Presentation with Intel

Jiali Duan

Jan. 19, 2022

Email: jialidua@usc.edu/duajiali@amazon.com

Phone: 1(213)2040380

Interested Domains

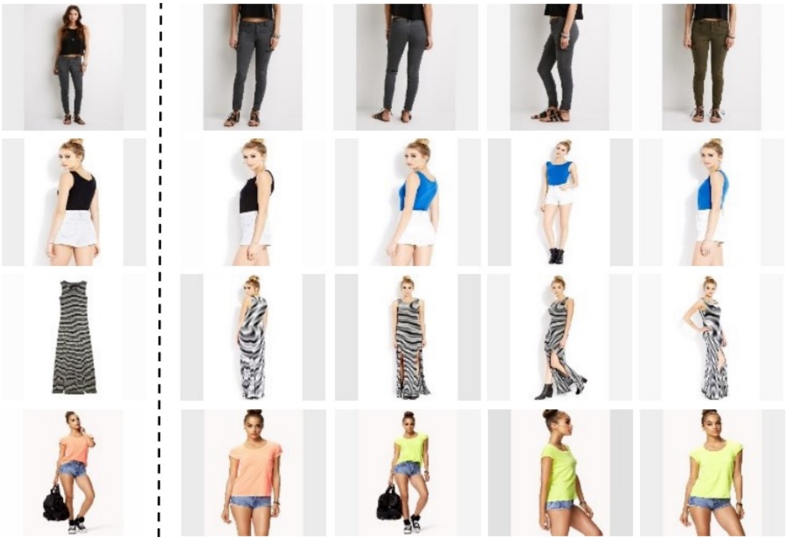
Generic CV Tasks

[Tero Karras et al., ICLR 2018]

Reinforcement Learning

[D Silver et al., Nature 2016]

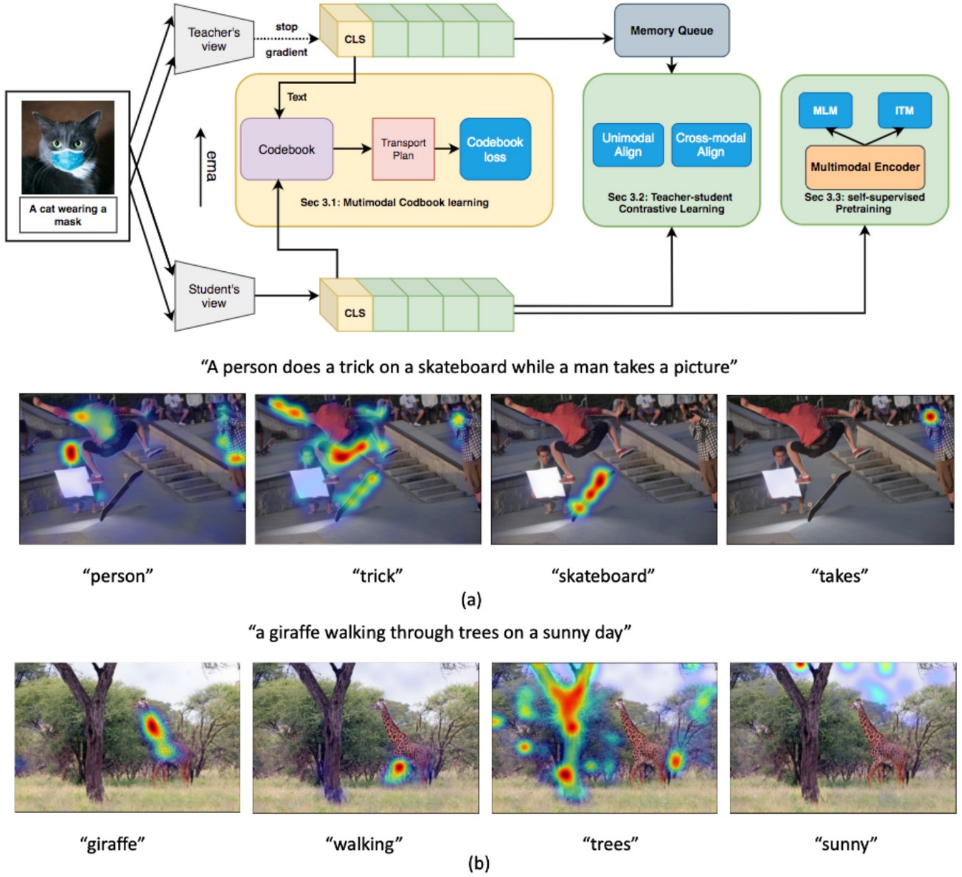
Representation Learning

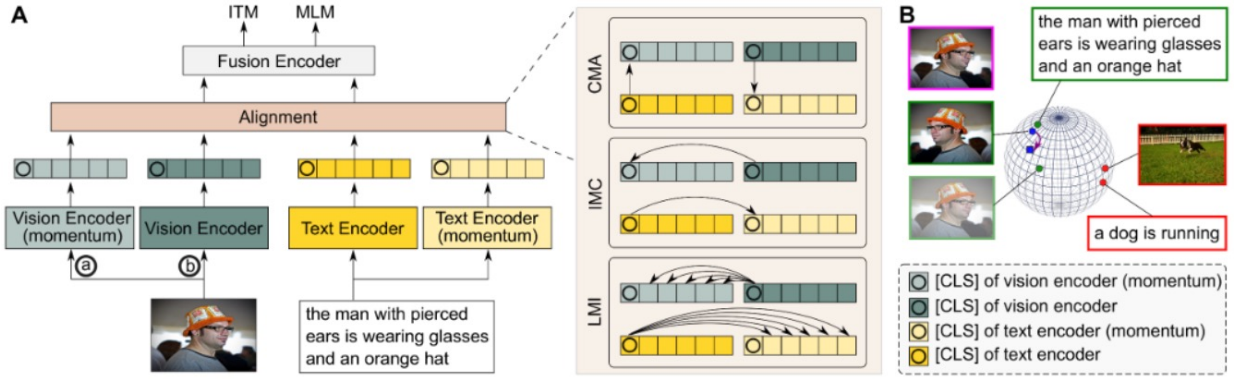


[Sungyeon Kim et al., CVPR 2020]

Recent Works

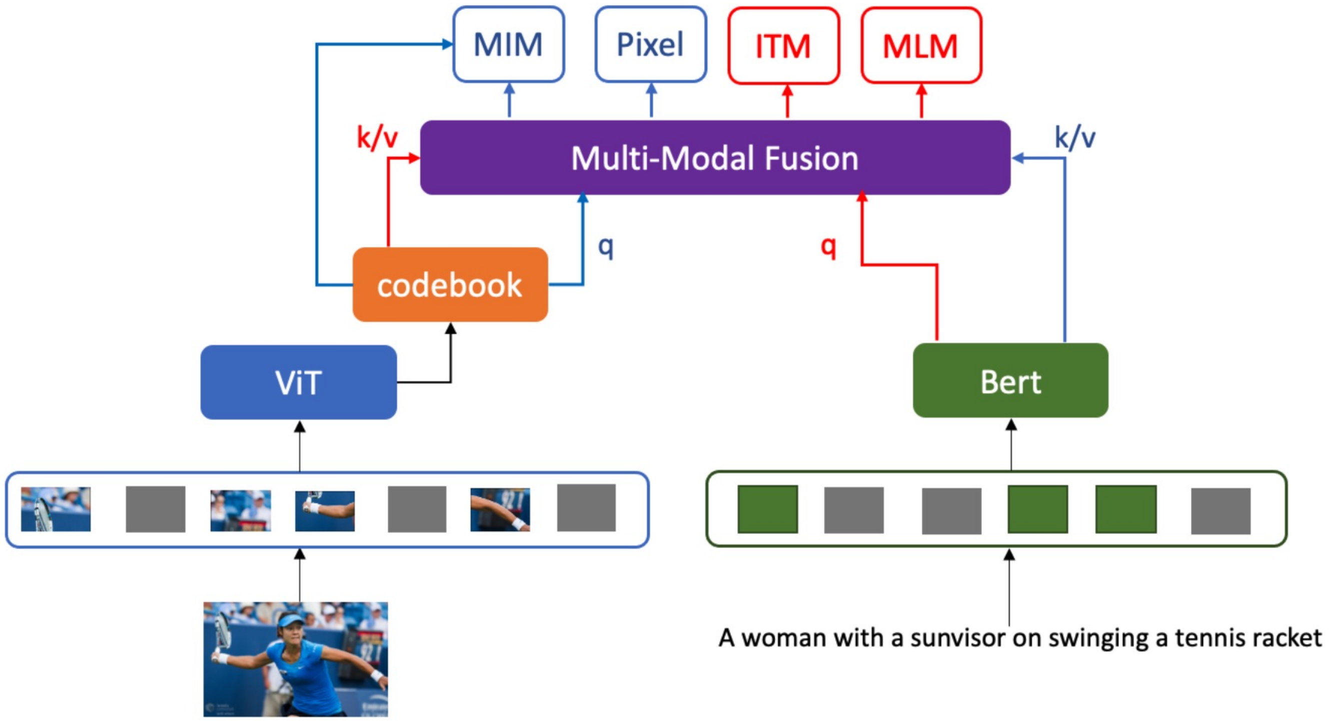
Vision Language Pretraining



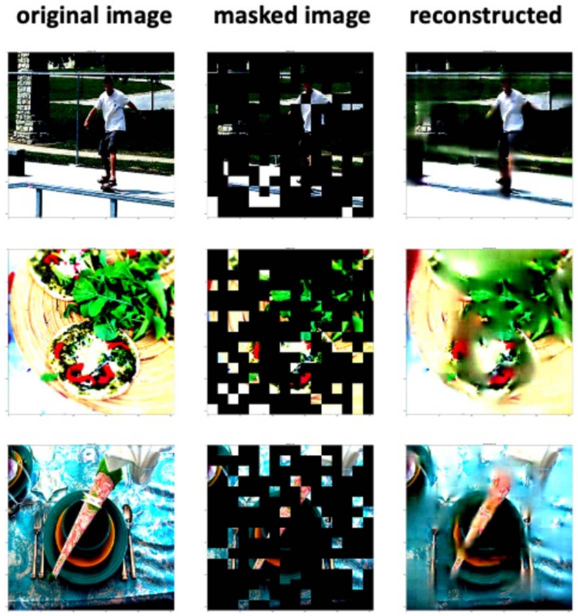


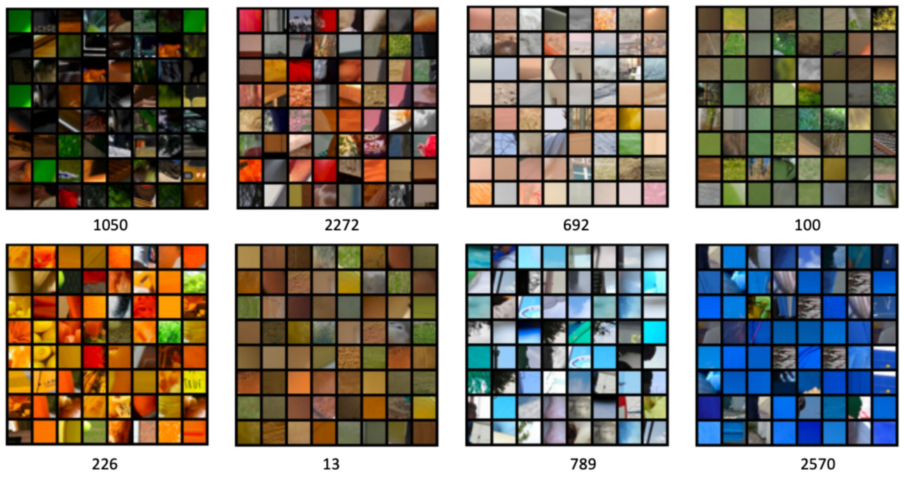
Method	#Images	MSCOCO (5K)						Flickr30K (1K)					
		Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ImageBERT [34]	6M	66.4	89.8	94.4	50.5	78.7	87.1	87.0	97.6	99.2	73.1	92.6	96.0
UNITER [7]	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA [13]	4M	✗	✗	✗	✗	✗	✗	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR [24]	4M	70.0	91.1	95.5	54.0	80.8	88.5	✗	✗	✗	✗	✗	✗
ViLT [20]	4M	61.5	86.3	92.7	42.7	72.9	83.1	83.5	96.7	98.6	64.4	88.7	93.8
UNIMO [23]	4M	✗	✗	✗	✗	✗	✗	89.7	98.4	99.1	74.7	93.47	96.1
SOHO [18]	200K	66.4	88.2	93.8	50.6	78.0	86.7	86.5	98.1	99.3	72.5	92.7	96.1
ALBEF [22]	4M	73.1	91.4	96.0	56.8	81.5	89.2	94.3	99.4	**99.8**	82.8	**96.7**	98.4
Ours	4M	**75.6**	**92.8**	**96.7**	**59.0**	**83.2**	**89.9**	**94.9**	**99.5**	**99.8**	**84.0**	**96.7**	**98.5**
ALIGN [19]	1.2B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6

Recent Works



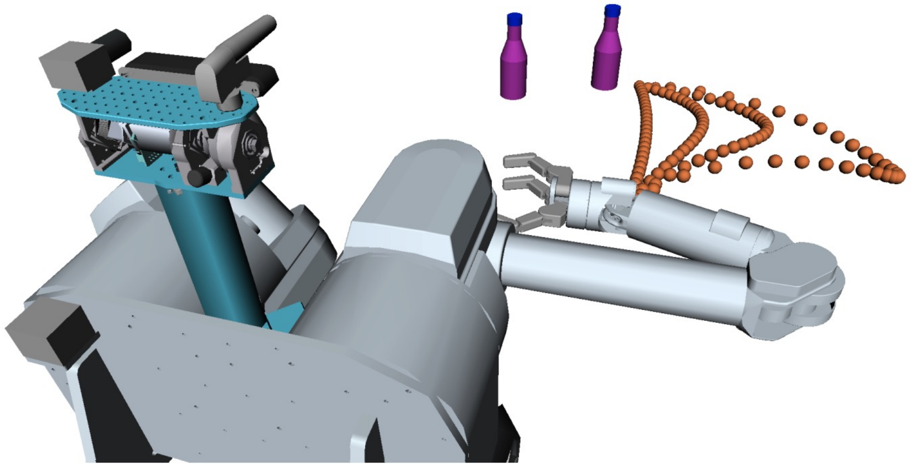
$$\mathcal{L}_{codebook} = -\log(p(x|q(x), l) - ||\mathrm{sg}[\mathrm{ViT}(x)] - e||_2^2 \\ - \beta||\mathrm{ViT}(x) - \mathrm{sg}[e]||_2^2$$





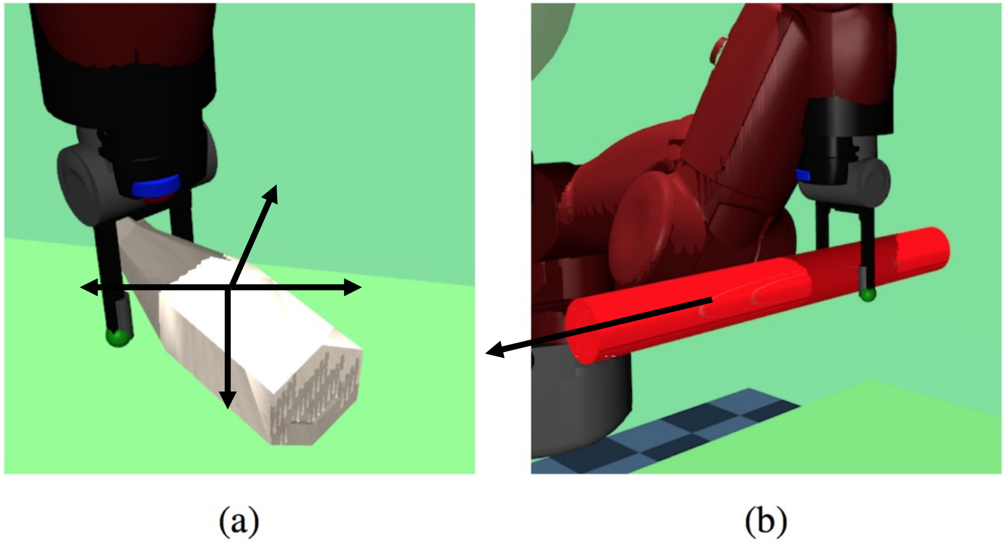
Adversarial Knowledge Learning from Human

Generate legible motion



[Anca Dragan et al., RSS 2013]

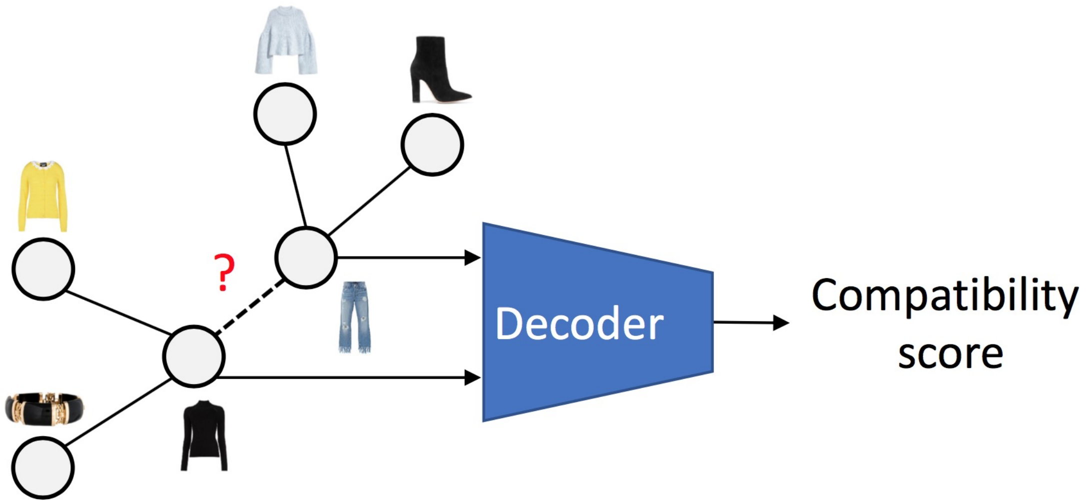
Human-robot Adversarial Learning



[J Duan, J Kuo, S Nikolaidis et al., IROS 2019]

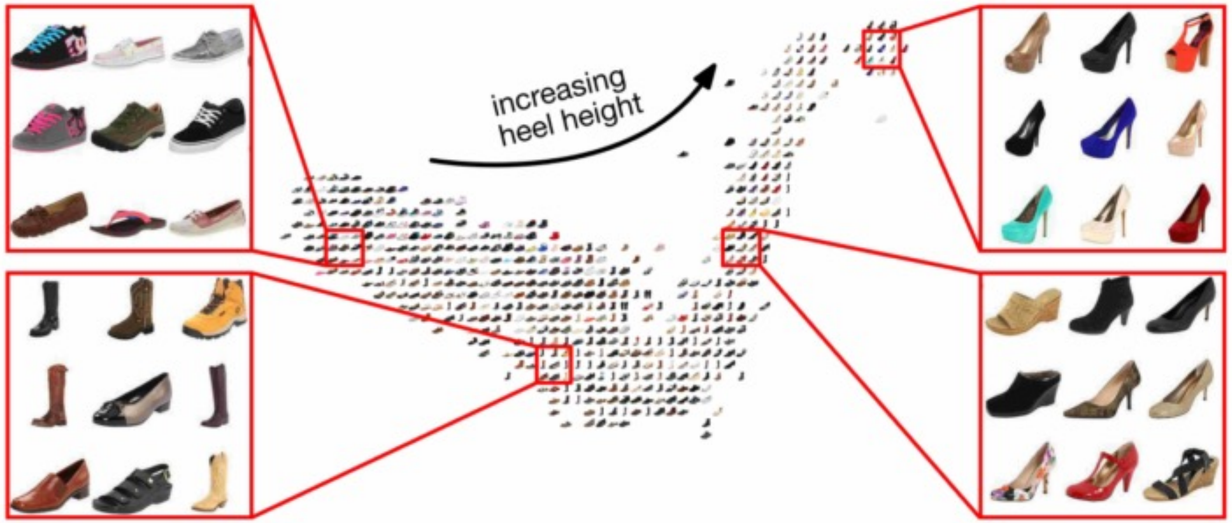
Structured Knowledge Learning via Metric Learning

Compatibility Graph Embedding



[J Duan, J Kuo et al., SCMLS 2020]

Conditional Similarity Embedding

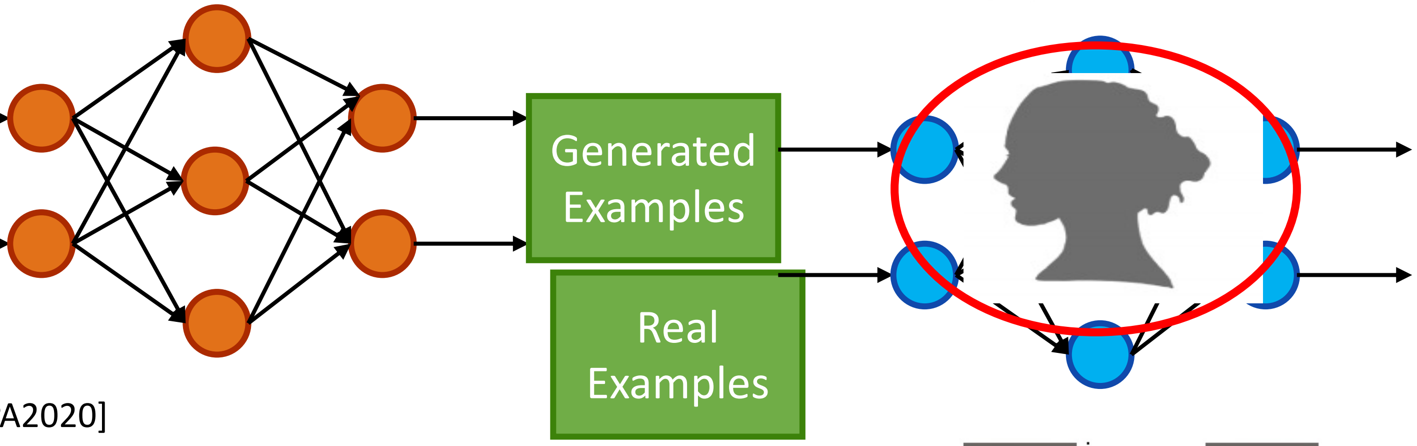


[Belongie, Serge et al., CVPR 2017]

Work Overview

Adversarial Knowledge Learning

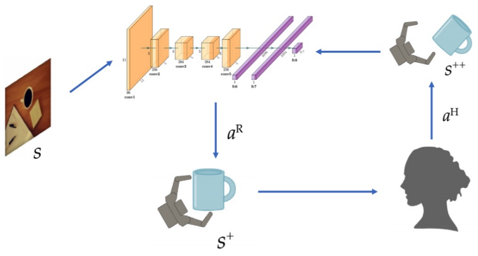
Generator



Portrait Manipulation

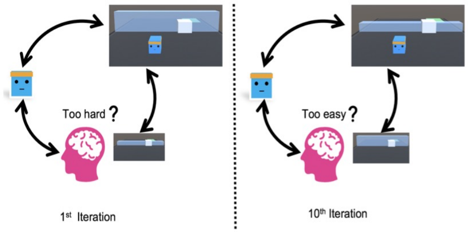
- [J Duan, J Kuo et al., APSIPA2020]

Structured Knowledge Learning



Human-robot Adversarial Games

- [J Duan, J Kuo, S Nikolaidis et al., IROS2019]
- [J Duan, J Kuo, S Nikolaidis et al., SCR 2019]

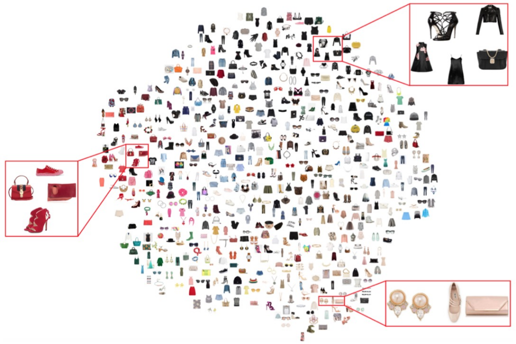


Human-guided Curriculum RL

- [J Duan, J Kuo, S Nikolaidis et al., 2020]

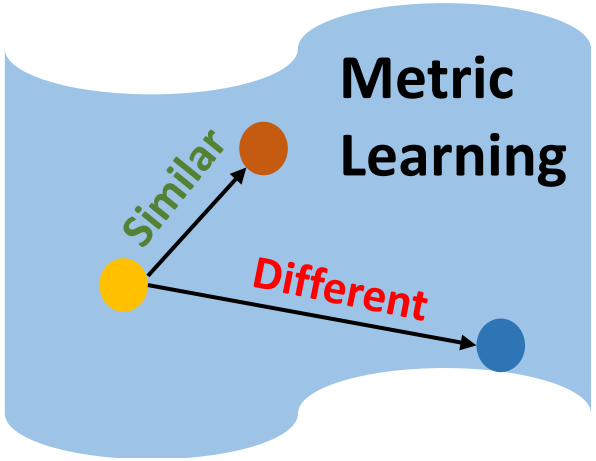
Work Overview

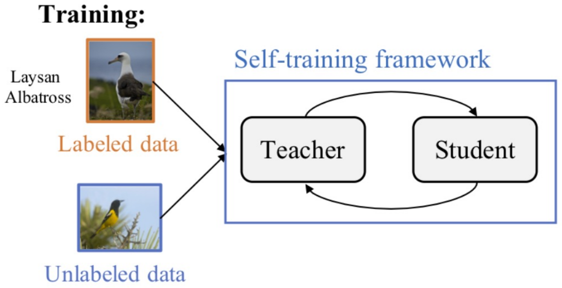
Adversarial Knowledge Learning



Compatible recommendation

- [J Duan, J Kuo et al., SCMLS 2020]
- [J Duan, J Kuo et al., JVCI 2021 (under review)]

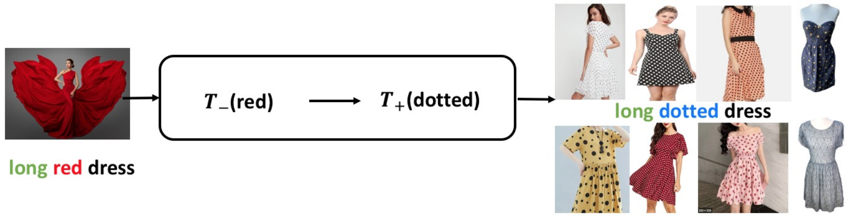




Self-training framework

- [J Duan, J Kuo et al., CVPR 2021]

Structured Knowledge Learning

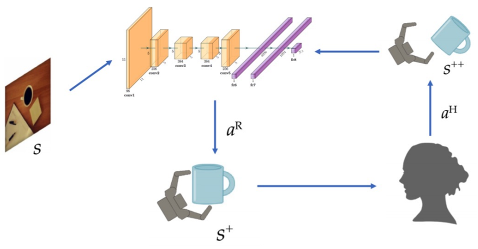


Compositional learning

- [J Duan, J Kuo et al., Preprint 2021]

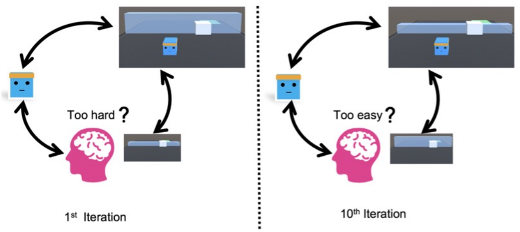
Work Overview

Adversarial Knowledge Learning



Human-robot Adversarial Games

- [J Duan, J Kuo, S Nikolaidis et al., IROS2019]
- [J Duan, J Kuo, S Nikolaidis et al., SCR 2019]



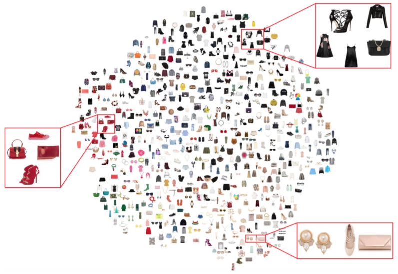
Human-guided Curriculum RL

- [J Duan, J Kuo, S Nikolaidis et al., 2020]

Portrait Manipulation

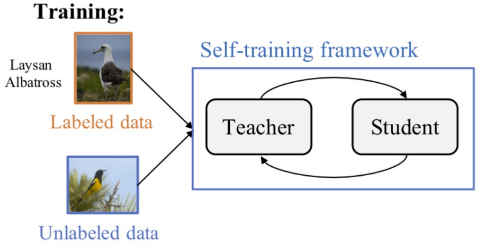
- [J Duan, J Kuo et al., APSIPA2020]

Structured Knowledge Learning



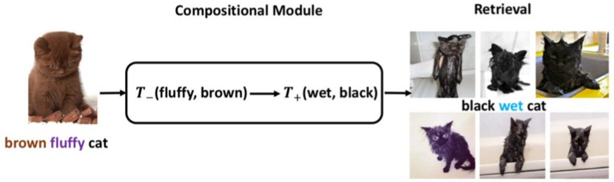
Compatible recommendation

- [J Duan, J Kuo et al., SCMLS 2020]
- [J Duan, J Kuo et al., JVCI 2021 (under review)]



Self-training framework

- [J Duan, J Kuo et al., CVPR 2021]

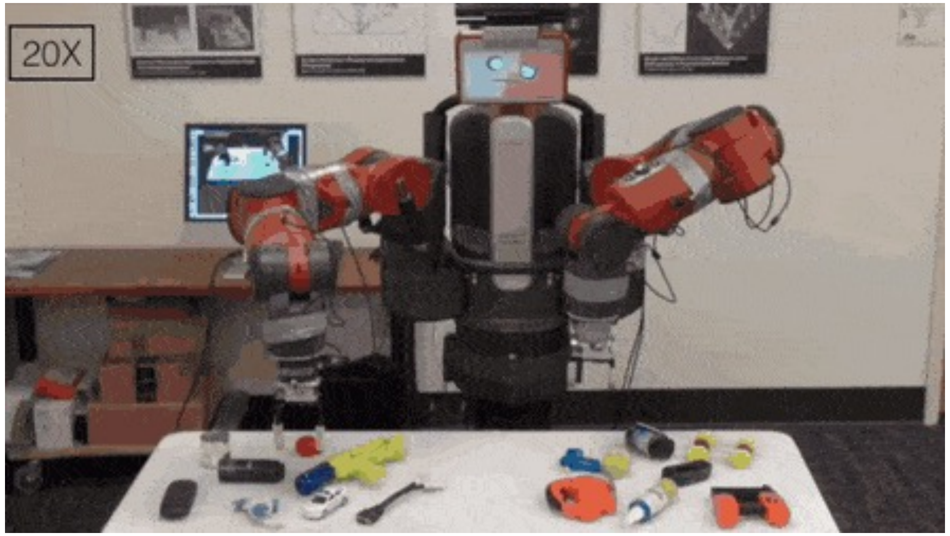


Compositional learning

- [J Duan, J Kuo et al., Preprint 2021]

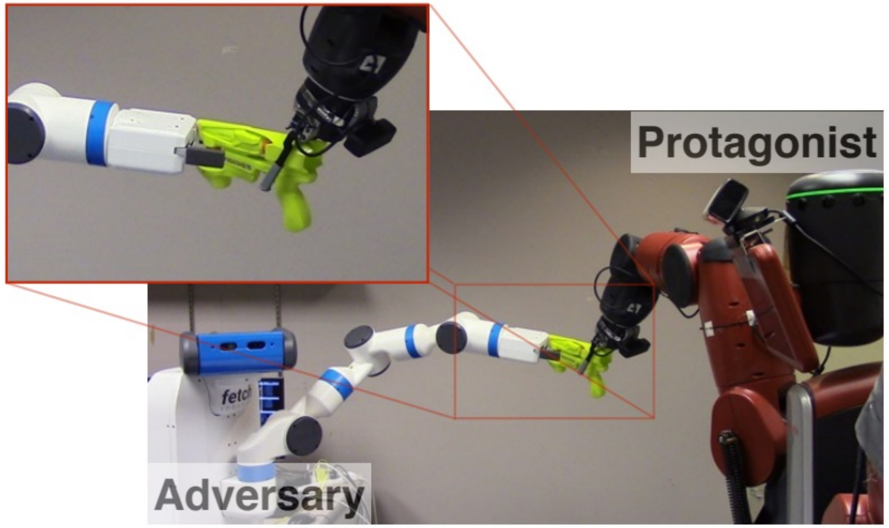
Learning Based Robotic Grasping

- Self-Supervised Learning [1]
  - Collect annotated data
  - Training and repeat



- Require combination of sensors
- Time consuming

- Simulated-Adversary [2]



1. Pinto, Lerrel, et al. "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours." ICRA 2016
2. Pinto, Lerrel et al. "Supervision via competition: Robot adversaries for learning tasks." ICRA 2017

Why Adversarial Human?

- Human has good prior over success/failure
- Collaborative human is not always true

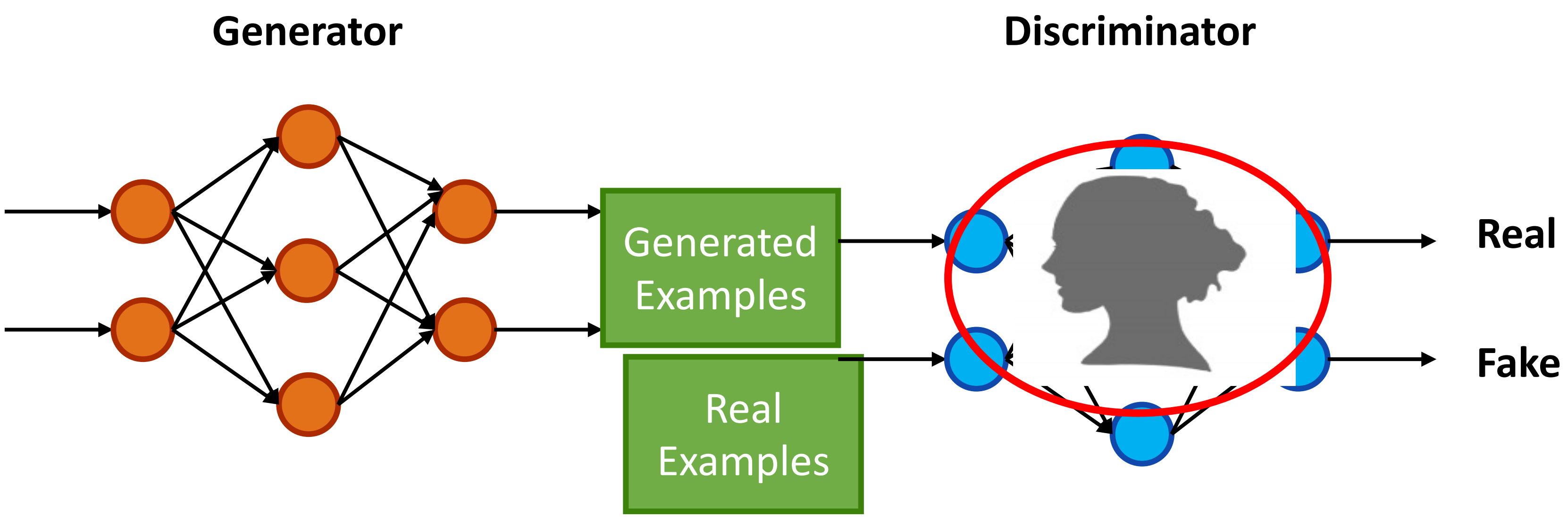
- *Human Robot Adversary > Self-Supervised Method*
- *Human Robot Adversary > Simulated adversary*

Dražen Brščić et al., 2015

Human-Robot Adversarial Learning as Gaming

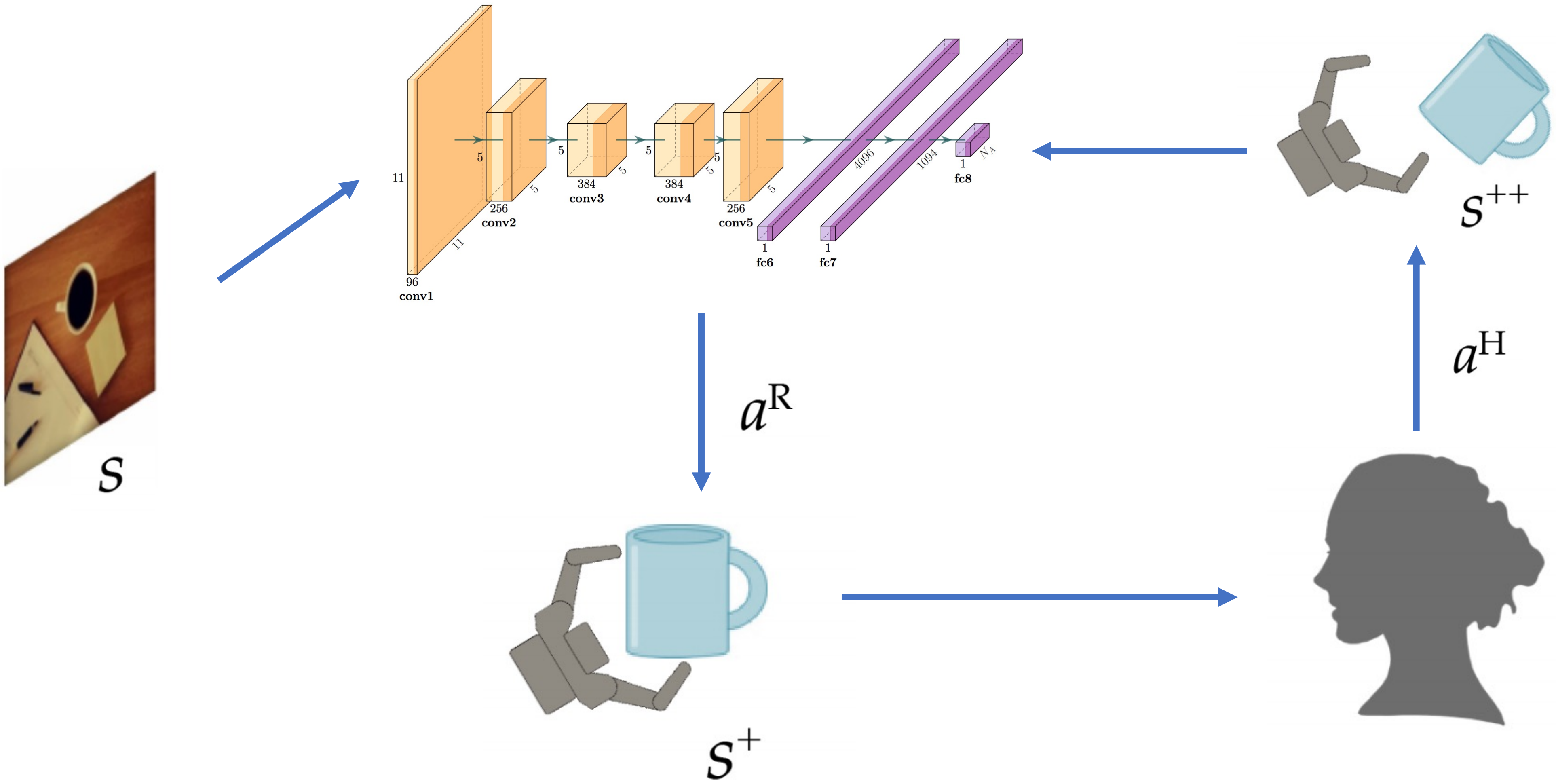
- Human-robot adversarial game
  - Robot Policy: $\pi^R : s \rightarrow a^R$
  - Human Policy: $\pi^H : s^+ \rightarrow a^H$
  - Reward: $r = R^{\mathrm{R}}(s, a^{\mathrm{R}}, s^+) - \alpha R^{\mathrm{H}}(s^+, a^{\mathrm{H}}, s^{++})$

$$\pi_*^{\mathrm{R}} = \underset{\pi^{\mathrm{R}}}{\operatorname{argmax}} \, \mathbb{E}\left[r(s, a^{\mathrm{R}}, a^{\mathrm{H}}) | \pi^{\mathrm{H}}\right]$$



Robot Learning via Human Adversarial Games. J Duan, J Kuo, S Nikolaidis et al. IROS 2019, Best Paper Finalist

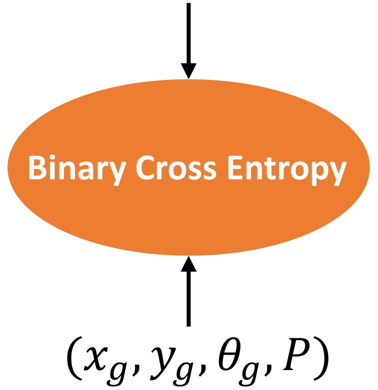
Overview of Our Framework

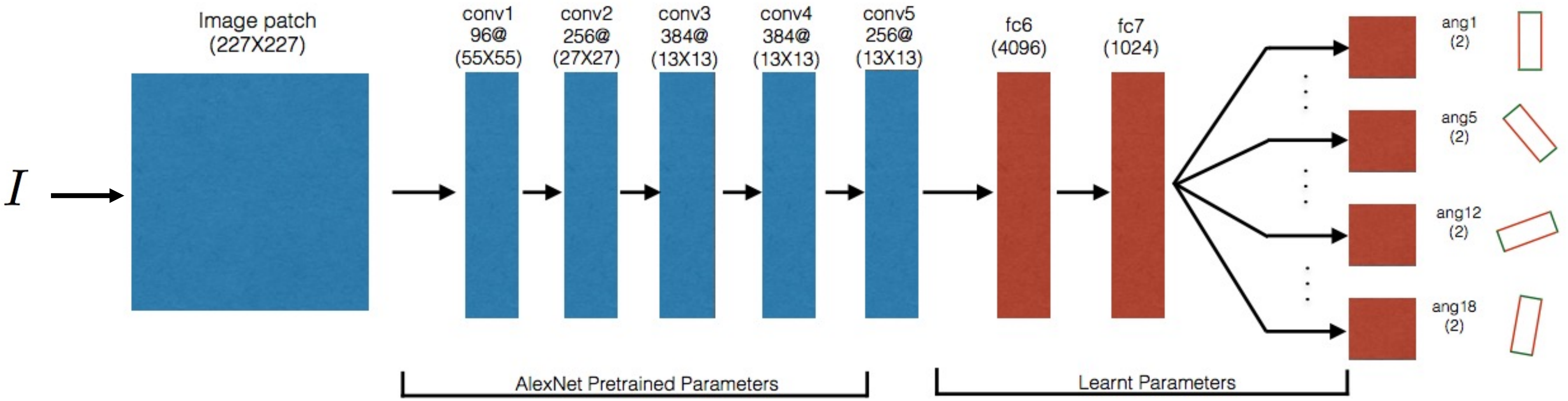


Training of Framework

$$r = \begin{cases} 0 & \text{if robot fails to grasp} \\ 1 & \text{if robot succeeds and human fails} \\ 1-\alpha & \text{if human succeeds} \end{cases}$$

- Grasping prediction
  - Input: I
  - Architecture: Alex-Net
  - Robot Policy $\pi^{\mathrm{R}} : I \mapsto (x_g, y_g, \theta_g)$
  - Confidence: P





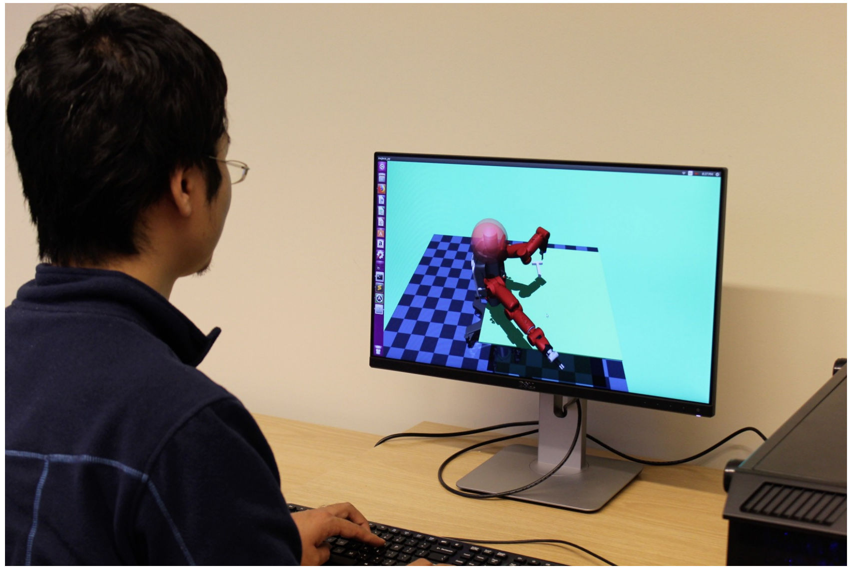
User Study Design

- Protocol
  - Goal: Maximize the failure of robot
  - Trials of 10 times before start
  - Between-subject

    21 Male/4 Female; one object per user
- Adversarial Disturbances
  - Interactive Mujoco [1]
  - Discrete action with 6 actions

    Up/down, left/right, inward/outward
  - Outcome

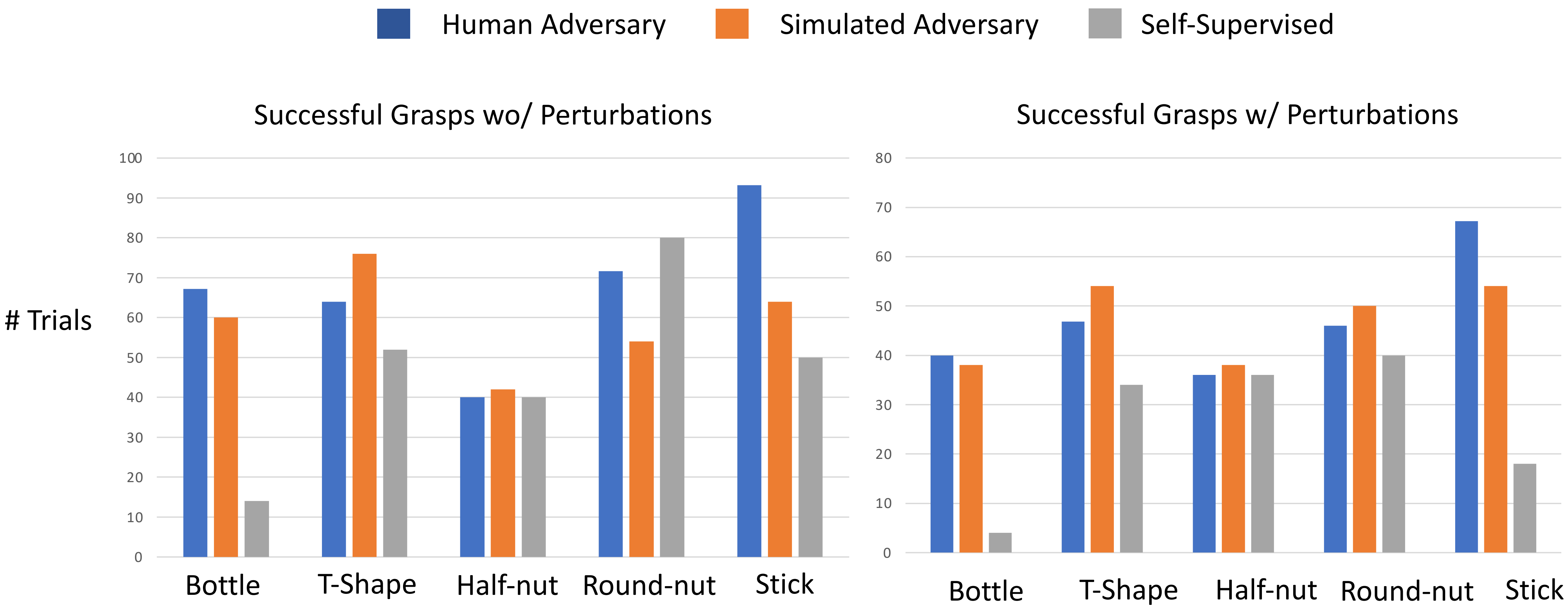
    remain on gripper/drop



Participants interacted with a simulated Baxter robot in the customized Mujoco simulation environment.

1. https://github.com/davidsonic/Interactive-mujoco_py

Successful Grasps without & with Perturbations

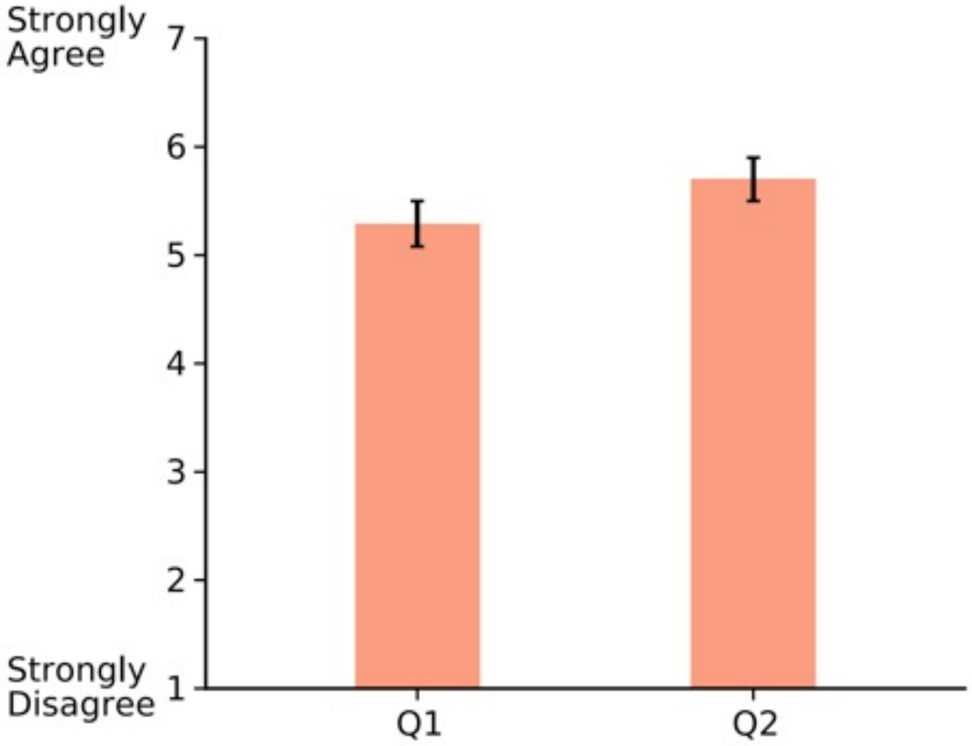


Subjective Metric

Questionnaire

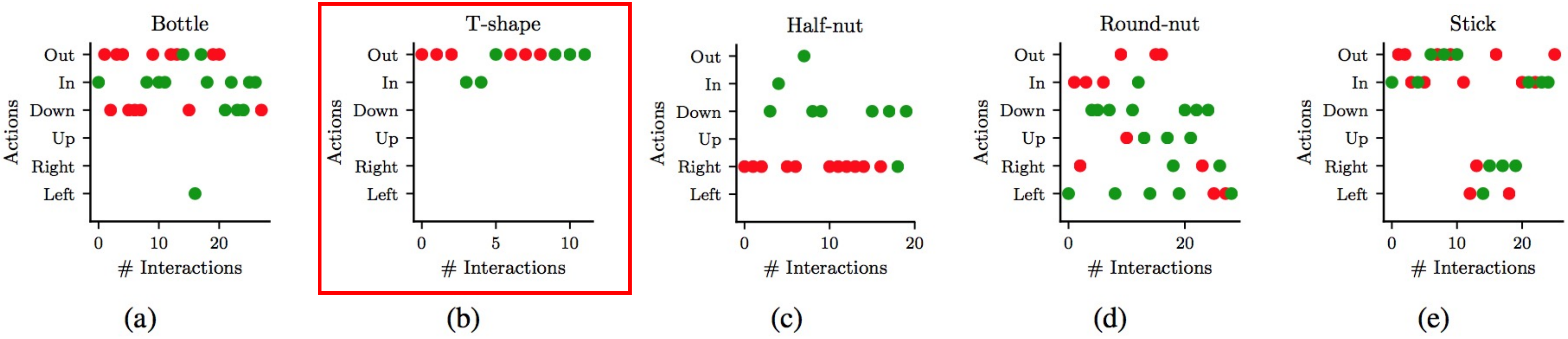
Q1: The robot learned throughout the study.

Q2: The performance of the robot improved throughout the study.



Adversarial Actions over Time

Transition from human success to robot success

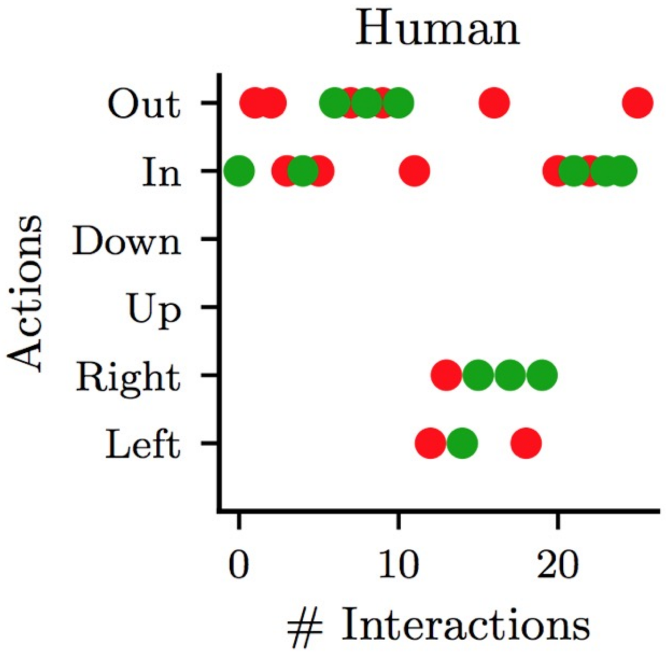


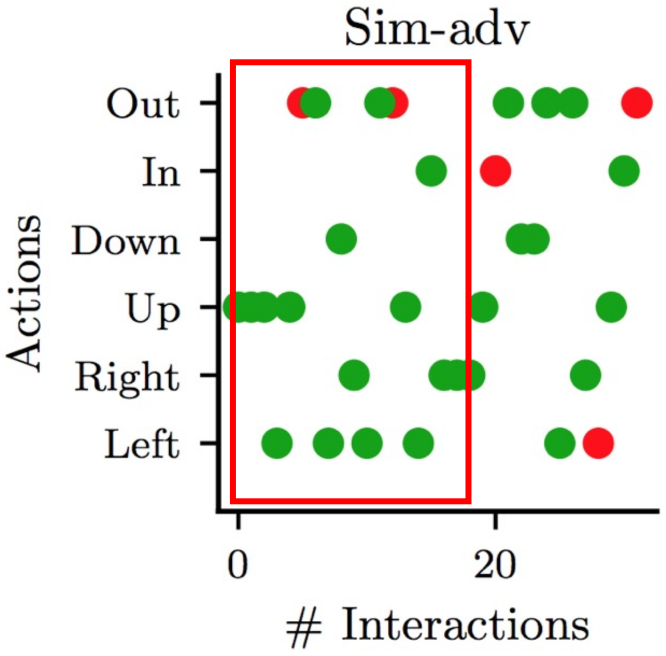
● human perturbation succeeds; ● robot succeeds.

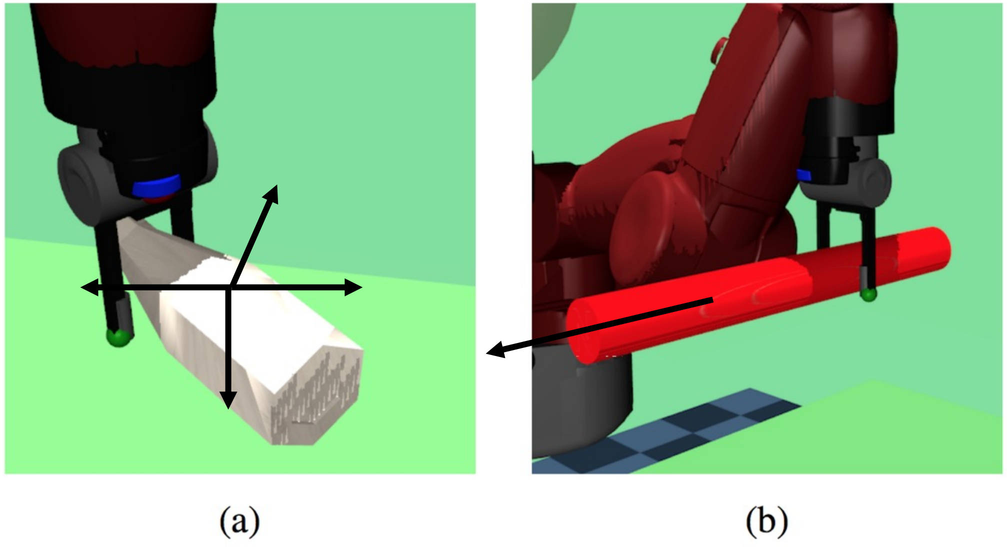
Expected Scenario: Human succeeds more at the beginning, then robot evolves to succeed more

Human vs Simulated Adversary

● human perturbation succeeds; ● robot succeeds.







Different geometry require different perturbations

Human prior over weakness of grasping and therefore more effective

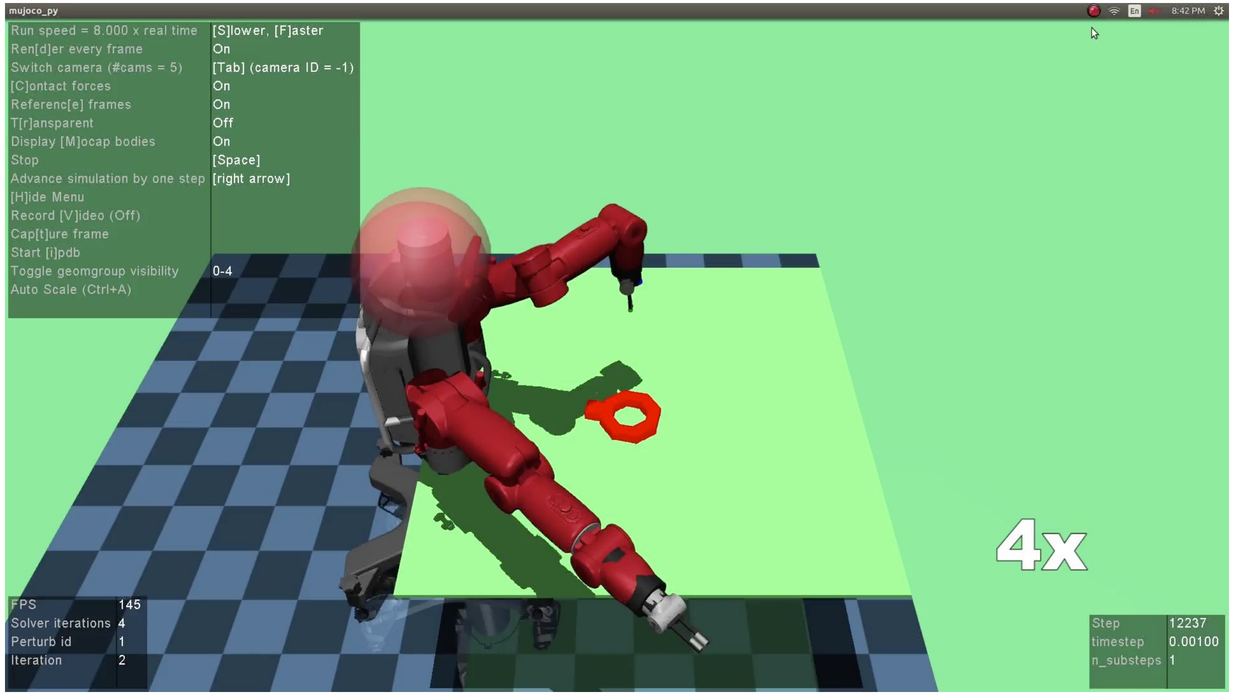
Experiment: Multiple Objects

- Task setting
  - Training 200 episodes
  - Different position & orientation
- Results

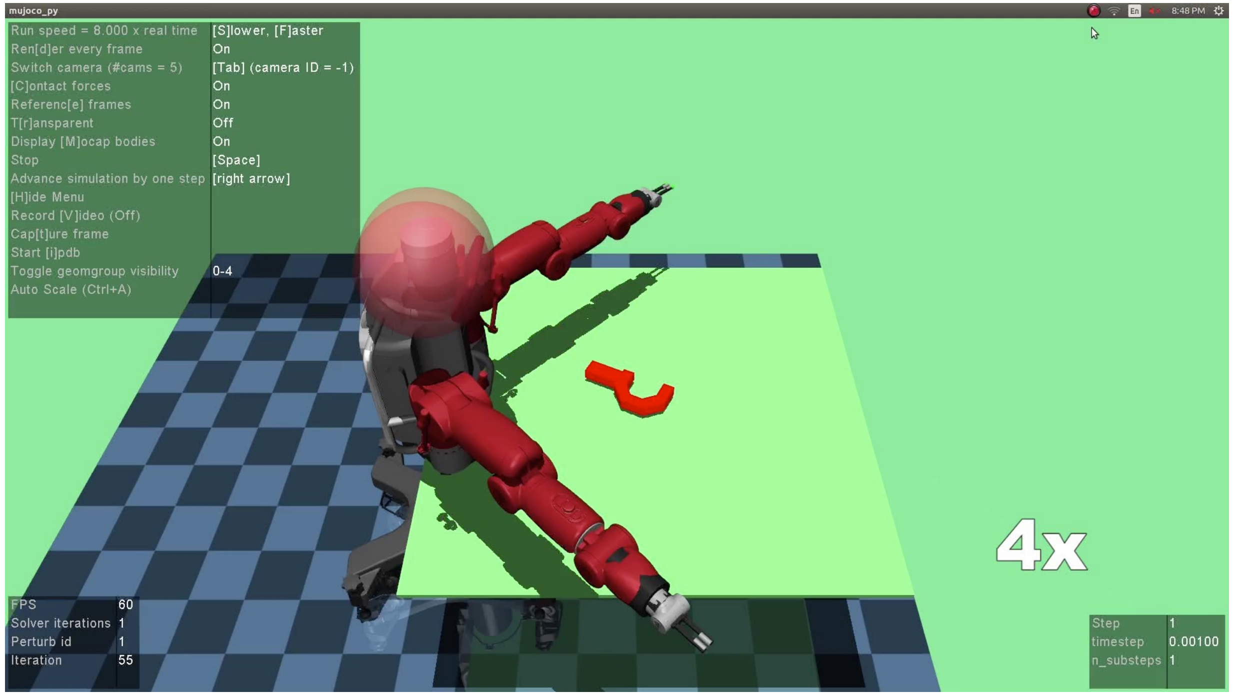
Success Rate	Without Disturbance	With Disturbance
Human-Robot Learning	52%	34%
Simulated Adversary	28%	22%

Comparison Before & After Training

Before training



After training



Can We Leverage Adversarial Human for RL?

Human-robot Adversarial Learning

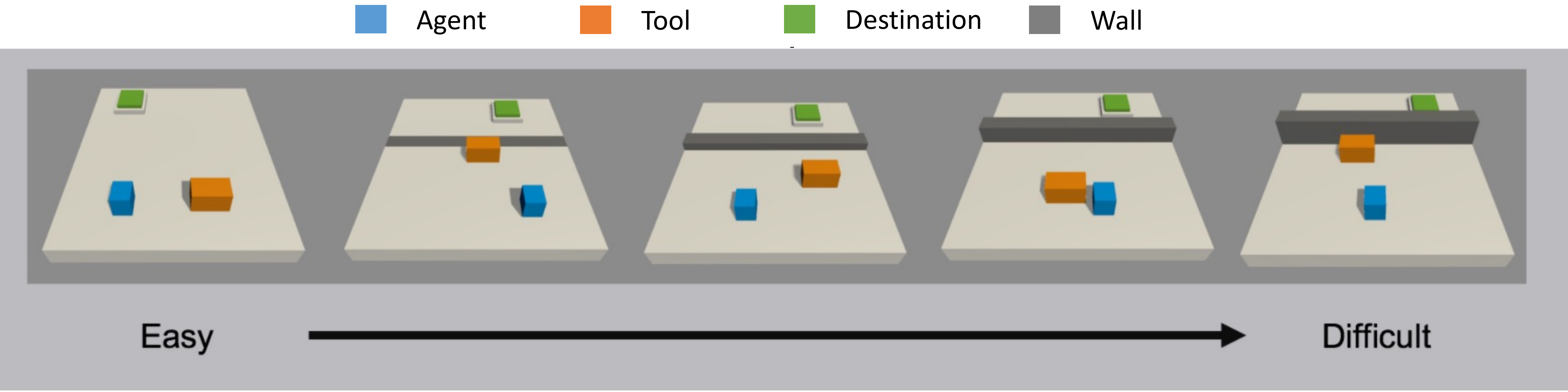
- Grasping prediction network: $\pi^R: I \rightarrow (x_g, y_g, \theta_g)$
- Reward: $r = R^R(s, a^R, s^+) - \boxed{\alpha R^H(S^+, a^H, S^{++})}$
- Loss: $Binary\ Cross\ Entropy(P, r)$

Reinforcement Learning

- Policy network: $\pi_\theta: s_t \rightarrow a_t$
- Reward: $r = \boxed{R_t(s_t, a_t)}$
- Policy gradient: $\nabla_\theta J(\pi_\theta) = E_t[R_t \nabla_\theta log\pi_\theta(a_t|s_t)]$

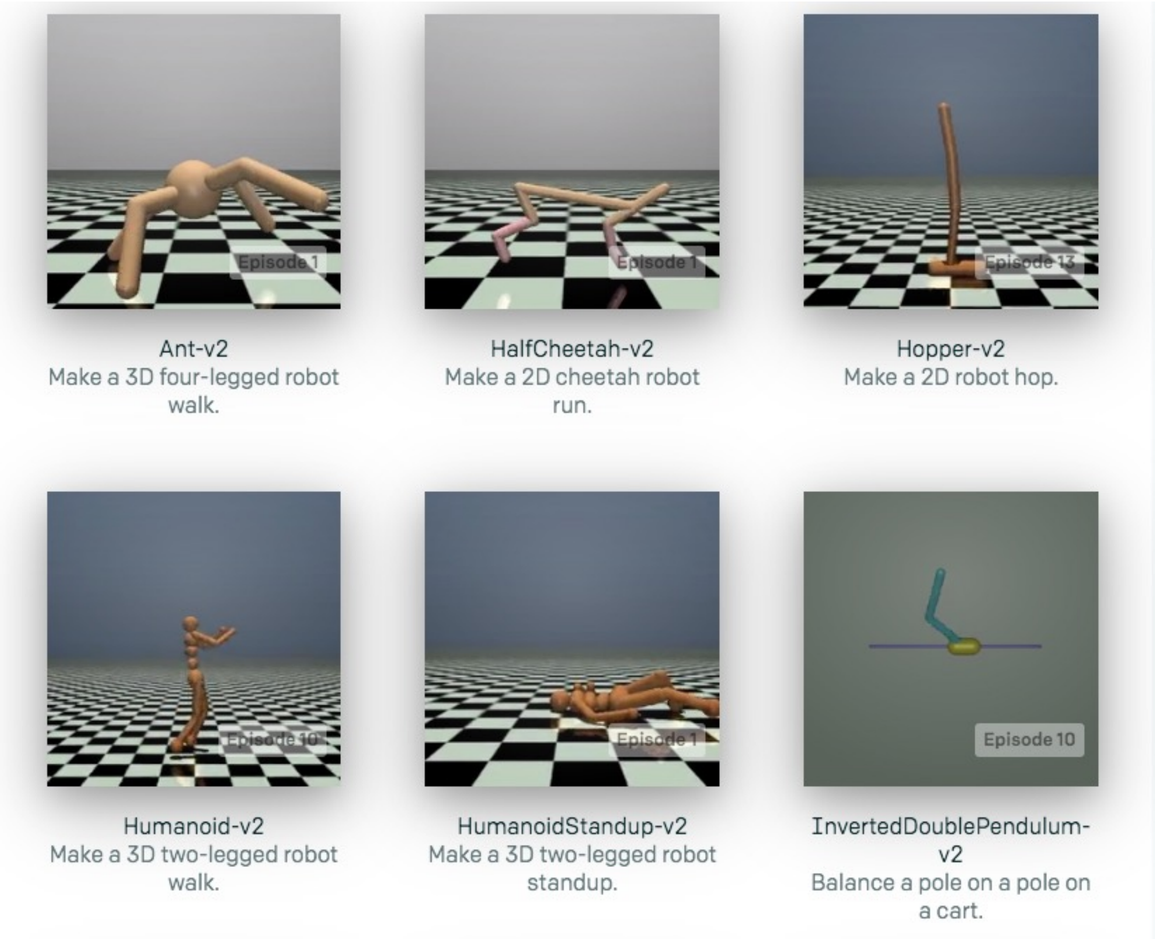
Adversarial Curriculum RL with Human-in-the-Loop

- Improve reinforcement learning with adversarial human
- Leverage adversarial human for curriculum design

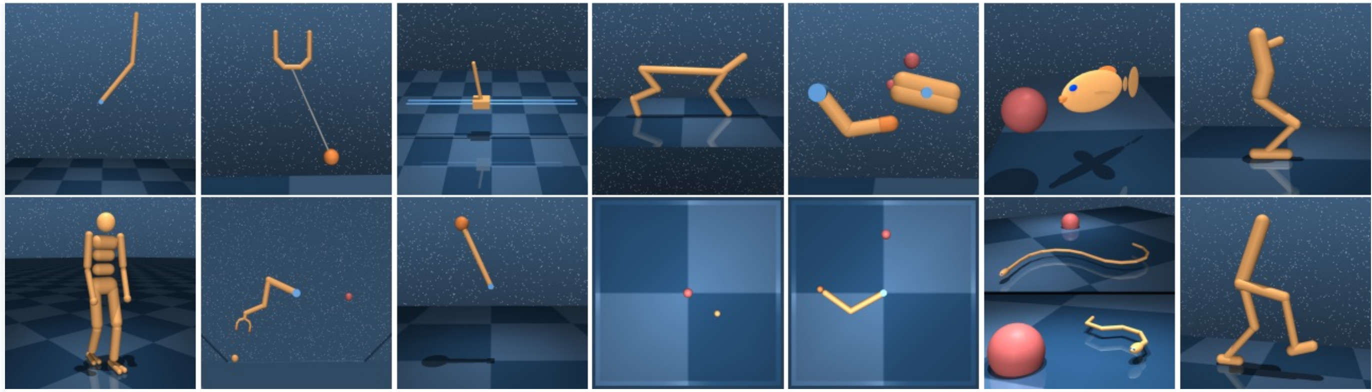


Curriculum Reinforcement Learning Guided by Human. J Duan, J Kuo, S Nikolaidis et al. 2020

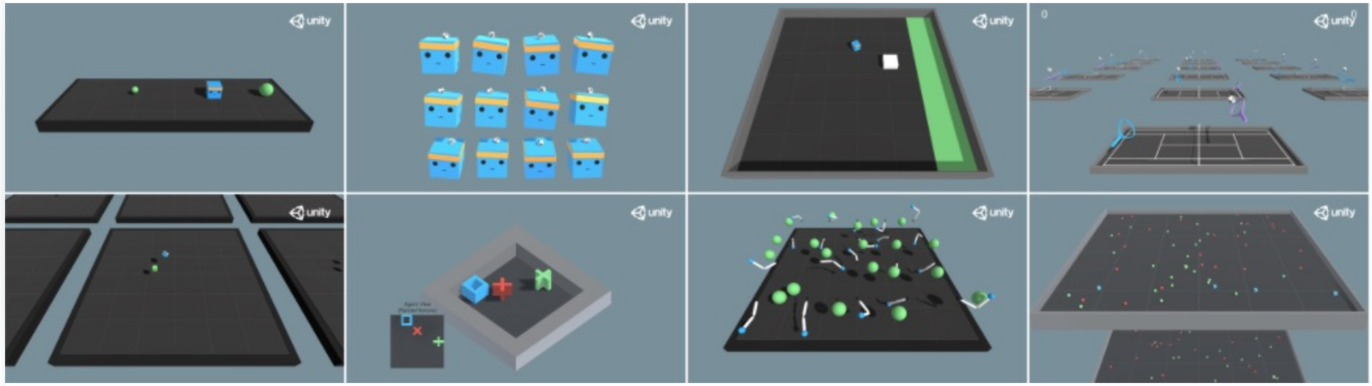
Existing Platform



OpenAI Gym

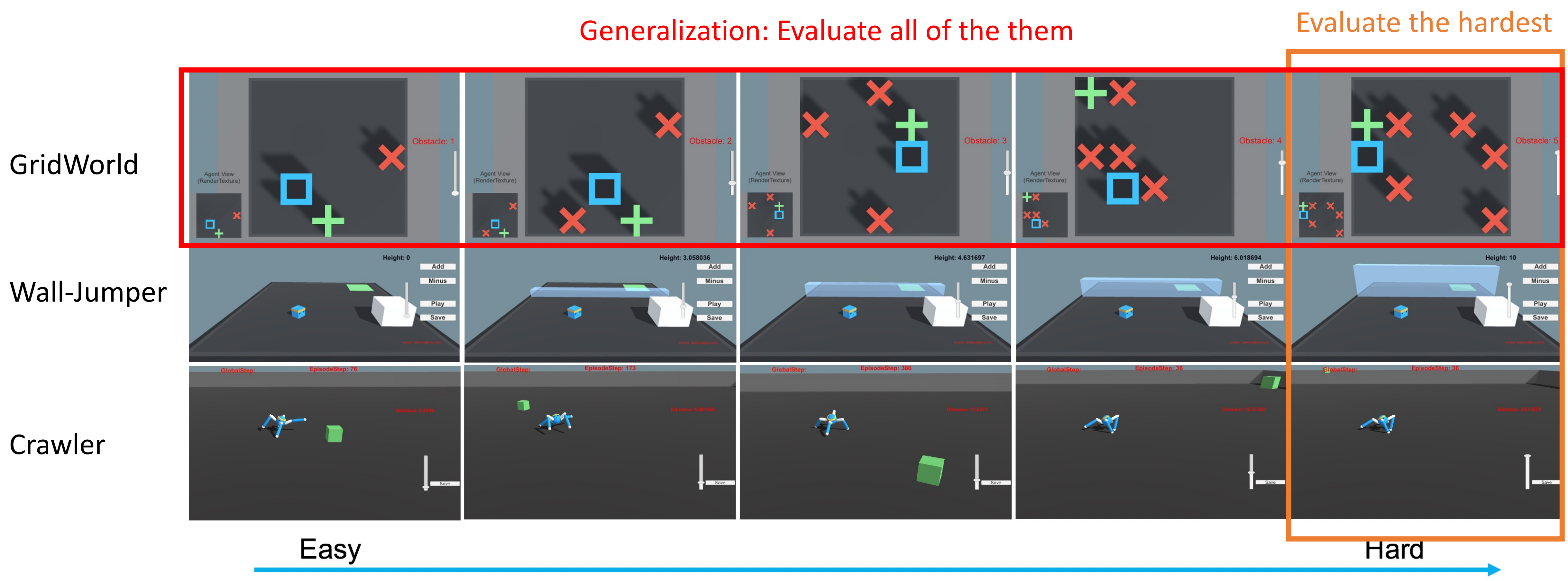


DeepMind Control Suite

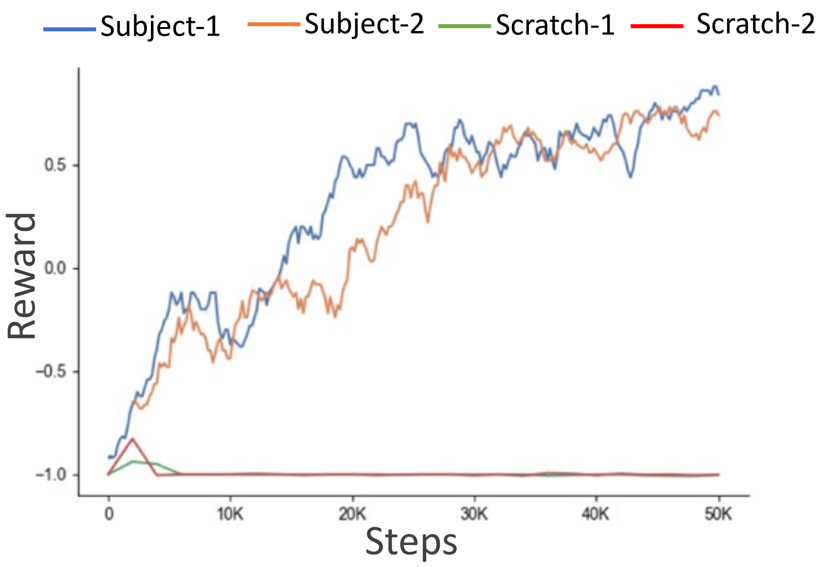


MLAgents

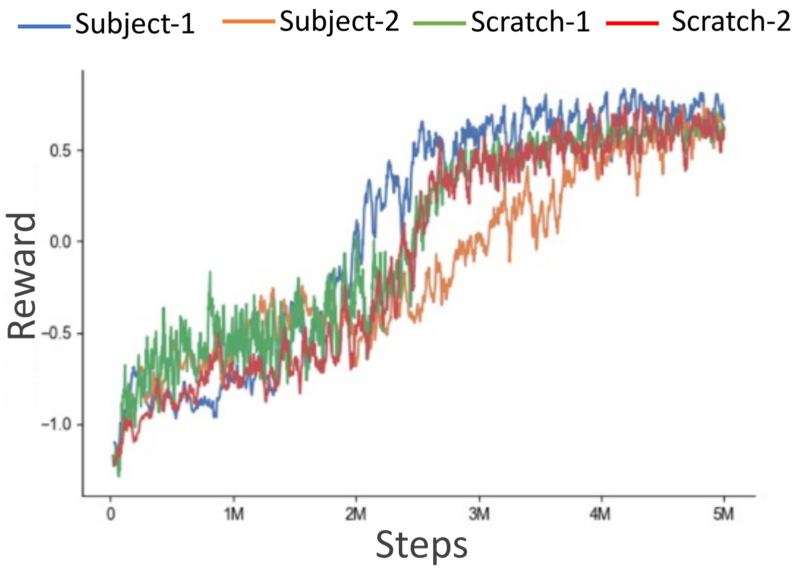
Our Platform



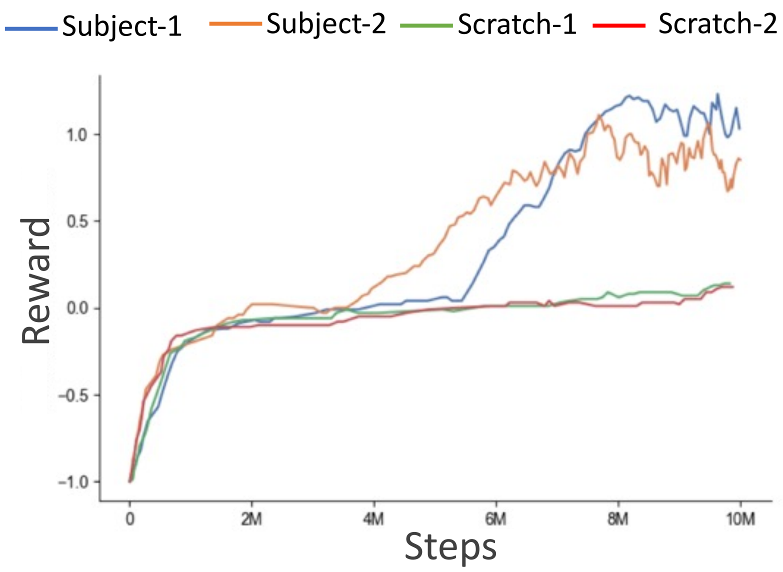
Evaluate on the Ultimate Task



(a) GridWorld (obstacles of 5)



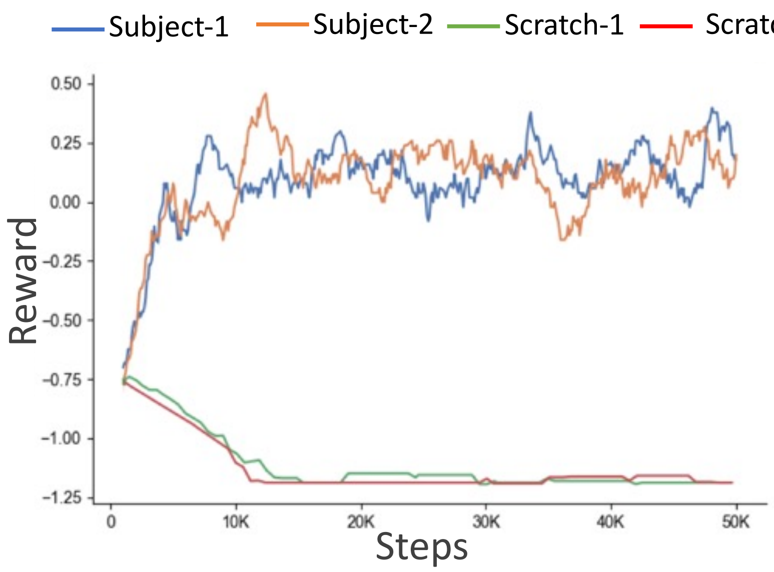
(b) Wall-Jumper (height of 8)



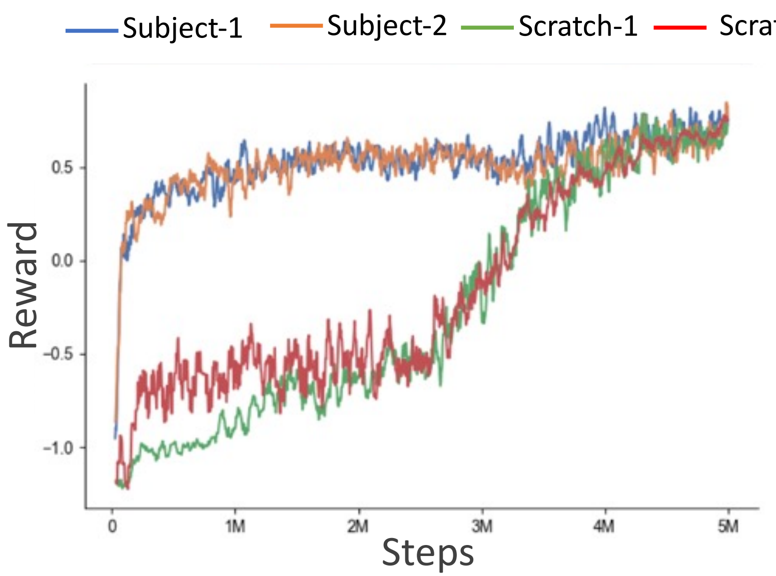
(c) SparseCrawler (radius of 40)

Higher expected accumulated reward while learning from scratch could fail completely

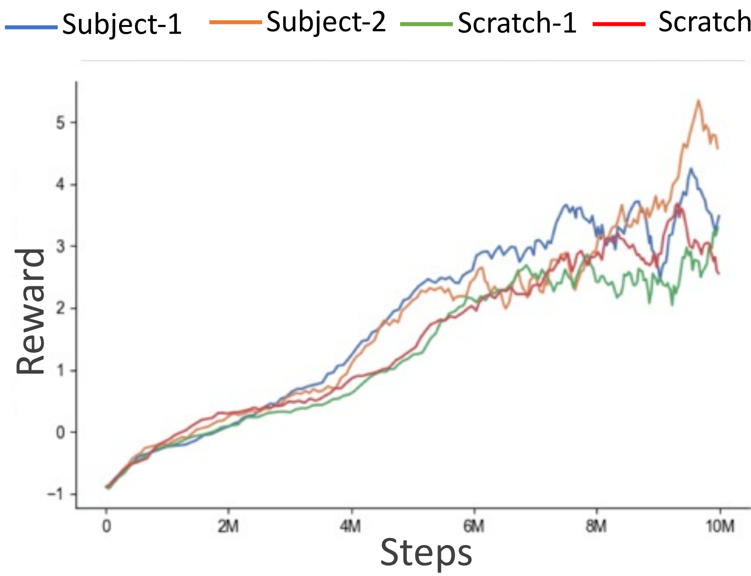
Evaluate across Tasks



(a) GridWorld (obstacles from 1 to 5)



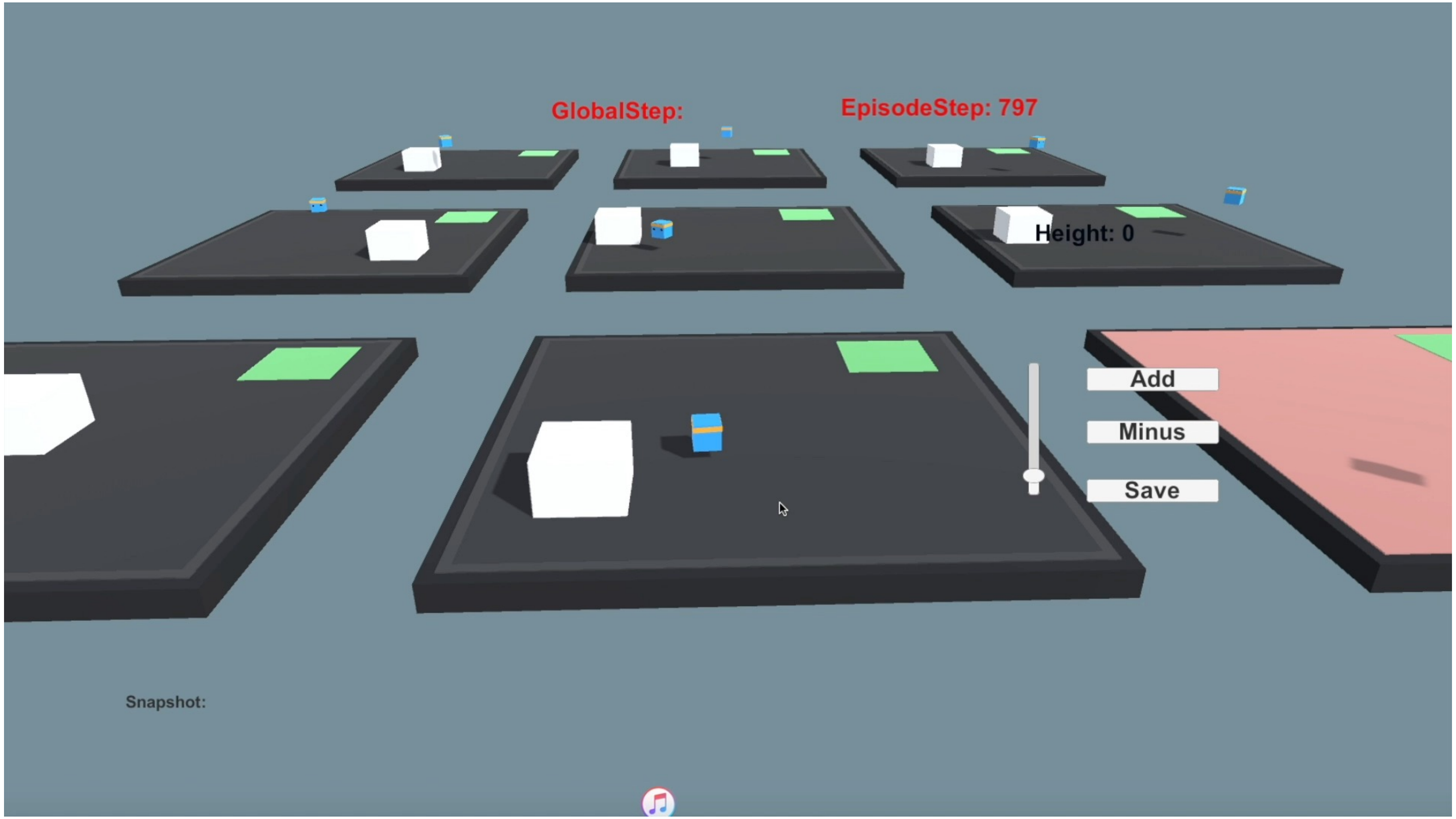
(b) Wall-Jumper (heights from 0 to 8)



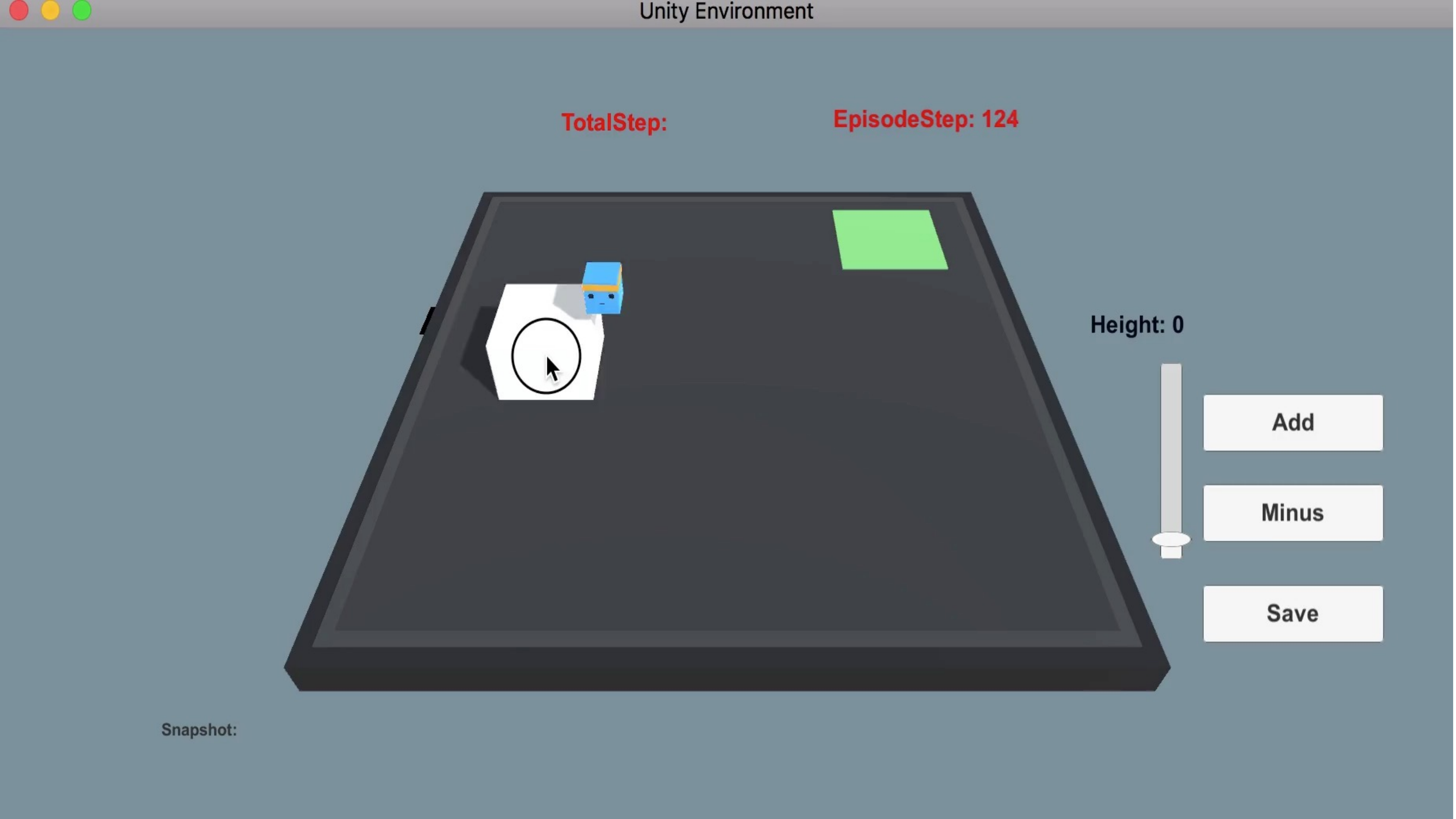
(c) SparseCrawler (radius from 5 to 40)

Better sampling efficiency with human in the loop

Agent Tool Destination Wall



Eval Eval Eval


Unity Environment
TotalStep:
EpisodeStep: 124
Height: 0
Add
Minus
Save
Snapshot:

- Challenge 1: Human has little control in deep learning
    - **How** do we improve robustness of deep learning?
    - **Can** we incorporate human prior into optimization?

- Incorporate human-in-the-loop as an adversary
- Exploit human adversarial knowledge to improve robustness
- Leverage human prior in reinforcement learning optimization

Summary of Contributions

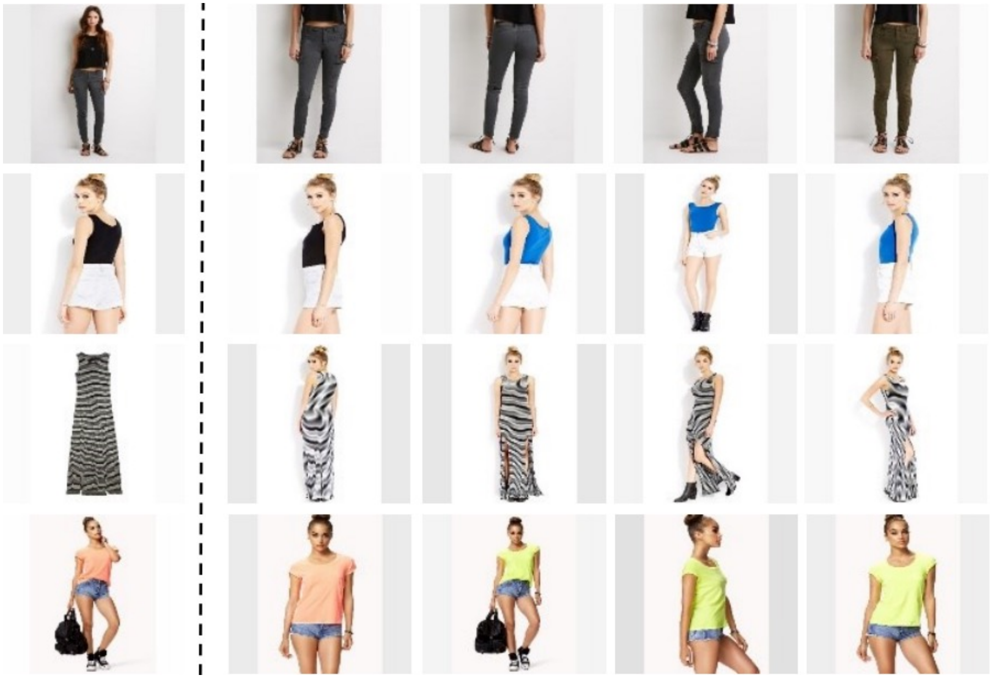
- **Adversarial Knowledge Learning from Human and Data**
  - Propose a general framework for human-robot adversarial learning
  - Interactive human for curriculum reinforcement learning
  - Exploit human prior for portrait manipulations

- **Practical and Technical Contributions**
  - An interactive human-robot adversarial learning platform [1]
  - An interactive curriculum RL platform [2]
  - A web application for portrait manipulation [3]

1. https://github.com/davidsonic/Interactive-mujoco_py
2. https://github.com/davidsonic/interactive-curriculum-reinforcement-learning
3. https://github.com/davidsonic/Flexible-Portrait-Manipulation

Structured Representation via Metric Learning

Learn structured embedding space



Query Ranking

t-SNE visualization [1] of the embedding space

1. Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *JMLR* 2008.

Applications

Information Search

Text-to-image retrieval

Blue,
Normal-sky,
With horizon

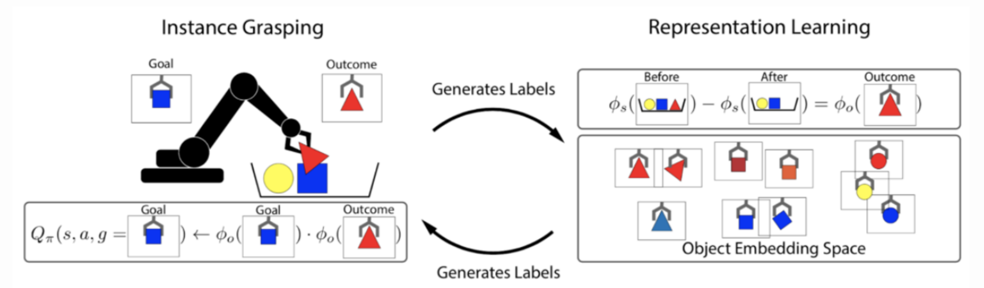
E-Commerce

Visual similarity search

Distance Metric Learning

Robotics

Representation Learning

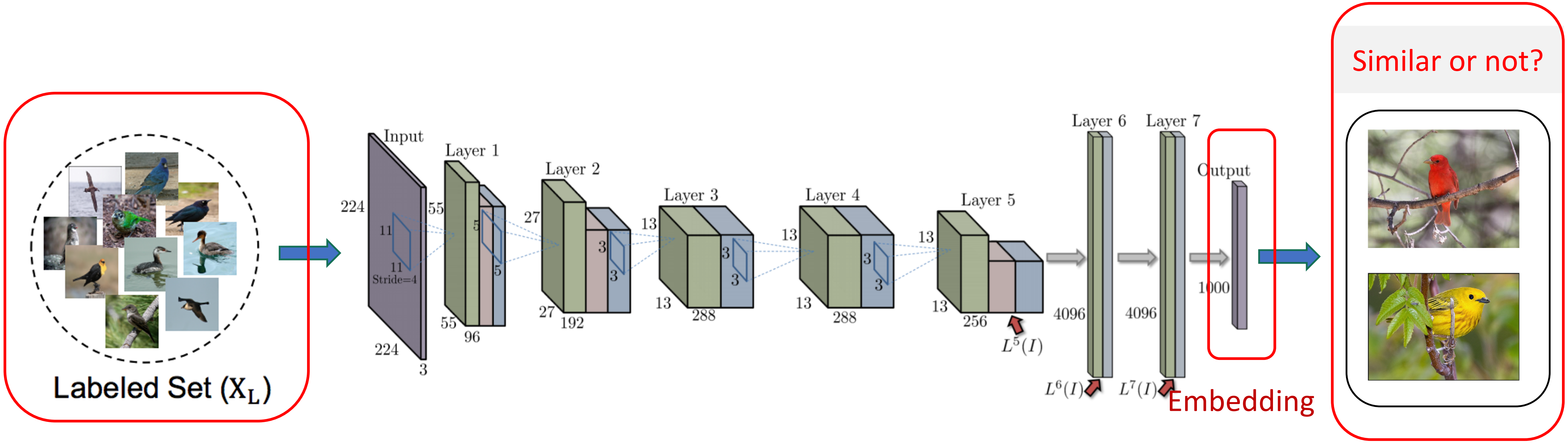


Security/Surveillance

Person-Reidentification

C.-C. Jay Kuo, "Resolving the Gap between Image Pixels and Semantics: Yesterday, Today and Tomorrow", ICSC Keynote. 2021

Intro: Supervised Metric Learning

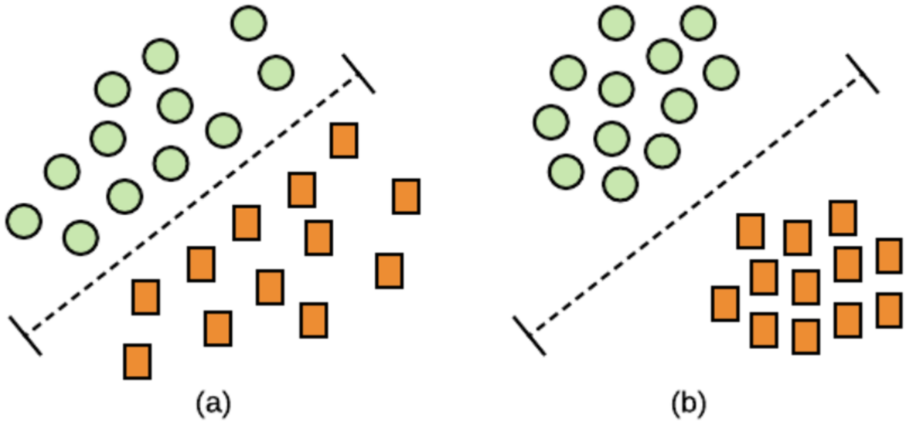


Input: Labeled, unlabeled, partially labeled

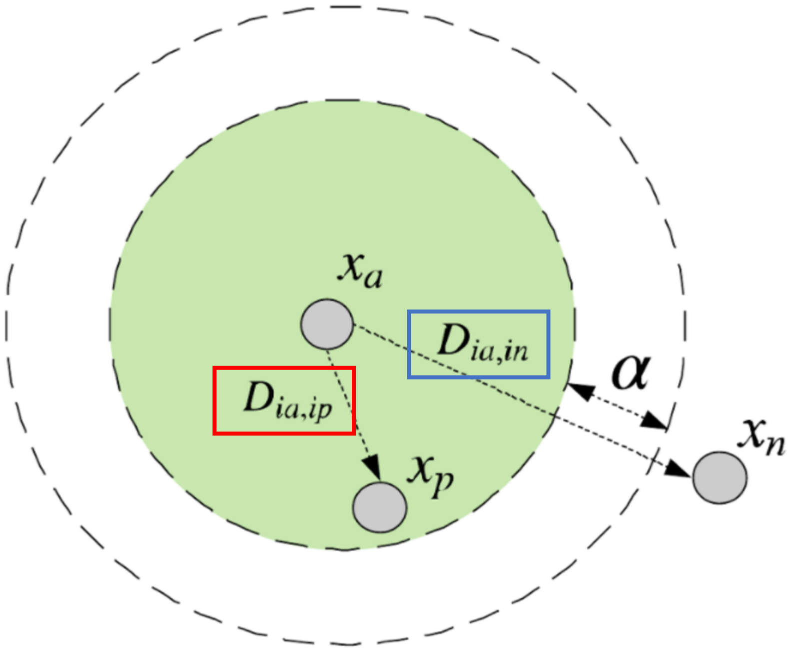
Various architectural design choices:
Siamese, U-Net, Skip-connections etc.,

Loss/Metric design

Intro: Contrastive Loss & Triplet Loss



(a) Separable Features (b) Discriminative Features

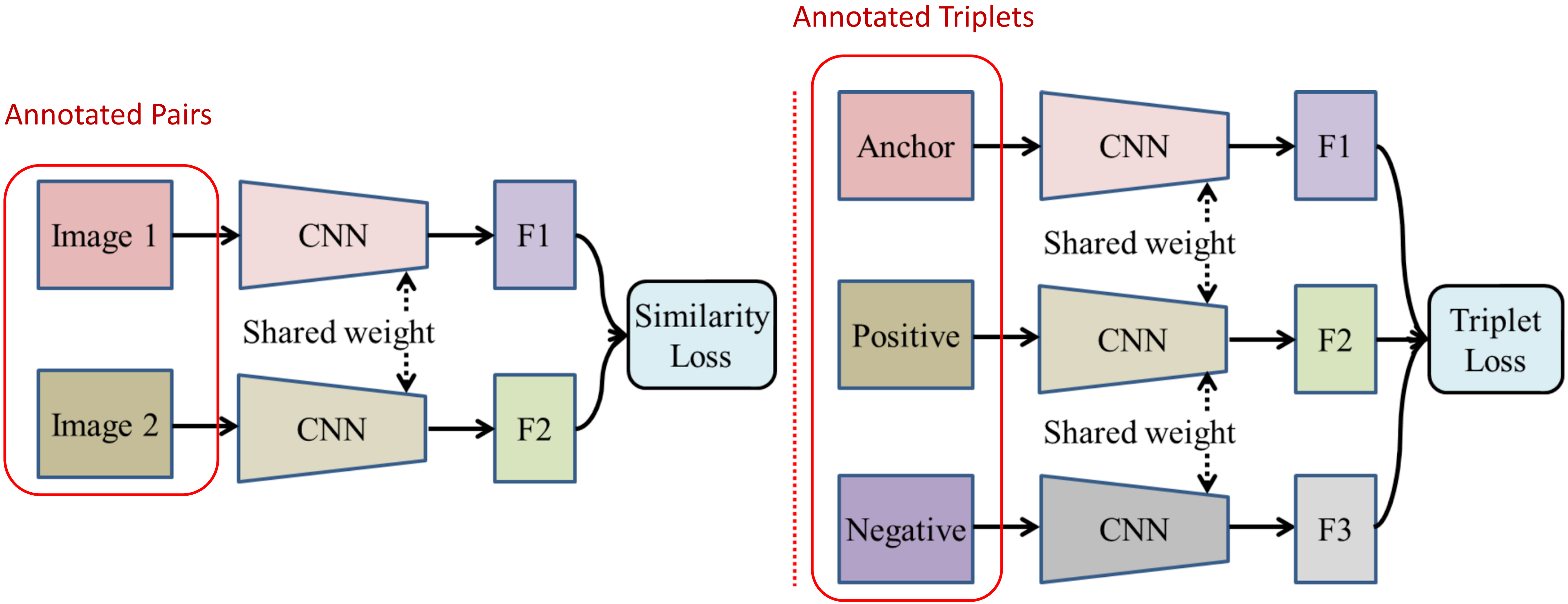


$$L_{contrastive} = [d_p - m_{pos}]_+ + [m_{neg} - d_n]_+$$

$$L_{triplet} = \boxed{[d_{ap}} - \boxed{d_{an}} + \alpha]_+$$

Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv, 2017.

Intro: Data Preparation



(a) Siamese Network

(b) Triplet Network

Motivation

- Existing methods require pairwise annotations

Same Class

Different Class

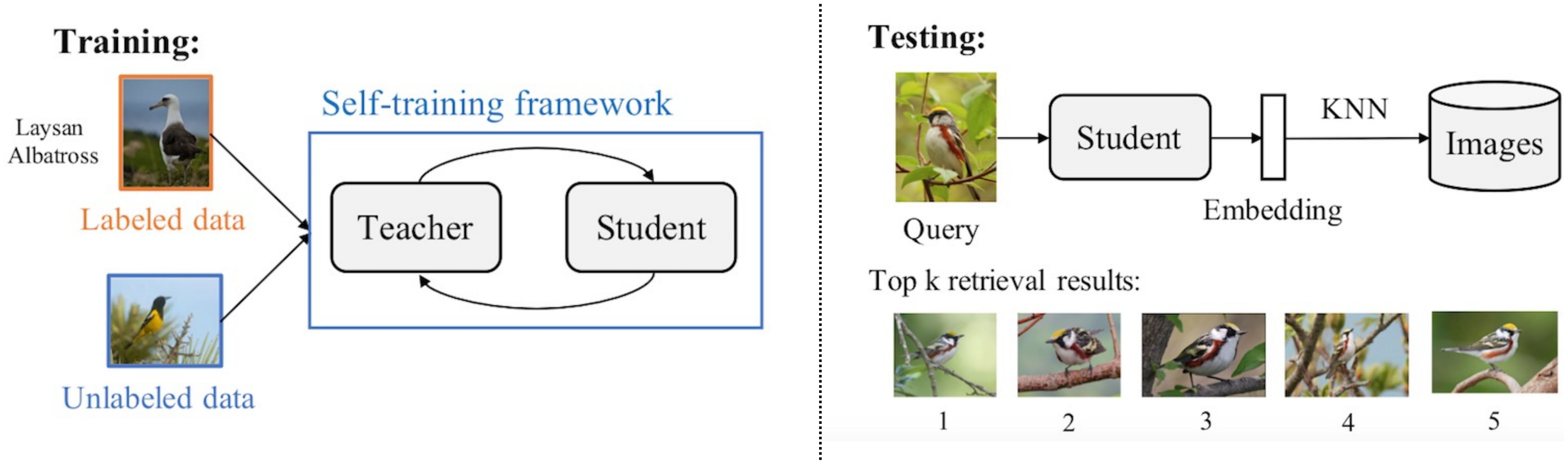
...

- Un-annotated data has not been leveraged

...

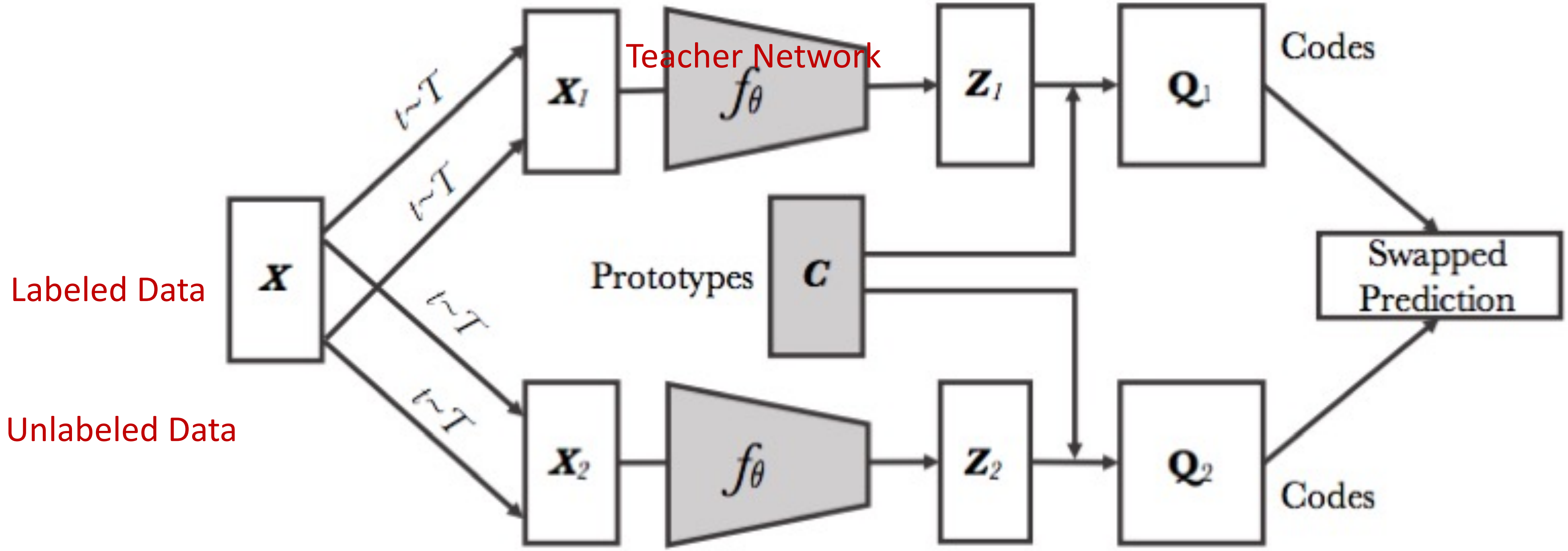
Goal: Leverage un-annotated data to improve deep metric learning

SLADE: A Self-Training Metric Learning Framework



J Duan, C.-C. Jay Kuo. SLADE: A Self-Training Framework For Distance Metric Learning. CVPR 2021.

Step 1: Self-supervised Teacher Pretraining

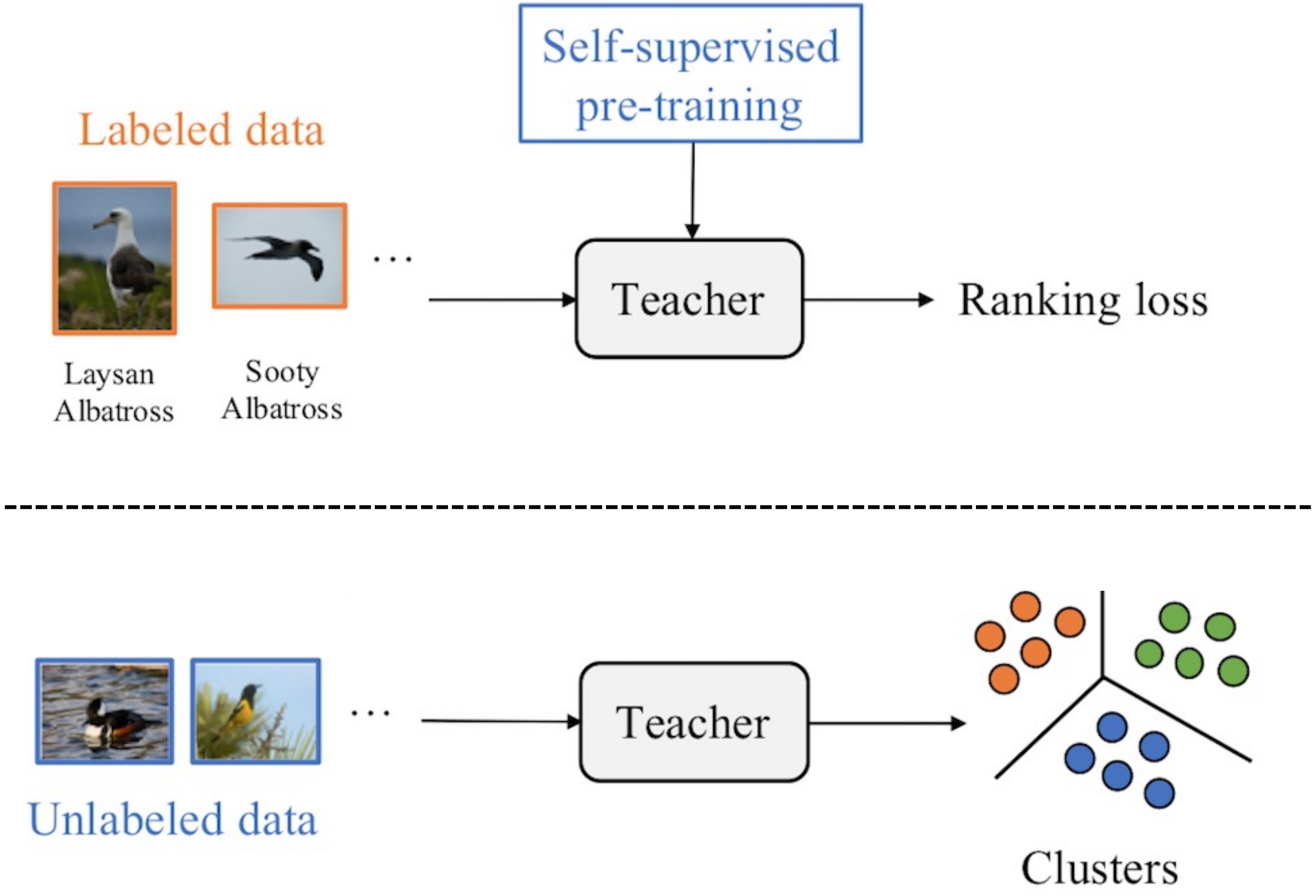


$$L(z_1, z_2) = l(z_1, q_2) + l(z_2, q_1)$$

$$L(z_1, z_2) = l(z_1, q_2) + l(z_2, q_1)$$

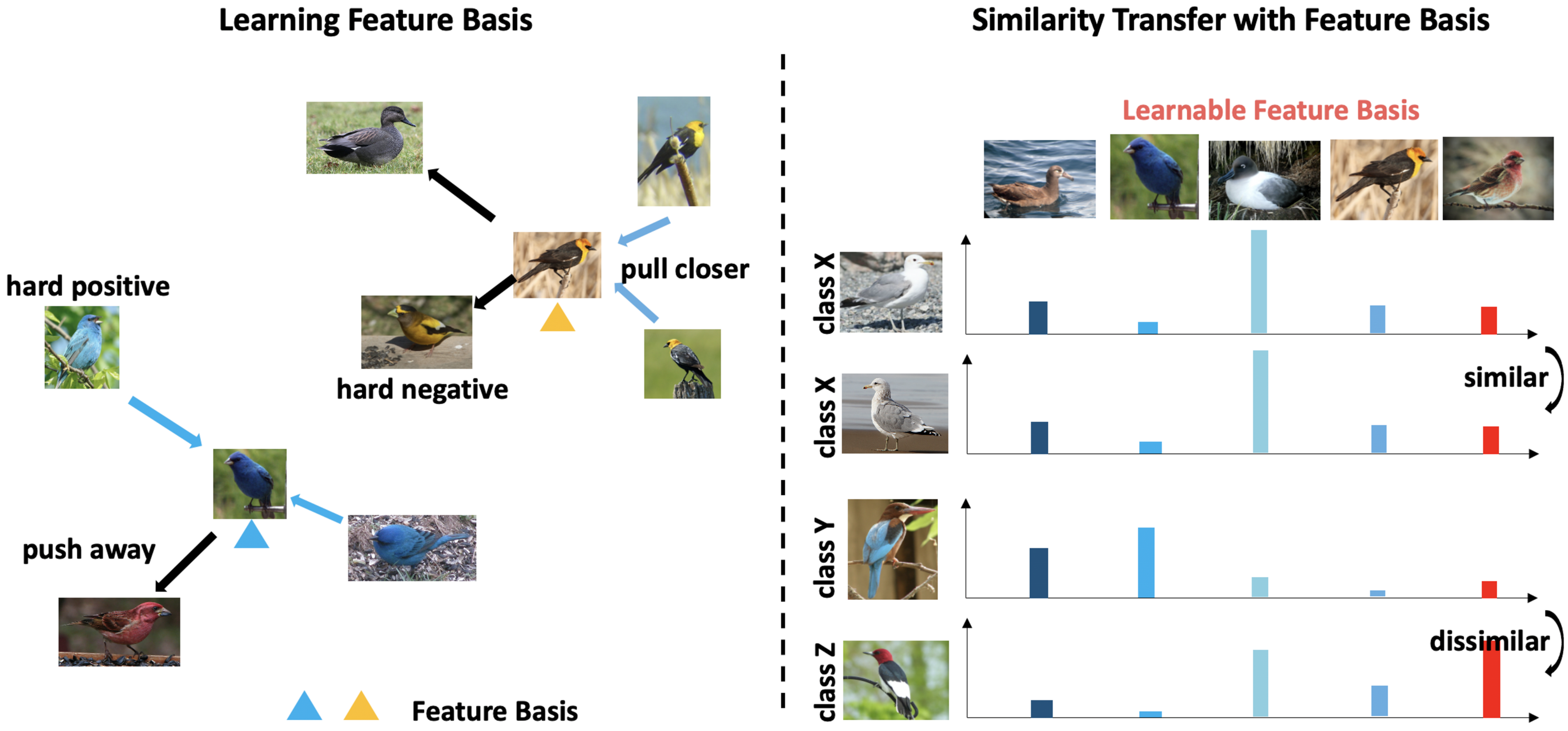
Mathilde Caron et al., Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. NeurIPS 2020.

Step 2: Pseudo Label Generation



Noisy!

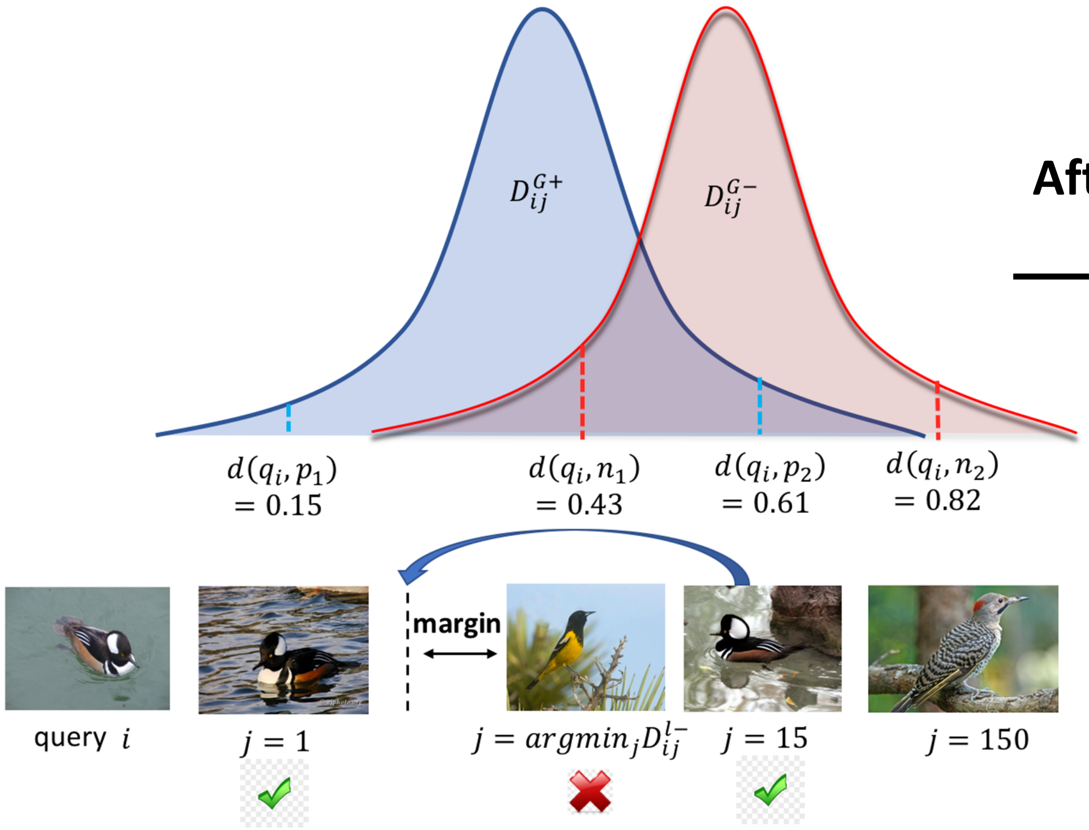
Step 3: Feature Basis Learning



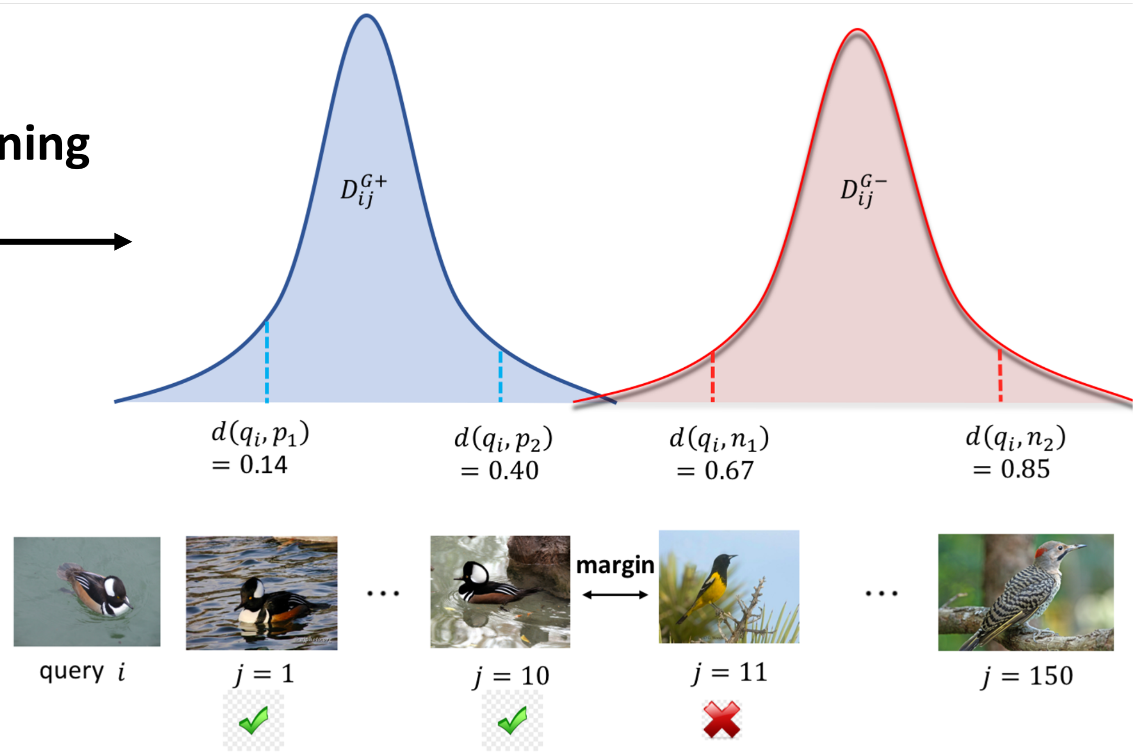
Step 3: Similarity Distribution Loss

Goal: reduce overlap between distributions

- Maximize distance between two means
- Reduce variances of two distributions



After Learning

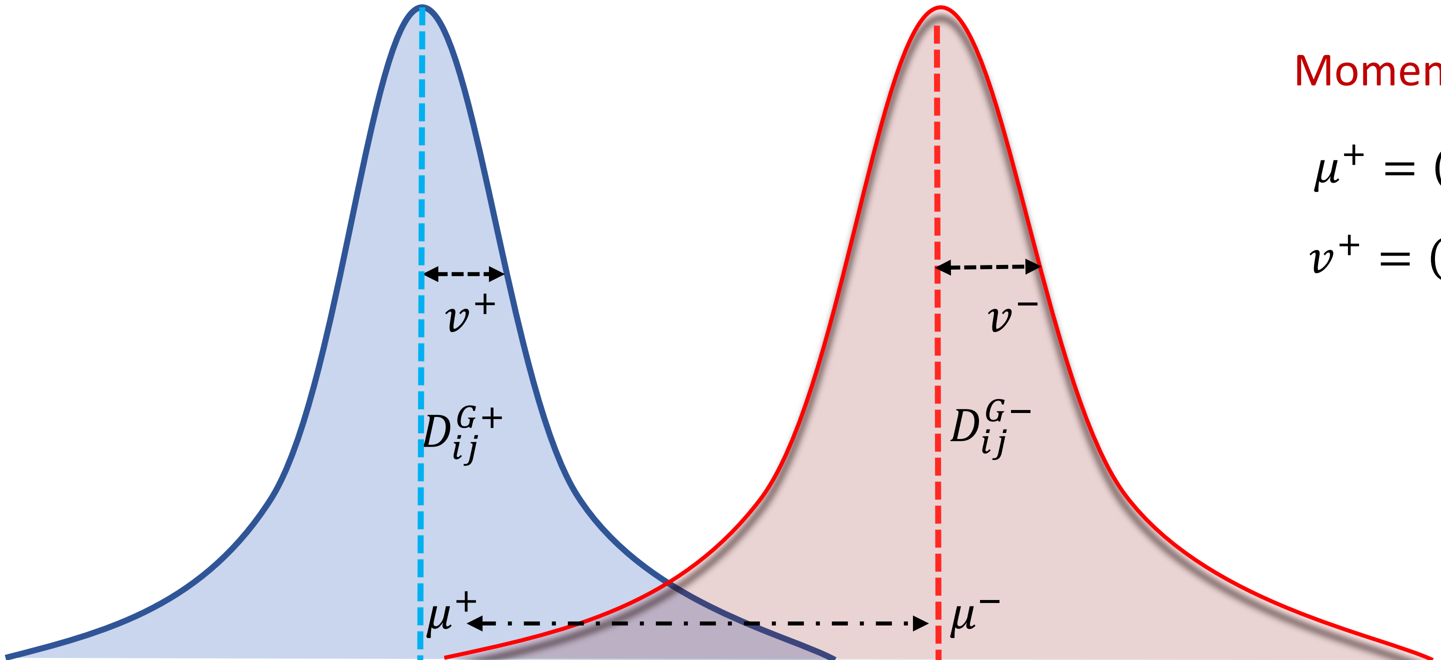


Step 3: Similarity Distribution Loss

Goal: reduce overlap between distributions

- Maximize distance between two means
- Reduce variances of two distributions

$$L_{SD}(G^+||G^-) = \max(\mu^- - \mu^+ + m, 0) + \lambda(v^+ + v^-)$$

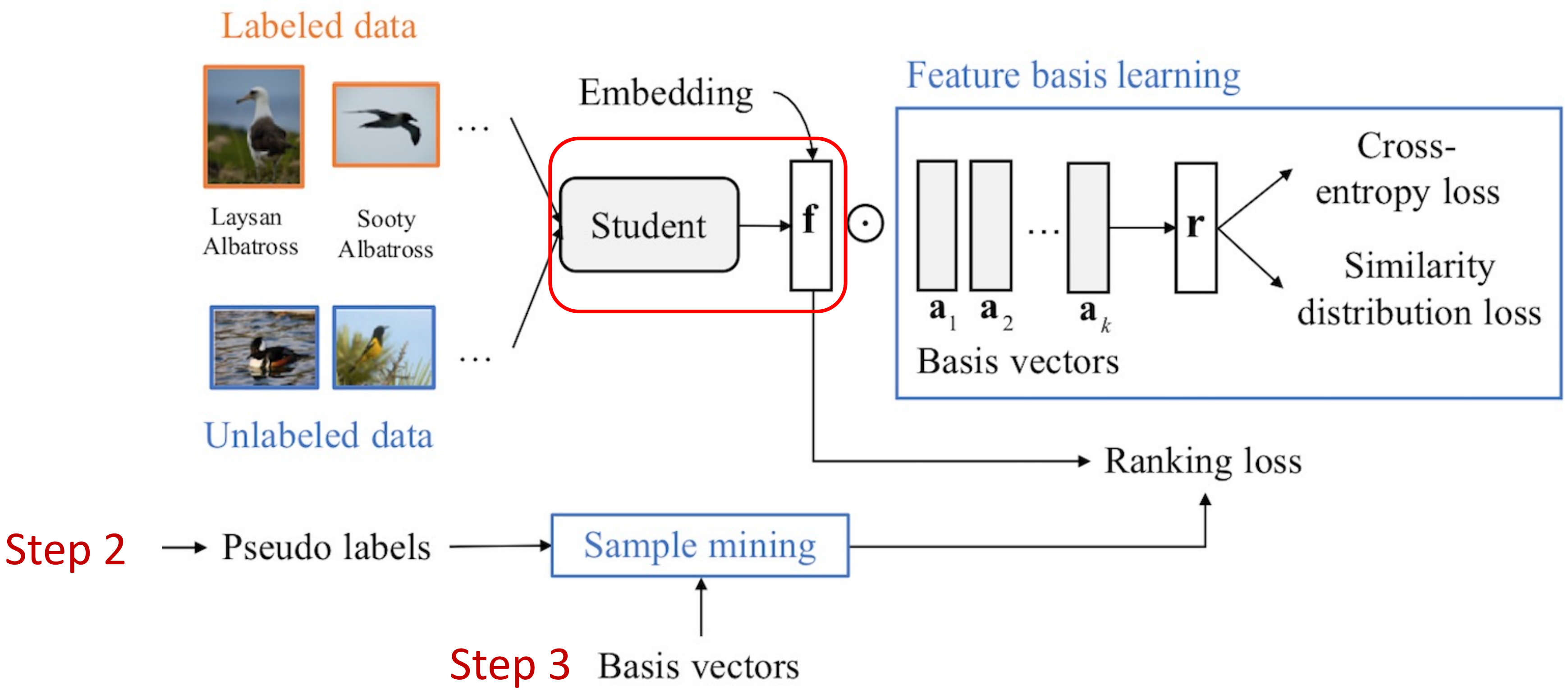


Momentum Update

$$\mu^+ = (1-\beta) \times \mu_b^+ + \beta \times \mu^+$$

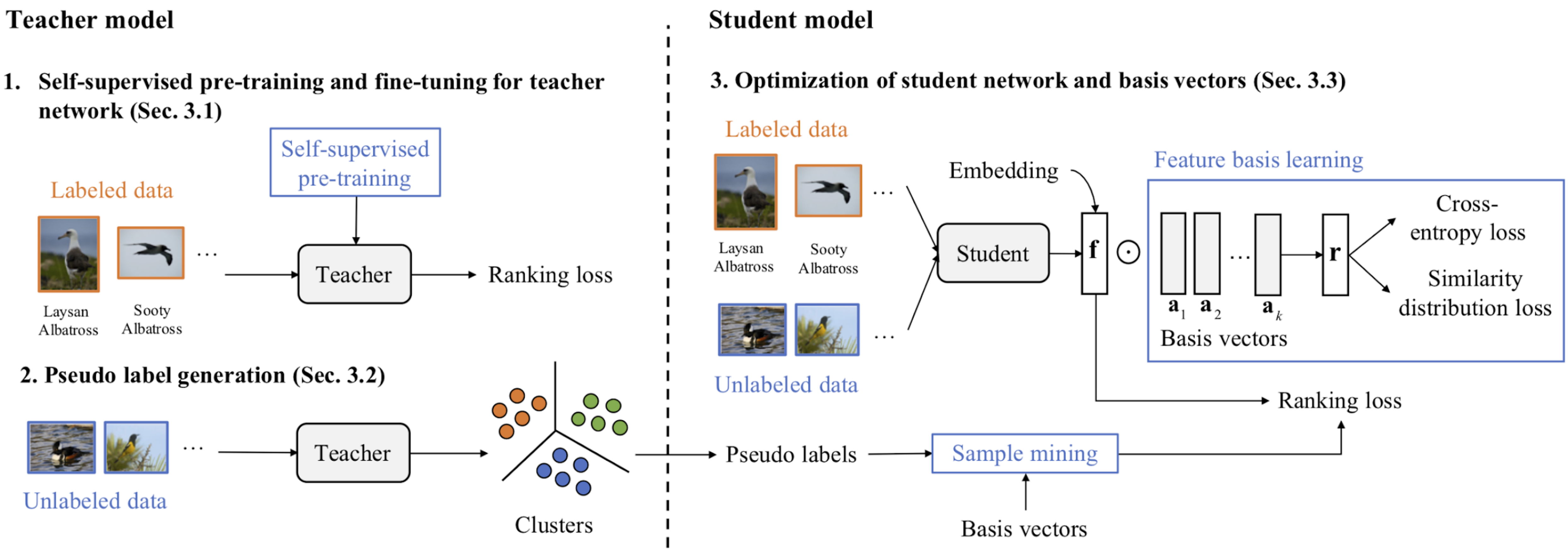
$$v^+ = (1-\beta) \times v_b^+ + \beta \times v^+$$

Training of Student Network



$$L = L_{rank}(D^l;\theta^s) + L_{rank}(D^u;\theta^s) + L_{basis}(D^l, D^u;\theta^s, W_a)$$

Putting It Together



J Duan, C.-C. Jay Kuo. SLADE: A Self-Training Framework For Distance Metric Learning. CVPR 2021.

Evaluation

CUB-200 (labeled): 200 species/ 12k images
NABIRDS (unlabeled): 400 species/ 48k images

Cars-196 (labeled): 196 brands/ 16k images
CompCars (unlabeled): 145 brands/ 16k images

In-shop (labeled): 8k instances/ 52k images
Fashion200k (unlabeled): 1k instances/200k images

Results: Performance Comparisons

Methods	Frwk	Init	Arc / Dim	CUB-200-2011			Cars-196		
				MAP@R	RP	P@1	MAP@R	RP	P@1
Contrastive [10]	[19]	ImageNet	BN / 512	26.53	37.24	68.13	24.89	35.11	81.78
Triplet [29]	[19]	ImageNet	BN / 512	23.69	34.55	64.24	23.02	33.71	79.13
ProxyNCA [18]	[19]	ImageNet	BN / 512	24.21	35.14	65.69	25.38	35.62	83.56
N. Softmax [35]	[19]	ImageNet	BN / 512	25.25	35.99	65.65	26.00	36.20	83.16
CosFace [25, 26]	[19]	ImageNet	BN / 512	26.70	37.49	67.32	27.57	37.32	85.52
FastAP [3]	[19]	ImageNet	BN / 512	23.53	34.20	63.17	23.14	33.61	78.45
MS+Miner [27]	[19]	ImageNet	BN / 512	26.52	37.37	67.73	27.01	37.08	83.67
Proxy-Anchor[1] [15]	[15]	ImageNet	R50 / 512	-	-	69.9	-	-	87.7
Proxy-Anchor[2] [15]	[19]	ImageNet	R50 / 512	25.56	36.38	66.04	30.70	40.52	86.84
ProxyNCA++ [22]	[22]	ImageNet	R50 / 2048	-	-	72.2	-	-	90.1
Mutual-Info [1]	[1]	ImageNet	R50 / 2048	-	-	69.2	-	-	89.3
Contrastive [10] (T_1)	[19]	ImageNet	R50 / 512	25.02	35.83	65.28	25.97	36.40	81.22
Contrastive [10] (T_2)	[19]	SwAV	R50 / 512	29.29	39.81	71.15	31.73	41.15	88.07
SLADE (Ours) (S_1)	[19]	ImageNet	R50 / 512	29.38	40.16	68.92	31.38	40.96	85.8
SLADE (Ours) (S_2)	[19]	SwAV	R50 / 512	**33.59**	**44.01**	**73.19**	**36.24**	**44.82**	**91.06**
MS [27] (T_3)	[19]	ImageNet	R50 / 512	26.38	37.51	66.31	28.33	38.29	85.16
MS [27] (T_4)	[19]	SwAV	R50 / 512	29.22	40.15	70.81	33.42	42.66	89.33
SLADE (Ours) (S_3)	[19]	ImageNet	R50 / 512	30.90	41.85	69.58	32.05	41.50	87.38
SLADE (Ours) (S_4)	[19]	SwAV	R50 / 512	**33.90**	**44.36**	**74.09**	**37.98**	**46.92**	**91.53**

J Duan, C.-C. Jay Kuo. SLADE: A Self-Training Framework For Distance Metric Learning. CVPR 2021.

Ablation Studies

Pre-trained weight	MAP@R	
	CUB-200	Cars-196
ImageNet [8]	29.38	31.38
Pre-trained SwAV [4]	32.79	35.54
Fine-tuned SwAV	33.59	36.24

Verification of the effect of Step 1

Regularization	CUB-200		
	MAP@R	RP	P@1
Local-CE	32.69	43.20	72.64
Global-CE	32.23	42.68	72.45
SD (Ours)	33.59	44.01	73.19

Effect of the proposed similarity distribution loss

Components	MAP@R	
	CUB-200	Cars-196
Teacher (contrastive)	29.29	31.73
Student (pseudo label)	30.81	31.99
+ Basis	32.45	35.78
+ Basis + Mining	33.59	36.24

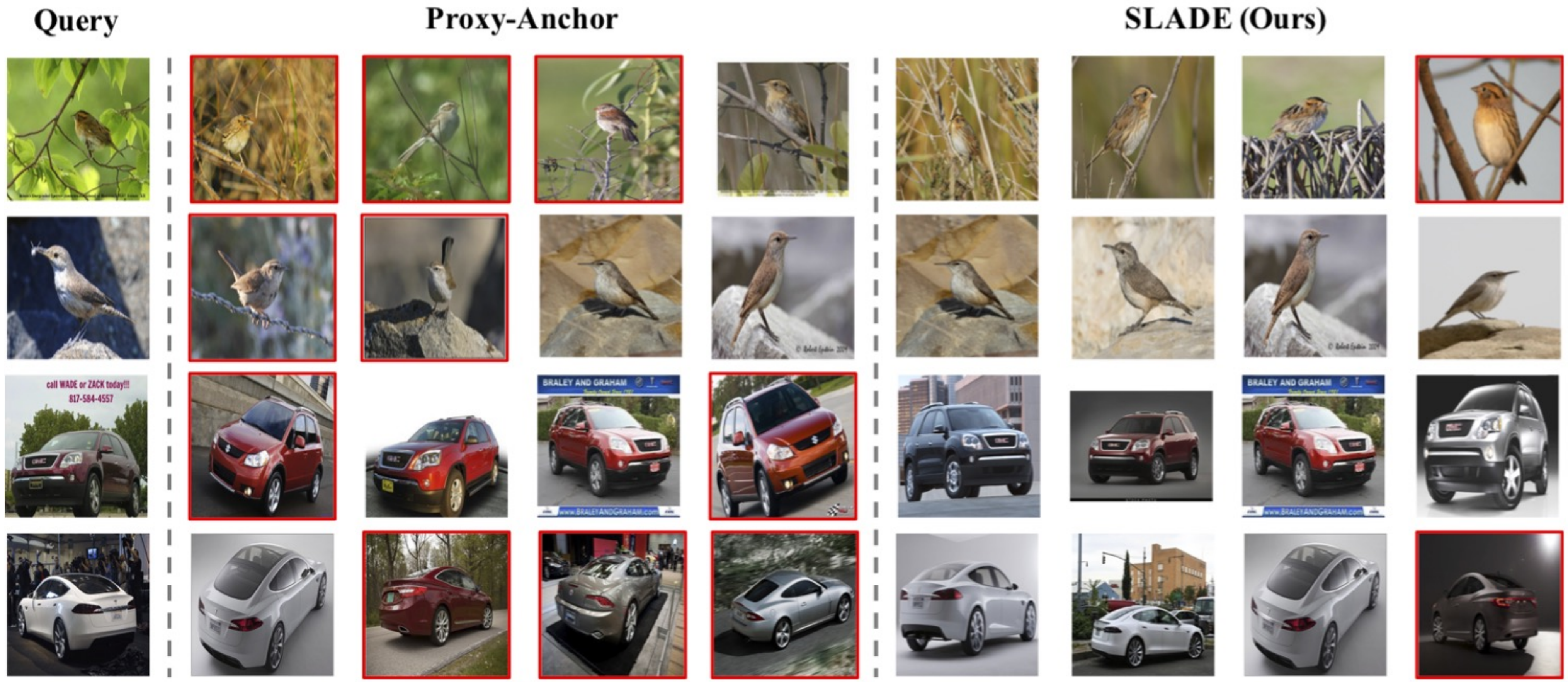
Verification proposed components in Step 3

k	NABirds		
	MAP@R	RP	P@1
100	31.83	42.25	72.19
200	32.61	43.02	72.75
300	32.81	43.18	72.21
400	33.59	44.01	73.19
500	33.26	43.69	73.26

Robustness to hyper-parameters

J Duan, C.-C. Jay Kuo. SLADE: A Self-Training Framework For Distance Metric Learning. CVPR 2021.

Qualitative Results



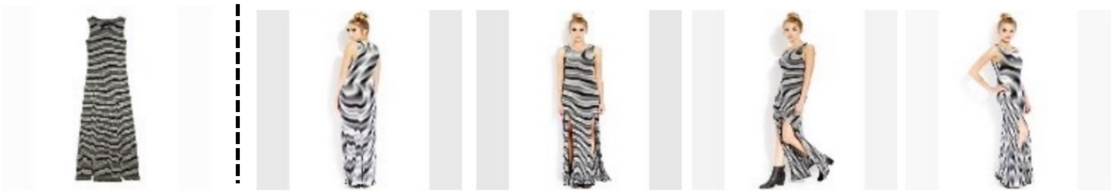
Comparison with SOTA [1] on CUB200 & Cars-196, more at BIRD & CARS

1. Kim S et al. Proxy anchor loss for deep metric learning. CVPR 2020

A Practical Need from Industry

- Existing search queries

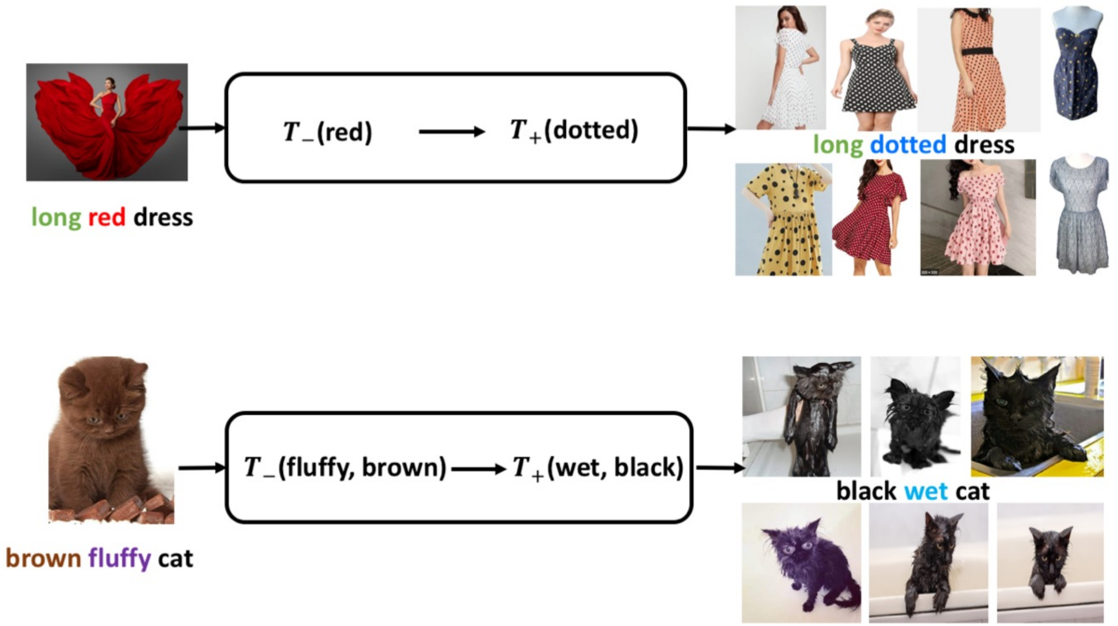
Image-to-Image search



Text-to-Image search

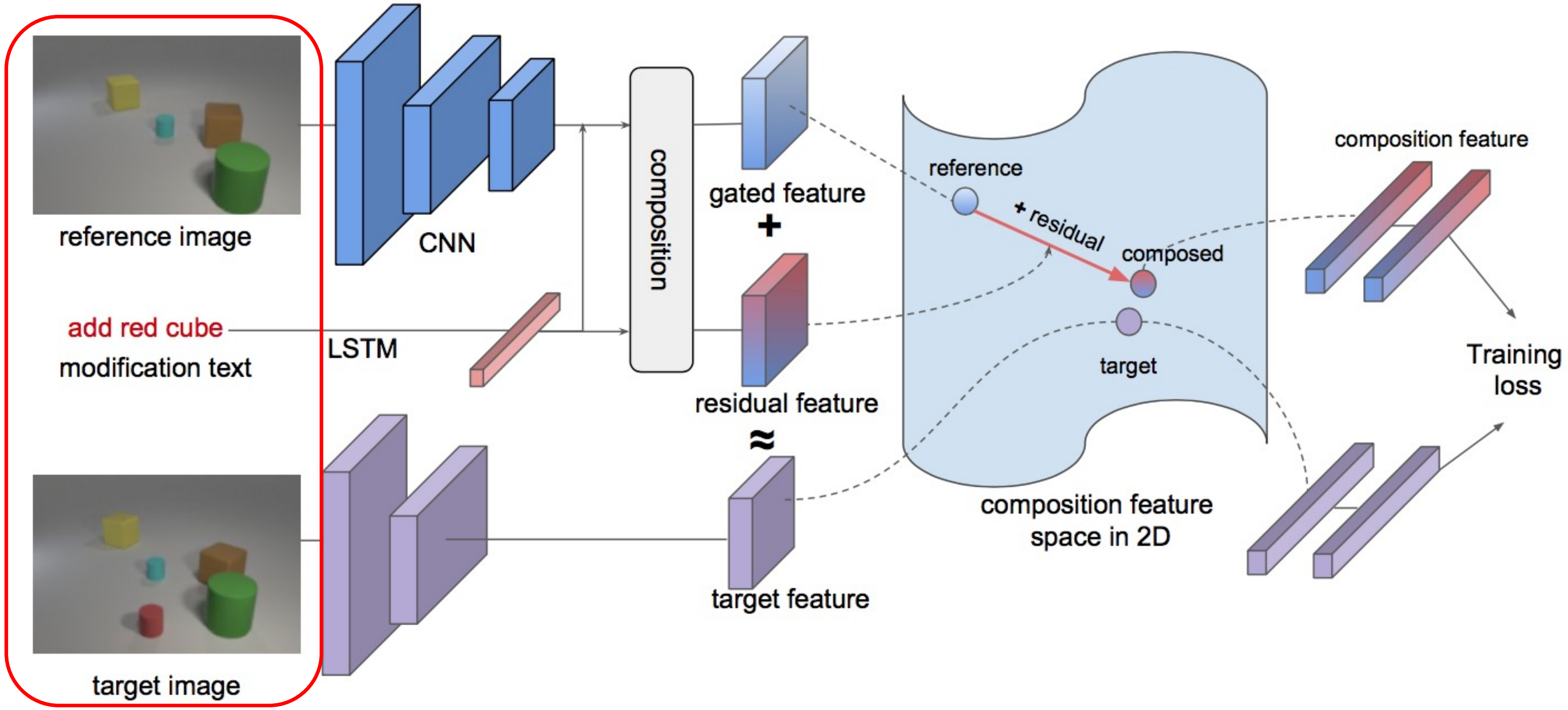
Blue,
Cloudy
Sky

- Image-attribute queries



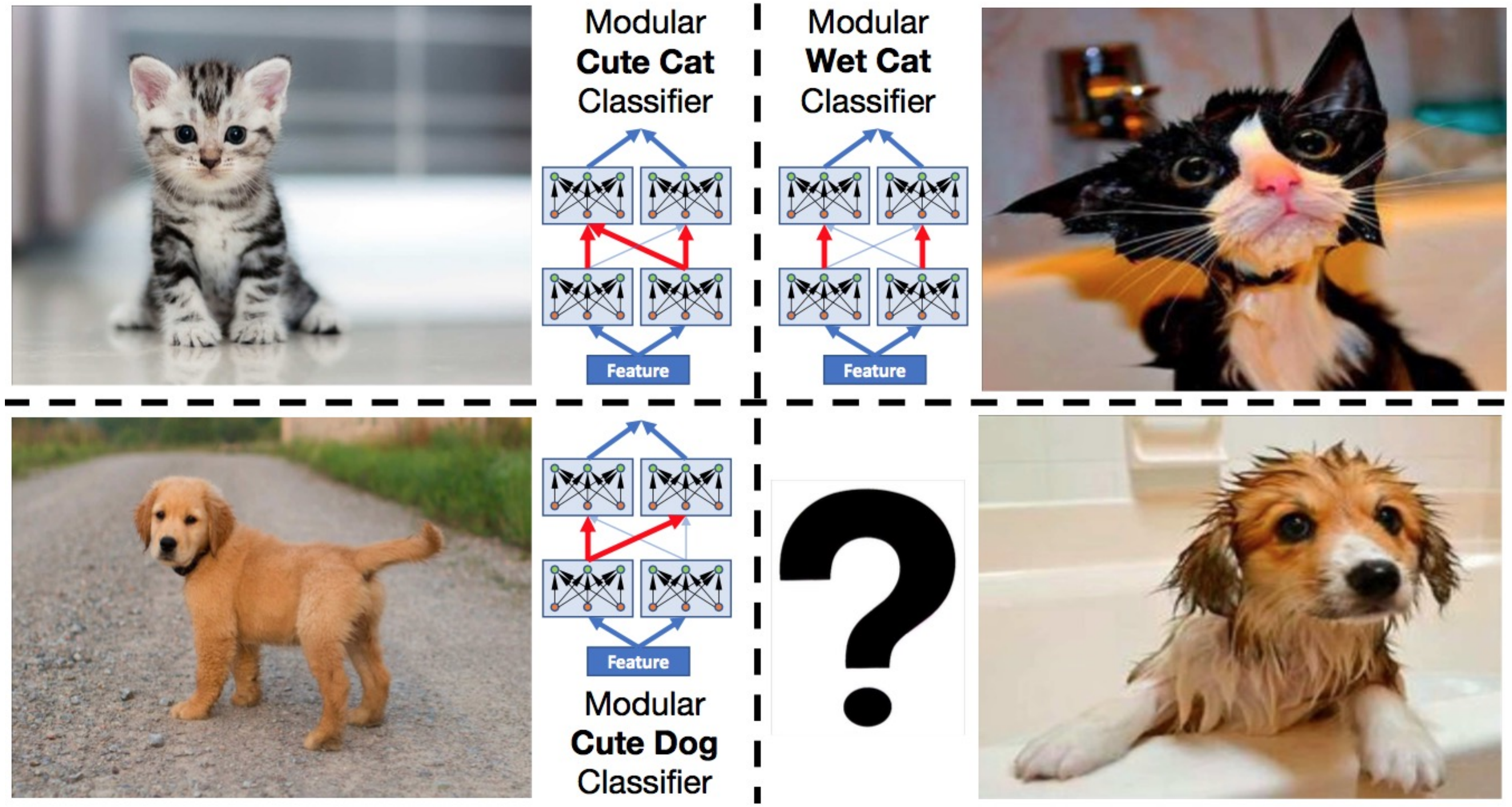
Related Work: Image-text Composition

Annotation Expensive



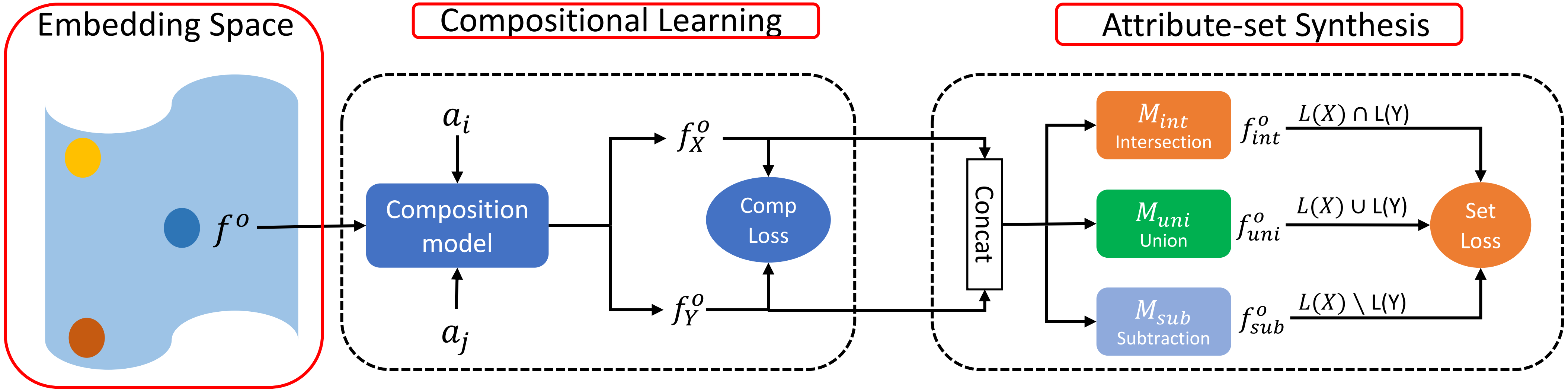
Li Fei-Fei et al., Composing Text and Image for Image Retrieval - An Empirical Odyssey . CVPR 2019.

Related Work: Zero-shot Classification



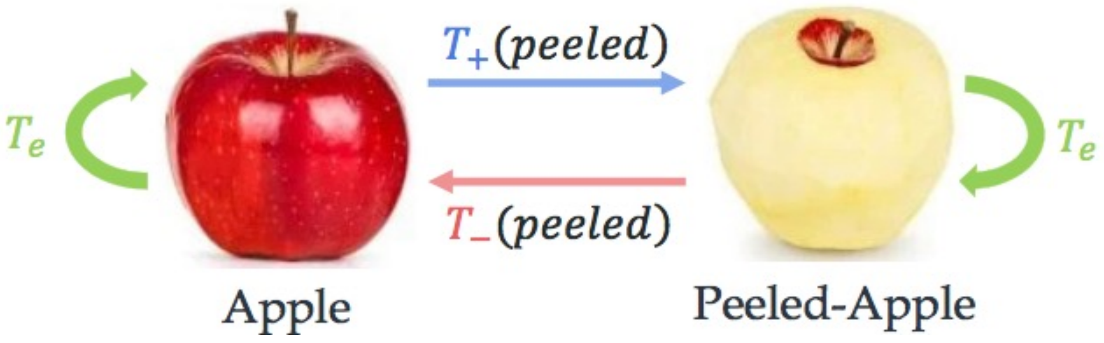
Senthil et al., Task-Driven Modular Networks for Zero-Shot Compositional Learning . ICCV 2019.

Compositional Attribute-based Metric Learning



J Duan, C.-C. Jay Kuo. Compositional Learning for Weakly Attribute-based Metric Learning. In Submission.

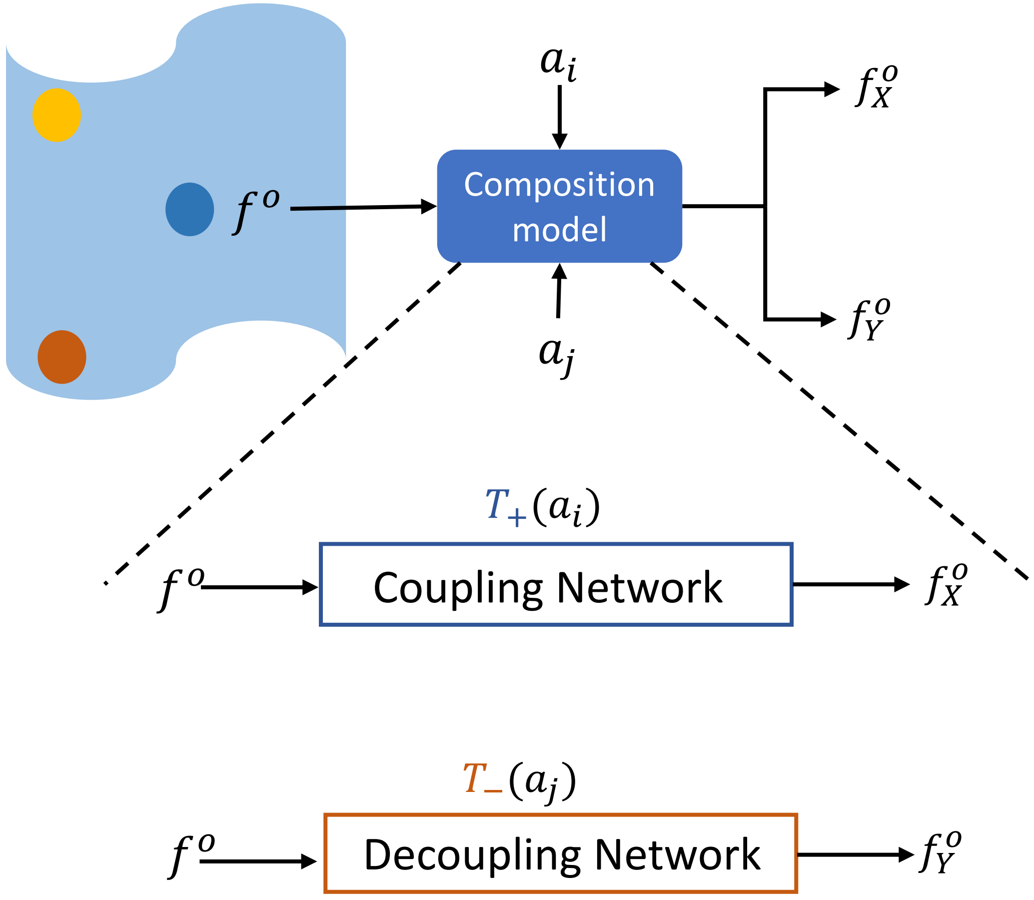
Compositional Module



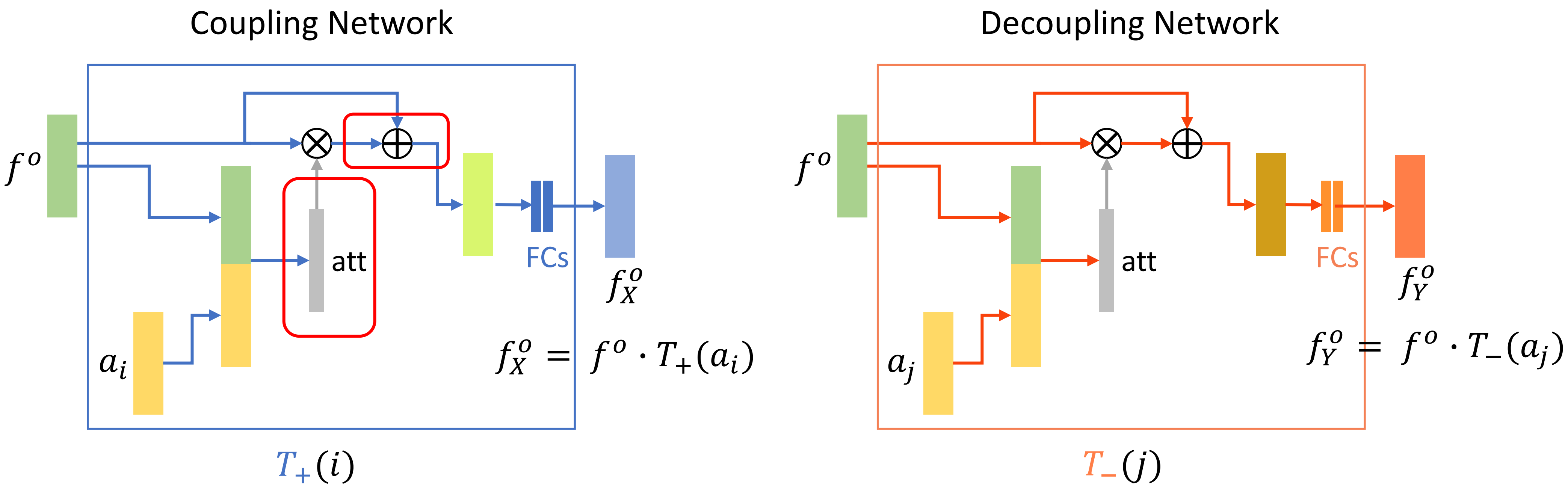
Attribute-Object Factorization

$$f_{red-apple} \cdot T_+\ (peeled) \rightarrow f_{peeled-apple}$$

$$f_{peeled-apple} \cdot T_-\ (peeled) \rightarrow f_{red-apple}$$



Implementation of Compositional Operations



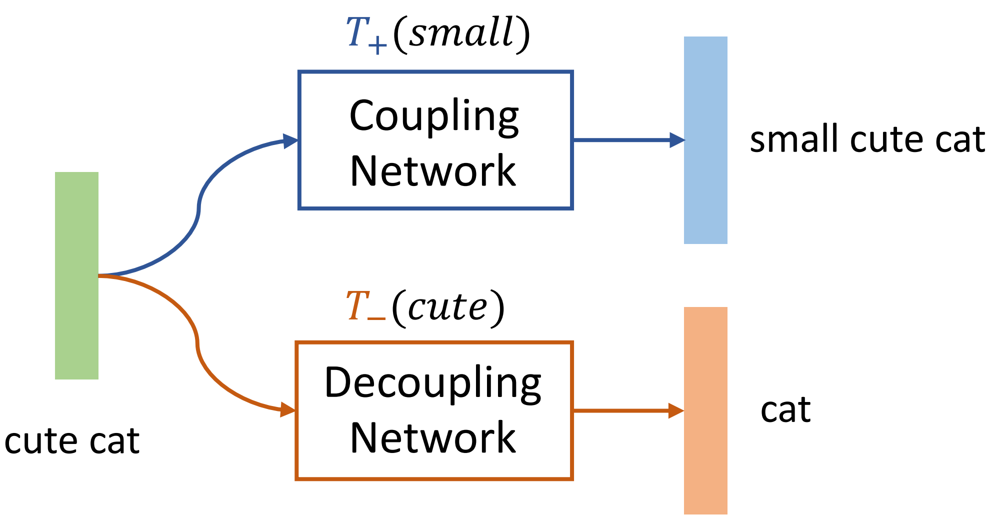
$$f^o_{gated} = \sigma(W_2 * RELU(W_1 * [f^o, a_i]))\odot f^o$$

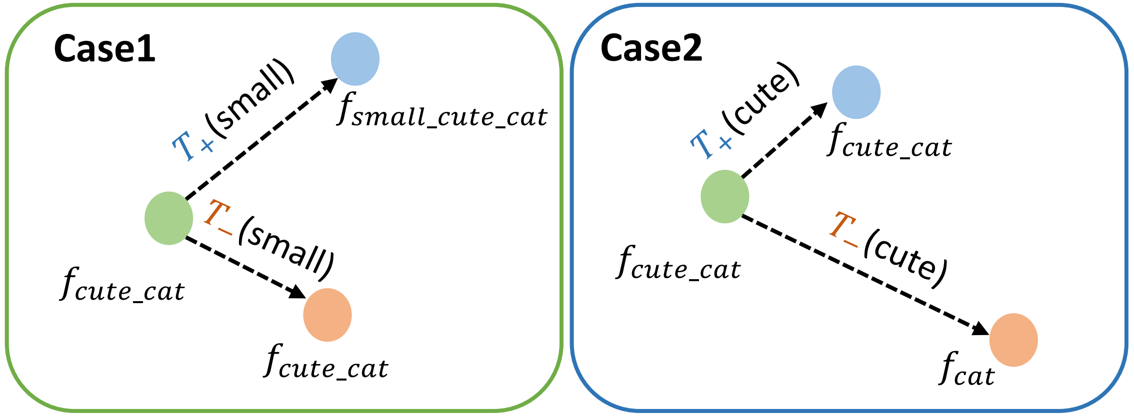
$$f_X^o = W_4 * RELU(W_3 * \boxed{(f^o + f^o_{gated})})$$

Compositional Loss Design

Metric Learning objective:

$$L_{metric} = \max(0, \text{m-d}(f^{o} \cdot T_{+}(a_i), f^{o'} \cdot T_{+}(a_j)), \forall o \neq o', \forall a_i, a_j \in \mathcal{A}$$





Relativity objective:

$$L_{rel} = \boxed{\max(0, \text{d}(f^{o}, f^{o} \cdot T_{-}(a_j)) - d(f^{o}, f^{o} \cdot T_{+}(a_j)) + \text{m})} + \boxed{\max(0, \text{d}(f^{o}, f^{o} \cdot T_{+}(a_i)) - d(f^{o}, f^{o} \cdot T_{-}(a_i)) + \text{m})}$$

Regularization Objectives

Commutativity:

$$L_{com} = ||f^o \cdot T_+(a_i) \cdot T_-(a_j) - f^o \cdot T_-(a_j) \cdot T_+(a_i)||_2$$

Invertibility:

$$L_{inv} = ||f^o \cdot T_+(a_i) \cdot T_-(a_i) - f^o||_2 + ||f^o \cdot T_-(a_i) \cdot T_+(a_i) - f^o||_2$$

Classification of objects and attributes:

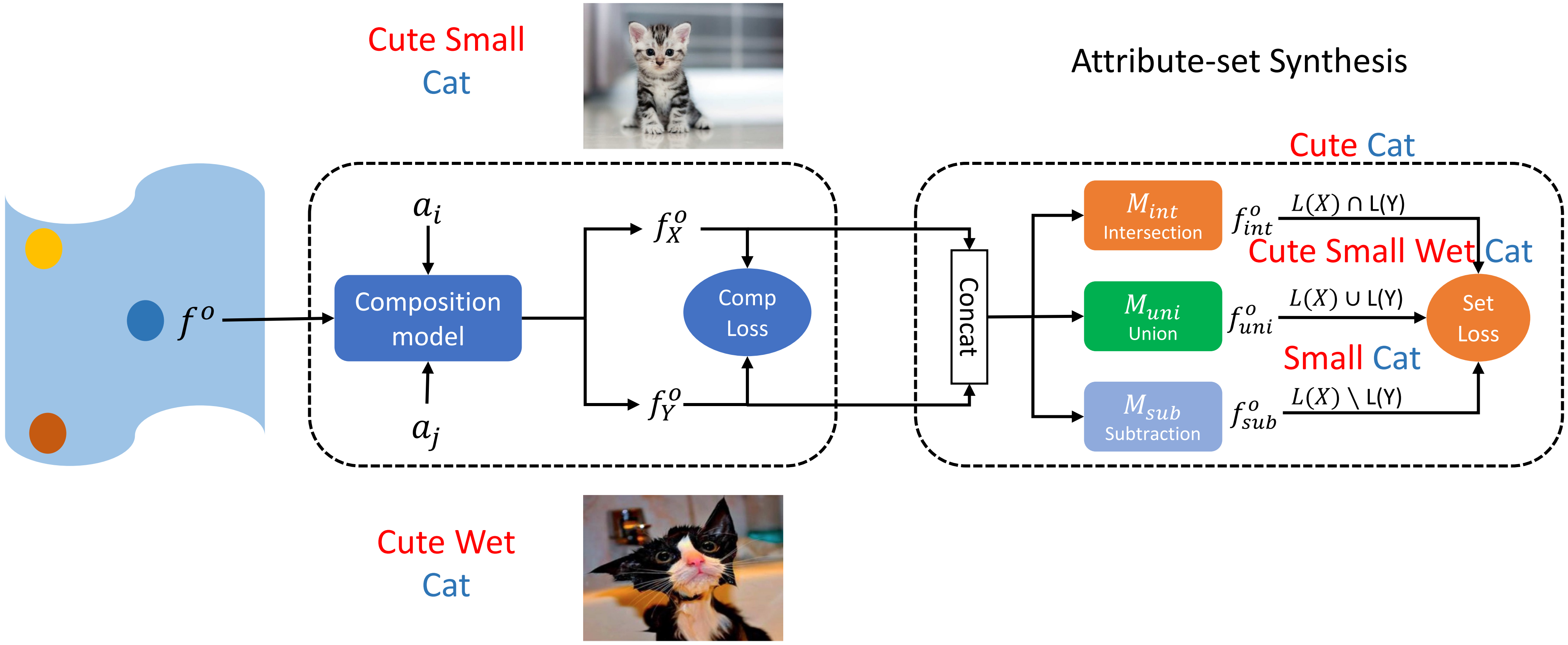
$$L_{cls} = BCE(f_X^o, L(X)) + BCE(f_Y^o, L(Y))$$

Final objective:

$$L_{comp} = \alpha_1 L_{metric} + \alpha_2 L_{rel} + \alpha_3 L_{com} + \alpha_4 L_{inv} + \alpha_5 L_{cls}$$

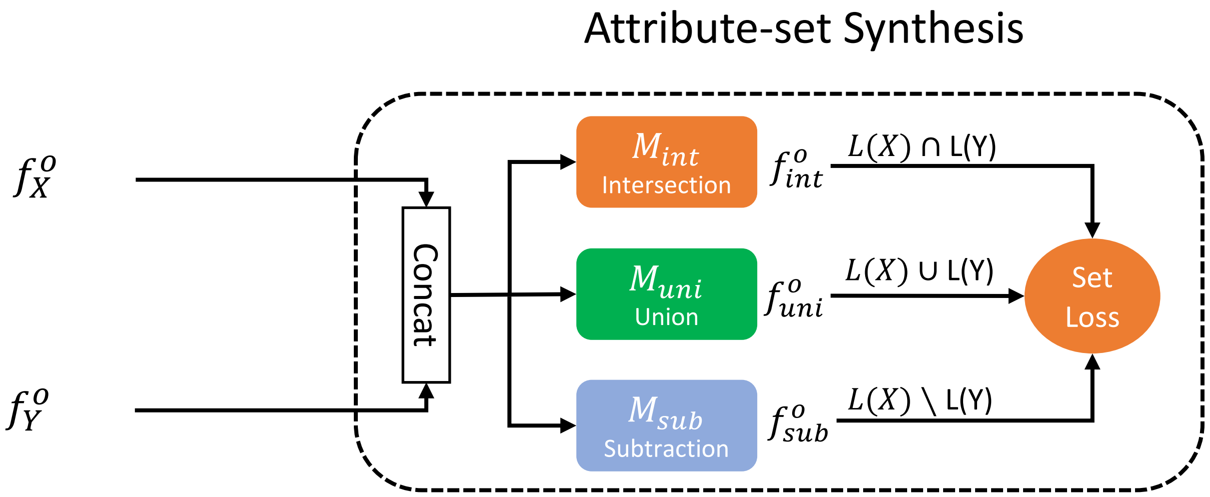
Kristen Grauman et al. "Attributes as operators: factorizing unseen attribute-object compositions." ECCV 2018.

Weakly-supervised Attribute Synthesis



Cats are used for illustration purposes, similar to ICCV19.

Attribute Set Operation Loss



$$f_{int}^o = M_{int}([f_X^o, f_Y^o])$$

$$f_{uni}^o = M_{uni}([f_X^o, f_Y^o])$$

$$f_{sub}^o = M_{sub}([f_X^o, f_Y^o])$$

Set Loss:

$$L_{int} = BCE(f_{int}^o, L(X) \cap L(Y))$$

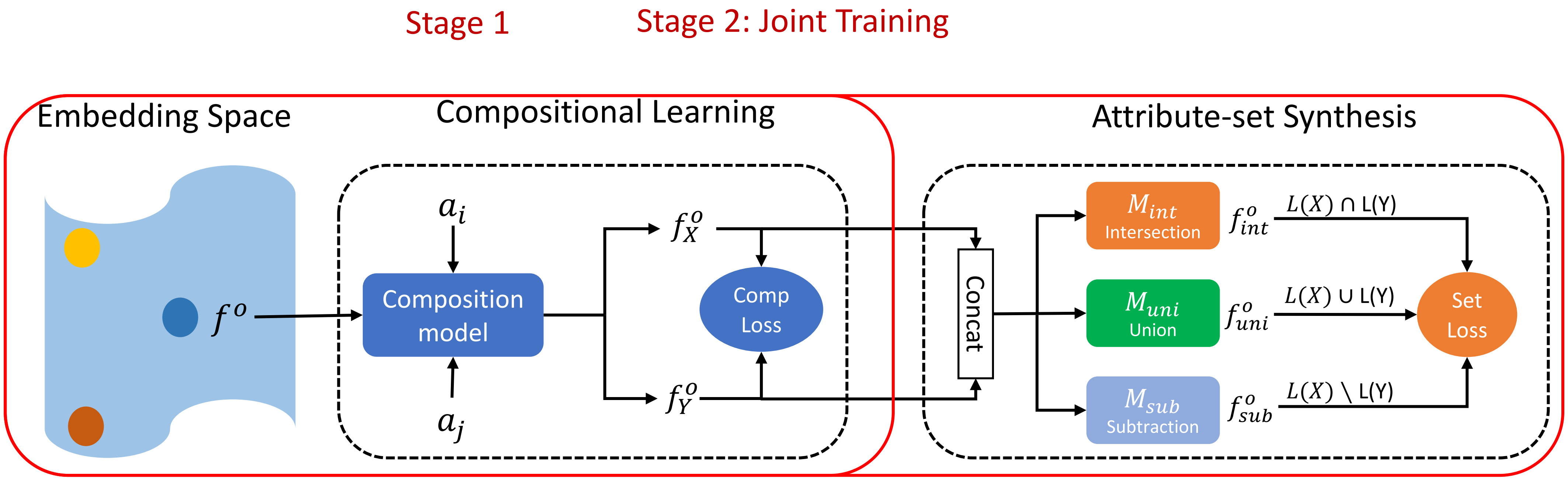
$$L_{uni} = BCE(f_{uni}^o, L(X) \cup L(Y))$$

$$L_{sub} = BCE(f_{sub}^o, L(X) \setminus L(Y))$$

Final objective:

$$L_{set} = \beta_1 L_{int} + \beta_2 L_{uni} + \beta_3 L_{sub}$$

Training

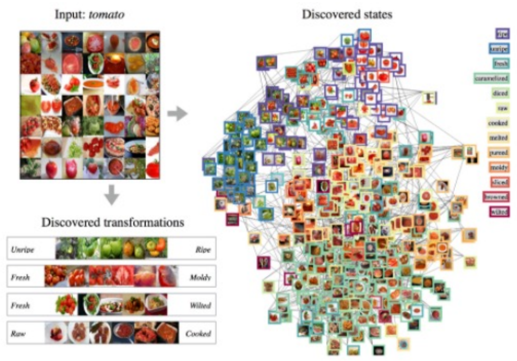


$$L = \lambda_1 L_{comp} + \lambda_2 L_{set}$$

J Duan, C.-C. Jay Kuo. Compositional Learning for Weakly Attribute-based Metric Learning. In Submission.

Evaluation

Zero-shot Retrieval: Top-k accuracy



MIT-States: 34,562 training / 19,191 test

UT-Zappos50k: 24,898 training/4,228 test

Multi-label Retrieval: Recall@K

Fashion200k: 200k images, 172k training/31,670 test

Quantitative Results

Method	MIT-States			UT-Zappos		
	Top-1	Top-2	Top-3	Top-1	Top-2	Top-3
Visual Product [23]	9.8/13.9*	16.1	20.6	49.9*	/	/
LabelEmbed (LE) [23]	11.2/13.4*	17.6	22.4	25.8*	/	/
- LEOR [23]	4.5	6.2	11.8	/	/	/
- LE + R [23]	9.3	16.3	20.8	/	/	/
- LabelEmbed+ [27]	14.8*	/	/	37.4*	/	/
AnalogousAttr [1]	1.4	/	/	18.3	/	/
Red Wine [23]	13.1	21.2	27.6	40.3	/	/
AttOperator [27]	14.2	19.6	25.1	46.2	56.6	69.2
TAFE-Net [43]	16.4	26.4	33.0	33.2	/	/
GenModel [27]	17.8	/	/	48.3	/	/
TIRG [40]	12.2	/	/	/	/	/
WAML (Ours)	**18.0**	**27.2**	**33.6**	**50.6**	**67.8**	**76.6**

Zero-shot retrieval results on MIT-States and UT-Zappos

Method	R@1	R@10	R@50
Image only [40]	3.5	22.7	43.7
Text only [40]	1.0	12.3	21.8
Concatenation [40]	$11.9^{\pm1.0}$	$39.7^{\pm1.0}$	$62.6^{\pm0.7}$
TIRG [40]	$14.1^{\pm0.6}$	$42.5^{\pm0.7}$	$63.8^{\pm0.8}$
Han *et al.* [7]	6.3	19.9	38.3
Show and Tell [39]	$12.3^{\pm1.1}$	$40.2^{\pm1.7}$	$61.8^{\pm0.9}$
Param Hashing [28]	$12.2^{\pm1.1}$	$40.0^{\pm1.1}$	$61.7^{\pm0.8}$
Relationship [35]	$13.0^{\pm0.6}$	$40.5^{\pm0.7}$	$62.4^{\pm0.6}$
FiLM [32]	$12.9^{\pm0.7}$	$39.5^{\pm2.1}$	$61.9^{\pm1.9}$
WAML	$\mathbf{22.8}^{\pm0.7}$	$\mathbf{50.3}^{\pm0.8}$	$\mathbf{71.6}^{\pm0.7}$

Retrieval performance on Fashion200k

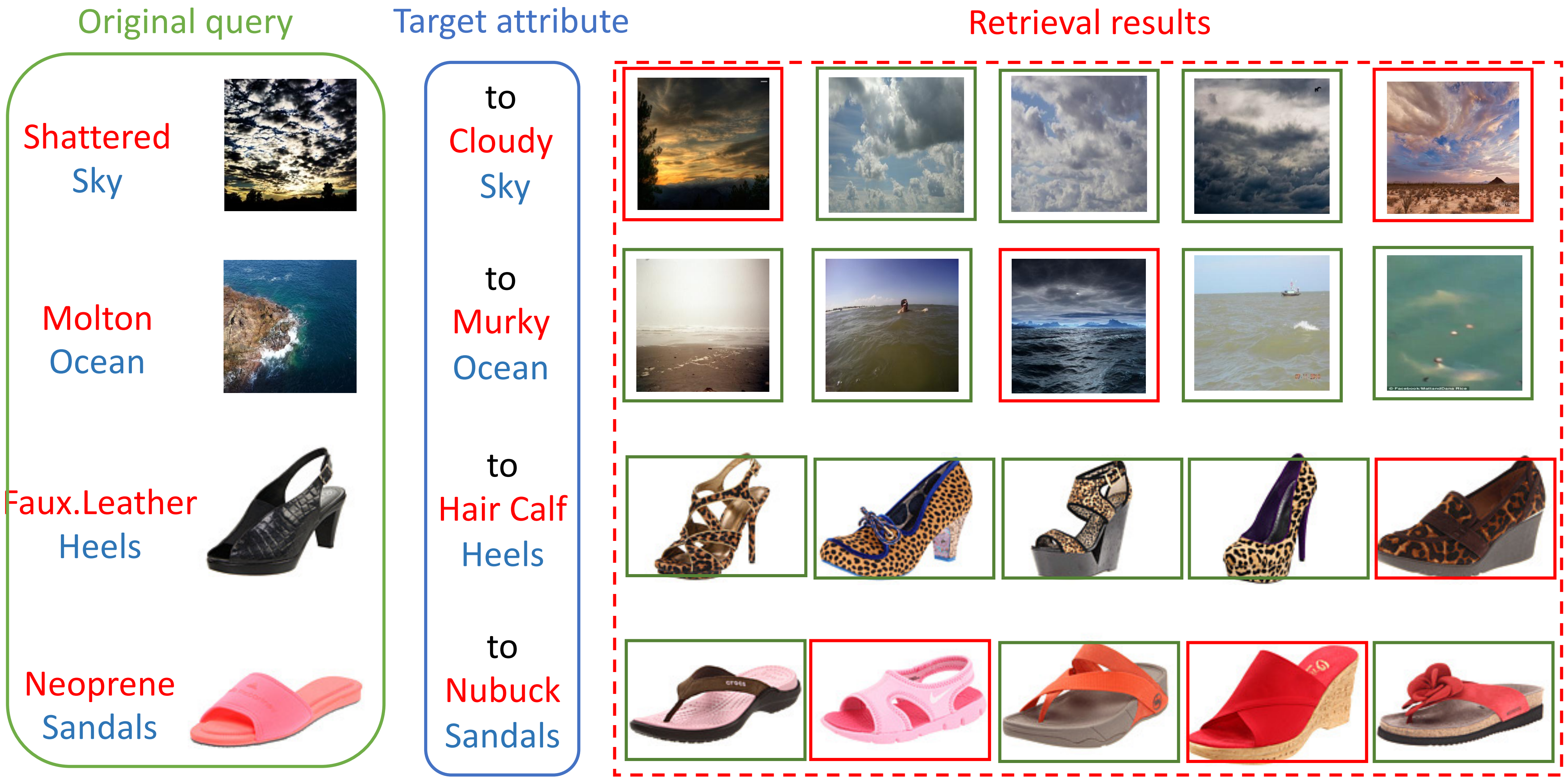
Method	MIT-States		UT-Zappos	
	Attribute	Object	Attribute	Object
AttrOperator [27]	14.6	20.5	29.7	67.5
GenModel [27]	15.1	27.7	18.4	68.1
WAML	**18.6**	**28.0**	**37.6**	66.2

Object-attribute recognition results on two benchmarks

Method	MIT-States			UT-Zappos		
	Top-1	Top-2	Top-3	Top-1	Top-2	Top-3
WAML	**18.0**	**27.2**	**33.6**	**50.6**	**67.8**	**76.6**
WAML w/o $\mathcal{L}_{comp}$	9.8	17.3	23.0	20.2	40.0	51.0
WAML w/o $\mathcal{L}_{set}$	16.9	24.5	30.9	47.6	64.4	73.0
WAML $\mathcal{L}_{metric}$ only	12.1	22.7	29.0	41.2	55.0	64.8
WAML Cos dist.	17.5	26.1	31.7	48.6	66.9	76.2

Ablation studies on different components

Qualitative Results



J Duan, C.-C. Jay Kuo. Compositional Learning for Weakly Attribute-based Metric Learning. In Submission.

- Challenge 2: Deep learning has mostly been used as black box
    - **Can** we enforce certain structures?
    - **How** do we learn more structured features?

- Learn structured representation via metric learning
- Model contextual information with graph structure
- Scale metric learning with a self-training framework
- Improve generalization by learning attribute compositions