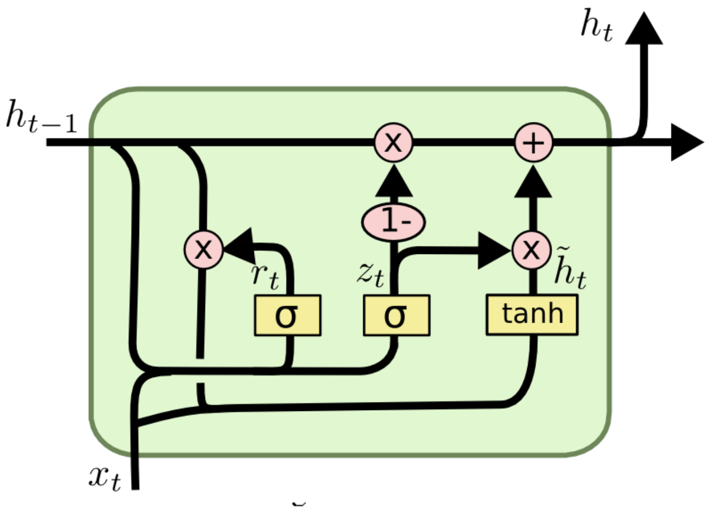
Temporal Continuity

Temporal Redundancy

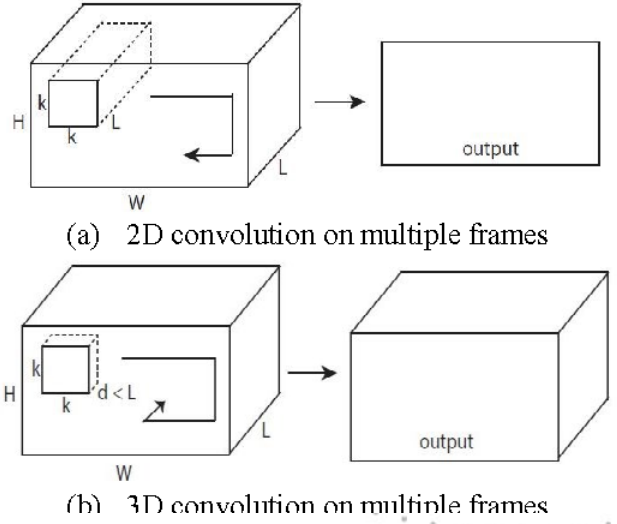
Arrow of time: LSTM vs C3D



LSTM



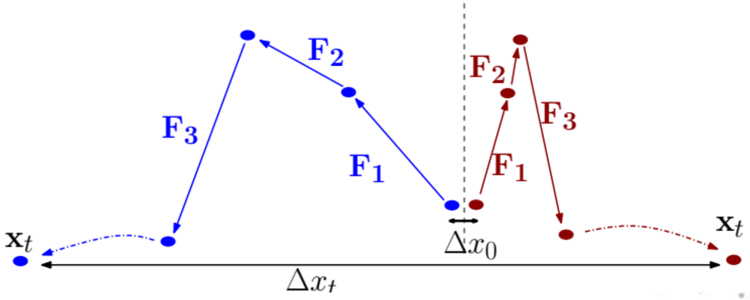
C3D



Revisiting recurrent neural networks

- Recurrent Nets are highly sensitive dynamical systems [1]
  - Even with highly discriminative symbolic (one-hot vector) inputs
  - Gradients very sensitive to initialization → Poor learning! → No generalization

- Visual features are
  - much noisier, less discriminative, much more redundant

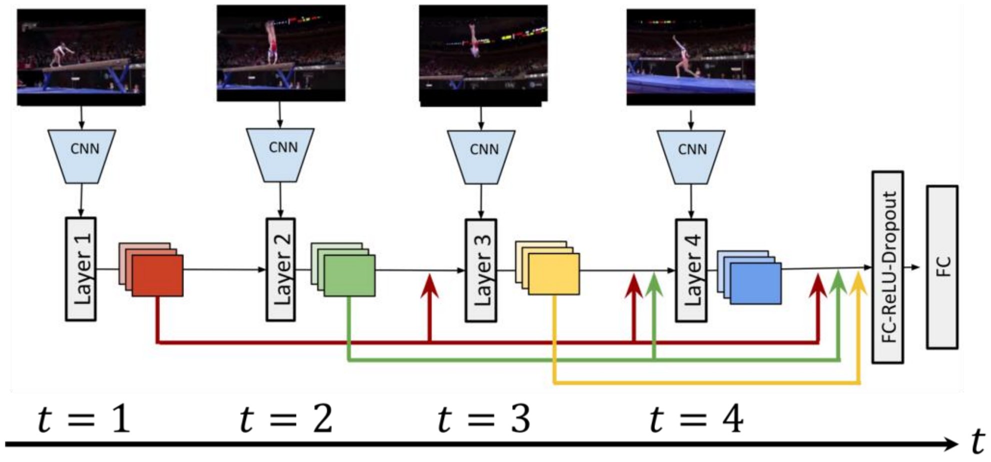


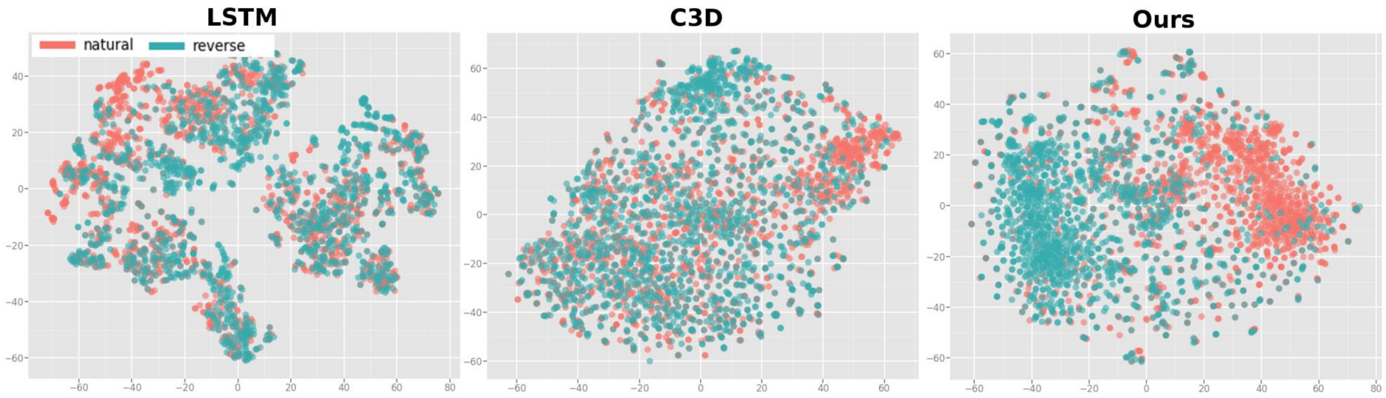
- Learning LSTM on videos is orders of magnitude harder
  - Chaotic regime → no useful gradients → no learning
  - Forward/Backward/Shuffling of frames → LSTM performs the same on arrow of time

[1] Pascanu and Bengio, On the difficulty of training recurrent neural networks, 2013

Time-aligned neural networks [1]

- Idea: Why not flip the ConvNet to align the layers with time steps?
- No vanishing/exploding gradients, no problems with noise/redundancy





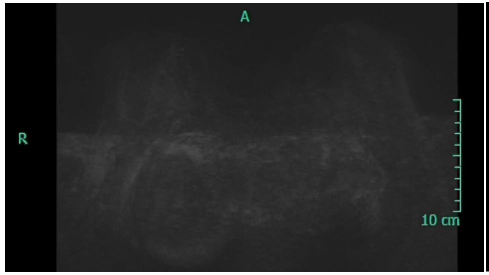
Conclusion: Poor temporal modelling could be due to hard –and thus unsuccessful- optimization

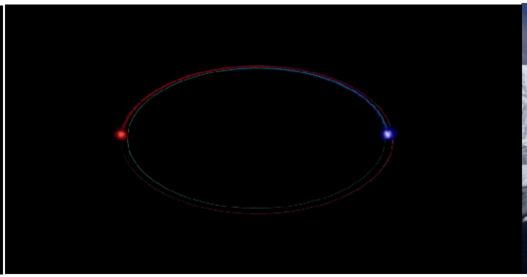
[1] Ghodrati, Gavves, Snoek, Video Time: Encoders, Properties, Evaluation, BMVC, 2018

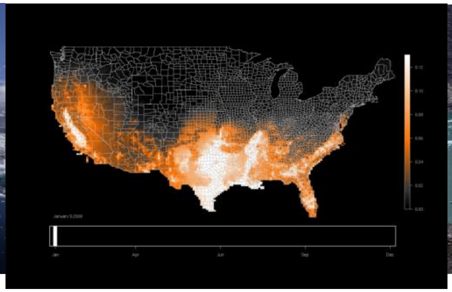
Time Evaluation

Engage-Generalize-Scale

- Current validation paradigm in video is hard to sustain
- Models for (*n*-dimensional + time) signals, e.g. scientific recordings
  - Particles through time, climate, astronomy, ecosystems, biology
- <u>Possible great advantage</u>: scientific knowledge as groundtruth

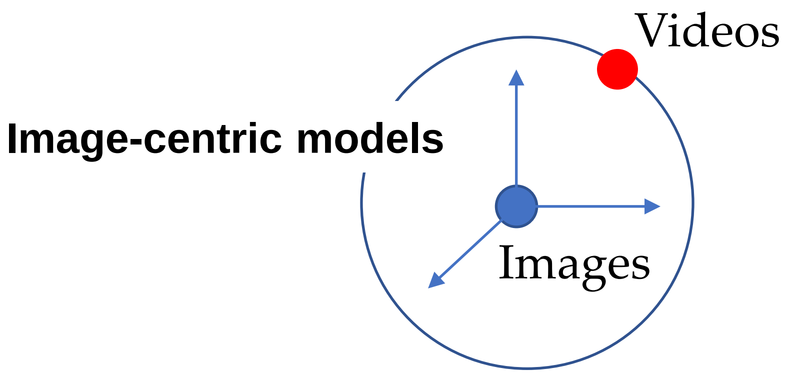






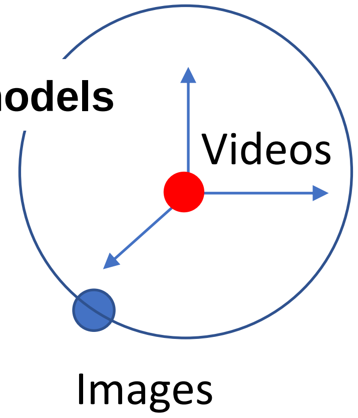

Conclusion

- Time largely ignored in model building and validation



Conclusion

Spatiotemporal-centric models



- Time largely ignored in model building and validation

Static → Temporal

- Geometry, geneartive & time supervision will be the key
- Hopefully, impact on any field with spatiotemporal complex data

AIDA

ARTIFICIAL INTELLIGENCE
DOCTORAL ACADEMY

The International AI Doctoral Academy (AIDA) is a joint initiative of the European R&D Projects: AI4media, VISION and HumanE-AI Net

AI4media | VISION | HUMANE AI NET

www.ai4media.eu | www.vision4eu.eu | www.humane-ai.eu

These projects have received funding from the European Union's Horizon 2020 research and innovation programme under the following Grant agreements: No 951911 (AI4media), No. 952070 (VISION) and No. 952026 (HumanE-AI Net)