

Movie Meets AI



Qingqiu Huang

26/03/2019

CONTENTS

1. Introduction
2. Tag-based Understand
3. Story-based Understand
4. Conclusion



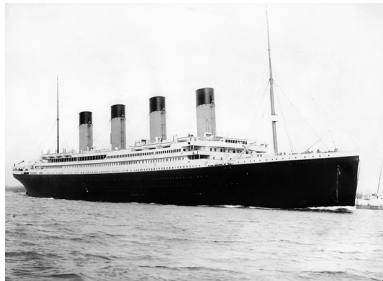


/01 Introduction

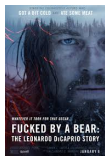


Why movies?

1. To understand movies is to understand our world




2. Cross-modal & rich resources





A Large Scale Dataset









Titanic

11/18/1997 ★ 7.5

Drama, Romance, Thriller

In 1996, treasure hunter Brock Lovett and his team aboard the research vessel *Akademik Mstislav Keldysh* search the wreck of *Titanic* for a necklace with a rare diamond, the Heart ...

 Leonardo DiCaprio Jack	 Kate Winslet Rose	 Frances Fisher Ruth	 Billy Zane Caledon	 Kathy Bates Molly
---	--	--	---	---



- **130K+** Movie Meta Data
 - Cast
 - Genres
 -
- **50K+** Trailers
- **45K+** Plot
- **100M+** Images
 - Poster
 - Profile
- **4000+** Movies
- **1000+** Script




/02

Tag-based Understand



Tag-based Understand









Titanic

11/18/1997 ★ 7.5

Drama, Romance, Thriller

In 1996, treasure hunter Brock Lovett and his team aboard the research vessel *Akademik Mstislav Keldysh* search the wreck of *Titanic* for a necklace with a rare diamond, the Heart ...

 Leonardo DiCaprio Jack	 Kate Winslet Rose	 Frances Fisher Ruth	 Billy Zane Caledon	 Kathy Bates Molly
---	--	--	---	---



Tag: genres, plot keywords



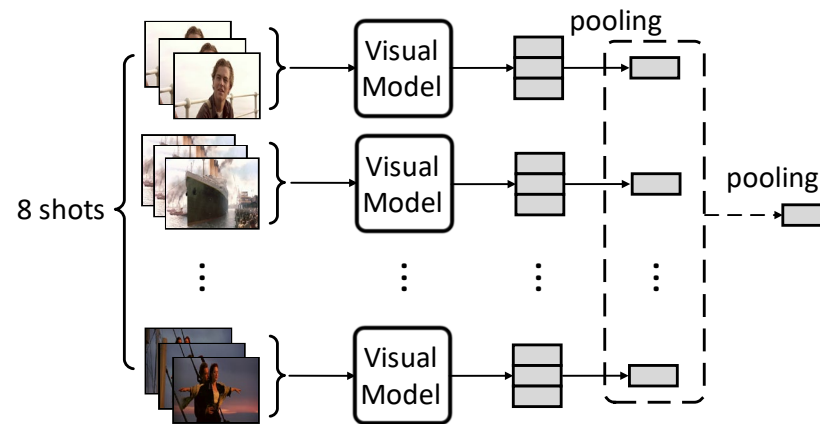
Tag-based Understand

Challenge

- **Movie is too long!** 90min vs. 1min
- **Only tag for the whole movie!**

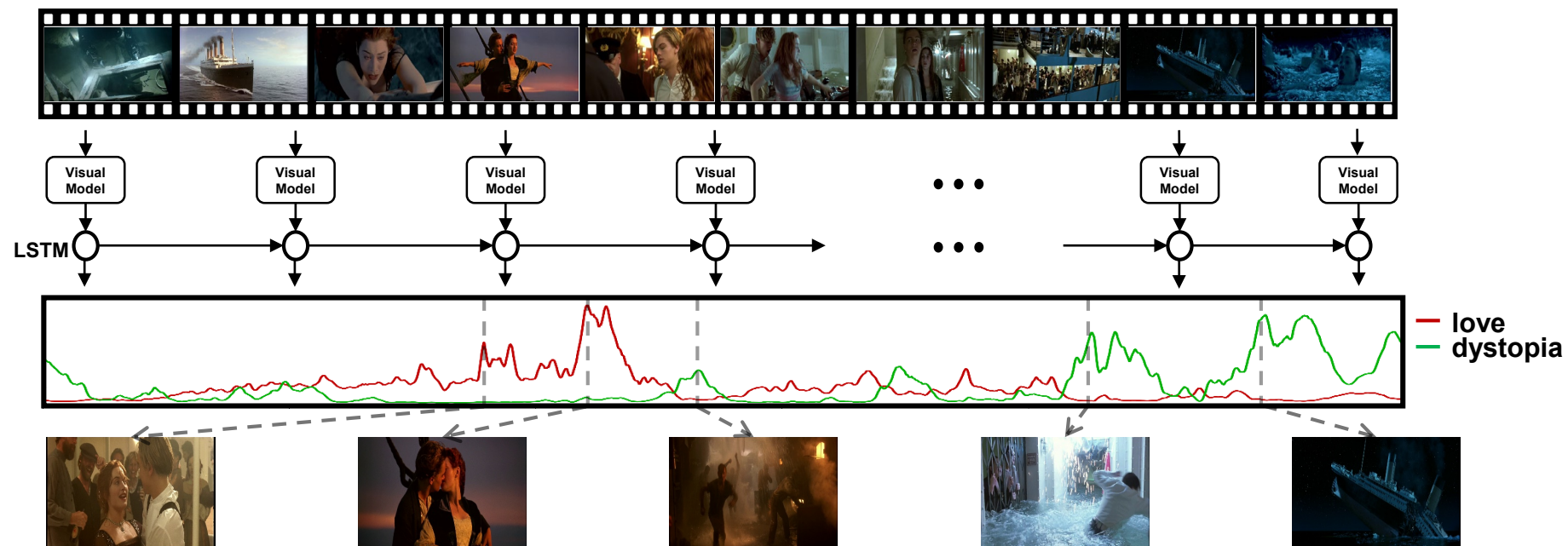
Solution

- **Take shot as unit**
- **Train on trailers**
- **Sparse sampling on training**





Tag-based Understand

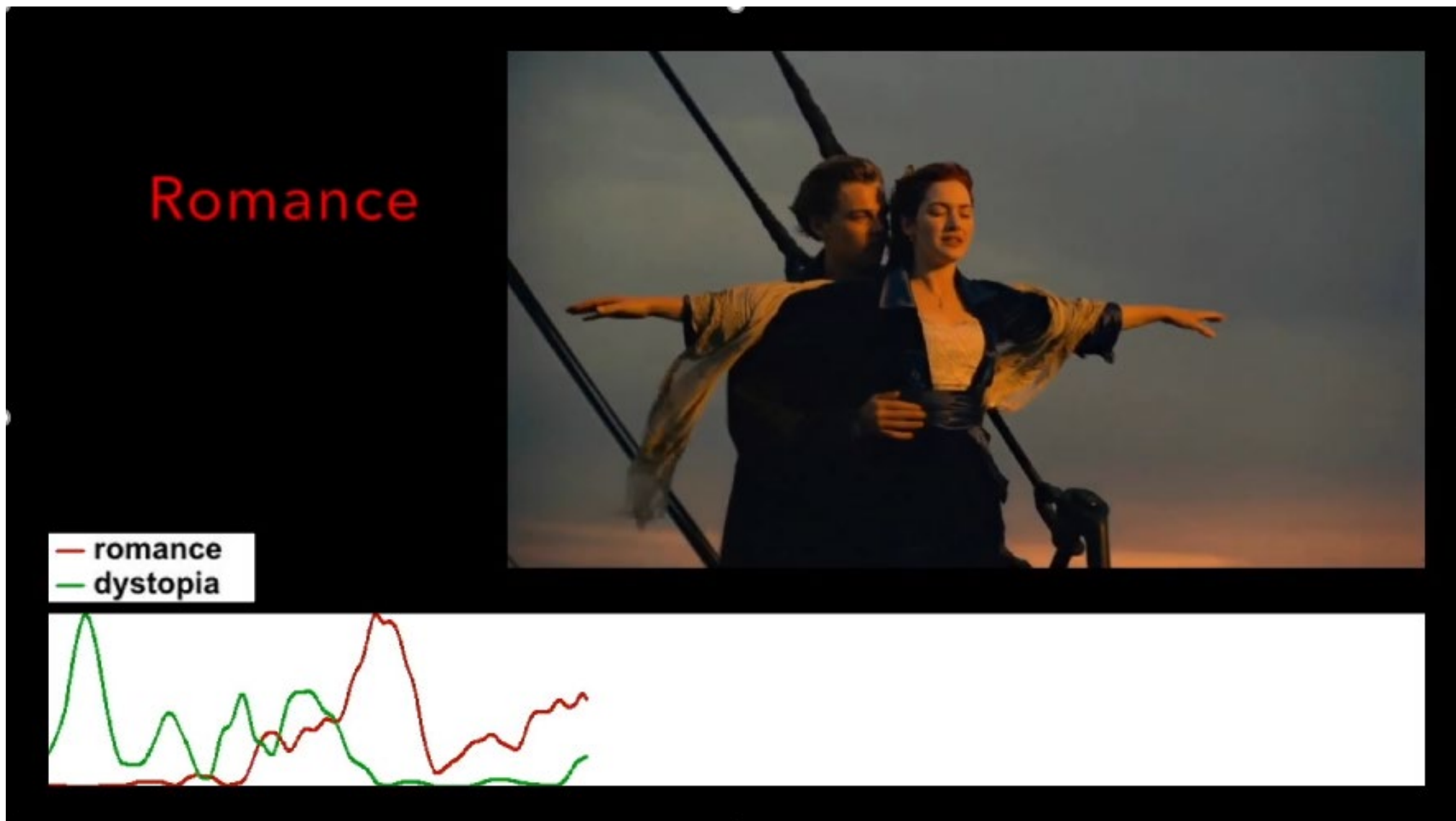


From Trailers to Storylines: An Efficient Way to Learn from Movies

Qingqiu Huang, Yuanjun Xiong, Yu Xiong, Yuqi Zhang, Dahua Lin



Tag-based Understand





Clips Retrieval by Tags



Teenager
Retrieval in Tomorrowland



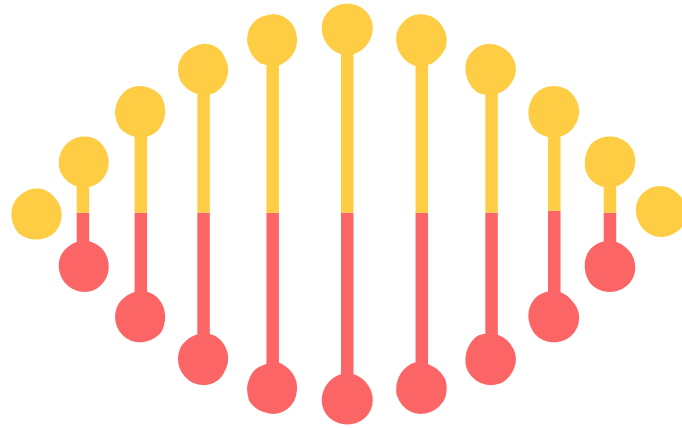
/03

Story-based Understand



Elements of Story



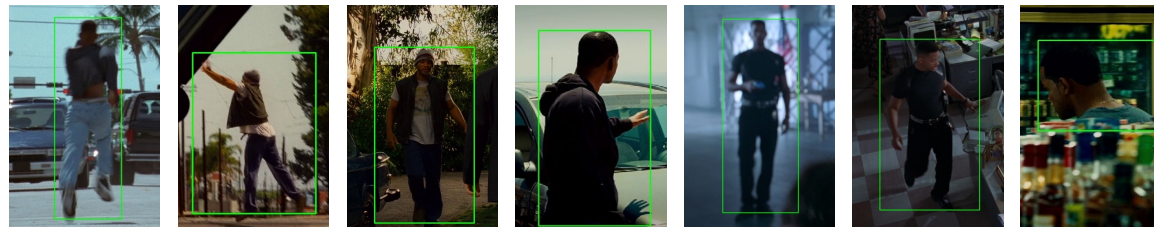
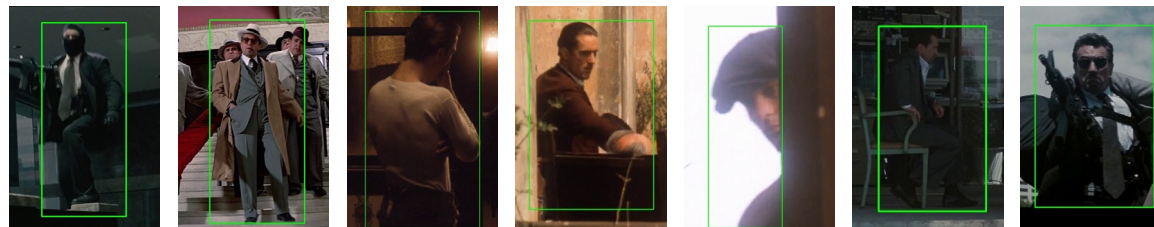
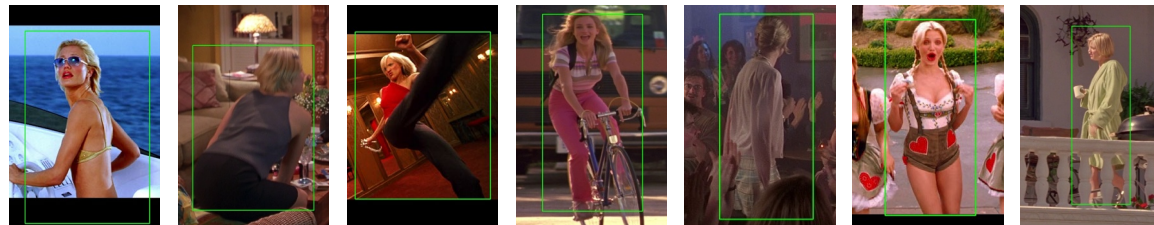


Cast



Cast in Movies (CIM)

- **3348** cast from **630** movies
- More than **1.2M** instances
- Bounding box and identity are manually annotated





Leonardo DiCaprio in CIM





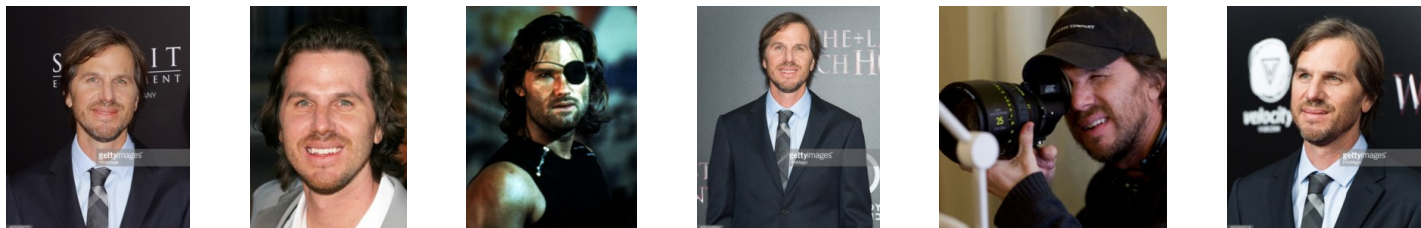
Kate Winslet in CIM





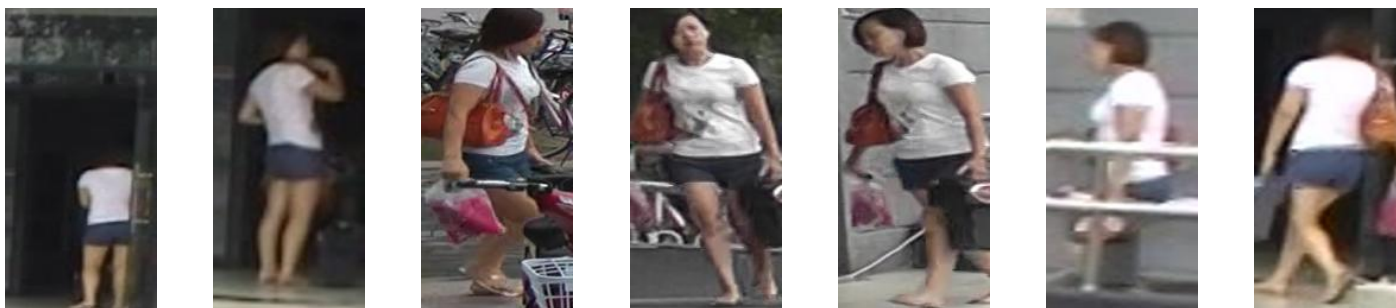
Cast Recognition

- Face Recognition



-- from MS-Celeb-1M

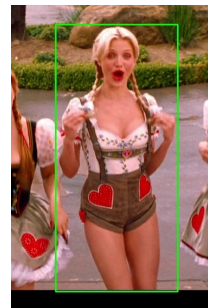
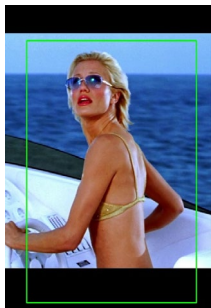
- Person Re-identification



-- from MARS



Cast Recognition



- Most of the instances in movie are without frontal faces -- **Face Recognition Failed**
- Clothing and makeup would change a lot -- **Person Re-id Failed**



Cast Recognition with Context



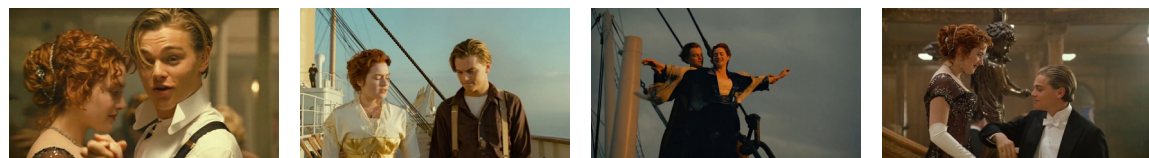
Who are they?

With Face + Visual Context + Social Context

Person-Event



Person-Person



Rose Jack Caledon Molly Ruth Edward Brock



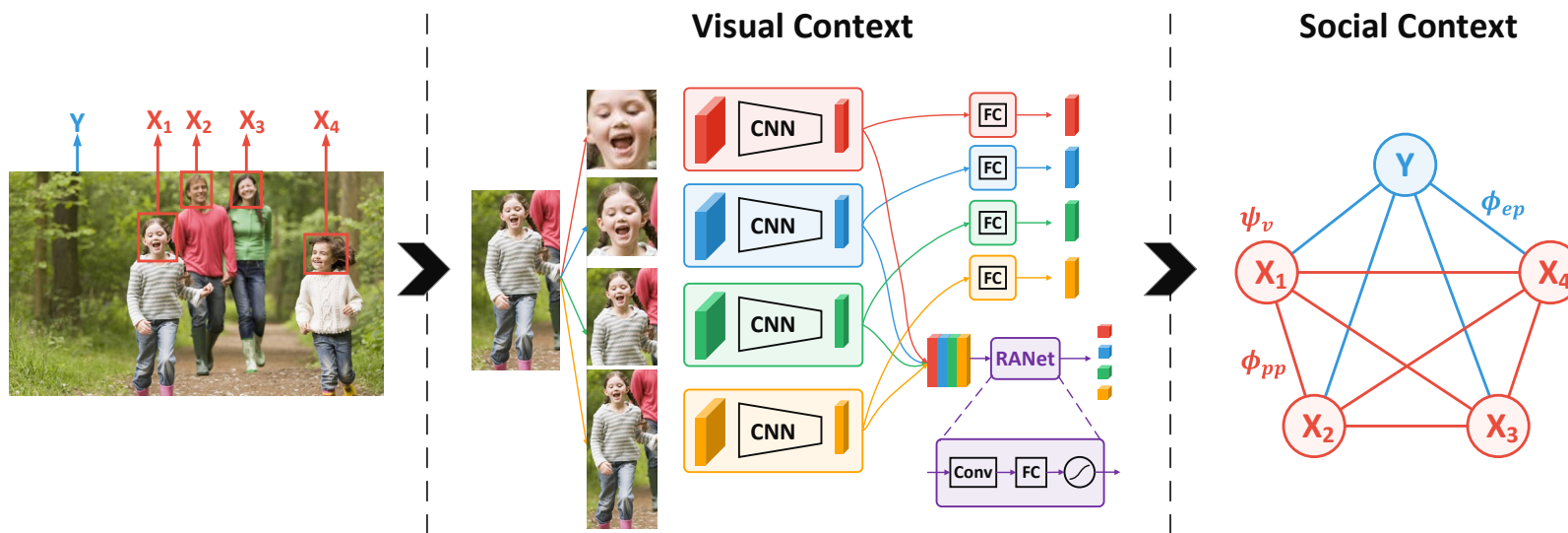
Cast Recognition with Context

- Learn instance-specific weights for different regions with a Region Attention Network (RANet)

$$s(i, j) = \sum_{r=1}^R w_i^r w_j^r s^r(i, j)$$

- Join person identification with social context learning, including person-person and event-person relations

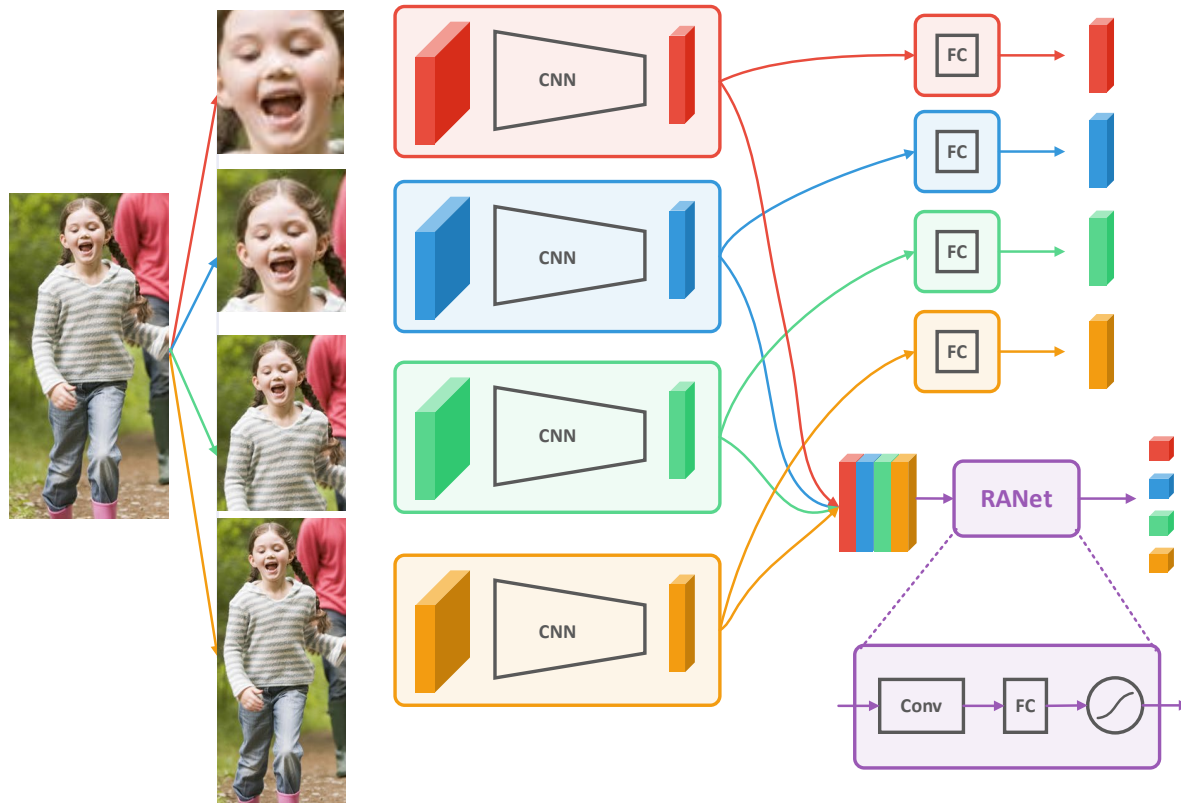
$$J(\mathbf{X}, \mathbf{Y}; \tilde{\mathbf{F}}, \mathbf{P}, \mathbf{Q} | \mathbf{S}, \mathbf{F}) = \psi_v(\mathbf{X} | \mathbf{S}) + \alpha \cdot \phi_{ep}(\mathbf{Y}, \mathbf{X}; \tilde{\mathbf{F}}, \mathbf{P} | \mathbf{F}) + \beta \cdot \phi_{pp}(\mathbf{X}; \mathbf{Q})$$



Unifying Identification and Context Learning for Person Recognition



Visual Matching

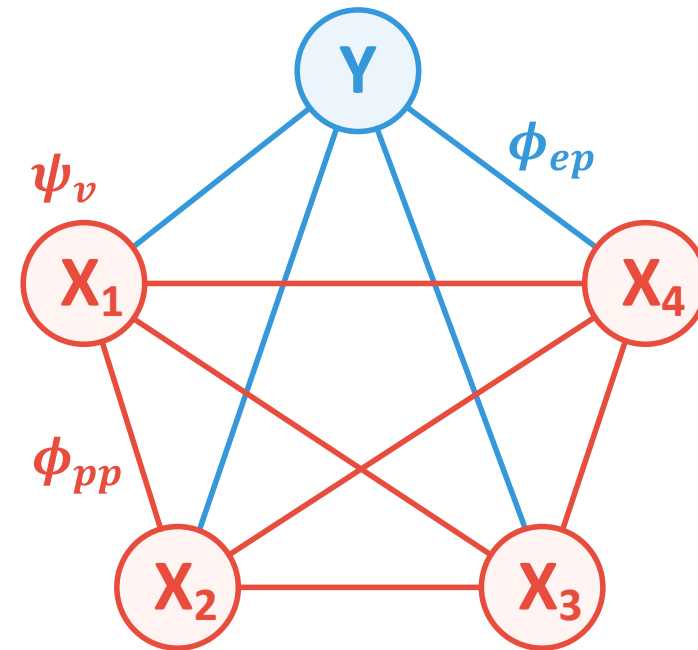
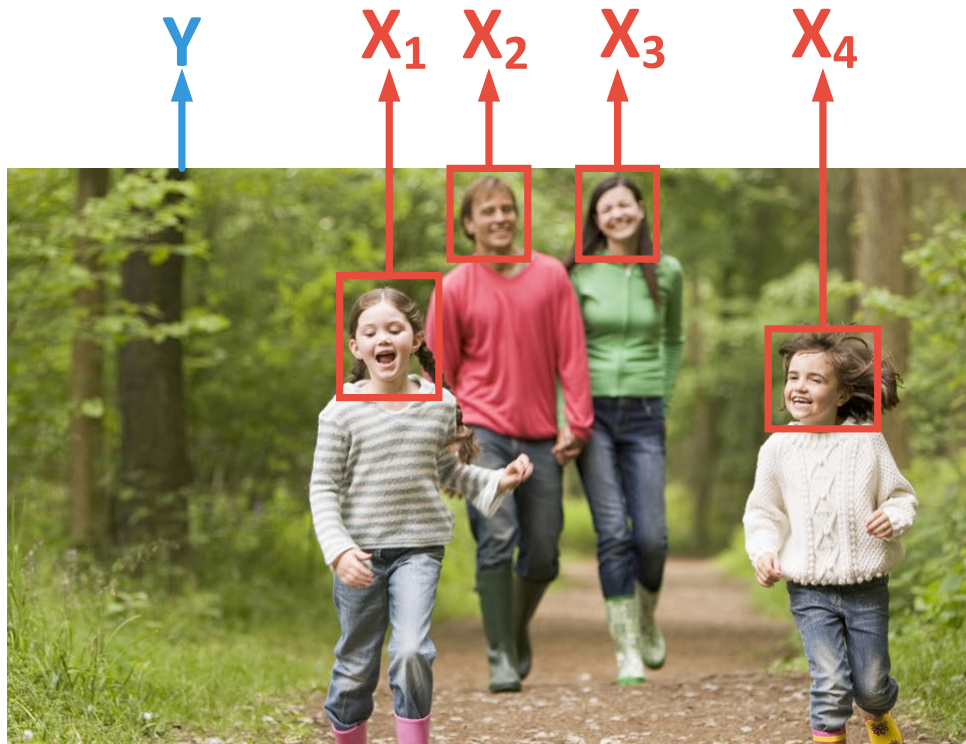


Region specific Weights

$$s(i, j) = \sum_{r=1}^R w_i^r w_j^r s^r(i, j).$$



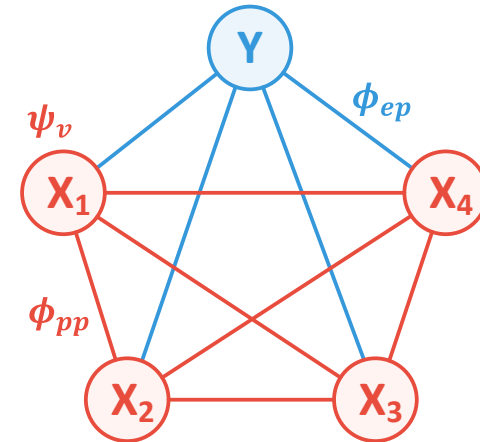
Unified Formulation with Social Context





Unified Formulation with Social Context

$$J(\mathbf{X}, \mathbf{Y}; \tilde{\mathbf{F}}, \mathbf{P}, \mathbf{Q} \mid \mathbf{S}, \mathbf{F}) = \psi_v(\mathbf{X} \mid \mathbf{S}) \\ + \alpha \cdot \phi_{ep}(\mathbf{Y}, \mathbf{X}; \tilde{\mathbf{F}}, \mathbf{P} \mid \mathbf{F}) + \beta \cdot \phi_{pp}(\mathbf{X}; \mathbf{Q}).$$



$$\phi_{ep}(\mathbf{Y}, \mathbf{X}; \tilde{\mathbf{F}}, \mathbf{P} \mid \mathbf{F}) = \sum_{i=1}^M \sum_{k=1}^K a_i^k y_i^k \\ \text{with } a_i^k = \sum_{j \in \mathcal{I}_i} \log(\mathbf{p}_k)^T \mathbf{x}_j - \|\mathbf{f}_i - \tilde{\mathbf{f}}_k\|^2,$$

$$\phi_{pp}(\mathbf{X}; \mathbf{Q}) = \sum_{i=1}^M \sum_{j \in \mathcal{I}_i} \sum_{j' \in \mathcal{I}_i: j' \neq j} \mathbf{x}_j^T \mathbf{Q} \mathbf{x}_{j'}.$$



Experiments

Dataset	Split	Existing Methods on PIPA				Ours		
		PIPER	Naeil	RNN	MLC	Baseline	RANet Fusion	Full Model
PIPA	Original	83.05	86.78	84.93	88.20	82.79	87.33	89.73
	Album	-	78.72	78.25	83.02	75.24	82.59	85.33
	Time	-	69.29	66.43	77.04	66.55	76.52	80.42
	Day	-	46.61	43.73	59.77	47.09	65.49	67.16
CIM	-	-	-	-	-	68.12	71.93	72.56



Experiments

Experiments of Recognition Results

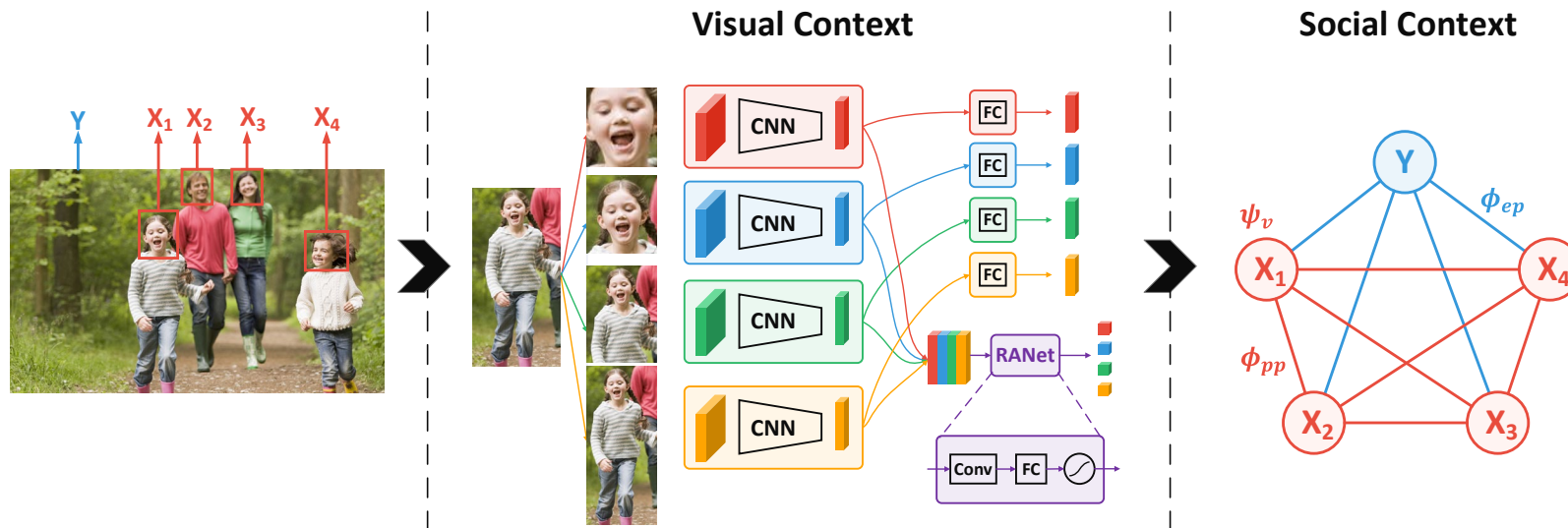


Events Discovered by Our Approach





Conclusion



- A new framework
 - Region Attention Network to adaptively combine visual cues
 - Unify person identification and context learning in joint inference
- Get state-of-the-art performance on PIPA and CIM



Cast Search with One Portrait

Query

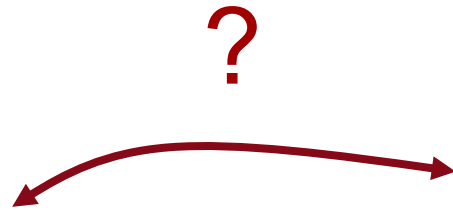


Database



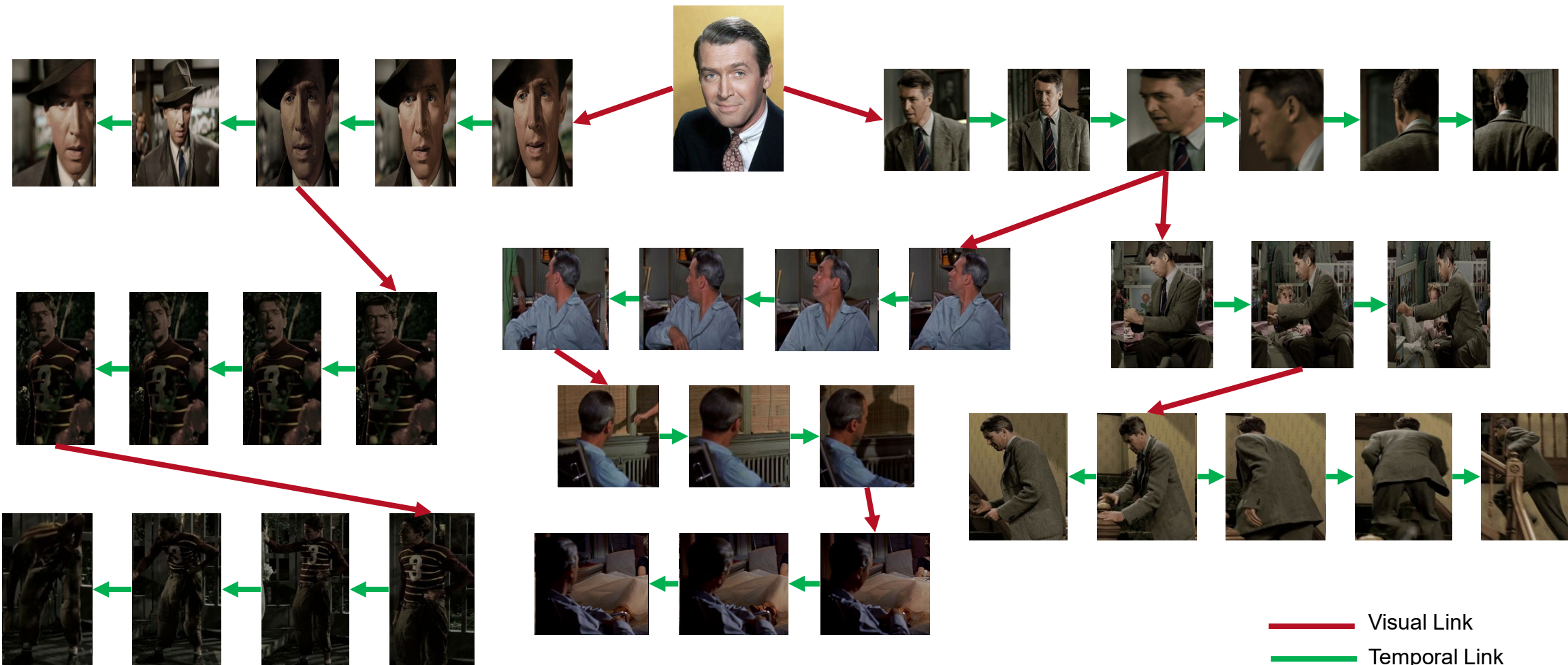


Cast Search with One Portrait





Cast Search with One Portrait



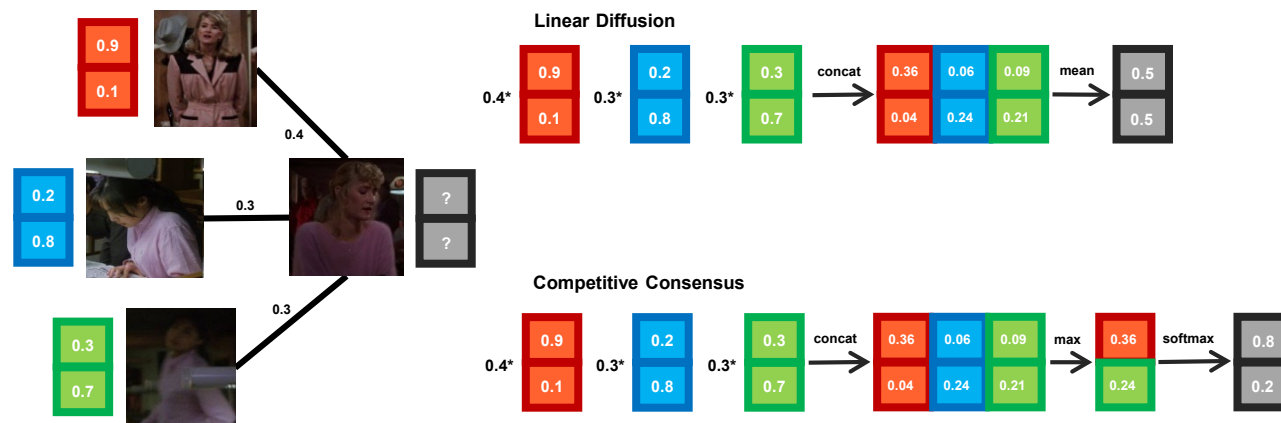
Person Search in Videos with one Portrait through Visual and Temporal Links

Qingqiu Huang, Wentao Liu, Dahua Lin *European Conference of Computer Vision (ECCV) 2018*



Cast Search with One Portrait

- Competitive Consensus



- Progressive Propagation

• mAP: 33.66% -> 47.41%



Experiments

	IN			ACROSS		
	mAP	R@1	R@3	mAP	R@1	R@3
FACE	53.55	76.19	91.11	42.16	53.15	61.12
LP	8.19	39.70	70.11	0.37	0.41	1.60
PPCC	63.49	83.44	94.40	62.27	62.54	73.86



Cast Search with One Portrait

	1 st Propagation	3 rd Propagation	5 th Propagation
			
			
			



Cast Search in a Whole Movie

Cast Search in Movies

The image displays a large grid of movie frames, likely from a single film, illustrating a cast search process. The frames are arranged in a grid, with a vertical blue line highlighting a specific frame in the top row. The frames show various scenes from the movie, with green bounding boxes around faces, indicating the search results for a specific actor. In the bottom left corner, there is a portrait of a woman with blonde hair, labeled "Kate".

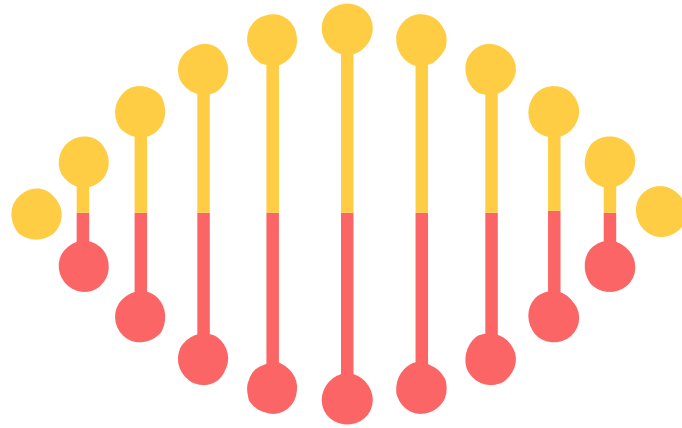


Future Work

- Memory

- Speech & Subtitle





Event



Event Retrieval and Localization by Natural Language

... Everyone looks up as a string of sand whizzes past like an express train. As the van doors are closed the sandstorm zooms in like a swarm of angry bees. The weight of the sand presses the accelerator on the van, picks up speed. ...



Everyone looks up as a string of sand whizzes past like an express train.



As the van doors are closed the sandstorm zooms in like a swarm of angry bees.

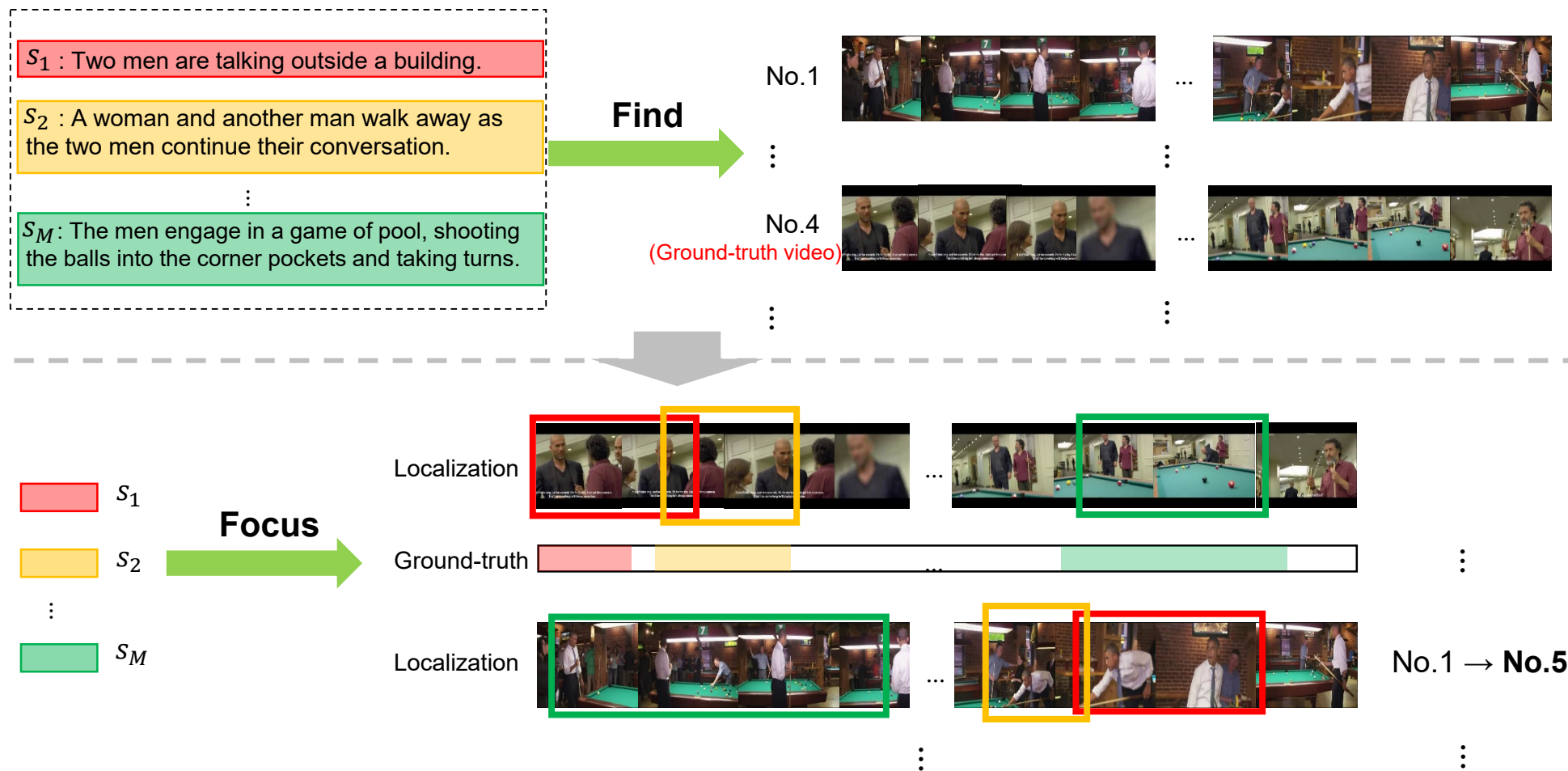


The weight of the sand presses the accelerator on the van, picks up speed.

Find and Focus: Retrieve and Localize Video Events with Natural Language Queries



Event Retrieval and Localization by Natural Language



Find and Focus: Retrieve and Localize Video Events with Natural Language Queries



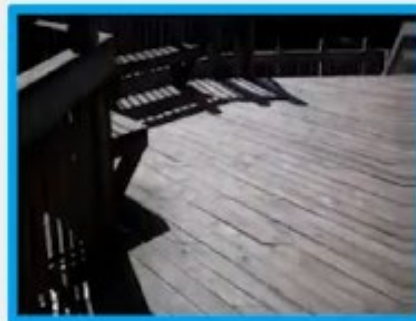
Event Retrieval and Localization by Natural Language

Find and Focus

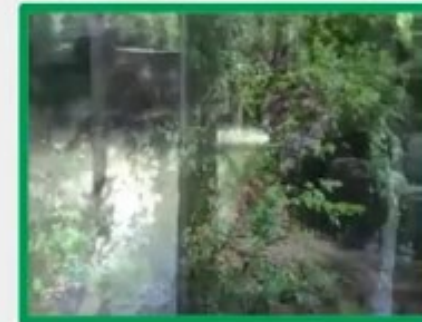
🔍 A river of water is shown. A dog is walking on a leash by the water. A bridge is shown above the water. ✕



A river of water is shown.



A dog is walking on a leash by the water.



A bridge is shown above the water.



Future Work

- Story-based Summary
- Caption (Story Telling)



/04 Conclusion



Conclusion

- A Large-scale Movie Dataset
- Tag-based Understand
 - Learn from trailers to get shot-level tag response
- Story-based Understand
 - Cast
 - Cast recognition with context
 - Cast search through visual and temporal links
 - Event
 - Hierarchical framework for video retrieval by natural language
 -



Thank You

Qingqiu Huang
25/03/2019