SRI International

SRI Center for Vision Technologies

Rakesh (Teddy) Kumar, Supun Samarasekera,
Ajay Divakaran, Michael Piacentino

Center for Vision Technologies
SRI International, Princeton NJ

October 31, 2018

Mission

World-changing solutions making
people safer, healthier, and more productive.

Mission

World-changing solutions making
people safer, healthier, and more productive.

Army
DARPA
Defense Threat
Reduction Agency

Dept. of Defense
Dept. of Education
Dept. of Energy
Dept. of Homeland Security

I-ARPA
National Guard
National Institutes of Health
National Science Foundation

Independent research center

$540 million annual revenues	2,100 staff members	21 locations worldwide

SRI spin-off ventures

Information Technology

CIC. · DESTI · discern · KASISTO · kuato studios · NUANCE

PACKETHOP · PSC · PYRAMID VISION · SARIF · SENSAR · Siri

SocialKinetics · teachscape · tempo · TOUT · trap!t · VideoBrush Corporation

Advanced Materials

ARTIFICIAL MUSCLE INCORPORATED · Averatek · colorep · lamina · LIGHTSCAPE materials, inc.

Princeton Lightwave · SP

Biomedical

DELSYS · intuity MEDICAL · LOCUS Pharmaceuticals · ORCHID CELLMARK · REDCOAT SOLUTIONS · Songbird

Robotics

GrabIT · INTUITIVE SURGICAL · redwood robotics

Information & Computing Sciences Division

- **$80M revenue**
- **250 staff members**
- **Four renowned laboratories**
  - Artificial Intelligence Center
  - Center for Vision Technologies
  - Speech Technology & Research Lab
  - Computer Science Lab
- **Leader in commercialization, ventures & licensing**

First computer mouse

First ARPANET nodes

.com
.gov
.org

First domain names

Emmy Awards for HDTV

NUANCE KASISTO dynaspeak® Pathway Tools

Siri tempo eduspeak® Sensay Analytics™

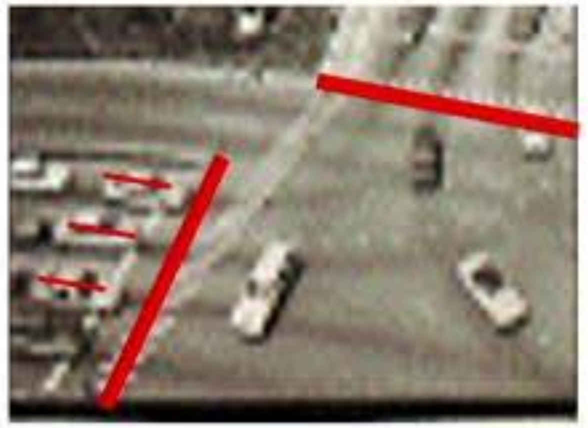
Center for Vision Technologies

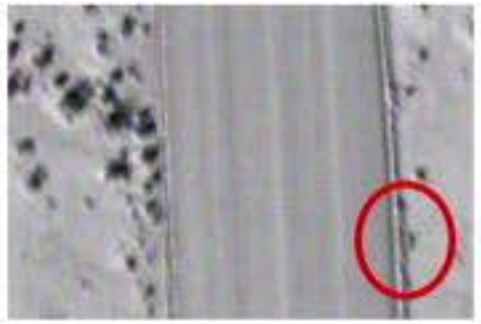
- 82 staff members
- 30 year history in Real Time Computer Vision
- 150+ patents

Some Accomplishments

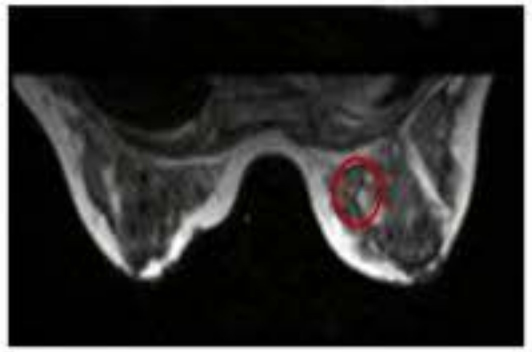
First real time AR broadcast on live TV 1994: Ads in Baseball Games >> 10 Yard Line in Football


VideoBrush: First ever live Video Mosaicing (now part of all Android phones)


Live traffic Monitoring, deployed all over the country


IED Detection

Currently saving lives in theatre

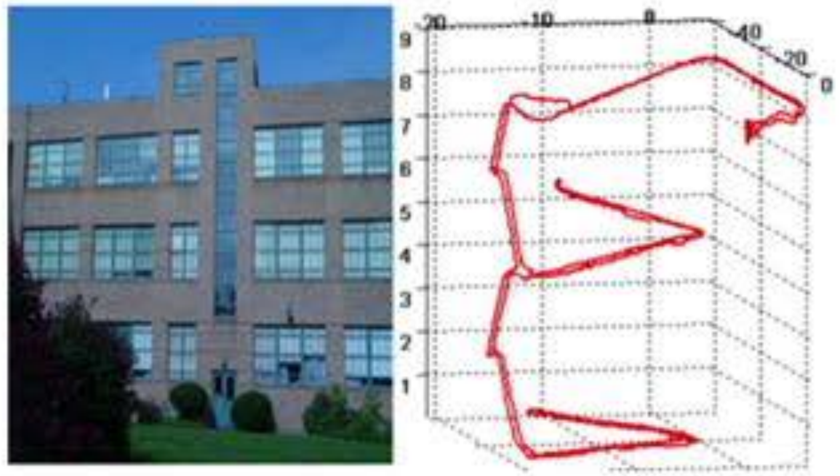

Breast Cancer: MRI based Tumor


Skin Measurement

Center for Vision Technologies

Leading Platforms

- **Computational Sensing**
  - Embedded Vision
- **2D/3D reasoning**
  - GPS denied navigation
  - 3D modeling/ mapping
  - Augmented reality
  - Surveillance
- **Data analytics**
  - Image search
  - Fine grain recognition
  - Activity understanding
  - Social Media reasoning
- **Human behavior modeling**
  - Emotion Detection
  - Biometrics
- **Machine Learning**
  - Explainable AI
  - Lifelong Learning



GPS Denied Navigation (Human, Robots, Vehicles, Aerial, Naval etc.)

Navigation & Mapping for Autonomous Vehicles

First ever Augmented Reality binoculars

Object detection, recognition, Search based on image/ video content

Human Behavior Modeling: Social interaction and communication with computers

Driver State Monitoring: Toyota Concept Car

Center for Vision Technologies

Leading Platforms

Intelligent Mobile Platforms

Real time edge based autonomous and augmented systems: robots, vehicles, people worn, augmented reality.

- **Computational Sensing**
  - Embedded Vision
- **2D-3D reasoning**
  - GPS-denied navigation
  - 3D modeling/mapping
  - Augmented reality
  - Surveillance
  - Change Detection

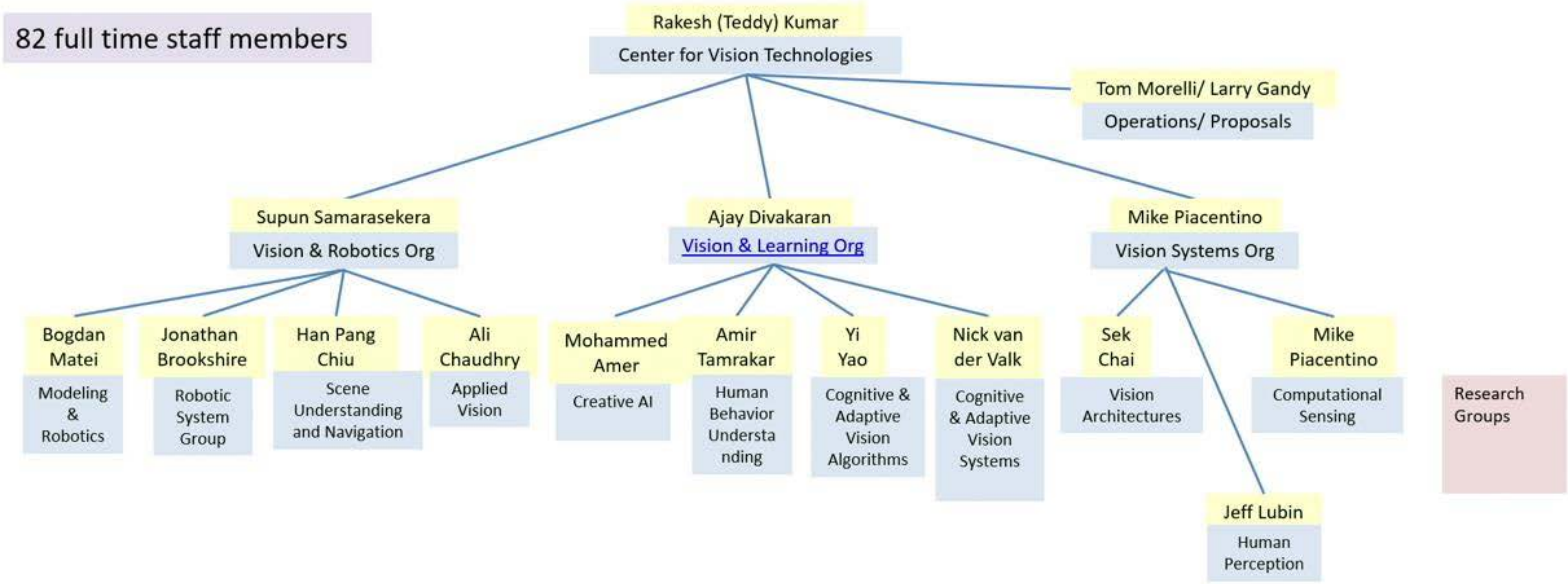
Human Understanding and Human Computer Interaction

- **Real-time Interactive Systems**
  - Operator State Assessment using multi-modal sensors (2D, 3D etc.)
    - Emotion Detection
  - Communicating with Computers using multi-modal sensors
  - Biometrics
  - Human activity understanding based on vision and other multi-modal sensors

Multi-modal Data Analytics and Machine Learning

- **Cloud-based Processing**
  - Image and Video search, Activity Recognition
  - Fine grain recognition using 2D and 3D sensors
  - Multi-modal Social Media Analytics
  - Explainable AI
  - Lifelong Learning
  - Creative AI

Center for Vision Technologies Organization Chart



SRI International

Vision and Learning:CVT Major Projects

Presenter: Ajay Divakaran

SRI International,

Princeton, NJ

October 31st, 2017

Content Understanding vs. Reaction

"A riot took place two days back in connection after *abc* event"

UNDERSTANDING

Semantic (visual)
1. People
2. Police
3. Fight
4. Camera

Text
1. Riot
2. abc event

Audio
1. Shouting
2. Angry

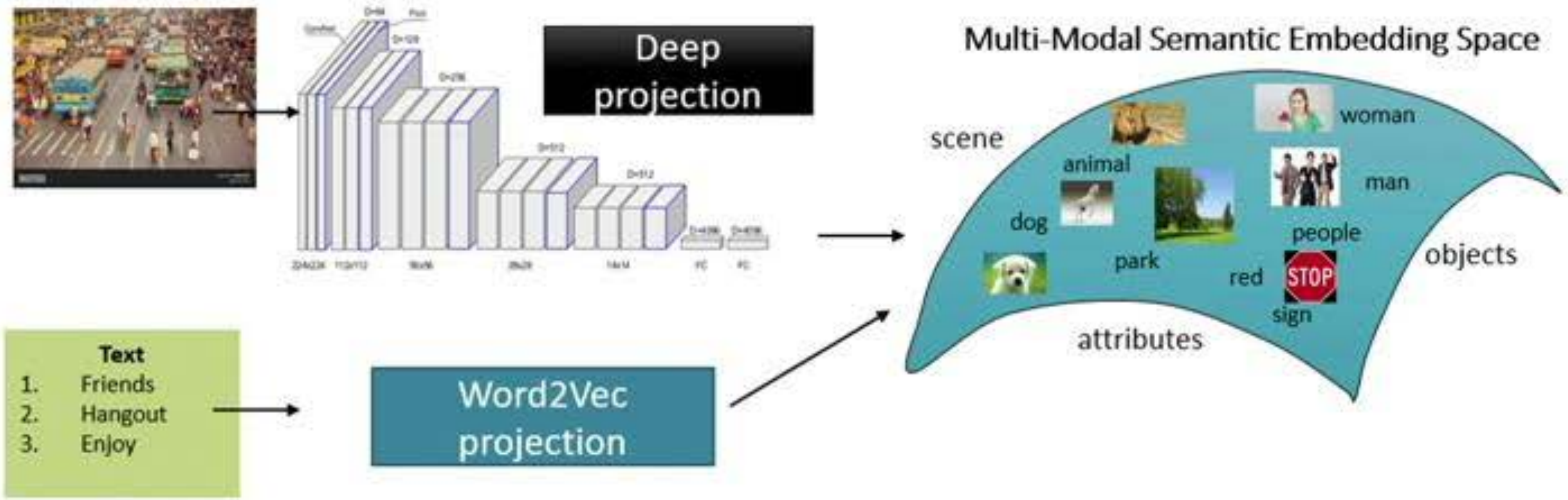
Sentiment (visual)
1. Anger
2. Stress
3. Unhappy

Symbolic (visual)
1. Style
2. Popularity

REACTION

Positive Negative Groups

Multimodal Embeddings



- Jointly embed paired items from different modalities in a common space [1]
- Loss enforces that co-occurring pairs are pulled closer and vice-versa [2]. Learning is loosely unsupervised
- Advantage: Leverage continuity of label space to handle new concepts by situating them among known concepts

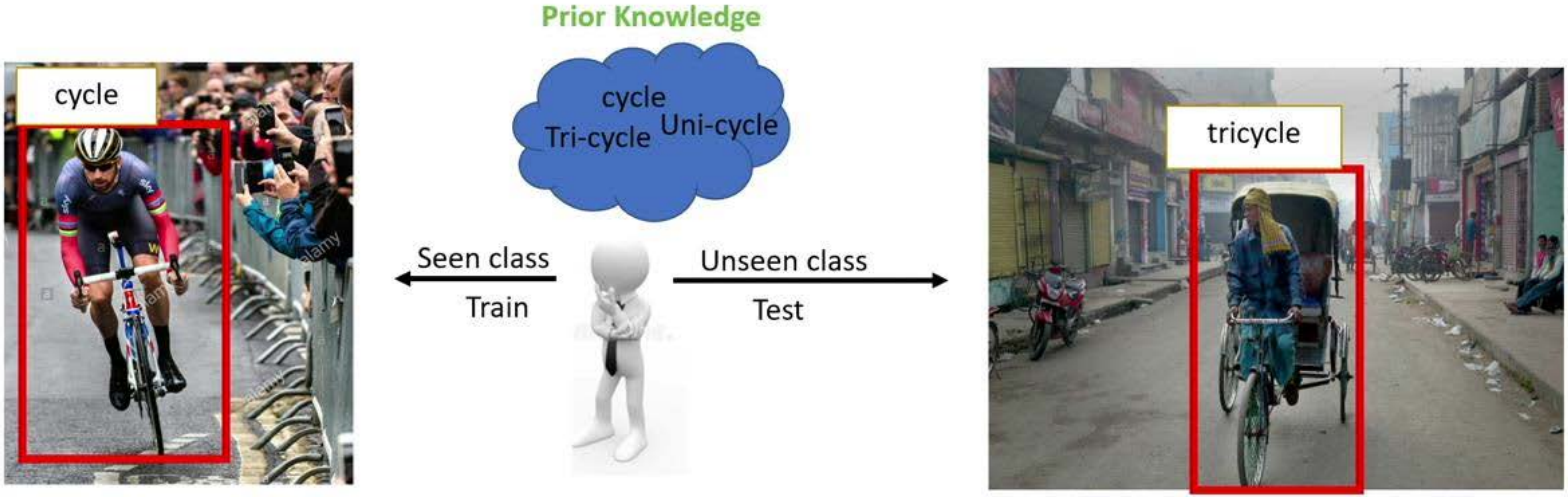
1. Facenet: A unified embedding for face recognition and clustering
2. Devise: A deep visual-semantic embedding model

Multimodal Embeddings Research Questions

- Recently multimodal embeddings leveraged for multiple tasks
  - Zero-shot learning
  - Captioning and VQA
  - Learn better word embeddings
- How far can we push the limits of learning in multimodal space (quantity and quality of data)? Push the tasks that are currently possible
- Is it possible to learn more than 2 modalities and how do they support each other [1]?
- Can we embed users and content within the same space?

1. TED- Can We Create New Senses For Humans, David Eagleman

Zero-Shot Object Detection



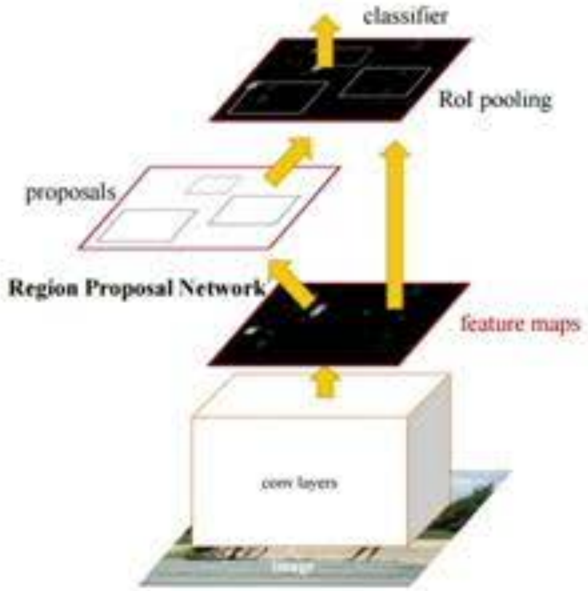
Joint work with Ankan Bansal*, Gaurav Sharma, Rama Chellappa and Ajay Divakaran

European Conference on Computer Vision 2018

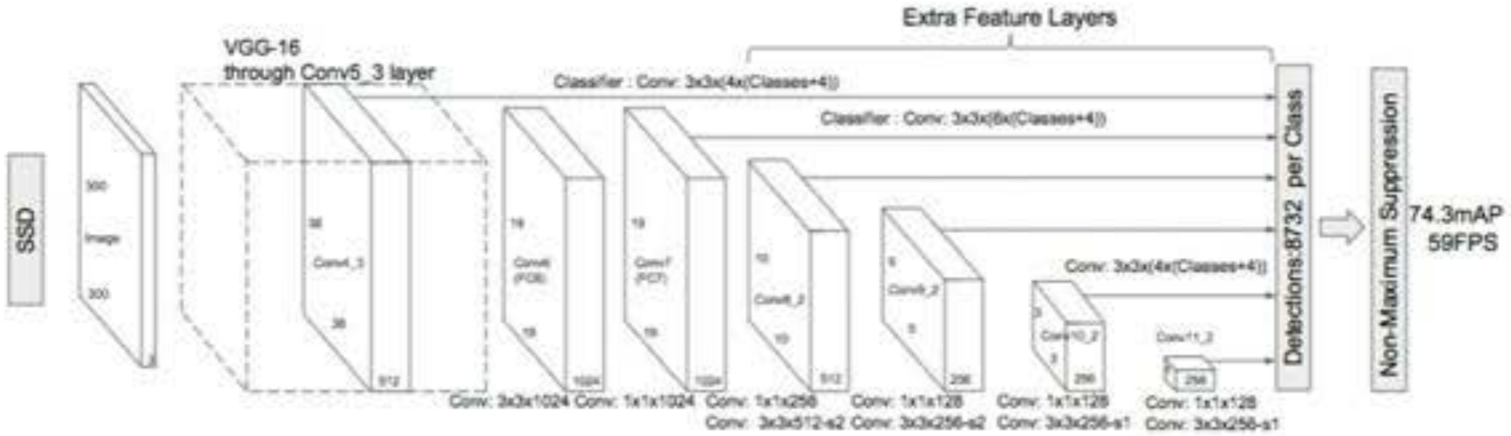
* Ankan Bansal was an intern at SRI

Overview

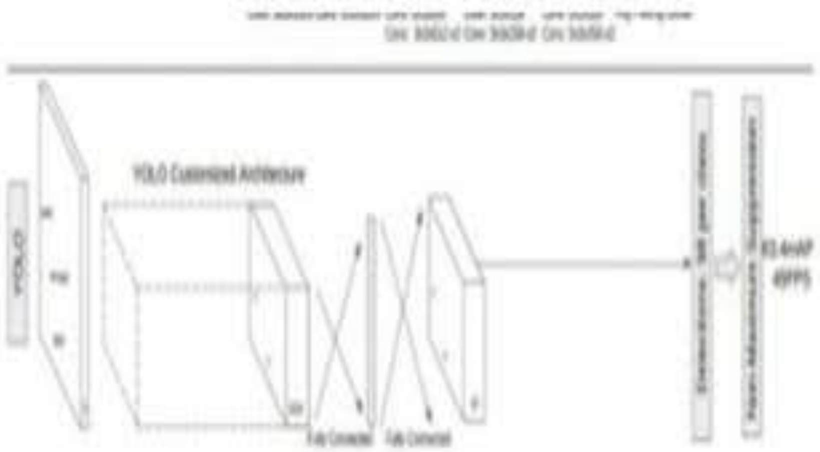
- Deep learning has resulted in significant progress in object detection
- But current methods require a few thousand instances per class for training
- Currently not possible to scale beyond few 100 object classes and impossible to detect novel objects- zero-shot learning



Faster RCNN, Ren et al.



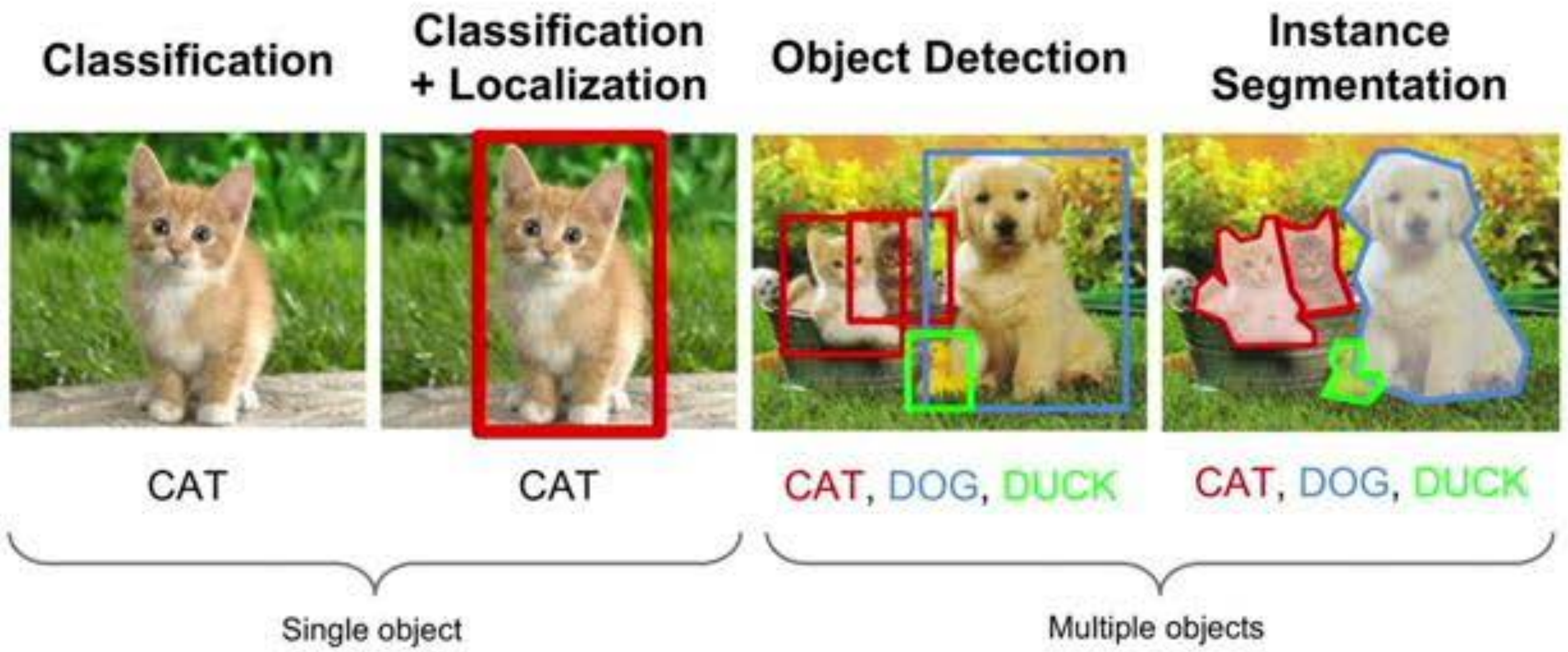
SSD, Liu et al.


YOLO, Redmon et al.

What is Zero-Shot (ZS) Learning

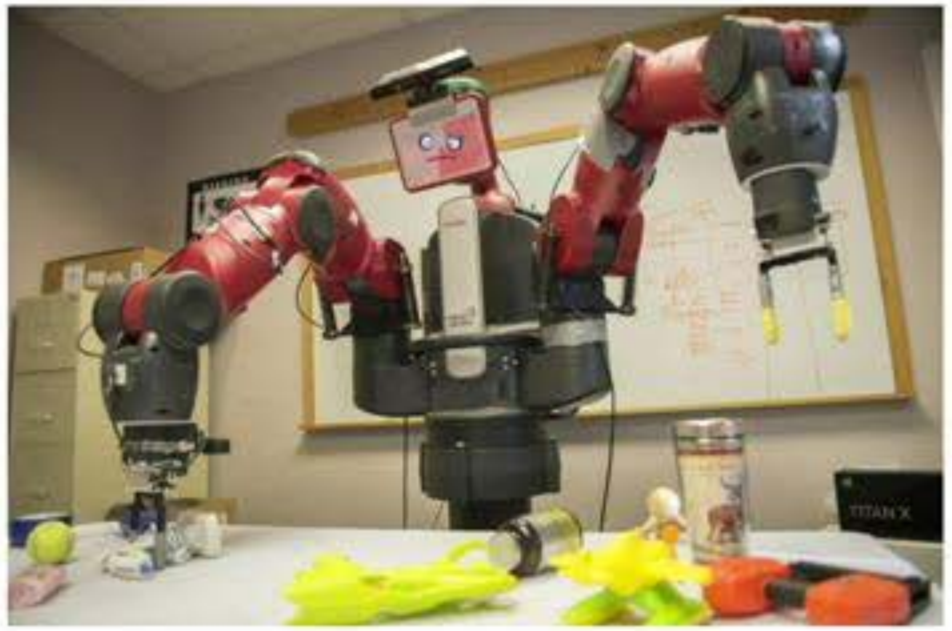
- Training: Learn models on example from "seen" classes
- Testing: Make predictions on examples from "unseen" classes
- Assumption: Unseen classes are related to seen classes semantically. For example "tri-cycle" is related to "cycle"
    - Relationships used to transfer models from seen to unseen classes
- Prior works have focused largely on zero-shot classification

From ZS Classification to Detection



- Detection is harder compared to classification:
  - Requires localization of all object instances in an image
  - Classification can often be done with contextual cues- which may not work for detection
- Invariances to occlusion, viewpoint, clutter etc. is required for accurate detection

Real-World Applications



Robotics
Function in unknown settings

Surveillance
Detect new objects in new environments

- Humans can easily scale up to 1000s of categories
  - Can also build a mental image of a new object based on prior knowledge
- Do we really need 1000s of training examples for a new category?

Zero-Shot Detection (ZSD)

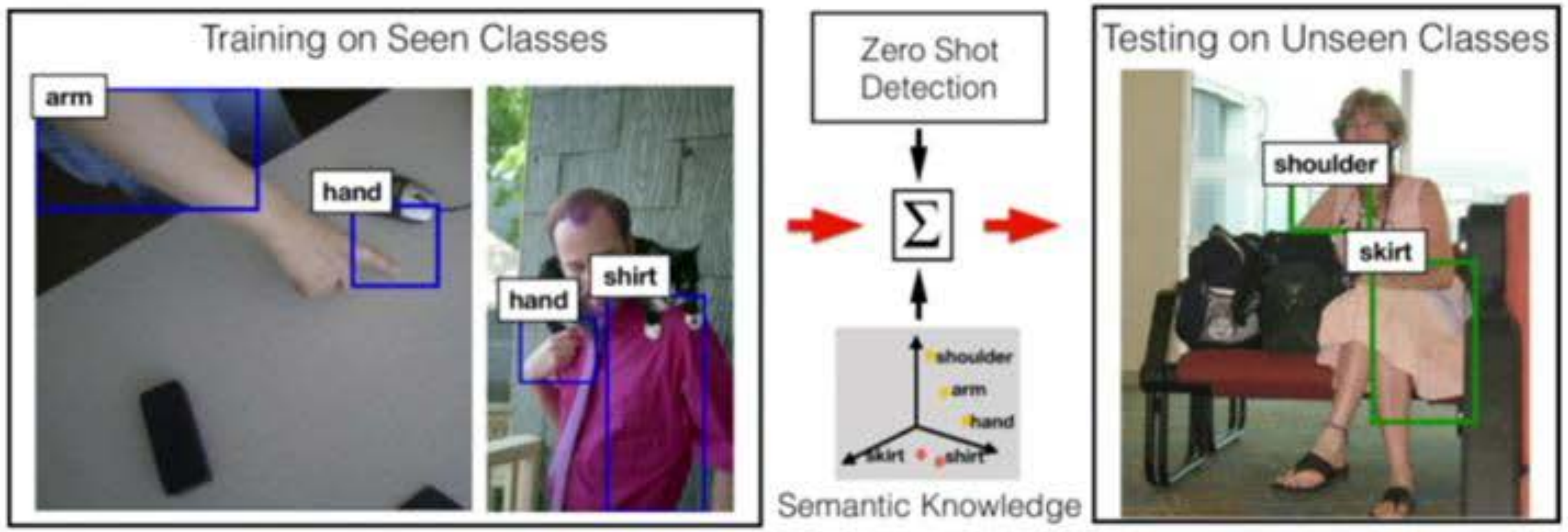
1. Introduce and target the challenging problem of ZSD
   - Extend prior work in ZSL for ZSD task
   - Working with real-world images with significant variations in views, clutter.

2. Modeling background for ZSD
   - Background class is added to improve performance in classicial detection models
   - But background in ZSD could be actual background (“stuff” classes) or unseen classes
   - Re-think and propose two methods

3. Propose a method to improve transfer via semantic knowledge by densely sampling the semantic space

Baseline Approach

- Build upon prior ZS methods that embed image features and class-labels in a common space
  - Knowledge is transferred via the semantic relatedness between class-labels
- Use RCNN architecture to compute features for a box and embed in word2vec space
  - Replace RCNN with any detection method



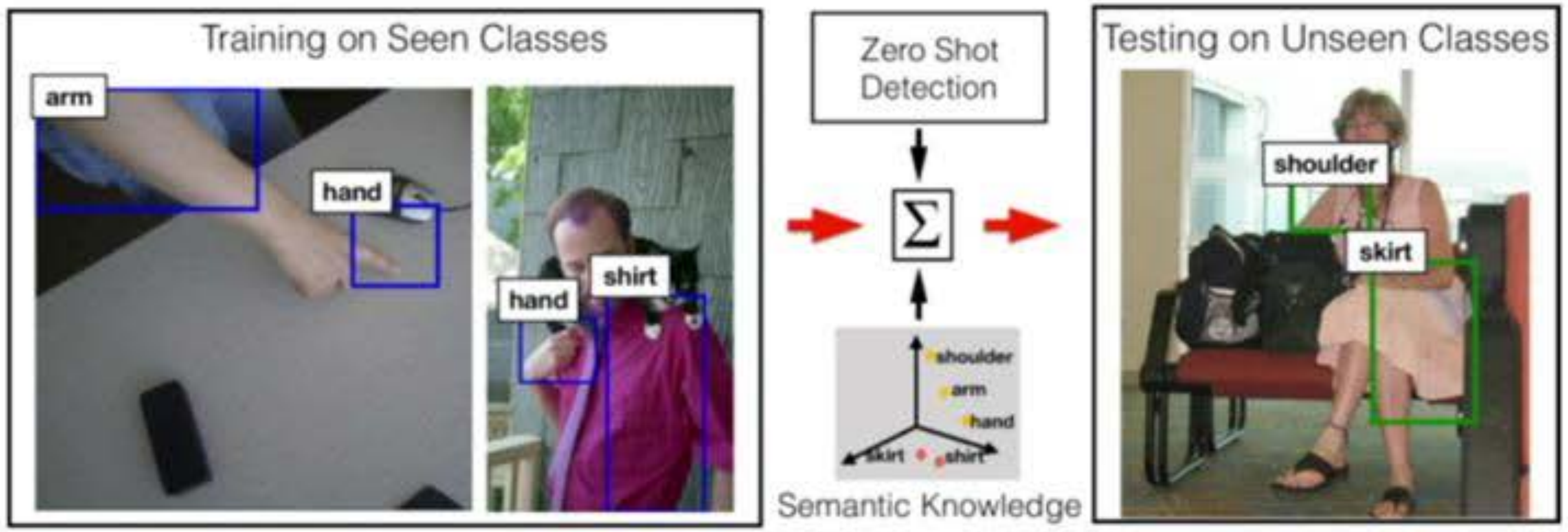
Leveraging Multimodal Embeddings for Social Media Analytics

- Interested in **UNDERSTANDING** posted content and their **REACTIONS** on social media platforms
- Why multimodal content
  - Posted content is increasingly multimodal e.g. 350 M photos uploaded daily on Facebook [1]
  - "A Picture is worth a thousand words" (image posts get 179% more interaction than an average post)
  - Multimodality can be used for improving understanding and filling the gaps in other modalities
- Why- Detect undesired content, identify communities of malicious users, track events, understand group dynamics

1. http://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9

Baseline Approach

- Build upon prior ZS methods that embed image features and class-labels in a common space
  - Knowledge is transferred via the semantic relatedness between class-labels
- Use RCNN architecture to compute features for a box and embed in word2vec space
  - Replace RCNN with any detection method



Baseline ZSD Approach

- Project deep features from boxes $\phi(b_i)$ using a linear projection

$$\psi_i = W_p \phi(b_i)$$

- Compute similarity between i[th] box and j[th] class label using cosine-similarity S_{ij}

- Ranking-loss to push embeddings for similar boxes and class labels together and vice-versa

$$\mathcal{L}(b_i, y_i, \theta) = \sum_{j \in \mathcal{S}, j \neq i} \max(0, m - S_{ii} + S_{ij})$$

- Predict test label of a bounding box by computing similarities with unseen classes

$$\hat{y}_i = \arg\max_{j \in \mathcal{U}} S_{ij}$$

Background-Aware ZSD

- Prior detection models with fixed number of classes add an additional background class to improve performance
  - Learn from proposals that do not contain a foreground class
  - Improves discrimination for hard-proposals (those which look similar to actual classes)
- Definition of background for ZSD is not clear
  - Does it contain “stuff” e.g. sky, ground etc.
  - Or Unseen objects
- Modeling background may help performance but how?

Statically Assigned Background (SB) based ZSD

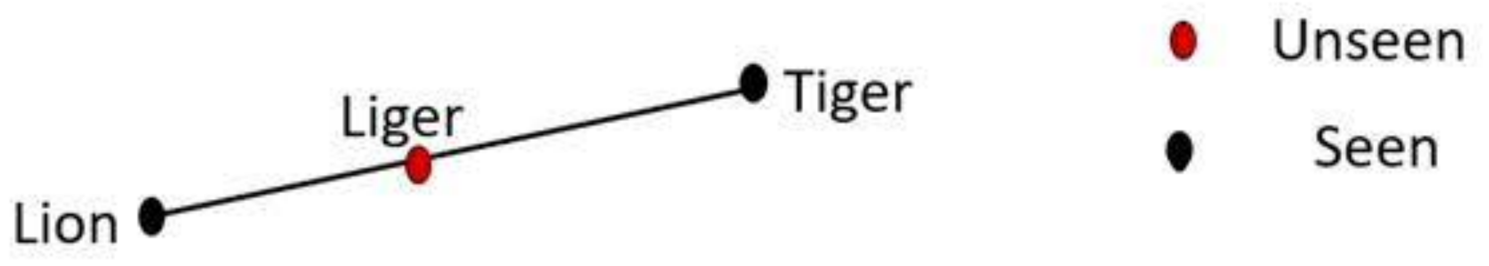
- Model as natural extension of prior detection method
- Added a fixed background vector [1……0] and assign background proposals to this class
- Limitations
  - Does not align with the structure imposed by semantic embeddings where each class is semantically related to other classes
  - Pushing all background boxes to a single monolithic vector is not optimal
- Propose a method based on latent assignments

Latent Assignment (LAB) based ZSD

- Spread background boxes across the embedding space instead to a single class
- Propose an EM style method that assigns latent classes to the background boxes:
  - Repeat – (1) latent assignment to background boxes, (2) model learning
  - Similar to semi-supervised learning
- Explicitly encode knowledge that background boxes do not belong to seen classes but to the set of remaining classes (background set)
  - Background set is obtained by removing seen classes from a larger set of classes in semantic embedding space

Densely Sampled Embedding Space (DES)

- Current methods piggyback on paired samples from seen classes to align visual example and class label
- Often lead to sparse sampling of the embedding space, resulting in weak alignments (continuous space)



- Propose to augment training dataset with samples from additional classes (no overlap with unseen) to densely sample the embedding space
  - Use large OpenImages dataset with bounding boxes for 545 classes

Experiments

- Use datasets with real-world images for training and testing
  - More than one object per image (different from most prior ZS works)
- Create splits* from MSCOCO and Visual Genome (VG)
  - Cluster semantic embeddings for classes (80% classes for training and 20% for testing)

Dataset	# Seen classes	# Unseen classes	Training samples
MSCOCO	48	17	73,774
Visual Genome	478	130	54,913

- For DSES we use OpenImages that contains 1.5M images spanning 545 objects

* Splits are public at http://ankan.umiacs.io/zsd.html

Experimental Details

- Use Inception-V3 as base CNN and edgeboxes for extracting proposals
- 300 dimensional pre-trained vectors as semantic embeddings
- Positive Boxes: IoU > 0.5 and Background boxes: 0 < IoU < 0.2 and few randomly chosen IoU = 0
- For LAB, we run 5 iterations of assign of background classes to background boxes and learning the model
- Report Recall@K: recall when only the top K detections (based on prediction score) are selected from an image

Results

ZSD Method	BG-aware	#classes			MSCOCO IoU			#classes			Visual Genome IoU		
		$\|\mathcal{S}\|$	$\|\mathcal{U}\|$	$\|\mathcal{O}\|$	0.4	0.5	0.6	$\|\mathcal{S}\|$	$\|\mathcal{U}\|$	$\|\mathcal{O}\|$	0.4	0.5	0.6
Baseline		48	17	0	34.36	22.14	11.31	478	130	0	8.19	5.19	2.63
SB	✓	48	17	1	34.46	24.39	12.55	478	130	1	6.06	4.09	2.43
DSES		378	17	0	**40.23**	**27.19**	**13.63**	716	130	0	7.78	4.75	2.34
LAB	✓	48	17	343	31.86	20.52	9.98	478	130	1673	**8.43**	**5.40**	**2.74**

- LAB performs best on VG
  - Latent assignments help spread the background boxes leading to better model
- SB performs better on MSCOCO (not on VG)
  - Due to our splits, the background boxes in MSCOCO didn't include unseen objects
  - Not possible for VG due to large number of objects. Leading to performance loss.

Results

ZSD Method	BG-aware	MSCOCO #classes $\lvert\mathcal{S}\rvert$	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{O}\rvert$	MSCOCO IoU 0.4	0.5	0.6	Visual Genome #classes $\lvert\mathcal{S}\rvert$	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{O}\rvert$	Visual Genome IoU 0.4	0.5	0.6
Baseline		48	17	0	34.36	22.14	11.31	478	130	0	8.19	5.19	2.63
SB	✓	48	17	1	34.46	24.39	12.55	478	130	1	6.06	4.09	2.43
DSES		378	17	0	**40.23**	**27.19**	**13.63**	716	130	0	7.78	4.75	2.34
LAB	✓	48	17	343	31.86	20.52	9.98	478	130	1673	**8.43**	**5.40**	**2.74**

- DSES performs best for MSCOCO
  - Significant gains
  - No of training classes increases by a factor 7.8 for MSCOCO
- DSES doesn't help for VG since no of classes are high apriori for VG
  - Leading to overfitting

Insights

MSCOCO

Good Classes		Bad Classes	
bus	couch	scissors	cat
52.70	47.52	0	3.86
cow	elephant	umbrella	tie
43.33	35.89	4.52	7.69

VisualGenome

Good Classes				Bad Classes		
laptop	skirt	car	cattle	bicycle	gravel	vent
48.54	35.00	33.56	29.41	0.19	0.80	0
kitten	building	cake	chair	garden	plant	zebra
33.33	32.41	29.93	28.67	0	0.22	0

- Trend for best performing classes same for standard object detectors
  - Mostly structured and well-defined objects like bus and cow
- Bottom classes such as vent, plant etc. are not usually well-defined and are more of “stuff” than “things” classes
- Some classes e.g. “zebra” not detected due to insufficient information during knowledge transfer
  - “zebra” is related to “giraffe” in semantic space. But model doesn’t know it has a lower neck and white-black stripes
  - Additional knowledge such as attributes might be helpful

Insights

MSCOCO

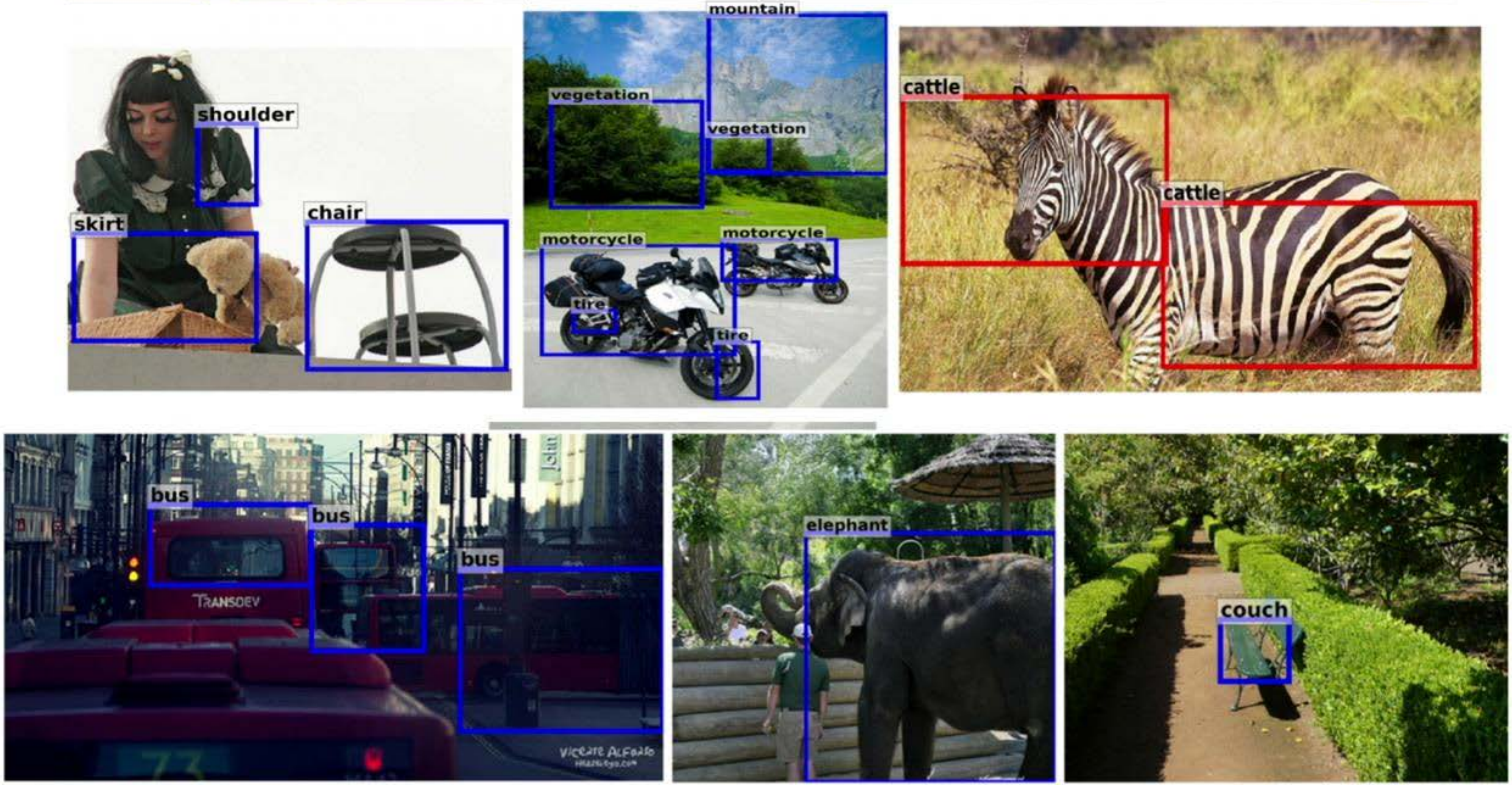
	Baseline			SB		
K↓ IoU→	0.3	0.4	0.5	0.3	0.4	0.5
All	47.91	37.86	24.47	43.79	35.58	**25.12**
100	43.62	34.36	22.14	42.22	**34.46**	**24.39**
80	41.69	32.64	21.01	41.47	**33.98**	**24.01**
50	36.19	27.37	17.05	**39.82**	**32.6**	**23.16**

VisualGenome

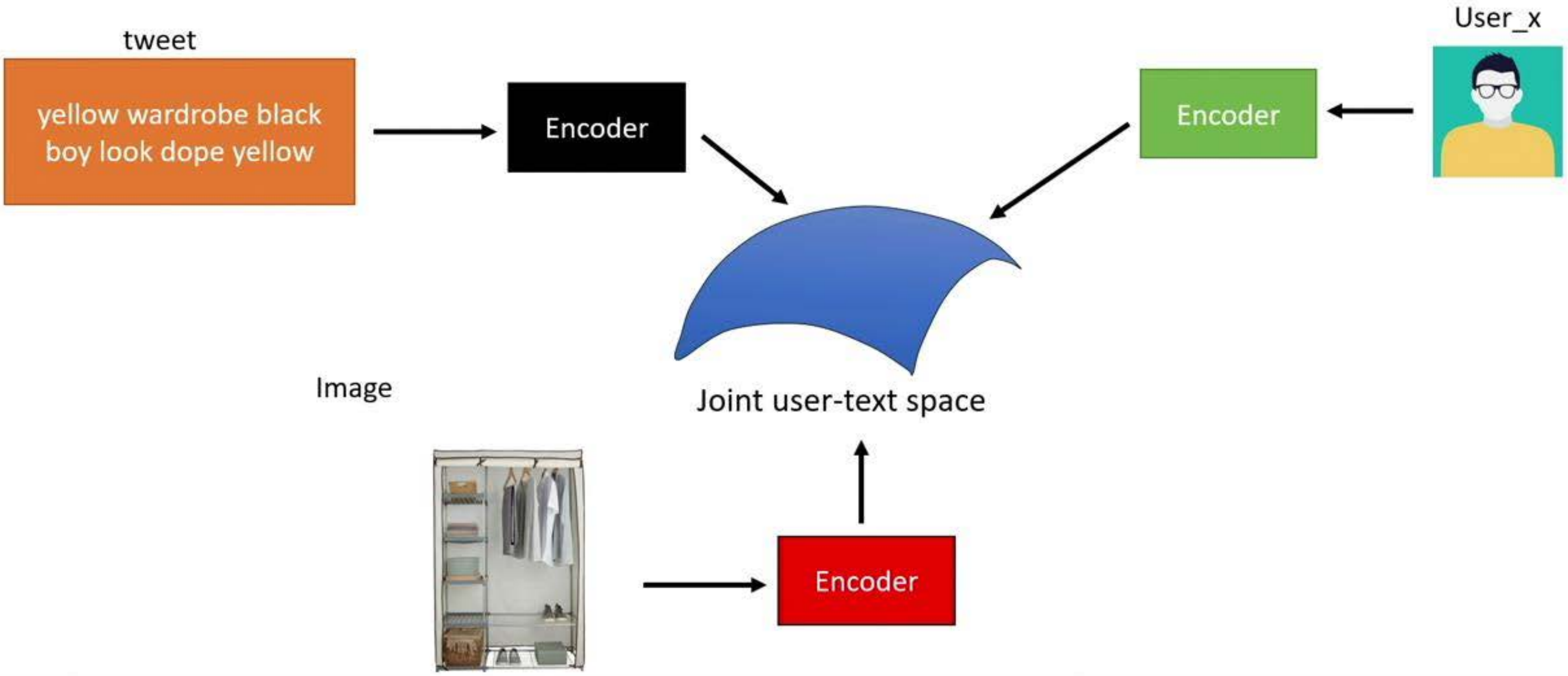
Baseline			LAB		
0.3	0.4	0.5	0.3	0.4	0.5
13.88	9.98	6.45	12.75	9.61	6.22
11.34	8.19	5.19	11.20	**8.43**	**5.40**
10.41	7.55	4.75	**10.45**	**7.86**	**5.06**
7.98	5.79	3.68	**8.54**	**6.44**	**4.14**

- Our background aware models performs better than baseline while predicting high-quality detections (higher performance in bottom right corner)
  - High quality detections = higher IoU and lower K
- Less difference between K=All and K=100
  - Top detections by our model are high quality


building
building
building
building
car
laptop
car
laptop
building
building
car
vegetation
car
vegetation
vegetation
car
car
poster
poster
window
building
poster
poster
poster
poster
road
writing
writing
floor
chair
cow
cow
cow
cow
cow
bus
Not In Service
dog


shoulder
skirt
chair
mountain
vegetation
vegetation
motorcycle
motorcycle
tire
tire
cattle
cattle
bus
bus
bus
TRANSDEV
elephant
couch

Understanding Social Media Content and their Reactions using Multimodal Embeddings



Motivation

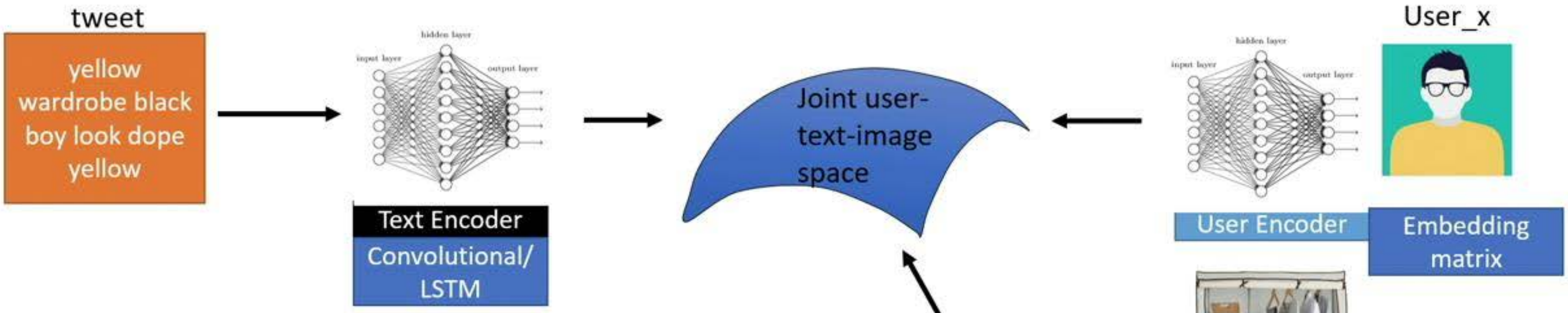
- Lots of unstructured content- text, images and videos- posed on social media
  - New language of self-expression [1]
- Can we simultaneously characterize users and understand the posted content
  - Prior works are limited and generally tackle a proxy task e.g. measuring persuasiveness of content using text/visual cues and might need curated labels [3]
  - Do not explicitly understand content i.e. what are the underlying semantics and sentiments
  - Limited in the number of semantic concepts- social media topics can be wide ranging
- How far can we reason about specific concepts such as particular leader or event

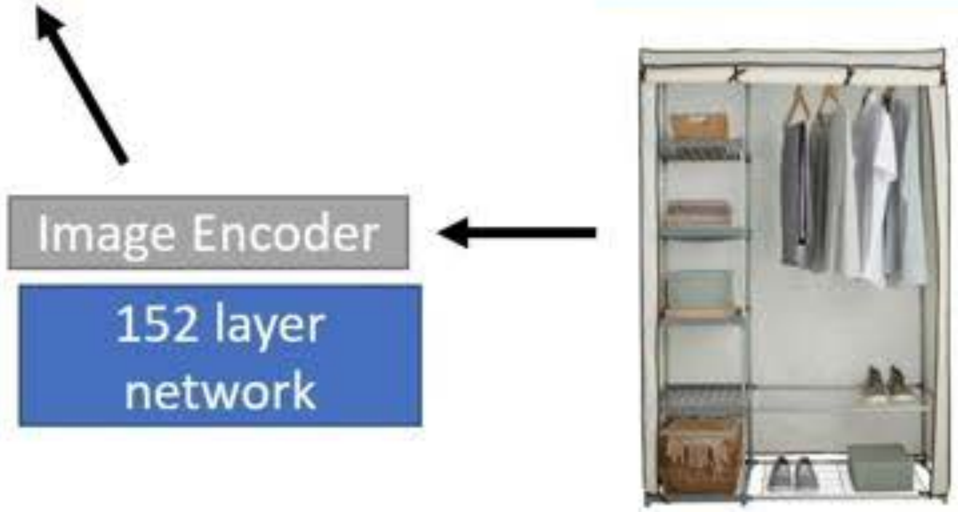
1. Self-Expression on Social Media: Do Tweets Present Accurate and Positive Portraits of Impulsivity, Self-Esteem, and Attachment Style?
2. The Role of Multimedia Content in Determining the Virality of Social Media Information
3. Exploiting multimodal affect and semantics to identify politically persuasive web videos

Unified Multimodal Embeddings (UME)

- Issue: Prior works are limited and generally tackle a proxy task
  - Solution: We propose an unsupervised method to learn the underlying content and its reaction
- Do not explicitly understand content
  - Solution: Learned using multimodal embeddings
- Limited in the number of semantic concepts
  - Solution: Do not restrict to specific concepts and let the model discover them on large-scale data

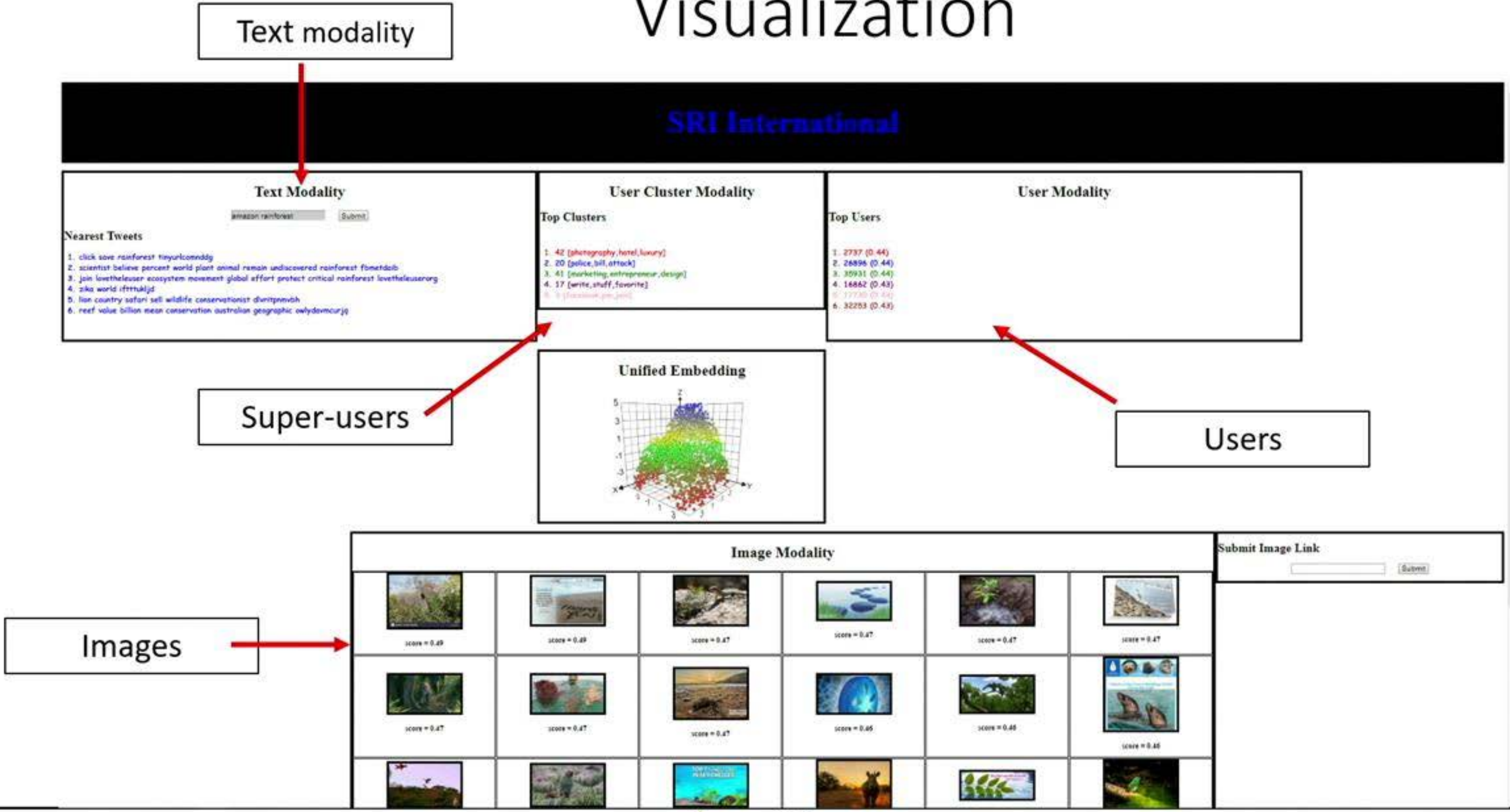
Deep Unified Embedding Model System





- Use ranking loss to jointly embed image, text and users in the same vectorial space
- Learn on a Twitter corpus of ~10M tweets, ~40K users and ~1M images

Visualization



Intrinsic Metrics

legendary singer yvonne chaka chaka activist ilwad elwan receive bet global power award

- Training Data
  - Twitter data with ~10M tweets and ~1M images
  - Cleaning done to remove duplicates from both tweets and images
  - Min tweets = 1 and max tweets = 1K.
  - Distribution is skewed
- Show results for retrieval on a held-out version of the dataset (10K text tweets and 1K image-text pairs)
- Report median-rank of the correct retrieval for each modality

$$\boxed{L = \lambda_1 L_{T-U} + \lambda_2 L_{I-T} + \lambda_3 L_{I-U}}$$

- Our joint loss function can be used to set the relative importance of each paired modality
- Lambda's (regularization parameters) allow us to control the contribution of each modality to the final embedding

Results with Individual Modality

Method	Median Rank						
	T -> U	I -> T	I -> U	λ_1	λ_2	λ_3	Comments
Random	20313	500	20313				
Only T-U	388	-	-	1	0	0	Individual
Only I-T	-	32	-	0	1	0	Individual
Only I-U	-	-	918	0	0	1	Individual
I-T + T-U	409	35	1972	1	1	0	
T-U + I-U	607	46	415	1	0	1	
I-T + I-U	4560	40	1133	0	1	1	
Variations	410	30	632	1	1	1	All same
Variations	470	33	518	1	1	5	Variations
Variations	609	29	596	1	5	5	Variations

- Generally results with individual modality are the best since the corresponding metric e.g. I->T is being optimized directly

Results with 2 pairs of modality

Method	Median Rank						
	T -> U	I -> T	I -> U	λ_1	λ_2	λ_3	Comments
Random	20313	500	20313				
Only T-U	388	-	-	1	0	0	Individual
Only I-T	-	32	-	0	1	0	Individual
Only I-U	-	-	918	0	0	1	Individual
I-T + T-U	409	35	1972	1	1	0	
T-U + I-U	607	46	415	1	0	1	
I-T + I-U	4560	40	1133	0	1	1	
Variations	410	30	632	1	1	1	All same
Variations	470	33	518	1	1	5	Variations
Variations	609	29	596	1	5	5	Variations

- We are able to reason about the modality pair that **we did not see** during training.
- Moreover, when training on T-U and I-U, the model is able to learn I-T automatically with good performance.
- Result highlight the benefits of multimodal space which is able to fill in the gaps for unseen modality pairs

Results with Other Variations

	T -> U	I -> T	I -> U	λ_1	λ_2	λ_3	Comments
Random	20313	500	20313				
Only T-U	388	-	-	1	0	0	Individual
Only I-T	-	32	-	0	1	0	Individual
Only I-U	-	-	918	0	0	1	Individual
I-T + T-U	409	35	1972	1	1	0	
T-U + I-U	607	46	415	1	0	1	
I-T + I-U	4560	40	1133	0	1	1	
Variations	410	30	632	1	1	1	All same
Variations	470	33	518	1	1	5	Variations
Variations	609	29	596	1	5	5	Variations

- The results while training all modalities together are quite strong. For I->U the results are even better than individual training.
- We observe advantages while using additional modality (corroboration)

Extrinsic Metrics

- Compare the learned word embeddings with joint training with standard embeddings on different word embedding benchmarks
- Establish the general quality of embeddings learned with joint training
- Green colored columns highlight the best algorithm
  - Glove: Pre-trained on very large corpus
  - Gensim: word2vec model learned on tweets
  - GRU_init_genism: Learn using our joint model

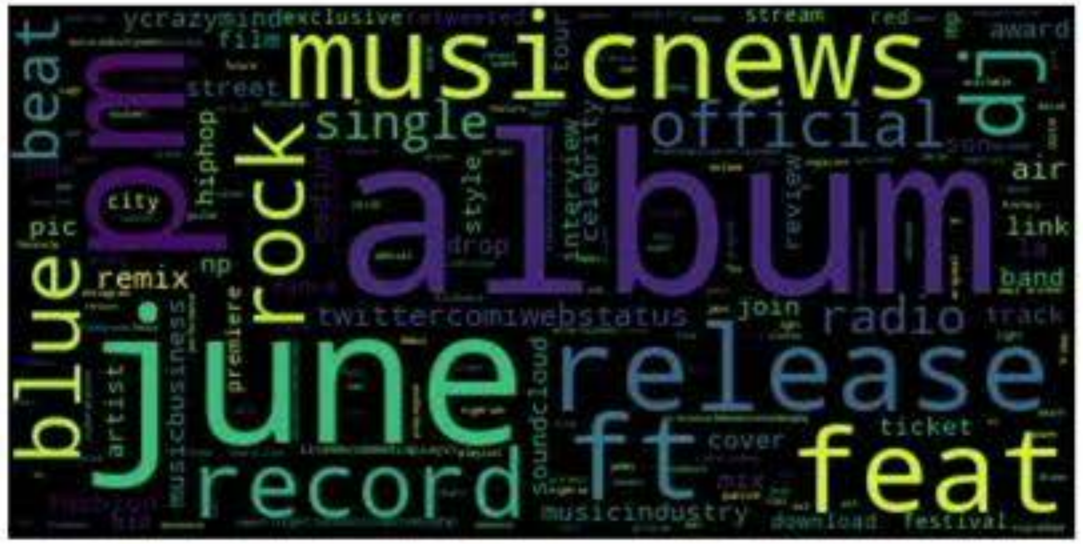
Task type		Glove GloVe	W2v on Twitter Gensim_model	Ours GRU_init_gensim 33-33-33
Cluster Purity	AP	0.644278607	0.5646766169	0.5597014925
	BLESS	0.78	0.785	0.785
	Battig	0.4203785127	0.3829095775	0.3840565857
	ESSLI_1a	0.75	0.75	0.8181818182
	ESSLI_2b	0.775	0.75	0.75
	ESSLI_2c	0.5777777778	0.5555555556	0.6
Spearman Correlation	MEN	0.6809145945	0.7726109311	0.7809875175
	MTurk	0.6193399527	0.5663841608	0.5534003108
	RG65	0.6761810088	0.6437696046	0.6521838936
	RW	0.3132708338	0.2098743607	0.19988206
	SimLex999	0.2975524357	0.3484608345	0.3486203273
	TR9856	0.0918082149	0.1139838909	0.1139256871
	WS353	0.4770989097	0.5380201177	0.5358684389
	WS353R	0.4150385259	0.4609326786	0.457403311
	WS353S	0.6037195454	0.6621465858	0.6610199523
Analogy Prediction	Google	0.6310888252	0.2738436349	0.2649918133
	MSR	0.55	0.087375	0.086375
	SemEval2012_2	0.1653455121	0.138981106	0.1500476972

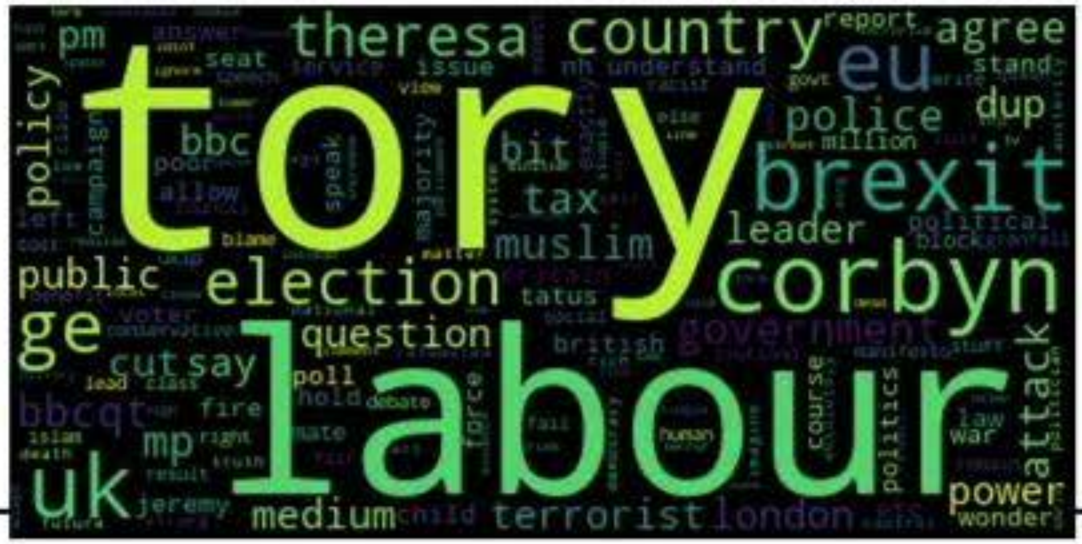
Results

- GloVe vector expected to generally outperforms other models due to training on large amounts of data (billions of documents)
- Our model performs better than GloVe and Gensim on certain tasks e.g. ESSLI_1a: Clustering nouns into semantic categories
- Joint model performs best when text network is GRU rather than convolutional
- The joint model is trained on extremely noisy data with a vocabulary that may not match the types of tasks the evaluation library performs
  - Currently evaluation on twitter sentiment prediction task

Cluster Visualizations

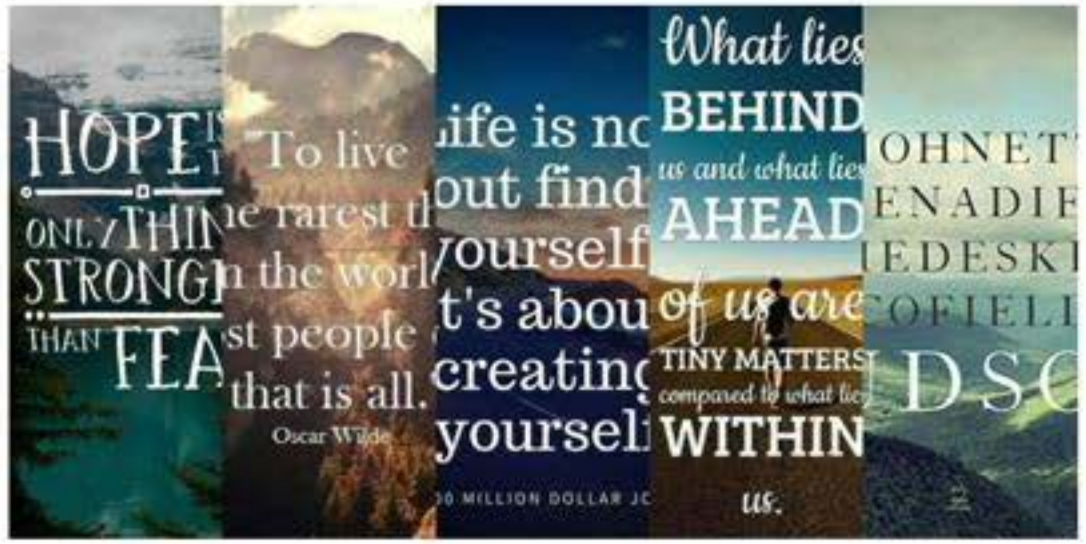
- We extract super-users by clustering the user embeddings into 50 clusters. We visualize wordclouds for some clusters here.

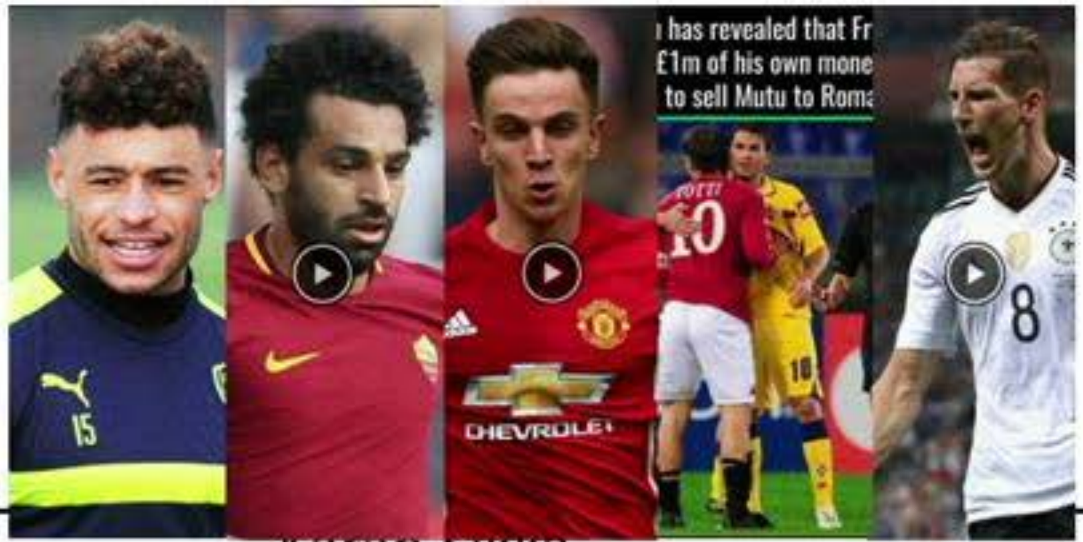




Cluster Visualizations - Images

- Representative images of the clusters for each wordcloud are shown below.



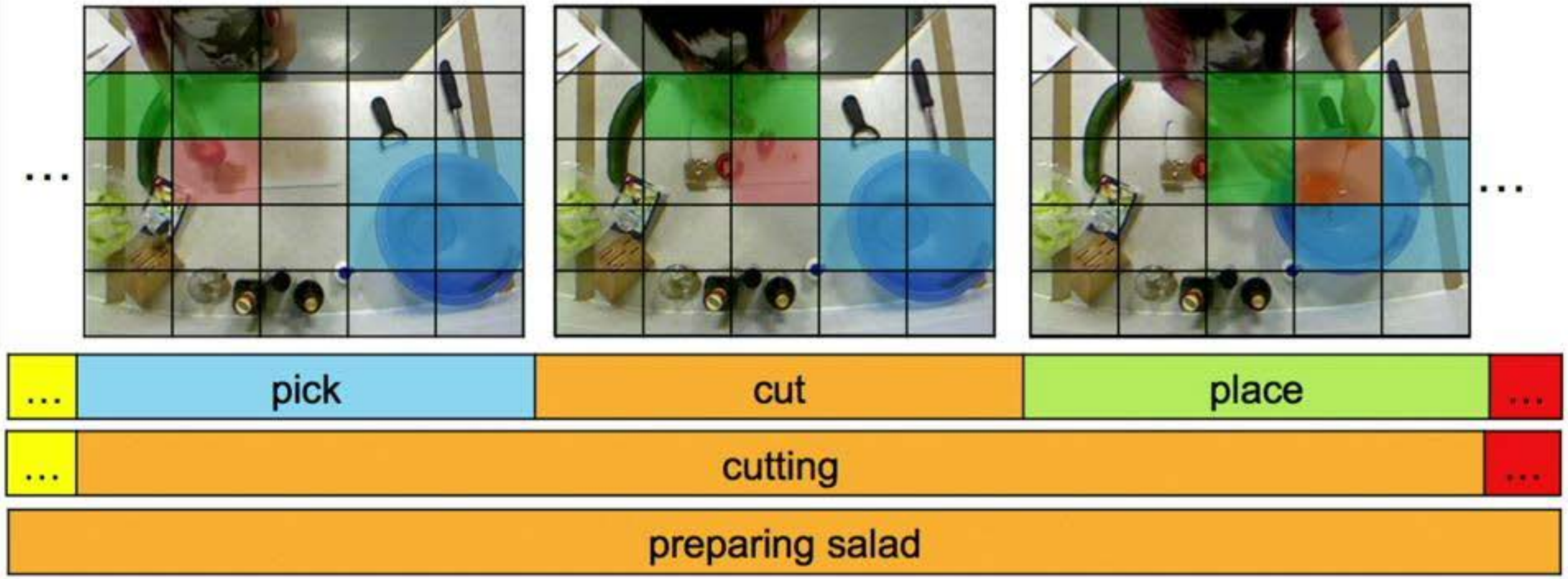


Discussion

- Compositionality
    - Composition is very important to disambiguate between different word senses. Provides context.
    - Our model effectively learns generic semantic concepts and few specific concepts without any additional supervision and with noisy data
- Currently doing more experiments and working on writing initial publications
    - Very important to factorize improvements and effects of different modalities esp. with deep learning
    - Glimpse of demo


Multimodal
Embedding Demo
SRI International
DARPA SocialSim
Top Tags
Car
Flag
Riot
Fight
Top Tags
Fight
Demonstration
Fire
Gathering

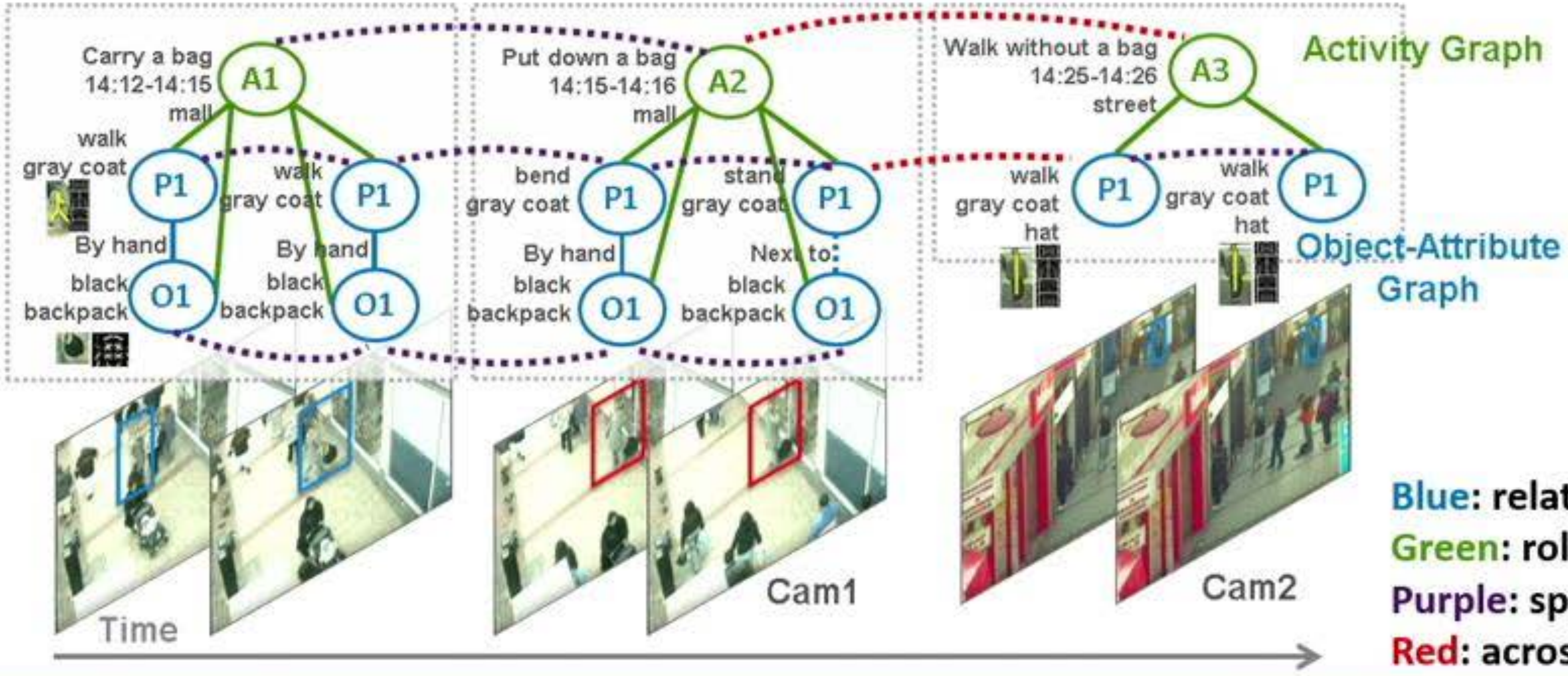
Task: Action Segmentation



SRI International

Representation: Activity-Object-Attribute Graph

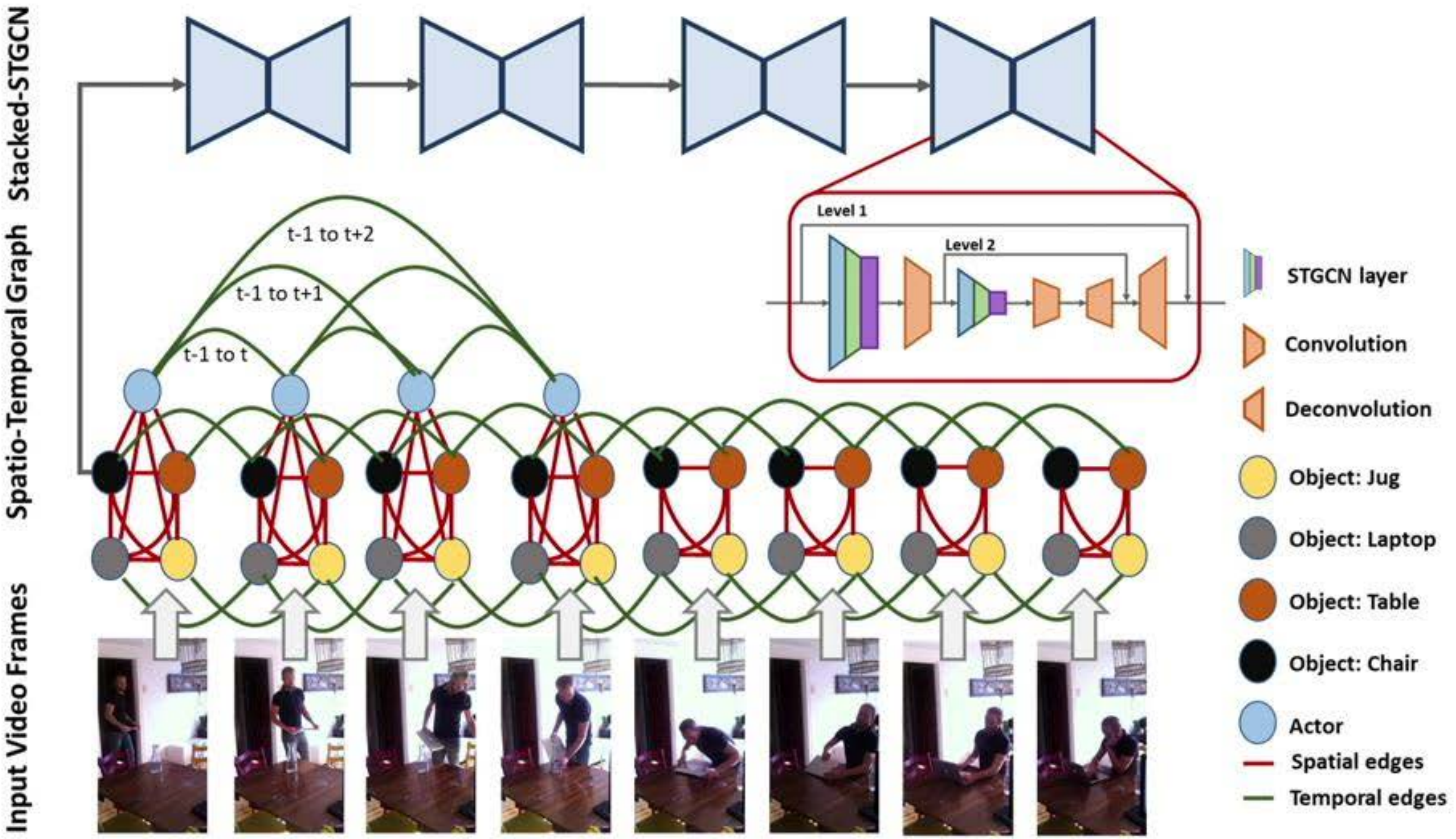
- A model connecting activity to its components over space and time
- **Activity Graph** tracks multiple threads of activities
- **Object-Attribute Graph** captures the state and state change of involved entities
- Resolve relationships among activities and objects and infer explicit observables and implicit consequentials



Blue: relationships
Green: roles
Purple: spatiotemporal
Red: across cameras

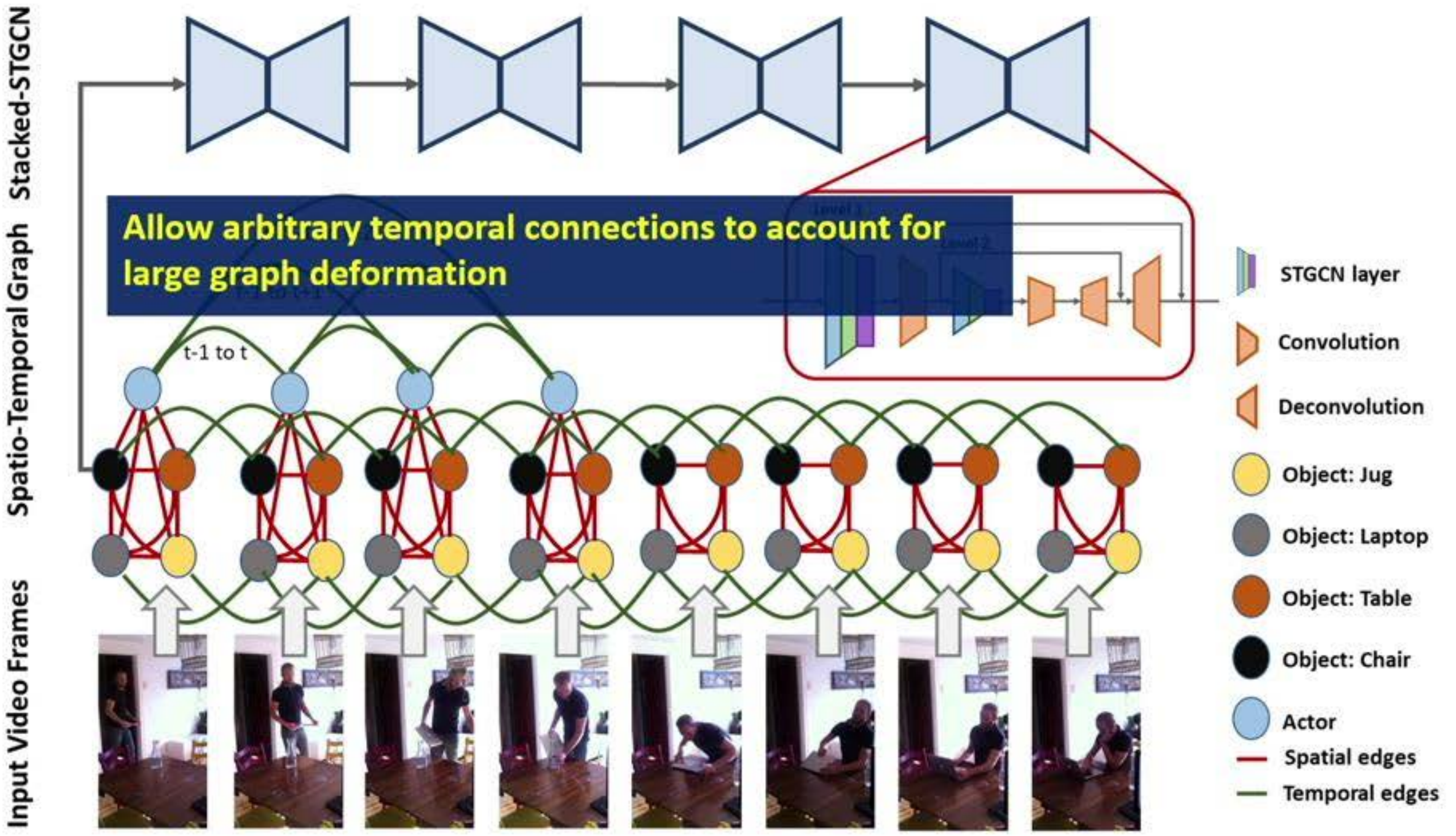
SRI International

Analytic: Graph Convolutional Network



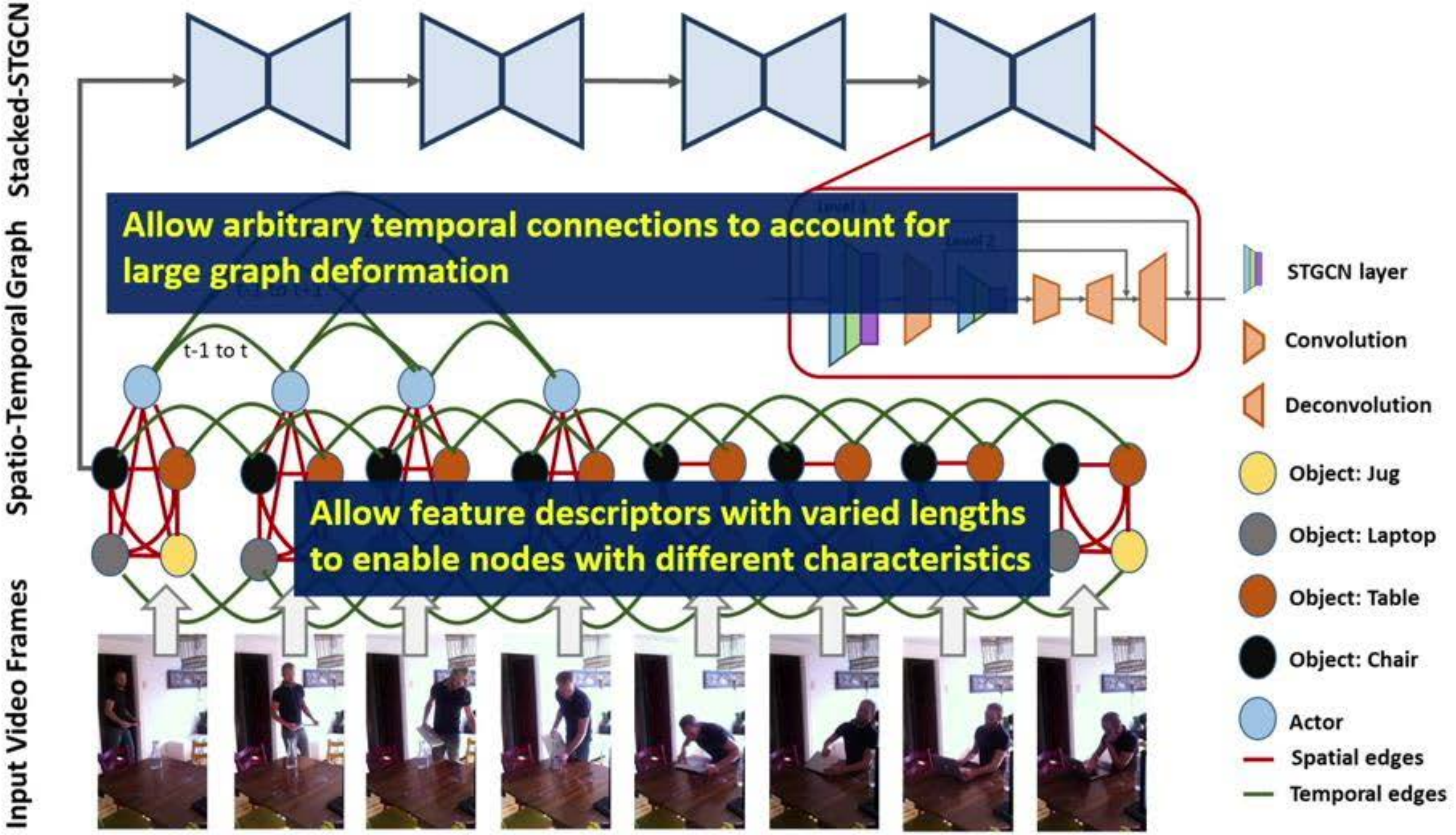
SRI International

Analytic: Graph Convolutional Network

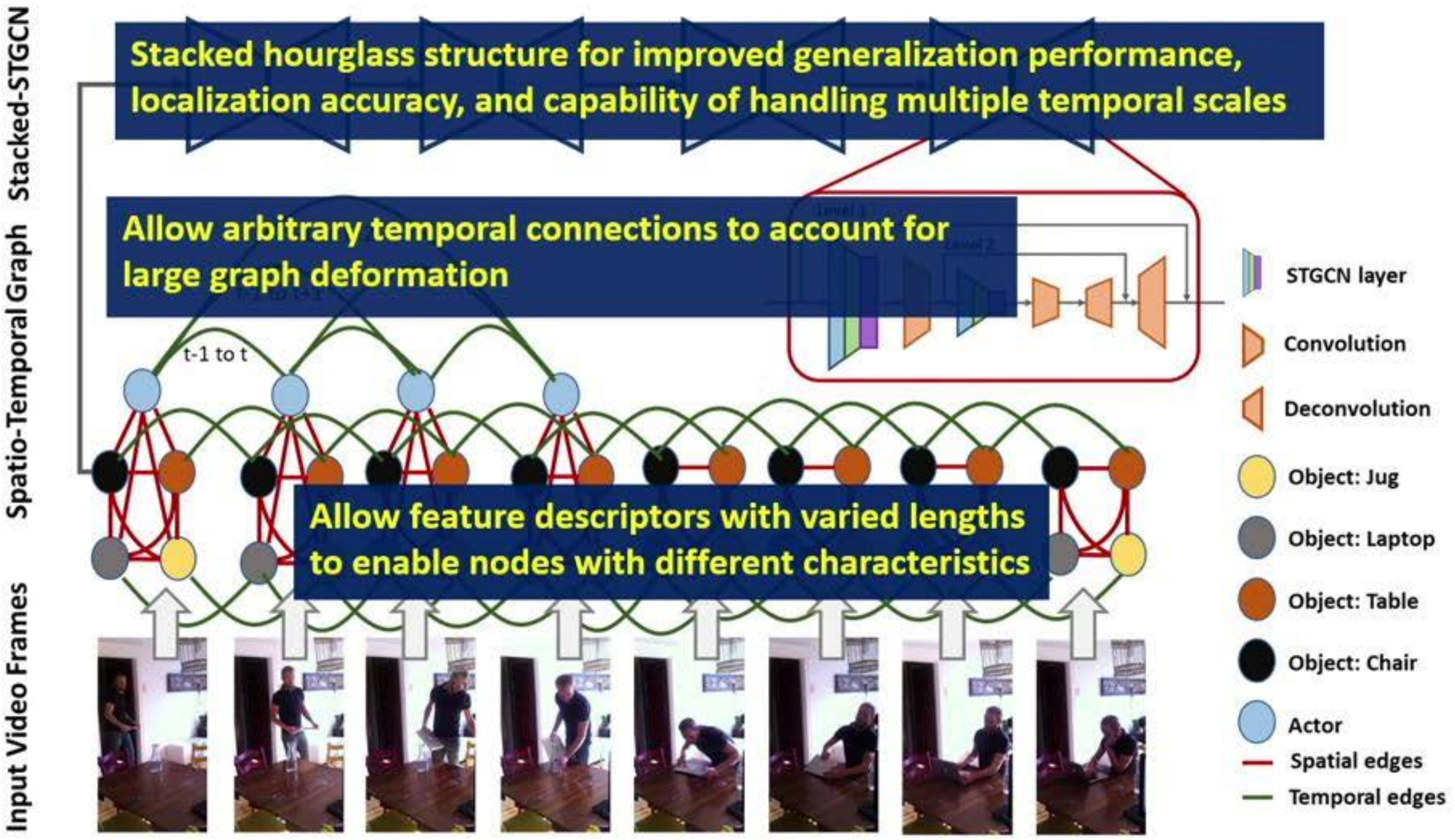


SRI International

Analytic: Graph Convolutional Network

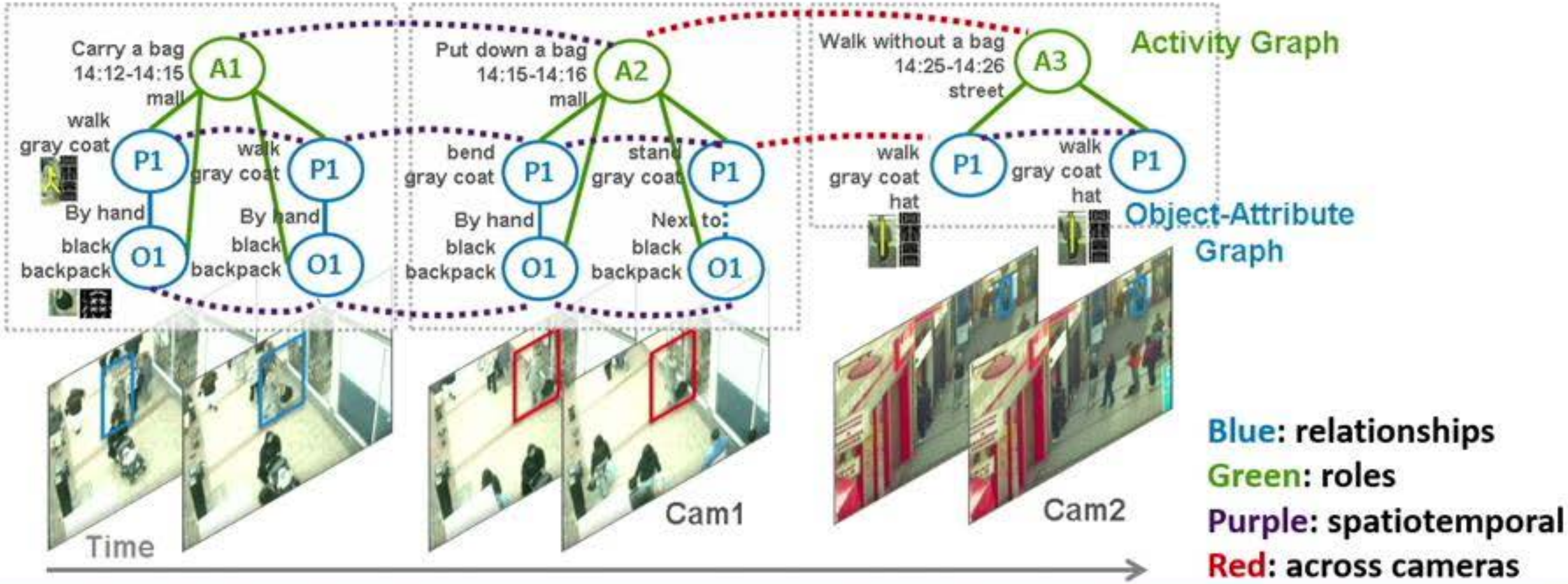


Analytic: Graph Convolutional Network



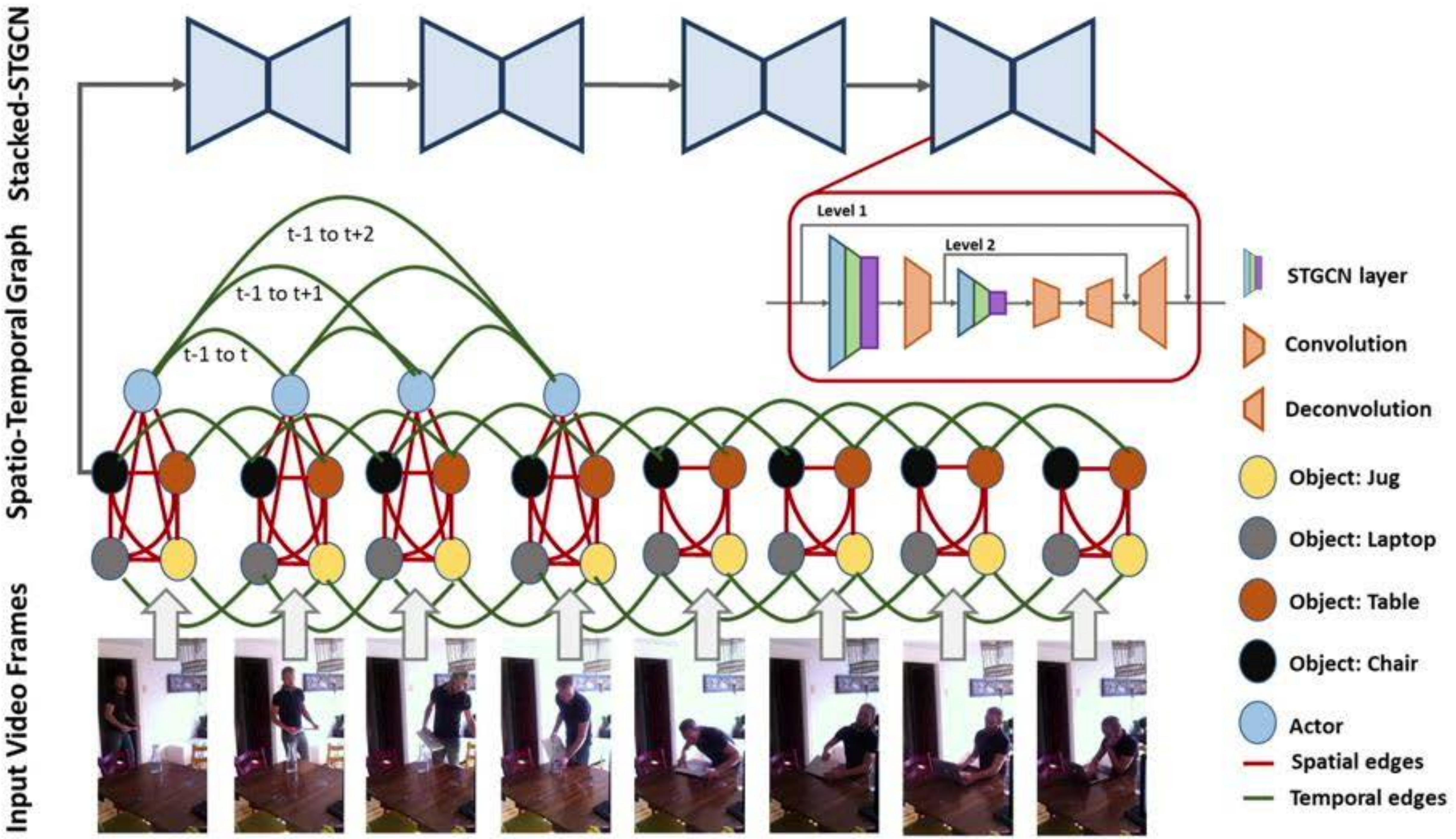
Representation: Activity-Object-Attribute Graph

- A model connecting activity to its components over space and time
- **Activity Graph** tracks multiple threads of activities
- **Object-Attribute Graph** captures the state and state change of involved entities
- Resolve relationships among activities and objects and infer explicit observables and implicit consequentials



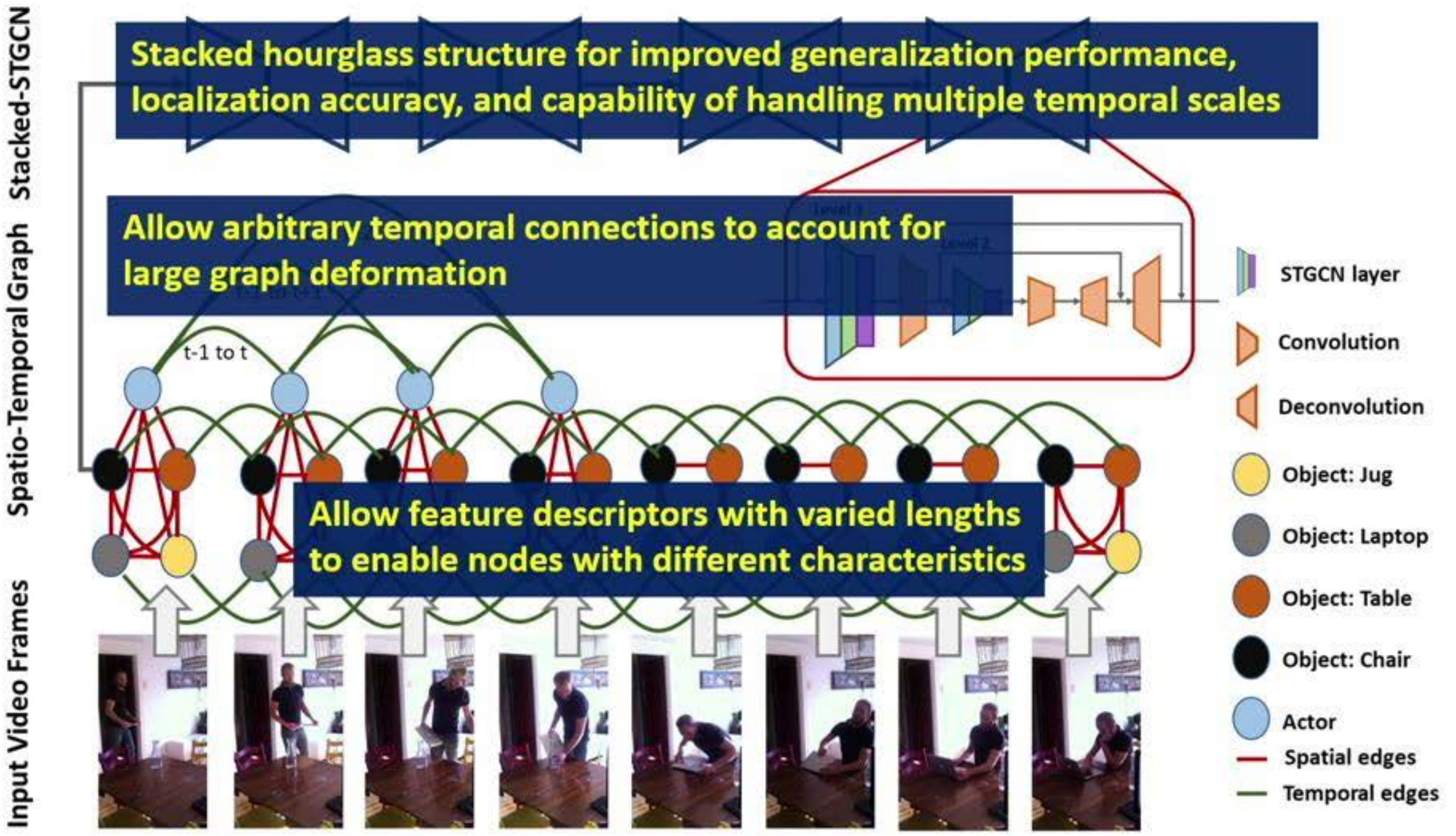
SRI International

Analytic: Graph Convolutional Network

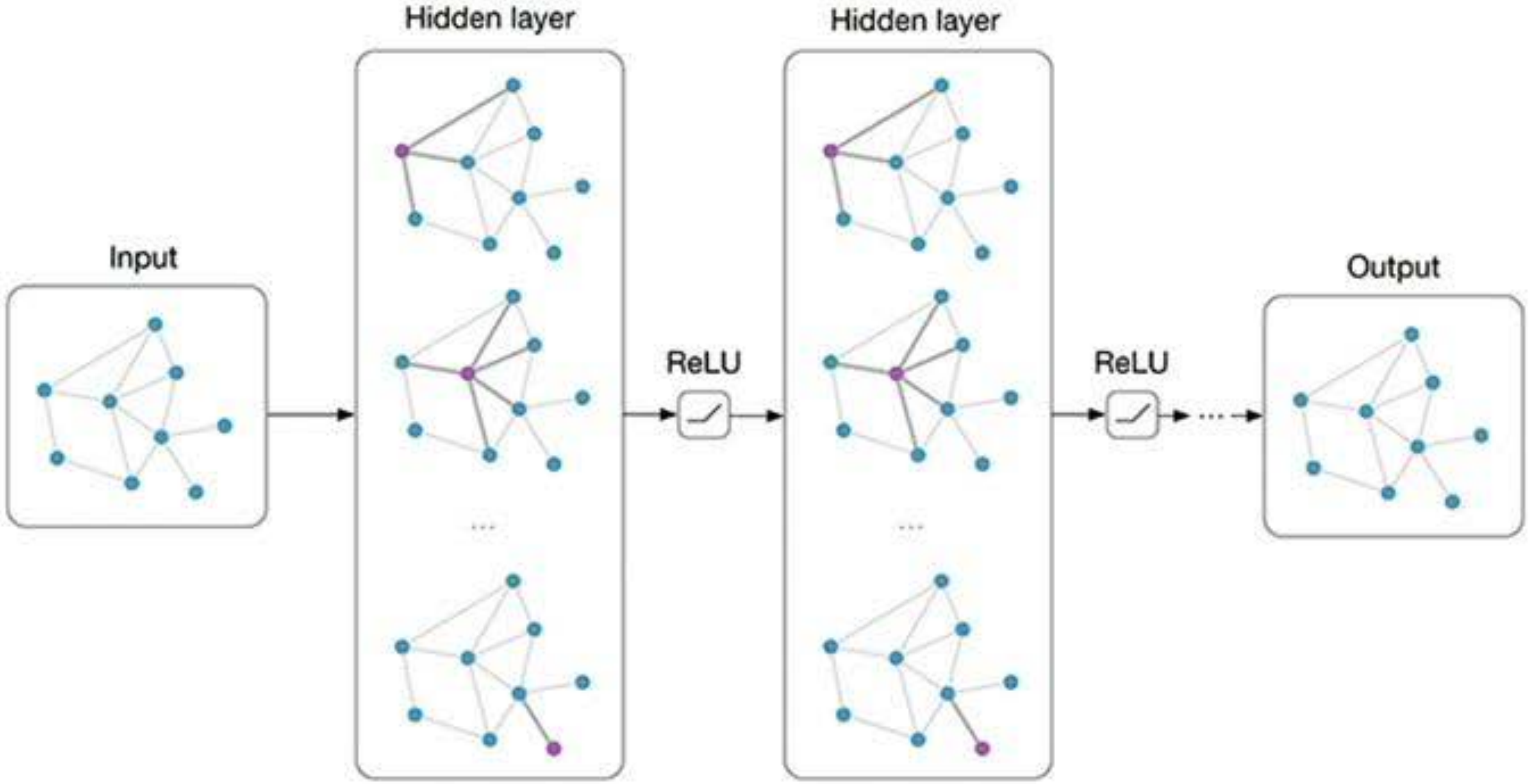


SRI International

Analytic: Graph Convolutional Network



Graph Convolutional Networks



$$f(H^{(l)}, A) = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

$\hat{A} = I + A$, $A = [e_{i,j}]$ is the adjacency matrix, $\widehat{D}$ is the diagonal node degree matrix of $\hat{A}$, $H^{(l)}$: $N \times d^l$ input matrix of the l^{th} layer, $W^{(l)}$: $d^l \times d^{l+1}$ weight matrix of the l^{th} layer, σ: nonlinear activation function

SRI International

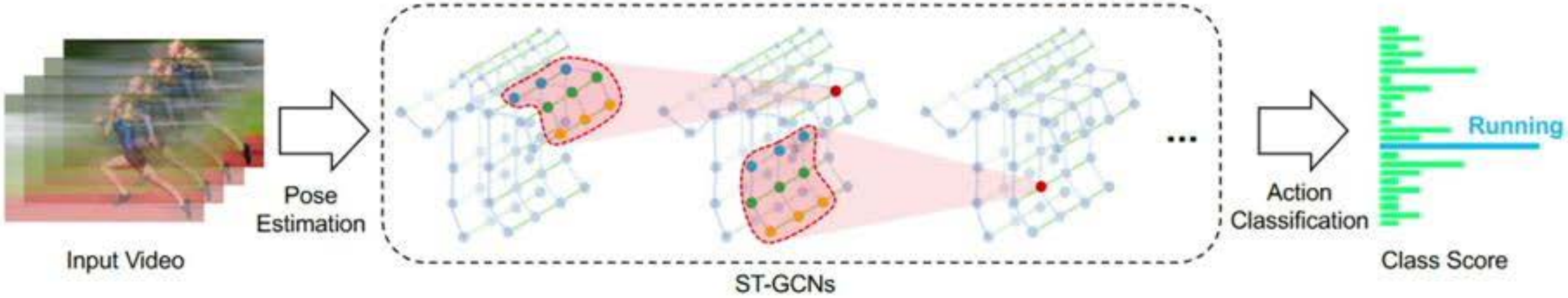
Stacked Spatio-Temporal Graph Convolutional Networks for Action Segmentation

Pallabi Ghosh, Yi Yao, Larry D. Davis, and Ajay Divakaran

2019/02/15

SRI International

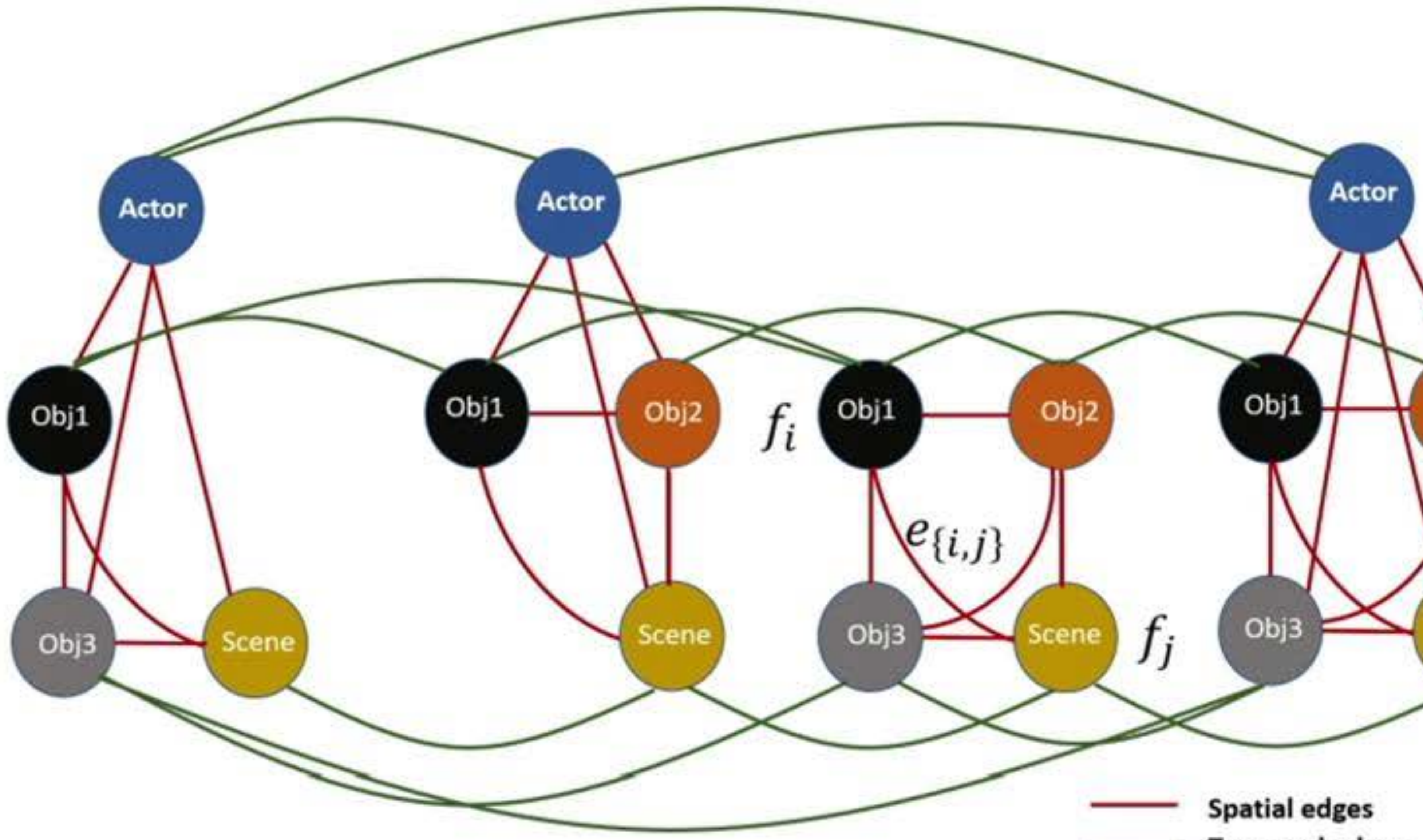
Spatiotemporal Graph Convolutional Network



- Temporal connection: connect the same joint in consecutive frames
  - $H^{l+1} = g\left(H^l, A_s\right) = \sigma(\widehat{D}_s^{-\frac{1}{2}} \hat{A}_s \widehat{D}_s^{-\frac{1}{2}} H^l W_s^l W_t^l)$
- Feature: (X, Y, C) for each joint

S. Yan, Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, AAAI 2018.

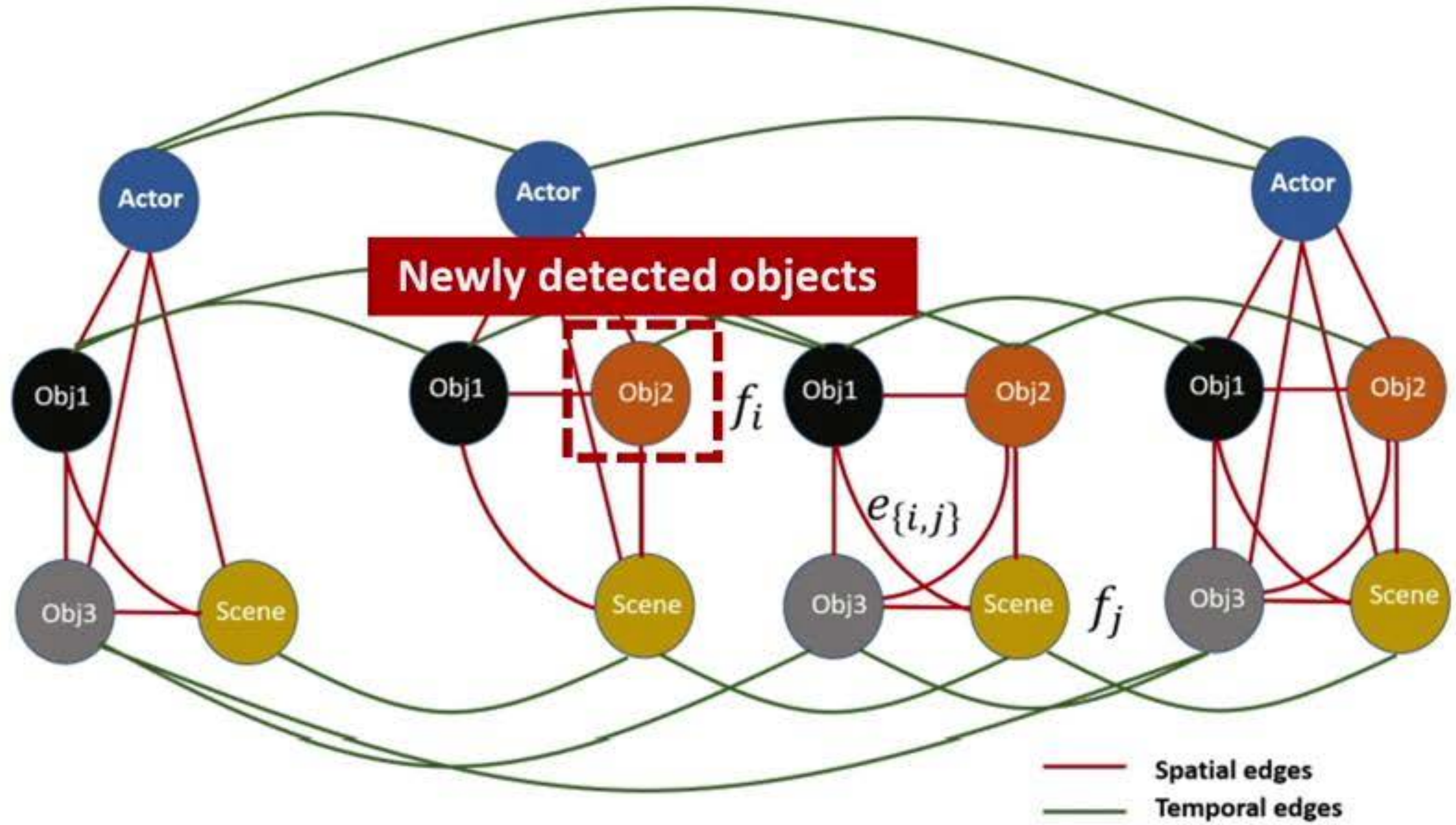
Arbitrary Temporal Connection



Allow arbitrary temporal connections to account for large graph deformation

SRI International

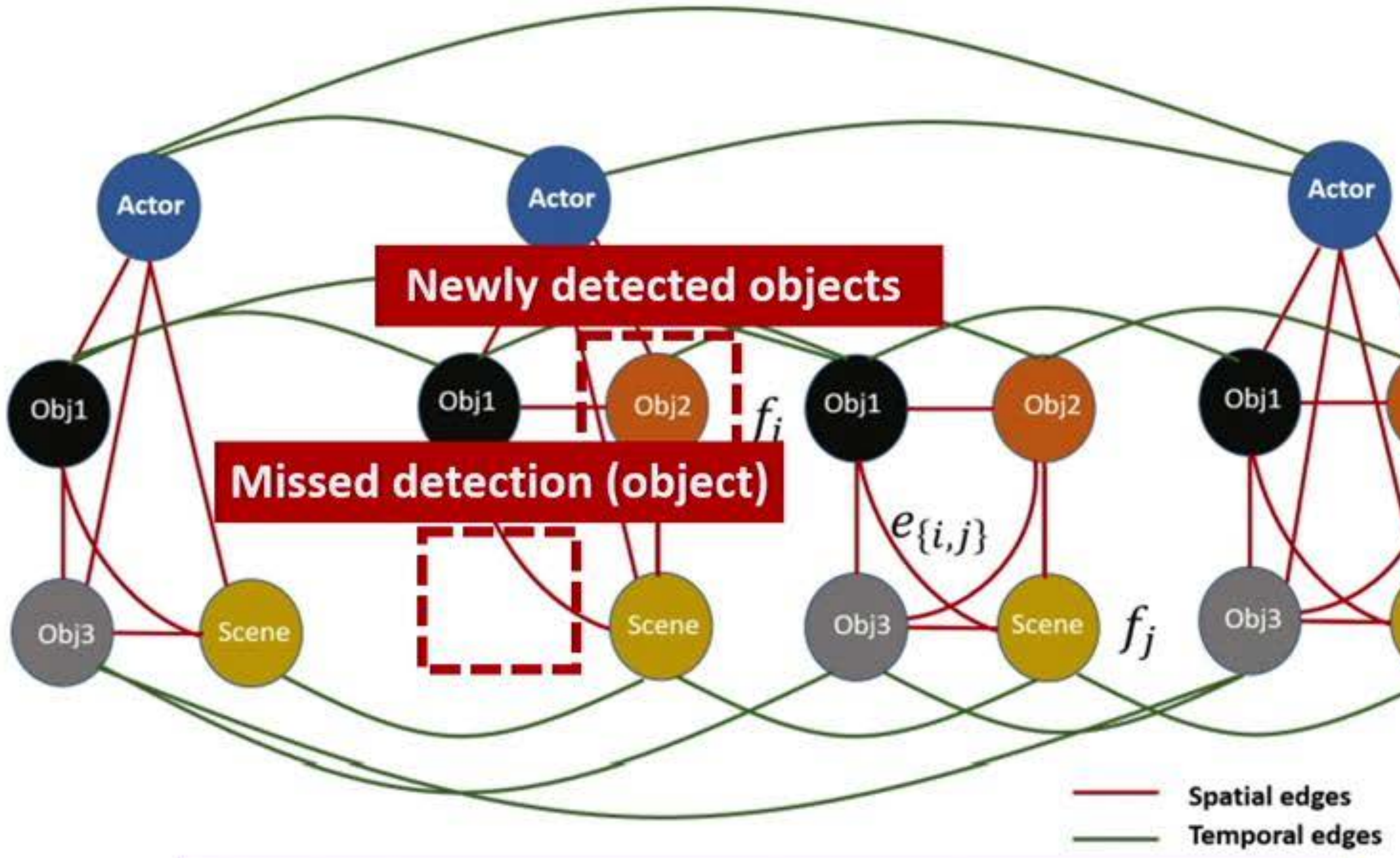
Arbitrary Temporal Connection



Allow arbitrary temporal connections to account for large graph deformation

SRI International

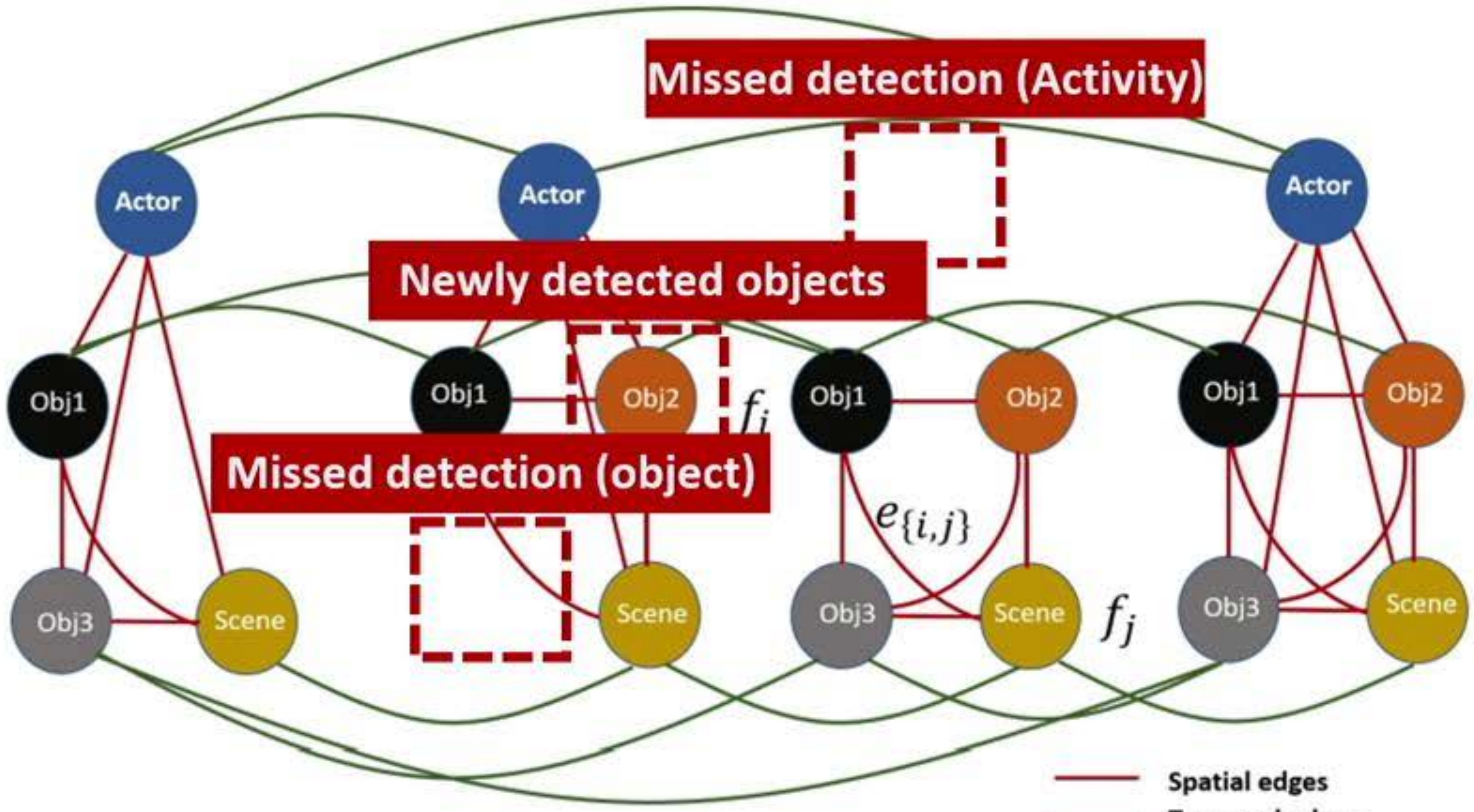
Arbitrary Temporal Connection



Allow arbitrary temporal connections to account for large graph deformation

SRI International

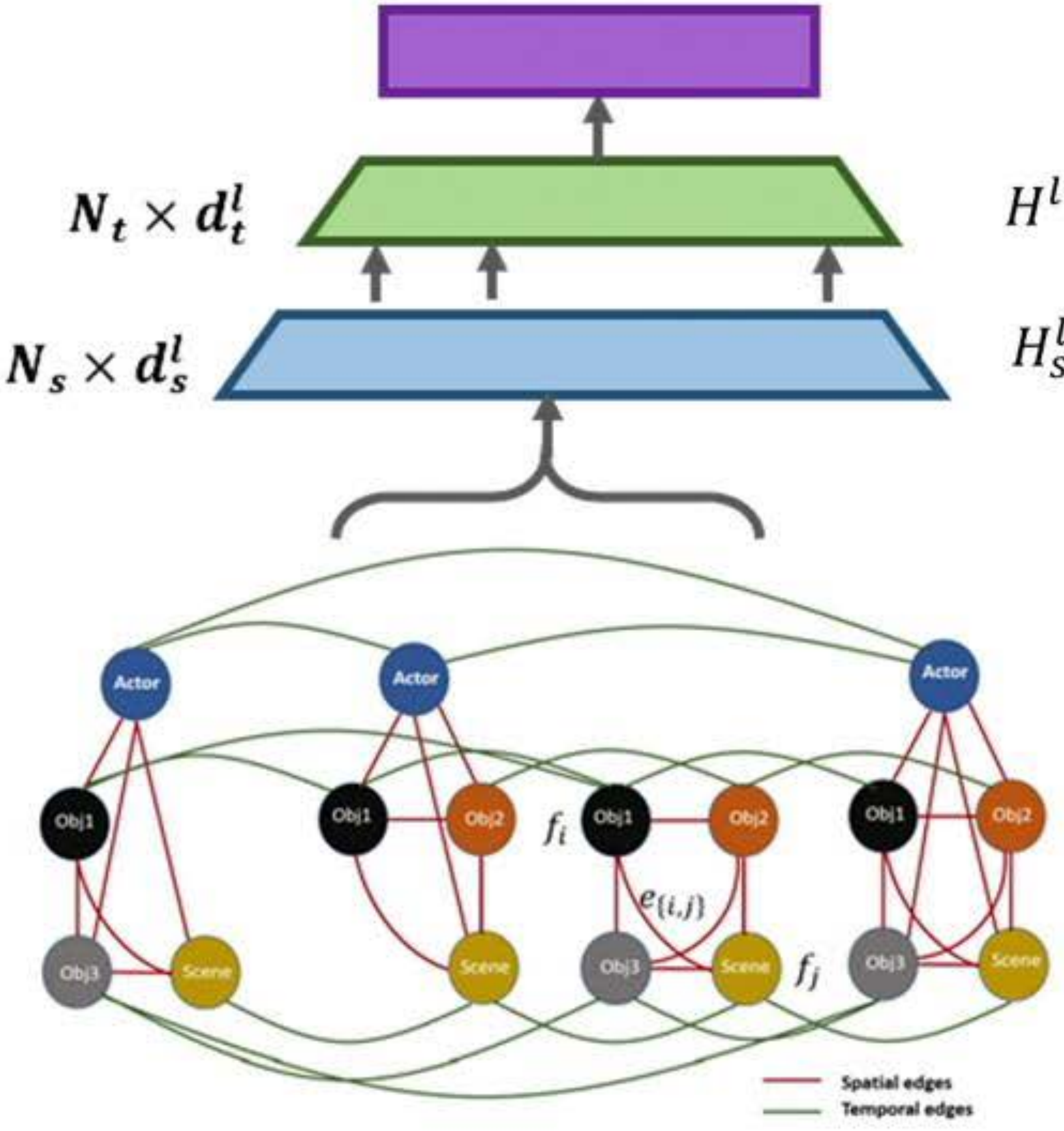
Arbitrary Temporal Connection



Allow arbitrary temporal connections to account for large graph deformation

SRI International

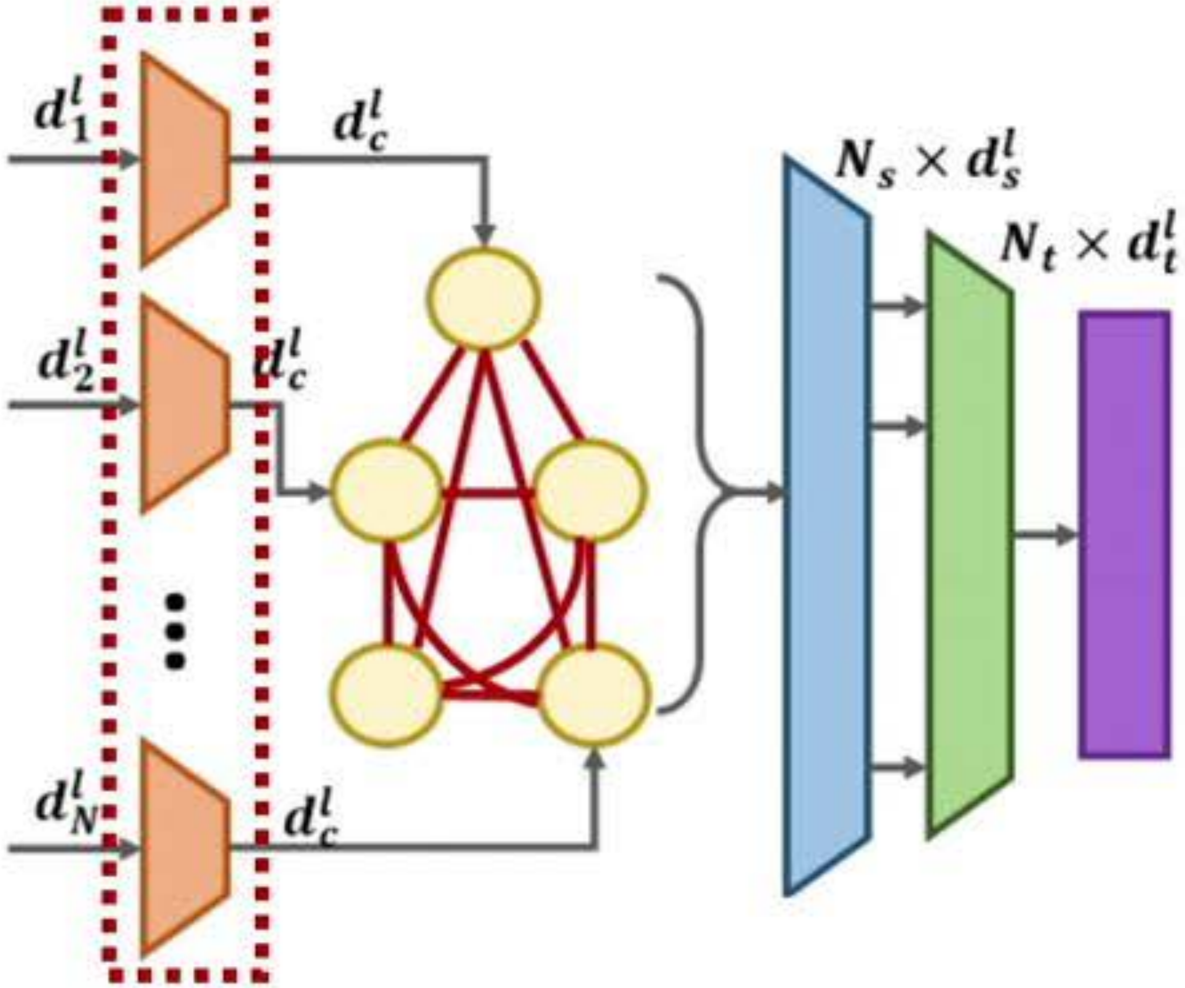
Arbitrary Temporal Connection



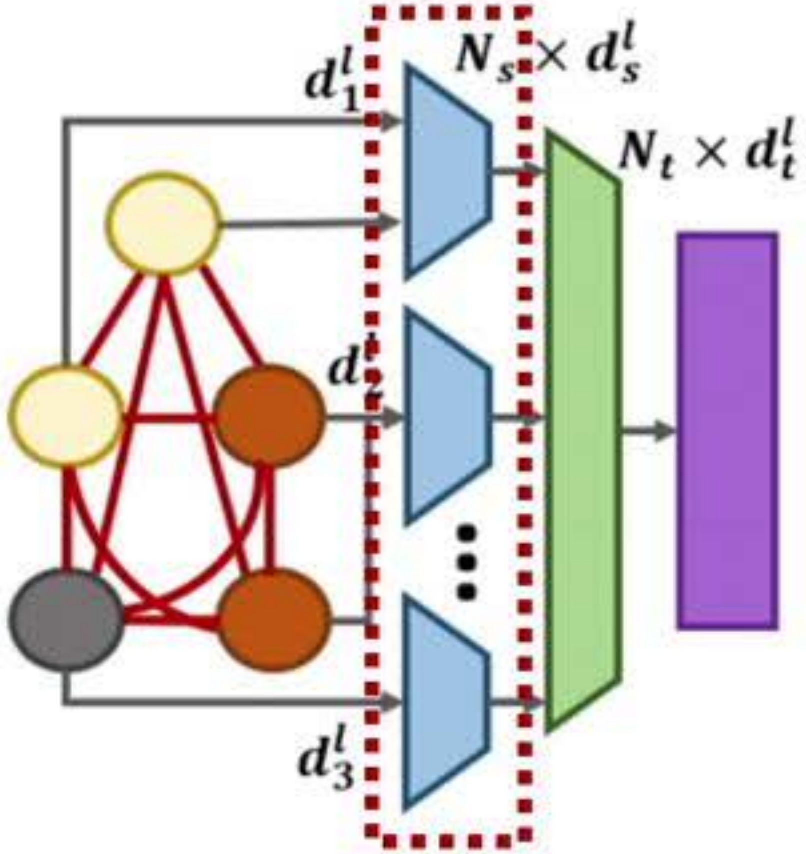
$$H^{l+1} = g_t(H_s^l, A_t) = \sigma(\widehat{D}_t^{-\frac{1}{2}} \hat{A}_t \widehat{D}_t^{-\frac{1}{2}} H_s^l W_t^l)$$

$$H_s^l = g_s(H^l, A_s) = \widehat{D}_s^{-\frac{1}{2}} \hat{A}_s \widehat{D}_s^{-\frac{1}{2}} H^l W_s^l$$

Features with Varied length



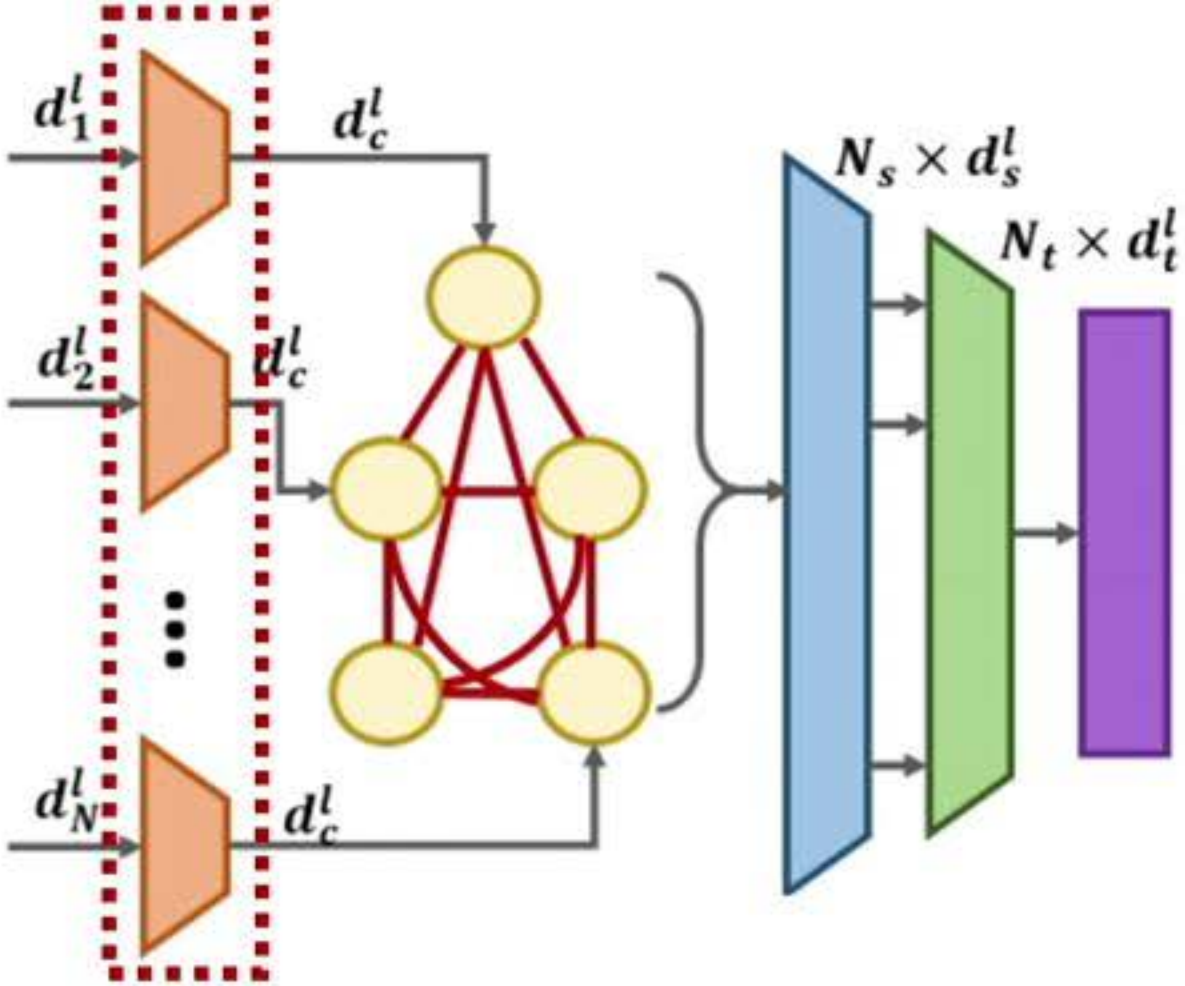
Initial convolution layers convert descriptors with varied length to the same feature space with a fixed length

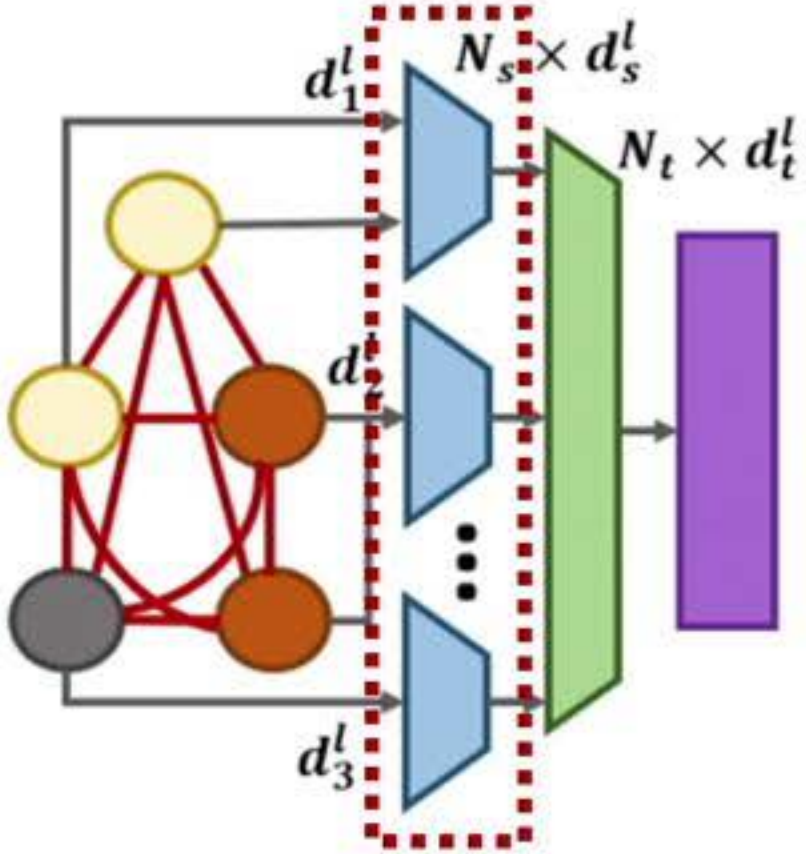


Group descriptors of the same feature; use multiple spatial GCNs for each group; the output of these spatial GCNs have the same dimension

Convolution **Spatial Graph Convolution** **Temporal Graph Convolution** **Non-Linear**

Features with Varied length



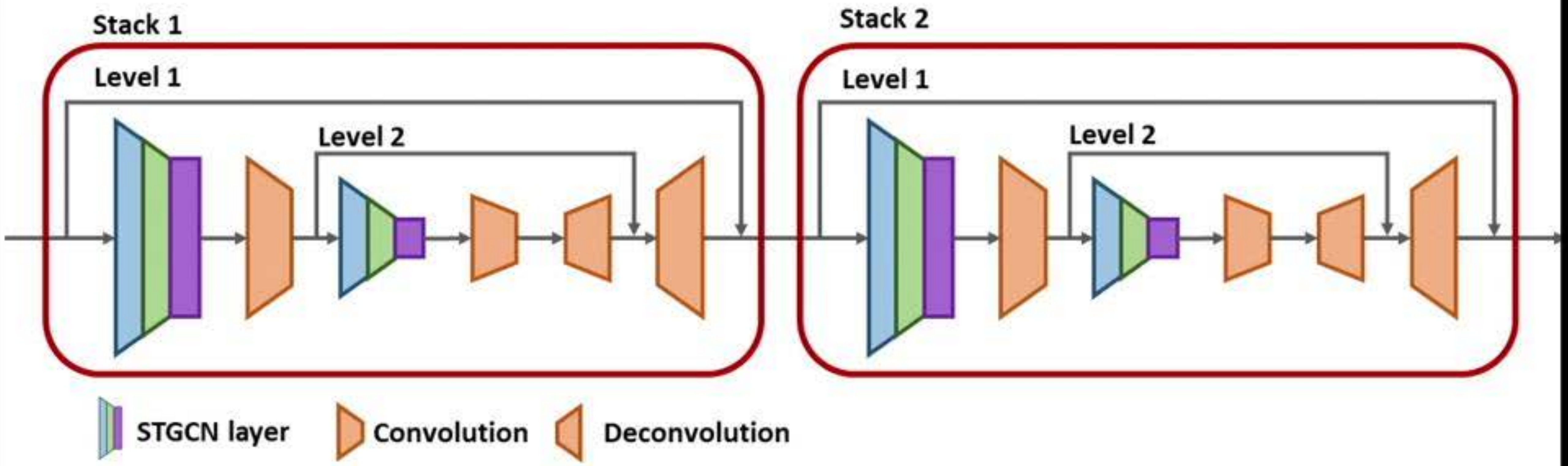


- Pros: Smaller network
- Cons: Possible loss of data

- Pros: Grouping of data reduces data loss
- Cons: More complicated and larger network

SRI International

Hourglass Architecture



- Stacked hourglass structure for improved generalization performance, localization accuracy, and capability of handling multiple temporal scales
- Non-symmetric encoding and decoding since feature pooling on graphs is only required in encoding
- The dimensions of the spatial and temporal adjacency matrices need to be adjusted accordingly

SRI International

Datasets

CAD120

(10 classes, Single-label):

Description	Count
Object Features	**18**
N1. Centroid location	3
N2. 2D bounding box	4
N3. Transformation matrix of SIFT matches between adjacent frames	6
N4. Distance moved by the centroid	1
N5. Displacement of centroid	1
Sub-activity Features	**103**
N6. Location of each joint (8 joints)	24
N7. Distance moved by each joint (8 joints)	8
N8. Displacement of each joint (8 joints)	8
N9. Body pose features	47
N10. Hand position features	16
Object-object Features (computed at start frame, middle frame, end frame, max and min)	**20**
E1. Difference in centroid locations ($\Delta x, \Delta y, \Delta z$)	3
E2. Distance between centroids	1
Object–sub-activity Features (computed at start frame, middle frame, end frame, max and min)	**40**
E3. Distance between each joint location and object centroid	8
Object Temporal Features	**4**
E4. Total and normalized vertical displacement	2
E5. Total and normalized distance between centroids	2
Sub-activity Temporal Features	**16**
E6. Total and normalized distance between each corresponding joint locations (8 joints)	16

Charades

(157 classes, Multi-label):

Features used in the graph nodes:

- Image level VGG features
  - RGB for scene; flow for motion
- Segment level I3D features
- Fast-RCNN for object
- Situation recognition for action

SRI International

Results – CAD120

These results are based on 4 fold cross validation. There are 4 different humans doing each activity and each of them form one of the folds meaning it is the test dataset for that fold. The rest of the 3 humans form the training set.

Method	Sub-Activity Detection F1 Score
Koppula et al	80.4
S-RNN w/o edge RNN	82.4
S-RNN	83.2
STGCN (Ours)	**87.3**

SRI International

Results – Charades

	VGG	I3D
Baseline	6.56	17.22
LSTM	7.85	18.12
Super-Event	8.53	**19.41**
Stacked-STGCN	**10.94**	18.51

Method	mAP
Random	2.42
RGB	7.89
Predictive-corrective	8.9
Two-stream	8.94
Two-stream +LSTM	9.6
R-C3D	12.7
Sigurdsson et. al.	12.8
I3D	17.22
I3D + LSTM	18.1
I3D + temporal pyramid	18.2
I3D + super-events	**19.41**
VGG + Stacked-STGCN (ours)	**10.94**
VGG + Stacked-STGCN all (ours)	**11.73**
I3D + Stacked-STGCN (ours)	**19.09**

SRI International

Ablation Study

- Baseline (no GCN)
  - Features are passed through a single Fully Connected layer outputting class probabilities
  - The final decision is based on the average of these probabilities.
  - **Improvement: 4.06 in mAP**

Experiments	mAP
All features; Baseline	7.67
All features; STGCN	9.22
VGG-RGB; STGCN; 1 time step	6.33
VGG-RGB; STGCN	6.54
All features; Stacked-STGCN; 1 time step	10.93
VGG-RGB; Stacked-STGCN	7.91
VGG-RGC+VGG-flow; Stacked-STGCN	10.94
All Features; Stacked-STGCN	**11.73**

SRI International

Ablation Study

- Baseline (no GCN)
  - Features are passed through a single Fully Connected layer outputting class probabilities
  - The final decision is based on the average of these probabilities.
  - **Improvement: 4.06 in mAP**
- Hourglass structure
  - A GCN with the same number of convolutional layers as the encoder of Stacked STGCN.
  - **Improvement: 2.51 in mAP**

Experiments	mAP
All features; Baseline	7.67
All features; STGCN	9.22
VGG-RGB; STGCN; 1 time step	6.33
VGG-RGB; STGCN	6.54
All features; Stacked-STGCN; 1 time step	10.93
VGG-RGB; Stacked-STGCN	7.91
VGG-RGC+VGG-flow; Stacked-STGCN	10.94
All Features; Stacked-STGCN	**11.73**

SRI International

Ablation Study

- Baseline (no GCN)
  - Features are passed through a single Fully Connected layer outputting class probabilities
  - The final decision is based on the average of these probabilities.
  - **Improvement: 4.06 in mAP**
- Hourglass structure
  - A GCN with the same number of convolutional layers as the encoder of Stacked STGCN.
  - **Improvement: 2.51 in mAP**
- Vanilla STGCN
  - Temporal connections across one time step
  - Nodes with the same type of features (VGG-RGB)
  - Pure graph convolutional operations (without hourglass)
  - **Improvements: 5.40 in mAP**

Experiments	mAP
All features; Baseline	7.67
All features; STGCN	9.22
VGG-RGB; STGCN; 1 time step	6.33
VGG-RGB; STGCN	6.54
All features; Stacked-STGCN; 1 time step	10.93
VGG-RGB; Stacked-STGCN	7.91
VGG-RGC+VGG-flow; Stacked-STGCN	10.94
All Features; Stacked-STGCN	**11.73**

SRI International

Ablation Study

- Input features
  - VGG-RGB: 7.91
  - VGG-RGB+VGG-flow: 10.94
  - All features: 11.73

Experiments	mAP
All features; Baseline	7.67
All features; STGCN	9.22
VGG-RGB; STGCN; 1 time step	6.33
VGG-RGB; STGCN	6.54
All features; Stacked-STGCN; 1 time step	10.93
VGG-RGB; Stacked-STGCN	7.91
VGG-RGC+VGG-flow; Stacked-STGCN	10.94
All Features; Stacked-STGCN	**11.73**

Ablation Study

- Input features
  - VGG-RGB: 7.91
  - VGG-RGB+VGG-flow: 10.94
  - All features: 11.73
- Temporal connections
  - All features; Stacked-STGCN
  - **Improvement: 0.80 in mAP**

Experiments	mAP
All features; Baseline	7.67
All features; STGCN	9.22
VGG-RGB; STGCN; 1 time step	6.33
VGG-RGB; STGCN	6.54
All features; Stacked-STGCN; 1 time step	10.93
VGG-RGB; Stacked-STGCN	7.91
VGG-RGC+VGG-flow; Stacked-STGCN	10.94
All Features; Stacked-STGCN	**11.73**

Ablation Study

- Input features
  - VGG-RGB: 7.91
  - VGG-RGB+VGG-flow: 10.94
  - All features: 11.73

- Temporal connections
  - All features; Stacked-STGCN
  - **Improvement: 0.80 in mAP**

  - VGG-RGB; STGCN
  - **Improvement: 0.21 in mAP**

  - Improvements depend on network architecture and application

Experiments	mAP
All features; Baseline	7.67
All features; STGCN	9.22
VGG-RGB; STGCN; 1 time step	6.33
VGG-RGB; STGCN	6.54
All features; Stacked-STGCN; 1 time step	10.93
VGG-RGB; Stacked-STGCN	7.91
VGG-RGC+VGG-flow; Stacked-STGCN	10.94
All Features; Stacked-STGCN	**11.73**

SRI International

Examples – CAD120

Example 1

GT Label	null	reaching	opening	reaching	moving	placing	reaching	closing
Prediction	null	reaching	opening	reaching	moving	placing	reaching	closing

Example 2

GT Label	null	reaching	moving	reaching	opening	reaching	moving	scrubbing	moving	placing	reaching	closing
Prediction	null	reaching	opening	reaching	opening	reaching	moving	scrubbing	reaching	placing	reaching	closing

Example 3

GT Label	reaching	opening	opening	reaching	moving	eating	reaching	reaching	moving	drinking	moving	placing
Prediction	null	opening	opening	reaching	moving	reaching	moving	reaching	moving	drinking	moving	placing

Ablation Study

- Input features
  - VGG-RGB: 7.91
  - VGG-RGB+VGG-flow: 10.94
  - All features: 11.73

- Temporal connections
  - All features; Stacked-STGCN
  - **Improvement: 0.80 in mAP**

  - VGG-RGB; STGCN
  - **Improvement: 0.21 in mAP**

  - Improvements depend on network architecture and application

Experiments	mAP
All features; Baseline	7.67
All features; STGCN	9.22
VGG-RGB; STGCN; 1 time step	6.33
VGG-RGB; STGCN	6.54
All features; Stacked-STGCN; 1 time step	10.93
VGG-RGB; Stacked-STGCN	7.91
VGG-RGC+VGG-flow; Stacked-STGCN	10.94
All Features; Stacked-STGCN	**11.73**

SRI International

Results – Charades

	VGG	I3D
Baseline	6.56	17.22
LSTM	7.85	18.12
Super-Event	8.53	**19.41**
Stacked-STGCN	**10.94**	18.51

Method	mAP
Random	2.42
RGB	7.89
Predictive-corrective	8.9
Two-stream	8.94
Two-stream +LSTM	9.6
R-C3D	12.7
Sigurdsson et. al.	12.8
I3D	17.22
I3D + LSTM	18.1
I3D + temporal pyramid	18.2
I3D + super-events	**19.41**
VGG + Stacked-STGCN (ours)	**10.94**
VGG + Stacked-STGCN all (ours)	**11.73**
I3D + Stacked-STGCN (ours)	**19.09**

SRI International