Learning Visual Representations

KHIMYA KHETARPAL

Traditionally

- Deep nets are trained with a LOT of labelled data
- Images and Videos

Alternatively

- Unsupervised learning with signal supervision
- Train on an auxiliary task

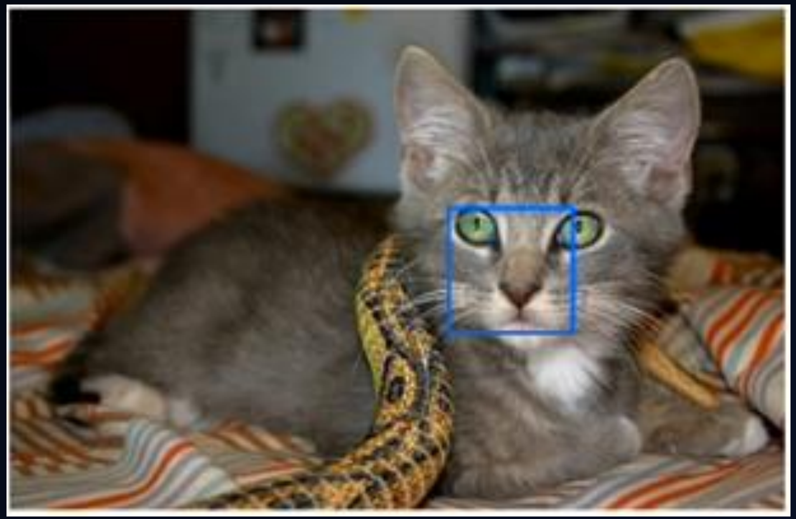
Unsupervised Learning – Signal Supervision

- By Context Prediction – Single Images[1]

[1] Doersch et al., 2016

Unsupervised Learning – Signal Supervision

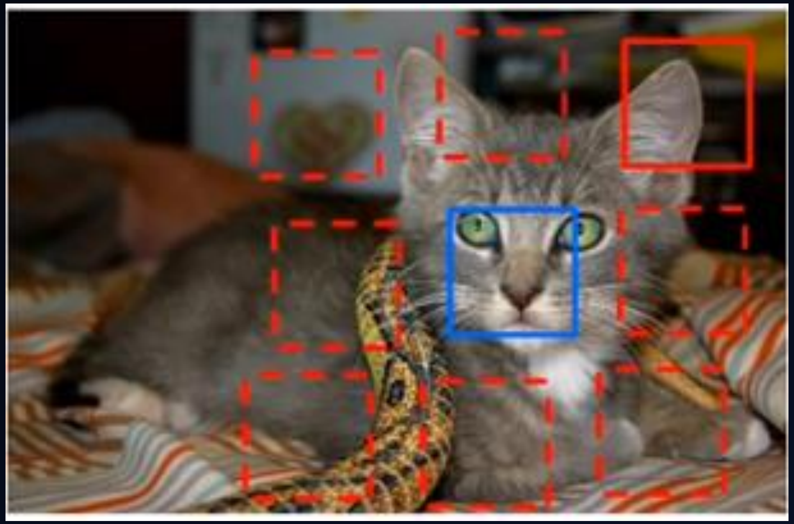
- By Context Prediction – Single Images[1]



[1] Doersch et al., 2016

Unsupervised Learning – Signal Supervision

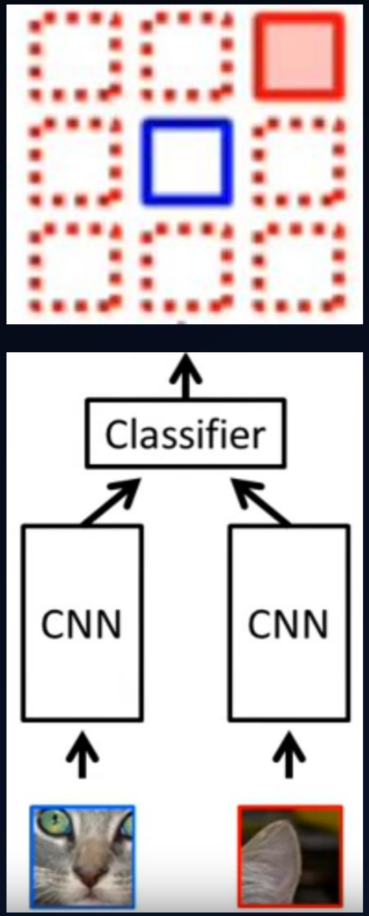
- By Context Prediction – Single Images[1]

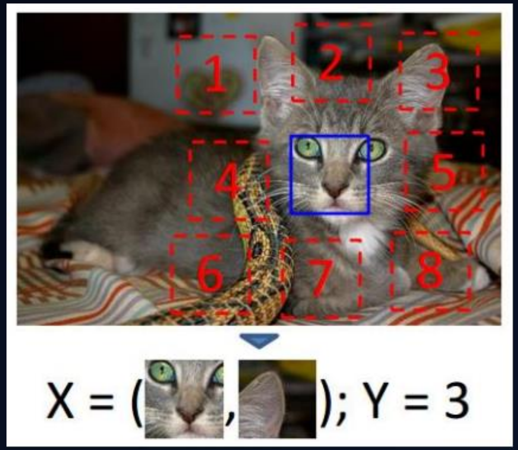


[1] Doersch et al., 2016

Unsupervised Learning – Signal Supervision

- By Context Prediction – Single Images[1]

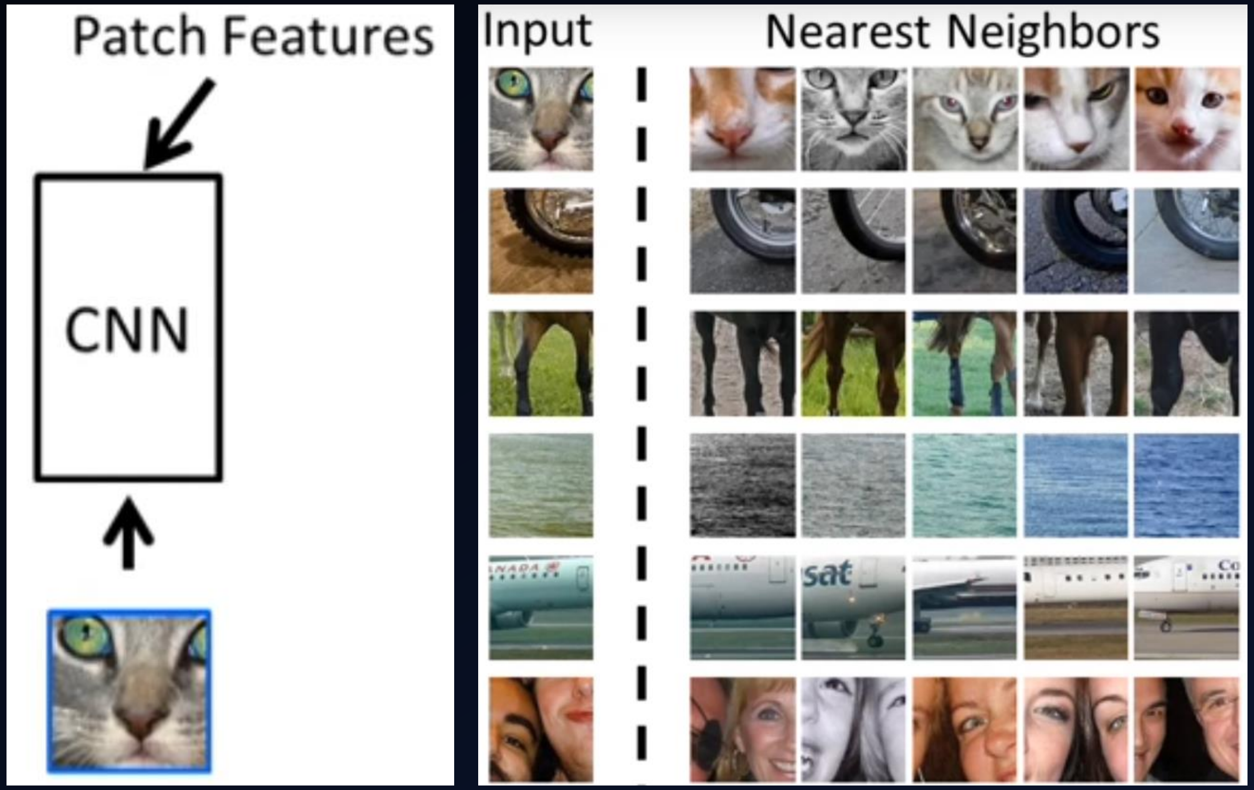




[1] Doersch et al., 2016

Unsupervised Learning – Signal Supervision

- By Context Prediction – Single Images[1]



[1] Doersch et al., 2016

Unsupervised Learning – Signal Supervision

- By Sound [2]

[2] Owens et al., 2016

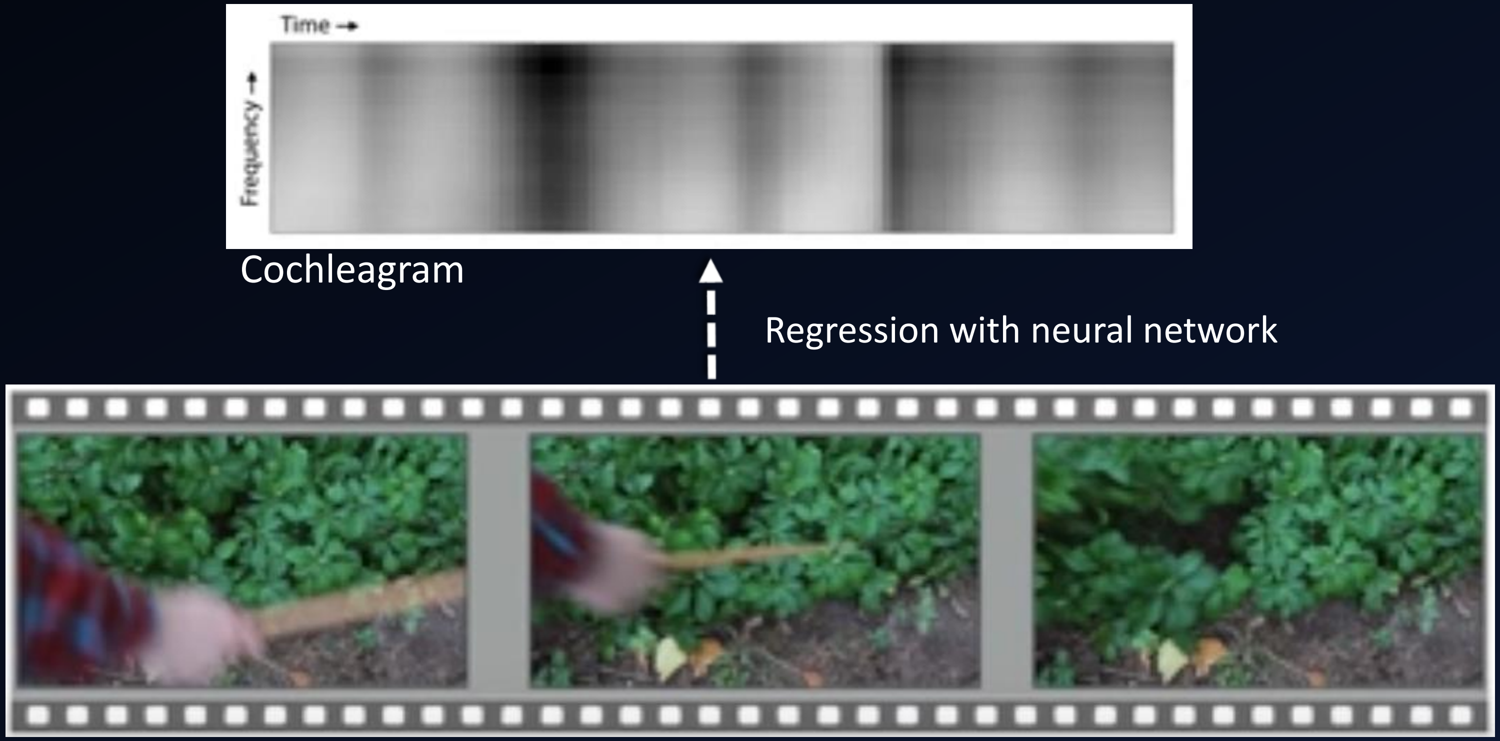
Unsupervised Learning – Signal Supervision

- By Sound [2]

[2] Owens et al., 2016

Unsupervised Learning – Signal Supervision

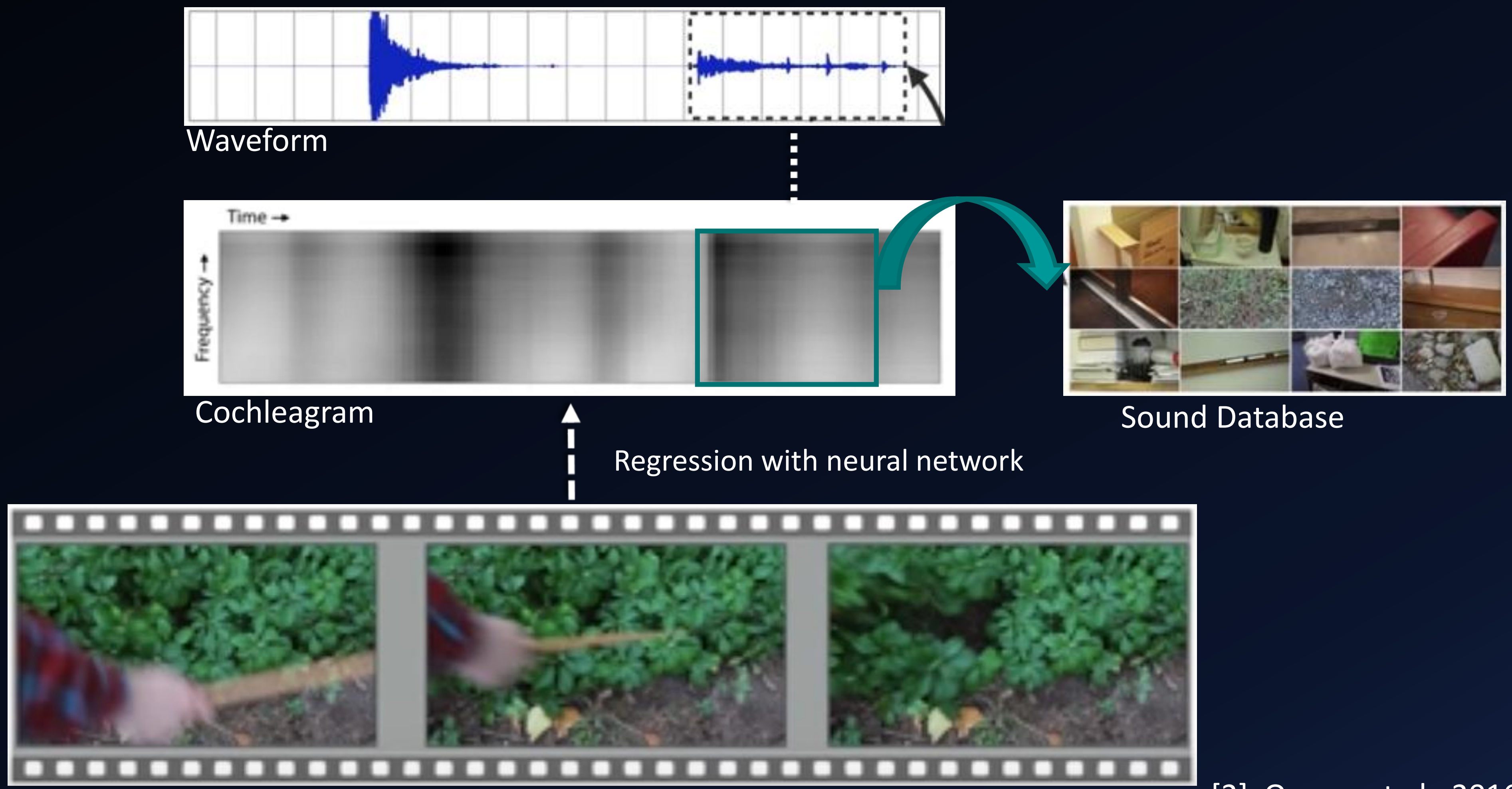
- By Sound [2]



[2] Owens et al., 2016

Unsupervised Learning – Signal Supervision

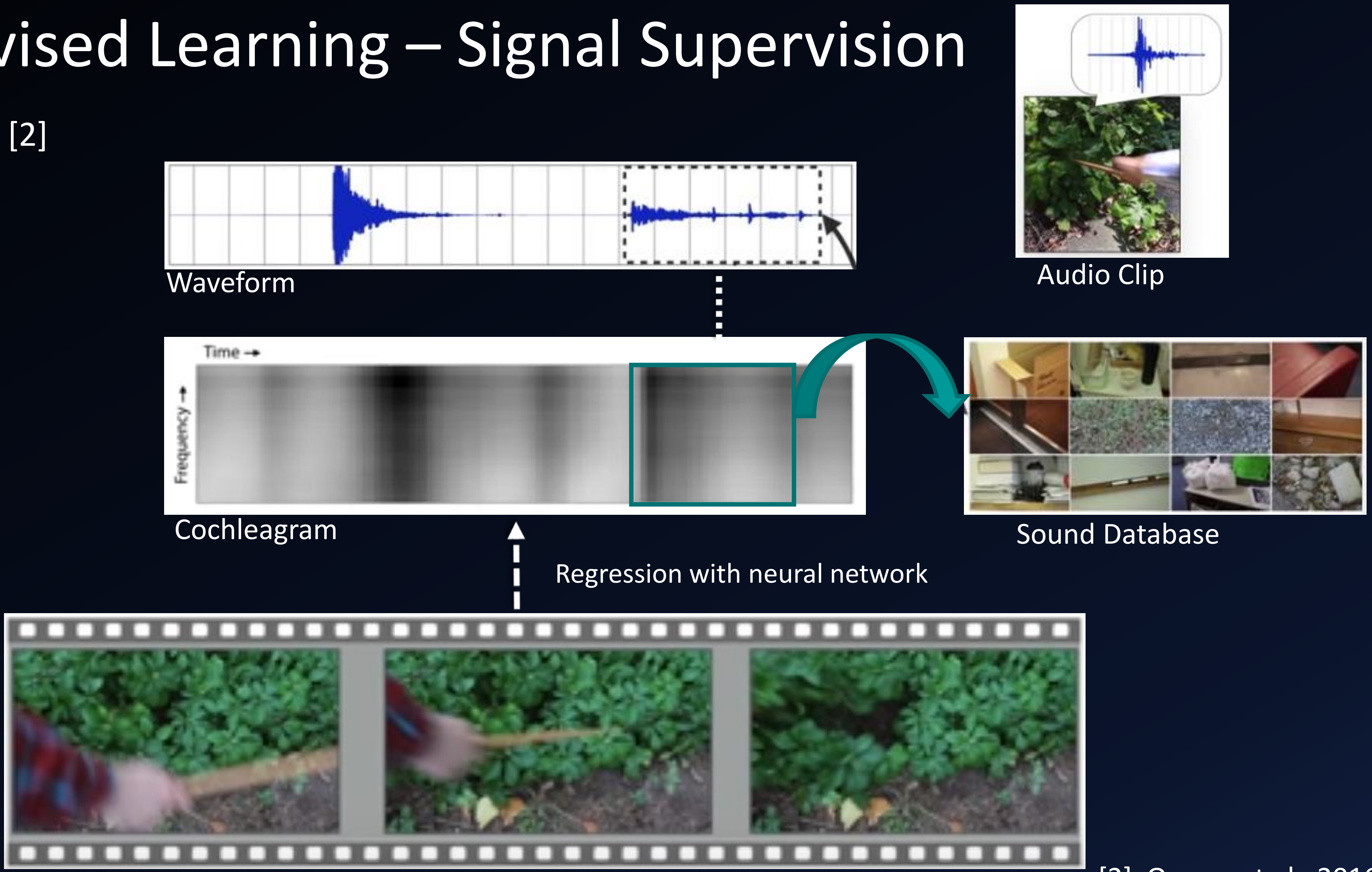
- By Sound [2]



[2] Owens et al., 2016

Unsupervised Learning – Signal Supervision

- By Sound [2]



[2] Owens et al., 2016

Unsupervised Learning – Signal Supervision

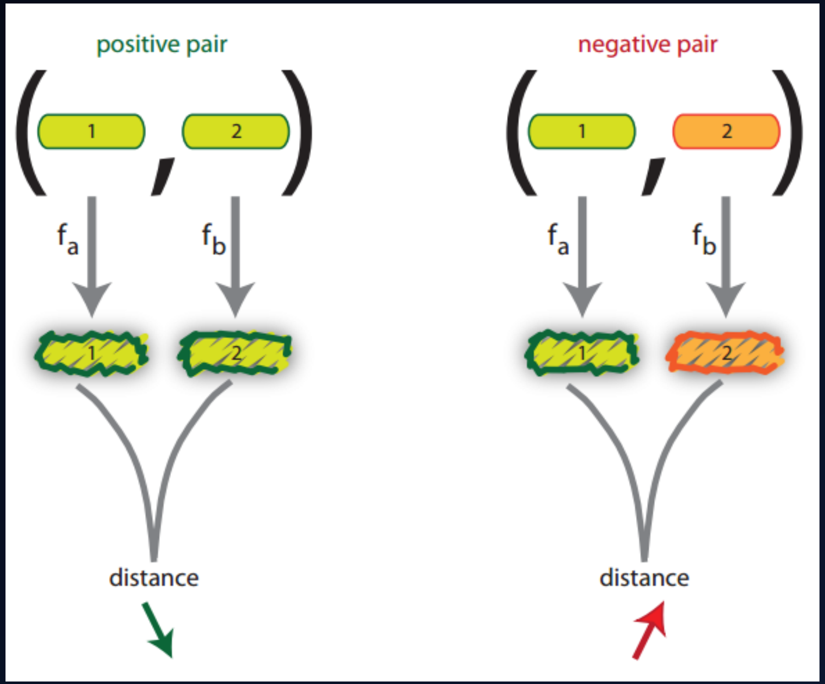
- By Temporal Coherence in Videos[3]
  - Enforcing the representation of two consecutive frames to be close
  - Preserves translations in consecutive frames

[3] Mobahi et al., 2009

[4] Chopra et al., 2005

Unsupervised Learning – Signal Supervision

- By Temporal Coherence in Videos[3]
  - Enforcing the representation of two consecutive frames to be close
  - Preserves translations in consecutive frames
  - Leverage temporal structure in data with *embedding algorithm* [4]

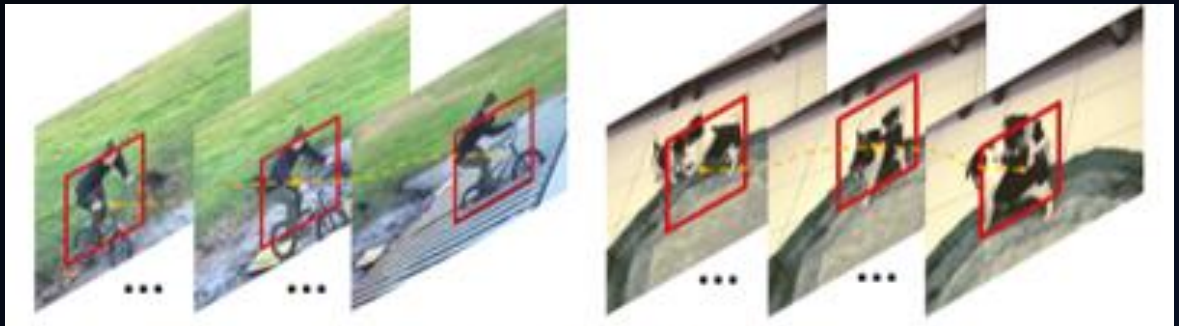


[3] Mobahi et al., 2009

[4] Chopra et al., 2005

Unsupervised Learning – Signal Supervision

- By Tracking Patches in Videos[5]

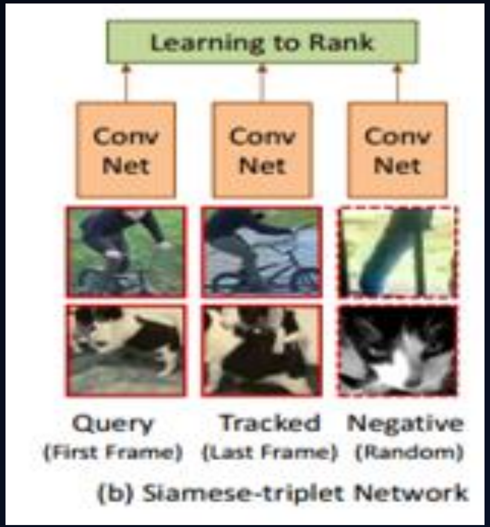


(a) Unsupervised Tracking in Videos

[5] Wang et al., 2015

Unsupervised Learning – Signal Supervision

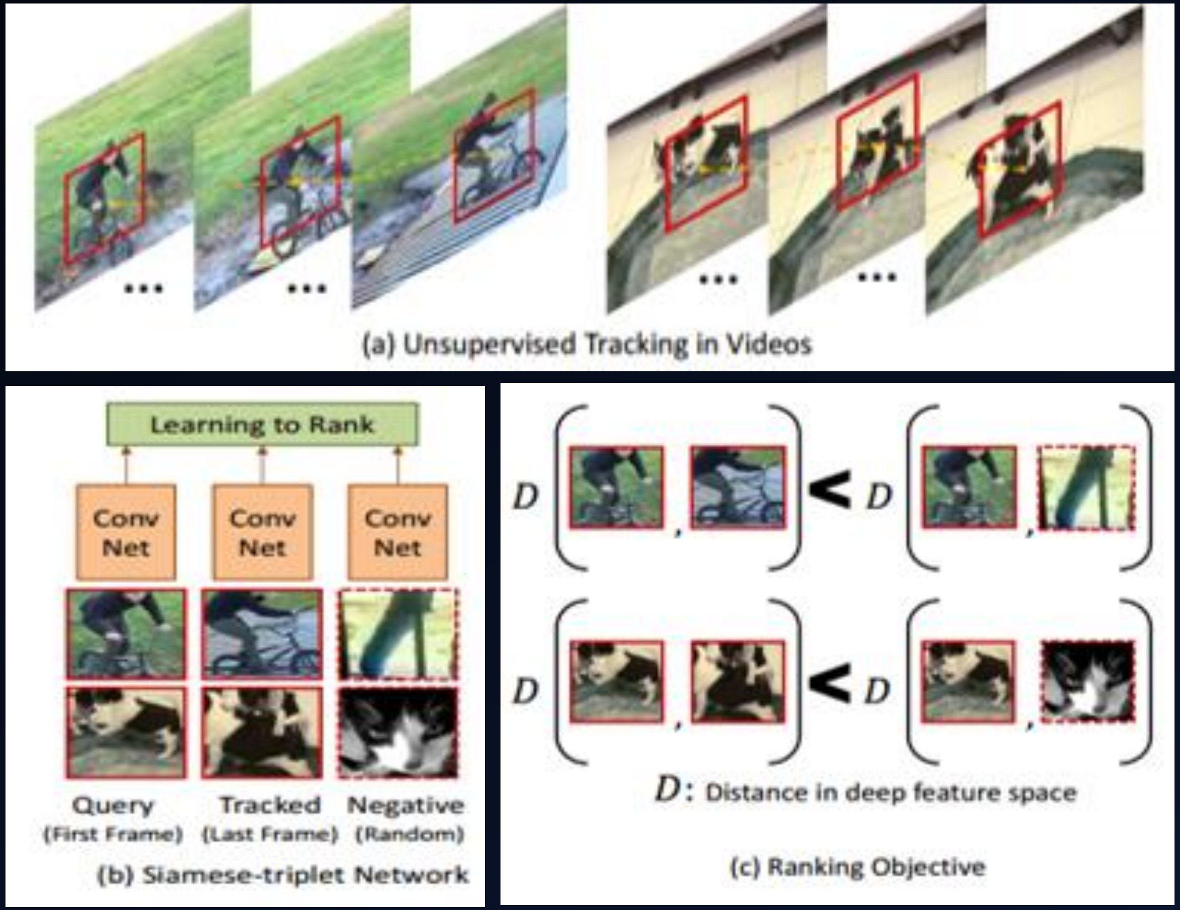
- By Tracking Patches in Videos[5]



[5] Wang et al., 2015

Unsupervised Learning – Signal Supervision

- By Tracking Patches in Videos[5]



[5] Wang et al., 2015

Unsupervised Learning – Signal Supervision

- By Tracking Patches in Videos[5]

[5] Wang et al., 2015

Until Now

- ***Passive*** Observations
- Does not involve any ***active*** observations

Until Now

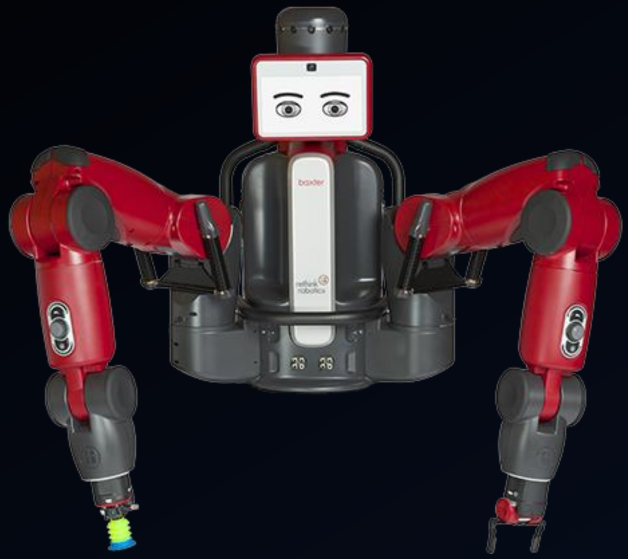
- Passive Observations
- Does not involve any kind of interactions

A new approach[6]

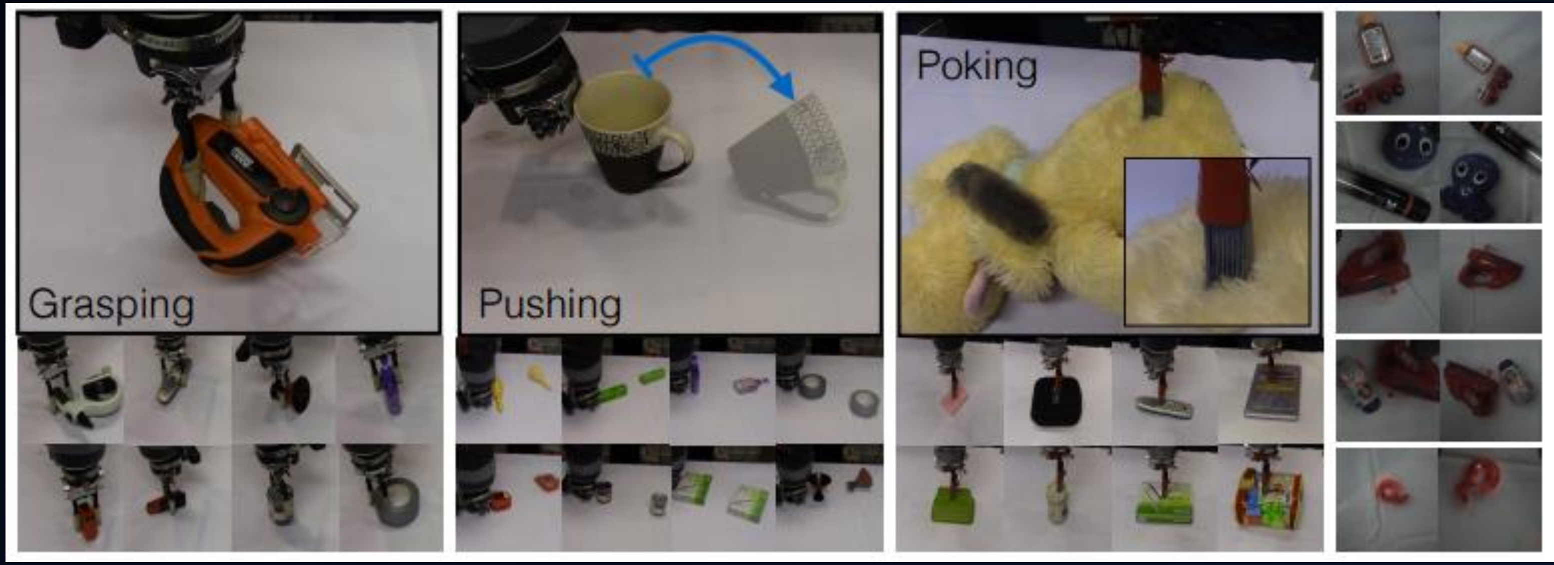
- Learning via ***physical interactions***
- ***Argument:*** Biological agents learn representations by pushing, poking, chewing, etc.

[6] Pinto et al., 2016

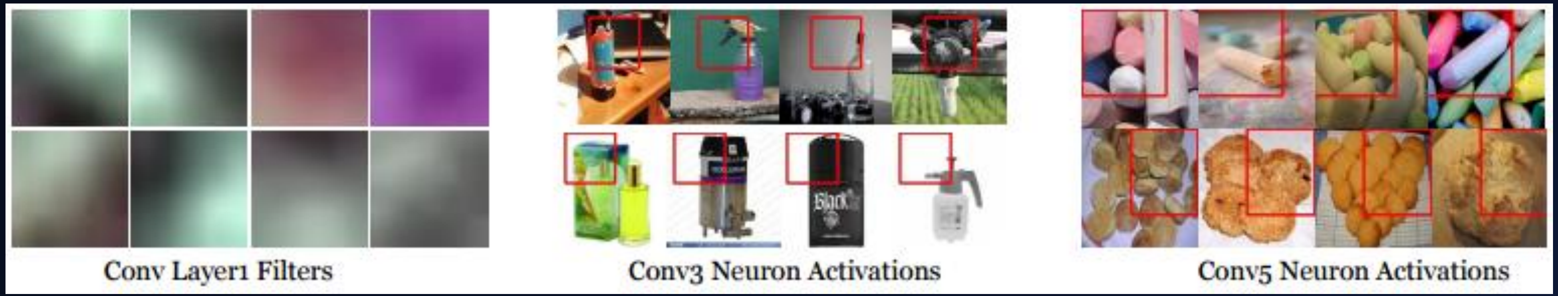
A new approach - The Curious Robot [6]



Baxter Robot

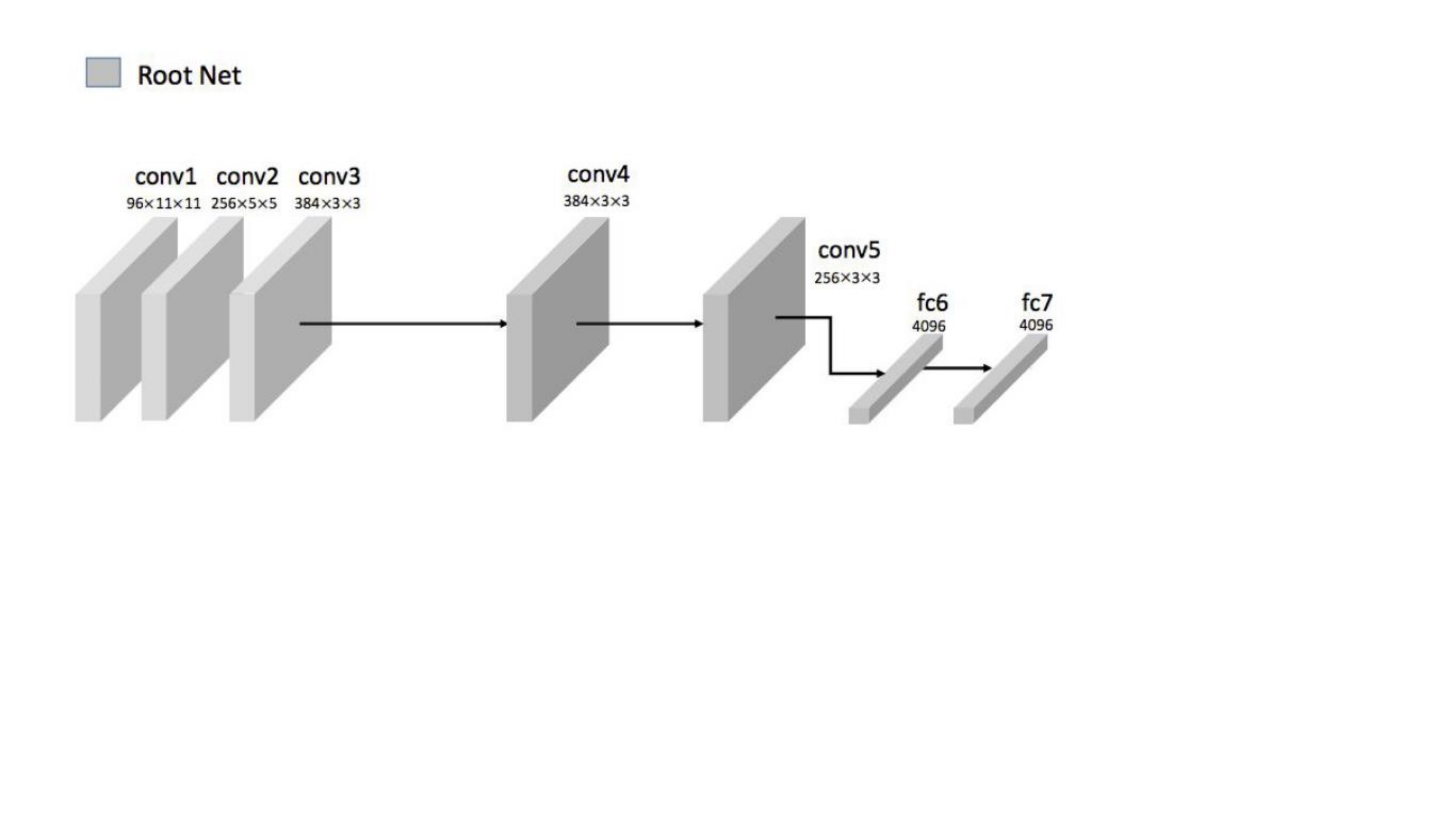


Physical Interactions Data

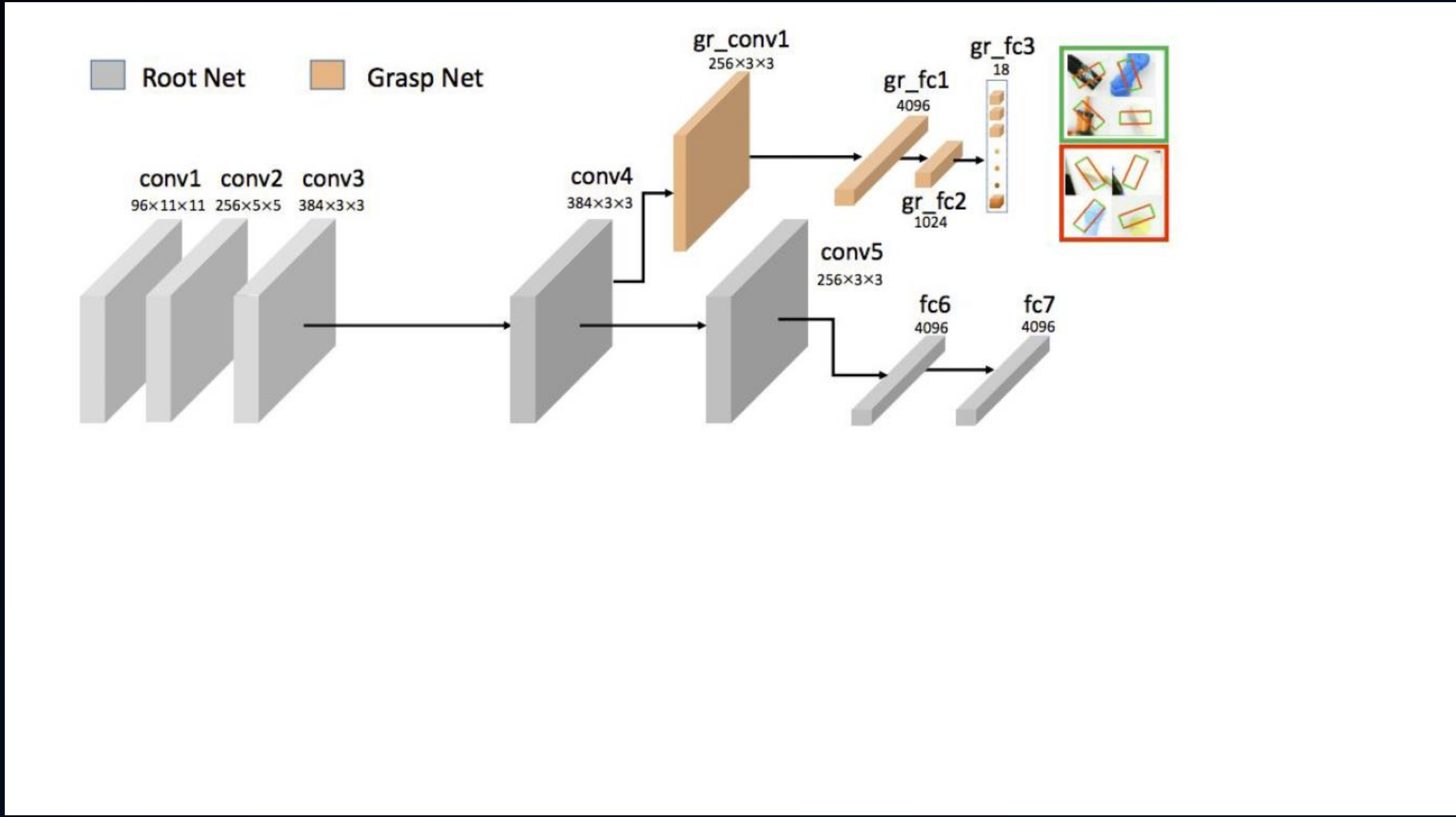


Learned Visual Representations

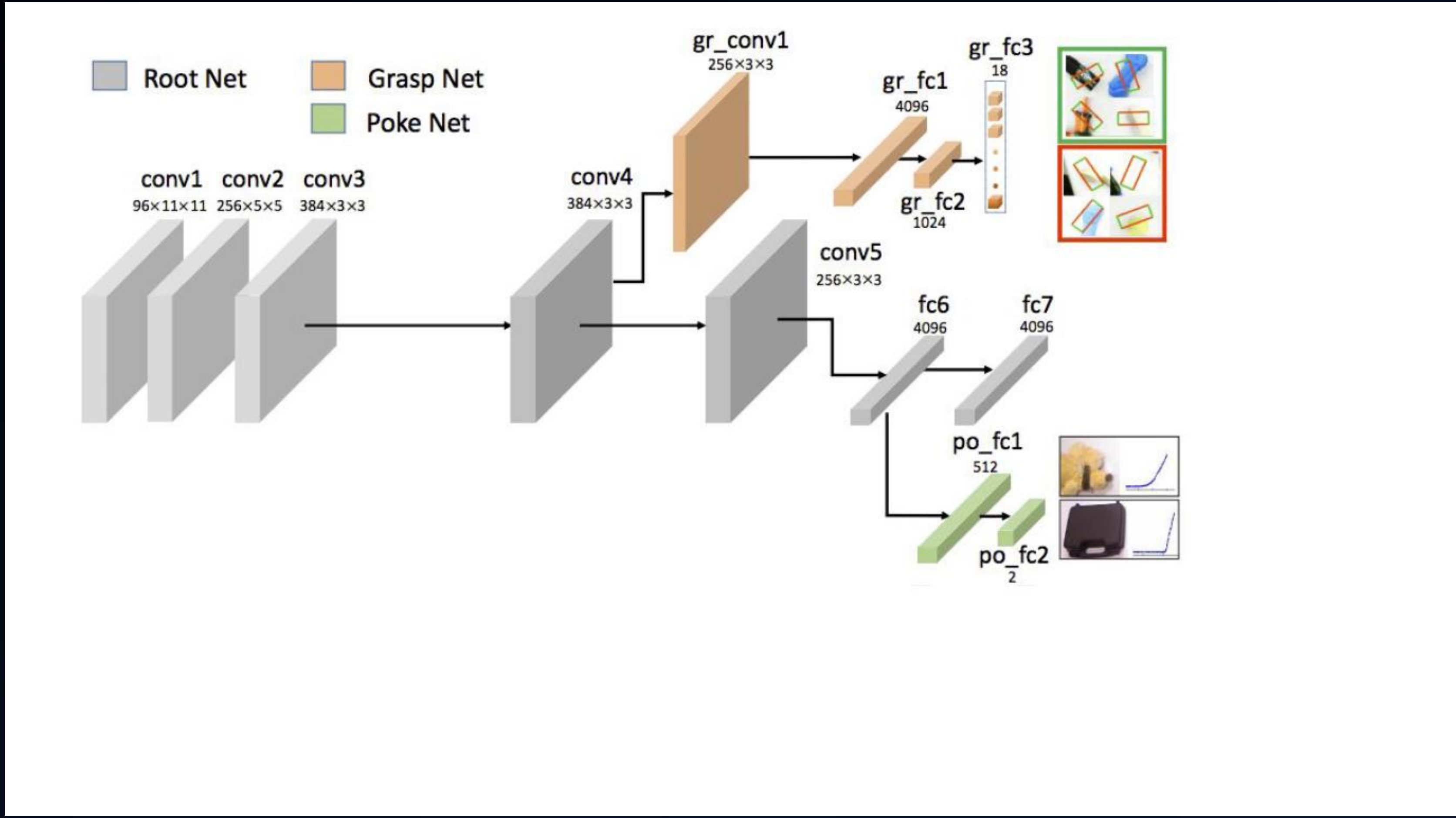
Network Architecture



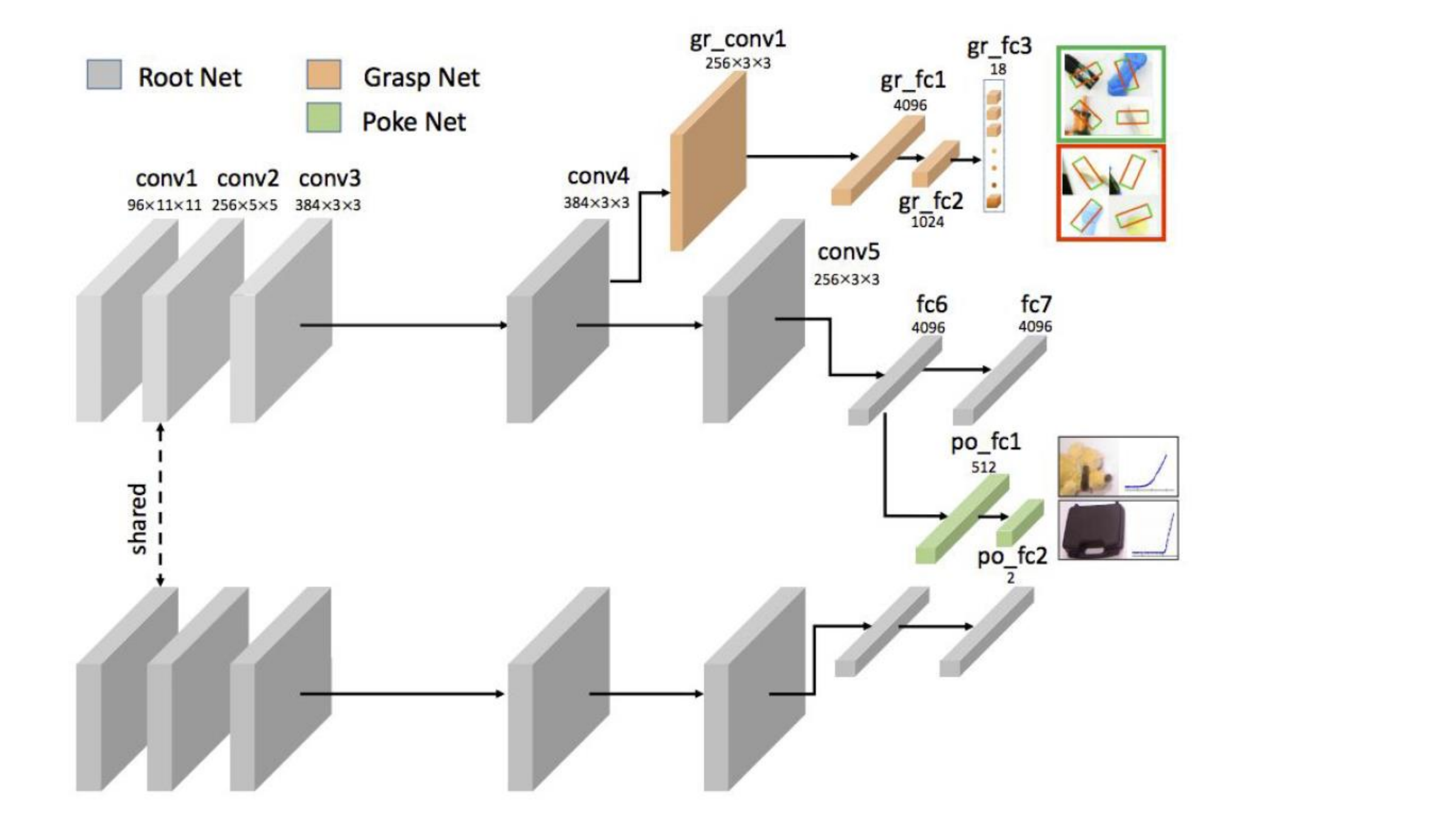
Network Architecture



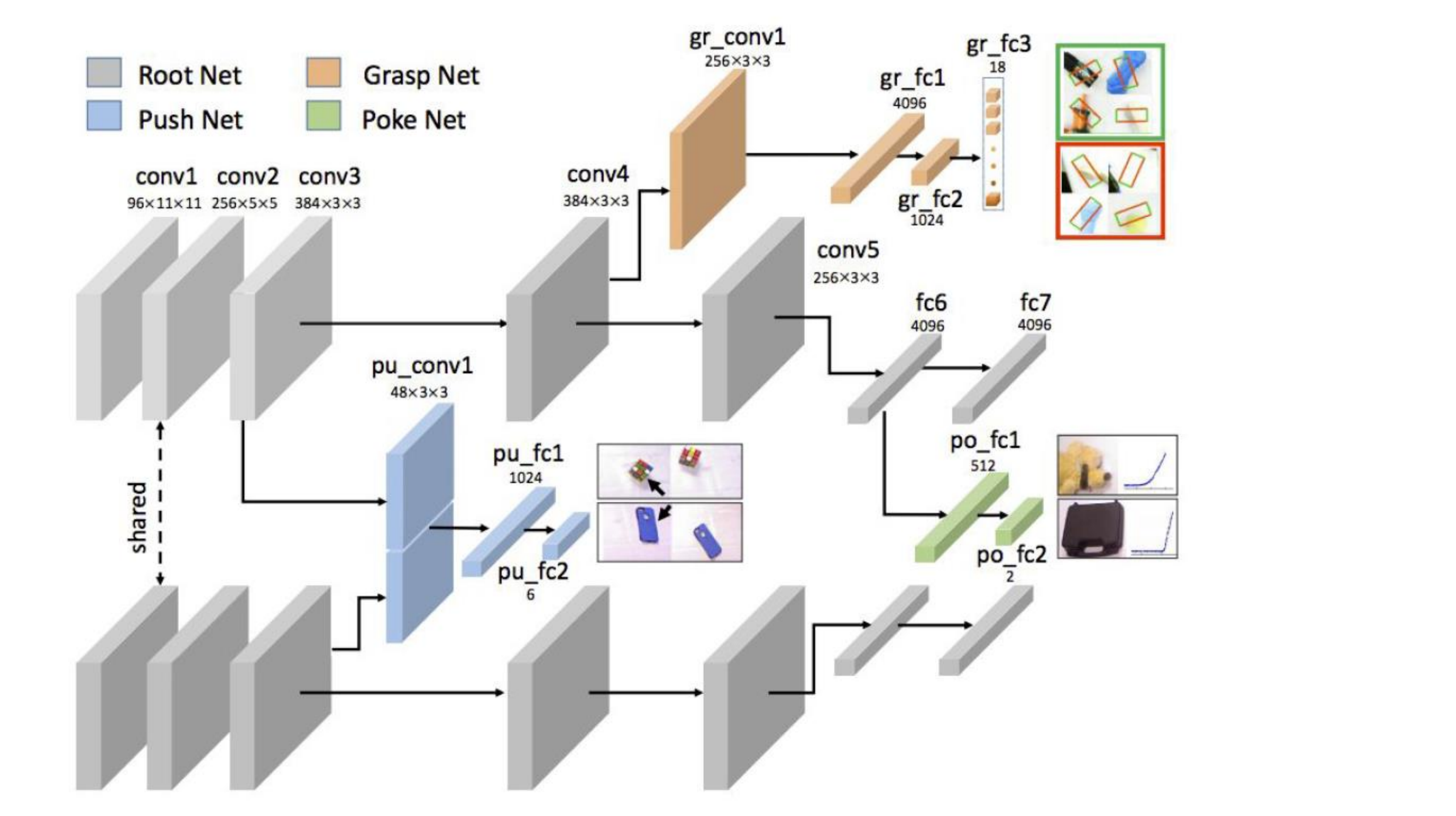
Network Architecture



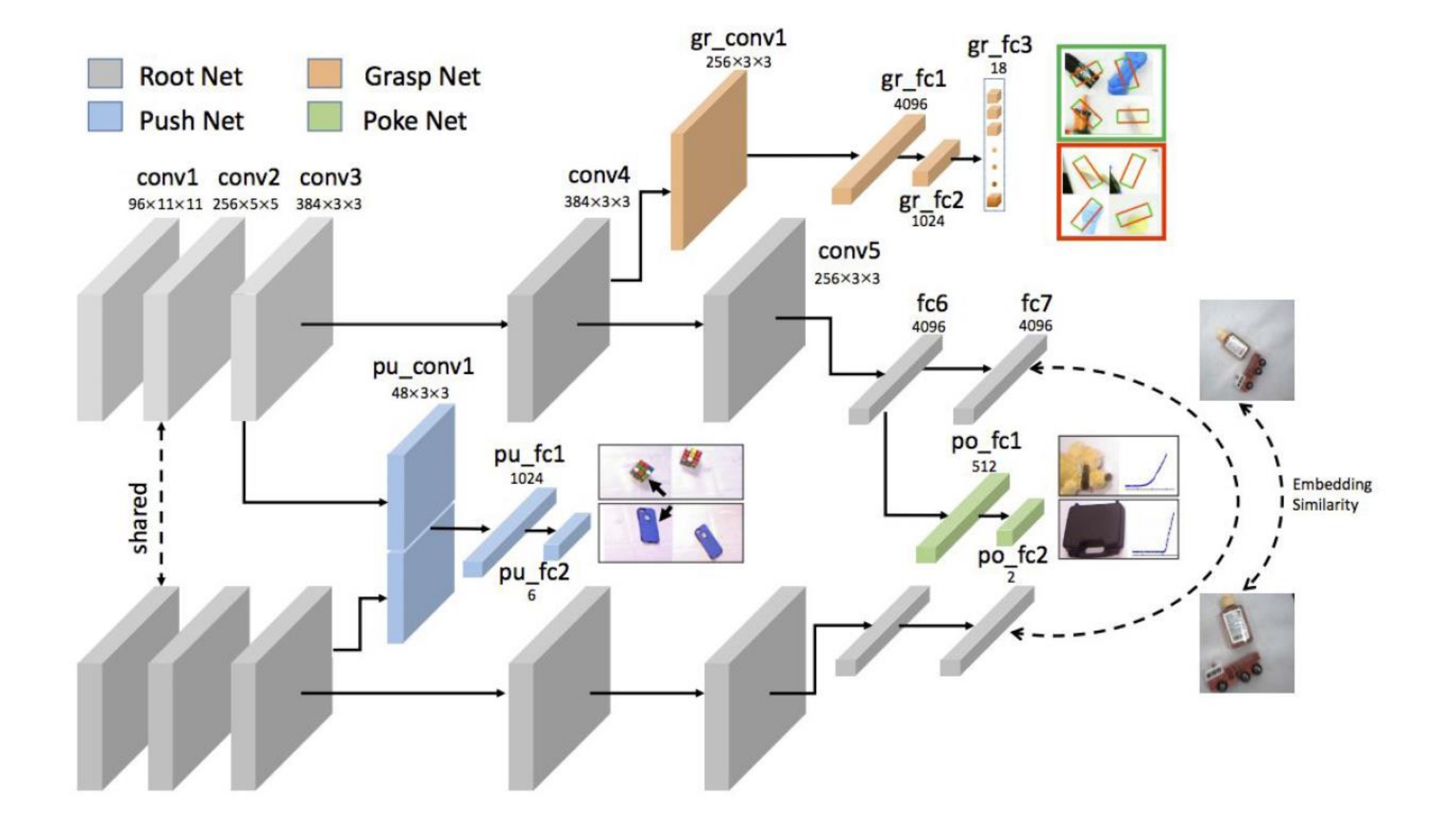
Network Architecture



Network Architecture

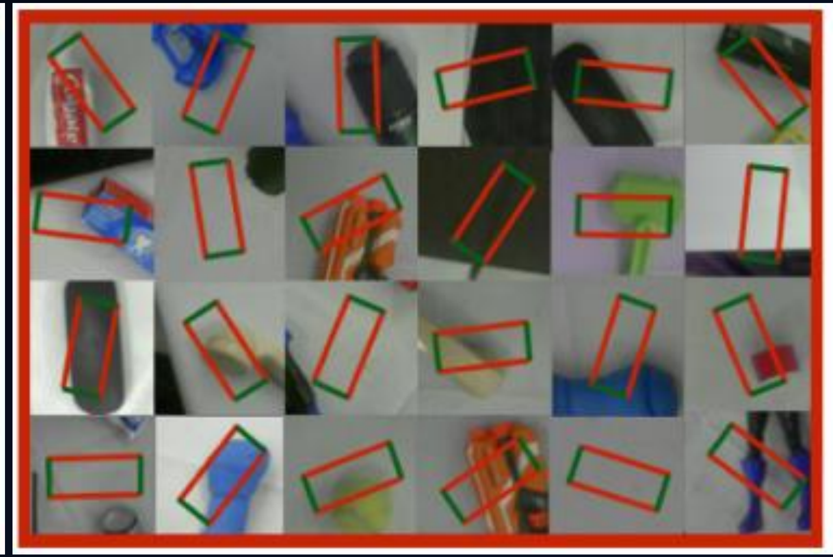


Network Architecture



Planar Grasps

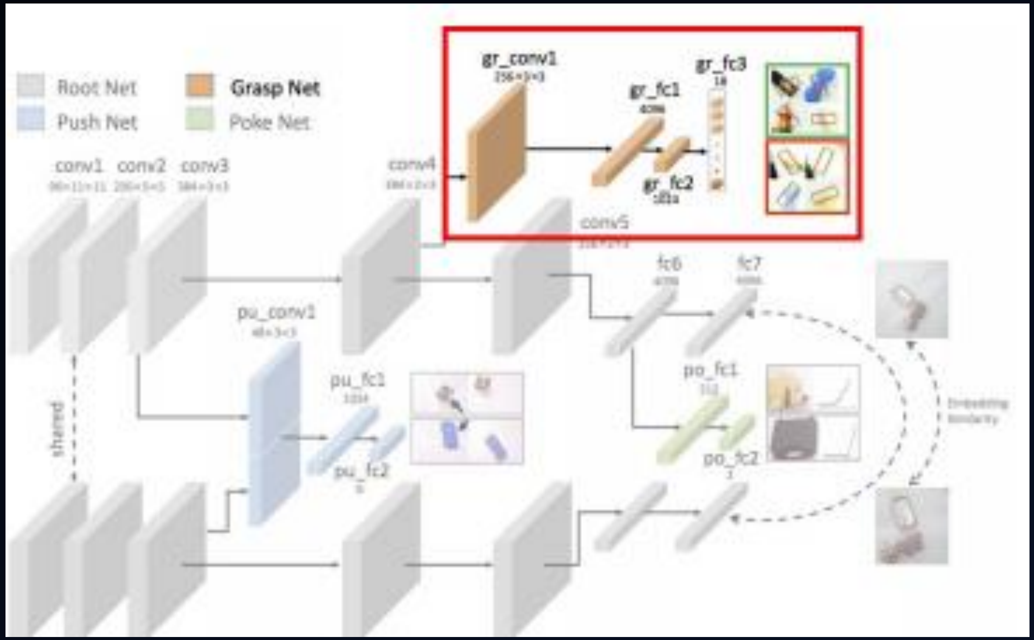
Successful Grasps



Unsuccessful Grasps

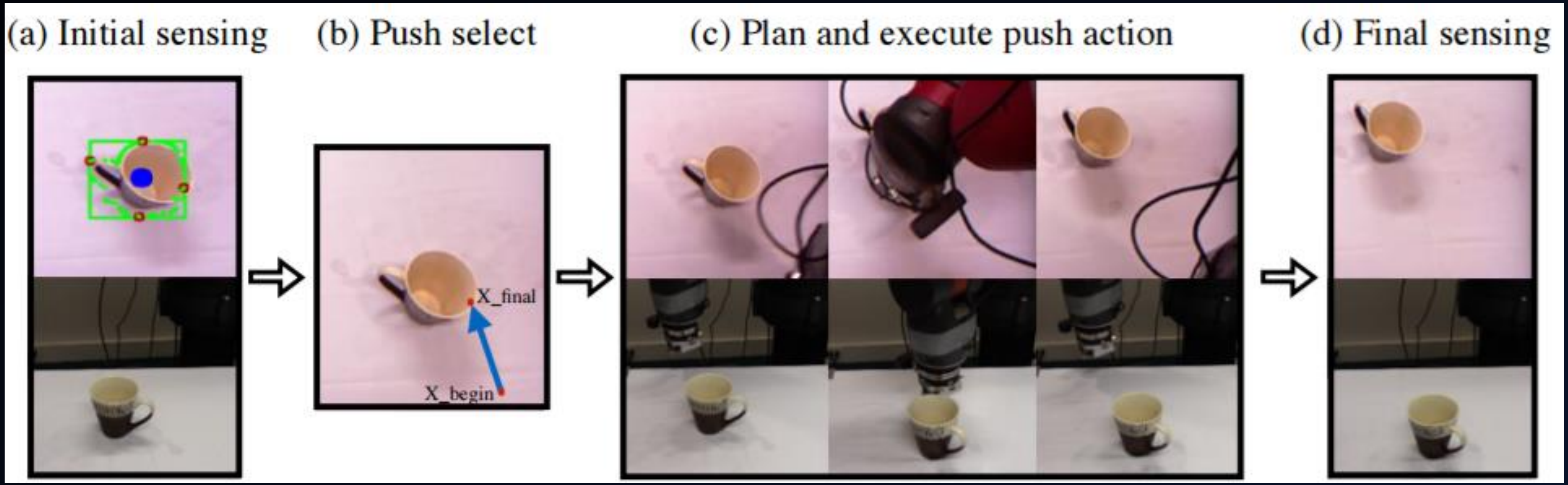
- Training
  - ~ 37K failed grasp interactions
  - ~ 3K successful grasps
- Testing
  - ~ 2.8K failed grasp interactions
  - ~ 0.2K successful grasps

Formulation - Planar Grasps



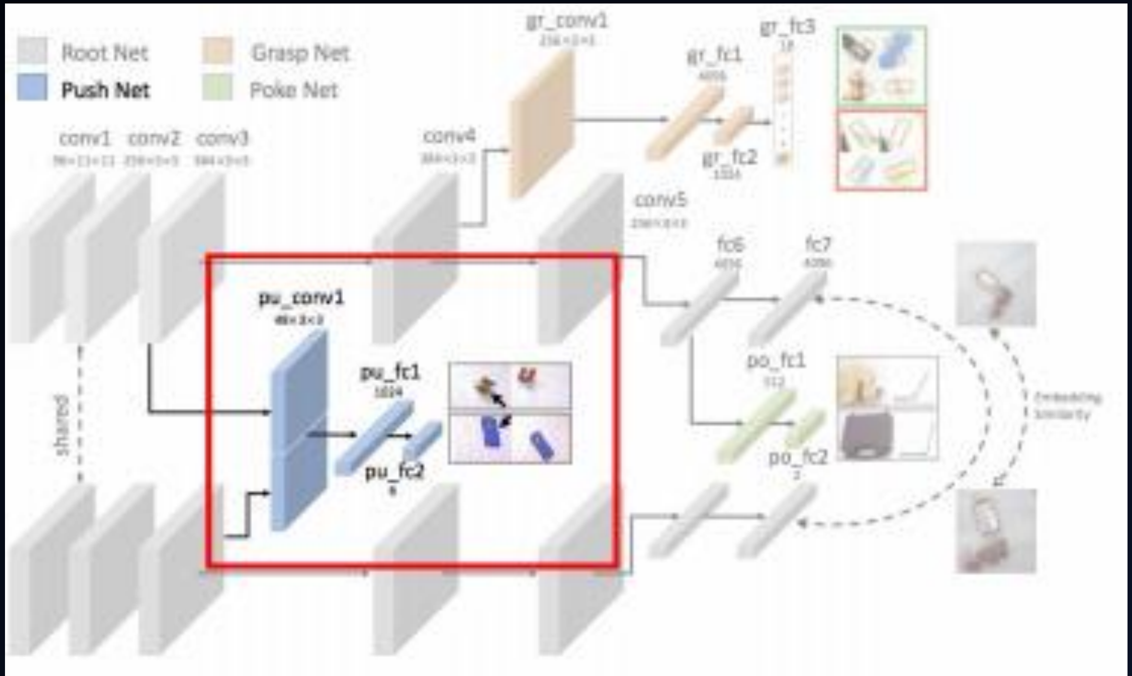
- Predicts whether the center location of the patch is graspable at 0°, 10°, .. 170°
- Loss: 18 way binary classification problems for 10 bins [0, 170]

Planar Push

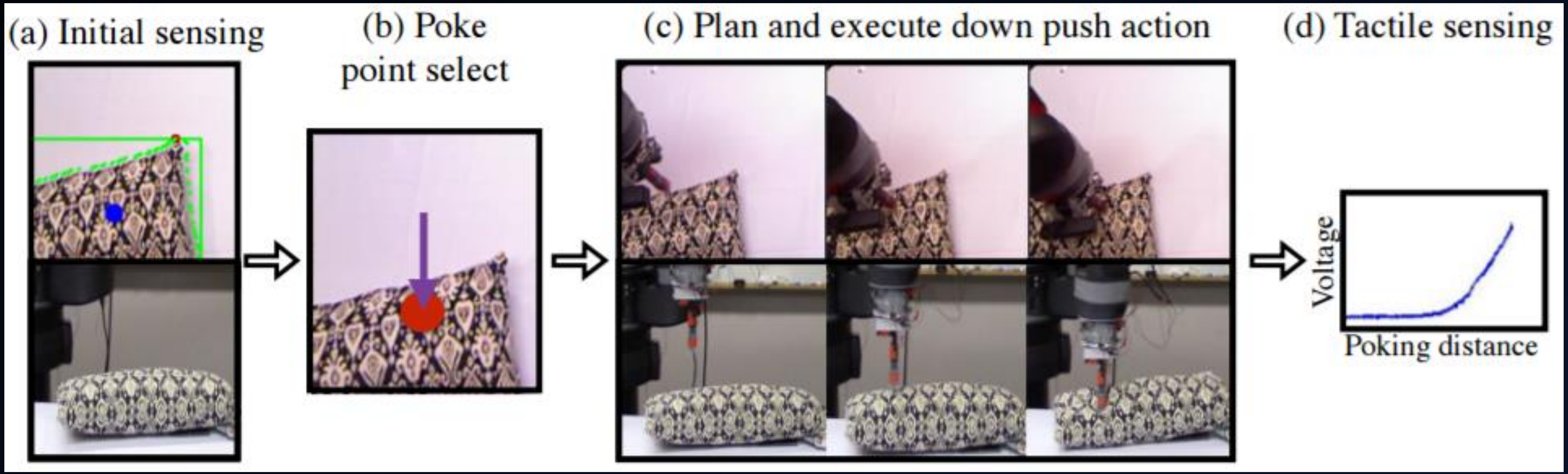


- Given I_{begin} I_{end}, Predict push parameters { X_{begin} X_{final} }
- What action caused this transformation
- Loss: Mean Squared Error (MSE)

Formulation - Planar Push

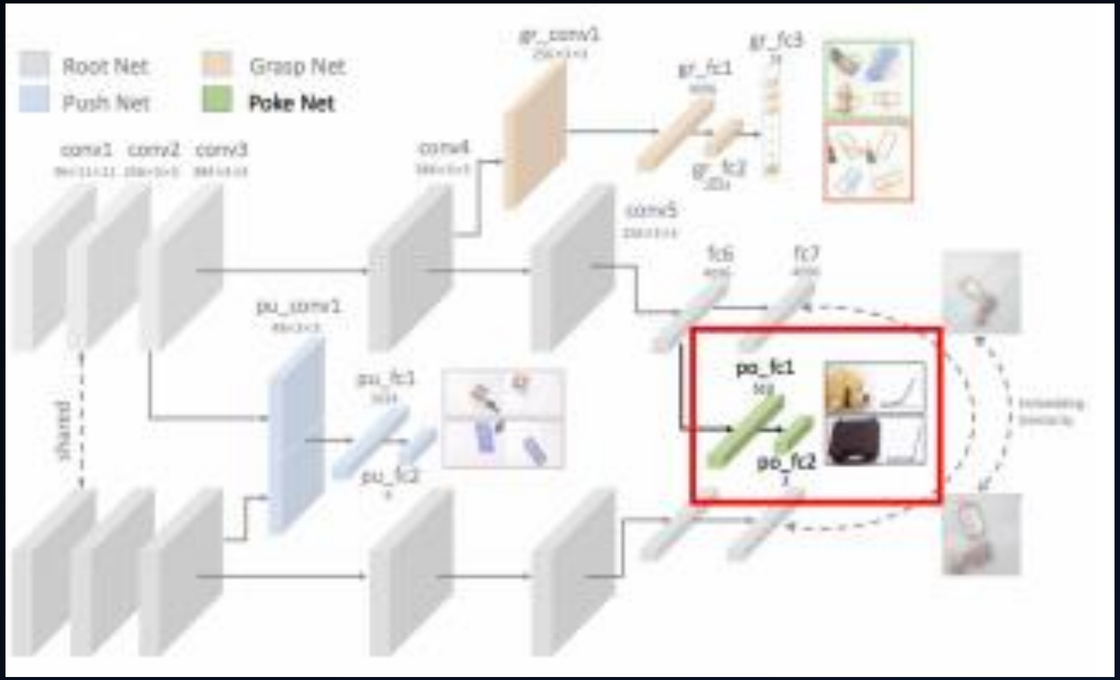


Poke



- Profile of the tactile graph – kind of object material
- Predicts the slope and intercept of the line parametrization of the voltage drop
- Loss: MSE

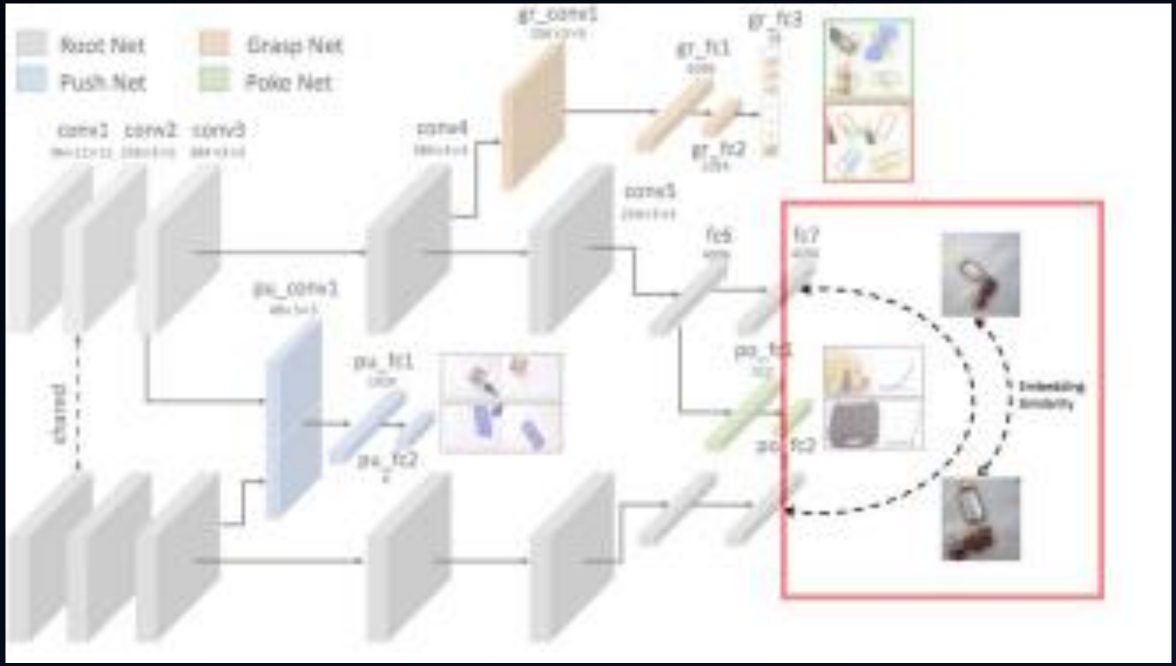
Formulation - Poke



Pose Invariance

- Multiple views of the same object
- Feature Embedding Problem

Formulation – Pose Invariance



- Enforces embedding similarity between images of the same object
- Loss: Cosine similarity

Training

- Stage I
  - Initializes the root network (up to conv4) with Gaussian
  - Train (20k iterations) on the grasp task only

- Stage II
  - Batch size of 128 per task is sequentially fed into the network
  - Weight update at respective back-prop cycles
  - Gradients are accumulated until one batch finishes
  - Mean aggregation of gradient -> Weight update step

Results

Maximally-Activating Images

conv5 conv4

Results

Nearest Neighbor

Results

Image Classification

	Household	UW RGB-D	Caltech-256
Root network with random init.	0.250	0.468	0.242
Root network trained on robot tasks [4]	0.354	0.693	0.317
AlexNet trained on ImageNet	0.625	0.820	0.656
Root network trained on identity data	0.315	0.660	0.252
Auto-encoder trained on all robot data	0.296	0.657	0.280

Task Ablation Analysis

Image Classification

	Household	UW RGB-D	Caltech-256
All robot tasks	0.354	0.693	0.317
Except Grasp	0.309	0.632	0.263
Except Push	0.356	0.710	0.279
Except Poke	0.342	0.684	0.289
Except Identity	0.324	0.711	0.297

Takeaways

- ✓ Active interaction with the world helps learn visual representations
- ✓ Ability to generalize to other tasks such as classification, retrieval

Takeaways

- ✓ Active interaction with the world helps learn visual representations
- ✓ Ability to generalize to other tasks such as classification, retrieval

- ➢ Heavily hand crafted architecture
- ➢ Dependency on the set up for tasks
- ➢ Task ablation analysis shows most dependency on grasp task for learning visual representations

Further Research – Brainstorming

Further Research – Brainstorming

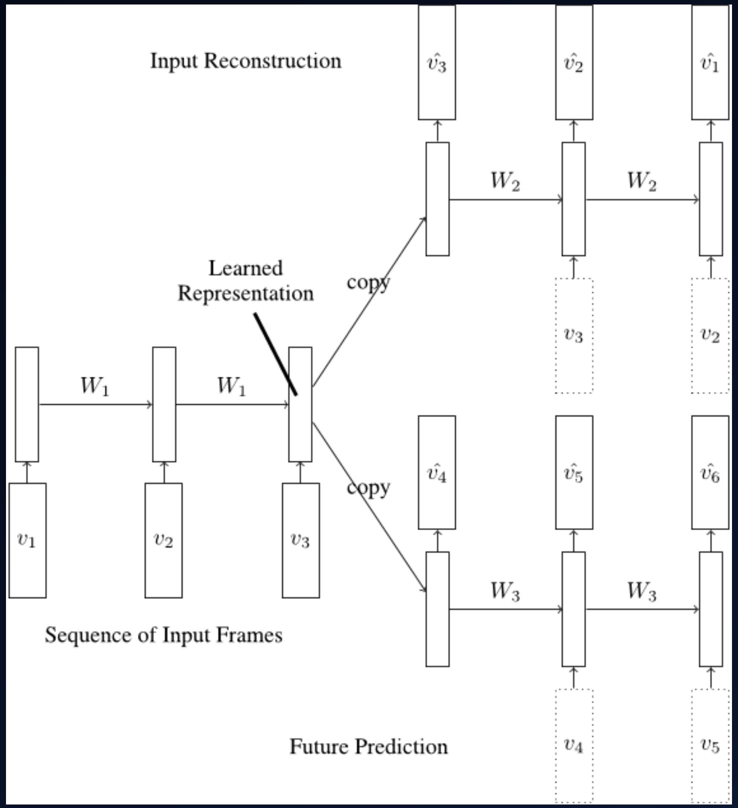
- *Argument*
  - Good Representation = Spatial + temporal
  - Human learning via interactions evolves with prior context from previous interactions

Further Research – Brainstorming

- *Argument*
  - Good Representation = Spatial + temporal
  - Human learning evolves with prior context from previous interactions
- *Explore*
  - Physical interactions as sequence of frames (temporal structure)
  - Consider a task which is a composition of {push, poke, grasp, pick, identify, ..}
  - *Would we be able to learn better visual representations* ?

Further Research – Brainstorming

- Unsupervised learning of video representations [7]
- An encoder LSTM to map an input sequence into a fixed length representation.
- Single or multiple decoder LSTMs to perform different tasks
- Reconstructing the input sequence, or predicting the future sequence



[7] Srivastava et al., 2016

Further Research – Brainstorming

- *Argument*
  - Humans learn from physical AND social interactions
  - A lot of context prior is embedded in human brain via language associations for the perceived world

Further Research – Brainstorming

- *Argument*
  - Humans learn from physical AND social interactions
  - A lot of context prior is embedded in human brain via language associations for the perceived world
- *Explore*
  - Using image captions, movie and it's text transcript, visual dialog, to serve as supervisory signal
  - Language based interactions (social) associated with sequential images (video) could help learn better representations ?

[8] Quattoni et al., 2007

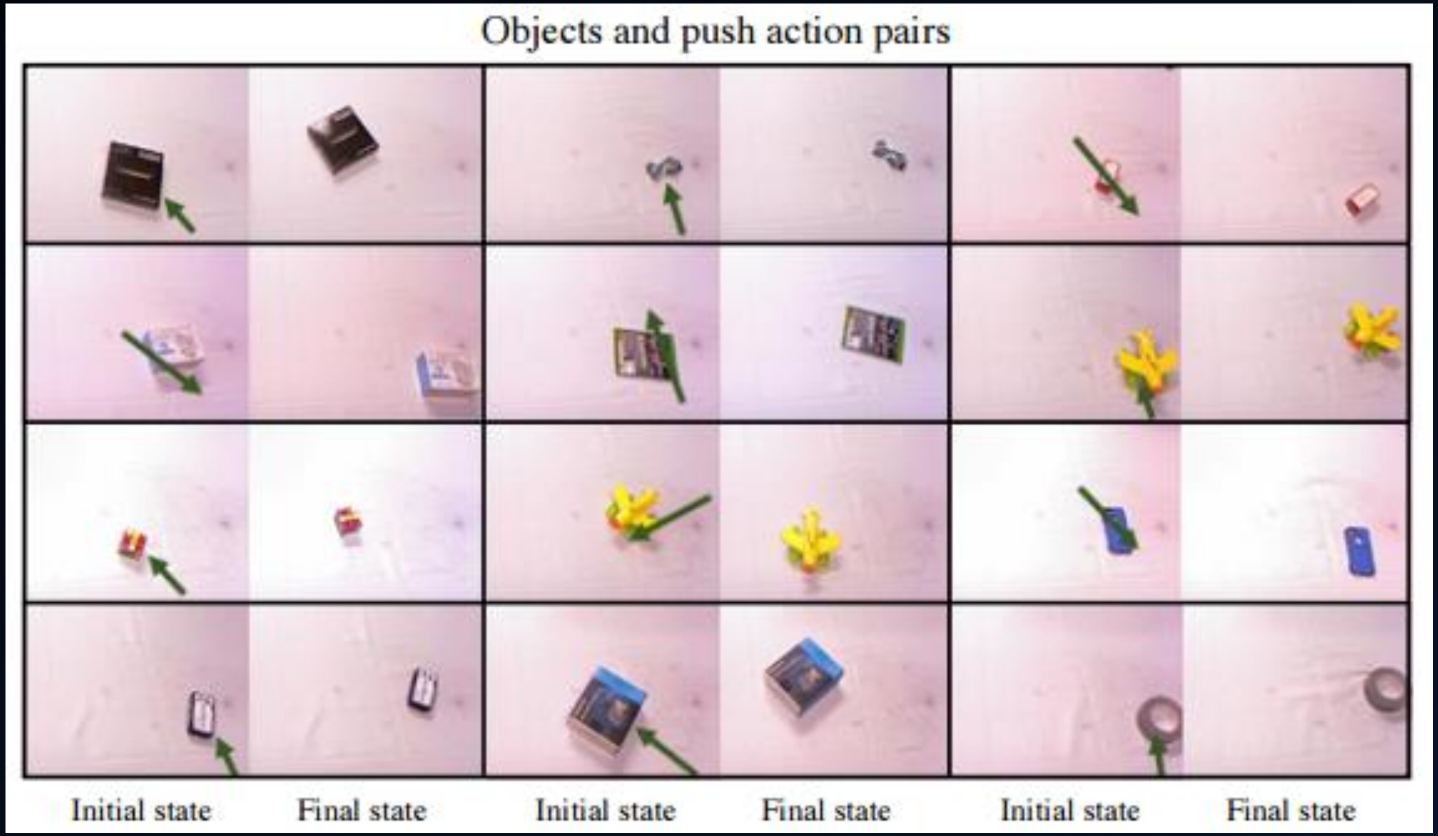
Thank You

Extra Slides

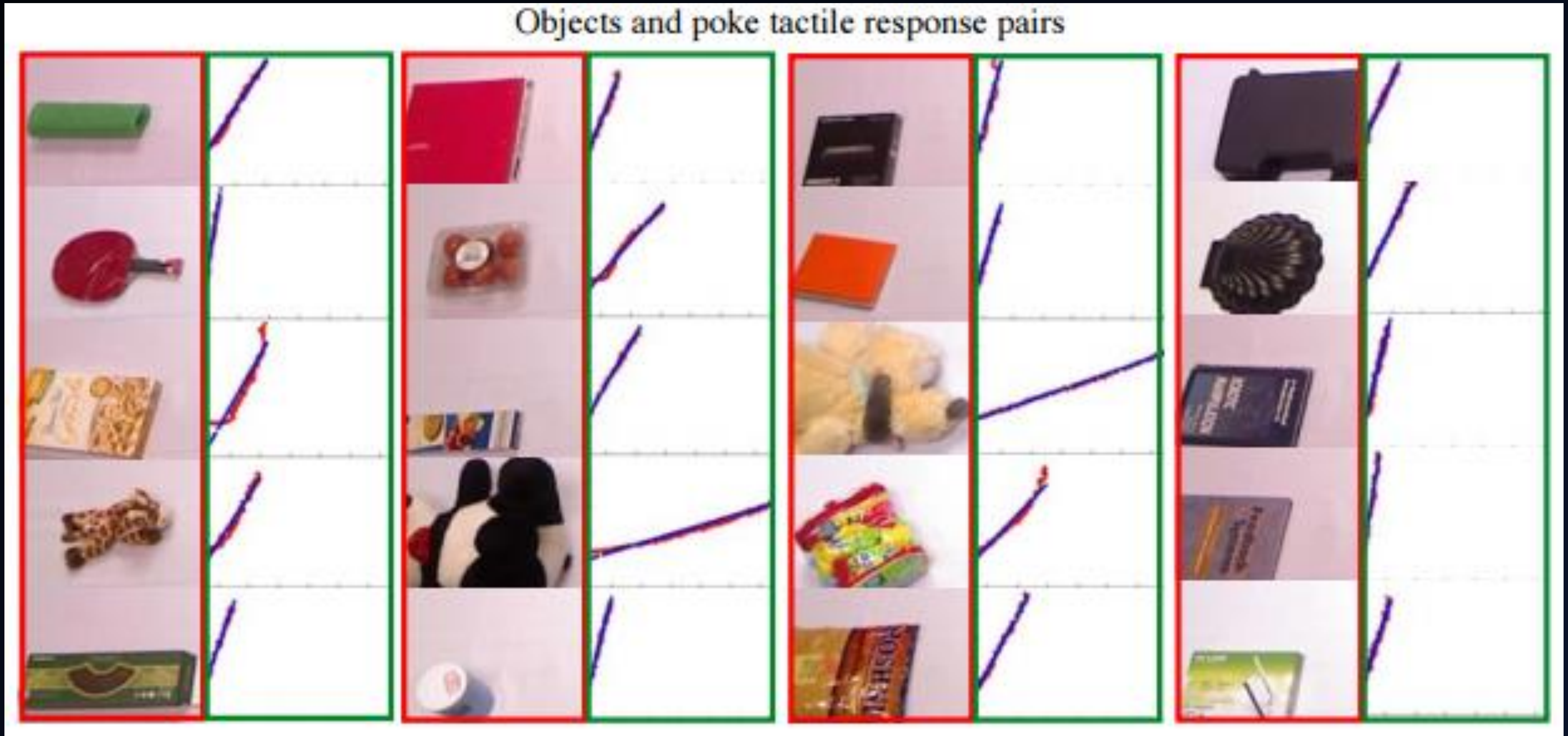
Dataset

Medium	Description	Data points
Grasping	5 images of the object grasped from multiple viewpoints in every successful grasp interaction	40287 grasps
Pushing	2 images of the object pushed in each interactions	5472 pushes/70 objects
Poking	A highly sensitive tactile optical sensors has been used to obtain skin sensor readings	1372 observations on 100 diverse objects
Identity Vision	Multiple images of the same object from different viewpoints	42K +/- image pairs

Push Task - Sample Datapoints



Poke Task - Sample Datapoints

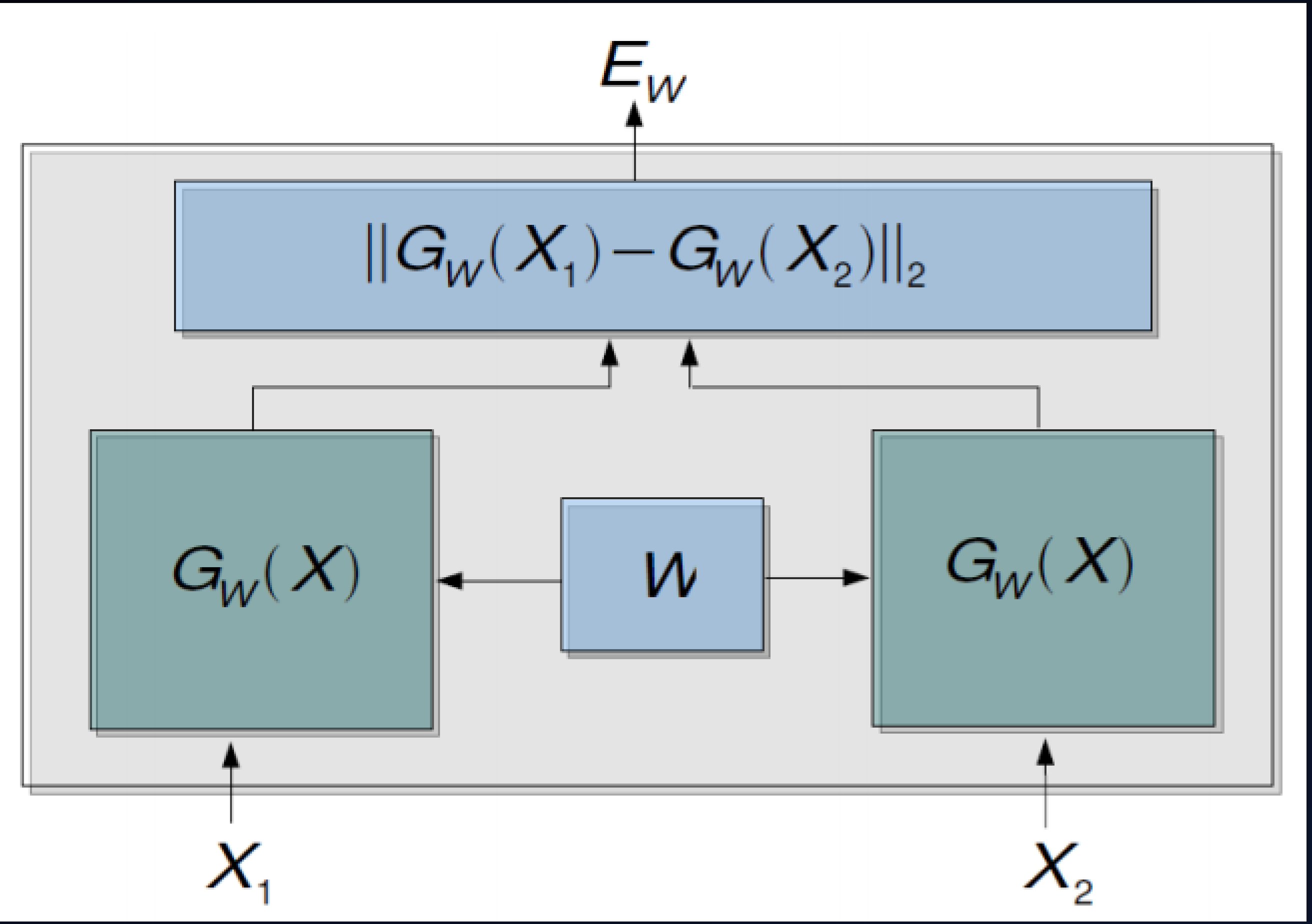


Grasp Task – Loss Formulation

- 18-way binary classifier
- Given a batch size B, with an input image Ii ,
- The label corresponding to angle θ_i defined by $l_i \in \{0, 1\}$
- Forward pass binary activations $\boldsymbol{A}_{ji}$ on the angle bin j the loss $\boldsymbol{L}$ is

$$L = \sum_{i=1}^{B} \sum_{j=1}^{N=18} \delta(j, \theta_i) \cdot \mathrm{softmax}(A_{ji}, l_i)$$

Siamese Architecture



Siamese Architecture