Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 10.2: New Directions

Louis-Philippe Morency

Objectives of today's class

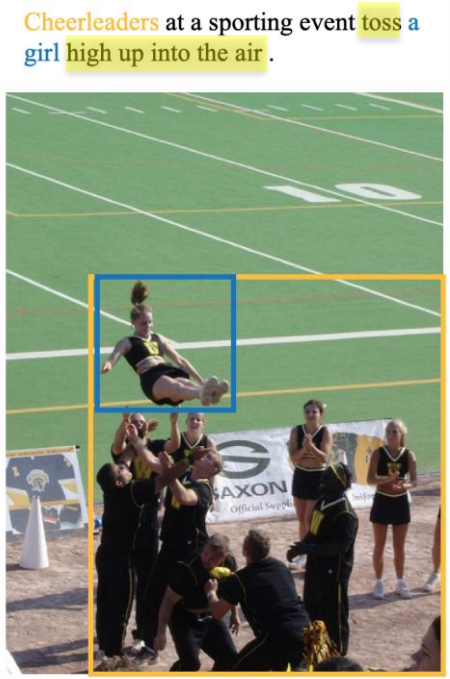
- New research directions in multimodal ML
  - Alignment
  - Representation
  - Fusion
  - Translation
  - Co-learning

New Directions: Alignment

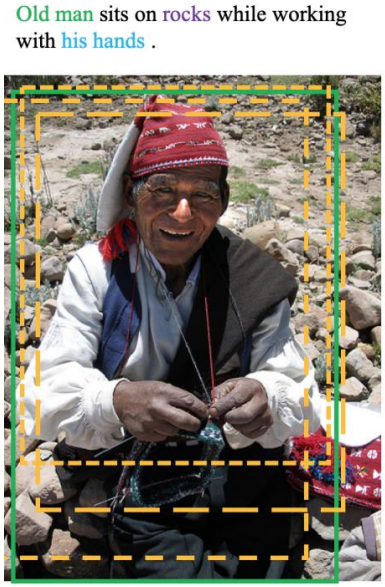
Phrase Grounding by Soft-Label Chain CRF

Two main problems:

(1) Dependencies between entities



(2) Multiple region proposals



Liu J, Hockenmaier J. “Phrase Grounding by Soft-Label Chain Conditional Random Field” EMNLP 2019

Phrase Grounding by Soft-Label Chain CRF

Two main problems:

(1) Dependencies between entities

Cheerleaders at a sporting event toss a girl high up into the air .

(2) Multiple region proposals

Old man sits on rocks while working with his hands .

Solution: Formulate the phrase grounding as a **sequence labeling task**

- ❑ Treat the candidate regions as **potential labels**
- ❑ Propose the Soft-Label Chain CRFs to model **dependencies** among regions
- ❑ Address the **multiplicity** of gold labels

Phrase Grounding by Soft-Label Chain CRF

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp s(\boldsymbol{y}, \boldsymbol{x})}{\sum_{\boldsymbol{y}'} \exp s(\boldsymbol{y}', \boldsymbol{x})}$$

- Input sequence: $\boldsymbol{x} = x^{1:T}$
- Label sequence: $\boldsymbol{y} = y^{1:T}$
- Score function: $s(\boldsymbol{x}, \boldsymbol{y})$

Standard CRF

- Cross-entropy Loss: $L = -\log p(\boldsymbol{y}|\boldsymbol{x}) = -s(\boldsymbol{y}, \boldsymbol{x}) + \log Z(\boldsymbol{x})$

➢ Each input x^i is associated to only one label y^i

Soft-Label CRF:

- KL-divergence between the model and target distribution:

$$L = \sum_{\boldsymbol{y}} \left\{ q(\boldsymbol{y}|\boldsymbol{x}) \log \frac{q(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y}|\boldsymbol{x})} \right\}$$

- Sequence of target distribution: $\boldsymbol{q} = q^{1:T}$
- **Label distribution** over all K possible labels for input x^t: $q^t \in \mathbb{R}^K$

➢ Each input x^i is associated to a distribution of labels y^i

Liu J, Hockenmaier J. “Phrase Grounding by Soft-Label Chain Conditional Random Field” EMNLP 2019

Phrase Grounding by Soft-Label Chain CRF

For efficiency: Reduce the model to a first-order linear chain CRF, whose scoring function factorizes as:

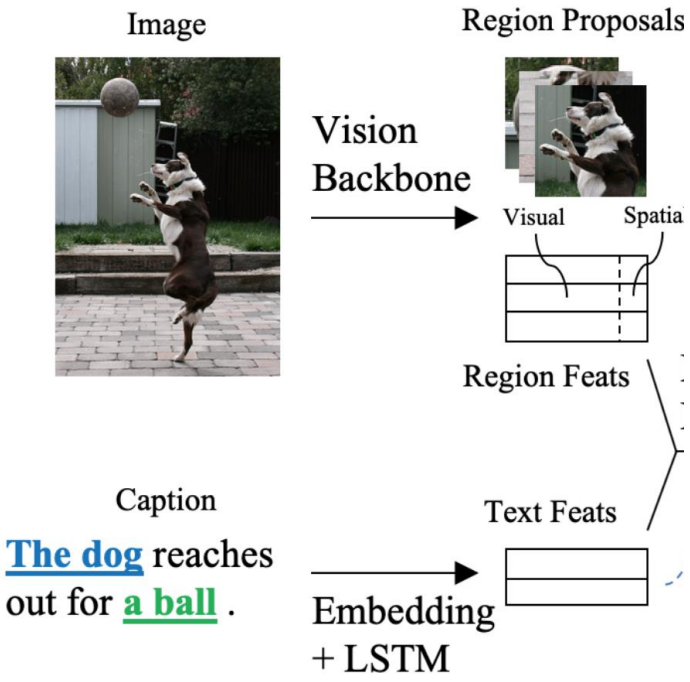
$$s(\boldsymbol{y}, \boldsymbol{x}) = \sum_t s(y^t, y^{t-1}, \boldsymbol{x})$$
$$= \sum_t \left\{ \tau(y^t, y^{t-1}, \boldsymbol{x}) + \varepsilon(y^t, \boldsymbol{x}) \right\}$$

where $\tau(\cdot, \cdot, \cdot)$ are the pairwise potentials between labels at $t-1$ and t

$\varepsilon(\cdot, \cdot)$ are the unary potentials between label and input at t

Liu J, Hockenmaier J. “Phrase Grounding by Soft-Label Chain Conditional Random Field” EMNLP 2019

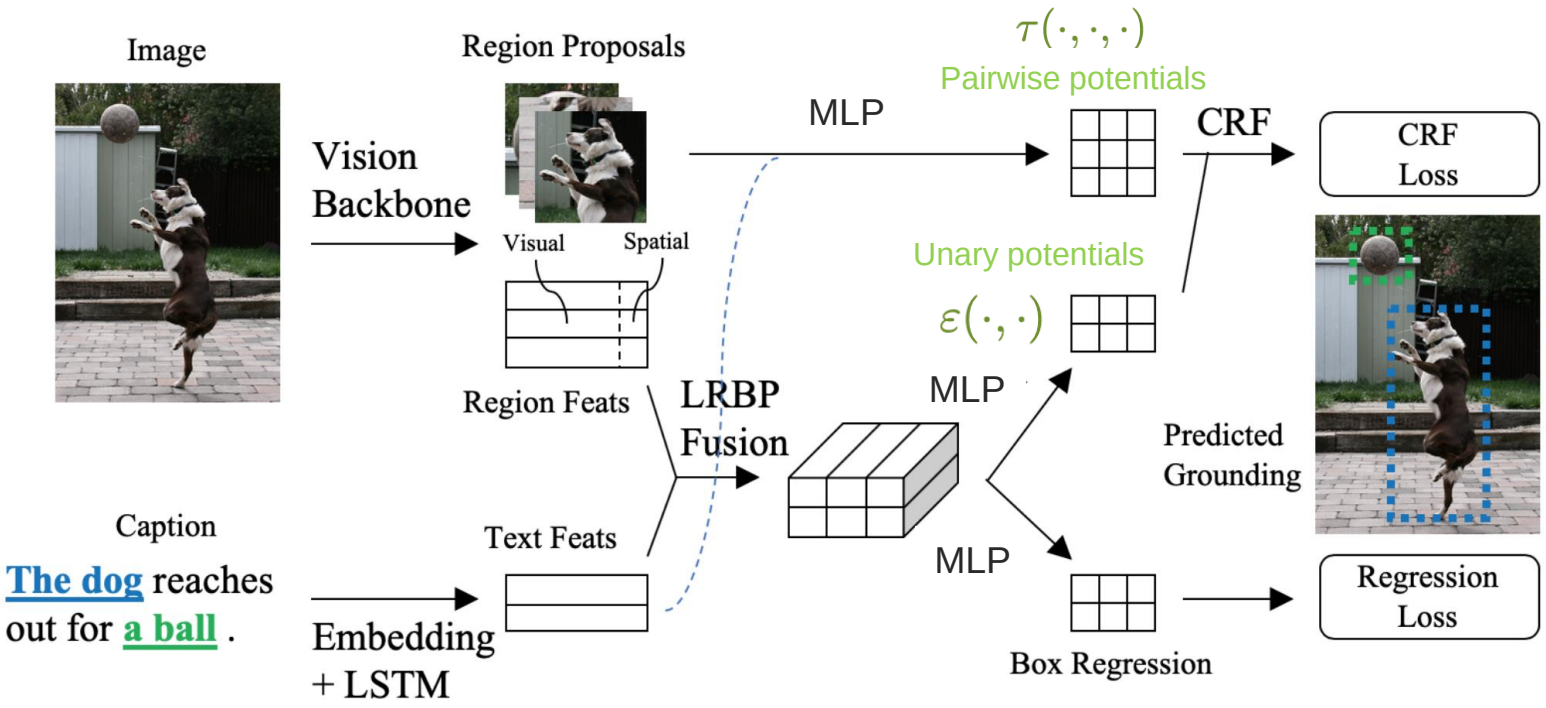
Phrase Grounding by Soft-Label Chain CRF



- Training Objective: $L = L_{\text{label}} + \gamma L_{\text{reg}}$

Liu J, Hockenmaier J. “Phrase Grounding by Soft-Label Chain Conditional Random Field” EMNLP 2019

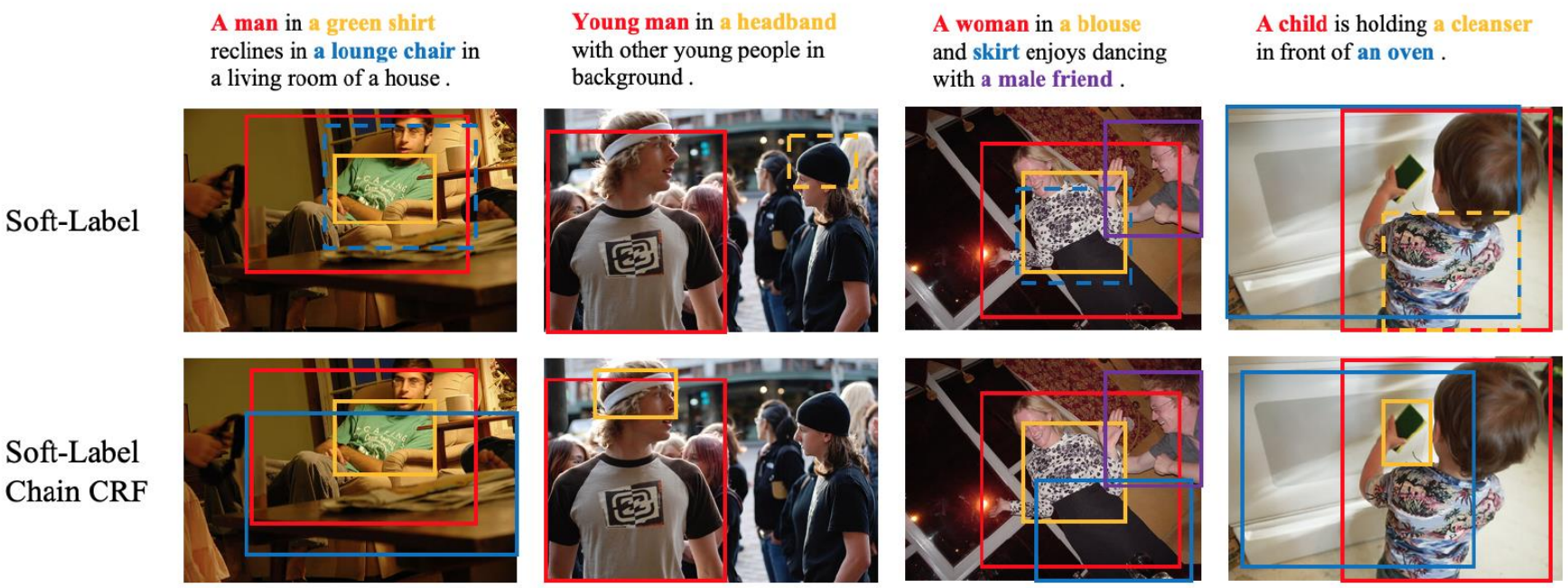
Phrase Grounding by Soft-Label Chain CRF



- Training Objective: $L = L_{\text{label}} + \gamma L_{\text{reg}}$

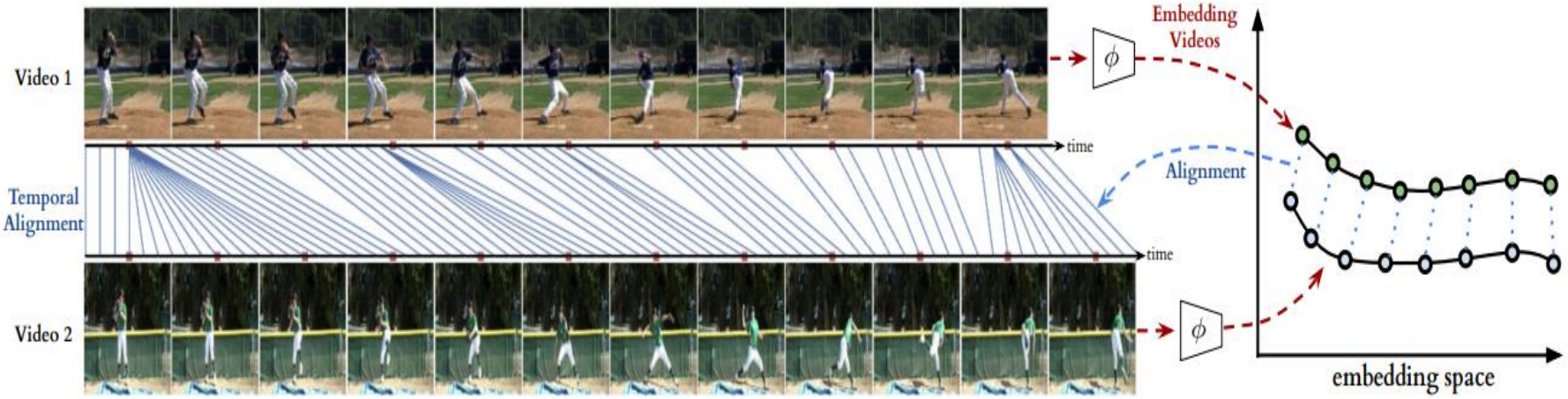
Liu J, Hockenmaier J. "Phrase Grounding by Soft-Label Chain Conditional Random Field" EMNLP 2019

Phrase Grounding by Soft-Label Chain CRF



Liu J, Hockenmaier J. “Phrase Grounding by Soft-Label Chain Conditional Random Field” EMNLP 2019

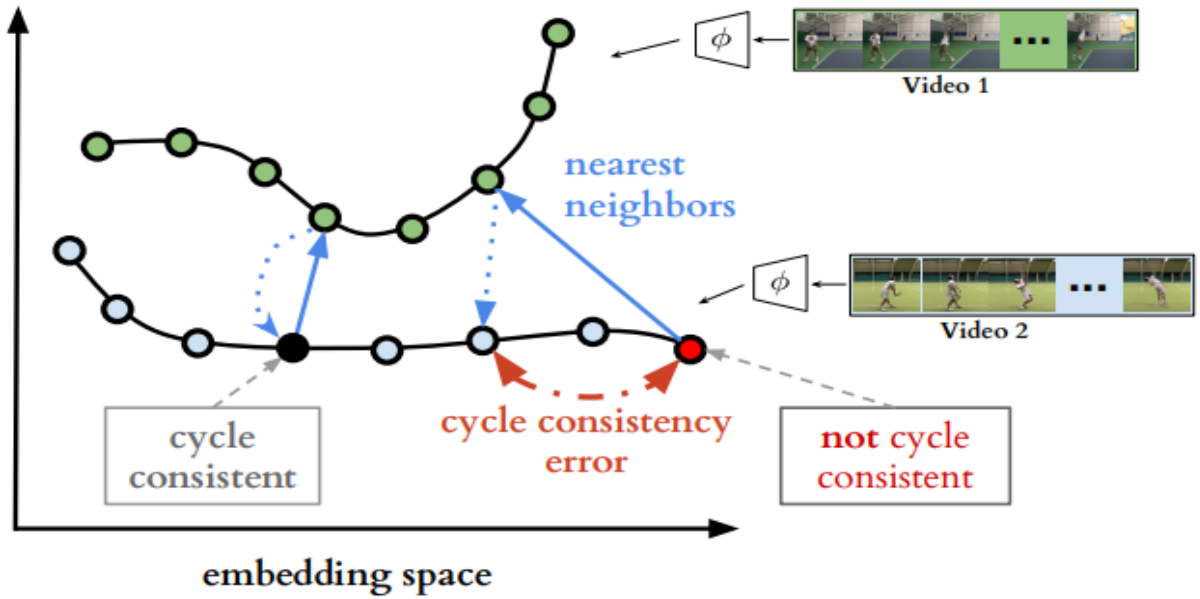
Temporal Cycle-Consistency Learning



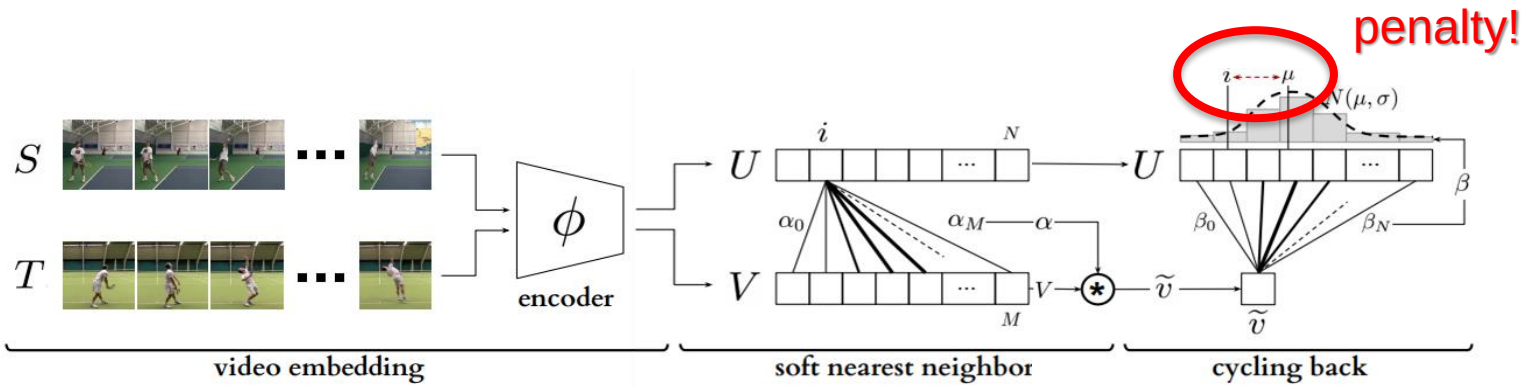
Self-supervised approach to learn an embedding space where two similar video sequences can be aligned temporally

Temporal Cycle-Consistency Learning

Representation Learning by enforcing **Cycle consistency**



Temporal Cycle-Consistency Learning



Compute “soft” / “weighted” nearest neighbour:

distances: $\alpha_j = \frac{e^{-||u_i - v_j||^2}}{\sum_k^M e^{-||u_i - v_k||^2}}$ Soft nearest neighbor: $\widetilde{v} = \sum_j^M \alpha_j v_j,$

Find the nearest neighbor the other way and then penalize the distance:

$$\beta_k = \frac{e^{-||\widetilde{v} - u_k||^2}}{\sum_j^N e^{-||\widetilde{v} - u_j||^2}}$$

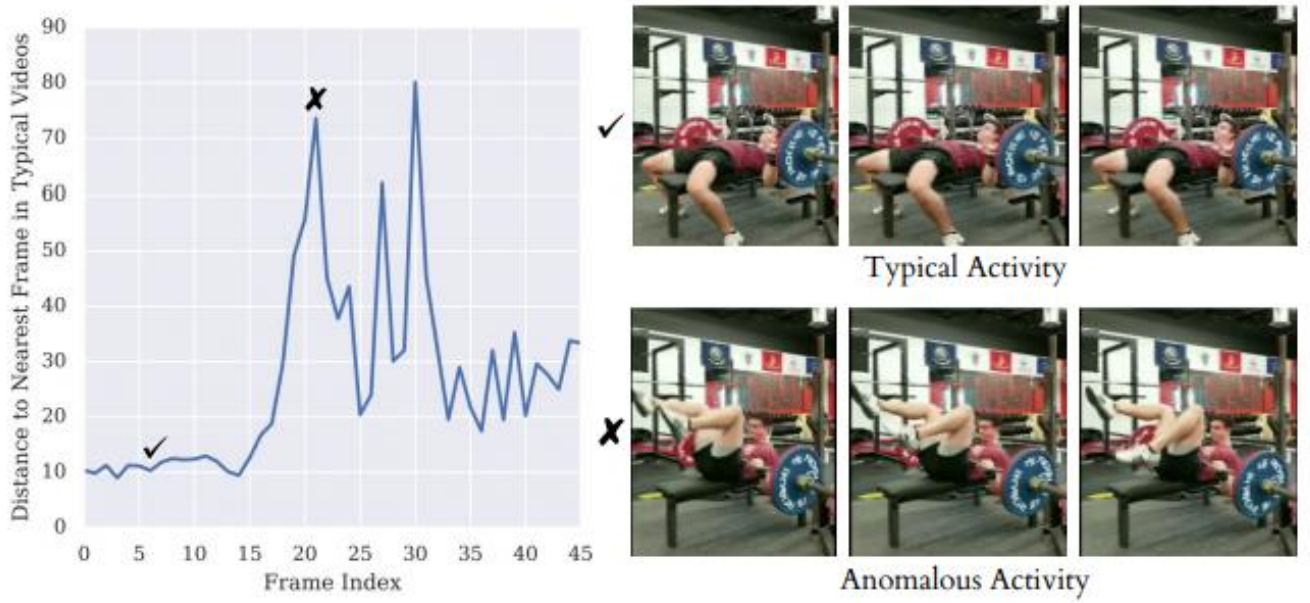
$$L_{cbr} = \frac{|i - \mu|^2}{\sigma^2} + \lambda \log(\sigma)$$

Temporal Cycle-Consistency Learning

Nearest Neighbour Retrieval

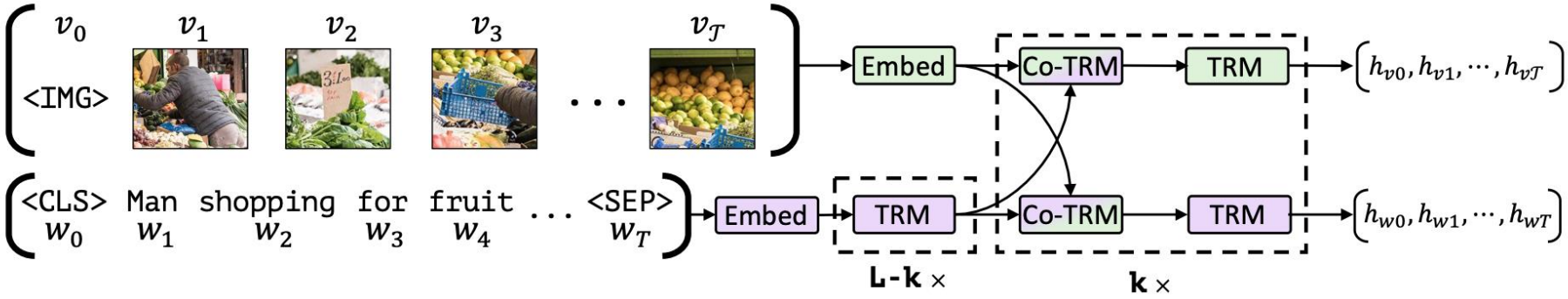
Temporal Cycle-Consistency Learning

Anomaly Detection



ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

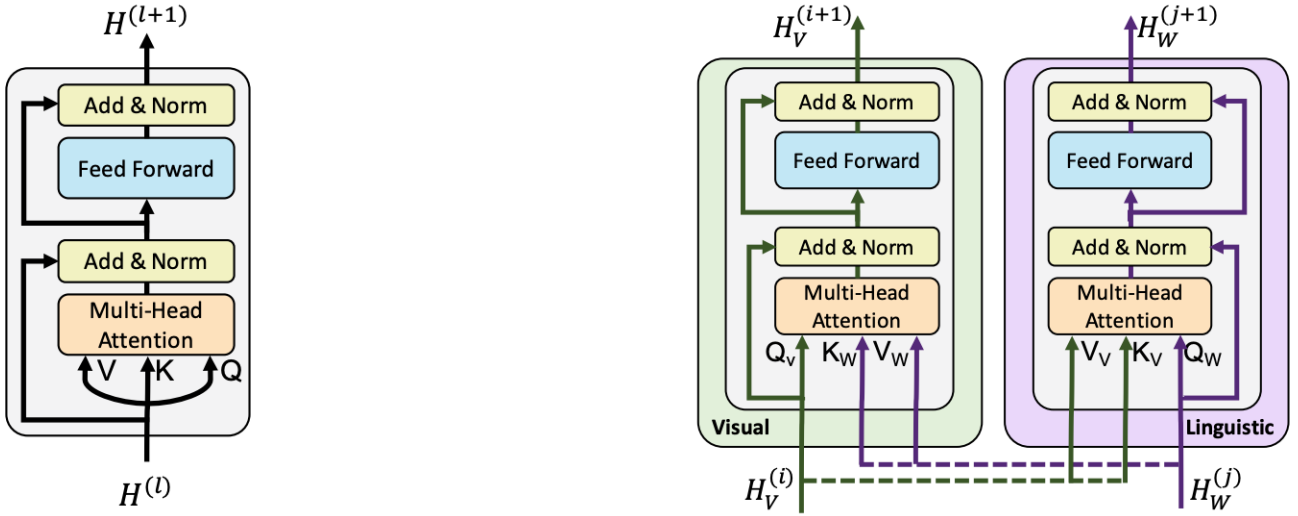
- ViLBERT: Extending BERT to jointly represent images and text
  - Two parallel BERT-style streams operating over image regions and text segments.
  - Each stream is a series of transformer blocks (TRM) and novel co-attentional transformer layers (Co-TRM).



Lu J, Batra D, Parikh D, et al. “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.” NeurIPS 2019

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

- Co-attentional transformer layers
  - Enable information exchange between modalities.
  - Provide interaction between modalities at varying representation depths.



(a) Standard encoder transformer block

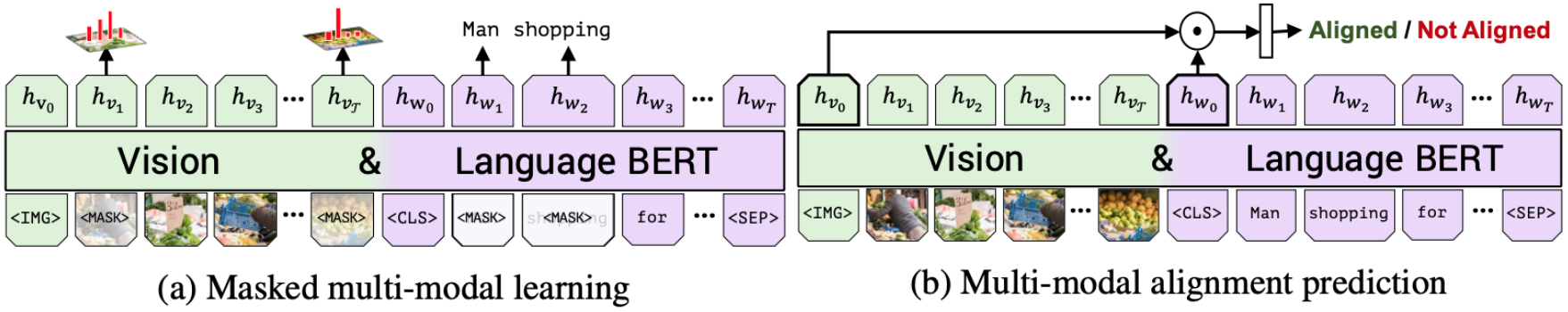
(b) Our co-attention transformer layer

Lu J, Batra D, Parikh D, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." NeurIPS 2019

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

Two pretraining tasks

1. masked multi-modal modelling
   - the model must reconstruct image region categories or words for masked inputs given the observed inputs
2. multi-modal alignment prediction
   - the model must predict whether or not the caption describes the image content.



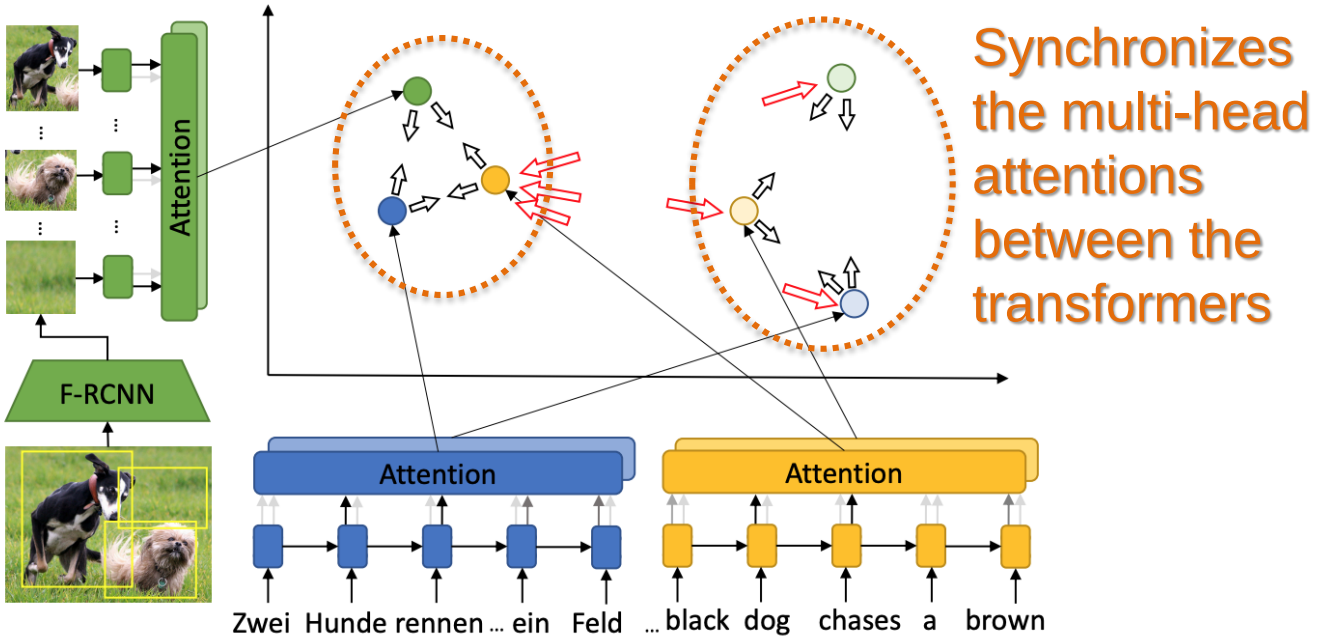
(a) Masked multi-modal learning

(b) Multi-modal alignment prediction

Lu J, Batra D, Parikh D, et al. “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.” NeurIPS 2019

Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations

- Introduce a new multi-head attention diversity loss to encourage diversity among attention heads.



Huang, Po-Yao, et al. “Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations.” EMNLP 2019

Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations

Multi-head attention diversity loss

- Taking Image-English instances {V, E} as an example

$$l_\theta^D(V,E) = \sum_p \sum_k \sum_r [\alpha_D - s(v_p^k, e_p^{k \neq r})]_+$$

If they are from the same k^{th} head, then they should be close to each other…

… within a certain margin

- α_D: diversity margin
- $s(a,b) = \frac{a^T b}{\|a\|\|b\|}$: cosine similarity
- e_p^k: the k-th attention head fo English sentence representation
- $[.]_+ = \max(0,.)$: the hinge function

Huang, Po-Yao, et al. “Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations.” EMNLP 2019

Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations

Multi-head attention diversity loss

- Taking Image-English instances {V, E} as an example

$$l_\theta^D(V,E) = \sum_p \sum_k \sum_r [\alpha_D - s(v_p^k, e_p^{k \neq r})]_+$$

- Diversity within-modalities and across-modalities:

$$l_\theta^D(V,E,G) = l_\theta^D(V,V) + l_\theta^D(G,G) + l_\theta^D(E,E) + l_\theta^D(V,E) + l_\theta^D(V,G) + l_\theta^D(G,E),$$

Huang, Po-Yao, et al. “Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations.” EMNLP 2019

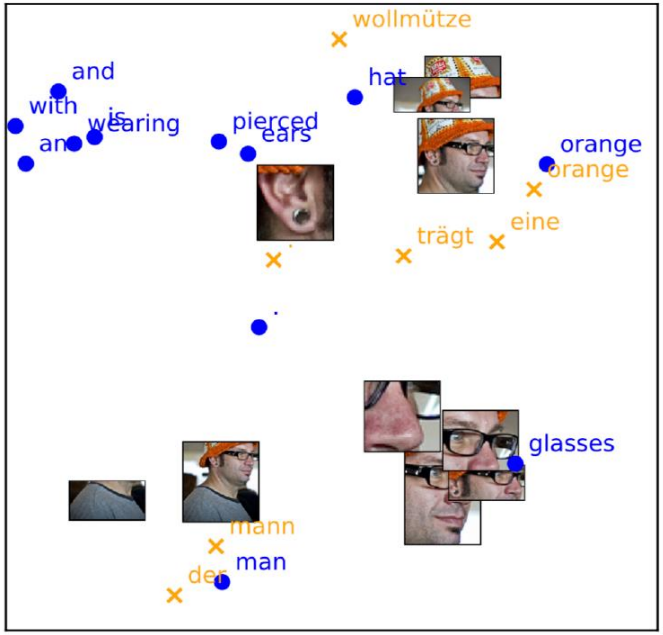
Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations

Learned multilingual multimodal embeddings:

(note the sentences are *without* translation pairs)

The man with pierced ears is wearing glasses and an orange hat

Der mann trägt eine orange wollmütze .



Huang, Po-Yao, et al. “Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations.” EMNLP 2019

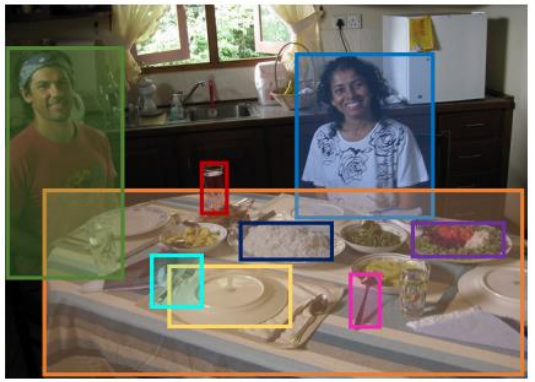
New Directions: Representation

ViCo: Word Embeddings from Visual Co-occurrences

Learn vector representations for text using visual co-occurrences

Four types of co-occurrences:

(a) Object - Attribute
(b) Attribute - Attribute
(c) Context
(d) Object-Hypernym



Region	Object Words	Attribute Words
	man, person, adult, mammal	muscular, smiling
	woman, person, adult, mammal	lean, smiling
	table, tablecloth, furniture	striped, oval
	rice, carbohydrates, food	white, grainy, cooked
	salad, roughage, food	leafy, chopped, healthy, red, green
	glass, glassware, utensil	clear, transparent, reflective, tall
	plate, crockery, utensil	ceramic, white, round, circular
	fork, cutlery, utensil	metallic, shiny, reflective
	spoon, cutlery, utensil	serving, metallic, shiny, reflective

ViCo: Word Embeddings from Visual Co-occurrences

Relatedness through Co-occurrences

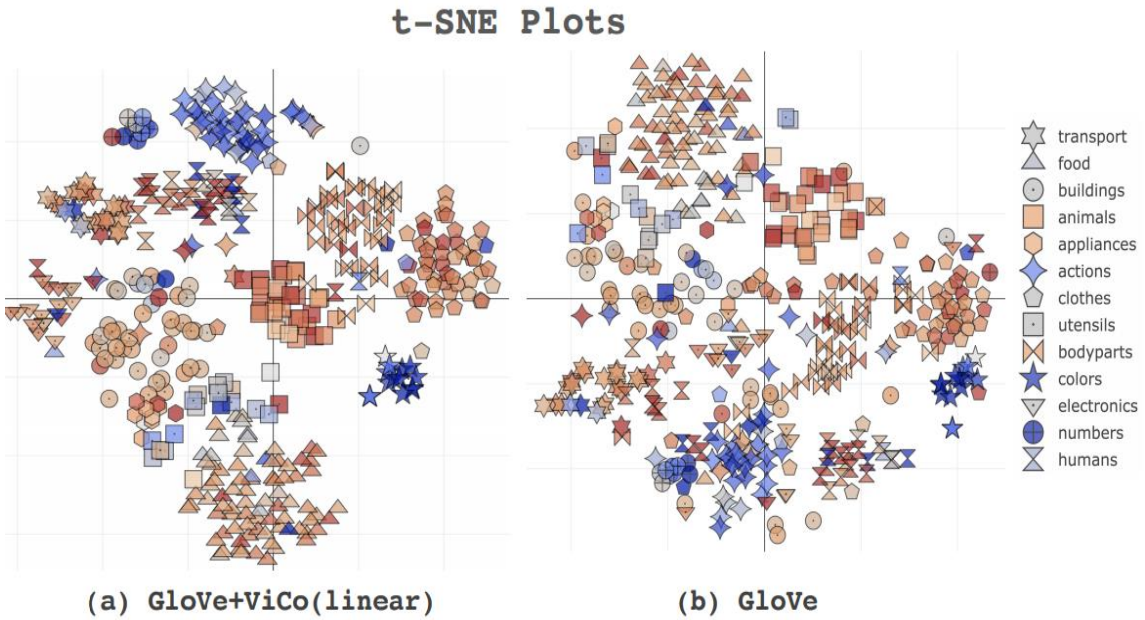
Word Pair	ViCo	Obj-Attr	Attr-Attr	Obj-Hyp	Context	GloVe
crouch / squat	0.61	0.74	0.72	0.18	0.25	0.05
sweet / dessert	0.66	0.78	0.76	0.56	0.79	0.43
man / male	0.71	0.98	0.8	0.38	1	0.34
purple / violet	0.75	0.93	1	0.24	0.03	0.52
hosiery / sock	0.52	0.27	0.18	0.87	0.07	0.23
aeroplane / aircraft	0.73	0.43	0.07	0.87	0.75	0.43
bench / pew	0.63	0.67	0.09	0.79	-0.14	0.1
keyboard / mouse	0.19	0.63	0.19	0.09	0.95	0.52
laptop / desk	0.39	0.23	0.24	0.1	0.94	0.28
window / door	0.59	0.46	0.35	0.53	0.93	0.67
hair / blonde	0.16	0.56	0.32	-0.15	0.17	0.51
thigh / ankle	0.09	0.19	0.03	0.01	0.39	0.74
garlic / onion	0.36	-0.03	0.3	0.37	0.56	0.77
driver / car	0.27	0.16	0.26	0.12	0.53	0.71
girl / boy	0.41	0.38	0.22	0.44	0.74	0.83

Since ViCo is learned from multiple types of co-occurrences, it is hypothesized to provide a richer sense of relatedness

- Learned using a multi-task Log-Bilinear Model

ViCo: Word Embeddings from Visual Co-occurrences

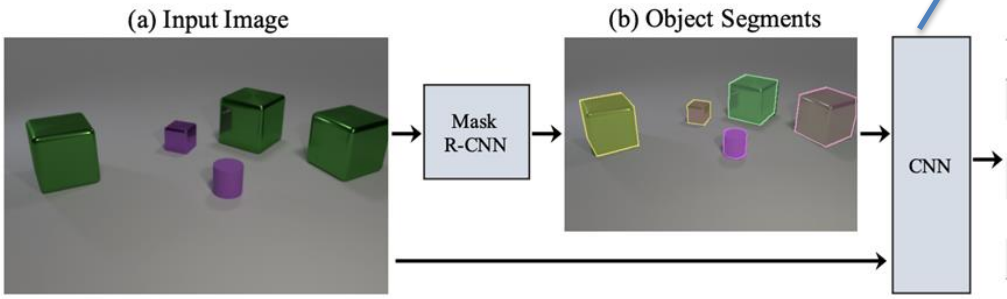
ViCO leads to more homogenous clusters compared to GloVe



Neural-symbolic VQA

1) Image de-rendering

Previously trained in a supervised way



(c) Structural Scene Representation

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

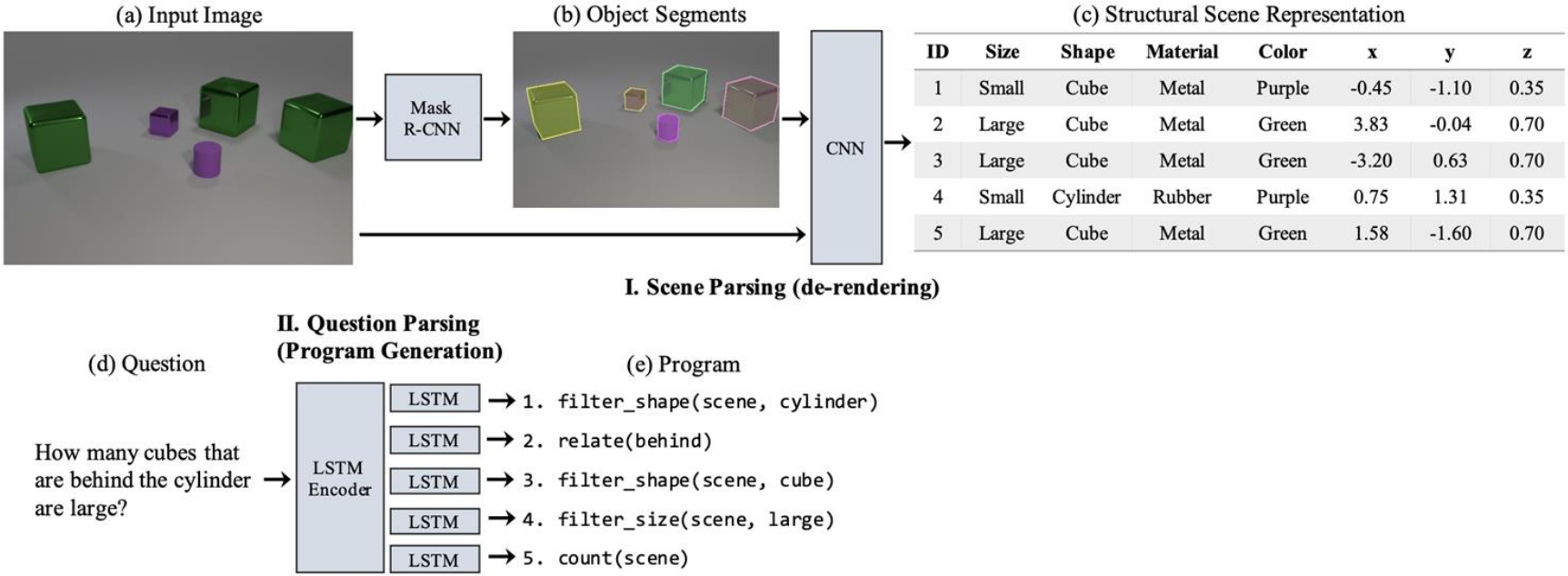
I. Scene Parsing (de-rendering)

Kexin Yi, et al. “Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding.” Neurips 2018

Neural-symbolic VQA

2) Parsing questions into programs

Similar to neural module networsk



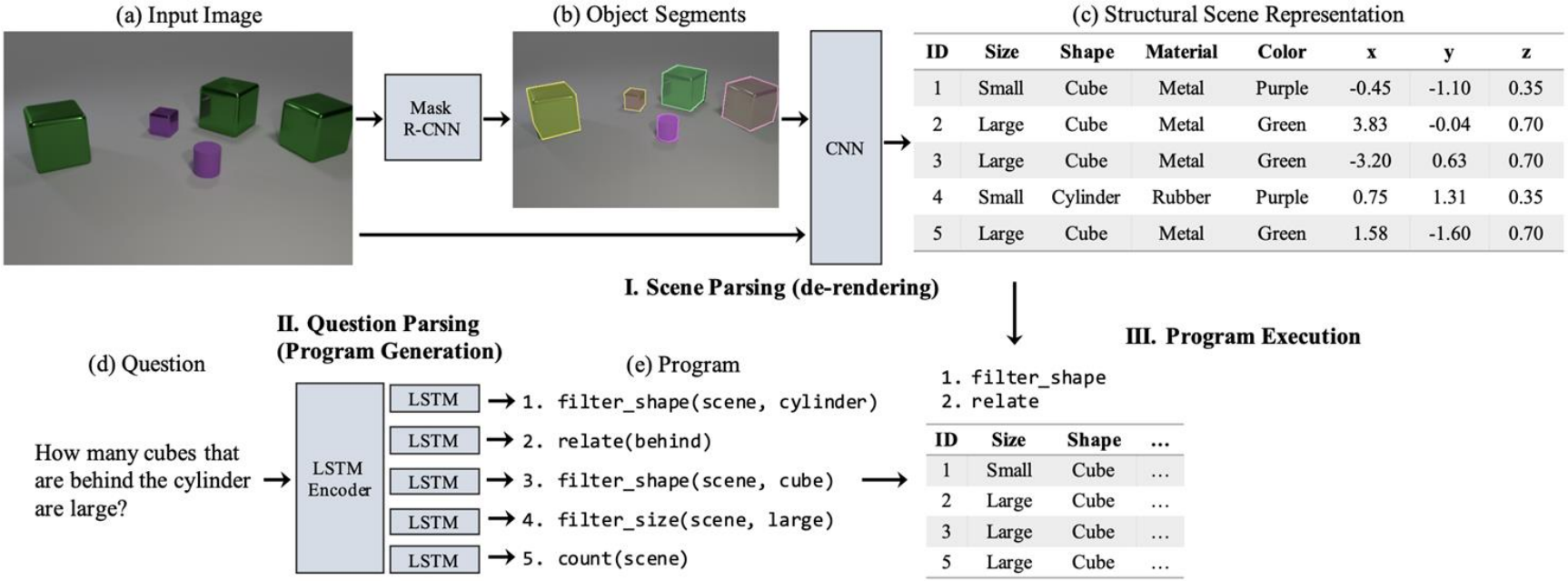
ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

Kexin Yi, et al. “Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding.” Neurips 2018

Neural-symbolic VQA

3) Program execution

Execution of the program is somewhat easier given the “symbolic” representation of the image



ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

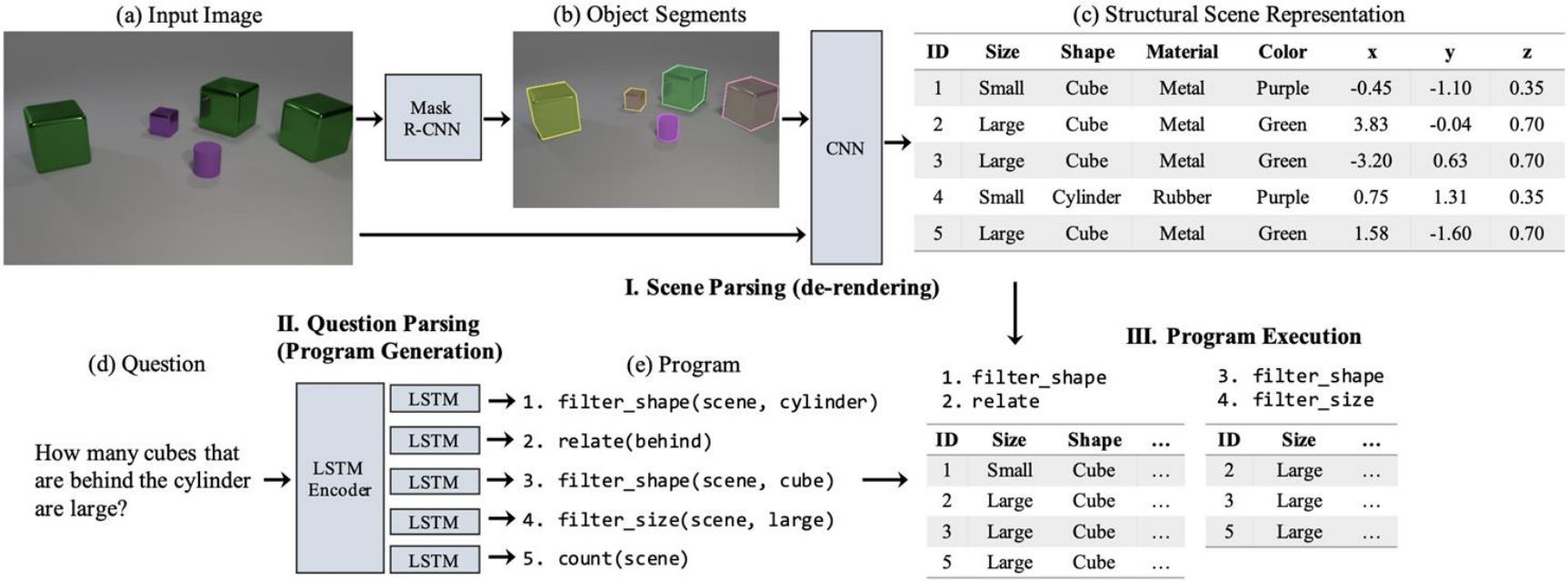
ID	Size	Shape	...
1	Small	Cube	...
2	Large	Cube	...
3	Large	Cube	...
5	Large	Cube	...

Kexin Yi, et al. “Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding.” Neurips 2018

Neural-symbolic VQA

3) Program execution

Execution of the program is somewhat easier given the “symbolic” representation of the image



ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

ID	Size	Shape	...
1	Small	Cube	...
2	Large	Cube	...
3	Large	Cube	...
5	Large	Cube	...

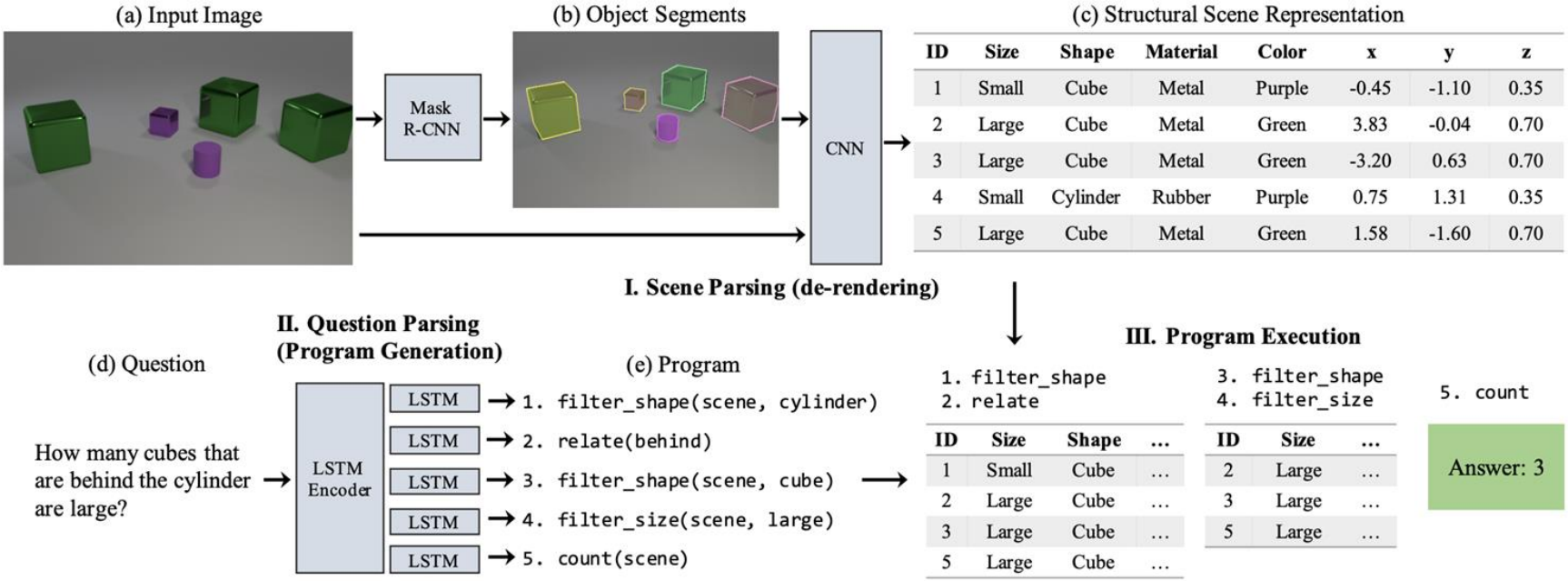
ID	Size	...
2	Large	...
3	Large	...
5	Large	...

Kexin Yi, et al. “Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding.” Neurips 2018

Neural-symbolic VQA

3) Program execution

Execution of the program is somewhat easier given the "symbolic" representation of the image



ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

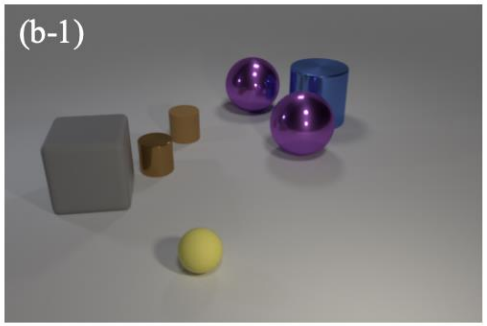
ID	Size	Shape	...
1	Small	Cube	...
2	Large	Cube	...
3	Large	Cube	...
5	Large	Cube	...

ID	Size	...
2	Large	...
3	Large	...
5	Large	...

Kexin Yi, et al. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding." Neurips 2018

Neural-symbolic VQA

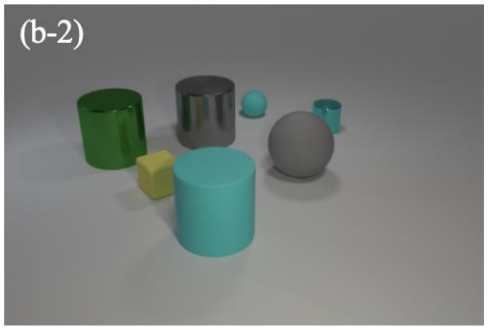
(b) 1K Programs



Q: What number of cylinders are gray objects or tiny brown matte objects?

Ours	IEP
scene	filter_small
filter_small	filter_brown
filter_brown	filter_large
filter_rubber	filter_cyan
scene	...(25 modules)
filter_gray	filter_metal
union	union
filter_cylinder	filter_cylinder
count	count

A: 1 **A:** 2



Q: Are there more yellow matte things that are right of the gray ball than cyan metallic objects?

Ours	IEP
scene	filter_small
filter_cyan	filter_cyan
filter_metal	union
Count	filter_brown
...(4 modules)	...(25 modules)
scene	filter_small
filter_yellow	filter_yellow
filter_rubber	filter_rubber
count	count
greater_than	greater_than

A: no **A:** no

Neural-symbolic programs give more accurate answers (shown in blue)

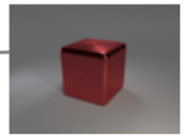
Kexin Yi, et al. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding." Neurips 2018

The Neuro-symbolic Concept Learner

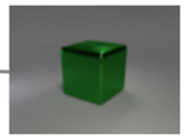
Extension from Neural-symbolic VQA:

Learns visual concepts, words, and semantic parsing of sentences without explicit supervision on any of them, but just by looking at **images and reading paired questions and answers**

I. Learning basic, object-based concepts.

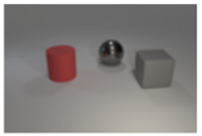


Q: What's the color of the object?
A: Red.
Q: Is there any cube?
A: Yes.



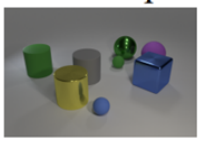
Q: What's the color of the object?
A: Green.
Q: Is there any cube?
A: Yes.

II. Learning relational concepts based on referential expressions.



Q: How many objects are right of the red object?
A: 2.
Q: How many objects have the same material as the cube?
A: 2

III. Interpret complex questions from visual cues.



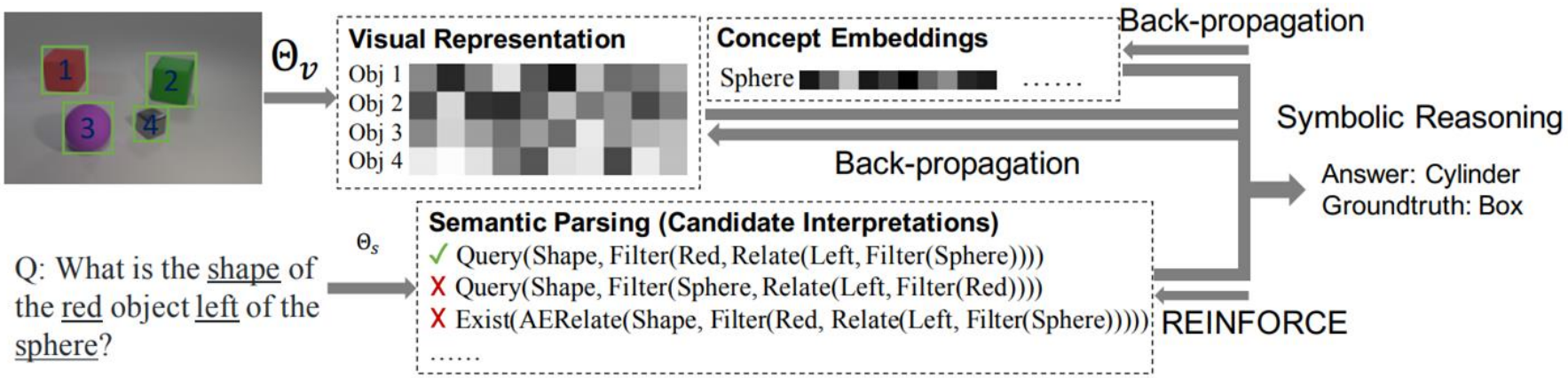
Q: How many objects are both right of the green cylinder and have the same material as the small blue ball?
A: 3

Jiayuan Mao , et al. "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision." ICLR 2019

The Neuro-symbolic Concept Learner

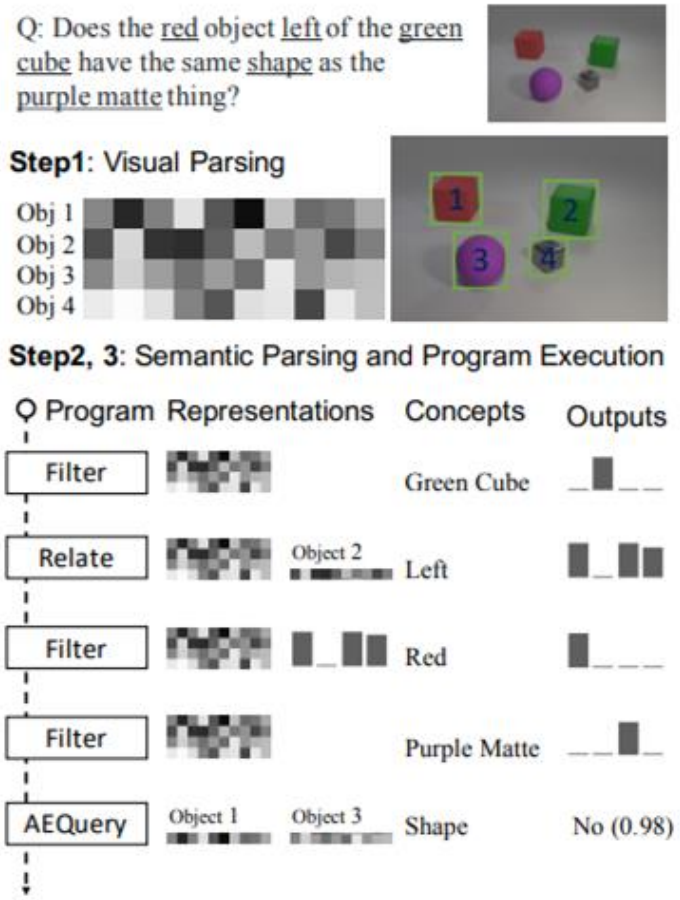
Extension from Neural-symbolic VQA:

Learns visual concepts, words, and semantic parsing of sentences without explicit supervision on any of them, but just by looking at **images and reading paired questions and answers**



Jiayuan Mao , et al. “The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision.” ICLR 2019

The Neuro-symbolic Concept Learner

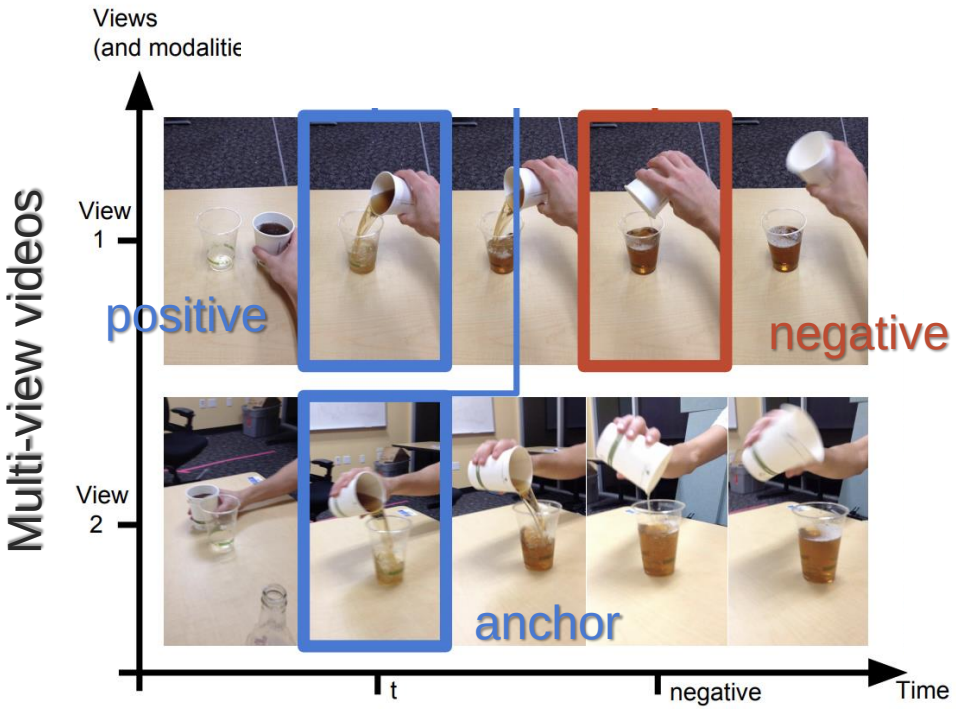


Jiayuan Mao , et al. “The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision.” ICLR 2019

Time-Contrastive Networks: Self-Supervised Learning from (Multi-View) Video

Goal: We want to observe and disentangle the world from many videos.

Main idea:
Embeddings should be close if from synchronized frames



Time-Contrastive Networks: Self-Supervised Learning from (Multi-View) Video

Let's learn an embedding function ***f*** for an sequence ***x***

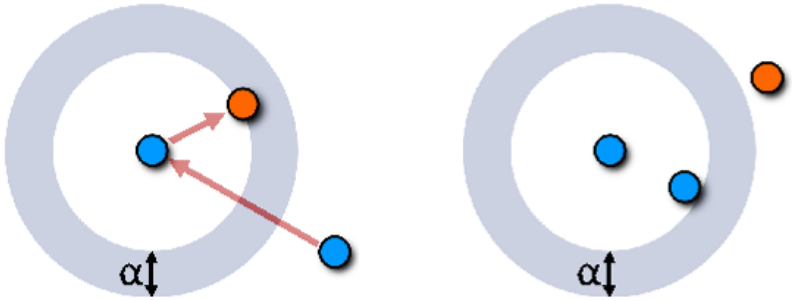
Set of all **triplets** in the training set

$$\forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in \mathcal{T}$$

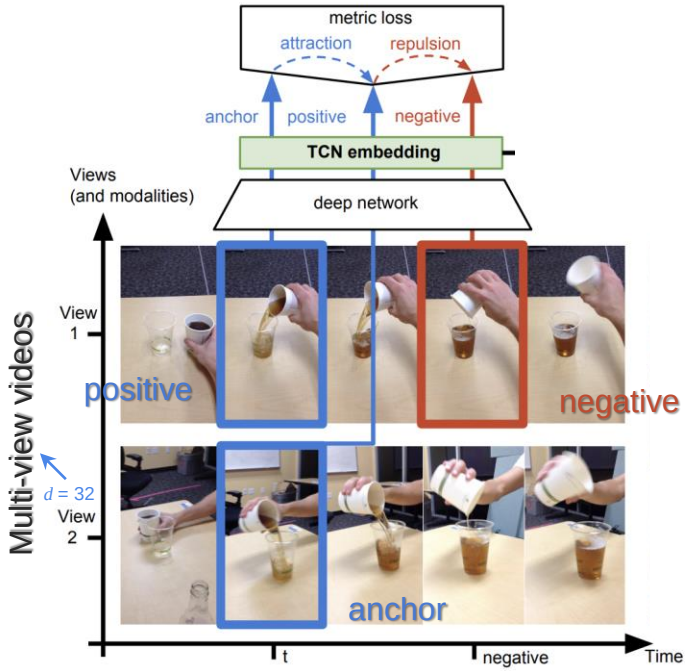
anchor positive negative

Margin enforced between positive/negative pairs:

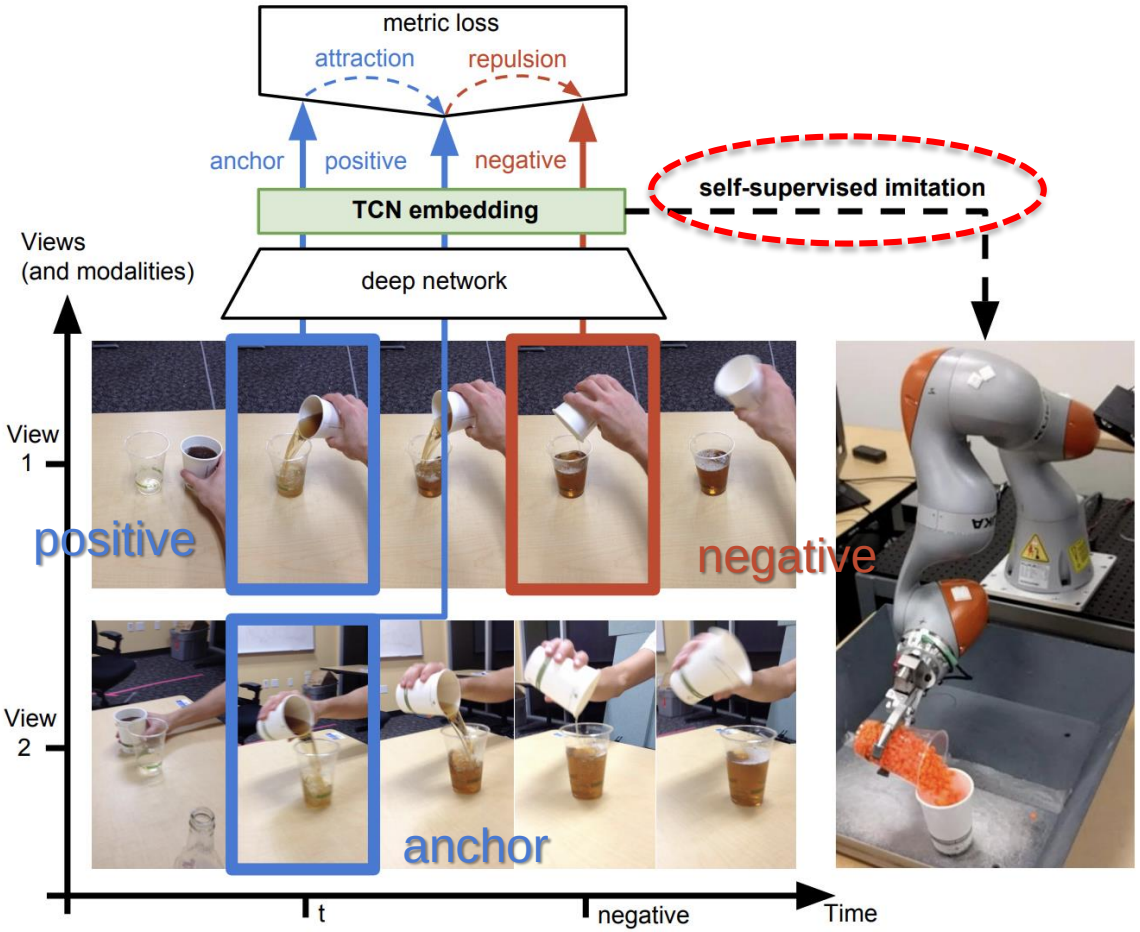
$$||f(x_i^a) - f(x_i^p)||_2^2 + \alpha < ||f(x_i^a) - f(x_i^n)||_2^2,$$



(a) Triplet: before. (b) Triplet: after.

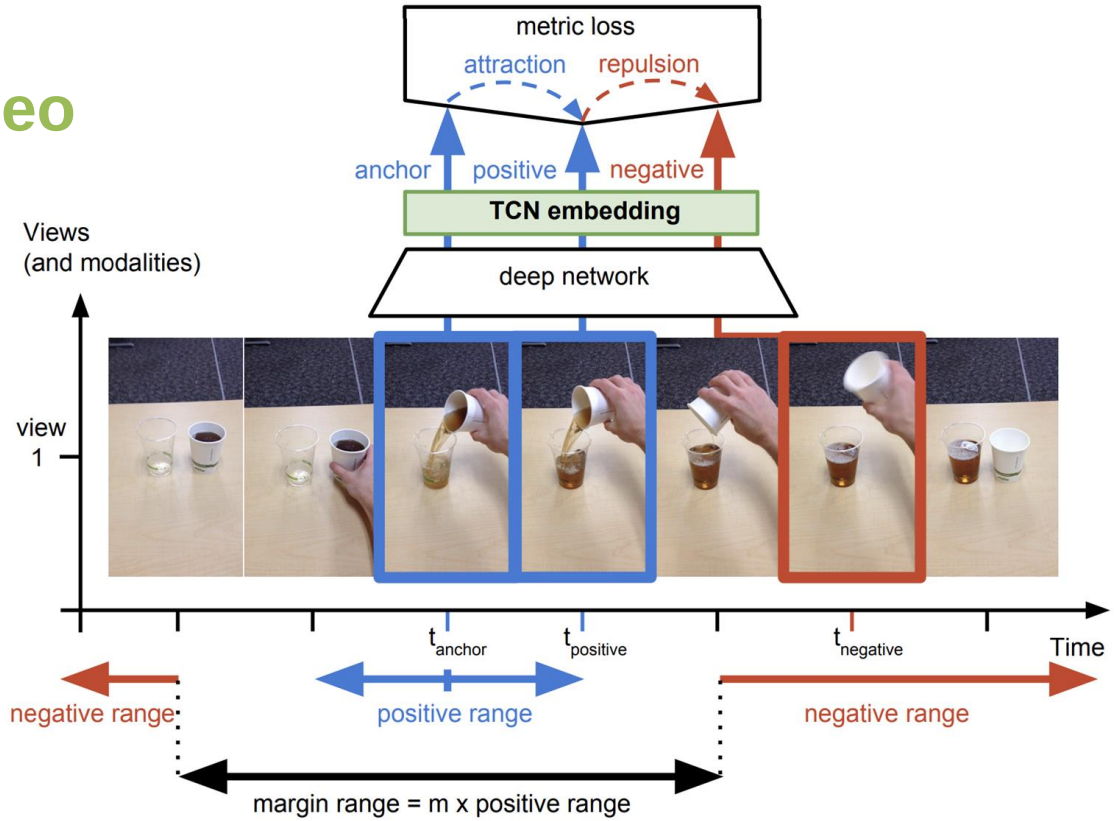


Time-Contrastive Networks: Self-Supervised Learning from (Multi-View) Video



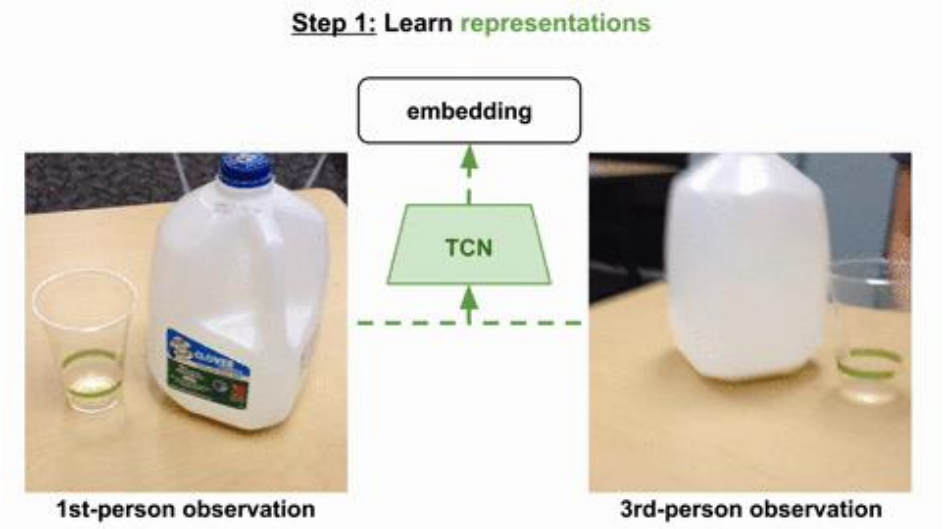
Time-Contrastive Networks: Self-Supervised Learning from Video

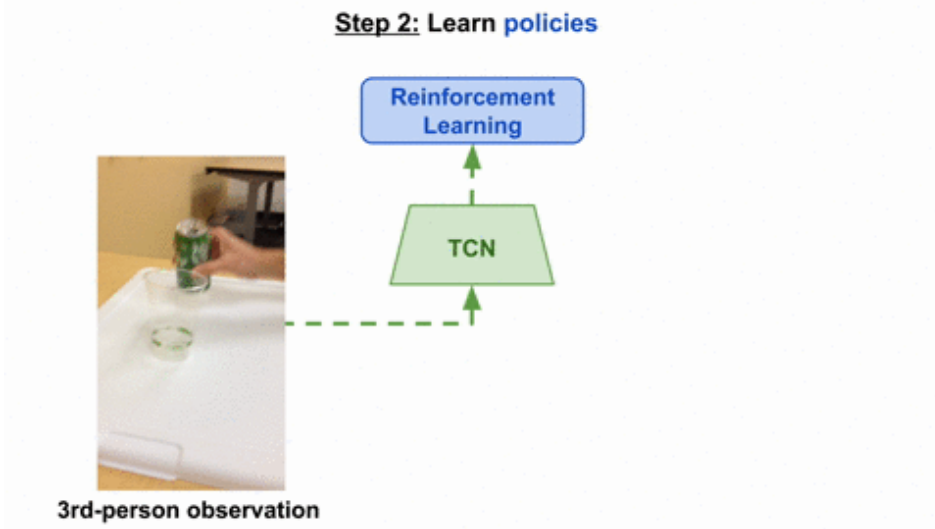
Learn RL policies from only one video



Time-Contrastive Networks: Self-Supervised Learning from Video

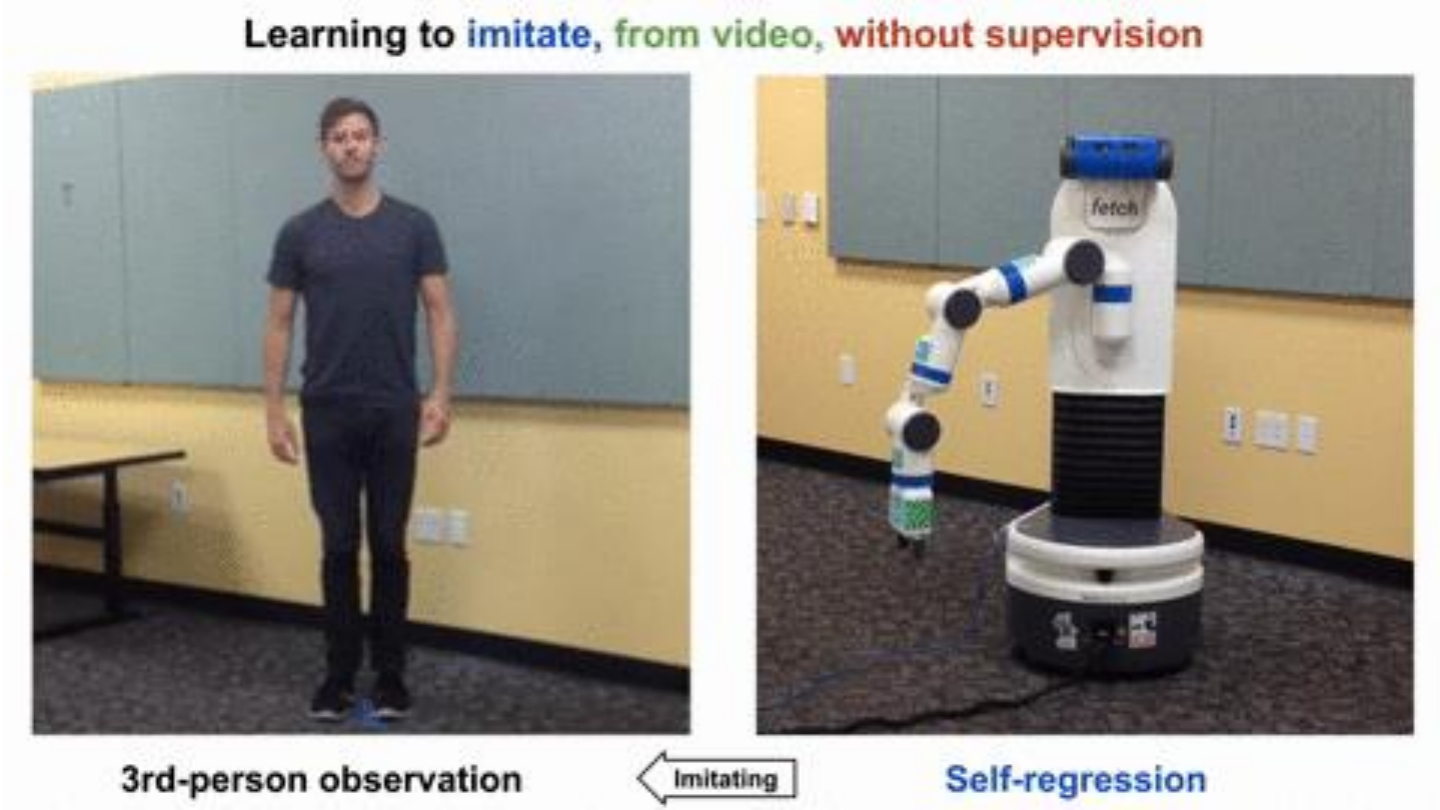
Demo: Pouring





Time-Contrastive Networks: Self-Supervised Learning from Video

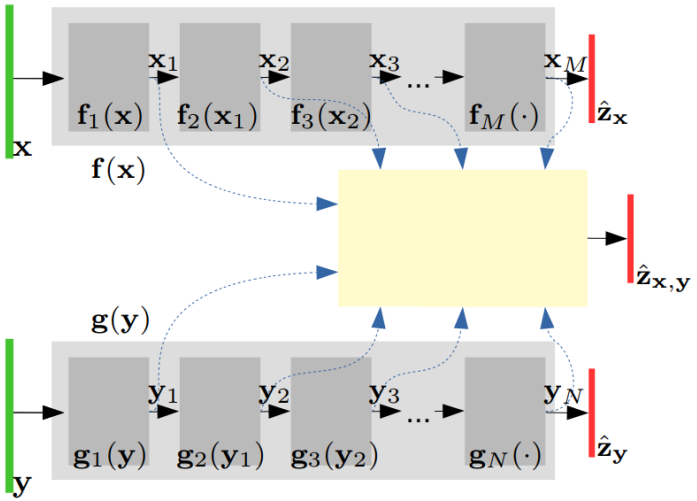
Demo: Pose Imitation



New Directions: Fusion

MFAS: Multimodal Fusion Architecture Search

Pose multimodal fusion as an architectural search problem

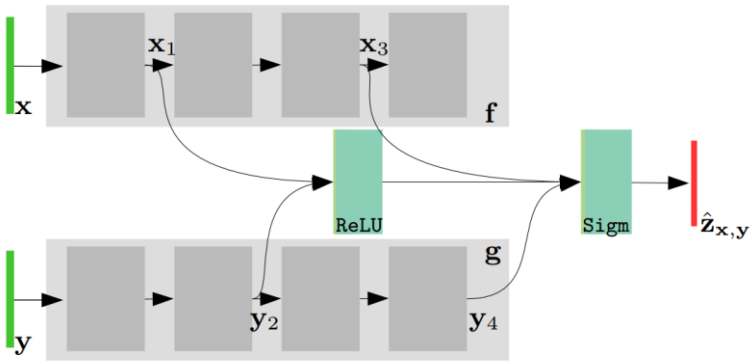


Each fusion layer combines three inputs:

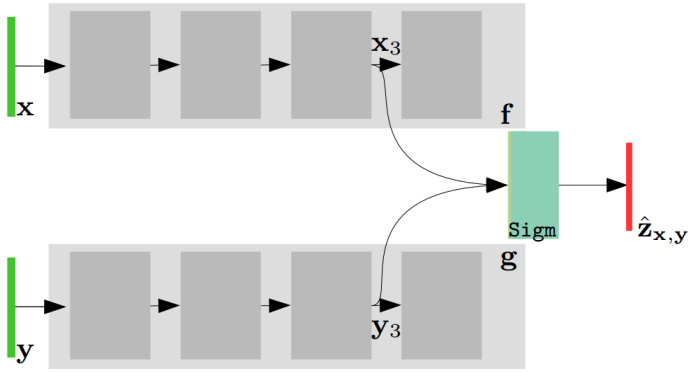
(a) Output from previous fusion layer
(b) Output from modality A
(c) Output from modality B

$$\mathbf{h}_l = \boldsymbol{\sigma}_{\gamma_l^p}\left(\mathbf{W}_l \begin{bmatrix} \mathbf{x}_{\gamma_l^m} \\ \mathbf{y}_{\gamma_l^n} \\ \mathbf{h}_{l-1} \end{bmatrix}\right)$$

MFAS: Multimodal Fusion Architecture Search



Realization 01



Realization 02

Enables the space to contain a large number of possible fusion architectures.

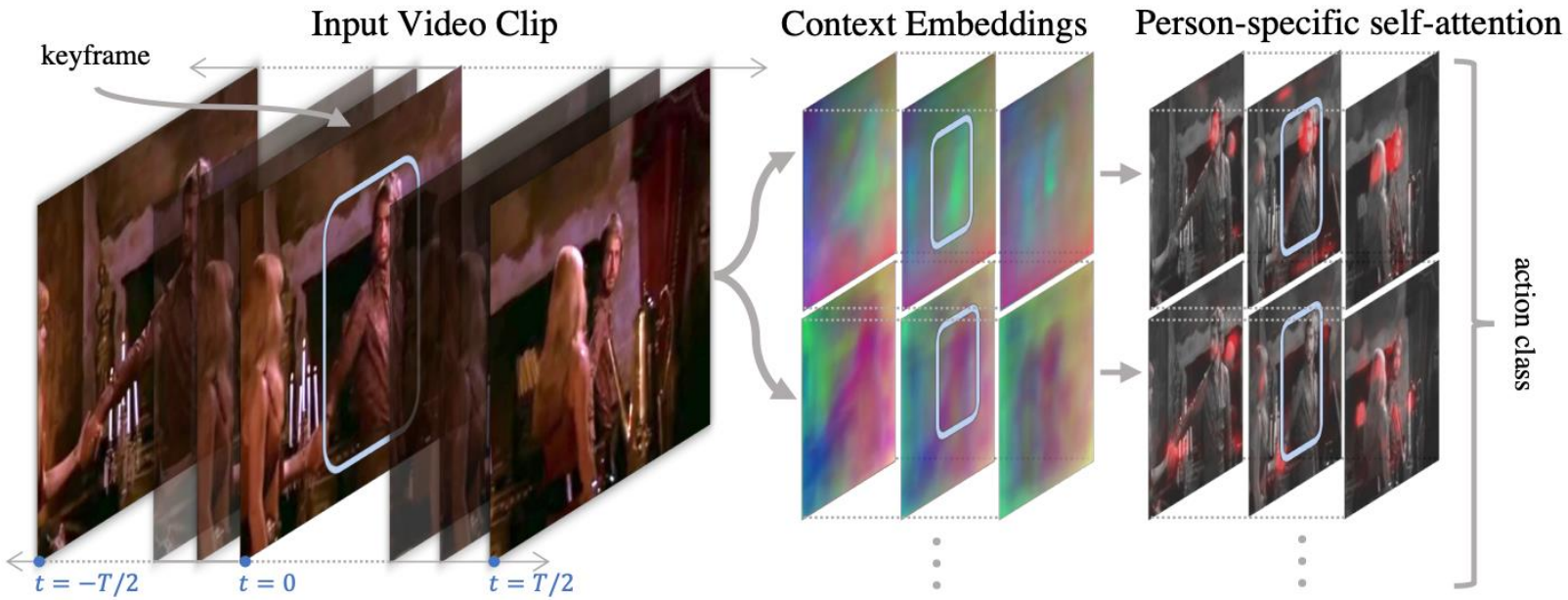
Space is naturally divided by complexity levels that can be interpreted as progression steps

Exploration performed by Sequential model-based optimization

Video Action Transformer Network

Recognizing and localizing human actions in video clips

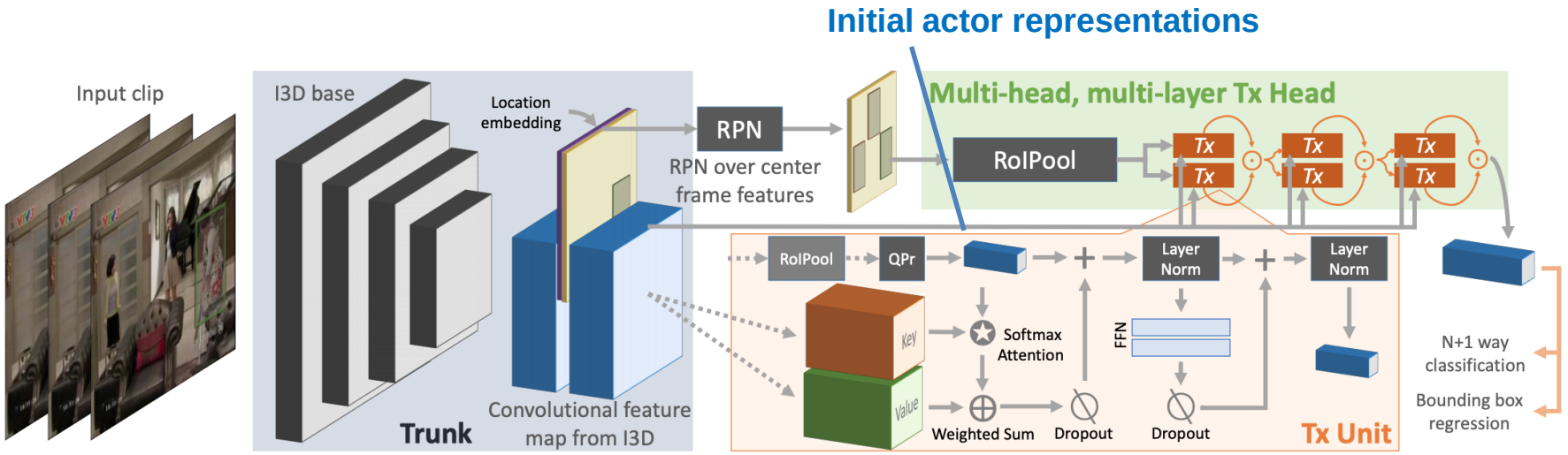
by attending person of interest and their context (other people, objects)



Rohit Girdhar, et al. “Video Action Transformer Network.” CVPR 2019

Video Action Transformer Network

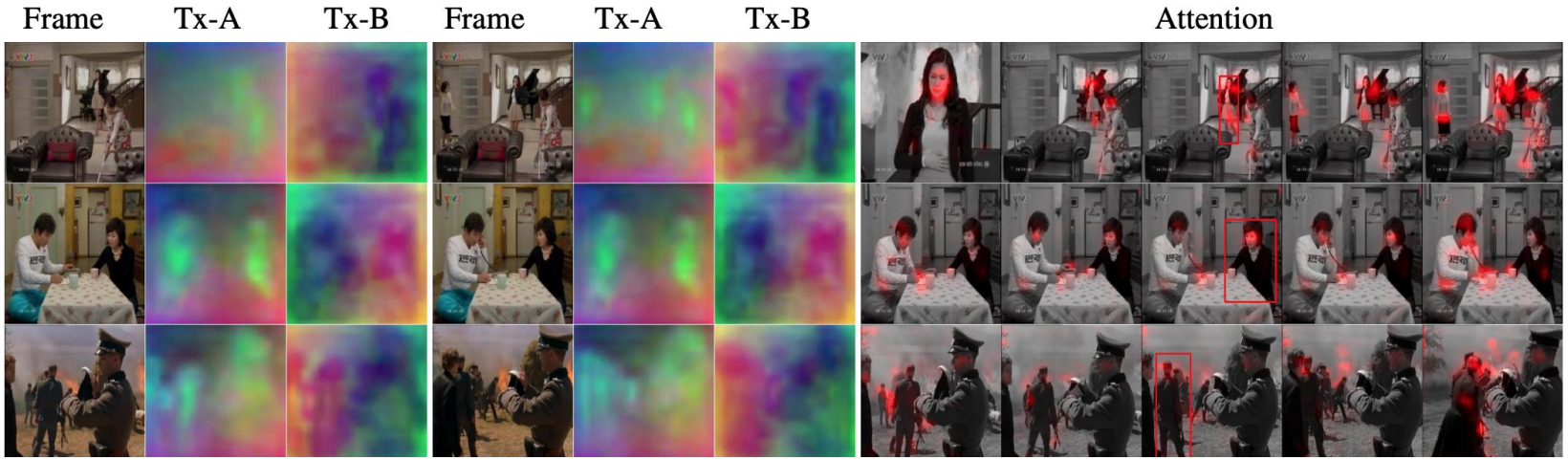
- **Trunk:** generate features and region proposals (RP) for the people present (using I3D)
- **Action Transformer Head:** use the person box from the RPN as a 'query' to locate regions to attend to, and aggregates the information over the clip to classify their actions



Rohit Girdhar, et al. "Video Action Transformer Network." CVPR 2019

Video Action Transformer Network

- Visualizing the key embeddings using color-coded 3D PCA projection
  - Different heads learn to track people at different levels.
  - Attends to face, hands of person, and other people/objects in scene



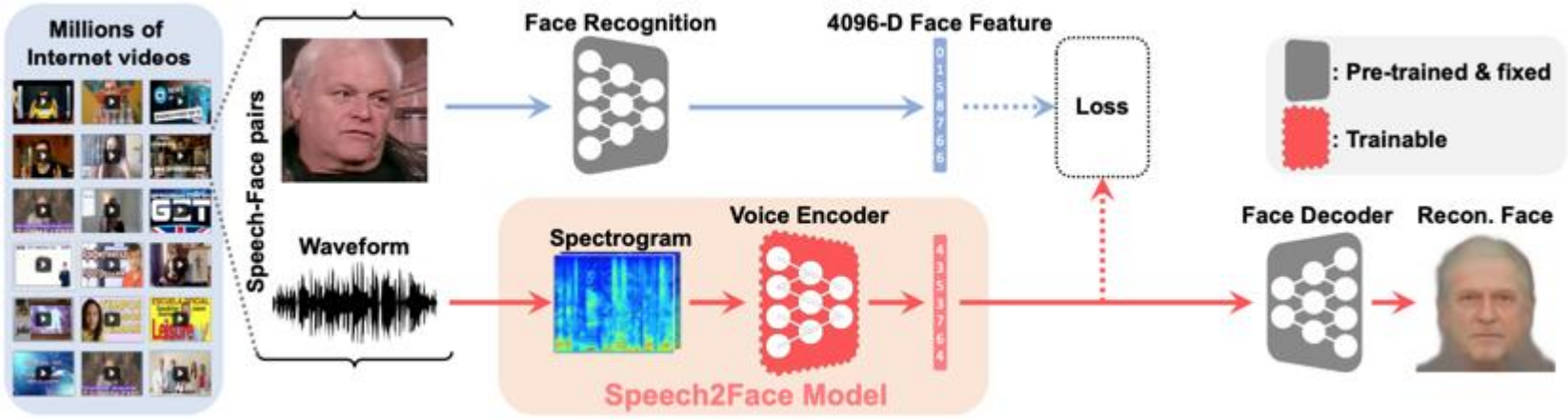
Rohit Girdhar, et al. “Video Action Transformer Network.” CVPR 2019

New Directions: Translation

Speech2face

Speech2face

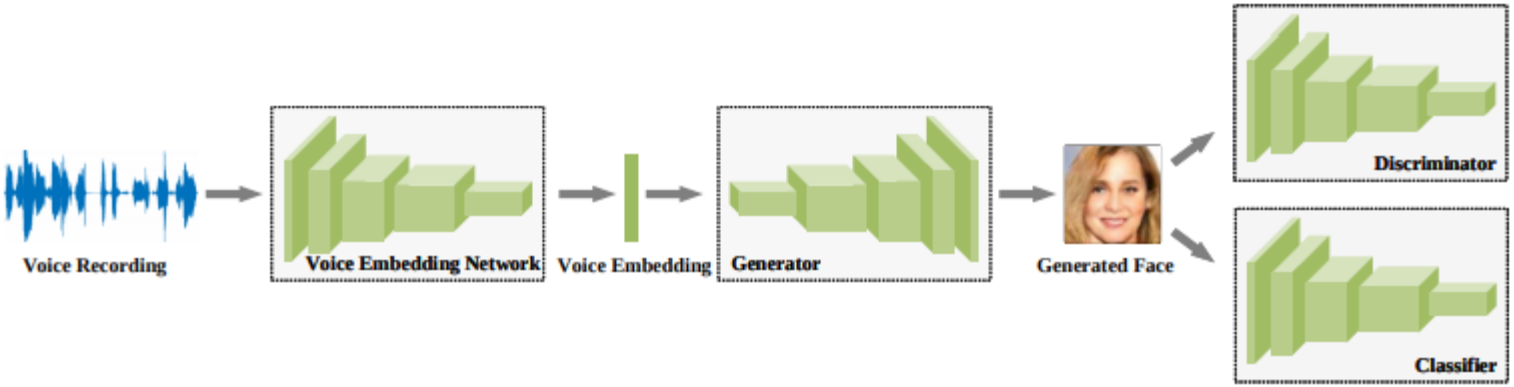
Voice encoder + face encoder + face decoder



Speech2face

Examples of reconstructed faces

Reconstructing faces from voices



- Introduce task of reconstructing face from voice
- Two adversaries:

  (a) Discriminator to verify the generated image is a face

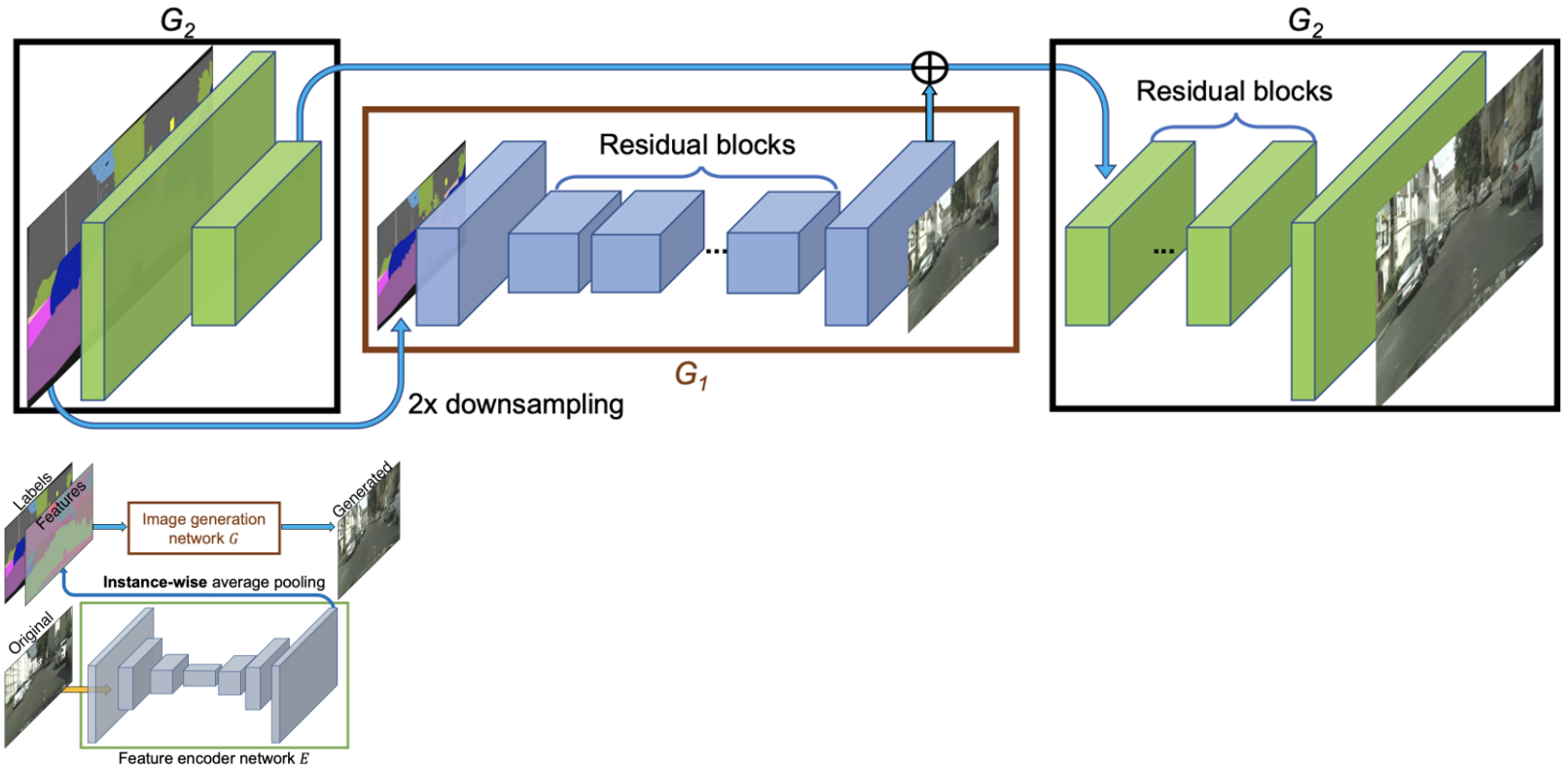
  (b) Classifier to assign a face image to the identity

Reconstructing faces from voices

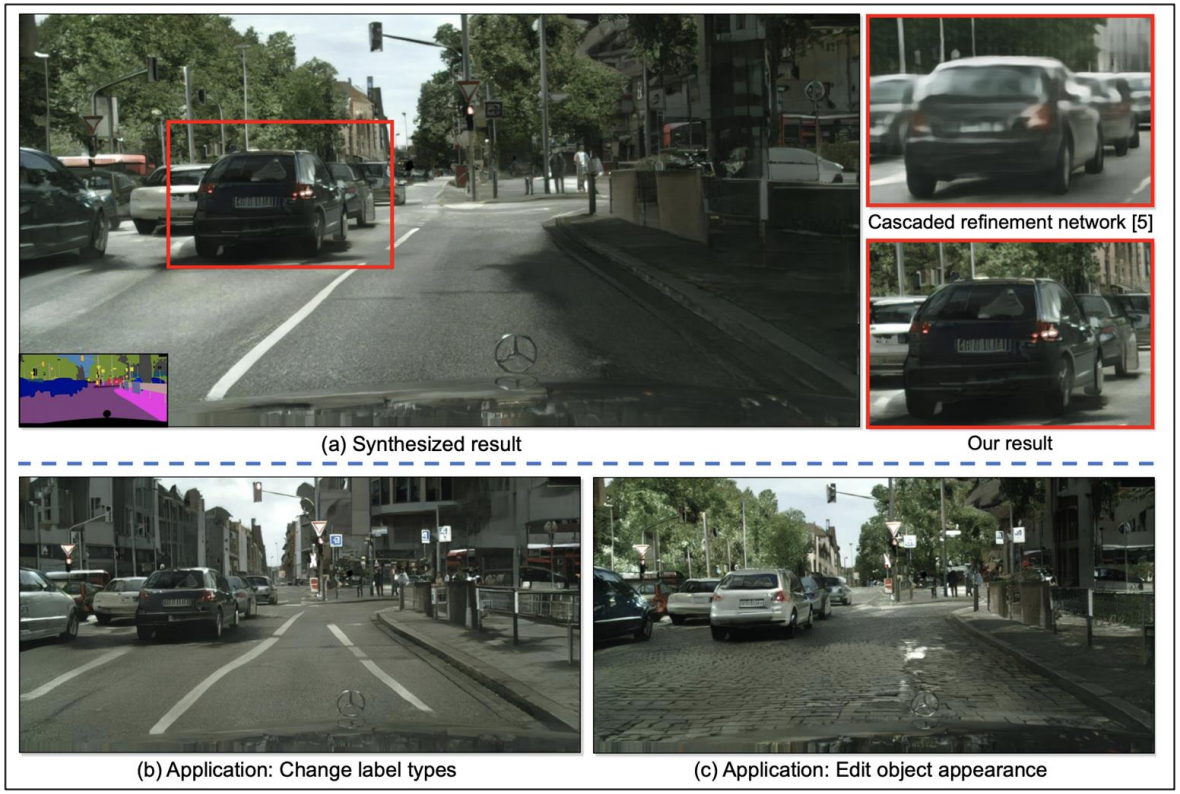
The generated face images have identity associations with the true speaker.

The produced faces have features(for ex. hair) that are presumably not predicted by voice, but simply obtained from their co-occurrence with other features

High-Resolution Image Synthesis and Semantic Manipulation with CGANs

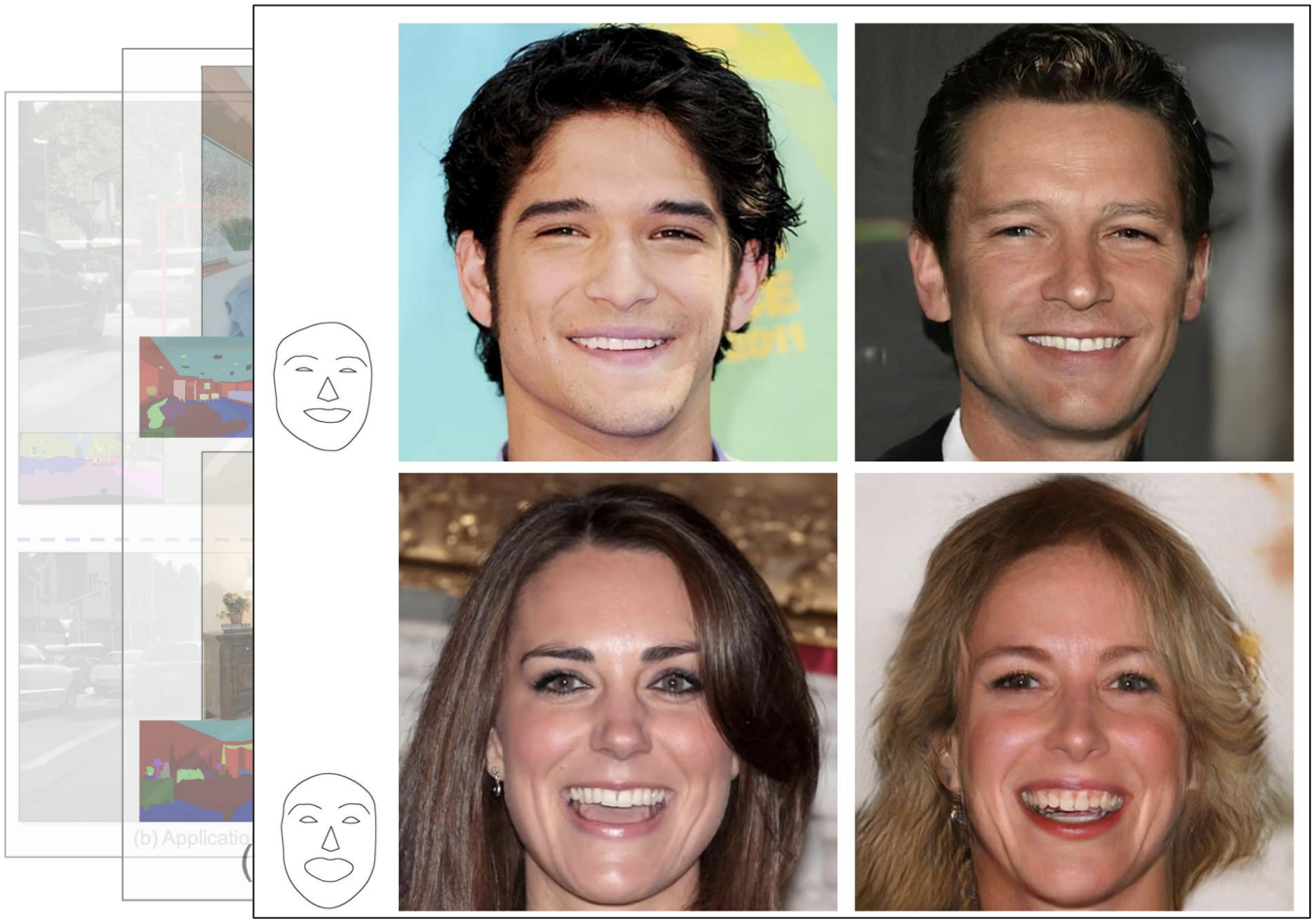


High-Resolution Image Synthesis and Semantic Manipulation with CGANs



High-Resolution Image Synthesis and Semantic Manipulation with CGANs

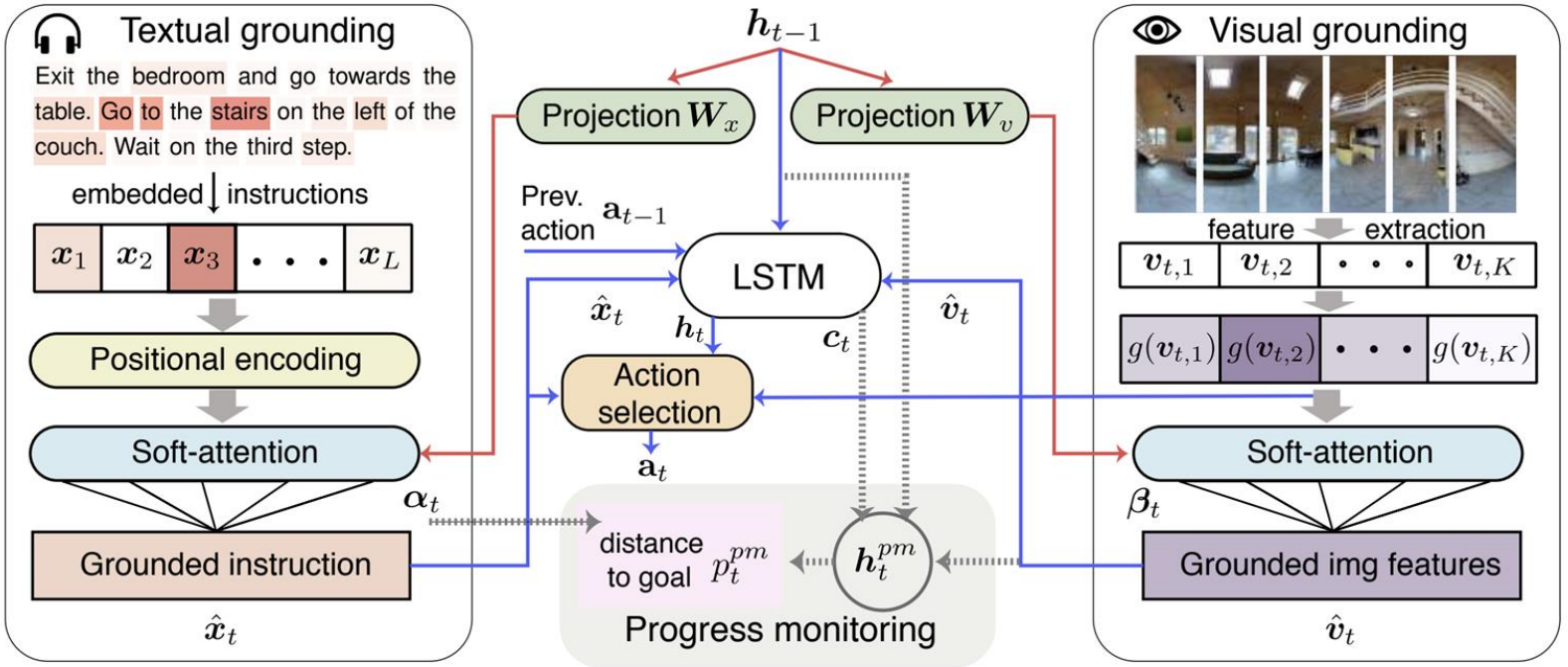
High-Resolution Image Synthesis and Semantic Manipulation with CGANs



Self-Monitoring Navigation Agent via Auxiliary Progress Estimation

Self-Monitoring Navigation Agent via Auxiliary Progress Estimation

Model: Co-grounded Attention Streams



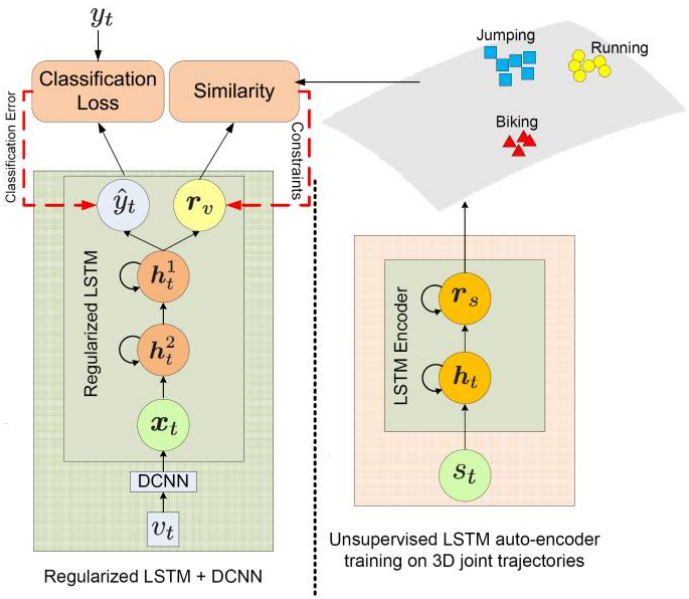
Self-Monitoring Navigation Agent via Auxiliary Progress Estimation

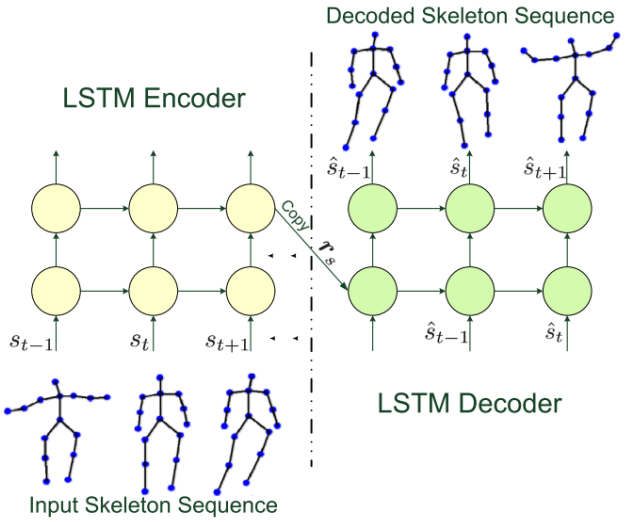
Results: Progress Monitoring

New Directions: Co-Learning

Regularizing with Skeleton Seqs

- Better unimodal representation by regularizing using a different modality

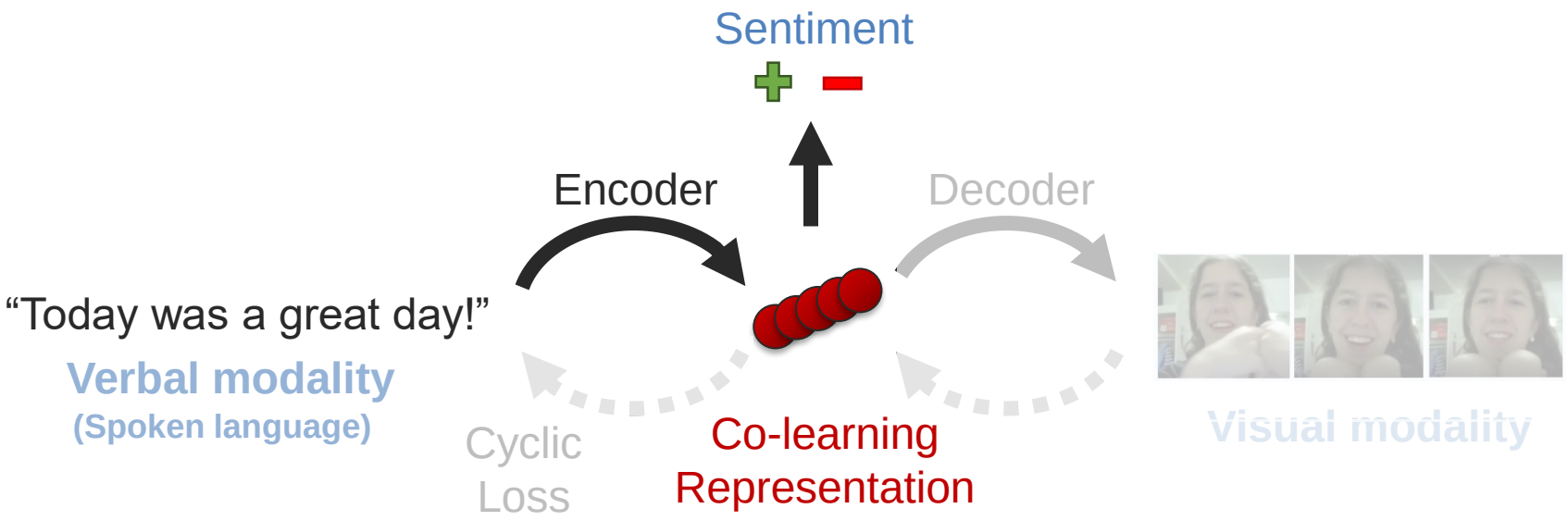




Non parallel data!

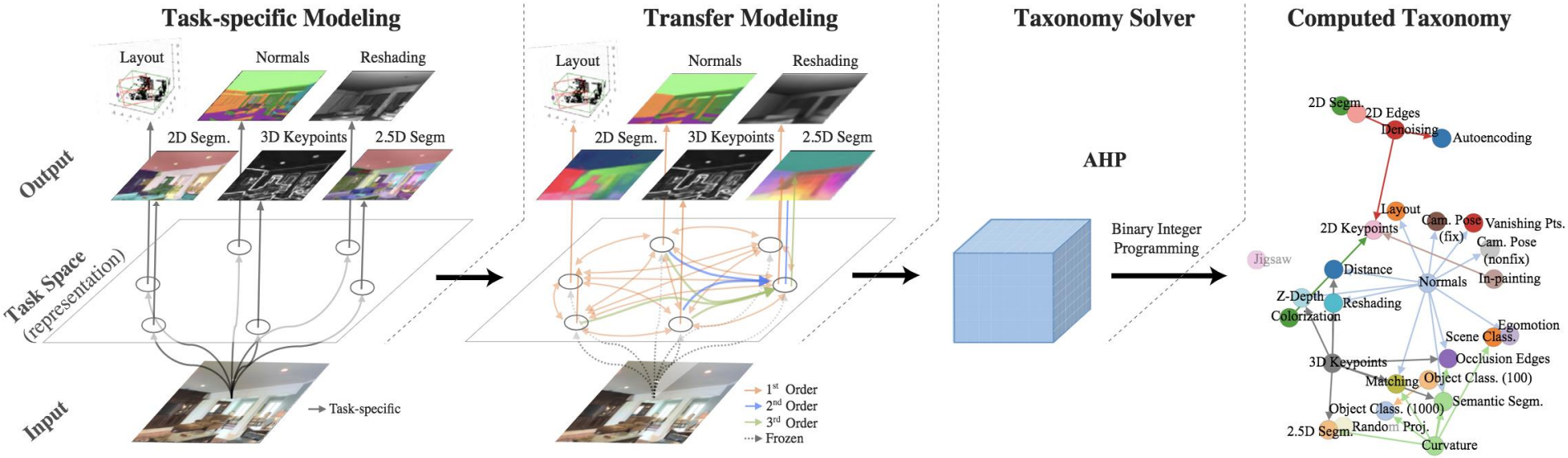
[B. Mahasseni and S. Todorovic, "Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition," in CVPR, 2016]

Multimodal Cyclic Translation



Paul Pu Liang*, Hai Pham*, et al., "Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities", AAAI 2019

Taskonomy

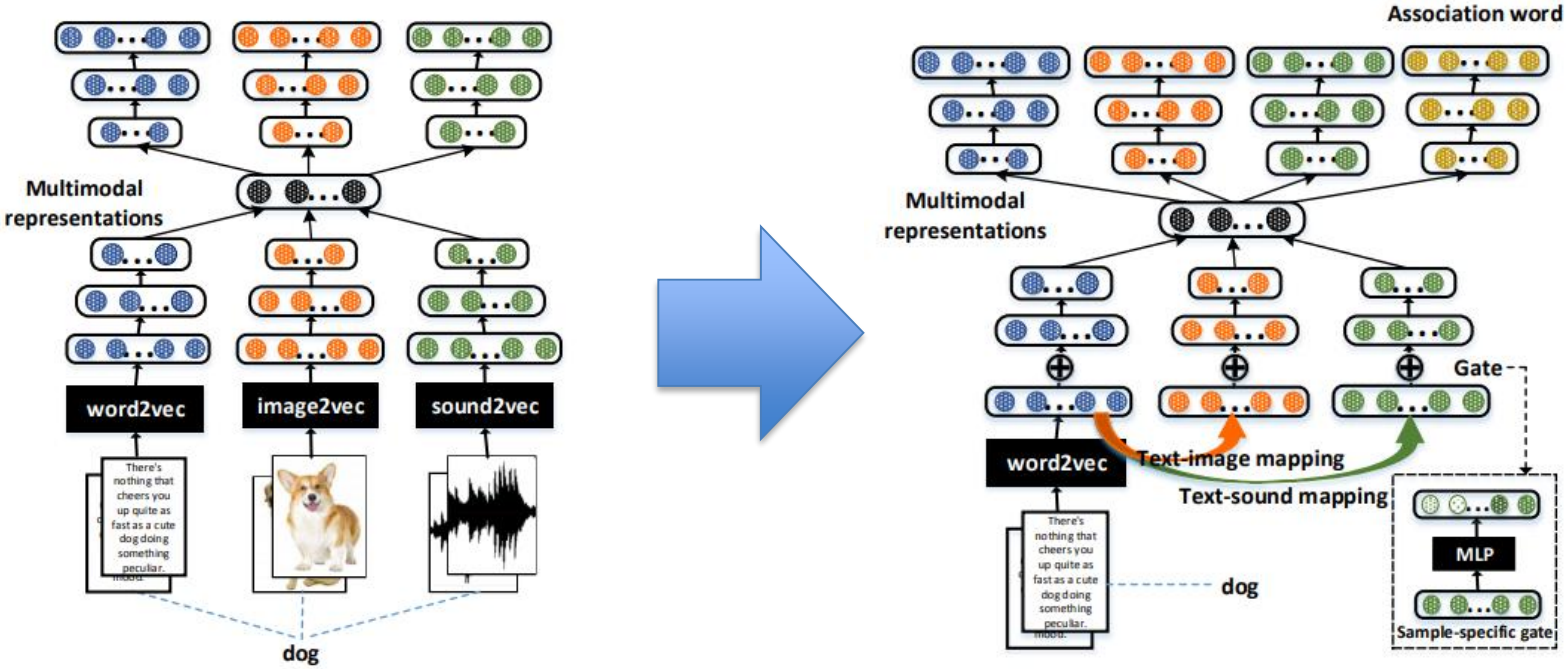


Process overview. The steps involved in creating the taxonomy.

Zamir, Amir R., et al. "Taskonomy: Disentangling Task Transfer Learning." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

Associative Multichannel Autoencoder

- Learning representation through fusion and translation
- Use associated word prediction to address data sparsity



[Wang et al. Associative Multichannel Autoencoder for Multimodal Word Representation, 2018]

Grounding Semantics in Olfactory Perception

- Grounding language in vision, sound, and **smell**

Olfactory-Relevant Examples					
MEN		sim	SimLex-999		sim
bakery	bread	0.96	steak	meat	0.75
grass	lawn	0.96	flower	violet	0.70
dog	terrier	0.90	tree	maple	0.55
bacon	meat	0.88	grass	moss	0.50
oak	wood	0.84	beach	sea	0.47
daisy	violet	0.76	cereal	wheat	0.38
daffodil	rose	0.74	bread	flour	0.33

	Chemical Compound				
Smell label	Phenethyl acetate	Isoamyl butyrate	Anisyl butyrate	Myrcene	Syringaldehyde
Melon	✓	✓			
Pineapple	✓				✓
Licorice			✓		
Anise			✓	✓	
Beer				✓	✓

Table 2: A BoCC model.

[Kiela et al., Grounding Semantics in Olfactory Perception, ACL-IJCNLP, 2015]