Time-Contrastive Networks: Self-Supervised Learning from Video

Presenters : Kshama Nitin Shah , Kemmannu Vineet Rao

Sermanet, Pierre, Corey Lynch, Jasmine Hsu, and Sergey Levine. ‘Time-Contrastive Networks: Self-Supervised Learning from Multi-View Observation’. *CoRR* abs/1704.06888 (2017)

Motivation

- Learning to perform tasks by imitating humans
- Robot needs to figure out whether behavior is similar to human behavior
- Self supervised learning from unlabeled videos without any provided correspondences
- Learning from third person views or observations

Sermanet, Pierre, Corey Lynch, Jasmine Hsu, and Sergey Levine. 'Time-Contrastive Networks: Self-Supervised Learning from Multi-View Observation'. *CoRR* abs/1704.06888 (2017)

Motivation

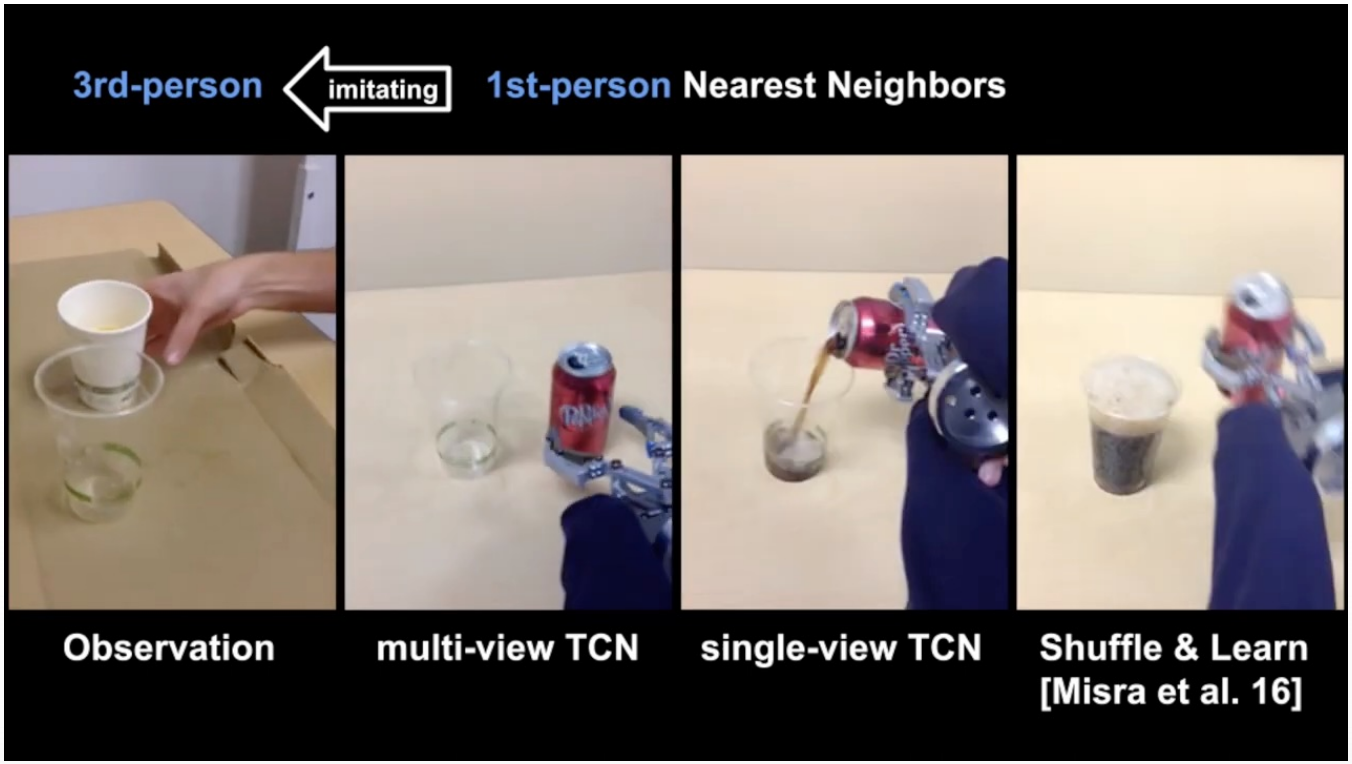
Imitation Learning
Intuition

https://www.standupcomedyclinic.com/5-reasons-to-use-imitation-and-emulation-to-learn-stand-up-comedy/

Introduction

- Modeling human dynamics using time contrastive networks
- Learning representations and robotic behaviours from unlabeled videos and multiple viewpoints
- Representations used to imitate human poses and object interactions
- Addressed through self supervision and multi-viewpoint representation learning

Introduction



Sermanet, Pierre, Corey Lynch, Jasmine Hsu, and Sergey Levine. 'Time-Contrastive Networks: Self-Supervised Learning from Multi-View Observation'. *CoRR* abs/1704.06888 (2017)

Related Work

Imitation Learning : Mirror Neurons

- Humans are able to learn by imitation
- Group of specialized neurons that mimic the behavior and actions of others
- Allows humans to mentally simulate an observation
- Viewpoint invariance crucial to imitation

Related Work

Imitation Learning : Mirror Neurons

https://www.youtube.com/watch?time_continue=11&v=qSRFvE0Z8Wg&embeds_euri=https%3A%2F%2Fdocs.google.com%2F&embeds_origin=https%3A%2F%2Fdocs.google.com&source_ve_path=MjM4NTE&feature=emb_title

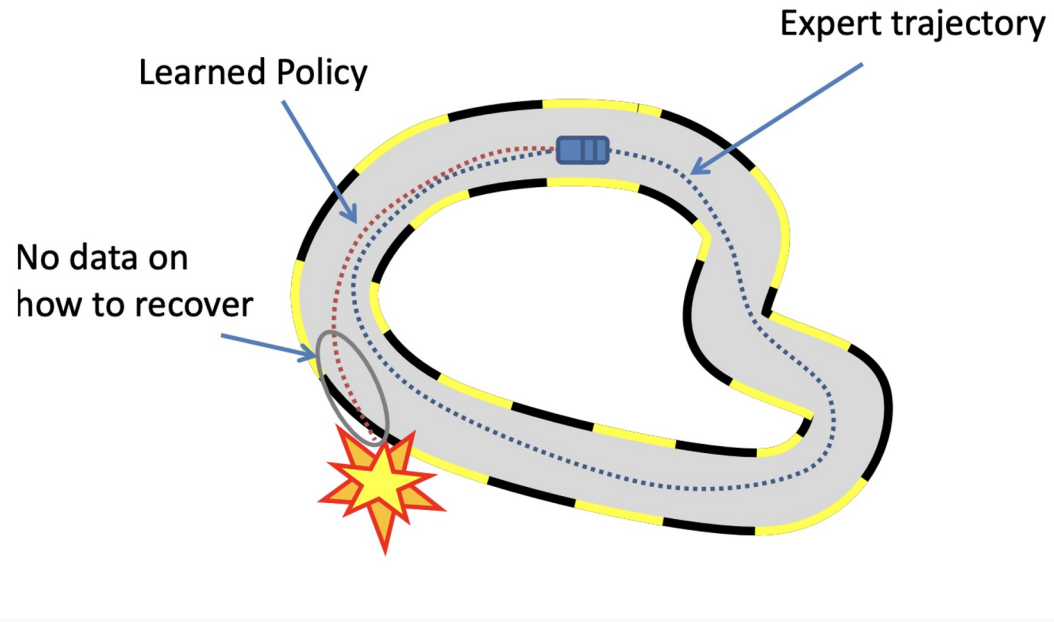
Related Work

Imitation Learning in robotics

Two Broad Approaches :

- Direct : Behavioral cloning - supervised training of policy by my learning to map state to actions.
- Indirect : Inverse RL - expert demonstrations are used to learn the unknown reward functions, then derive a optimal policy

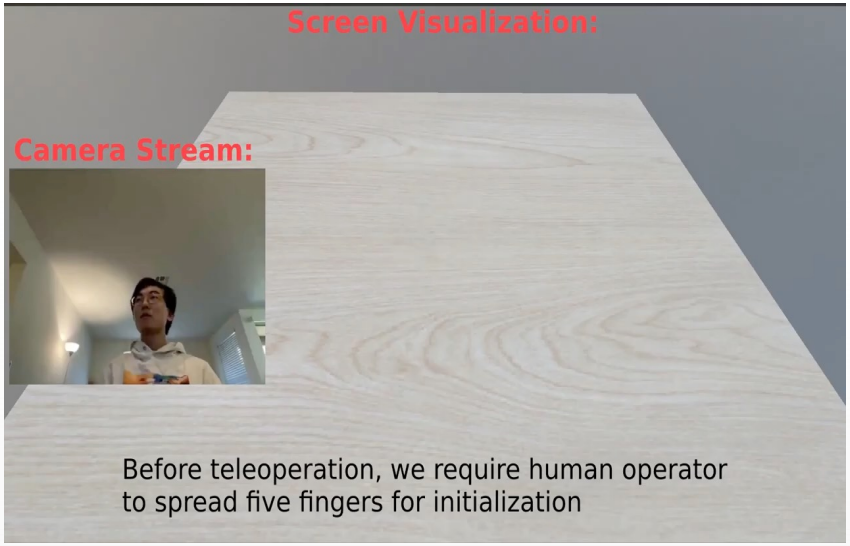
Behavior Cloning



Slide borrowed from : https://www.andrew.cmu.edu/course/10-703/slides/Lecture_Imitation_supervised-Nov-5-2018.pdf

Related Work

- Kinesthetic demonstrations or teleoperation

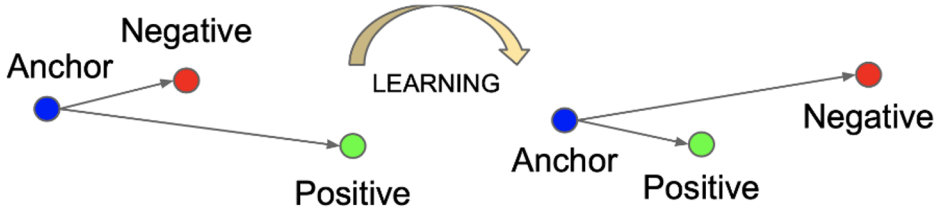


Calinon, Sylvain, Florent Guenter, and Aude Billard. 'On Learning, Representing, and Generalizing a Task in a Humanoid Robot'. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 37, no. 2 (2007): 286–98. https://doi.org/10.1109/TSMCB.2006.886952
Qin, Yuzhe, Hao Su, and Xiaolong Wang. 'From One Hand to Multiple Hands: Imitation Learning for Dexterous Manipulation from Single-Camera Teleoperation'. *ArXiv [Cs.RO]*, 2023. arXiv. http://arxiv.org/abs/2204.12490

Related Work

Self-Supervised Representation Learning

- Goal of contrastive learning : positive and negative pairs
- Triplet Loss

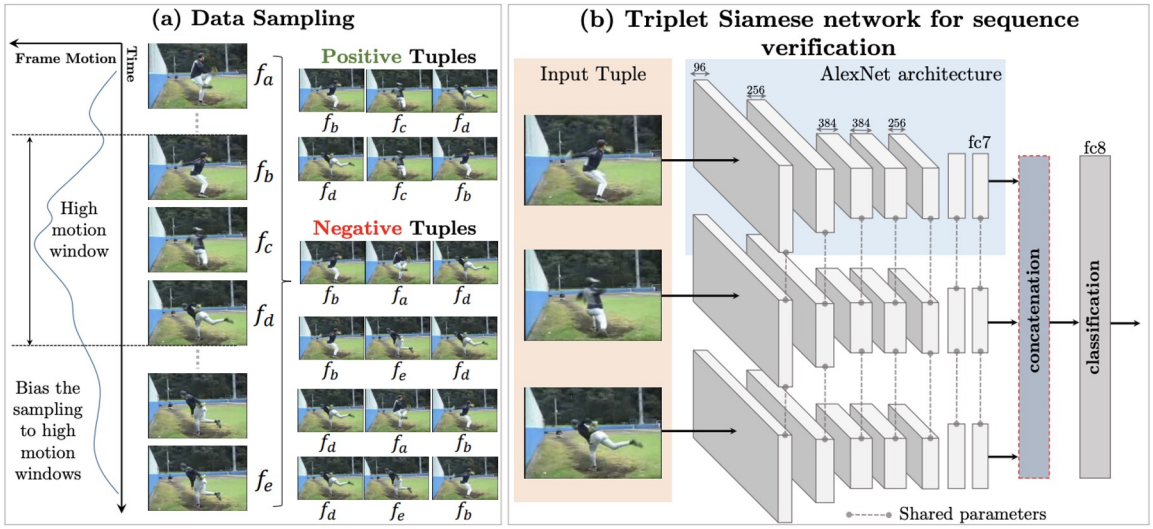


$$\mathcal{L}_{\text{triplet}}(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-) = \sum_{\mathbf{x} \in \mathcal{X}} \max\left(0, \|f(\mathbf{x}) - f(\mathbf{x}^+)\|_2^2 - \|f(\mathbf{x}) - f(\mathbf{x}^-)\|_2^2 + \epsilon\right)$$

Weng, Lilian. 'Contrastive Representation Learning'. *Lilianweng. Github. Io*, May 2021. https://lilianweng.github.io/posts/2021-05-31-contrastive/

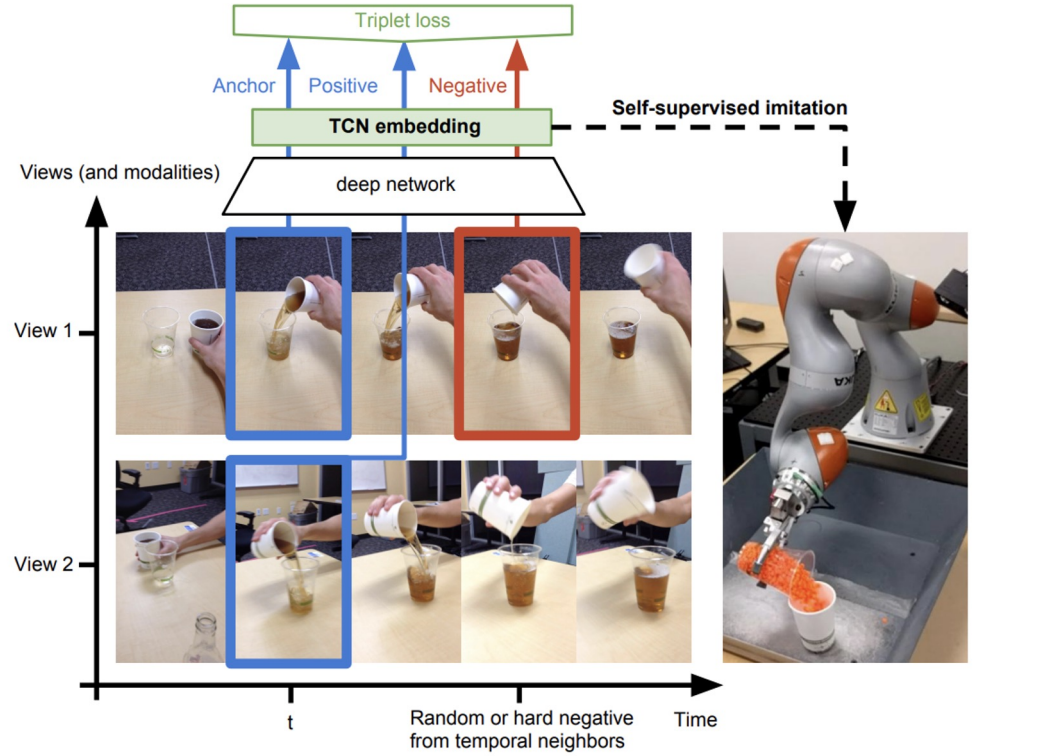
Related Work

Shuffle and Learn



Misra, Ishan, C. Lawrence Zitnick, and Martial Hebert. 'Unsupervised Learning Using Sequential Verification for Action Recognition'. *CoRR* abs/1603.08561 (2016)

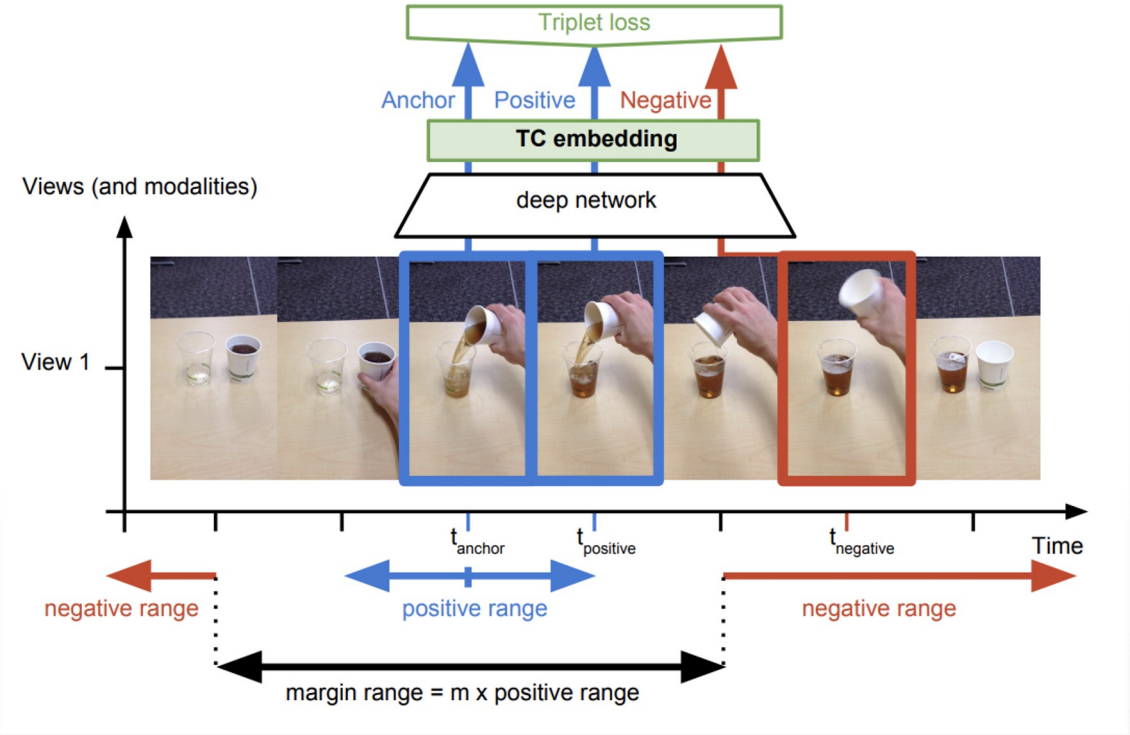
Methodology - Object Interaction Model - Multiview TCN



Methodology - Object Interaction Model - Multiview TCN

- Learns disentangled representations without labels
- Learns viewpoint, scale, occlusion, motion-blur, lighting and background invariance
- Visual changes over time modeled by temporal competition by neighboring frames
- Correspondences between different agents
- Attributes pertinent to the task

Methodology- Object Interaction Model - Single View TCN



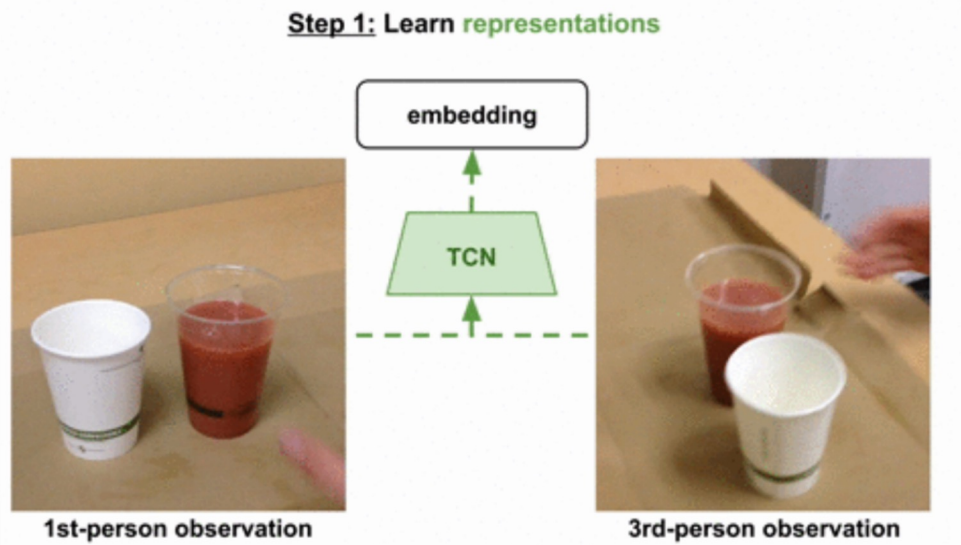
ImageNet Pre-Training

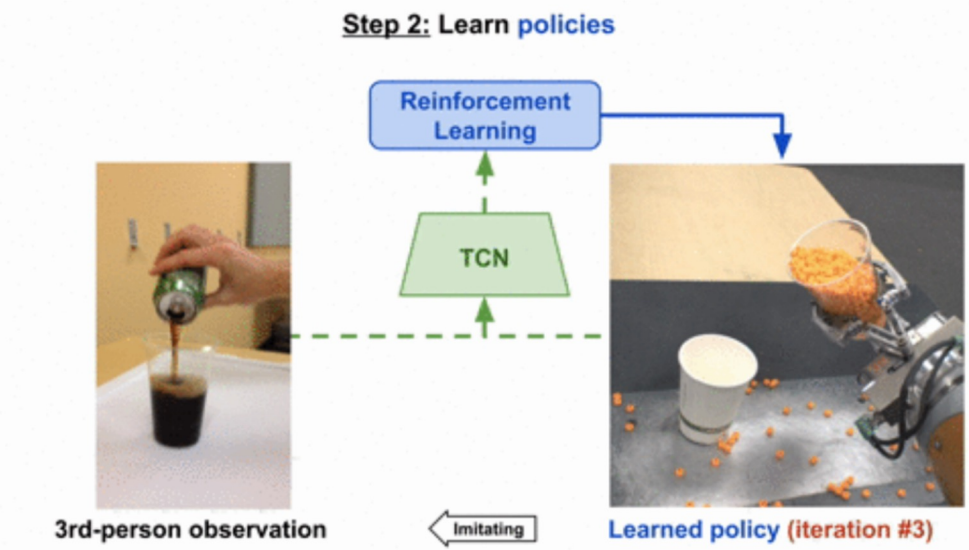
- Deep Network used - InceptionNet
- Pre-trained on ImageNet
- Tasks used objects that were present in ImageNet

Learning Robotic Behaviors using RL

- Reward Function - Distance between TCN embedding of human demonstration and robot camera images
- Policy update using implicit reward function
- Enables learning of object interaction skills directly from videos

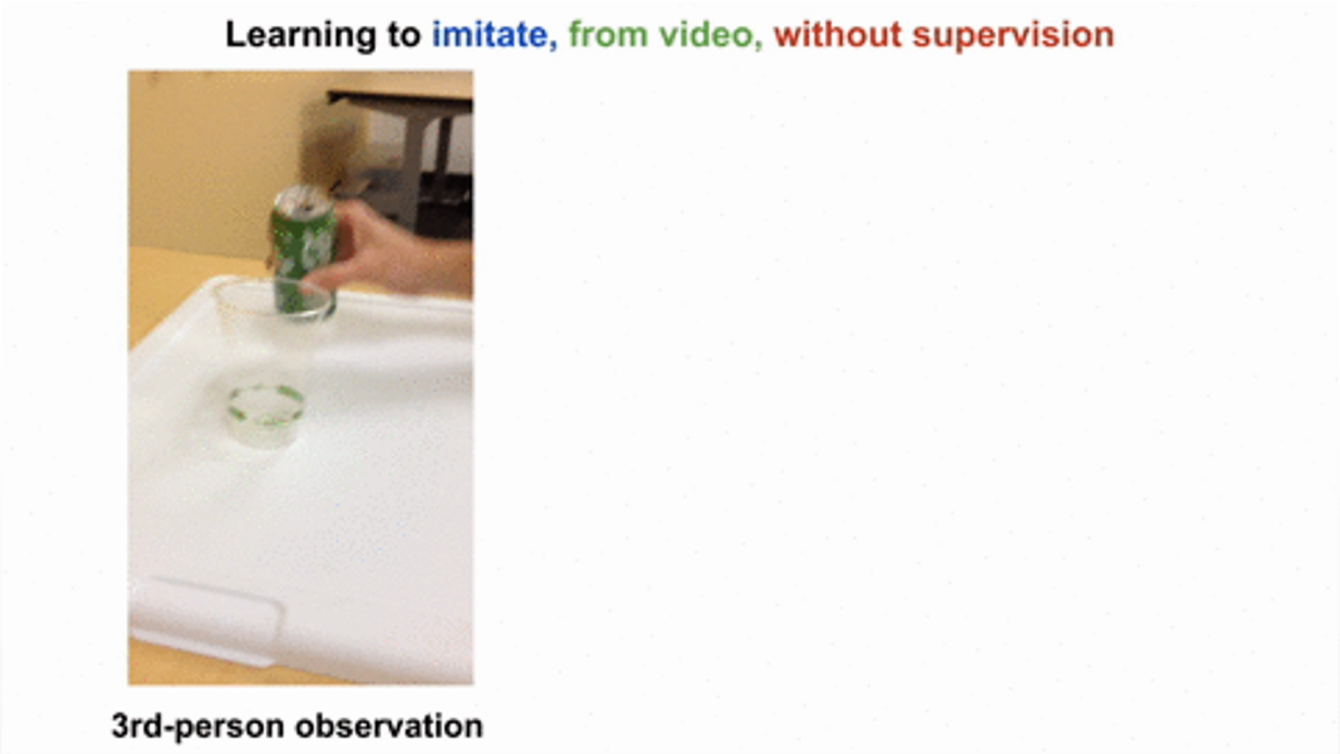
Methodology - Object Interaction Model - Multiview TCN





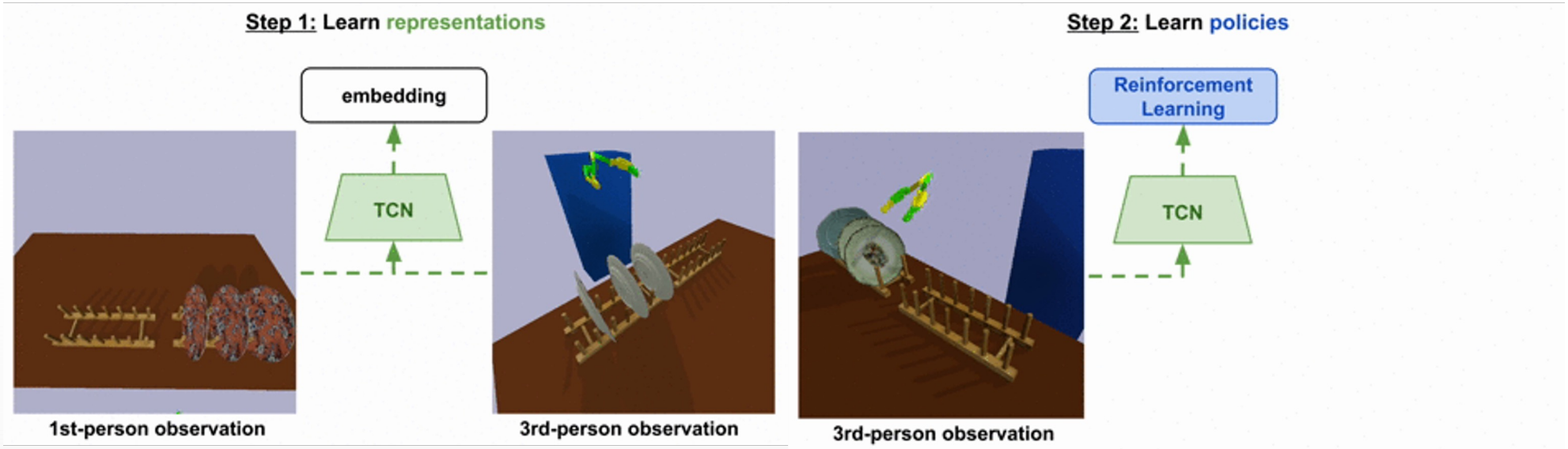
Resulting Policies

Pouring Task

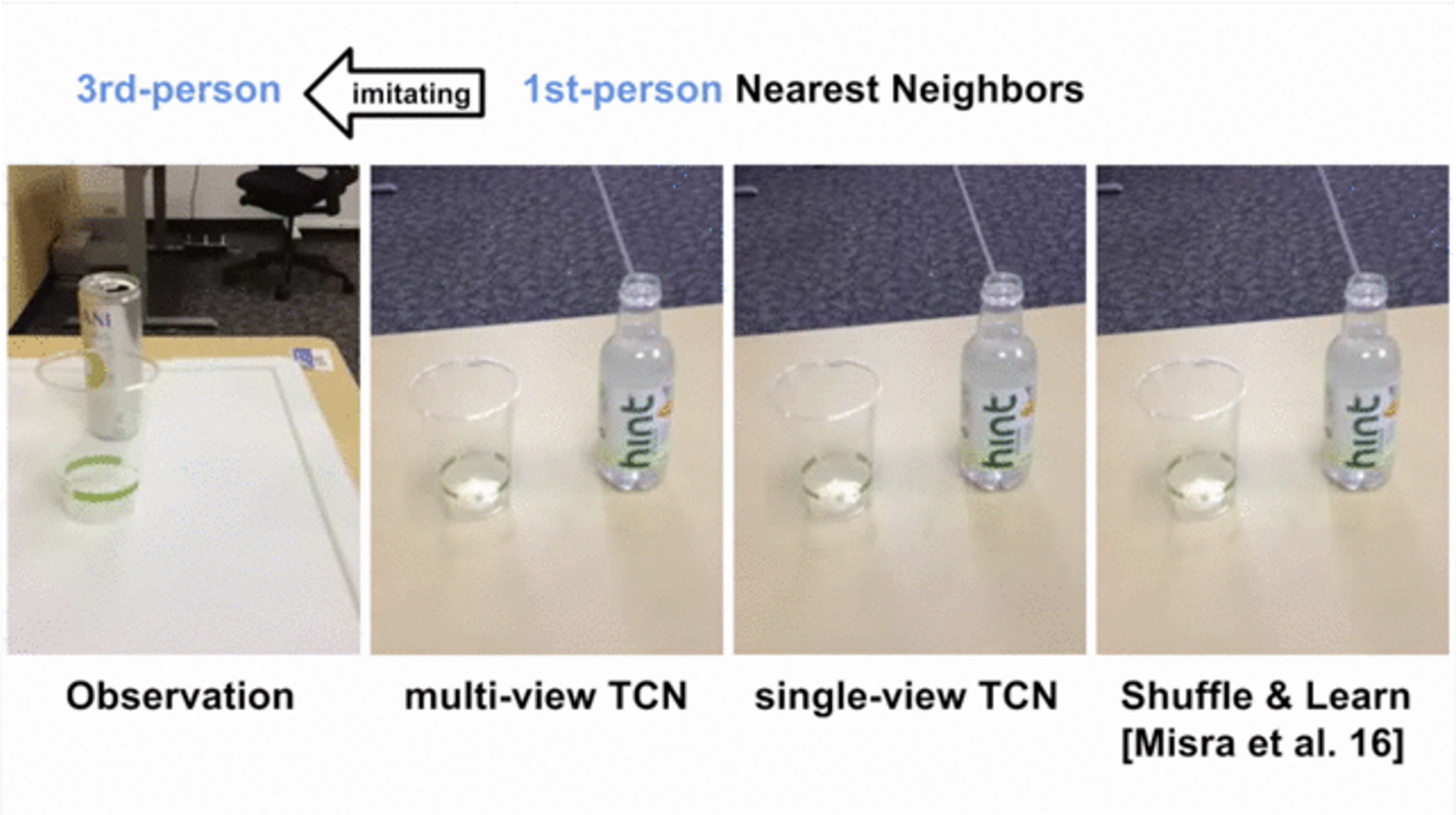


Resulting Policies

Simulated Dish Rack Task



Results : Qualitative (Nearest Neighbor Imitation)



Results:Quantitative

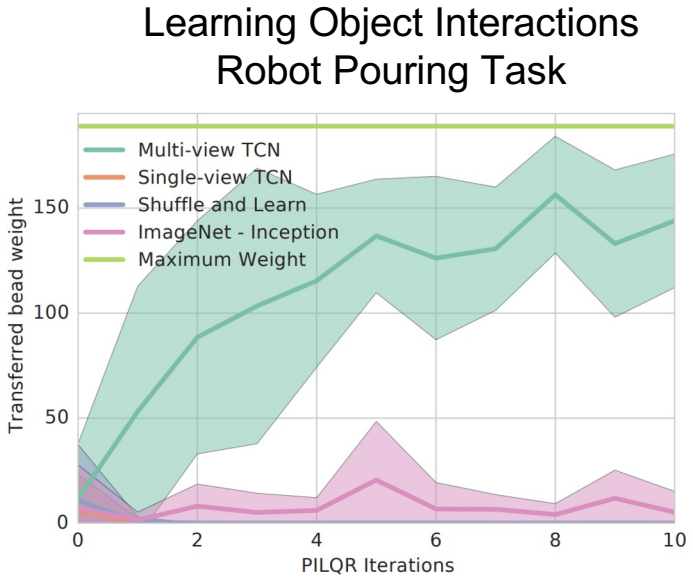


Fig. 7: **Learning progress of the pouring task,** using a single 3rd-person human demonstration, as shown in Fig. 6. This graph reports the weight in grams measured from the target recipient after each pouring action (maximum weight is 189g) along with the standard deviation of all 10 rollouts per iteration. The robot manages to successfully learn the pouring task using the multi-view TCN model after only 10 iterations.

Results:Quantitative

Model Selection

Method	alignment error	classif. error	training iteration
Random	28.1%	54.2%	-
Inception-ImageNet	29.8%	51.9%	-
shuffle & learn [31]	22.8%	27.0%	575k
single-view TCN (triplet)	25.8%	24.3%	266k
multi-view TCN (npairs)	18.1%	22.2%	938k
multi-view TCN (triplet)	18.8%	21.4%	397k
multi-view TCN (lifted)	18.0%	19.6%	119k

TABLE I: **Pouring alignment and classification errors:** all models are selected at their lowest validation loss. The classification error considers 5 classes related to pouring detailed in Table II.

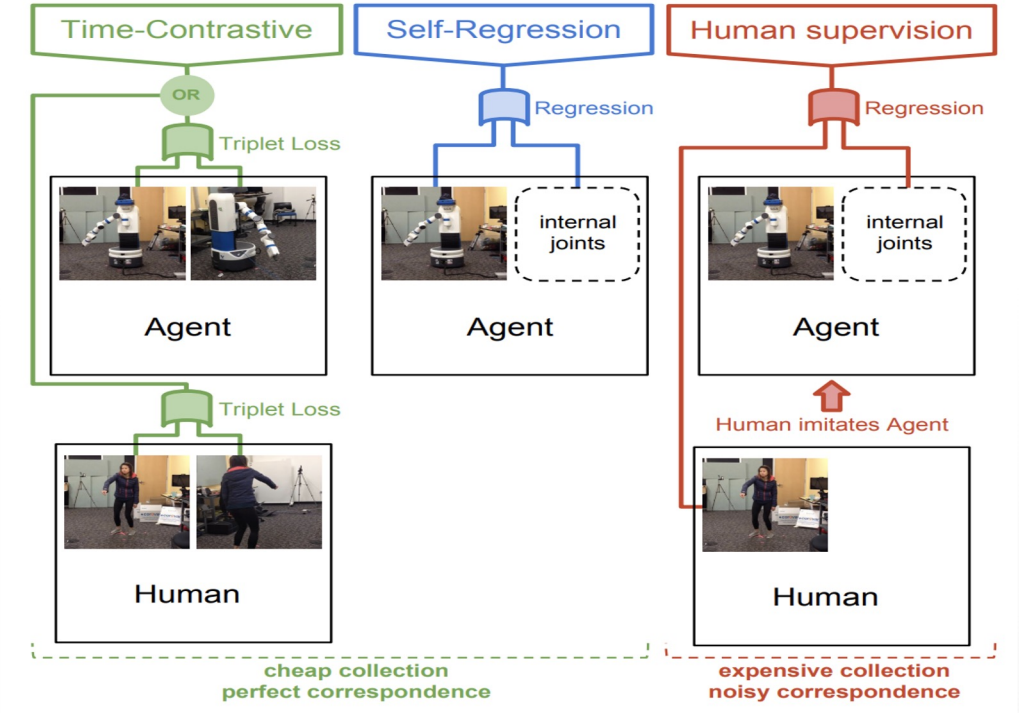
Results:Quantitative

Detailed Attribute Classification Errors

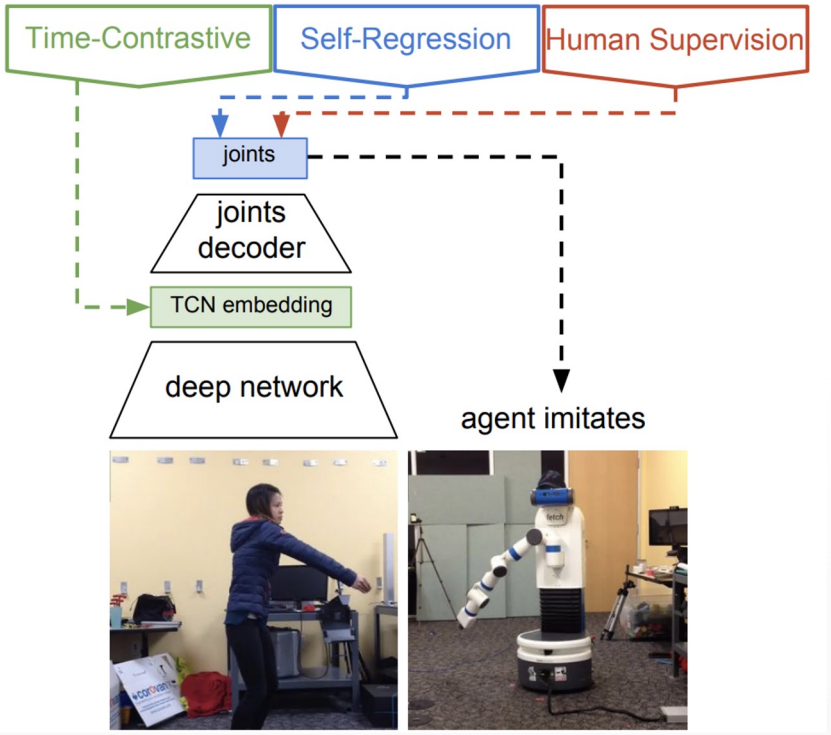
Method	hand contact with container container	within pouring distance	container angle	liquid is flowing	recipient has liquid
Random	49.9%	48.9%	74.5%	49.2%	48.4%
Imagenet Inception	47.4%	45.2%	71.8%	48.8%	49.2%
shuffle & learn	17.2%	17.8%	46.3%	25.7%	28.0%
single-view TCN (triplet)	12.6%	14.4%	41.2%	21.6%	31.9%
multi-view TCN (npairs)	8.0%	9.0%	35.9%	24.7%	35.5%
multi-view TCN (triplet)	7.8%	10.0%	34.8%	22.7%	31.5%
multi-view TCN (lifted)	7.8%	9.0%	35.4%	17.9%	27.7%

Direct Human Pose Estimation

- Implicit mapping between human and robot poses
- Uses self regression for learning the mapping



Methodology- Human Pose Model - MultiView TCN



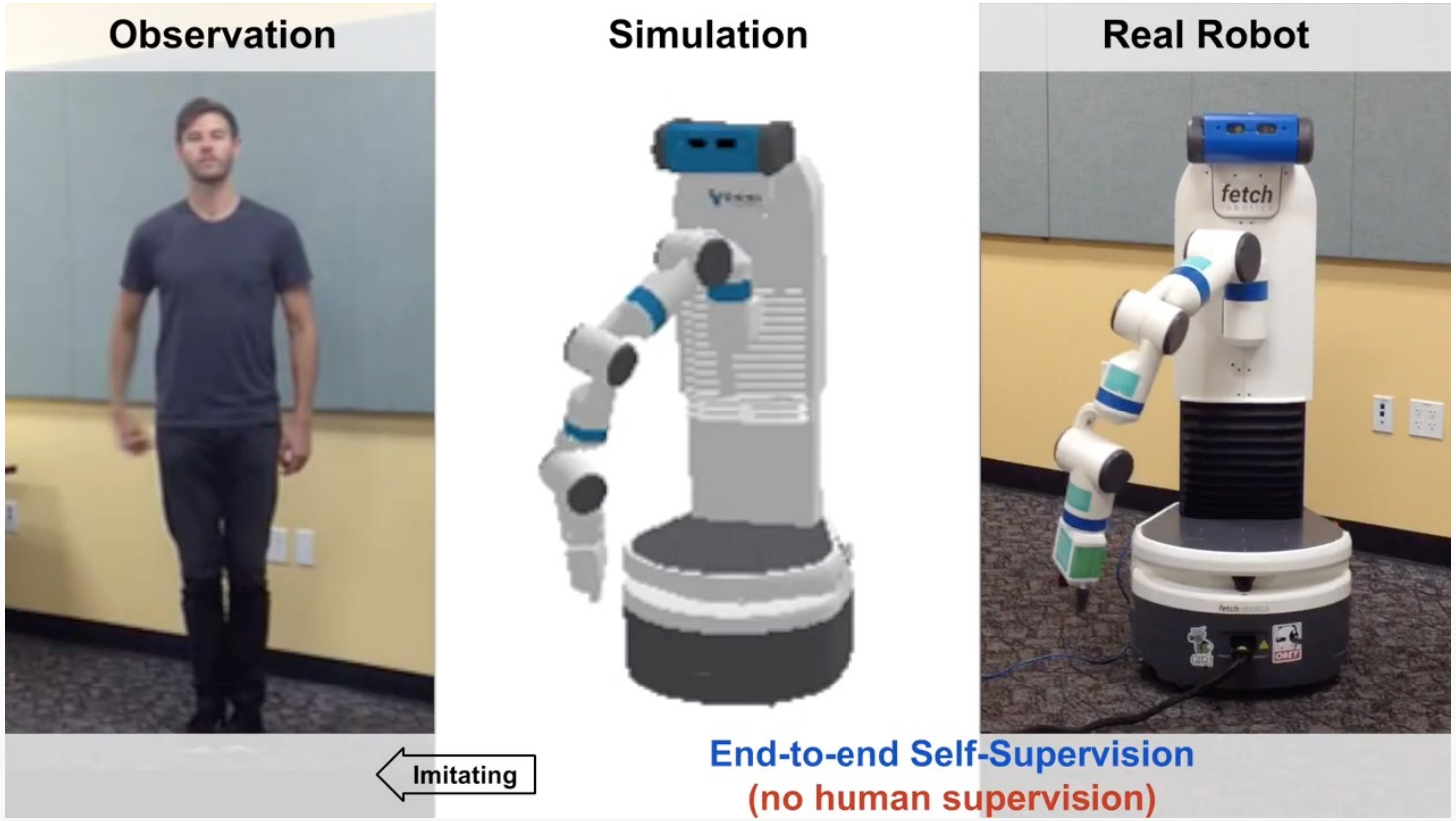
Results

Self-Regression for human pose estimation

Supervision	**Robot joints distance error %**
Random (possible) joints	42.4 ± 0.1
Self	38.8 ± 0.1
Human	33.4 ± 0.4
Human + Self	33.0 ± 0.5
TC + Self	32.1 ± 0.3
TC + Human	29.7 ± 0.1
TC + Human + Self	$\mathbf{29.5} \pm 0.2$

TABLE III: **Imitation error for different combinations of supervision signals.** The error reported is the joints distance between prediction and groundtruth. Note perfect imitation is not possible.

Results


Observation
Simulation
Real Robot
fetch
Imitating
End-to-end Self-Supervision
(no human supervision)

Discussion

Discussion #1

- @95_f1:
- Self-regression seems similar to the way in which humans can relate their actions to perception.
- Connection to Affordance Prediction?

Discussion #2

- Limitation : TCN Embedding is task specific.

- What kind of changes would we need to make this TCN embedding task agnostic? Or extend it for multiple tasks?

Discussion #3

- Why is there a disparity between the performance of single view and multiview TCN?

- @95_f6:

  Animals live entirely in first-person, yet are still able to handle this mapping without ever being provided with this type of data. I wonder what the difference or key factors are here that allow us to perform that mapping (at least to a point, after enough development)

Thank You!