Look-ahead before you leap:
End-to-end active recognition by forecasting the effect of motion

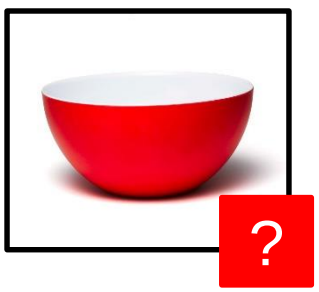
Dinesh Jayaraman and Kristen Grauman

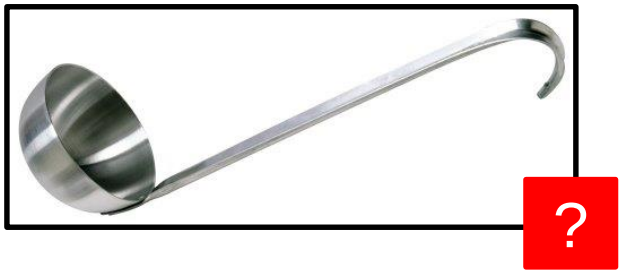
THE UNIVERSITY OF
TEXAS
AT AUSTIN

Status quo: passive snapshot recognition

Status quo: passive snapshot recognition

Object recognition

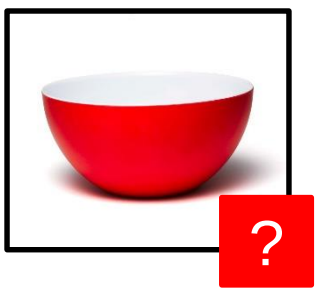


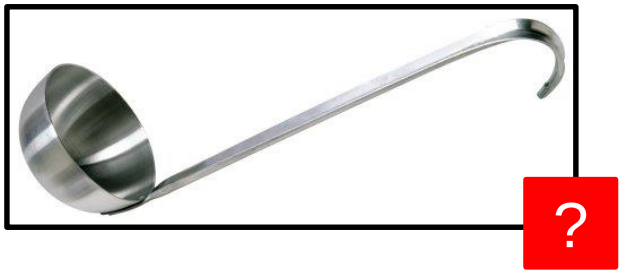


Scene recognition

Status quo: passive snapshot recognition

Object recognition





Scene recognition

Status quo: passive snapshot recognition

Video credit Yu-Chuan Su

The active recognition setting

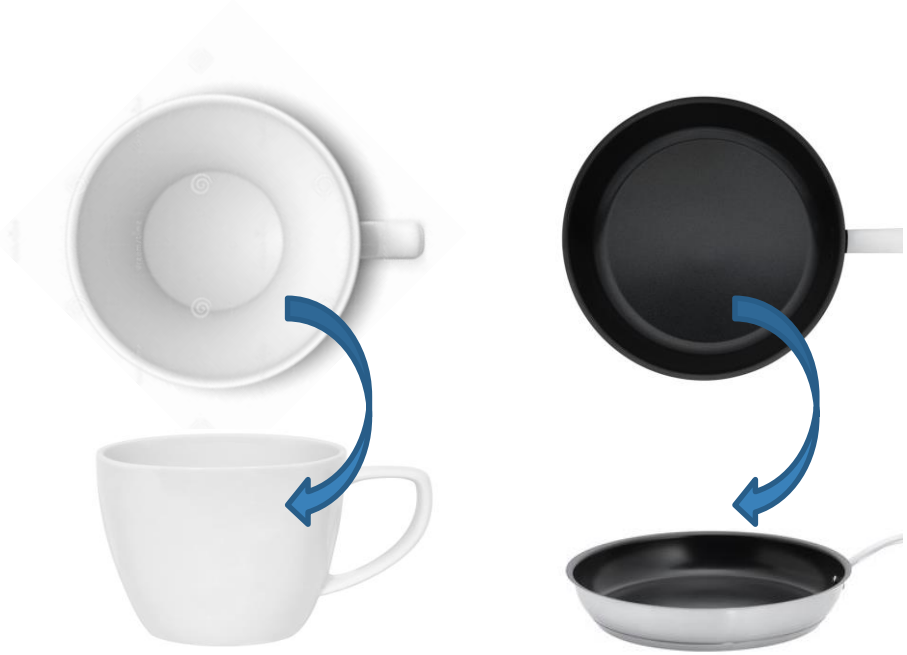
The active recognition setting

The active recognition setting

mug/bowl/pan?

mug/bowl/pan?

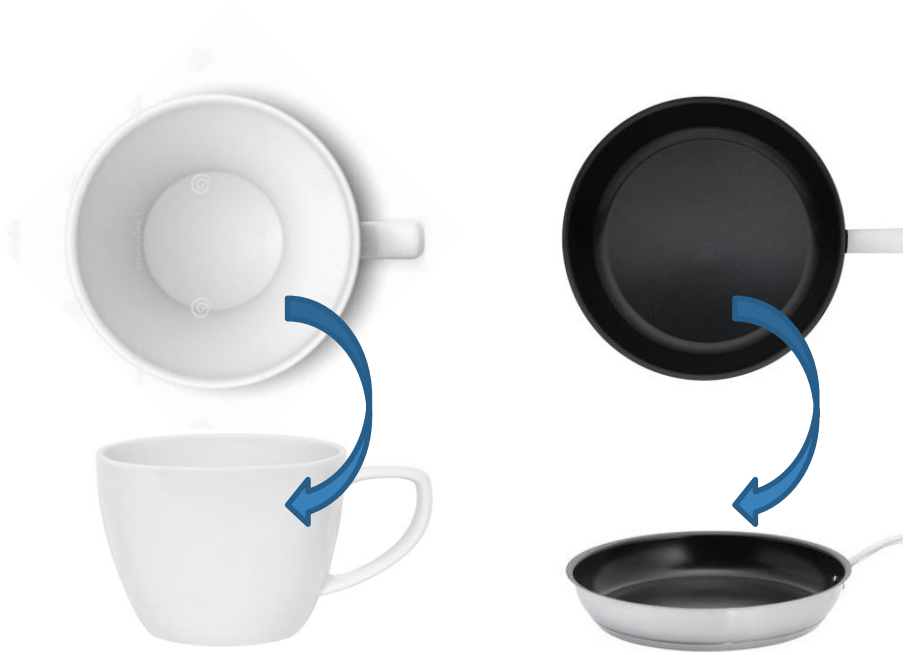
The active recognition setting



mug

frying pan

The active recognition setting



mug frying pan

"Active recognition":
The recognition system can ***select*** which views to see.

Active vs. passive recognition

Active vs. passive recognition

Difficulty: unconstrained visual input



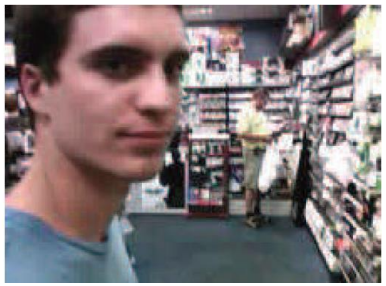



Image credit: Bo Xiong

Active vs. passive recognition

Difficulty: unconstrained visual input



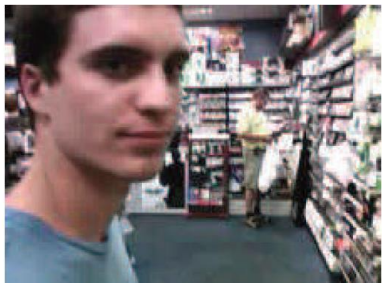
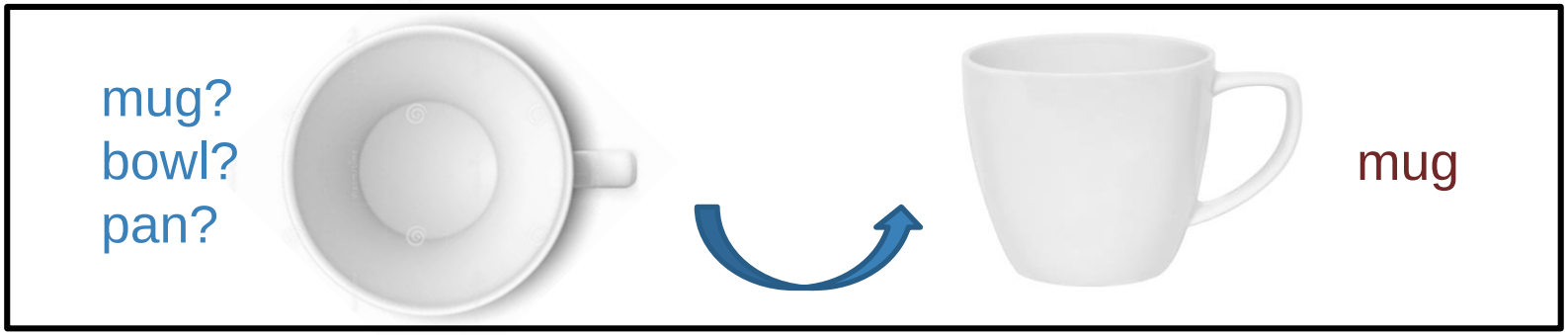



Image credit: Bo Xiong

Opportunity:

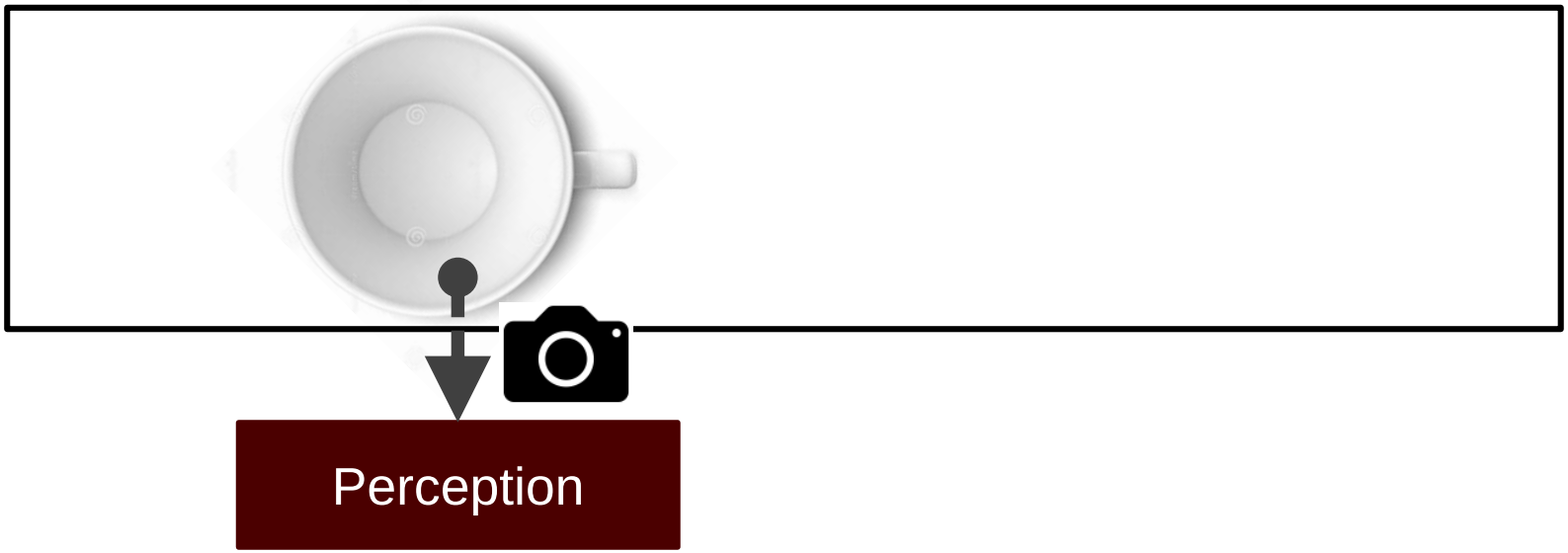


- Not restricted to a *single* snapshot.
- *Strategically acquiring* new views.

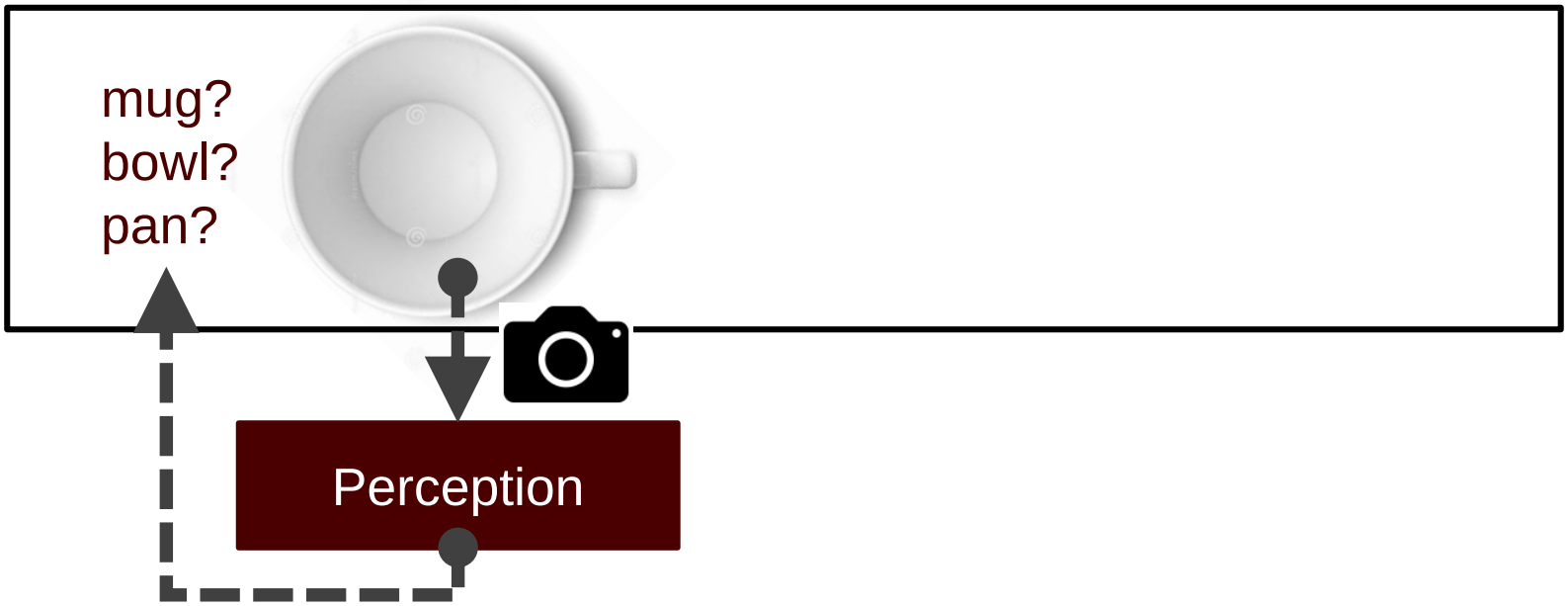
Components of active recognition

Components of active recognition

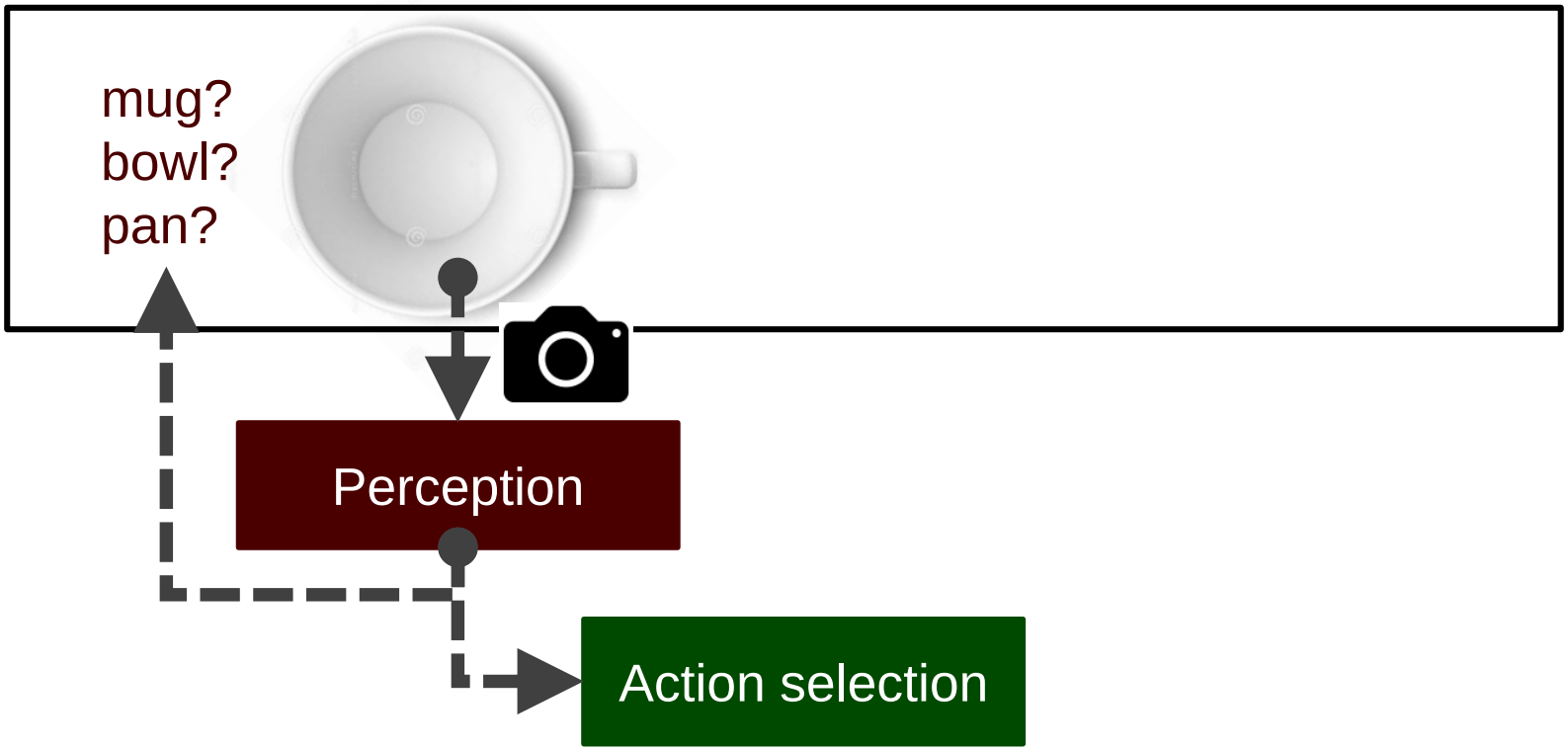
Components of active recognition



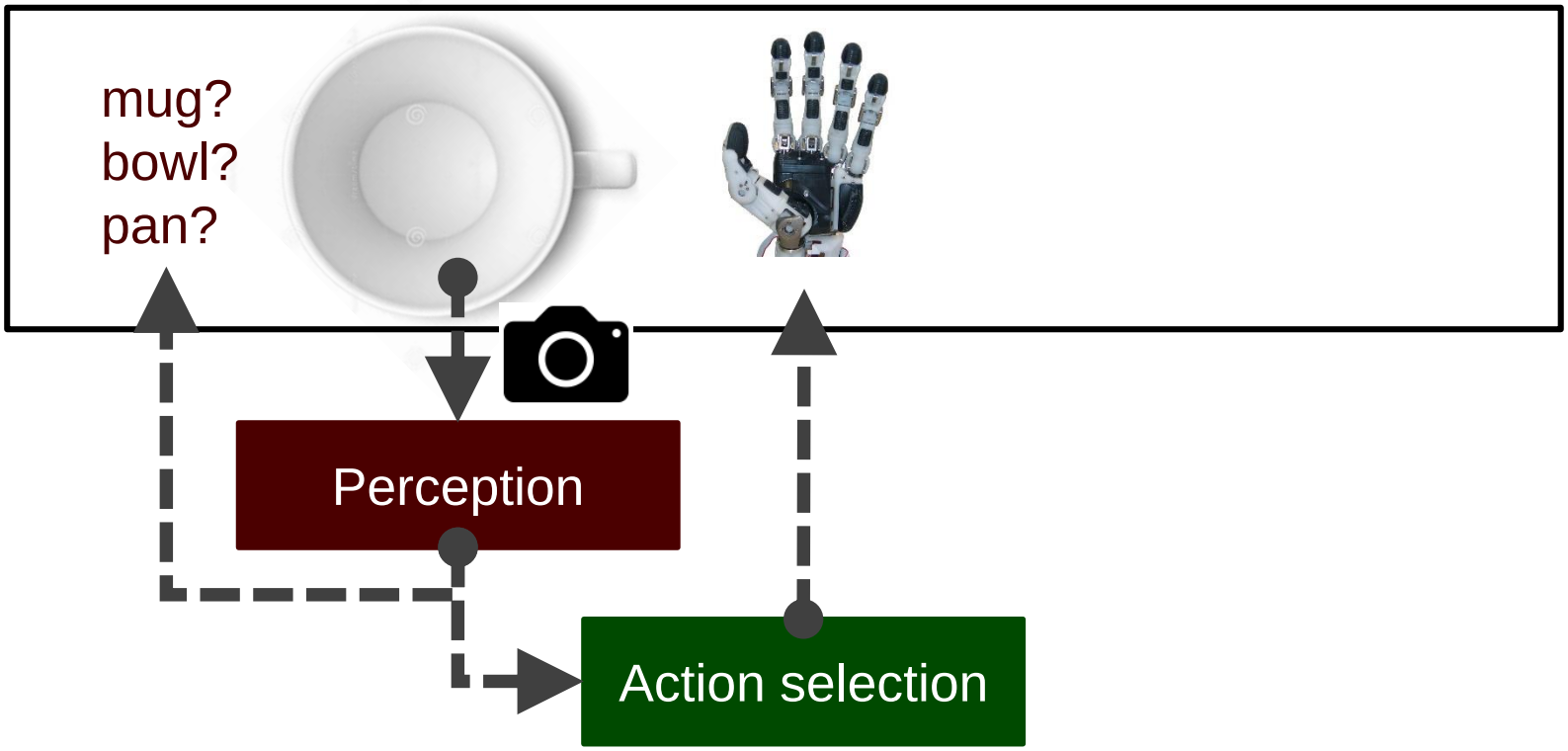
Components of active recognition



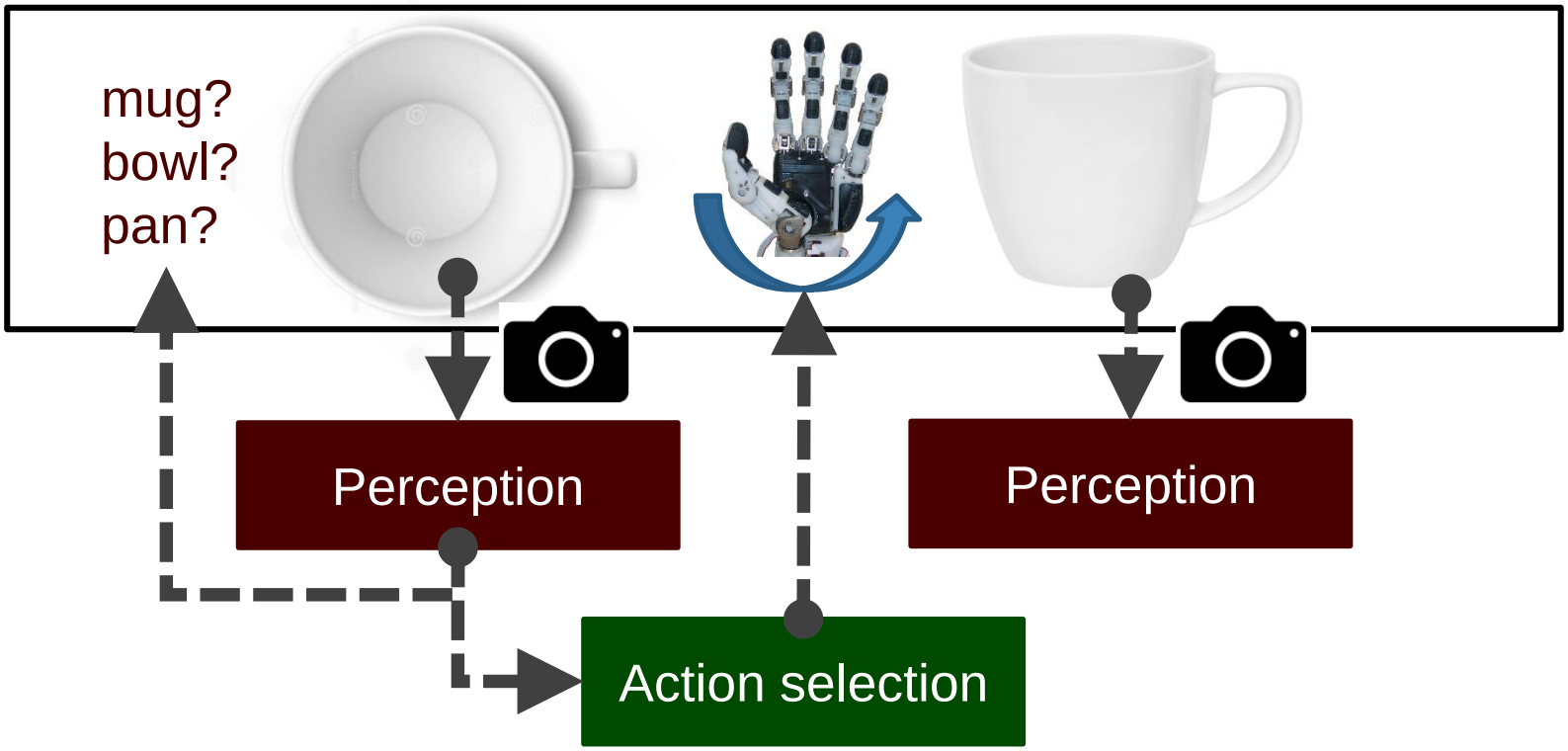
Components of active recognition



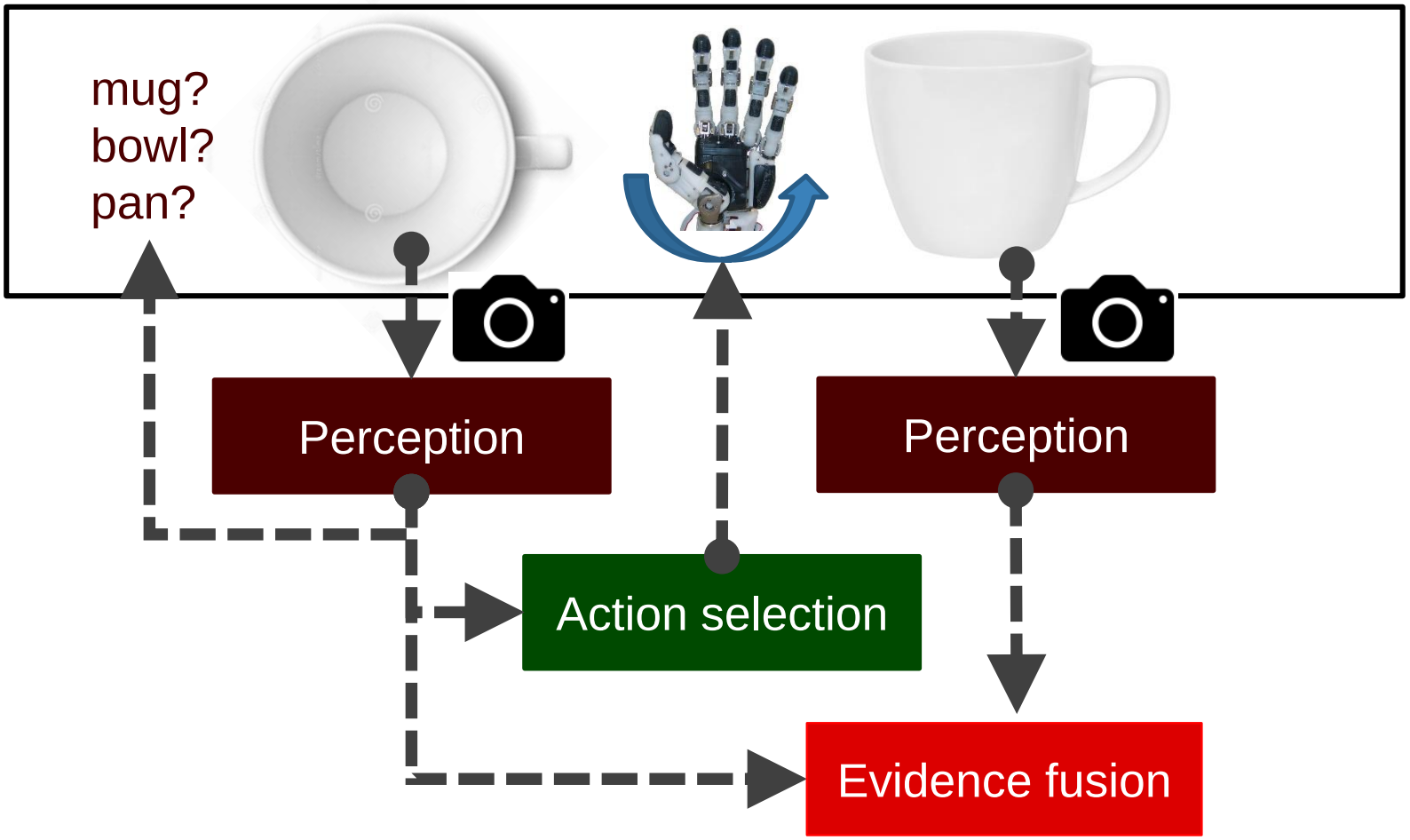
Components of active recognition



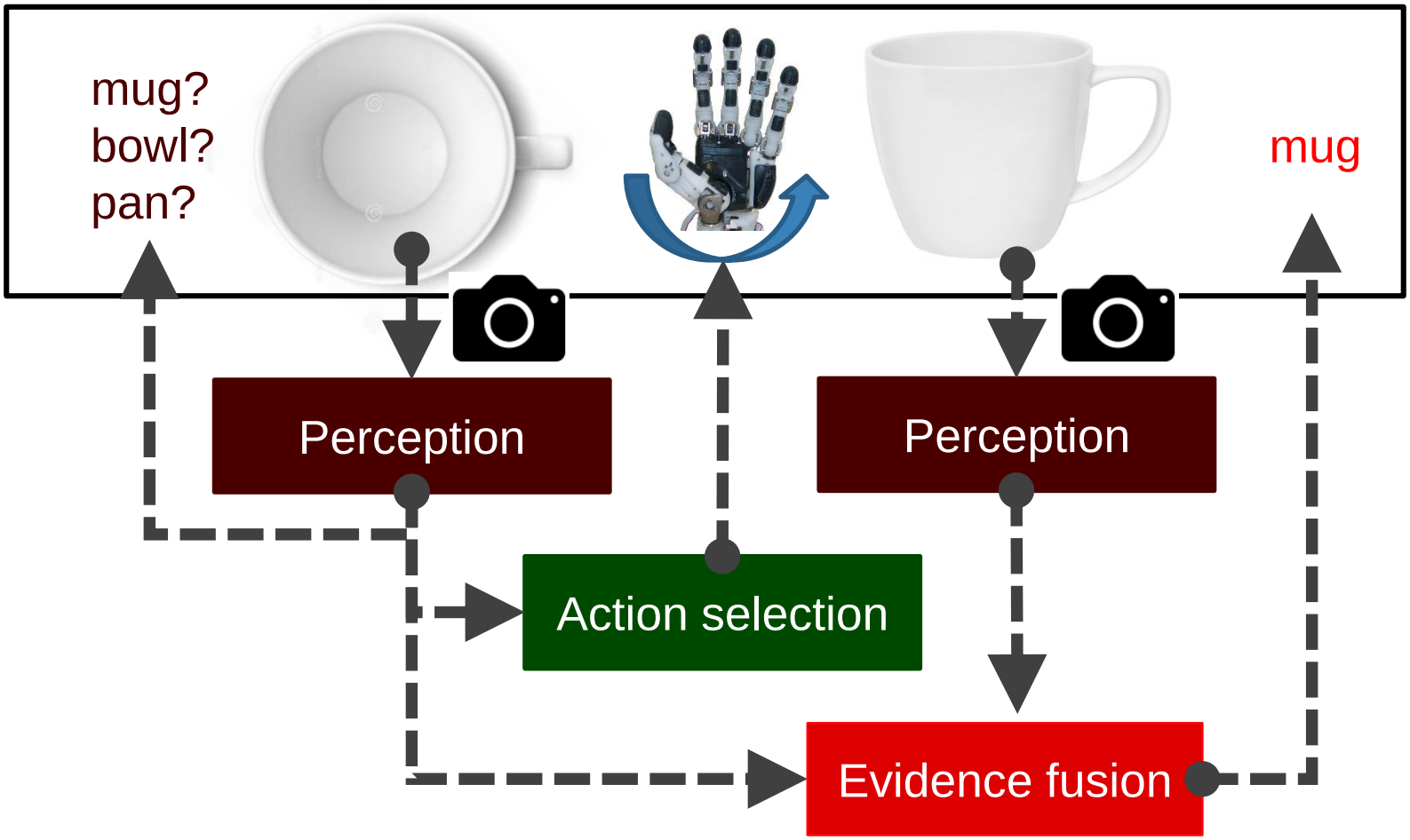
Components of active recognition



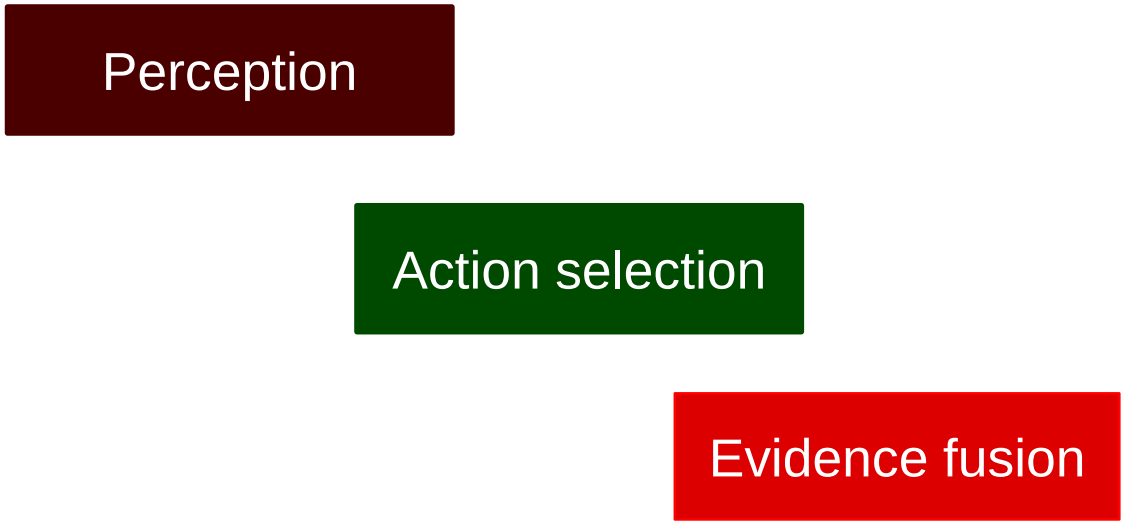
Components of active recognition



Components of active recognition



Components of active recognition



Prior active recognition approaches

Perception

- Train for 1-view recognition

  Bajcsy 1988
  Wilkes 1992
  Dickinson 1997
  Schiele 1998
  Denzler 2002
  Soatto 2009
  Ramanathan 2011
  Aloimonos 2011
  Borotschnig 2011
  Wu 2015
  Jayaraman 2015
  Johns 2016

Action selection

- Navigate to a pre-selected viewpoint

  Dickinson 1997
  Schiele 1998
  Denzler 2002

- Greedily maximize information gain

  Borotschnig 1998
  Ramanathan 2011
  Wu 2015
  Jayaraman 2015

- Reinforcement learning

  Paletta 2000,
  Malmir 2015

Evidence fusion

- Verification

  Dickinson 1997
  Schiele 1998

- Averaging

  Johns 2016

- Bayes/Naïve Bayes

  Paletta 2000
  Denzler 2002
  Ramanathan 2011
  Malmir 2015

Prior active recognition approaches

Perception

- Train for 1-view recognition

Bajcsy 1988
Wilkes 1992
Dickinson 1997
Schiele 1998
Denzler 2002
Soatto 2009
Ramanathan 2011
Aloimonos 2011
Borotschnig 2011

Action selection

- Navigate to a pre-selected viewpoint

Dickinson 1997
Schiele 1998
Denzler 2002

- Greedily maximize information gain

Borotschnig 1998
Ramanathan 2011
Wu 2015
Paletta 2000,
Malmir 2015

Evidence fusion

- Verification

Dickinson 1997
Schiele 1998

- Averaging

Johns 2016

- Bayes/Naïve Bayes

Paletta 2000
Denzler 2002
Ramanathan 2011

Weakness: Independent, often heuristic solutions for the three active recognition components.

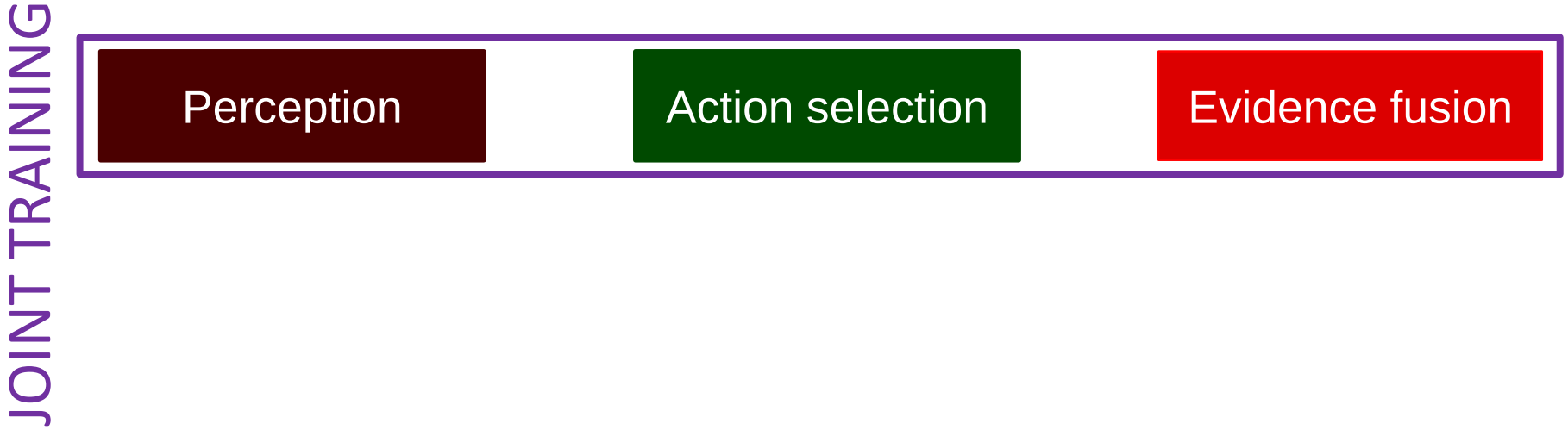
Our idea

Perception

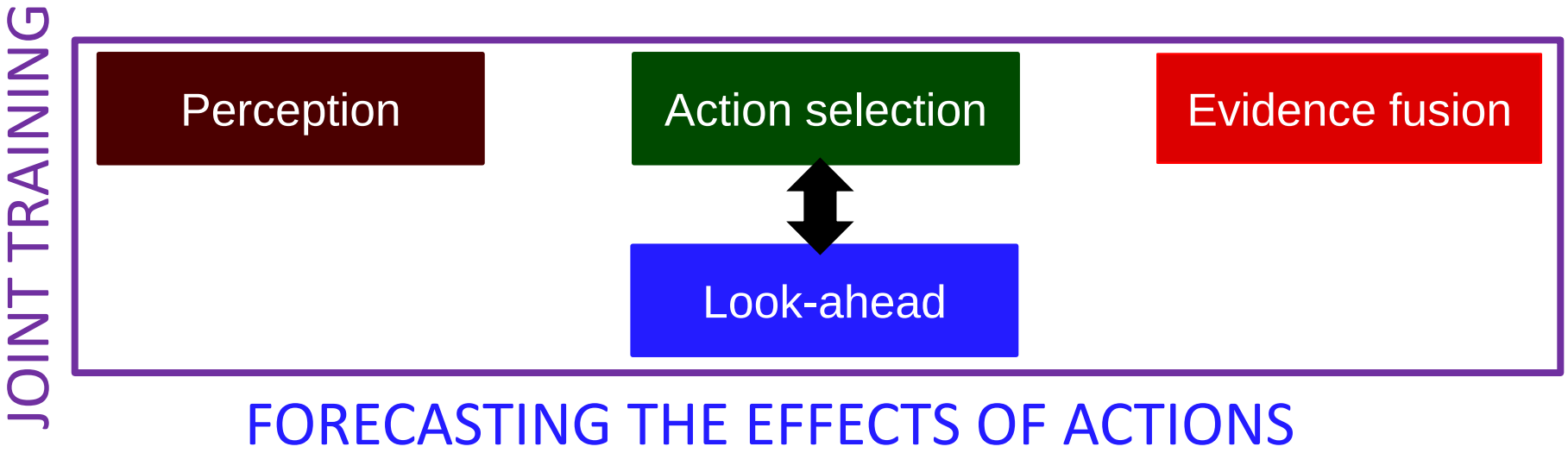
Action selection

Evidence fusion

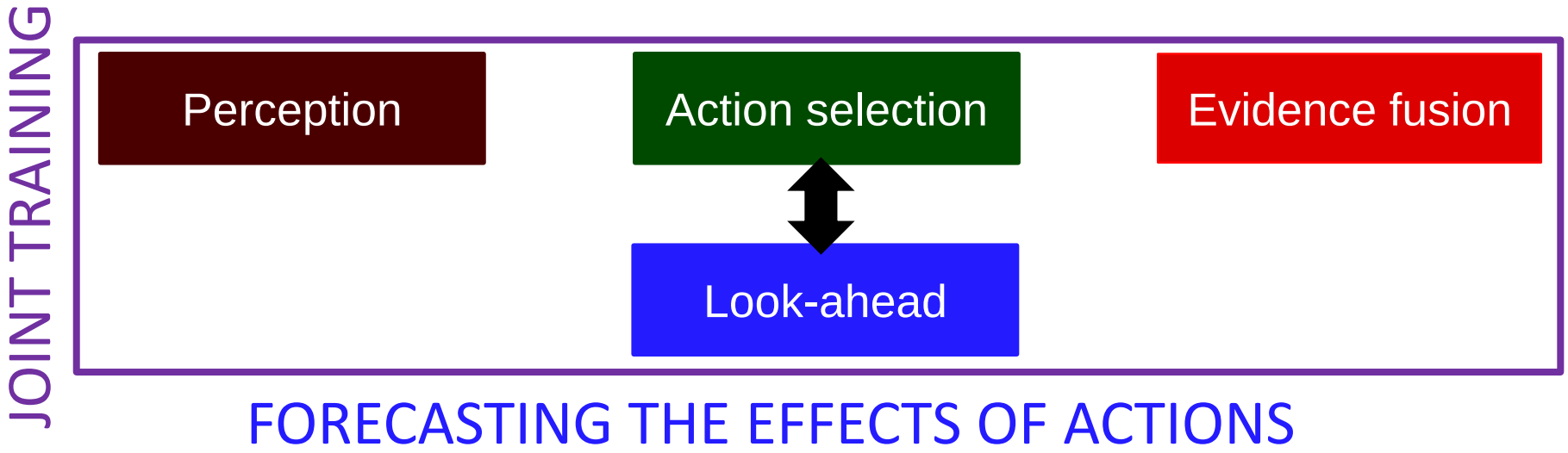
Our idea


JOINT TRAINING
Perception
Action selection
Evidence fusion

Our idea


JOINT TRAINING
Perception
Action selection
Evidence fusion
Look-ahead
FORECASTING THE EFFECTS OF ACTIONS

Our idea



Multi-task training of active recognition components + look-ahead.

Towards real-world active recognition

Towards real-world active recognition

Instance recognition from turn-table scans

Toy category recognition with custom robot

[Nene 1996, Schiele 1998, Denzler 2003, Ramanathan 2011…]

Towards real-world active recognition

Directing a camera for scene category recognition

Manipulation for object recognition

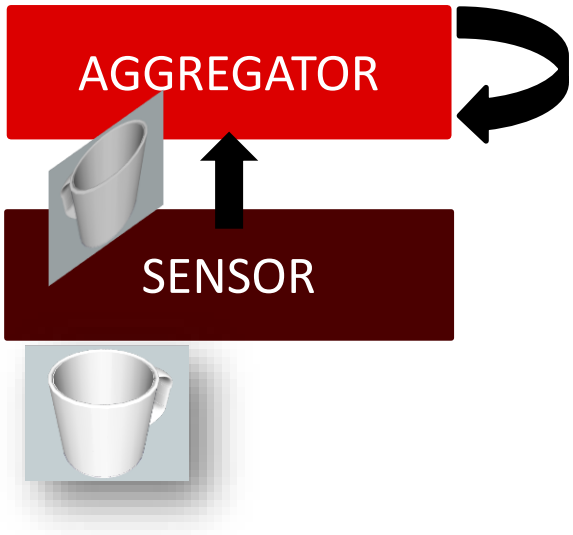
[Malmir et al, 2015]

High-level architecture

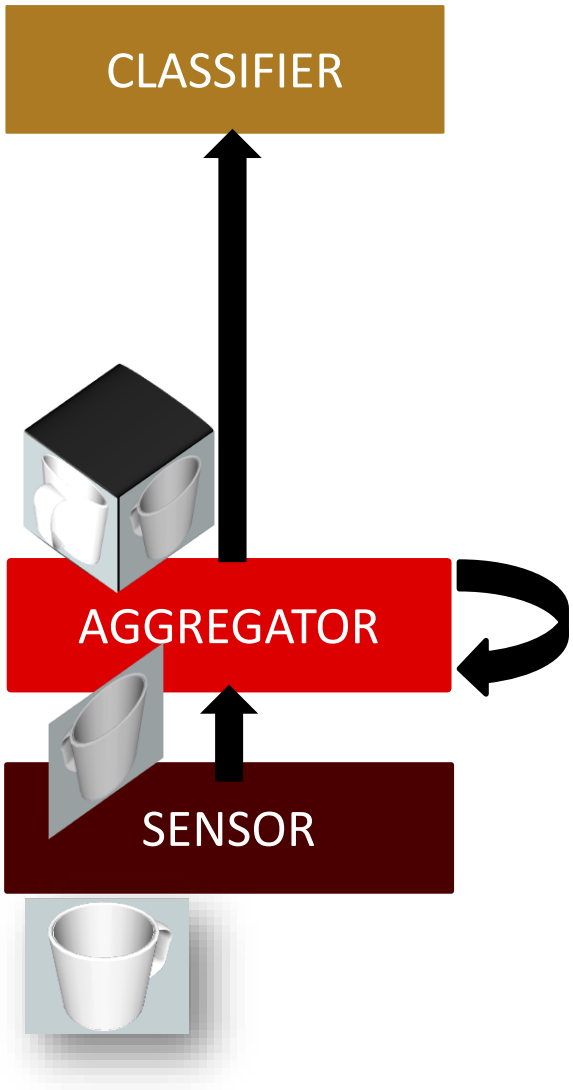
High-level architecture

SENSOR

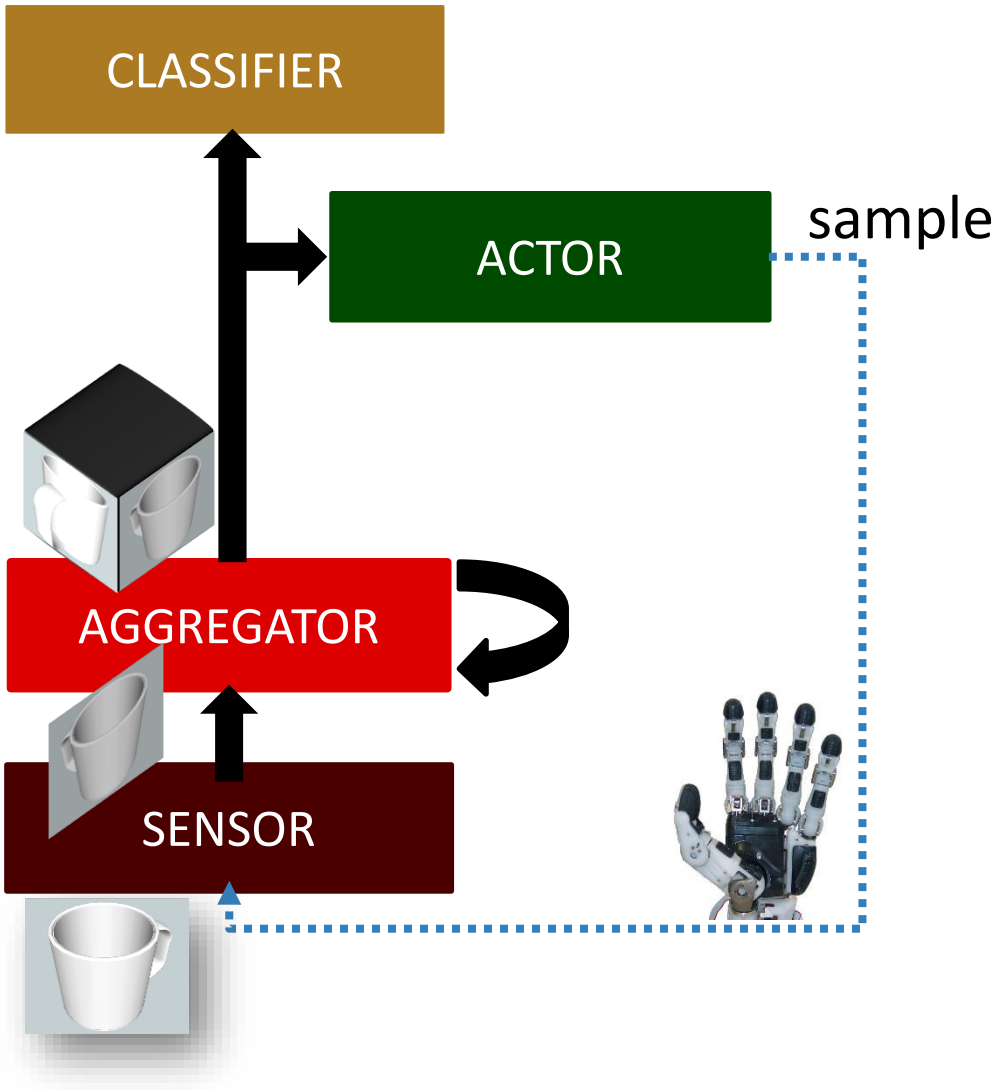
High-level architecture



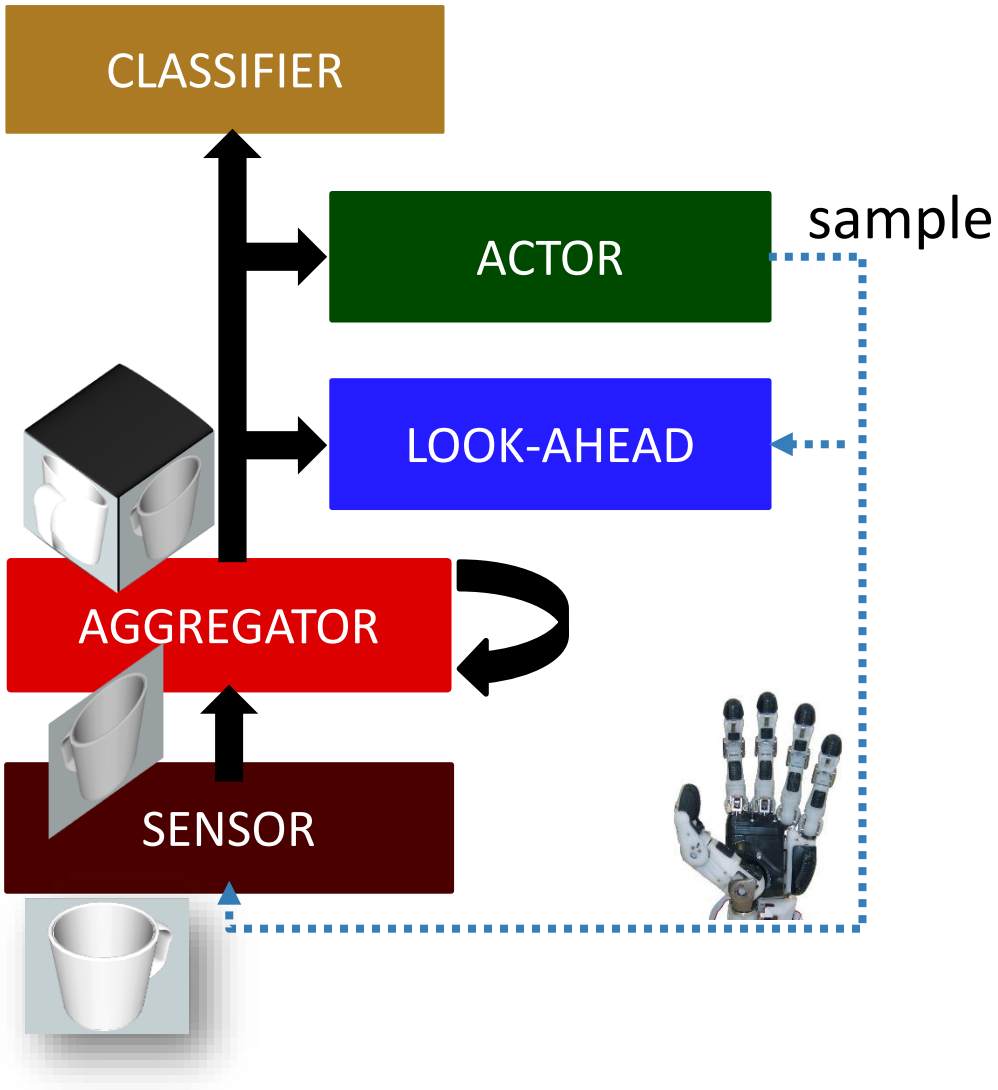
High-level architecture



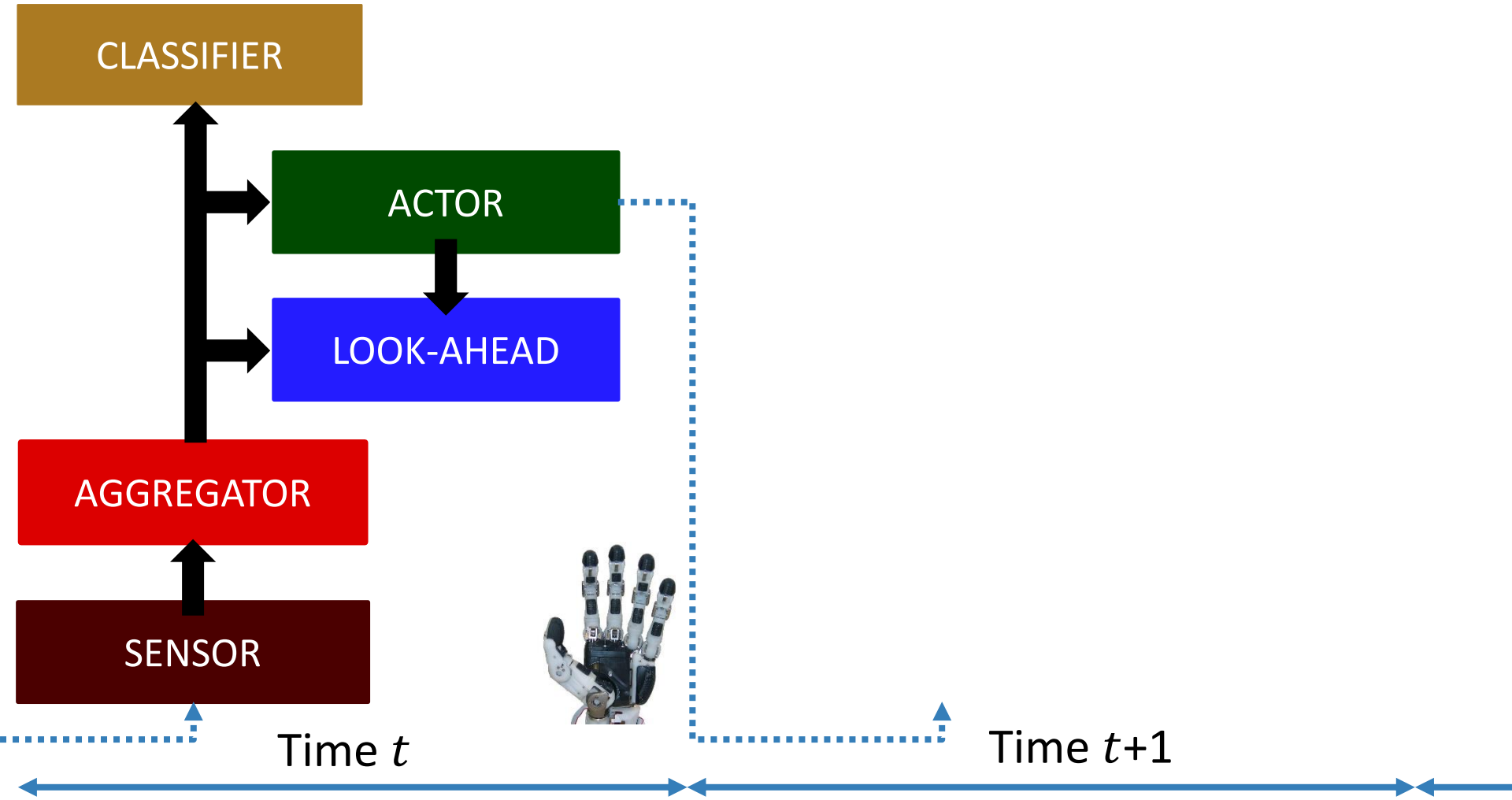
High-level architecture



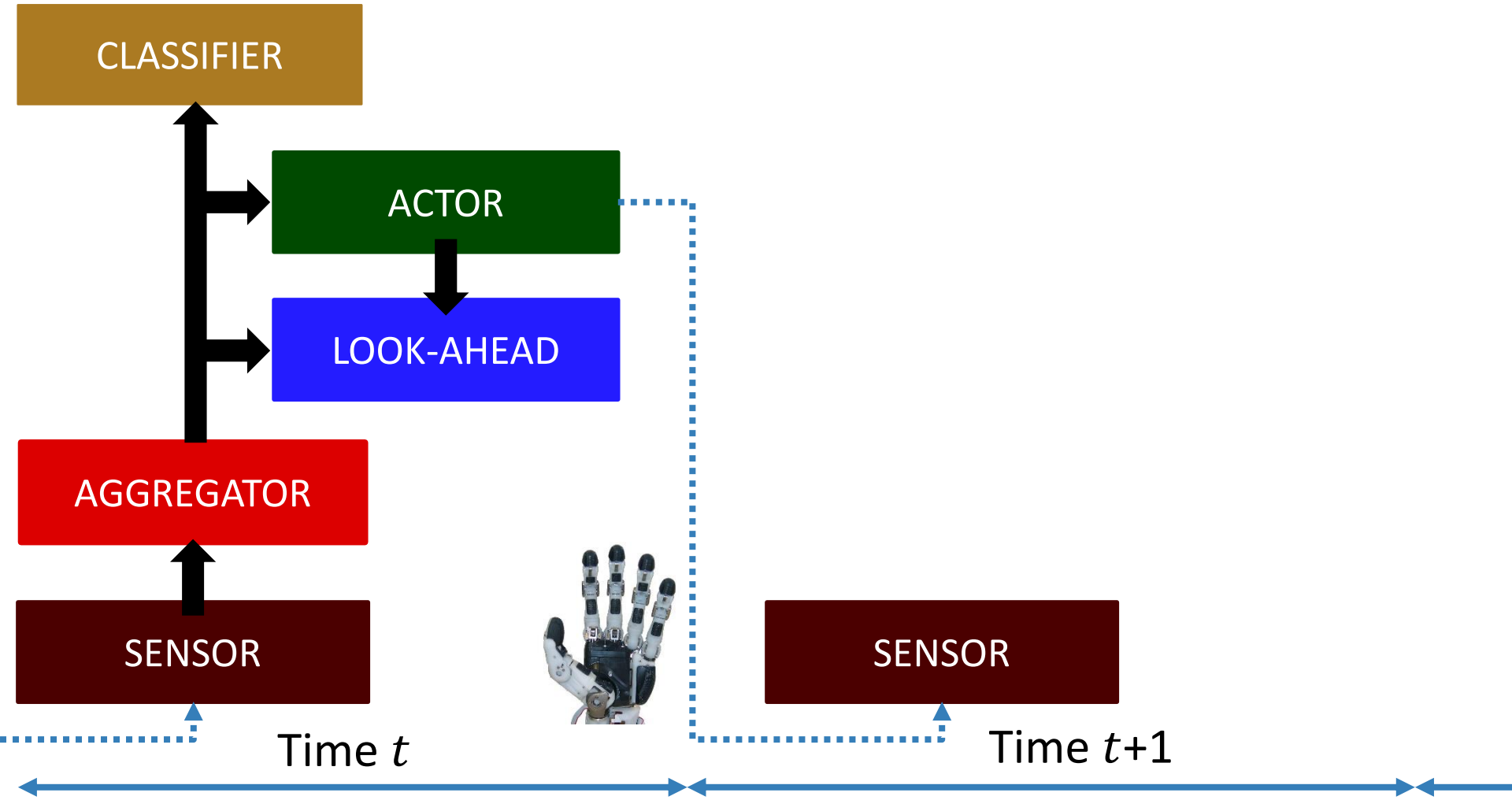
High-level architecture



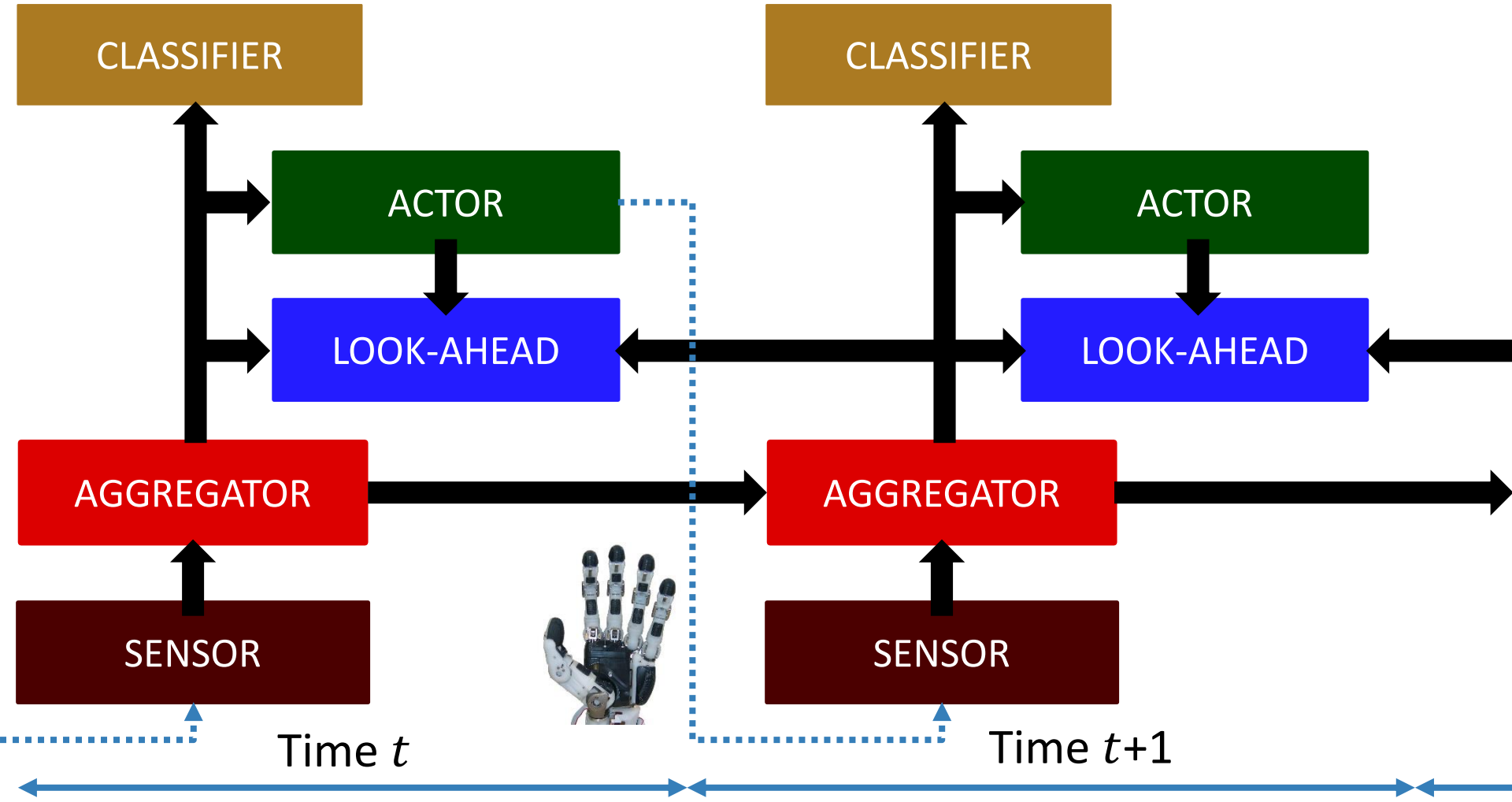
Unrolled architecture



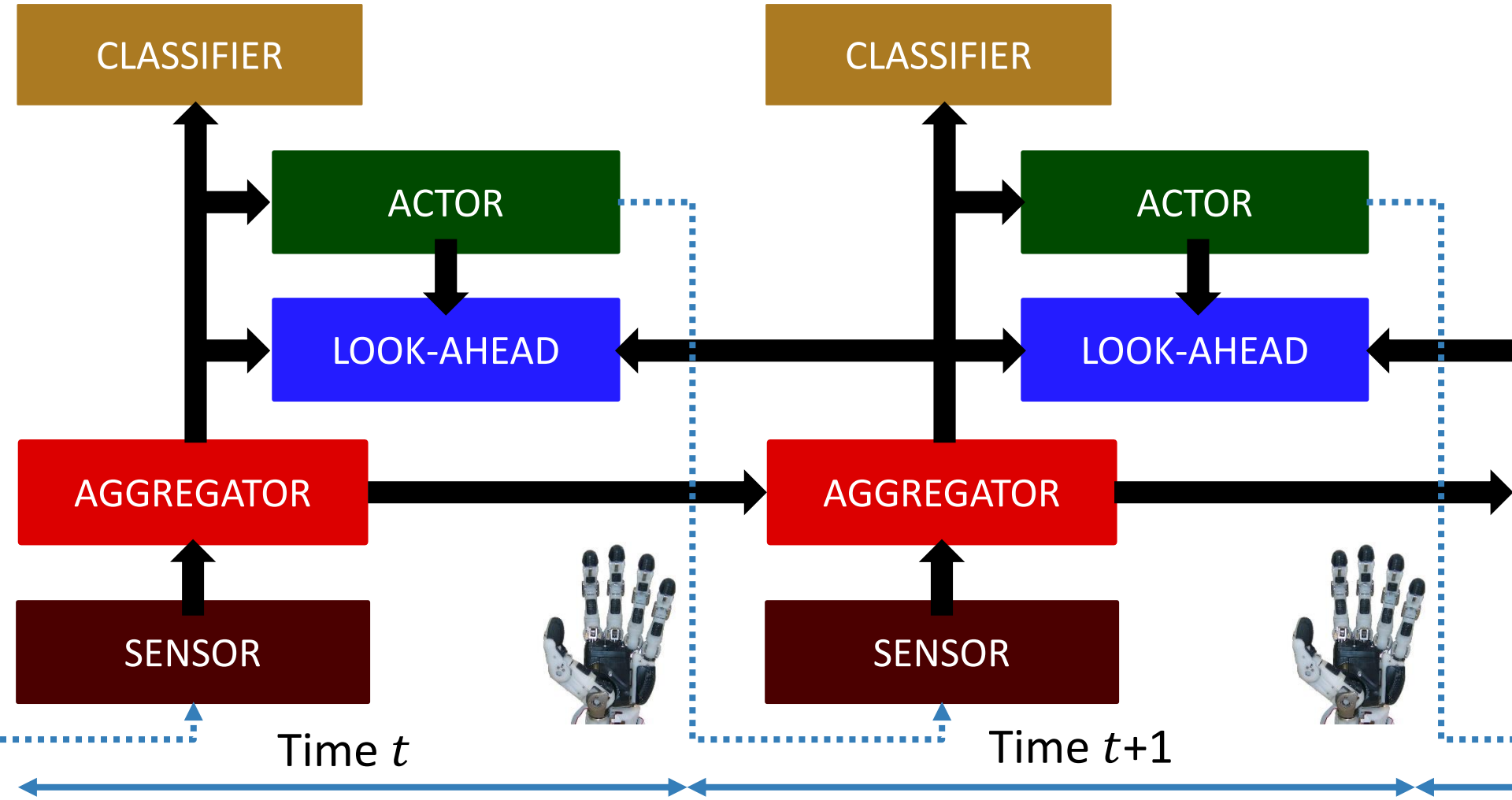
Unrolled architecture



Unrolled architecture



Unrolled architecture



The modules

CLASSIFIER

ACTOR

LOOK-AHEAD

AGGREGATOR

SENSOR

Perception module - SENSOR

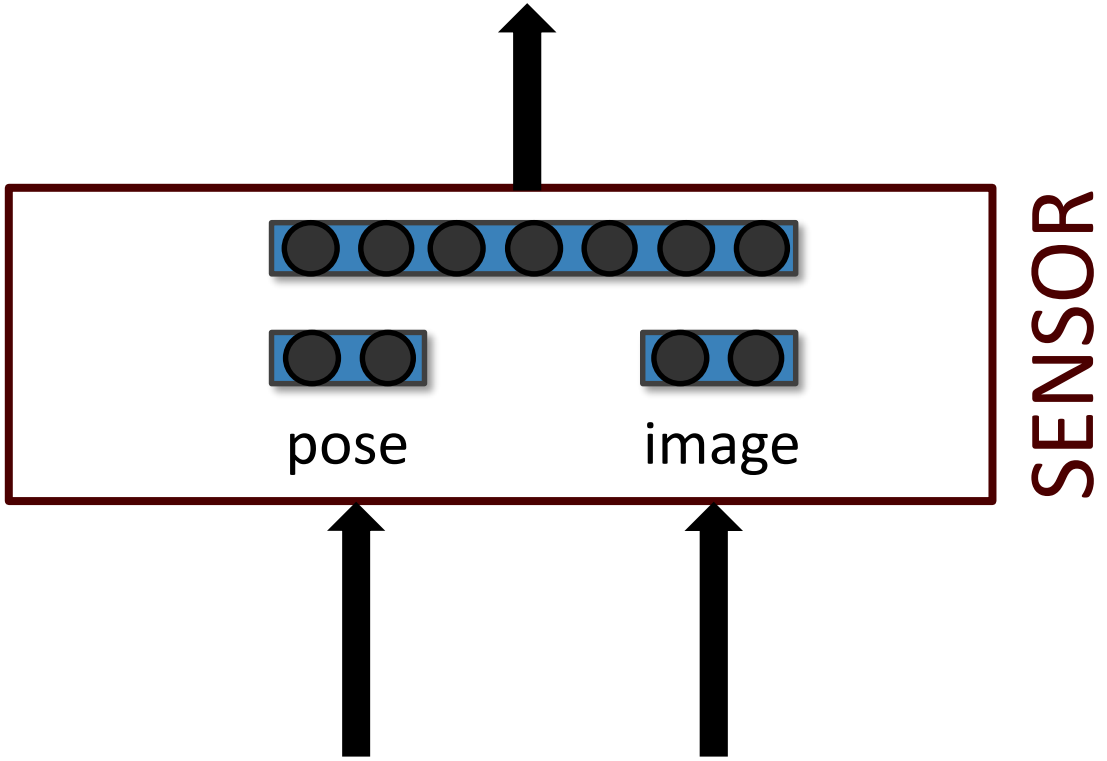
CLASSIFIER

ACTOR

LOOK-AHEAD

AGGREGATOR

SENSOR



Perception module - SENSOR

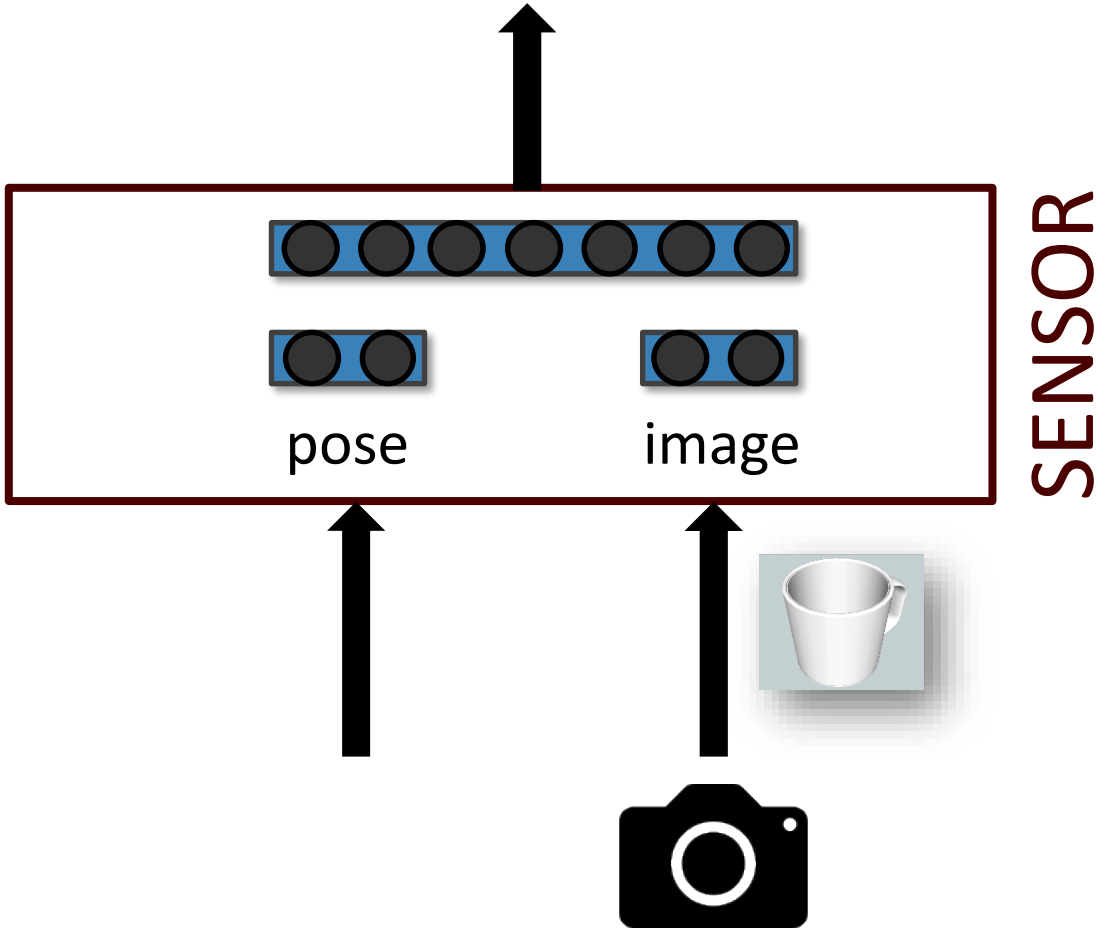
CLASSIFIER

ACTOR

LOOK-AHEAD

AGGREGATOR

SENSOR



Perception module - SENSOR

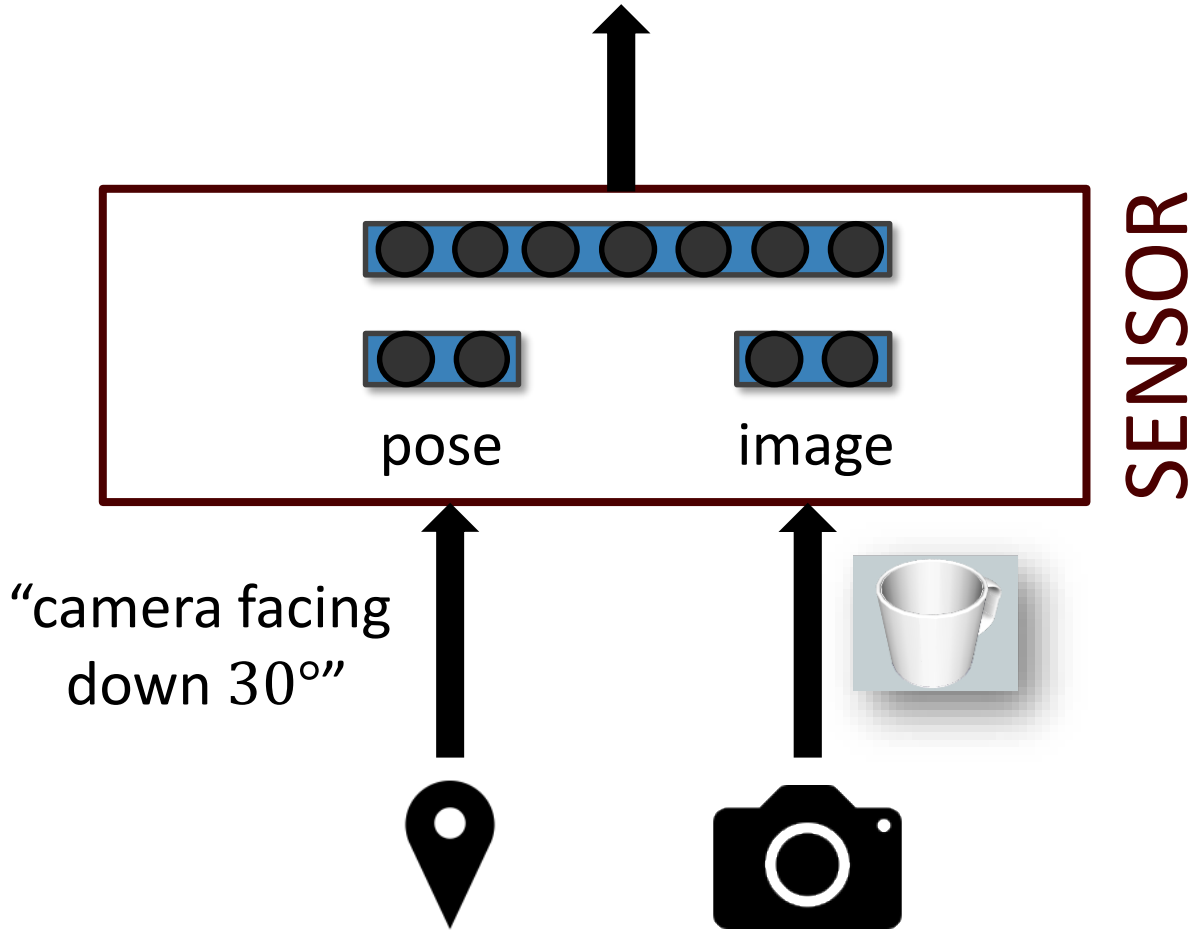
CLASSIFIER

ACTOR

LOOK-AHEAD

AGGREGATOR

SENSOR



Perception module - SENSOR

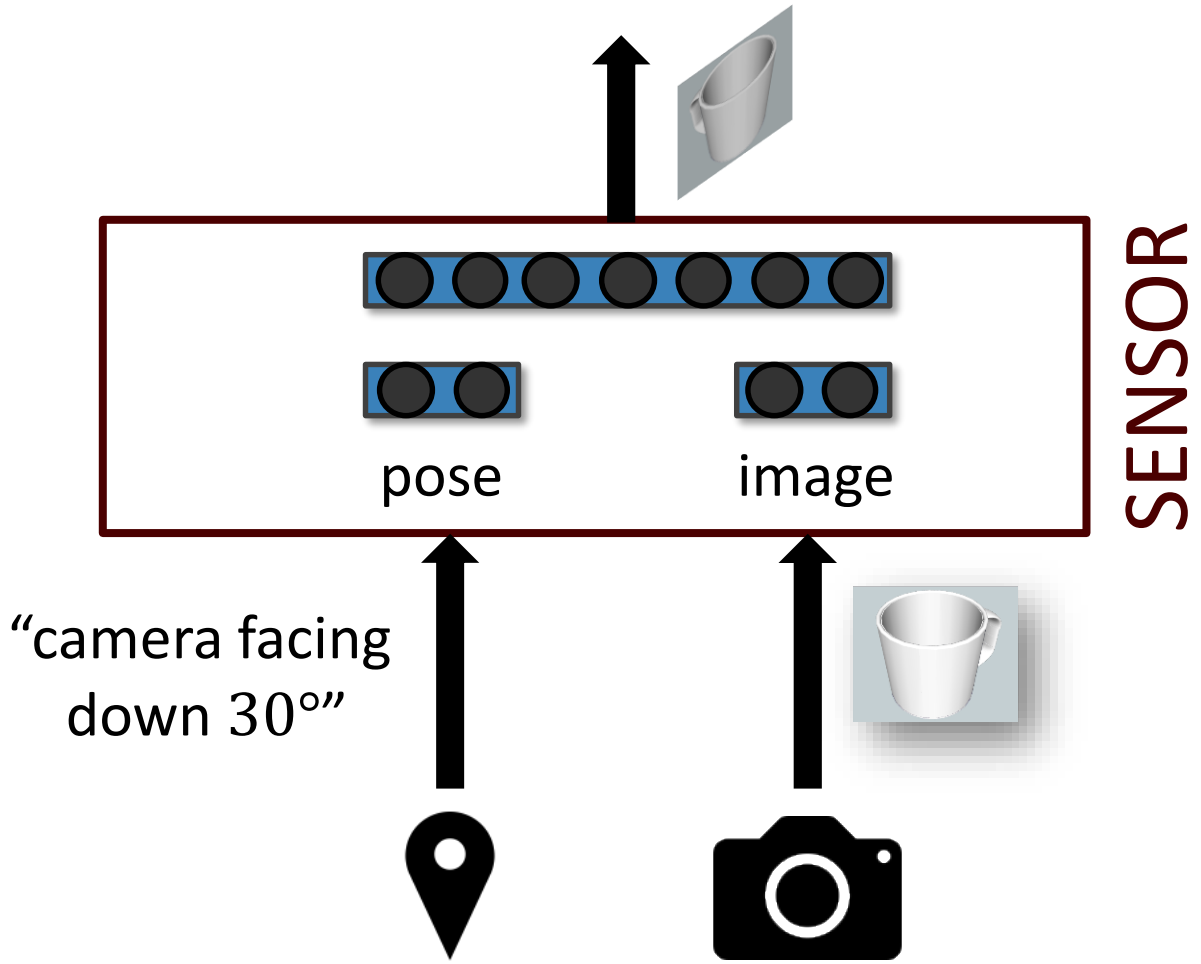
CLASSIFIER

ACTOR

LOOK-AHEAD

AGGREGATOR

SENSOR



Evidence fusion - AGGREGATOR

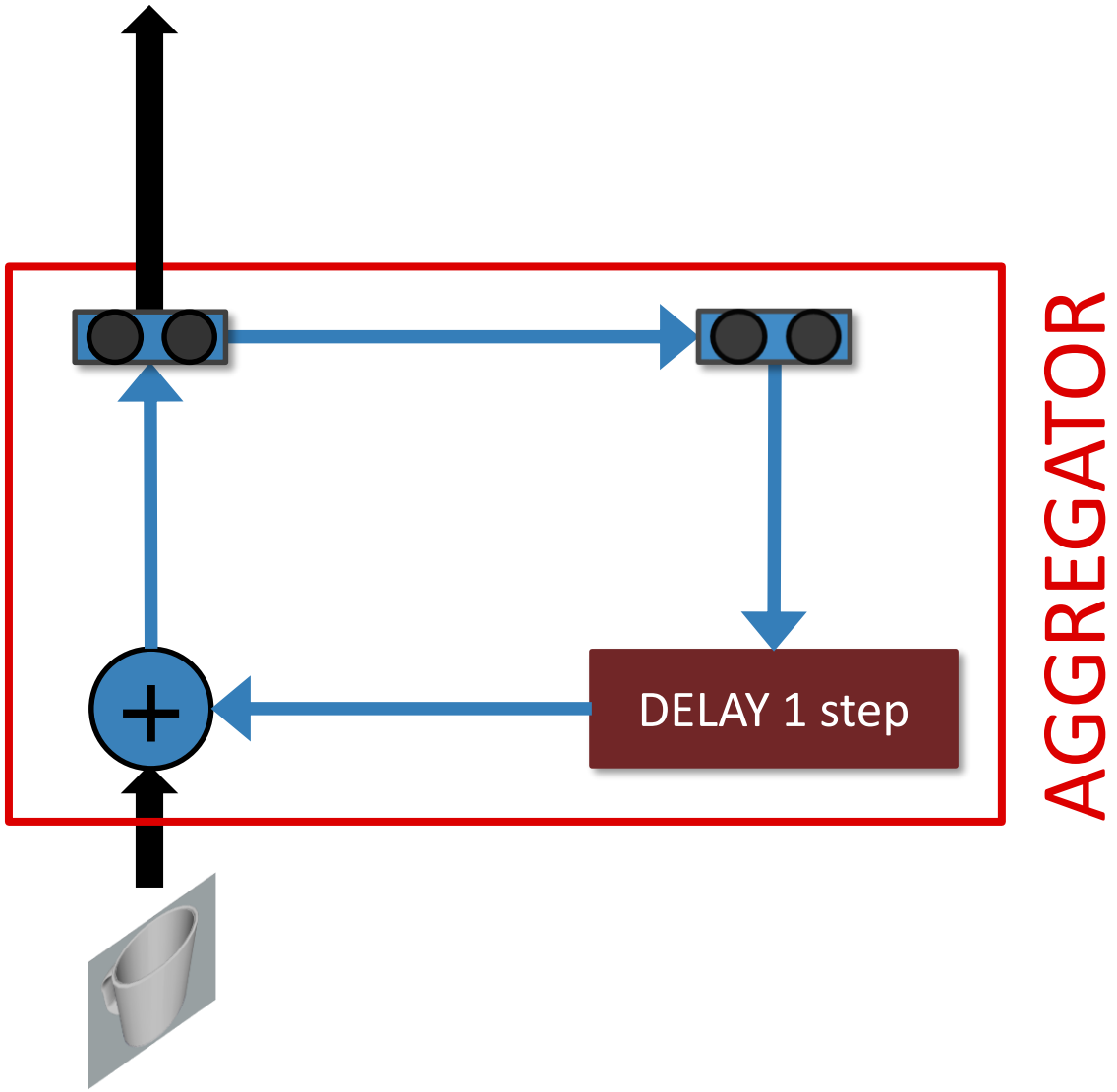
CLASSIFIER

ACTOR

LOOK-AHEAD

AGGREGATOR

SENSOR



Evidence fusion - AGGREGATOR

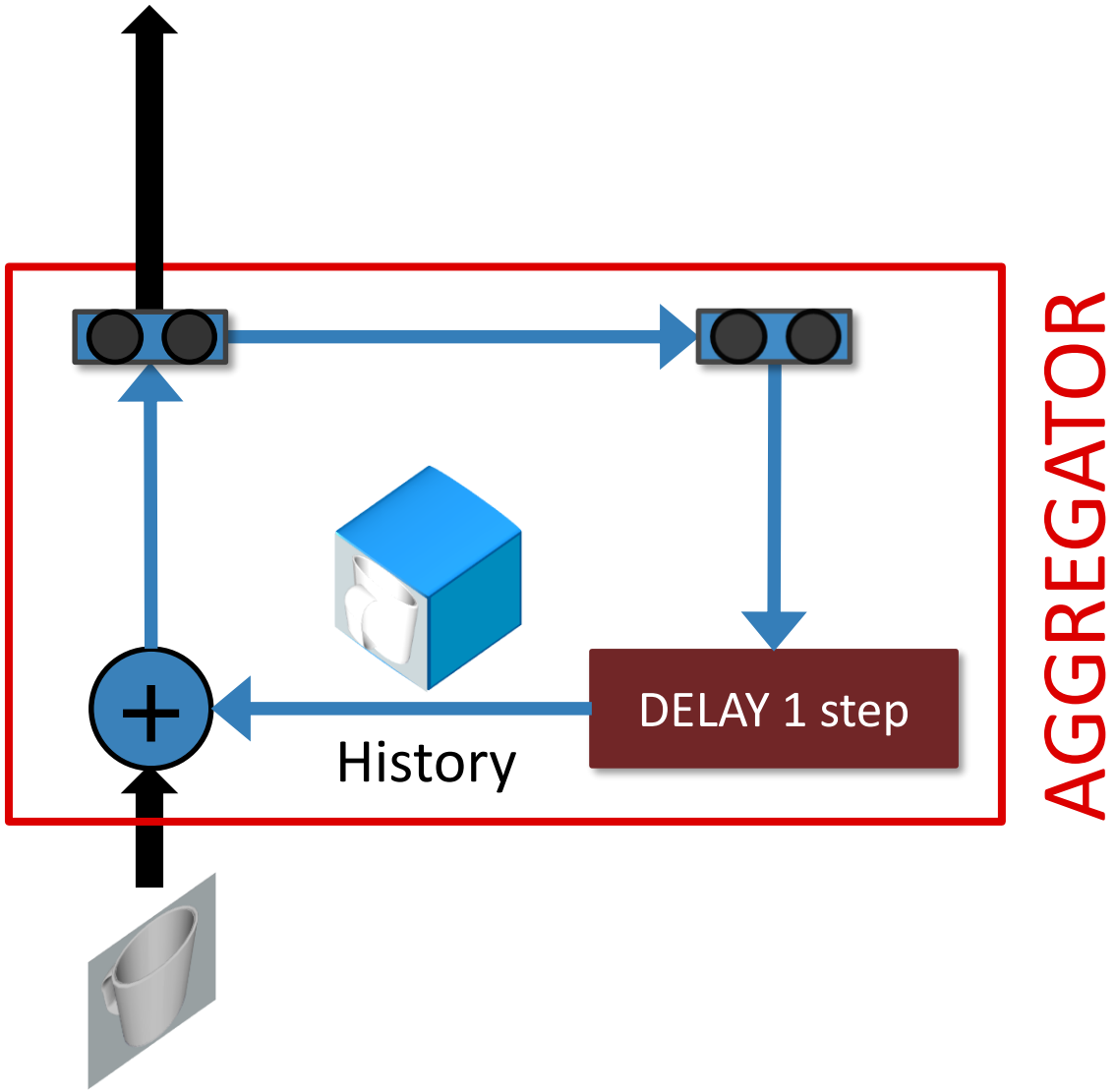
CLASSIFIER

ACTOR

LOOK-AHEAD

AGGREGATOR

SENSOR



Evidence fusion - AGGREGATOR

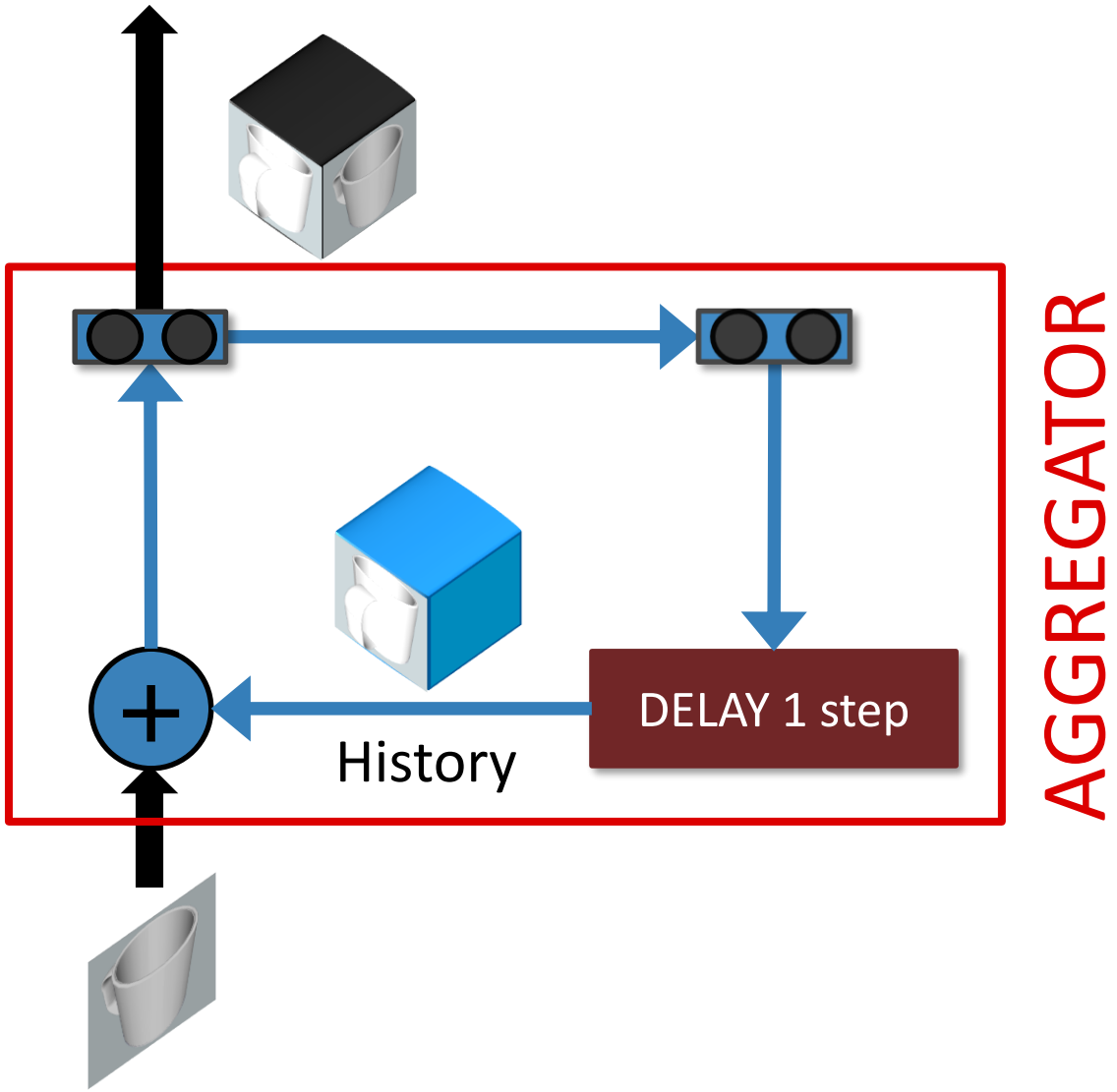
CLASSIFIER

ACTOR

LOOK-AHEAD

AGGREGATOR

SENSOR



Evidence fusion - AGGREGATOR

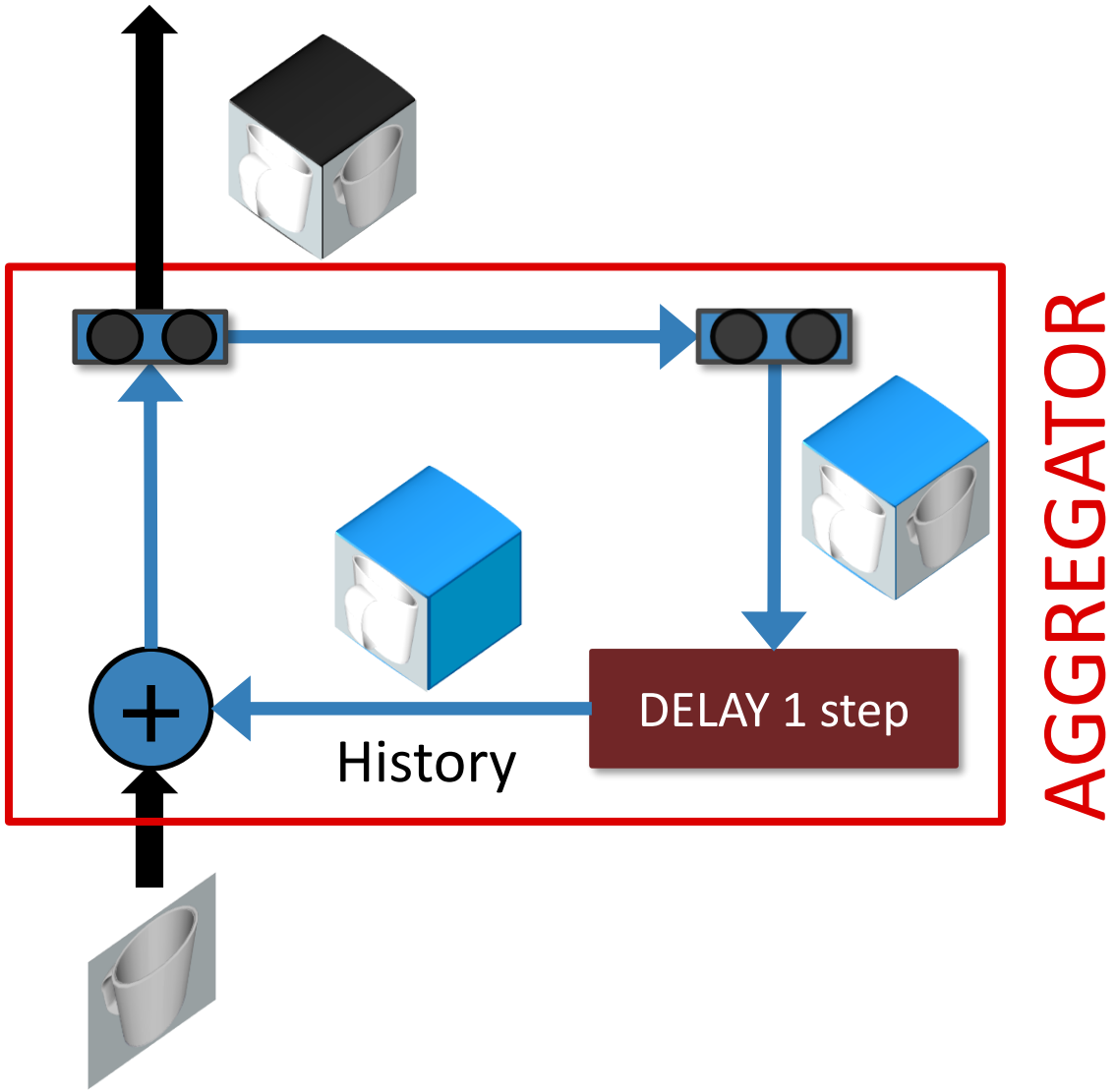
CLASSIFIER

ACTOR

LOOK-AHEAD

AGGREGATOR

SENSOR



Future prediction – LOOKAHEAD

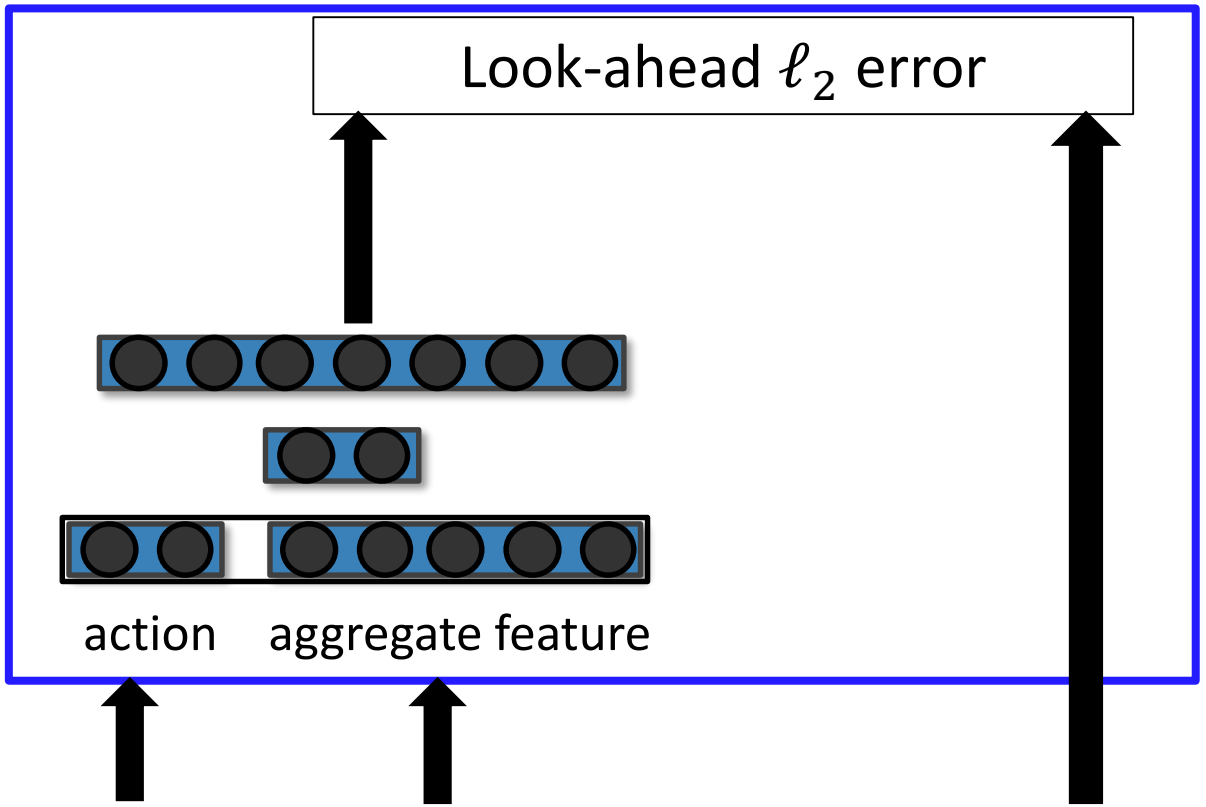
CLASSIFIER

ACTOR

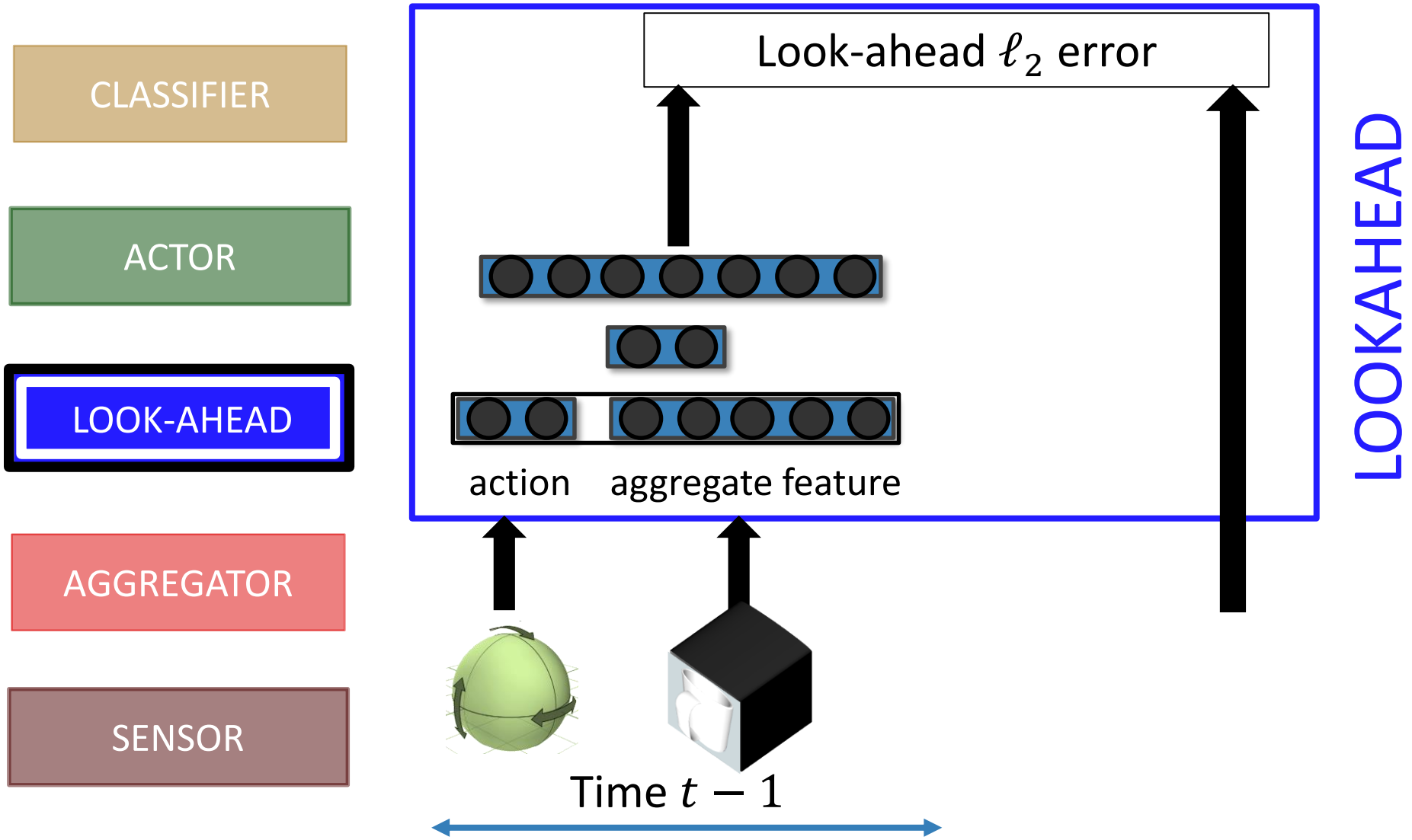
LOOK-AHEAD

AGGREGATOR

SENSOR



Future prediction – LOOKAHEAD



Future prediction – LOOKAHEAD

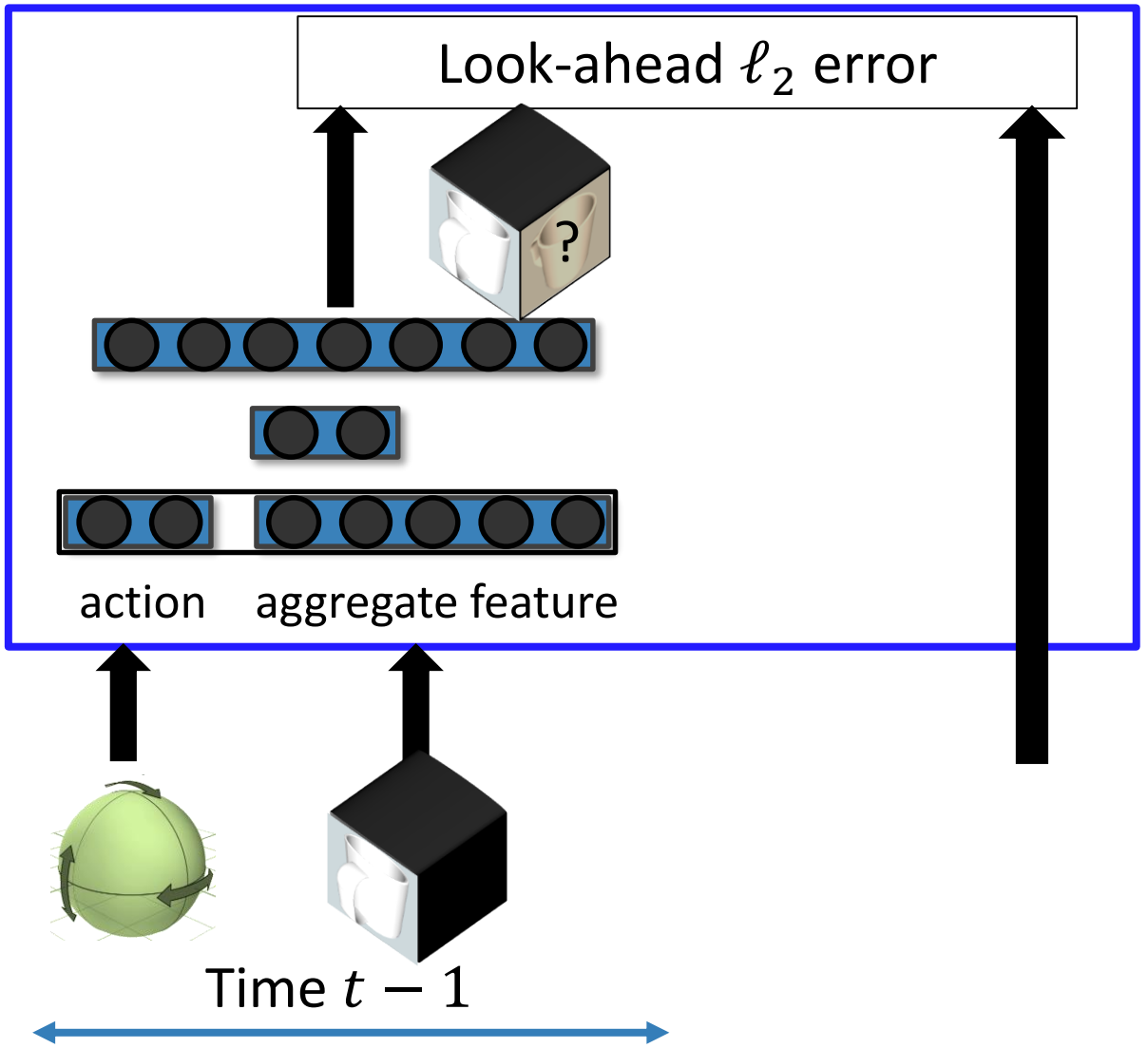
CLASSIFIER

ACTOR

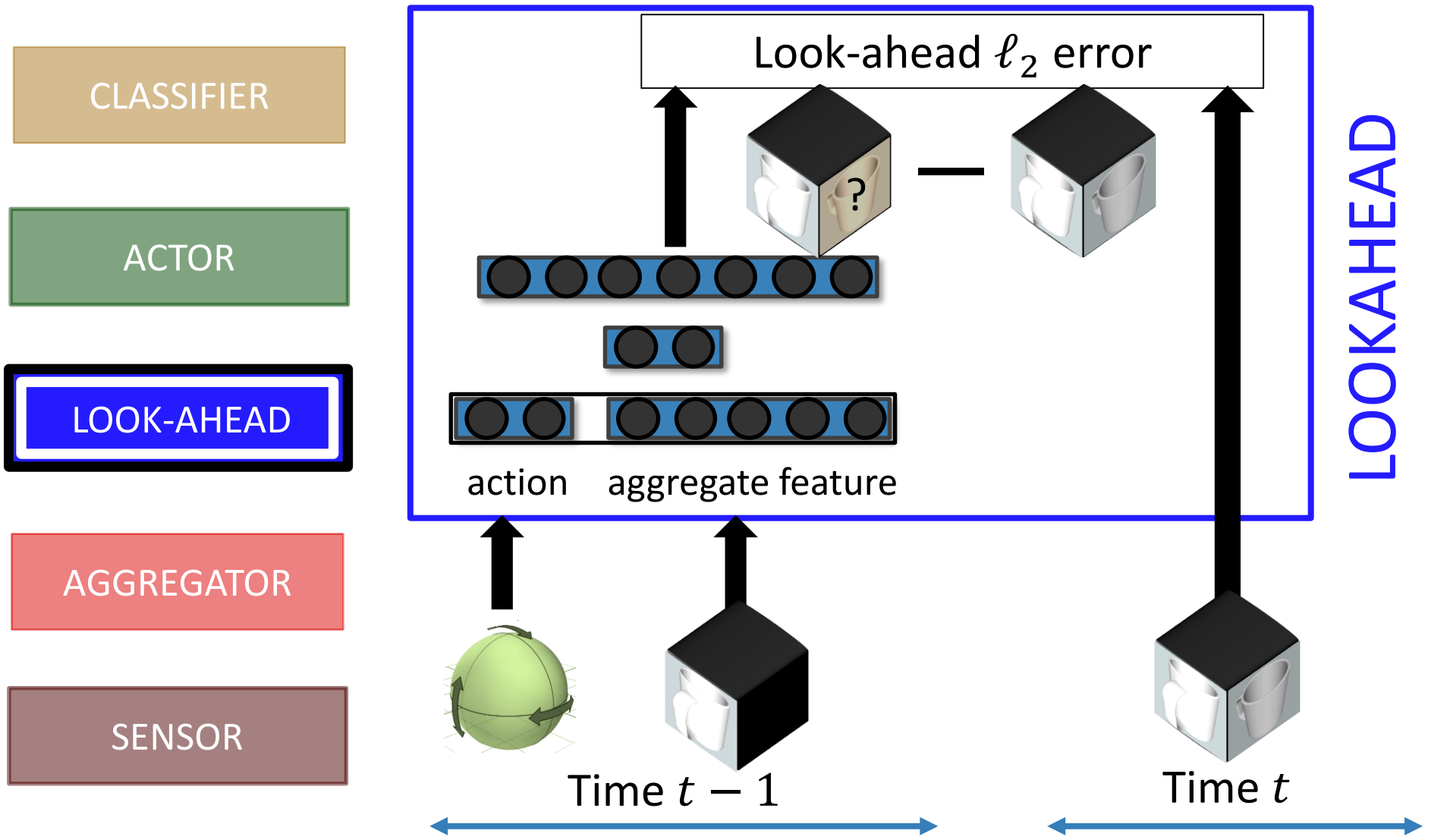
LOOK-AHEAD

AGGREGATOR

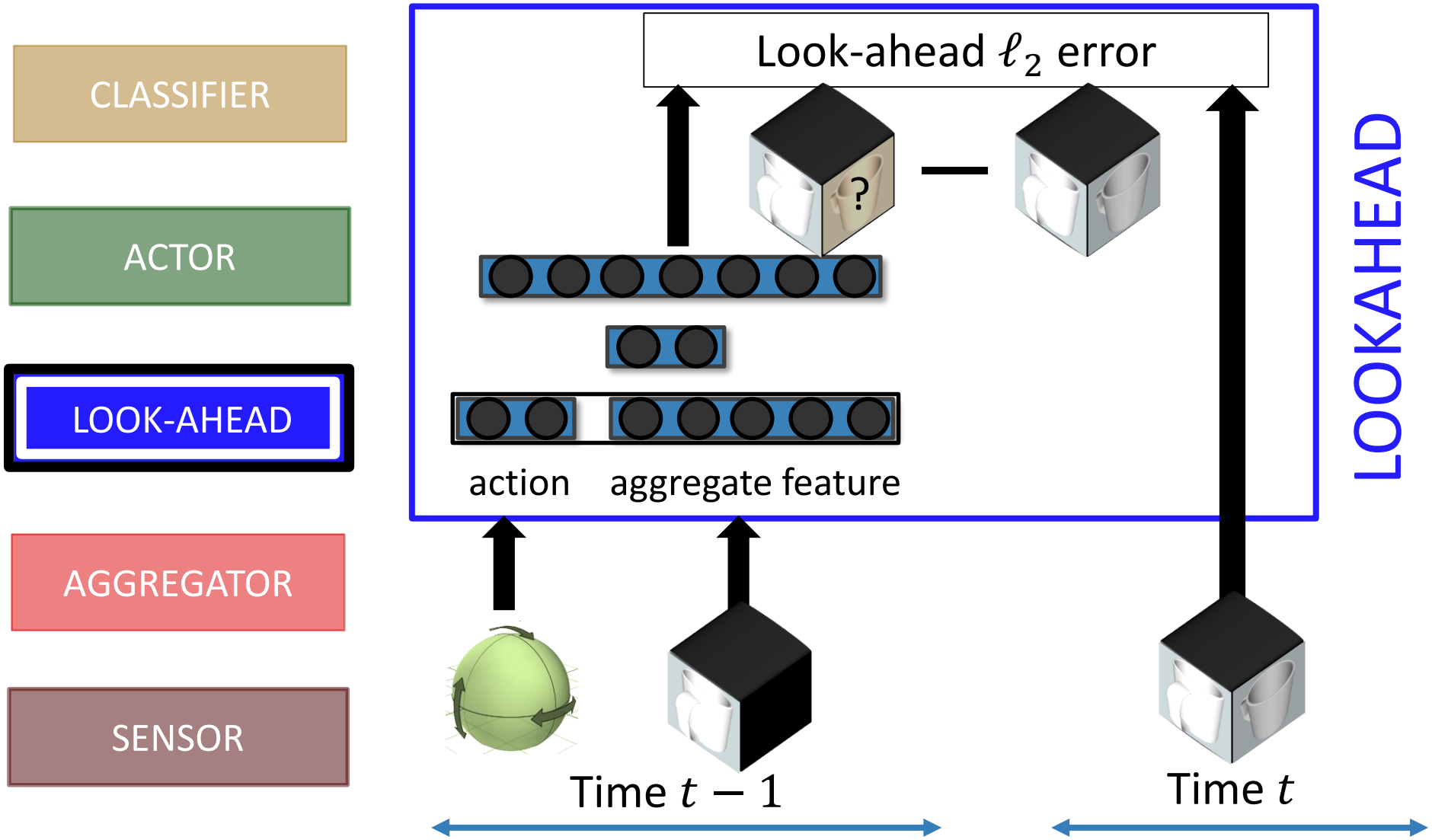
SENSOR



Future prediction – LOOKAHEAD



Future prediction – LOOKAHEAD



Jayaraman et al, Learning image representations tied to egomotion, ICCV 2015

Action selection - ACTOR

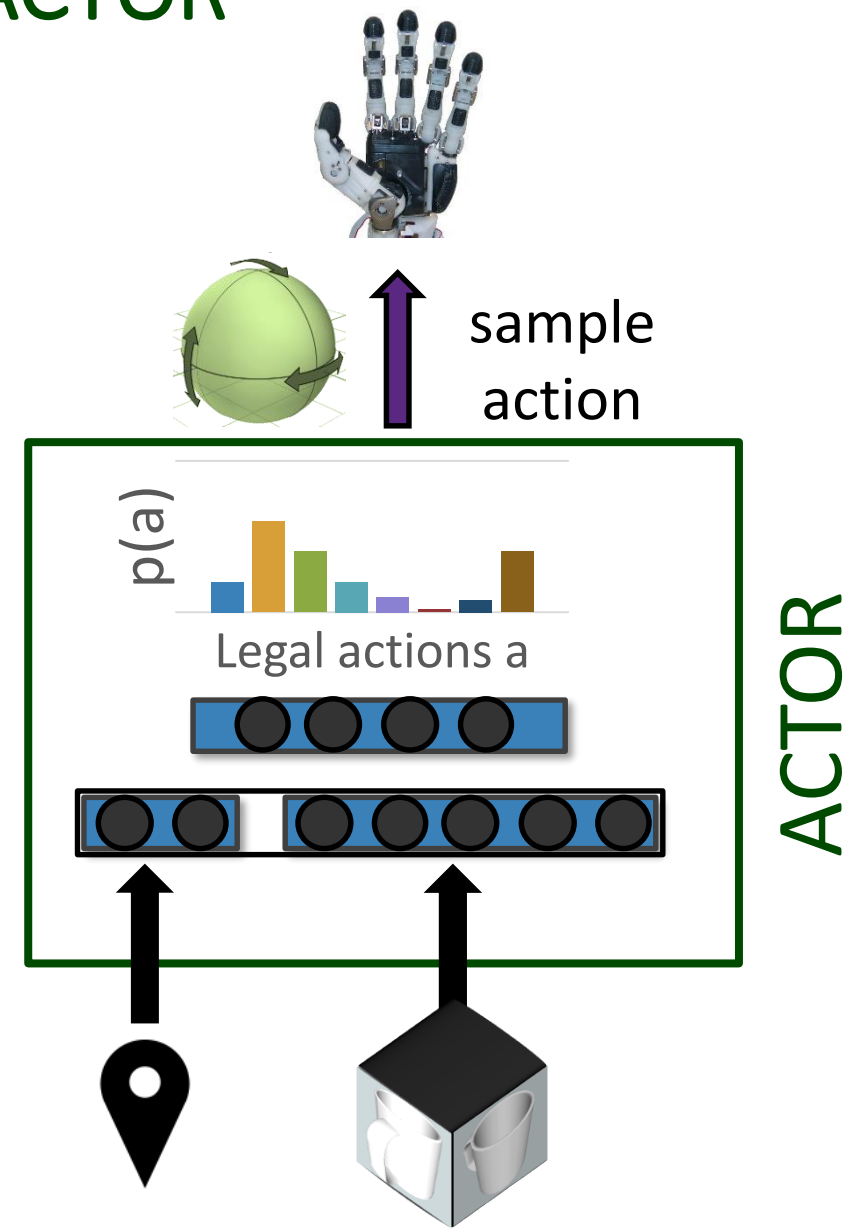
CLASSIFIER

ACTOR

LOOK-AHEAD

AGGREGATOR

SENSOR



Classification – CLASSIFIER

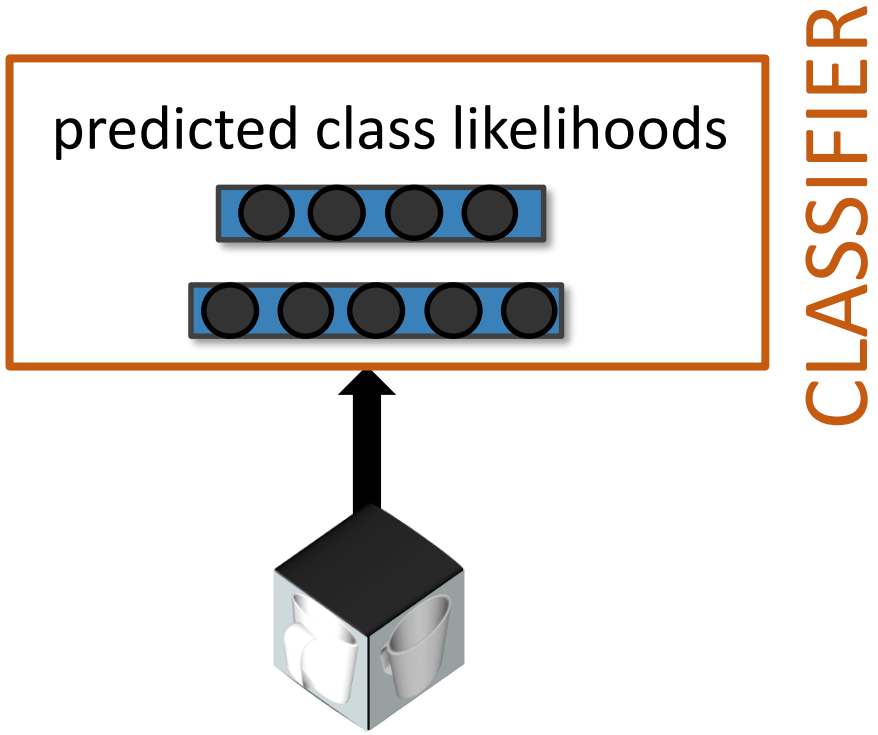
CLASSIFIER

ACTOR

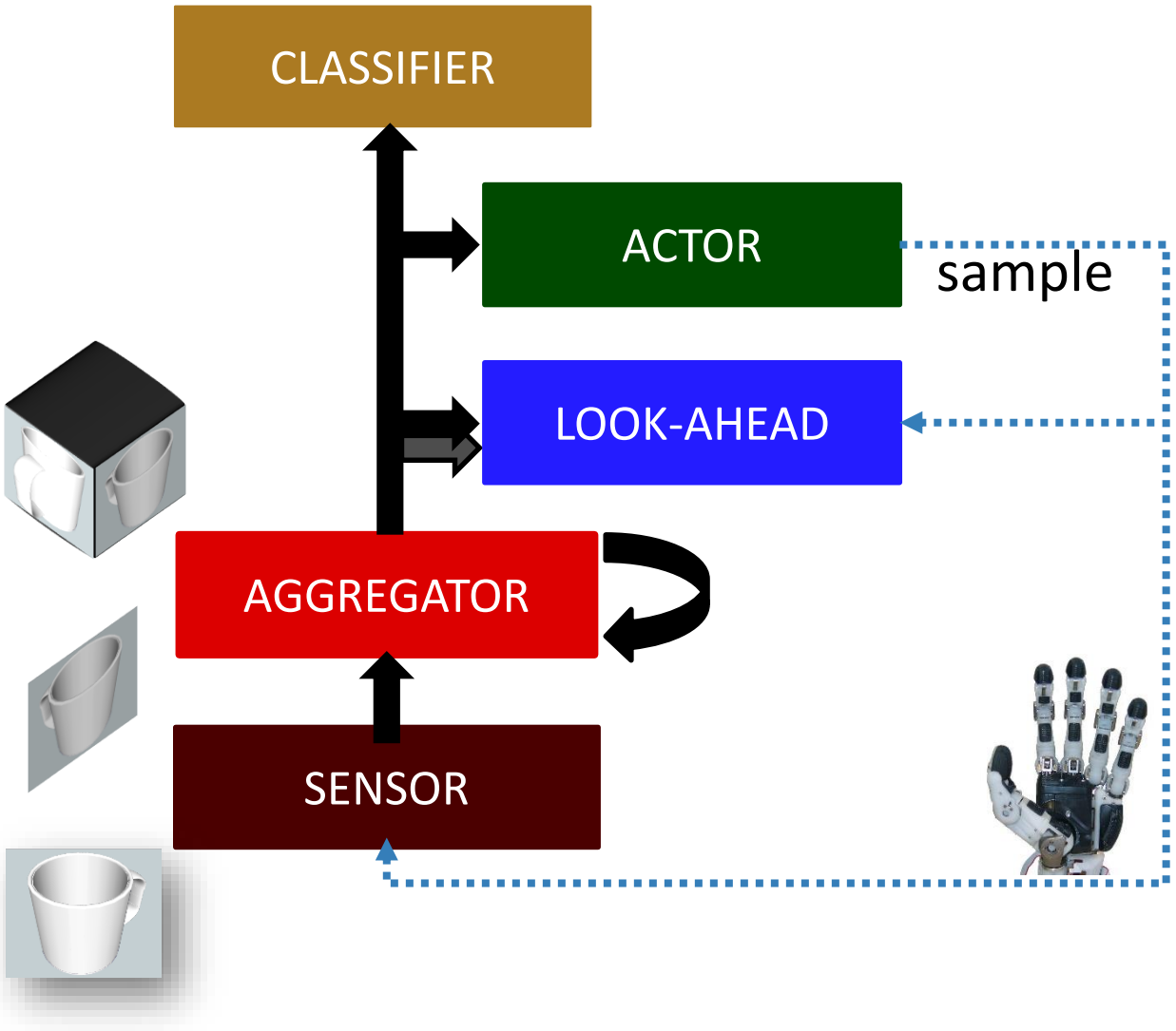
LOOK-AHEAD

AGGREGATOR

SENSOR

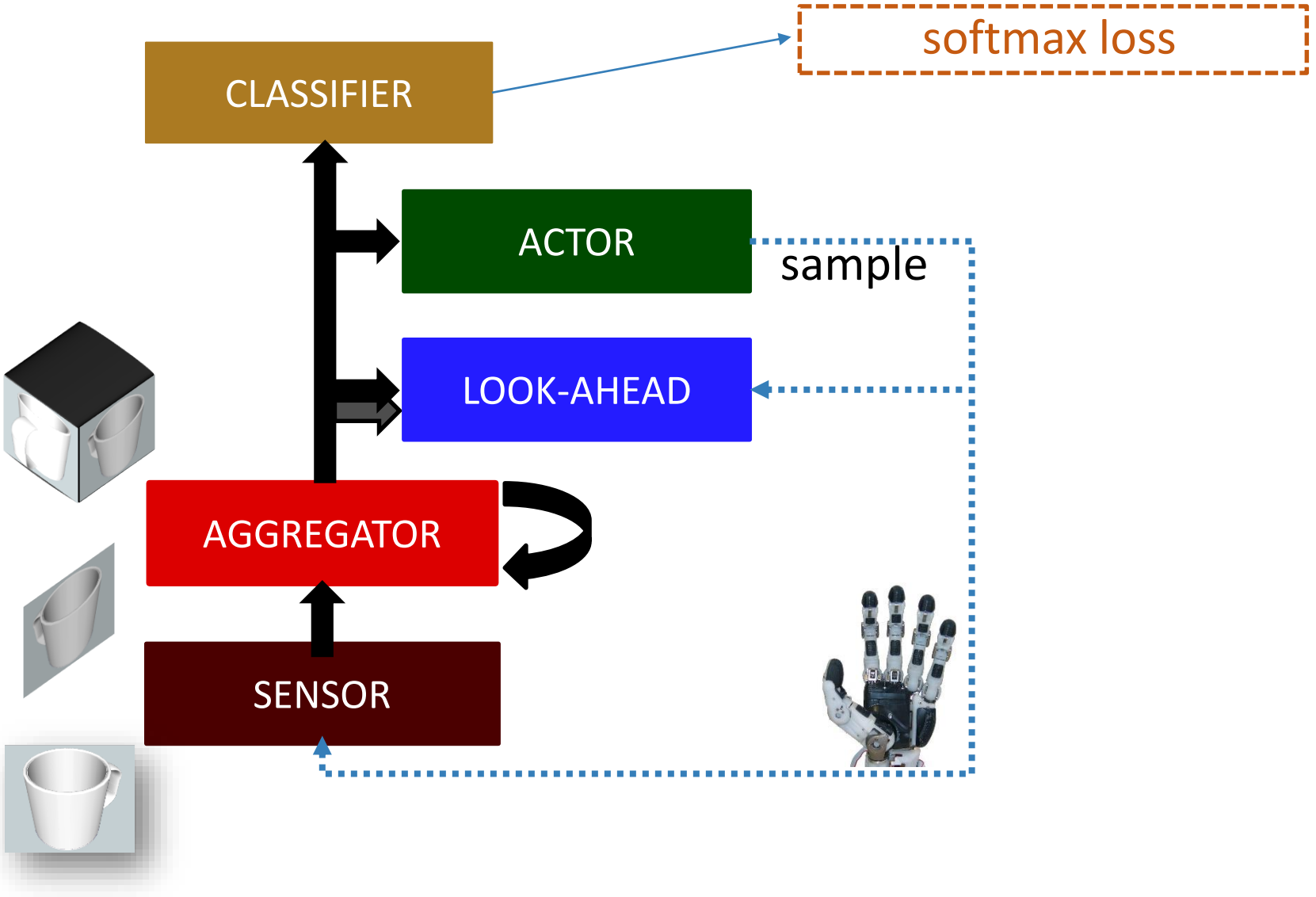


Training



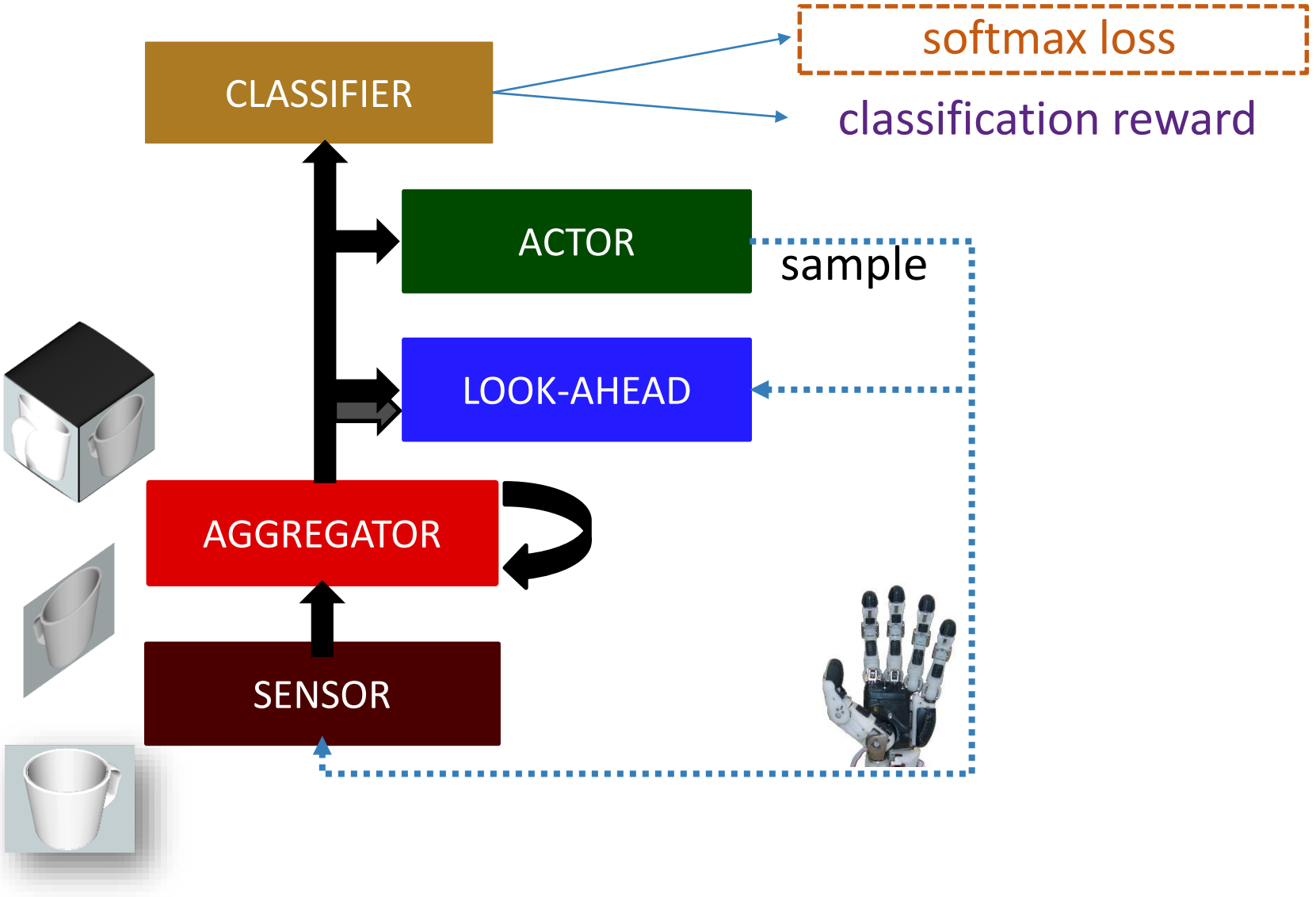
[Williams 1992, Mnih 2014]

Training



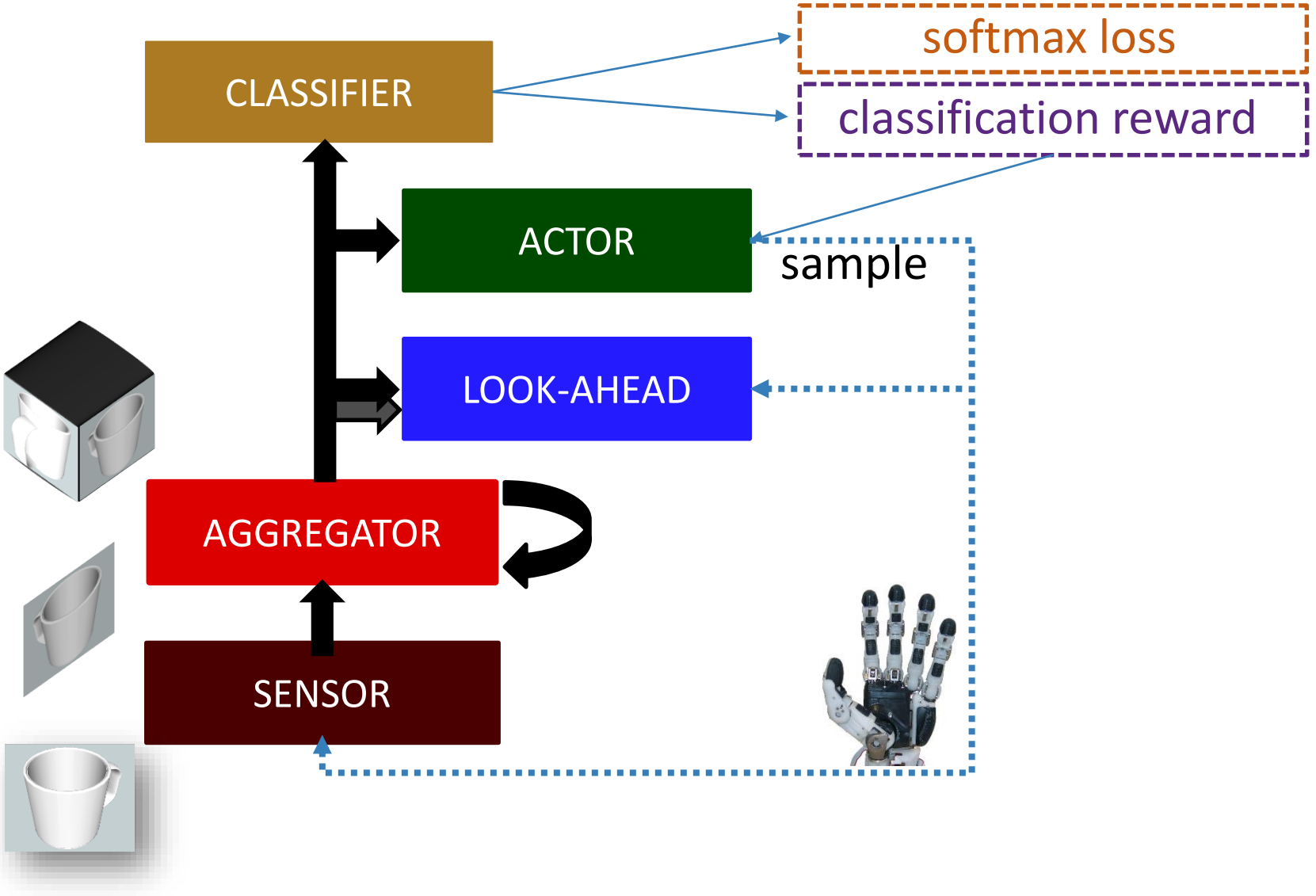
[Williams 1992, Mnih 2014]

Training



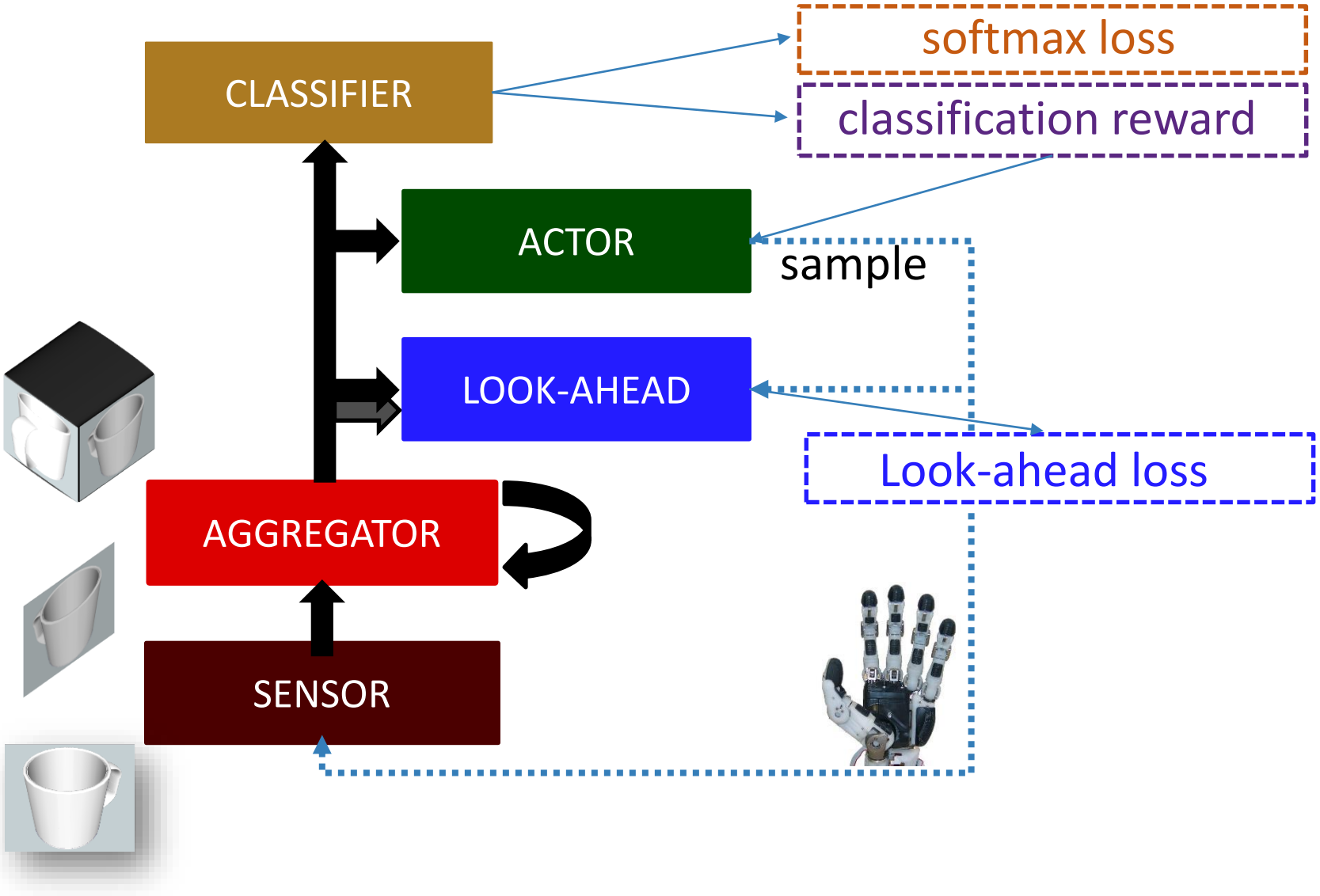
[Williams 1992, Mnih 2014]

Training



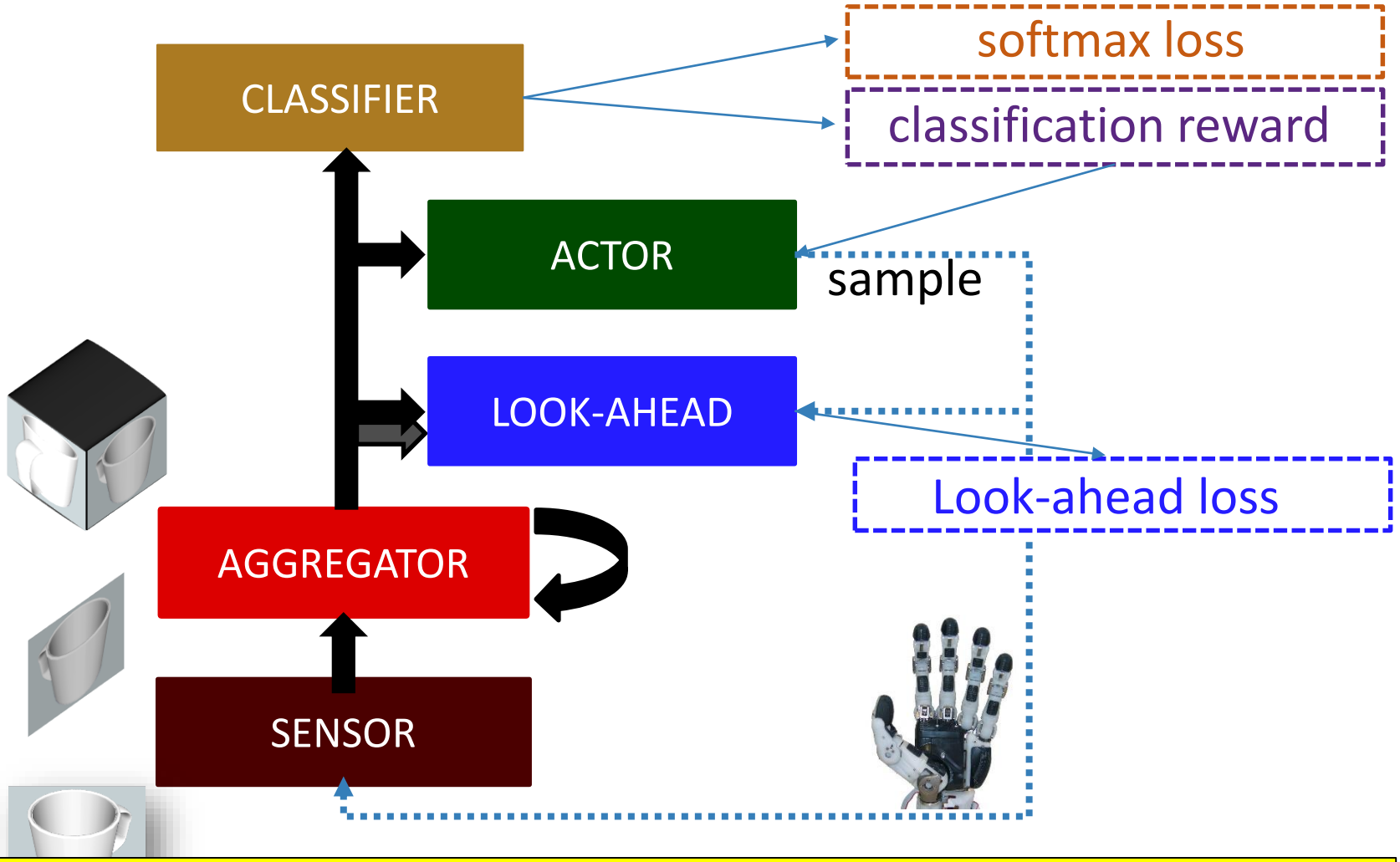
[Williams 1992, Mnih 2014]

Training



[Williams 1992, Mnih 2014]

Training



Trained through a combination of gradient descent and REINFORCE.

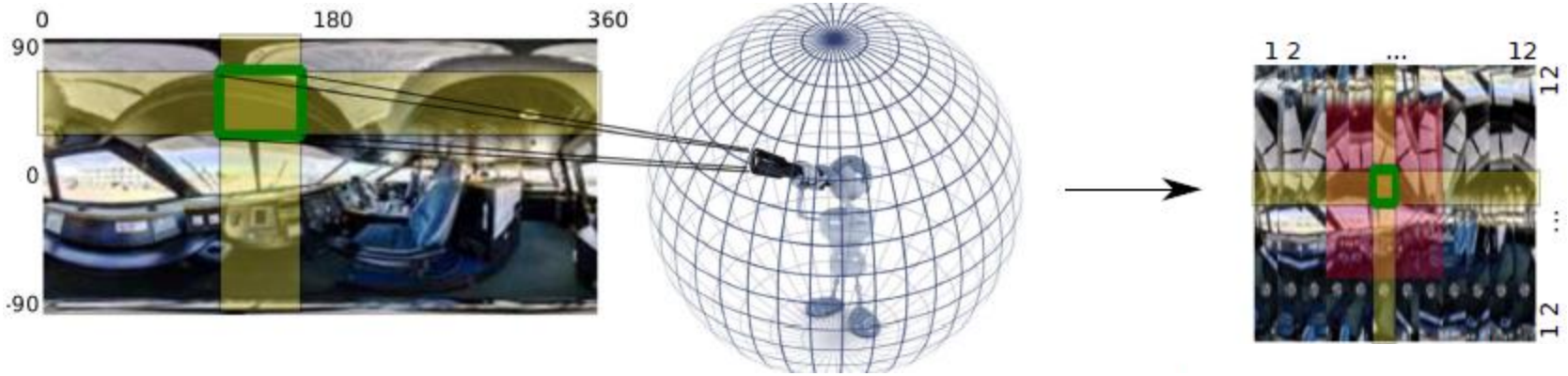
[Williams 1992, Mnih 2014]

Experiments

Experiments

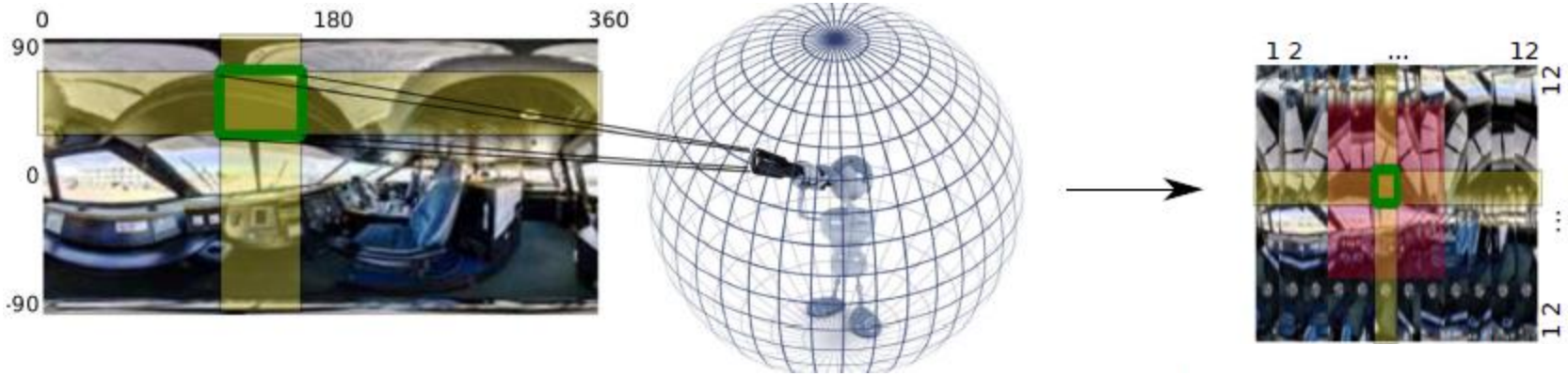
Experiments

SUN 360 panoramas [Xiao 2012]



Experiments

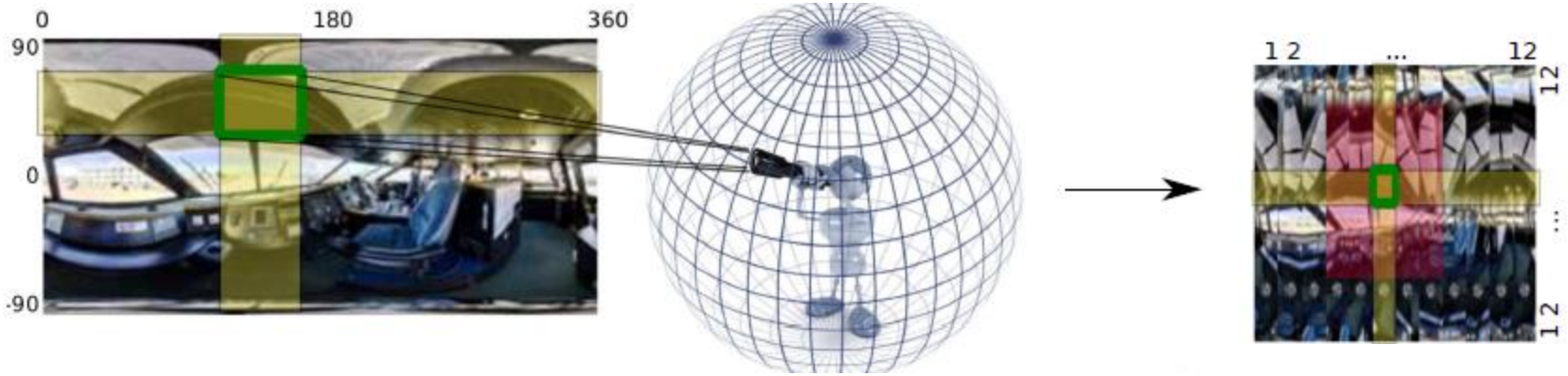
SUN 360 panoramas [Xiao 2012]



GERMS toy manipulation [Malmir 2015]

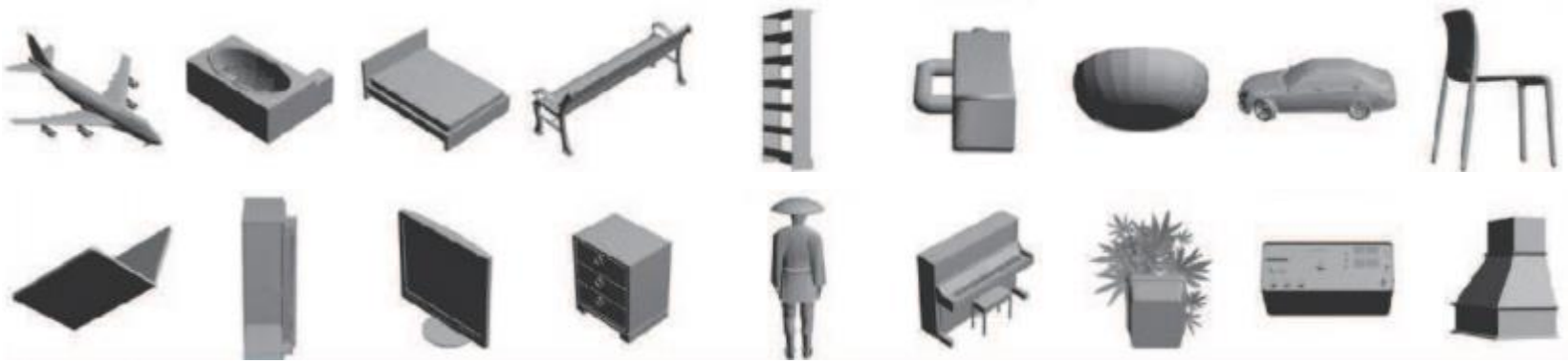
Experiments

SUN 360 panoramas [Xiao 2012]

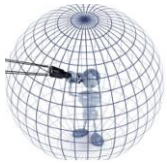


GERMS toy manipulation [Malmir 2015]

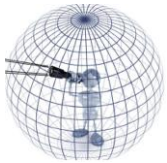
ModelNet-10 CAD models [Wu 2015]



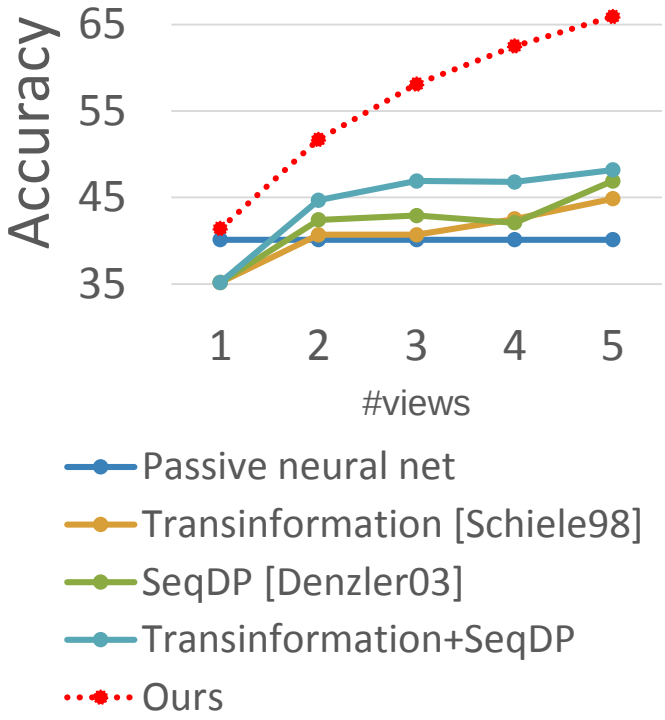
Quantitative results



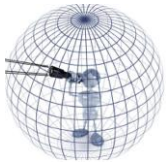
Quantitative results



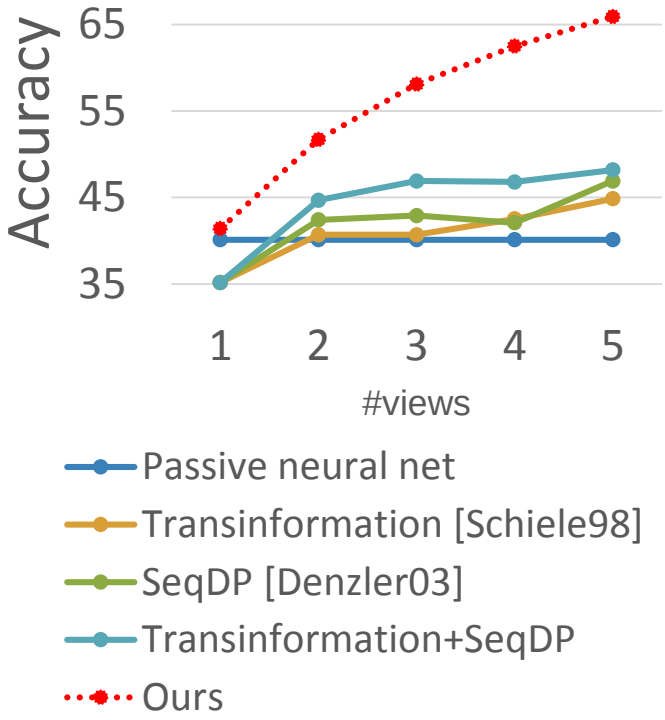
SUN 360



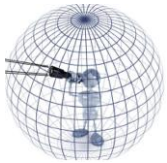
Quantitative results



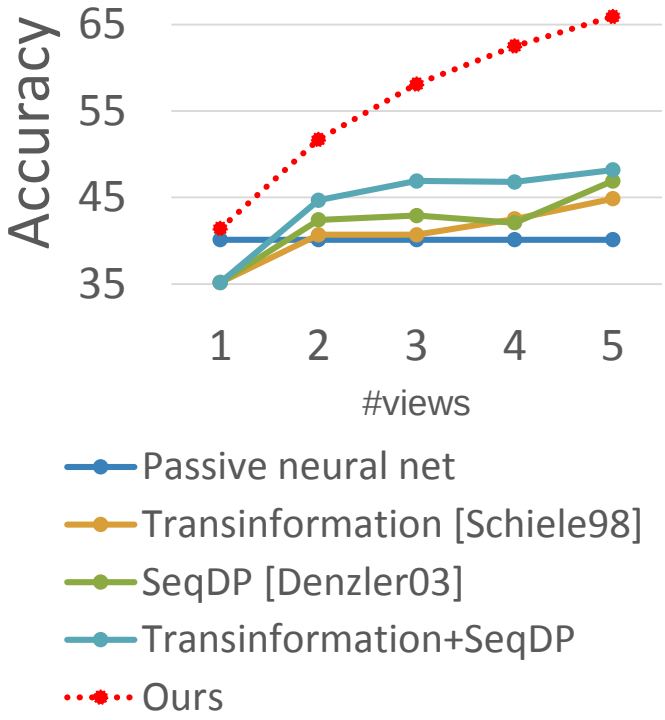
SUN 360



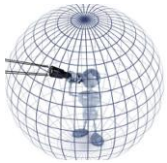
Quantitative results



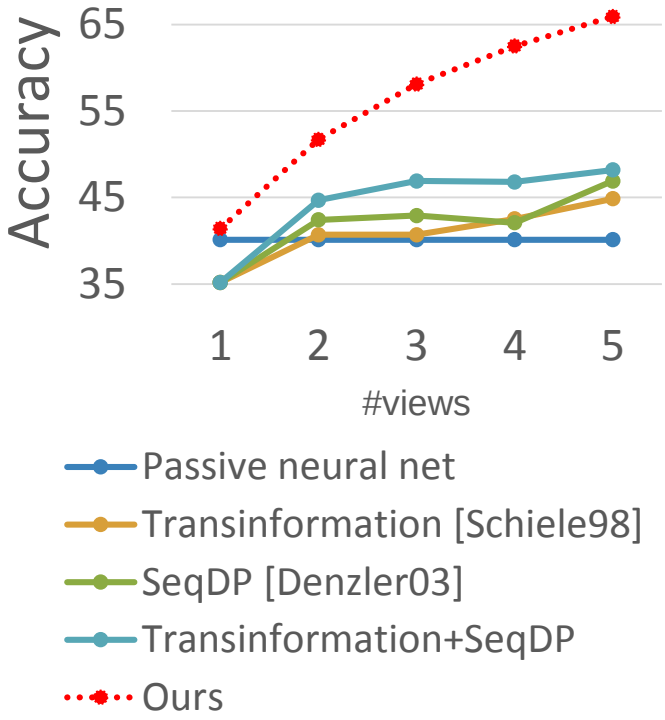
SUN 360



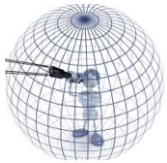
Quantitative results



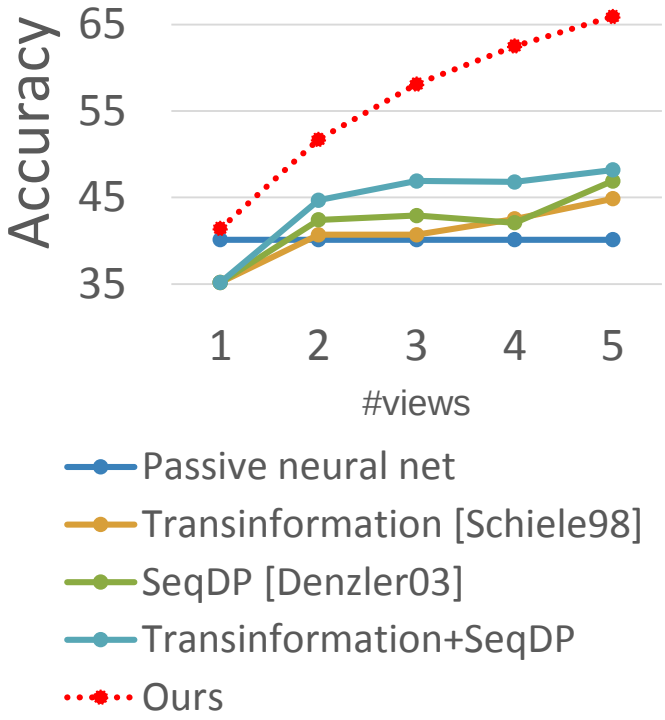
SUN 360



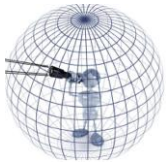
Quantitative results



SUN 360

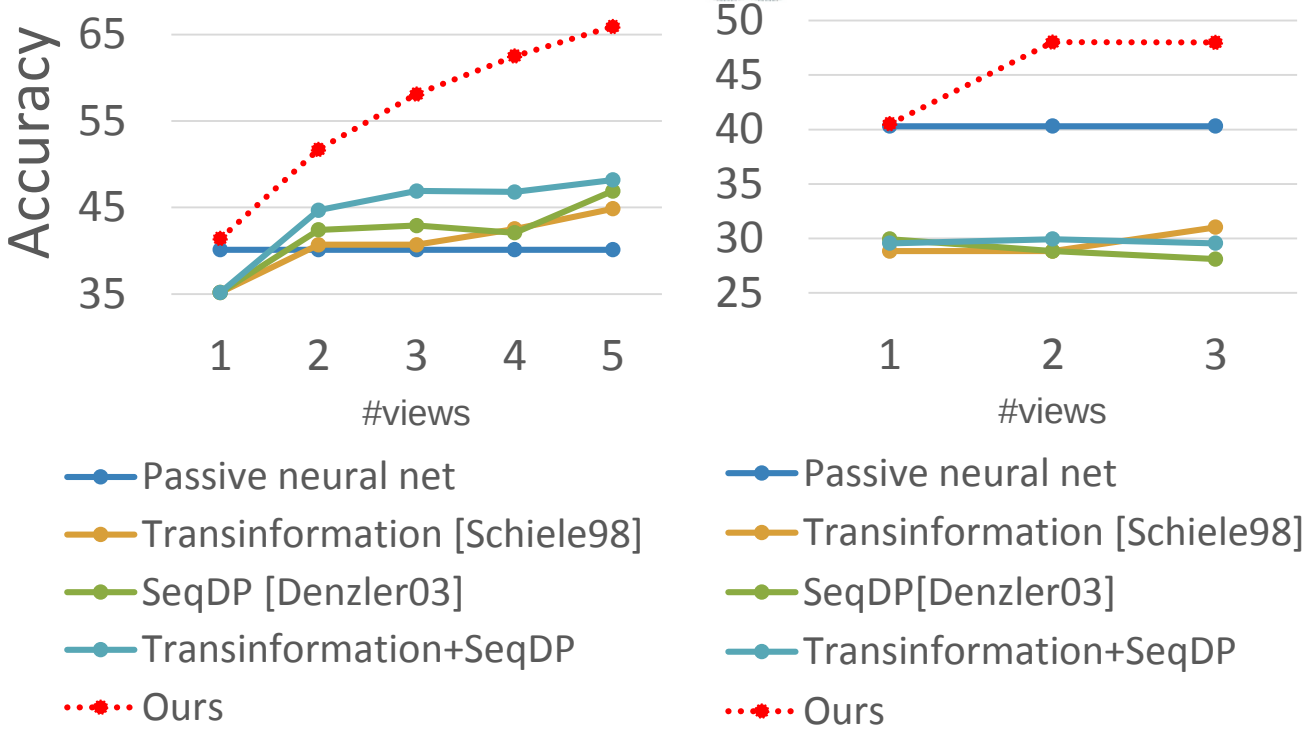


Quantitative results

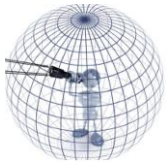


SUN 360

GERMS



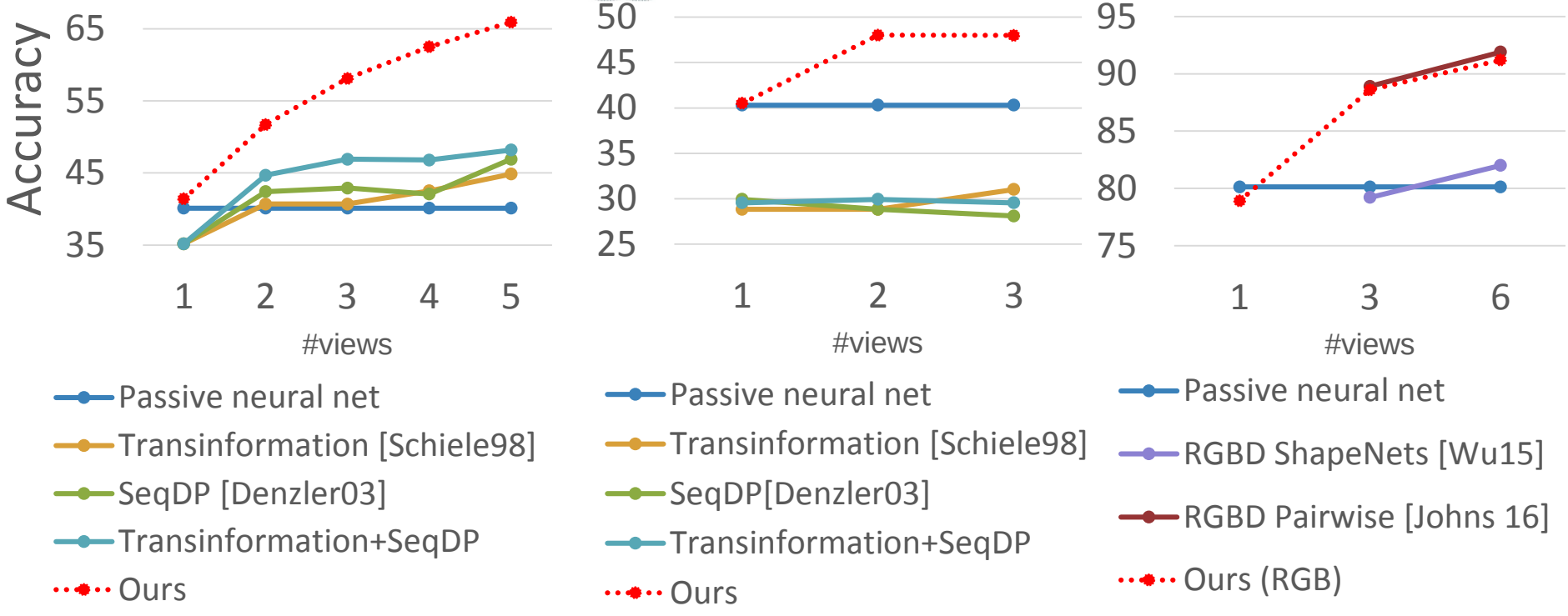
Quantitative results



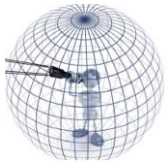
SUN 360

GERMS

ModelNet-10



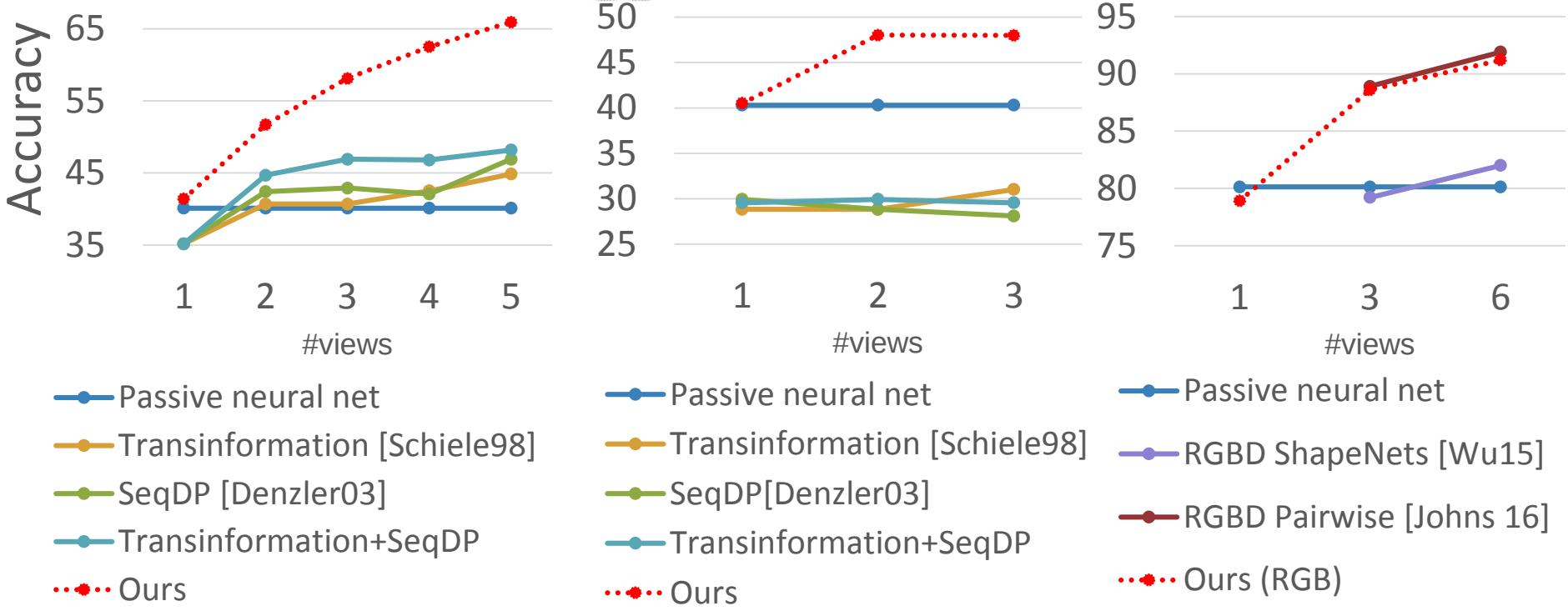
Quantitative results



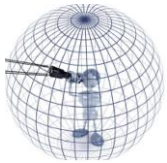
SUN 360

GERMS

ModelNet-10



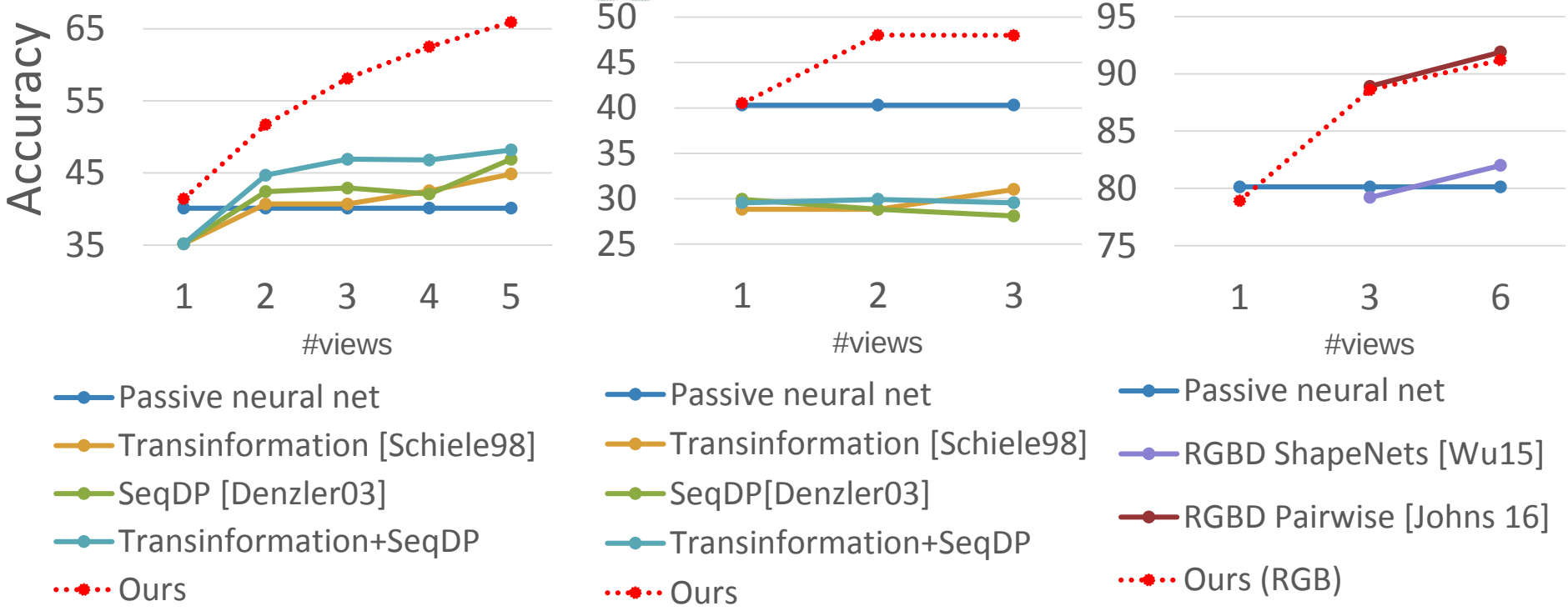
Quantitative results



SUN 360

GERMS

ModelNet-10



Our method strongly outperforms representative traditional active recognition approaches.

Active recognition: results

Active recognition: results

P(“Plaza courtyard”): (6.28%)

Top 3 guesses: Restaurant
Train interior
Shop

Active recognition: results

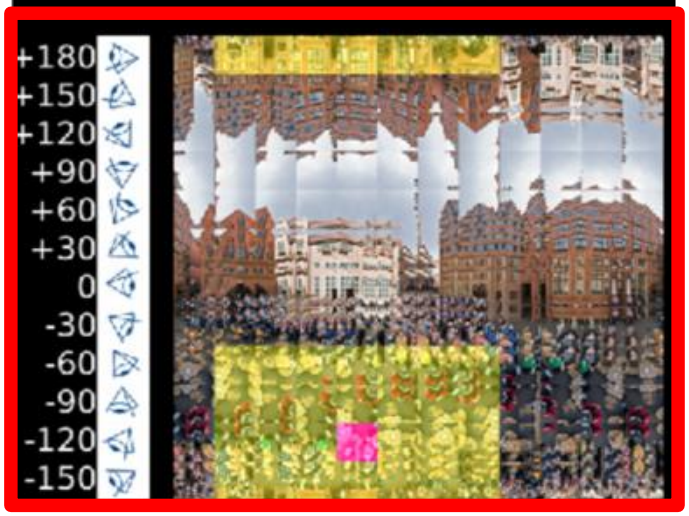
P(“Plaza courtyard”): (6.28%)

Top 3 guesses: Restaurant
Train interior
Shop

Active recognition: results

P(“Plaza courtyard”): (6.28%)

Top 3 guesses: Restaurant
Train interior
Shop



Active recognition: results

P(“Plaza courtyard”): (6.28%)

Top 3 guesses: Restaurant
Train interior
Shop

Active recognition: results

P(“Plaza courtyard”): (6.28%)

Top 3 guesses: Restaurant
Train interior
Shop

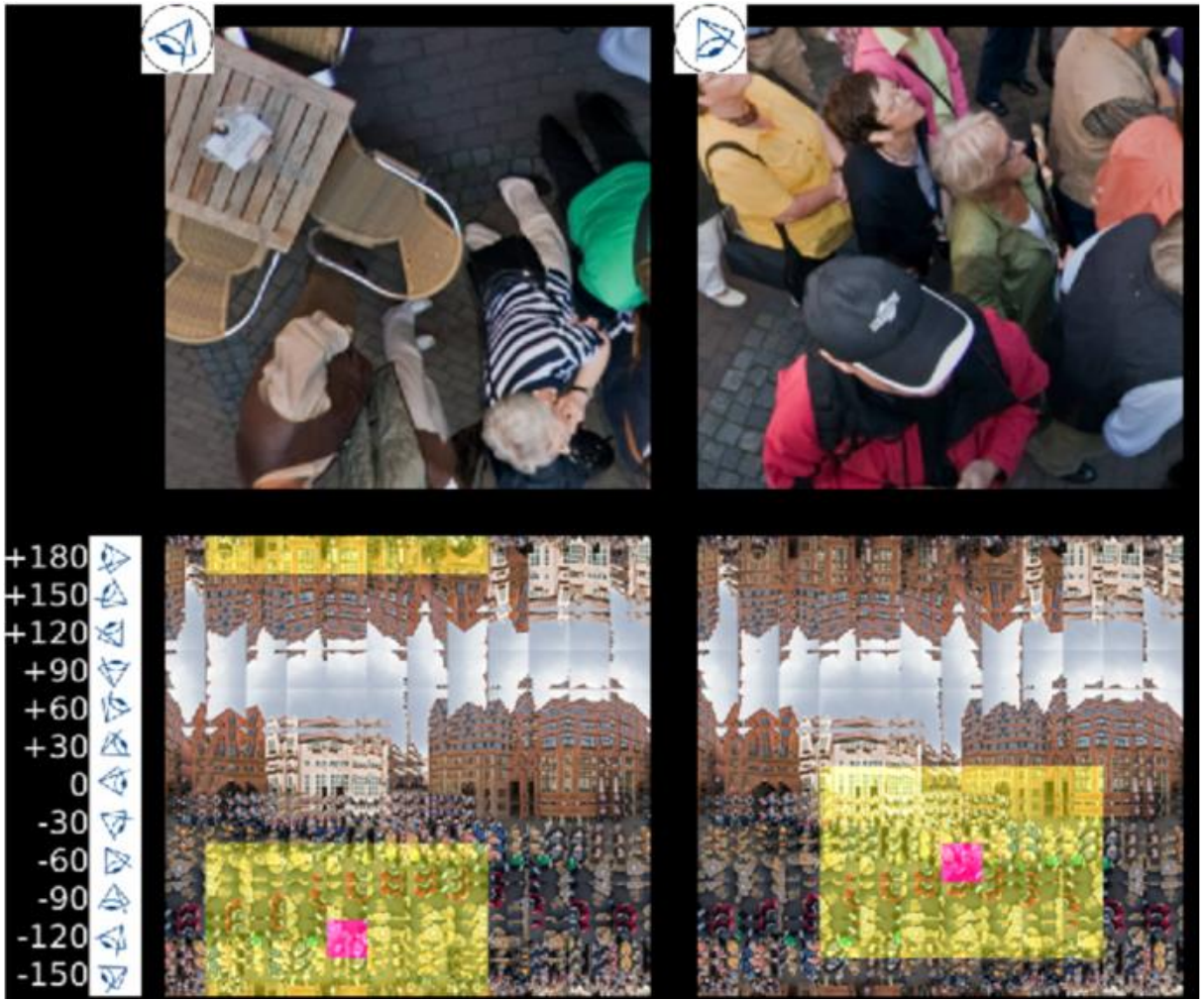
Active recognition: results

P(“Plaza courtyard”): (6.28%) (11.95)

Top 3 guesses: Restaurant Theater

Train interior Restaurant

Shop Plaza courtyard



Active recognition: results

Active recognition: results

Active recognition: results

Predicted label:

T=1

Active recognition: results

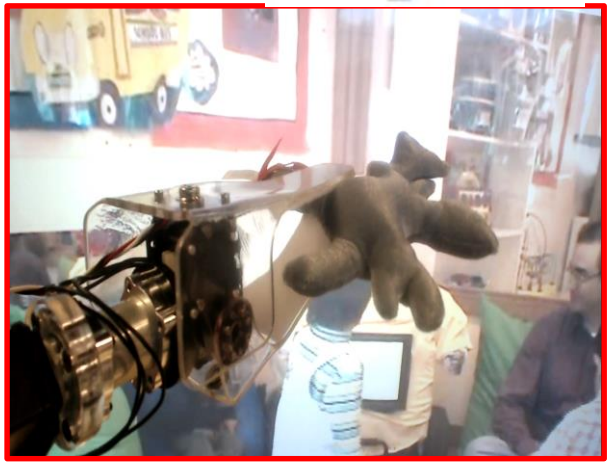
Predicted label:

T=1

T=2

Active recognition: results

Predicted label:



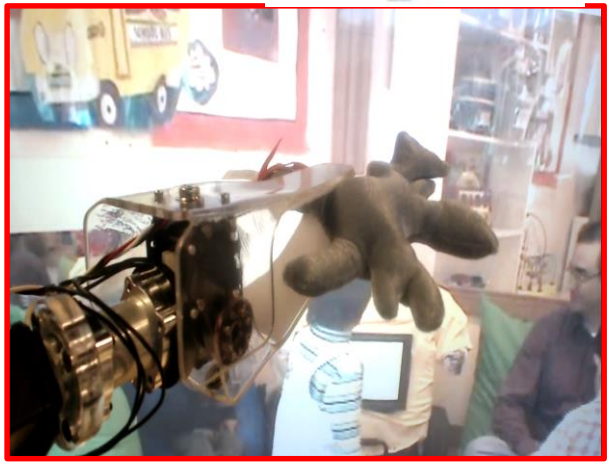
T=1

T=2

T=3

Active recognition: results

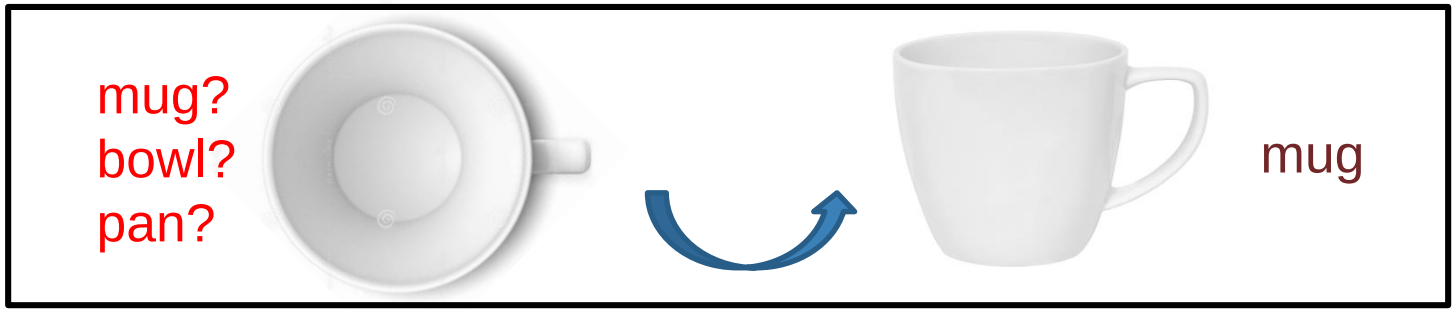
Predicted label:



T=1 T=2 T=3

More examples in paper and supplementary material!

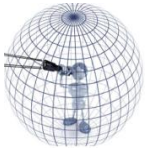
Summary



- Joint end-to-end active recognition.
- Improvement with auxiliary look-ahead task.
- Realistic but reproducible experimental settings.

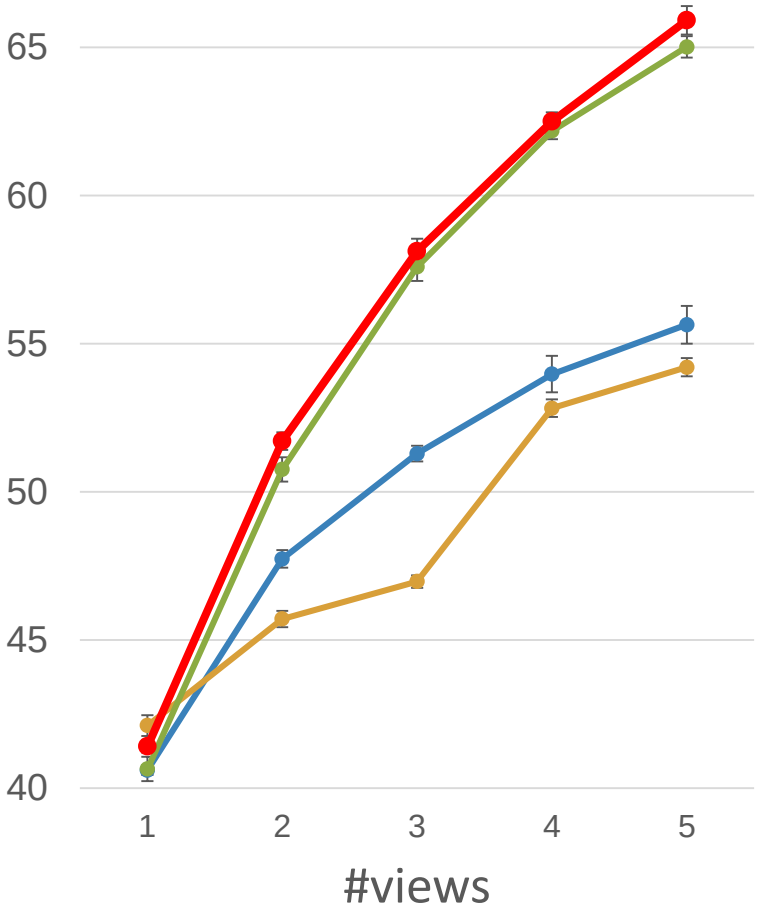
Data and code soon at
http://www.cs.utexas.edu/~dineshj/

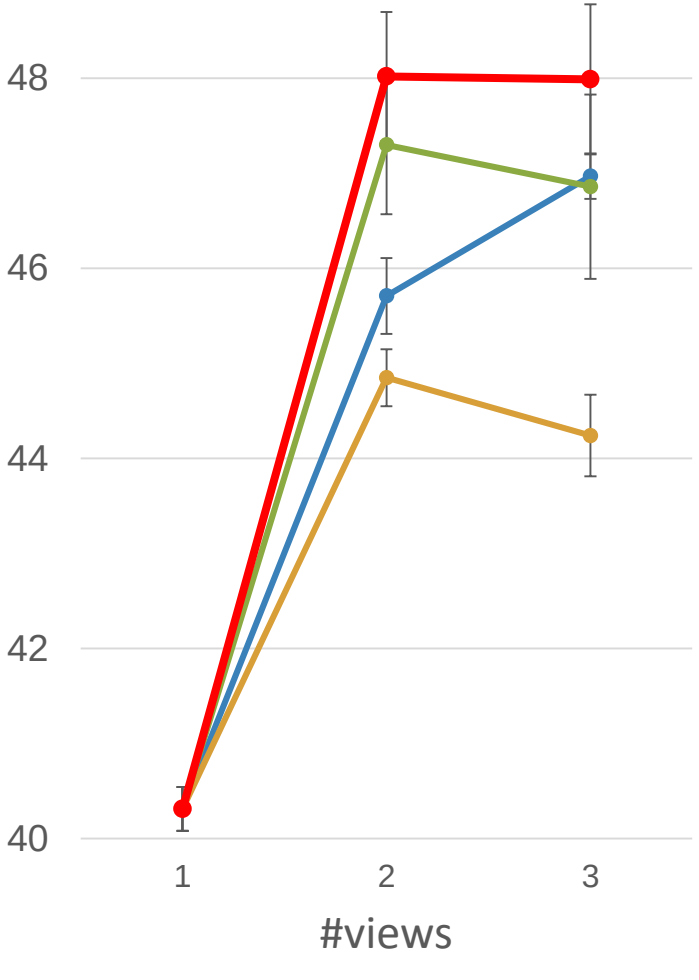
THE UNIVERSITY OF TEXAS AT AUSTIN



SUN 360

GERMS





Random views (avg)

Random views (recurrent)

Active RNN (ours)

Look-ahead active RNN (ours)

Training all 3 components jointly is most critical to performance.