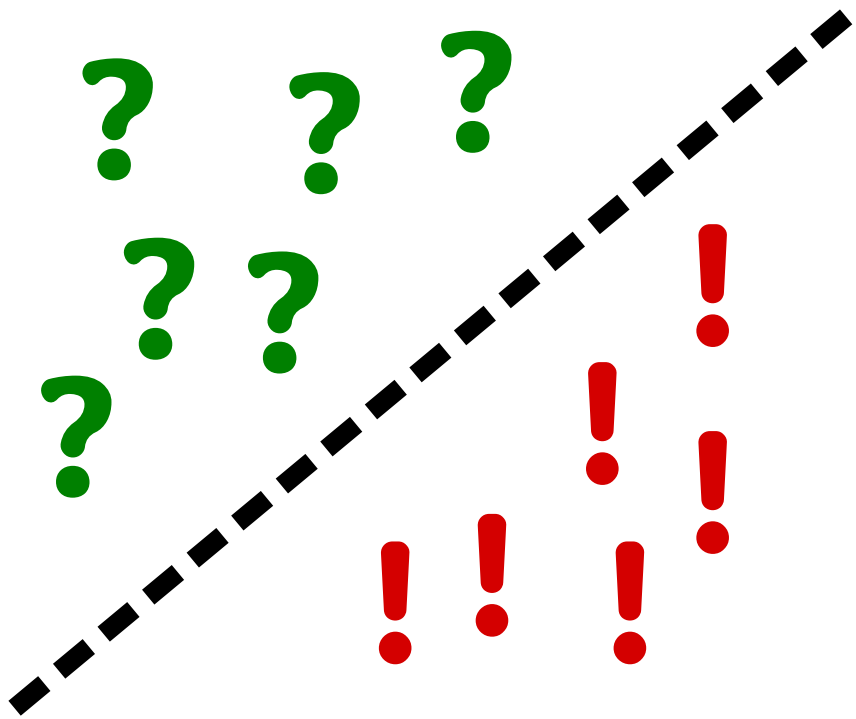
Deep House of Cards?

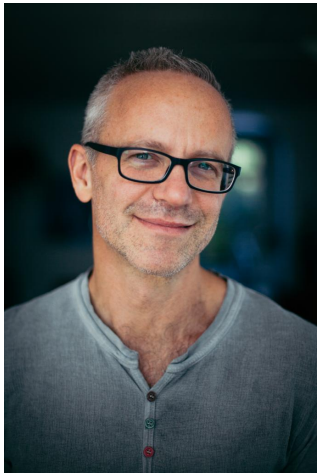
Testing the fundaments in image and video AI



TU Delft Computer Vision Lab
Jan van Gemert

Whoami: Jan van Gemert

Head of the Computer Vision Lab at TU Delft

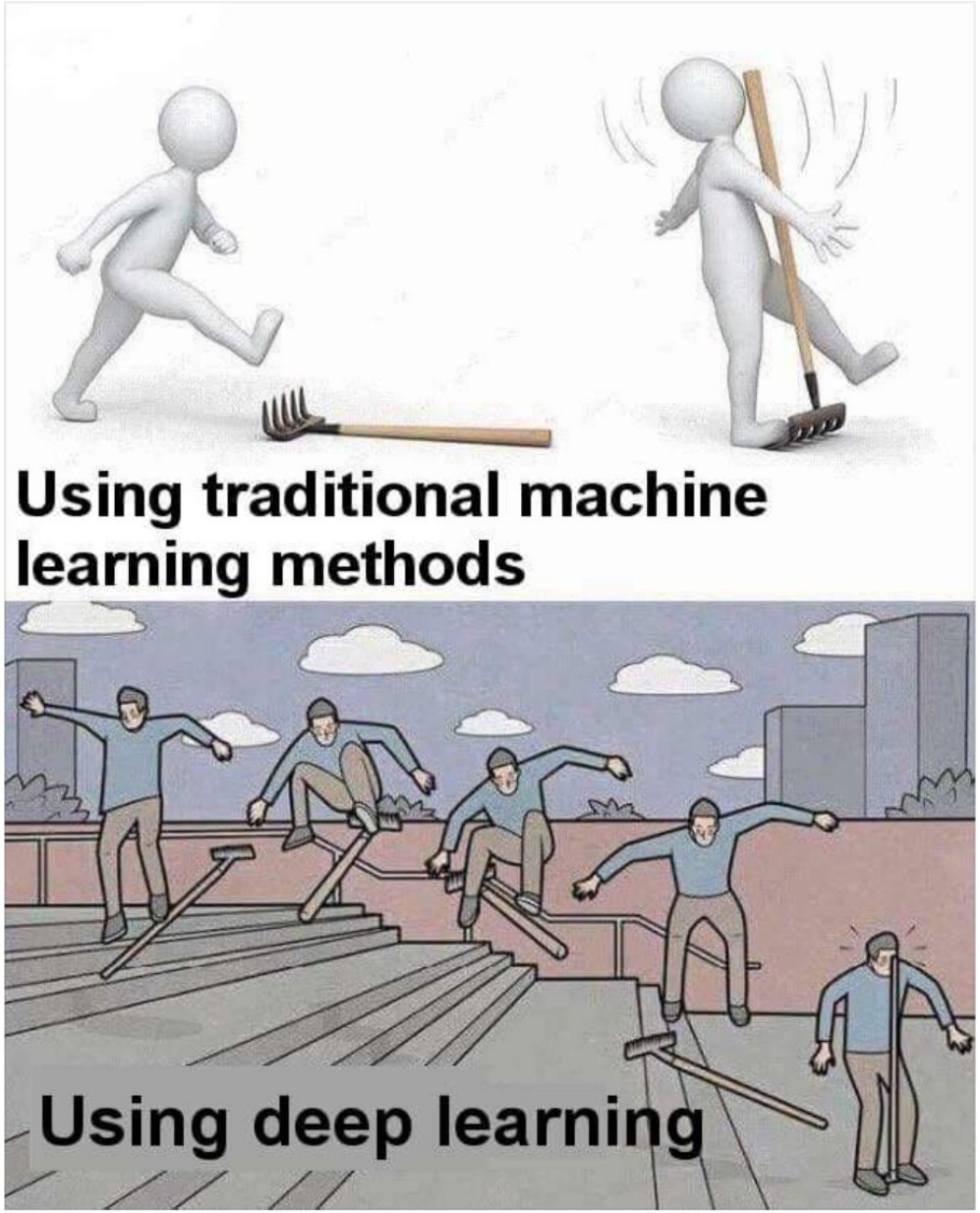


Two main research themes:

1. Fundamental, empirical, understanding-based deep learning research (to)
2. Find, evaluate, and incorporate powerful yet flexible physical priors for data efficient visual recognition AI.

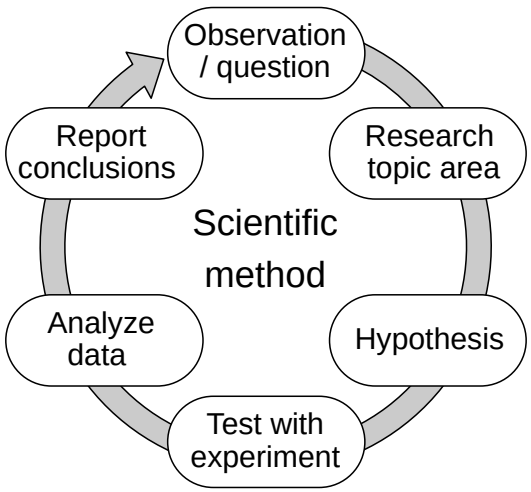
Applied on image, video, action, object, human analysis, ...

Deep learning and Machine learning



The scientific method[1] in times of deep learning

Deep learning is powering the AI revolution.
Yet, as a scientific field, it has growing pains[2,3]



- Improvement-driven (large compute/data)
- Opportunistic (career driven)
- Reviewer damage (Benchmark fetish; Mathiness)
- Confusing speculation with explanation
- Not identifying the reasons for empirical gains.

With bigtech dominating data/compute[4]; lets focus on fundaments.

---

[1]: https://en.wikipedia.org/wiki/Scientific_method

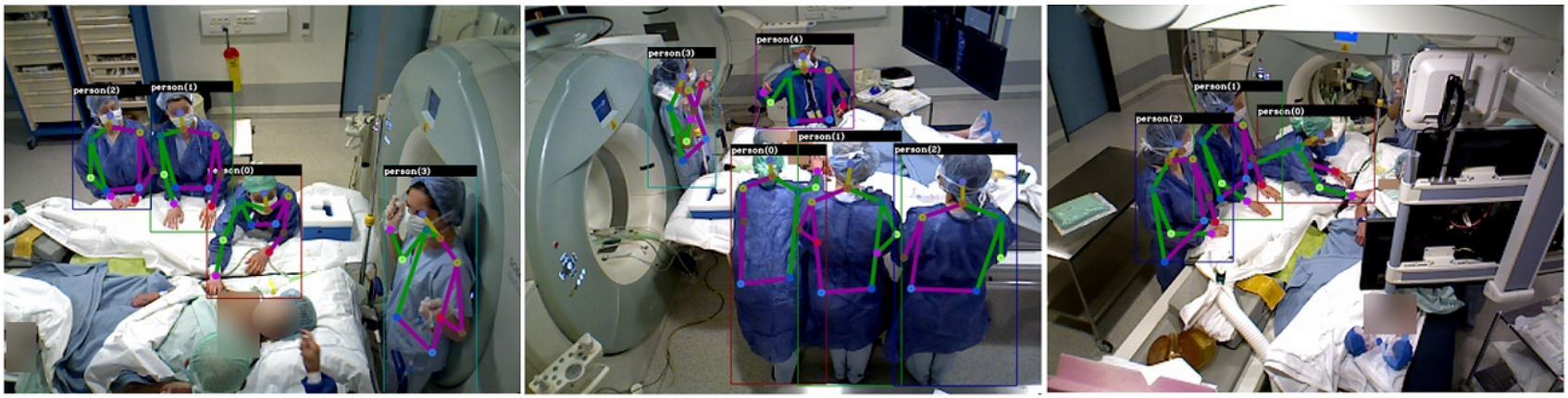
[2]: Lipton et al. "Troubling Trends in Machine Learning Scholarship", 2018.

[3]: Sculley, David, et al. "Winner's curse? On pace, progress, and empirical rigor." 2018.

[4]: Togelius, Julian, et al. "Choose your weapon: Survival strategies for depressed AI academics" Proc. of the IEEE (2024).

Video activity progress prediction

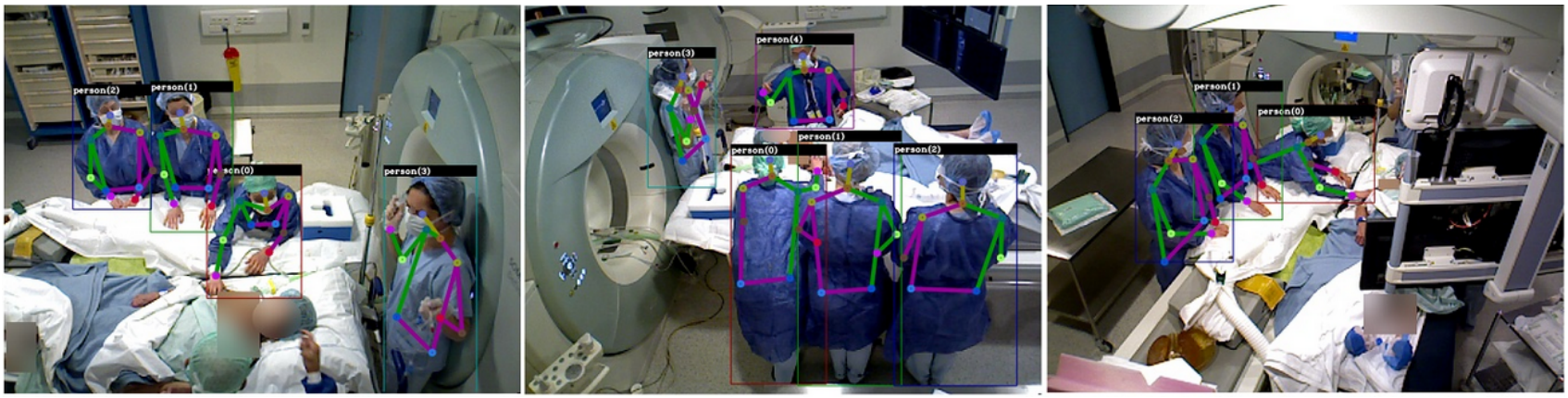
Useful for cooking, surgery scheduling, sports, video editing, etc.[5]



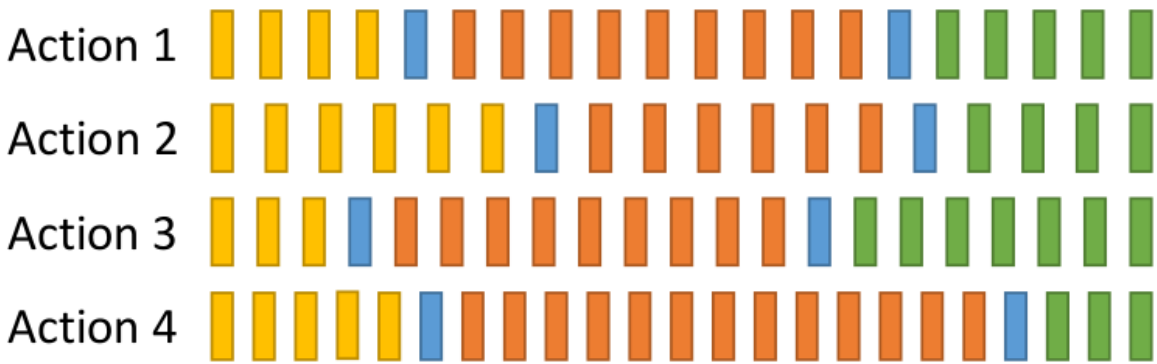
[5]: Becattini, Federico, et al. "Am I done? Predicting action progress in videos." ACM Trans. on Multimedia Computing 2020

Video activity progress prediction

Useful for cooking, surgery scheduling, sports, video editing, etc.[5]



Example of phases (colors) in activities[1]:



[5]: Becattini, Federico, et al. "Am I done? Predicting action progress in videos." ACM Trans. on Multimedia Computing 2020

Video activity progress prediction: Testing fundaments[6]

[6]: Boer et al. "Is there progress in activity progress prediction?" ICCV-w, 2023.

Video activity progress prediction: Testing fundaments[6]

(a) *UCF101-24* on *full-videos*.

(b) *Breakfast* on *full-videos*.

(c) *Cholec80* on *full-videos*.

3 datasets

[6]: Boer et al. "Is there progress in activity progress prediction?" ICCV-w, 2023.

Video activity progress prediction: Testing fundaments[6]

(a) *UCF101-24* on *full-videos*.

(b) *Breakfast* on *full-videos*.

(c) *Cholec80* on *full-videos*.

3 datasets, 5 learning-based methods

[6]: Boer et al. "Is there progress in activity progress prediction?" ICCV-w, 2023.

Video activity progress prediction: Testing fundaments[6]

(a) *UCF101-24* on *full-videos*.

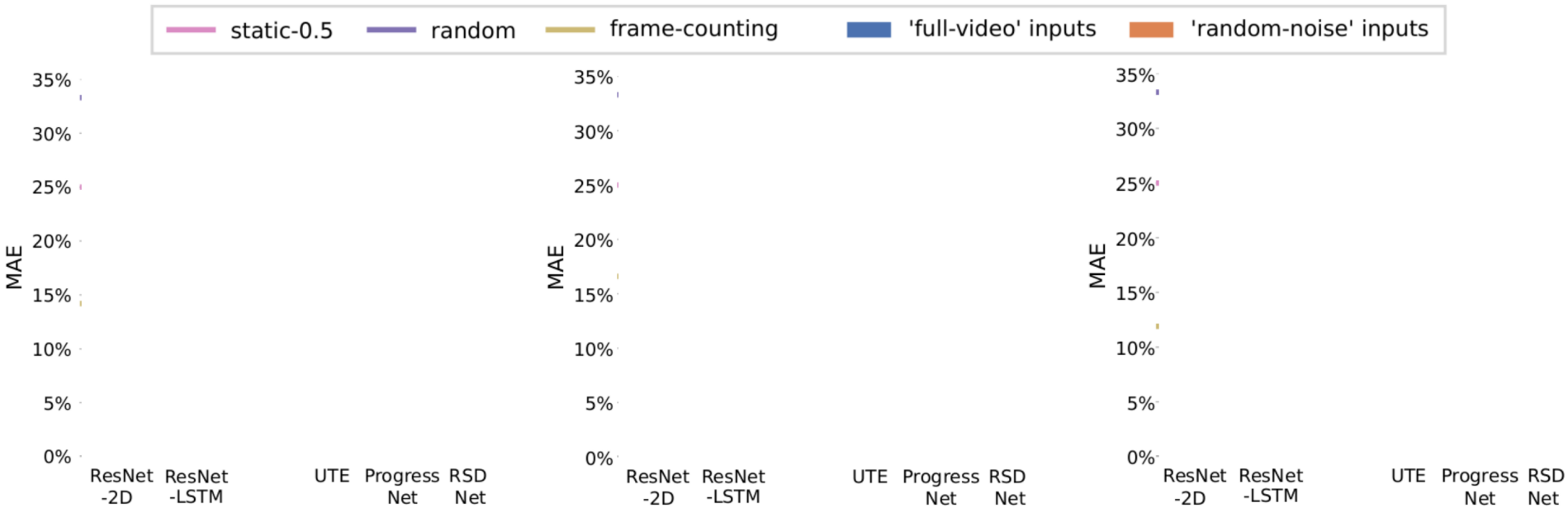
(b) *Breakfast* on *full-videos*.

(c) *Cholec80* on *full-videos*.

3 datasets, 5 learning-based methods, 3 naive baselines
2 inputs evaluated: random noise; or the actual full video.

[6]: Boer et al. "Is there progress in activity progress prediction?" ICCV-w, 2023.

Video activity progress prediction: Testing fundaments[6]

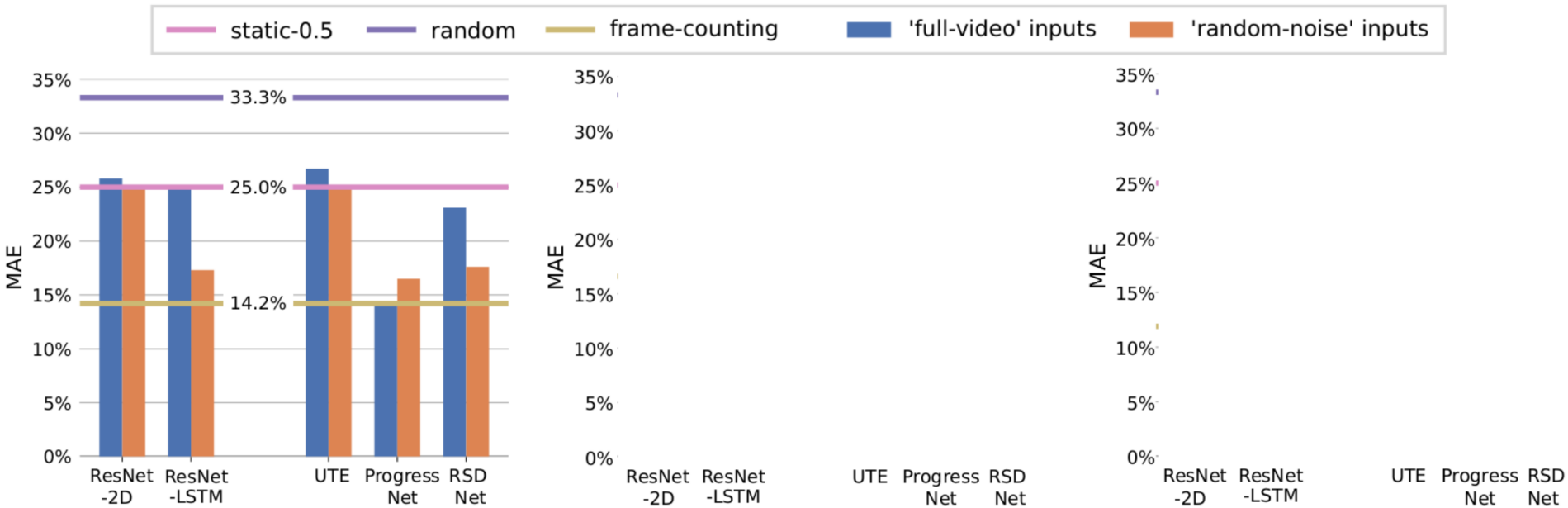


(a) *UCF101-24* on *full-videos*. (b) *Breakfast* on *full-videos*. (c) *Cholec80* on *full-videos*.

3 datasets, 5 learning-based methods, 3 naive baselines
2 inputs evaluated: random noise; or the actual full video.
Mean Average Error (MAE) evaluated.

[6]: Boer et al. "Is there progress in activity progress prediction?" ICCV-w, 2023.

Video activity progress prediction: Testing fundaments[6]

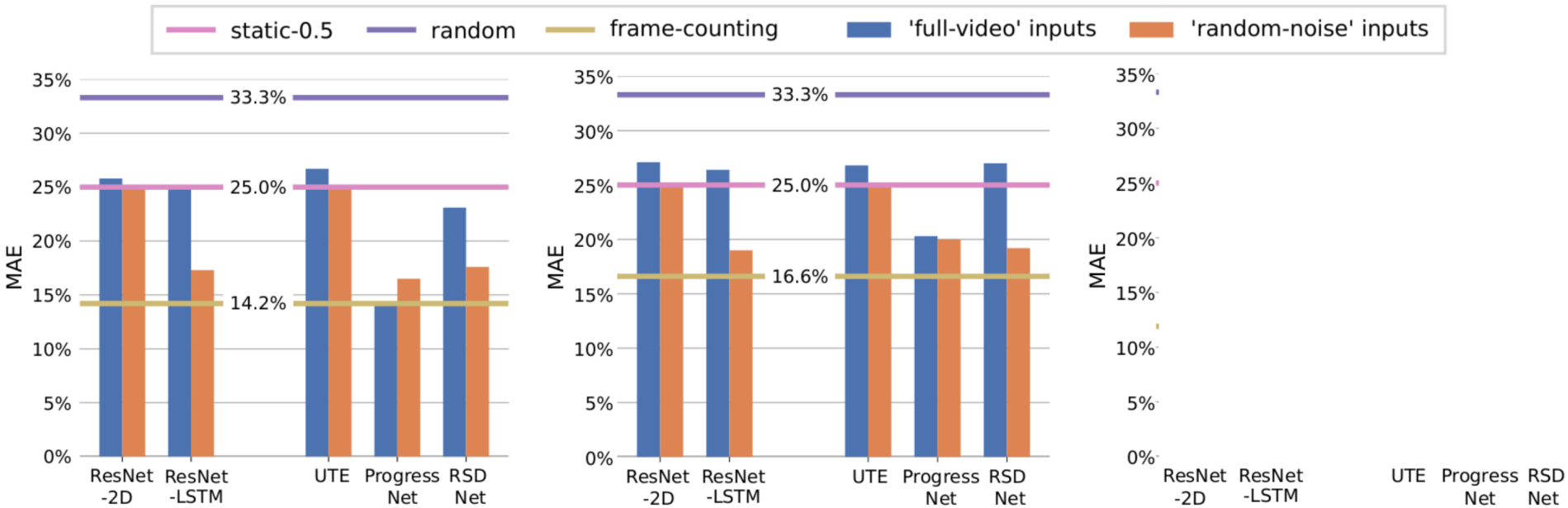


(a) *UCF101-24* on *full-videos*. (b) *Breakfast* on *full-videos*. (c) *Cholec80* on *full-videos*.

3 datasets, 5 learning-based methods, 3 naive baselines
2 inputs evaluated: random noise; or the actual full video.
Mean Average Error (MAE) evaluated.

[6]: Boer et al. "Is there progress in activity progress prediction?" ICCV-w, 2023.

Video activity progress prediction: Testing fundaments[6]

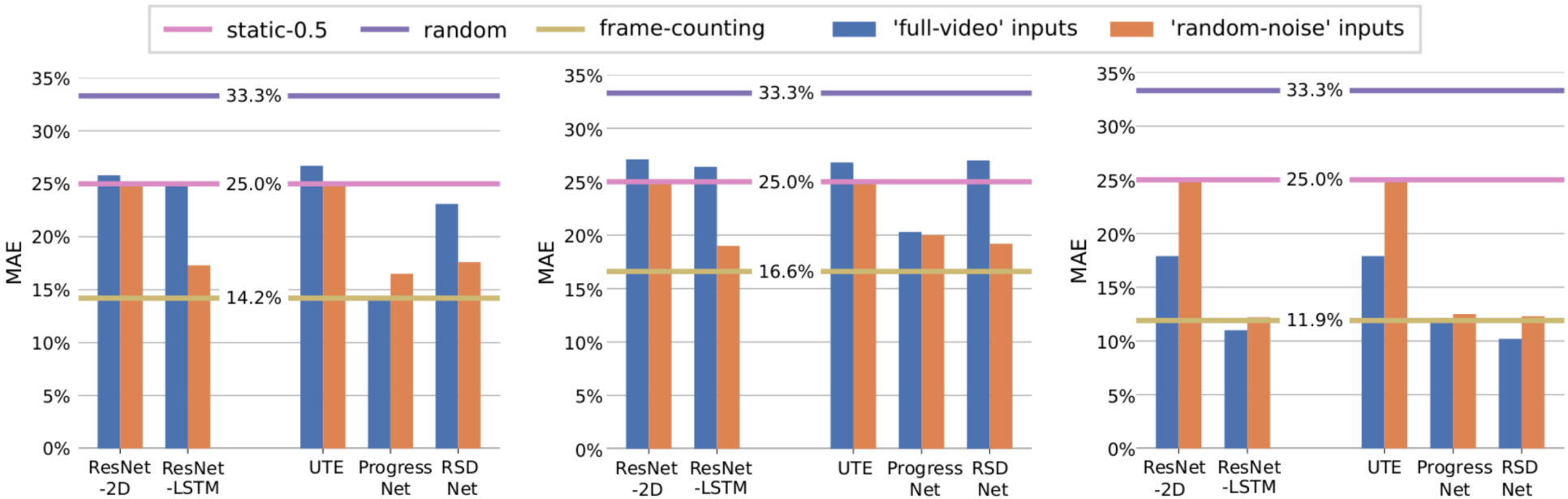


(a) *UCF101-24* on *full-videos*. (b) *Breakfast* on *full-videos*. (c) *Cholec80* on *full-videos*.

3 datasets, 5 learning-based methods, 3 naive baselines
2 inputs evaluated: random noise; or the actual full video.
Mean Average Error (MAE) evaluated.

[6]: Boer et al. "Is there progress in activity progress prediction?" ICCV-w, 2023.

Video activity progress prediction: Testing fundaments[6]

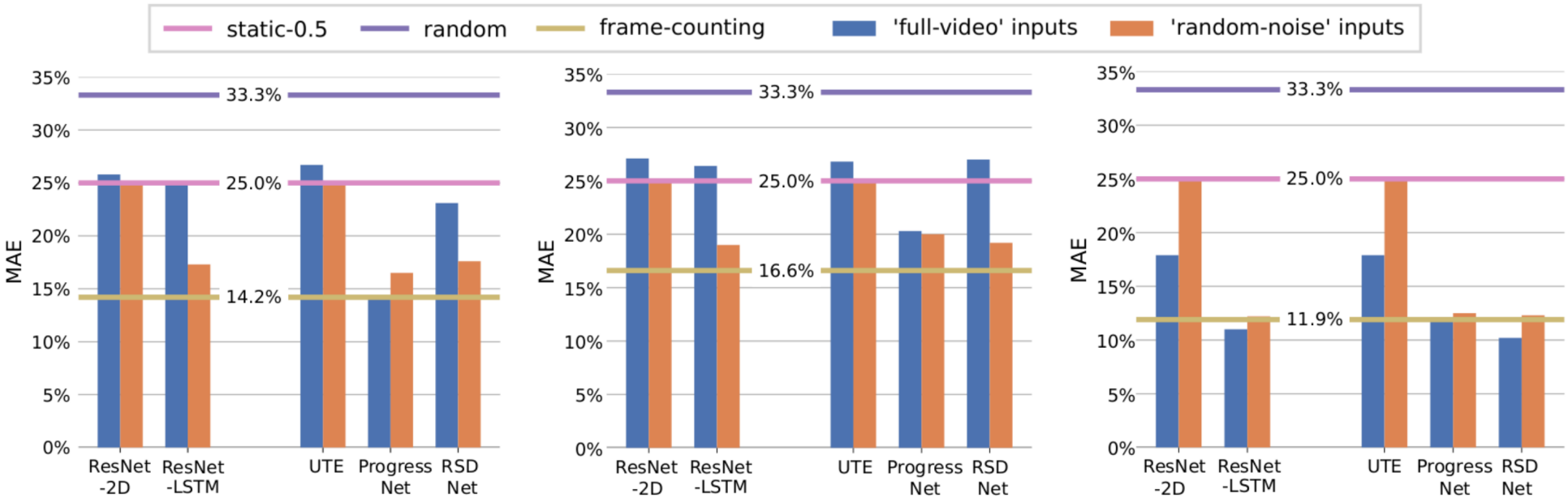


(a) *UCF101-24* on *full-videos*. (b) *Breakfast* on *full-videos*. (c) *Cholec80* on *full-videos*.

3 datasets, 5 learning-based methods, 3 naive baselines
2 inputs evaluated: random noise; or the actual full video.
Mean Average Error (MAE) evaluated.

[6]: Boer et al. "Is there progress in activity progress prediction?" ICCV-w, 2023.

Video activity progress prediction: Testing fundaments[6]



(a) *UCF101-24* on *full-videos*. (b) *Breakfast* on *full-videos*. (c) *Cholec80* on *full-videos*.

3 datasets, 5 learning-based methods, 3 naive baselines
2 inputs evaluated: random noise; or the actual full video.
Mean Average Error (MAE) evaluated.

- Random noise as input works well (combats visual overfitting?)
- Framecounting is hard to beat.

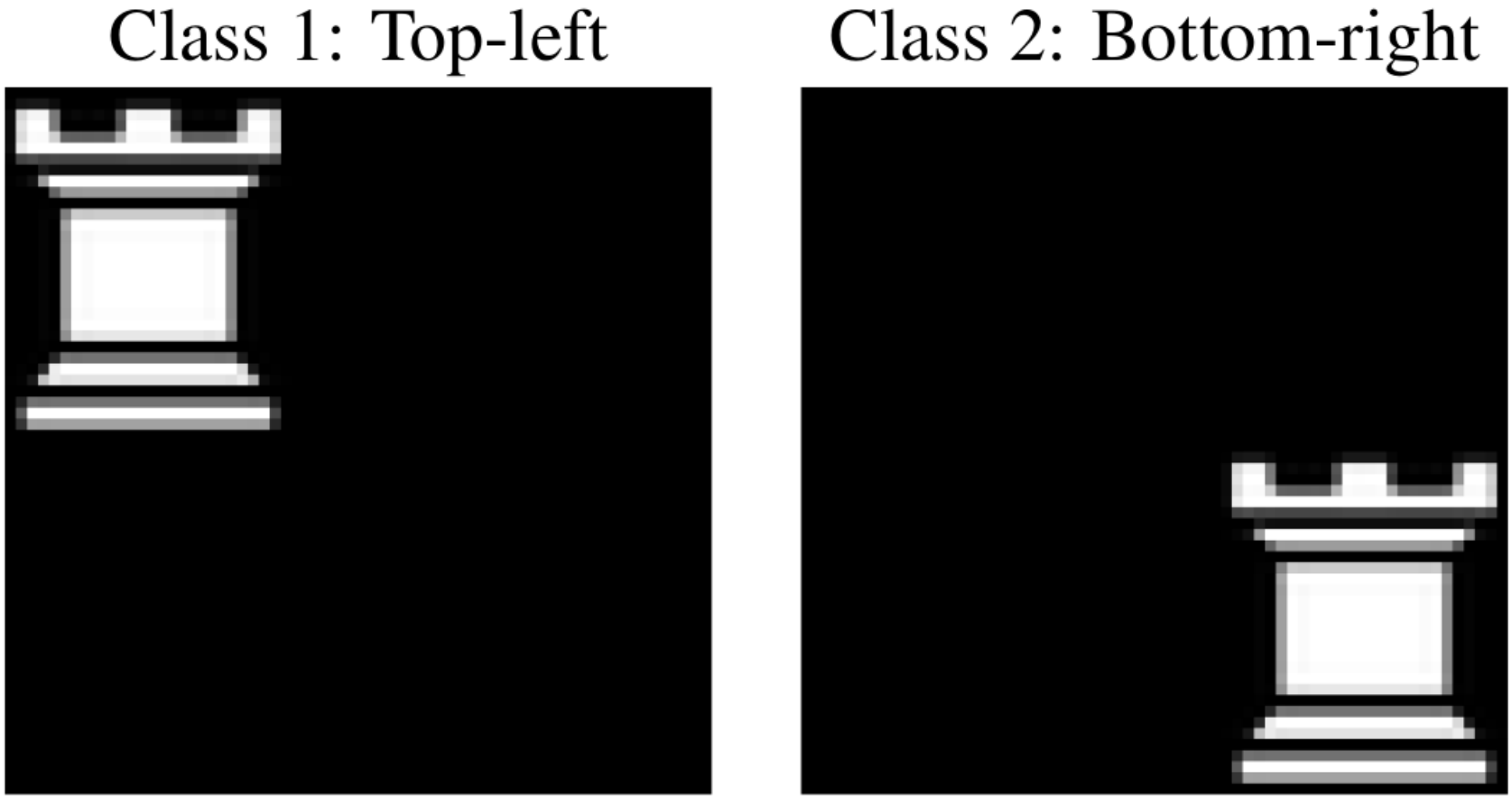
Testing fundaments gives insight!

[6]: Boer et al. "Is there progress in activity progress prediction?" ICCV-w, 2023.

Translation invariance in CNNs[7]

[7]: Kayhan et al. "On Translation Invariance in CNNs: Convolutional Layers can Exploit Absolute Spatial Location", CVPR, 2020.

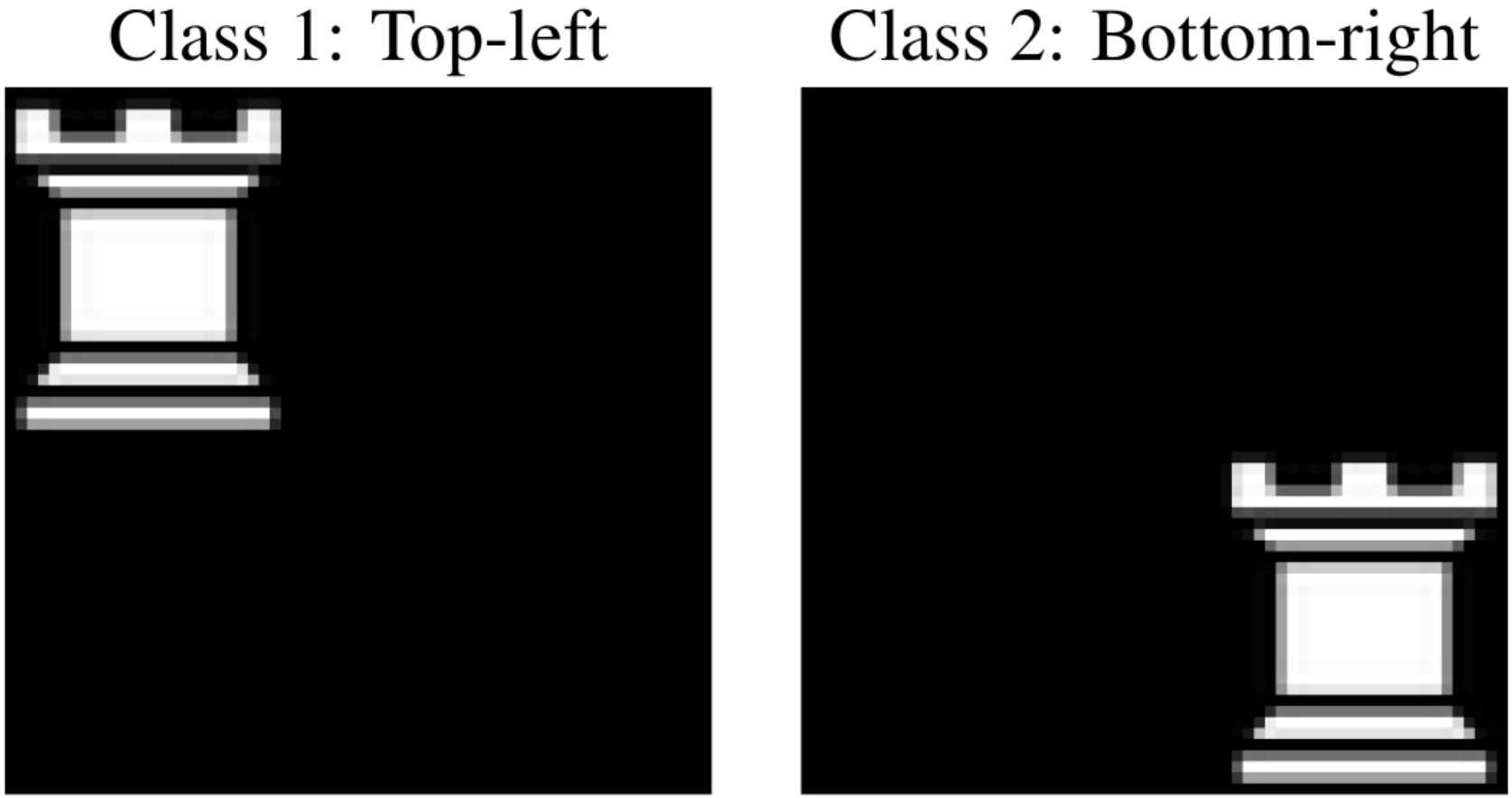
Translation invariance in CNNs[7]



Single conv layer, single 5x5 kernel, zero-padding, ReLu, global max pooling, SGD, and a soft-max loss.

[7]: Kayhan et al. "On Translation Invariance in CNNs: Convolutional Layers can Exploit Absolute Spatial Location", CVPR, 2020.

Translation invariance in CNNs[7]

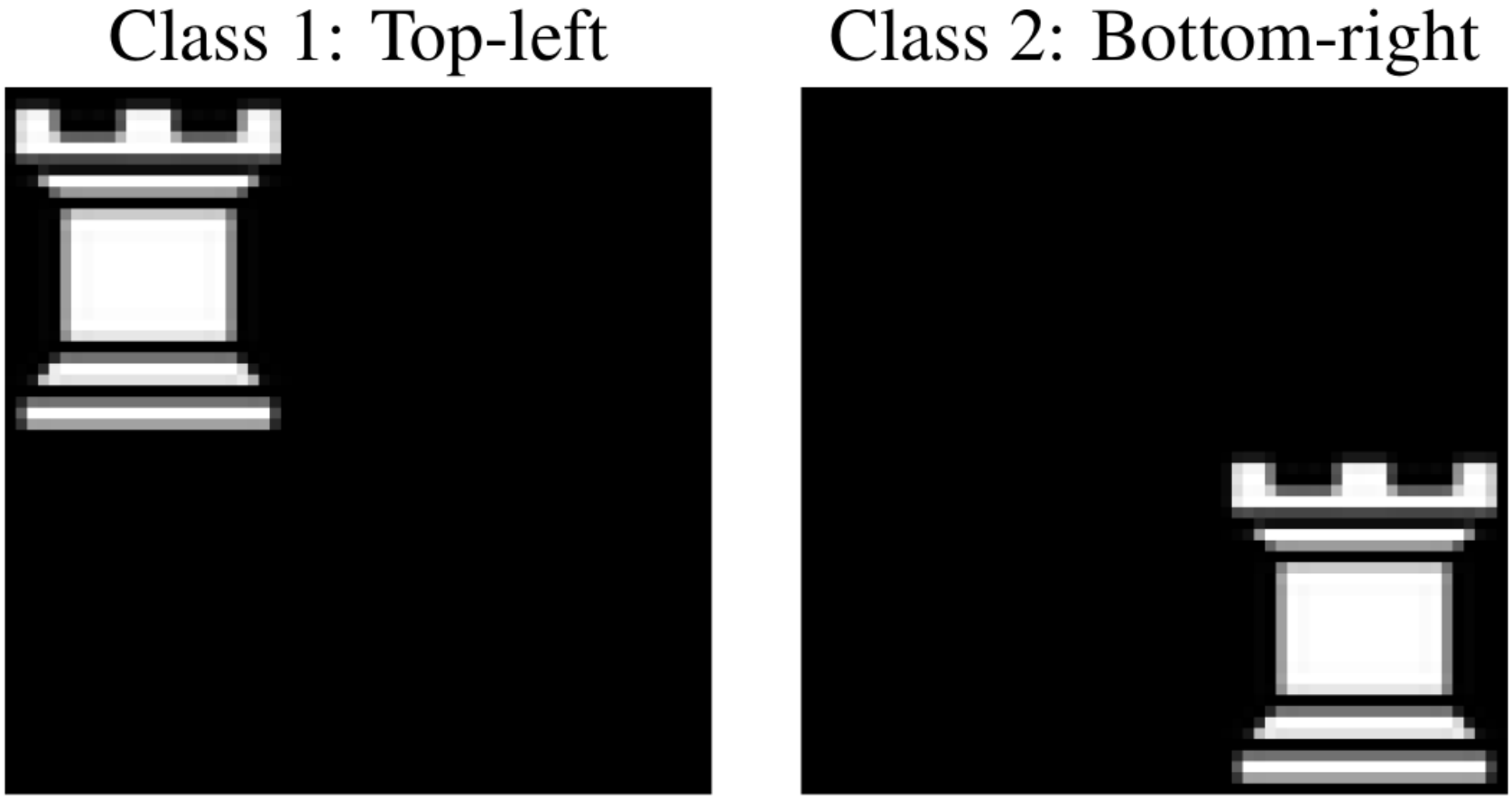


Single conv layer, single 5x5 kernel, zero-padding, ReLu, global max pooling, SGD, and a soft-max loss.

Can it predict the classes?

[7]: Kayhan et al. "On Translation Invariance in CNNs: Convolutional Layers can Exploit Absolute Spatial Location", CVPR, 2020.

Translation invariance in CNNs[7]

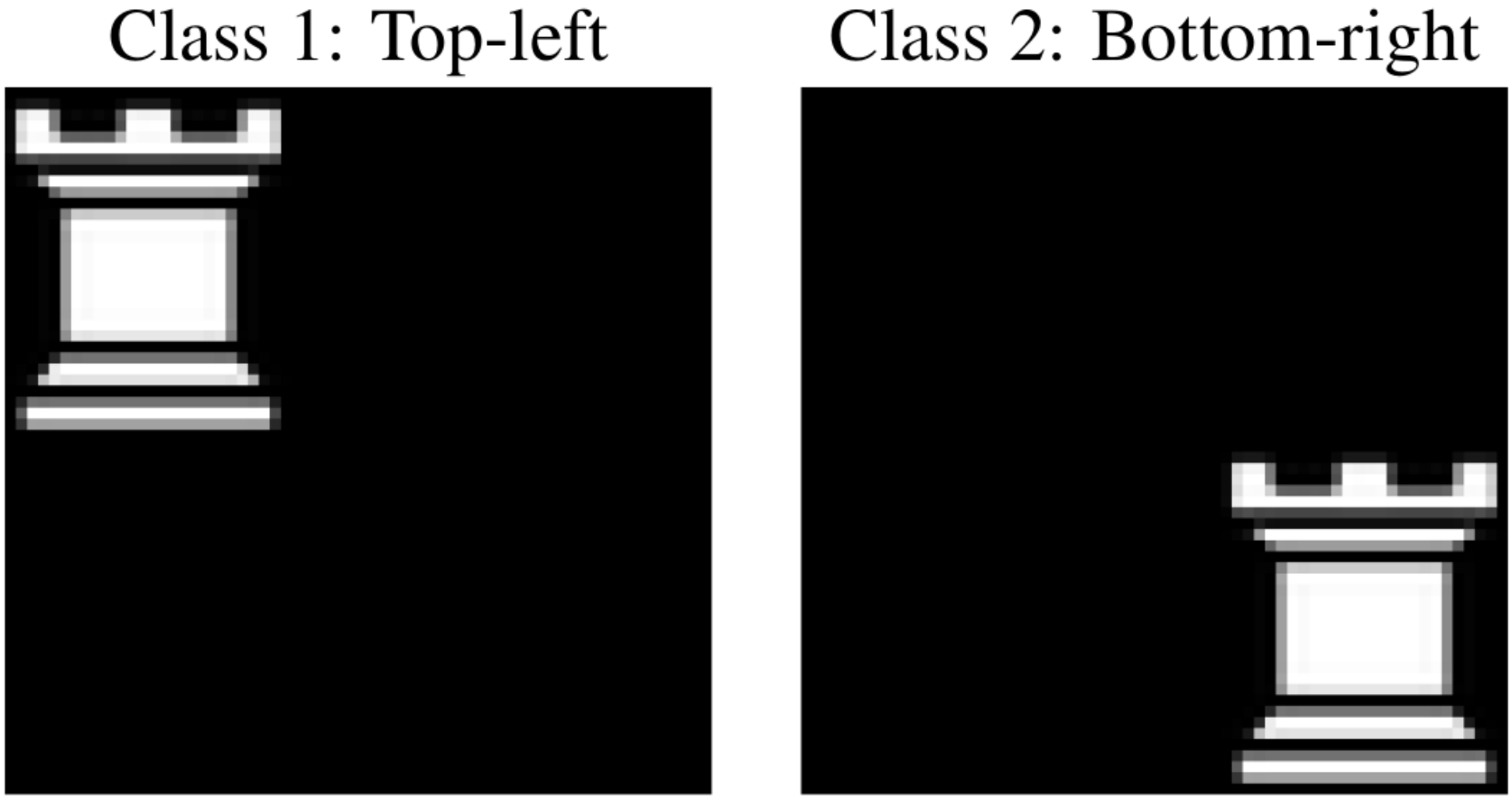


Single conv layer, single 5x5 kernel, zero-padding, ReLu, global max pooling, SGD, and a soft-max loss.

Can it predict the classes? Yes.

[7]: Kayhan et al. "On Translation Invariance in CNNs: Convolutional Layers can Exploit Absolute Spatial Location", CVPR, 2020.

Translation invariance in CNNs[7]



Single conv layer, single 5x5 kernel, zero-padding, ReLu, global max pooling, SGD, and a soft-max loss.

Can it predict the classes? Yes.

Even in standard architectures, commonly believed 'truths' may be subtle.

[7]: Kayhan et al. "On Translation Invariance in CNNs: Convolutional Layers can Exploit Absolute Spatial Location", CVPR, 2020.

Long-term video analysis[8]

Long-term understanding: temporal reasoning over short-term actions

Long term video analysis goes beyond short-term action recognition

[8]: Strafforello et al. "Are current long-term video understanding datasets long-term?", ICCV-w, 2023.

Long-term video analysis[8]: Testing fundaments

Long-term understanding: temporal reasoning over short-term actions

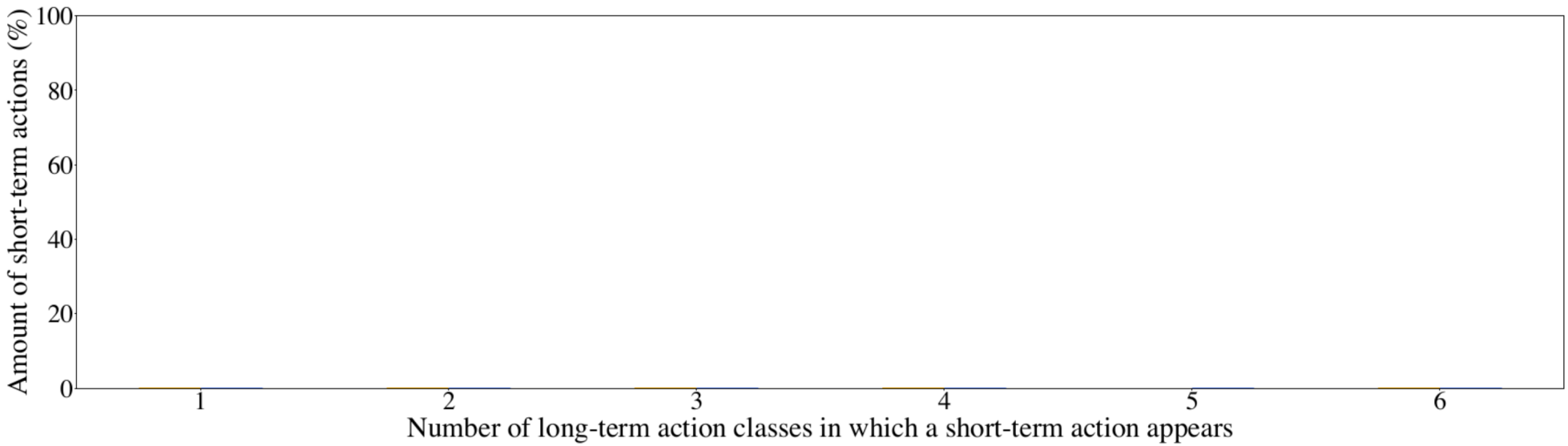
Analyze long-term vs short-term actions

[8]: Strafforello et al. "Are current long-term video understanding datasets long-term?", ICCV-w, 2023.

Long-term video analysis[8]: Testing fundaments

Long-term understanding: temporal reasoning over short-term actions

Analyze long-term vs short-term actions

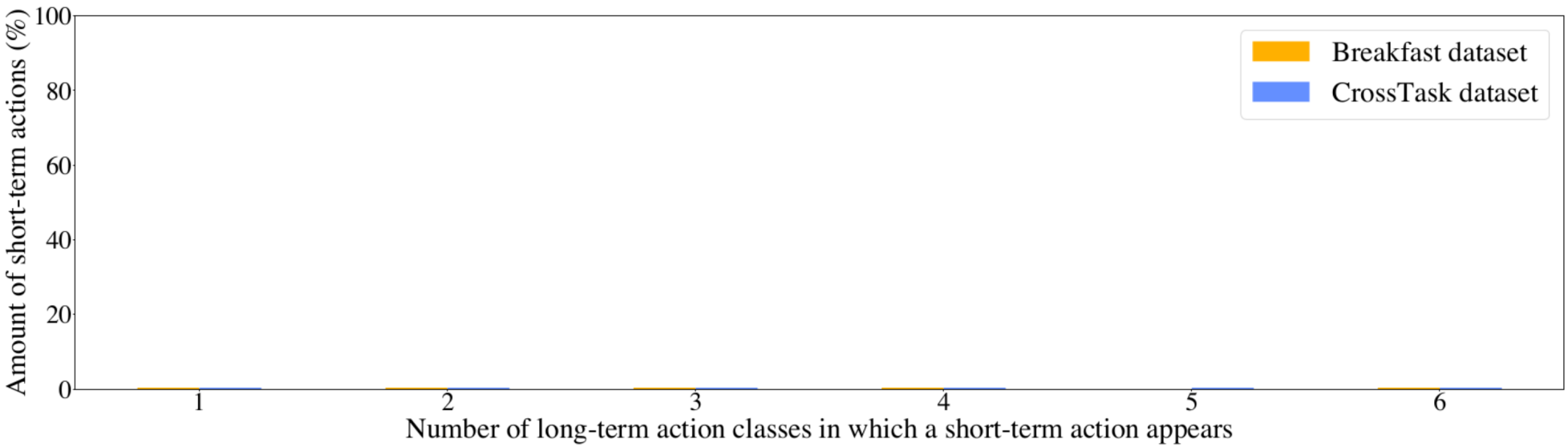


[8]: Strafforello et al. "Are current long-term video understanding datasets long-term?", ICCV-w, 2023.

Long-term video analysis[8]: Testing fundaments

Long-term understanding: temporal reasoning over short-term actions

Analyze long-term vs short-term actions

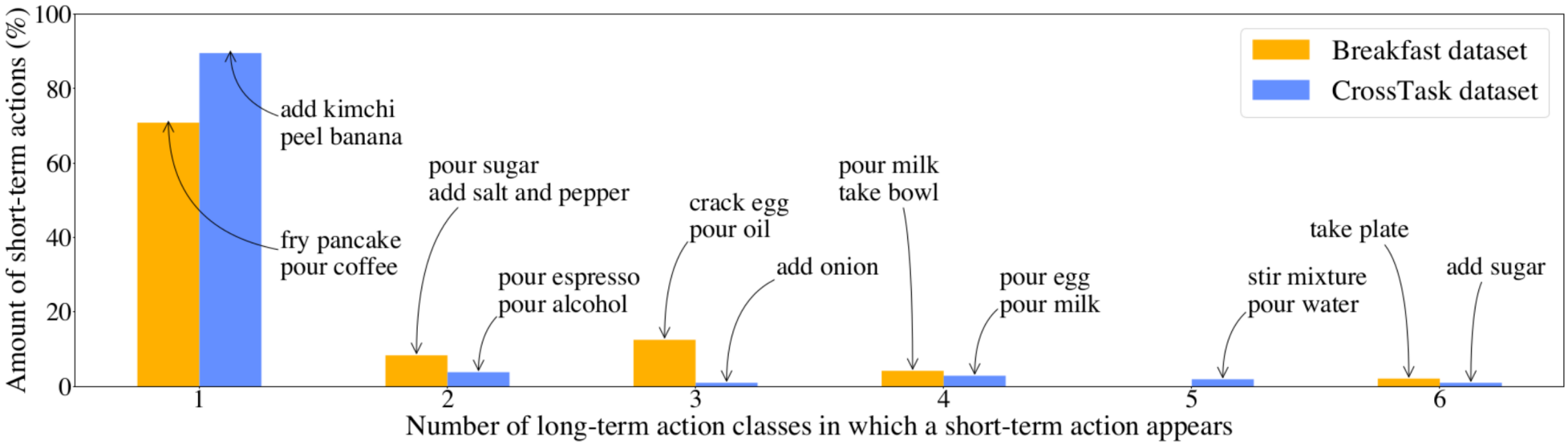


[8]: Strafforello et al. "Are current long-term video understanding datasets long-term?", ICCV-w, 2023.

Long-term video analysis[8]: Testing fundaments

Long-term understanding: temporal reasoning over short-term actions

Analyze long-term vs short-term actions



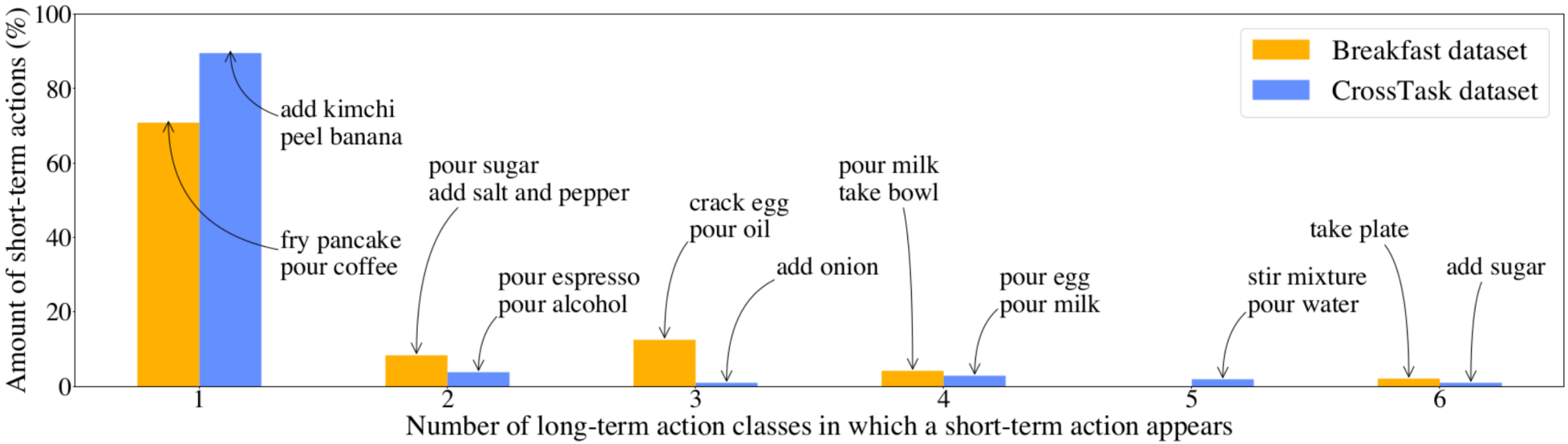
- Most short-term actions appear only in one long-term action class

[8]: Strafforello et al. "Are current long-term video understanding datasets long-term?", ICCV-w, 2023.

Long-term video analysis[8]: Testing fundaments

Long-term understanding: temporal reasoning over short-term actions

Analyze long-term vs short-term actions



- Most short-term actions appear only in one long-term action class
- Recognizing a single short-term action is sufficient: not long-term

[8]: Strafforello et al. "Are current long-term video understanding datasets long-term?", ICCV-w, 2023.

Long-term video analysis[8]

Long-term understanding: temporal reasoning over short-term actions

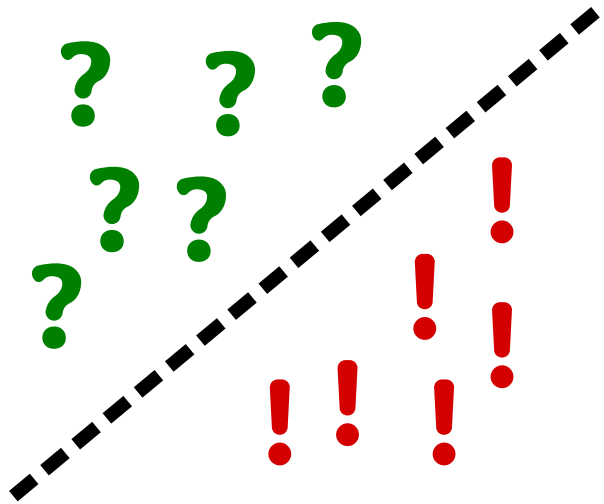
Dataset	Classification accuracy (%)	
	Full Videos	Video Segments
Breakfast	**93.33**	90.0
CrossTask	**100.0**	97.2
LVU – Relationship	**88.89**	**88.89**
LVU – Scene	**100.0**	**100.0**
LVU – Speaking	**80.0**	60.0

Table 2: Average video recognition accuracy obtained from the *Full Videos Survey* and *Video Segments Survey* on the Breakfast [24], CrossTask [49] and LVU [41] datasets. The results suggest that long-term information is helpful but not necessary in the majority of the evaluated datasets.

[8]: Strafforello et al. "Are current long-term video understanding datasets long-term?", ICCV-w, 2023.

Discussion

Deep learning research is great! Amazingly fast progress.
It's important to get results and ideas out, so we can build on them.



As researchers it's our job to understand;
and to rigorously evaluate scientific claims.

Keep testing the fundaments!

My fundamental, emperical, understanding-based deep learning research guidelines: `http://jvgemert.github.io/links.html`