跨模态内容分析

北航人工智能研究院
刘偲

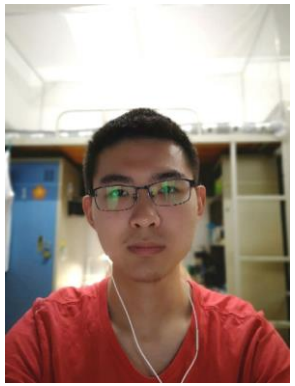
可乐实验室

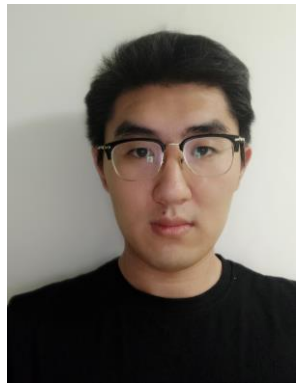
si liu

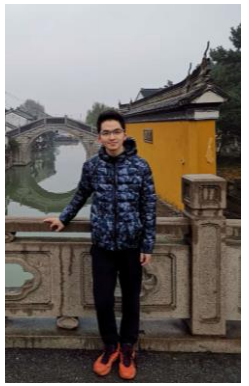
Beihang University
Verified email at buaa.edu.cn - Homepage
semantic segmentation vision + language GANs

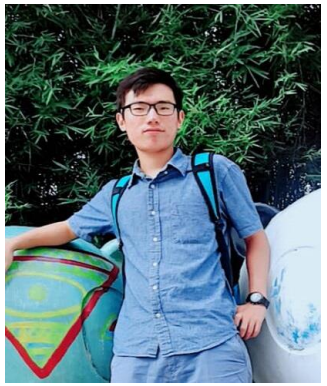
Cited by

	All	Since 2015
Citations	5665	5087
h-index	32	30
i10-index	50	47









Vision and Language

CV	+	NLP	=	Vision-to-Language (V2L)
Image Understanding	+	Language Generation	=	Image Captioning
Image Classification Object Detection Segmentation Object Counting Colour Analysis ….	+	Question Answering	=	Visual Question Answering
Image Understanding	+	Dialog	=	Visual Dialog

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

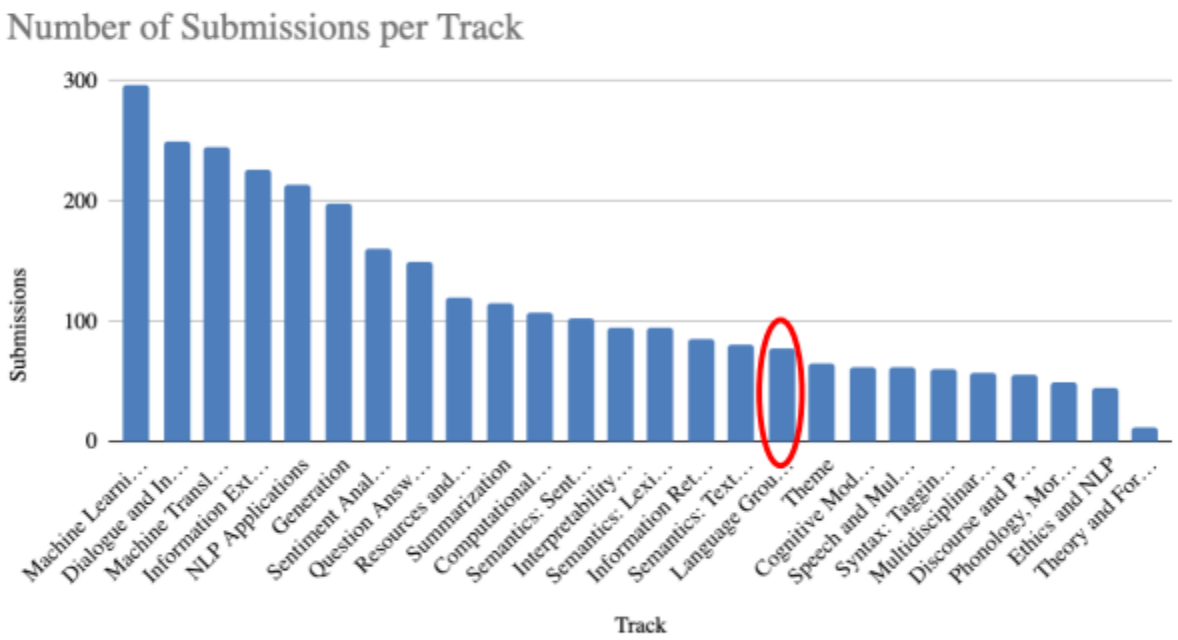
What color is the tablecloth?

Ours: white

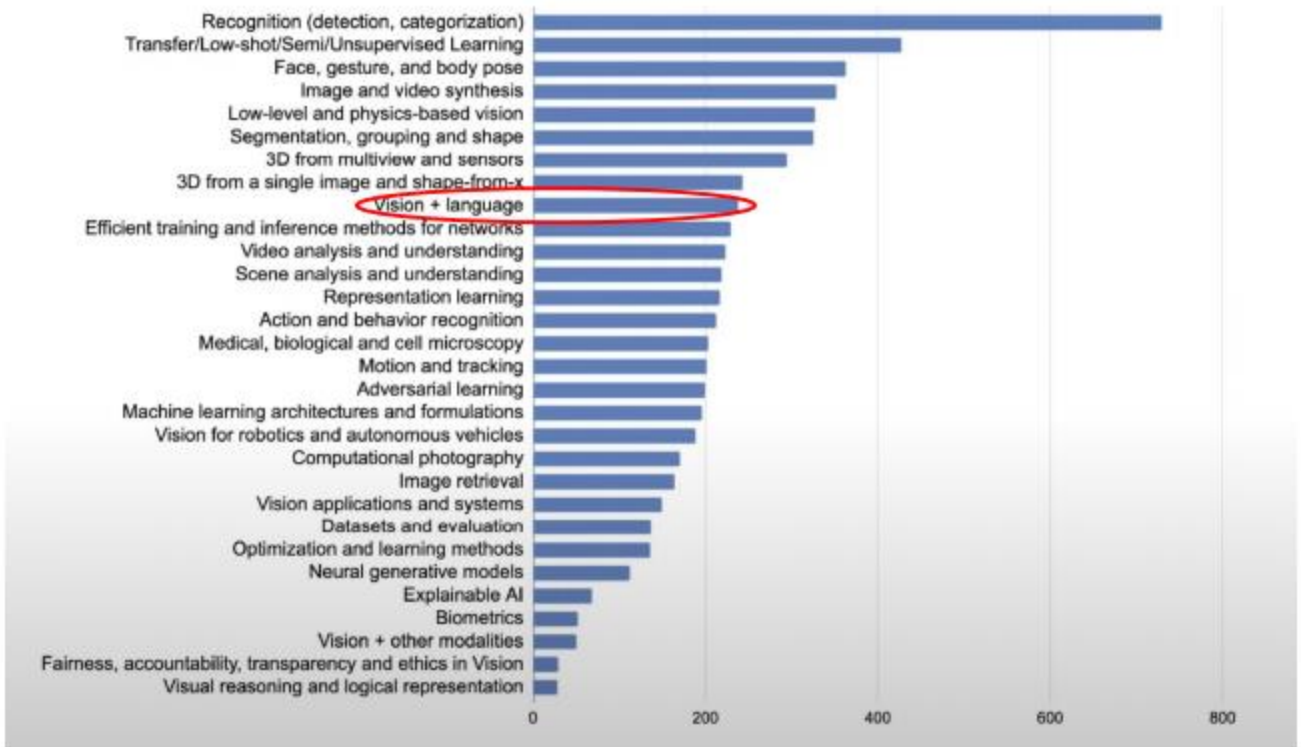
How many people in the photo?

2

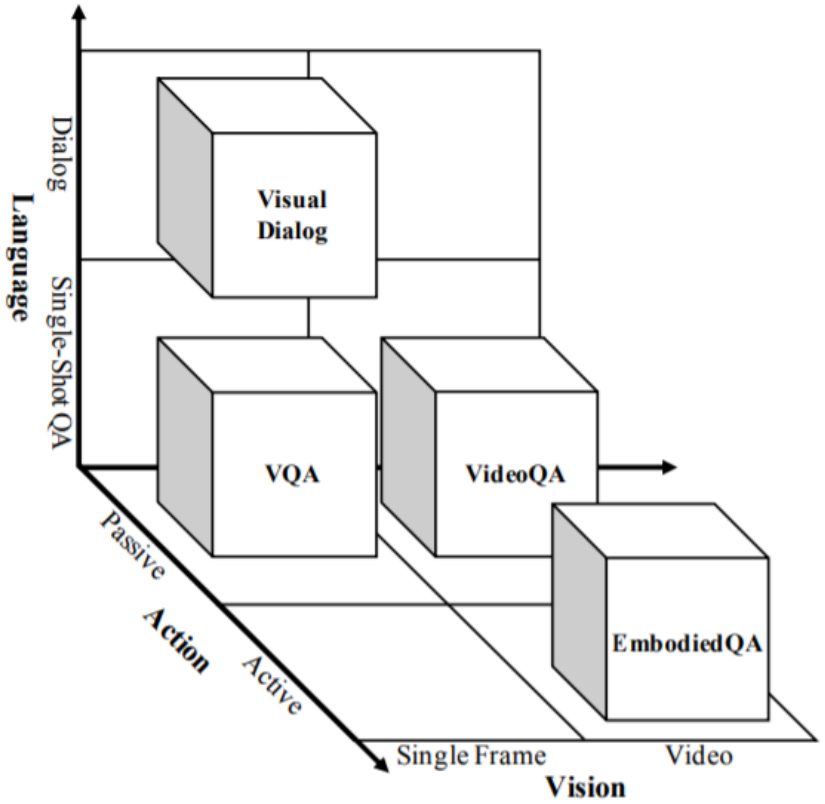
Language and Vision @ ACL 2020



Language and Vision @ CVPR 2020

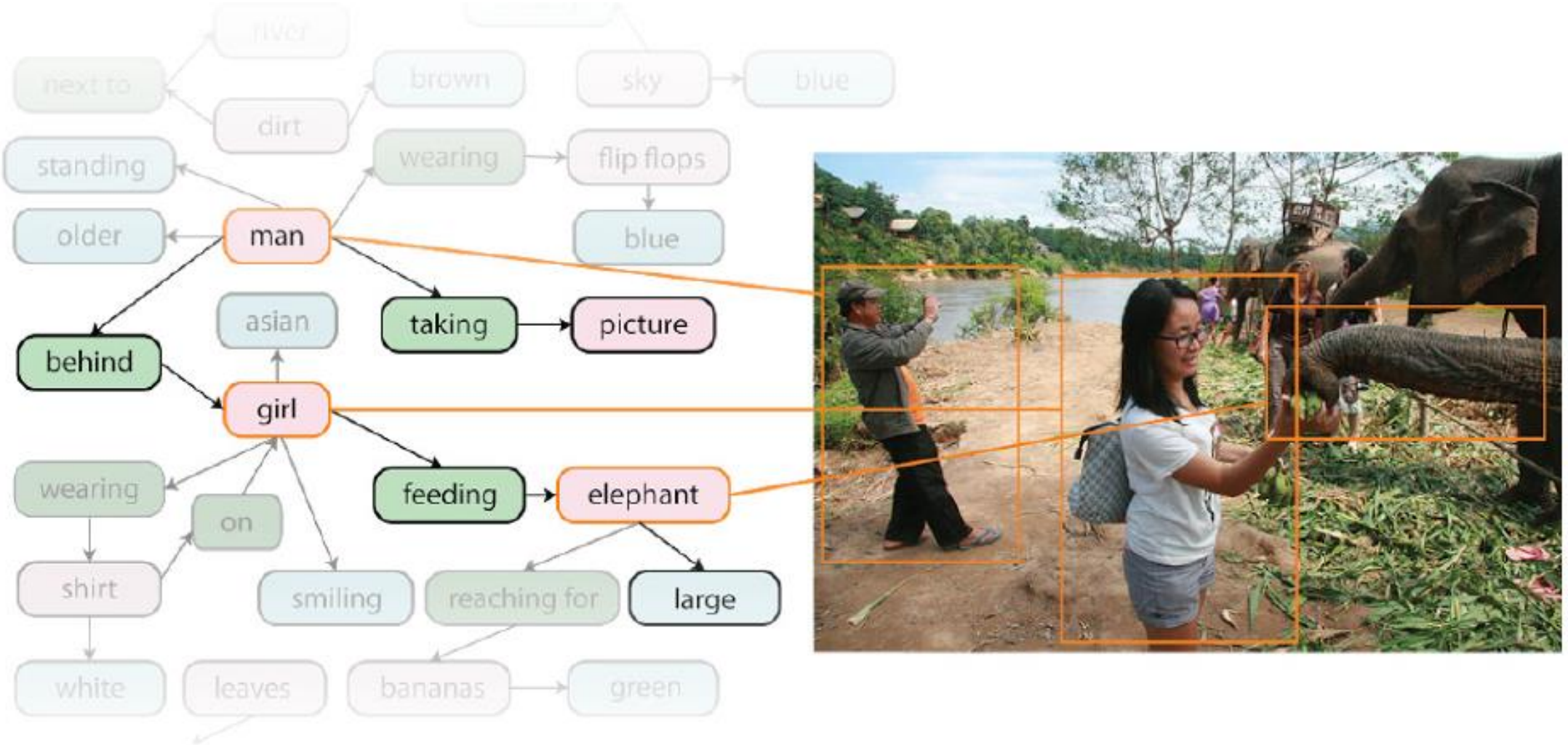


Relations with others



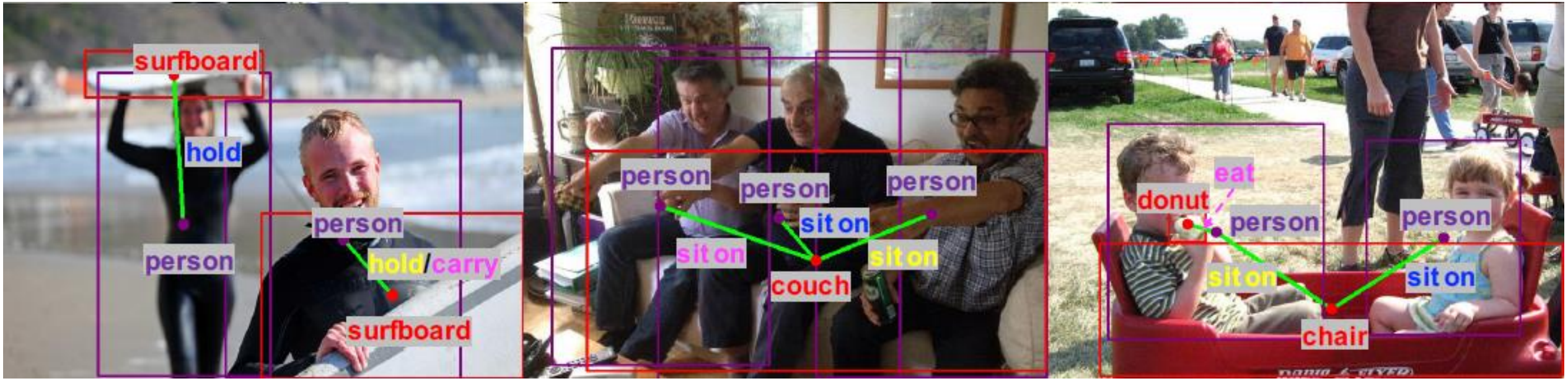
- Visual relation detection
- Referring expression

视觉关系检测



Guanghui Ren, Lejian Ren, Yue Liao, **Si Liu** et al,《Scene Graph Generation with Hierarchical Context》 TNNLS 2020

1.Human object interaction



挑战：算法复杂度高，无法实时应用

Yue Liao, **Si Liu** et al, PPDM: Parallel Point Detection and Matching for Real-time Human-Object Interaction Detection, CVPR 2020

挑战：现实场景常存在较多干扰实例，难以高效精准匹配有交互关系的实例

输入图片

前人方法：自底而上，轮询配对

思路：自顶而下，通过关系锚点来匹配实例

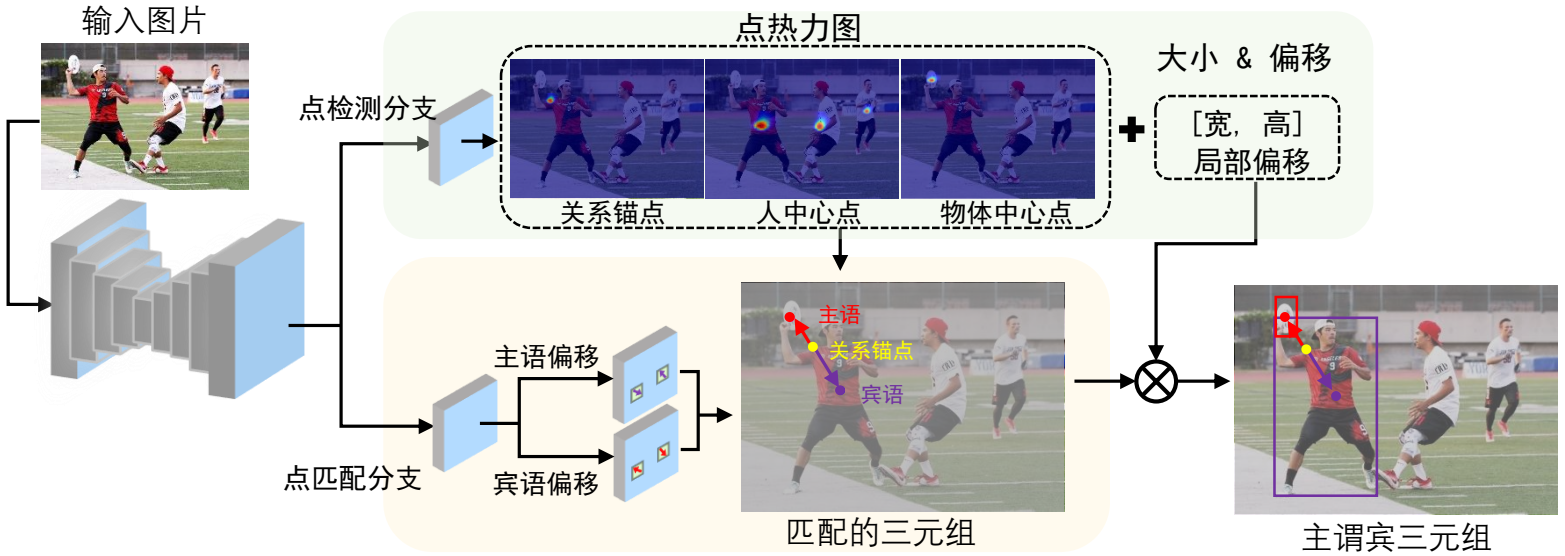
匹配主宾锚点

输出匹配主宾

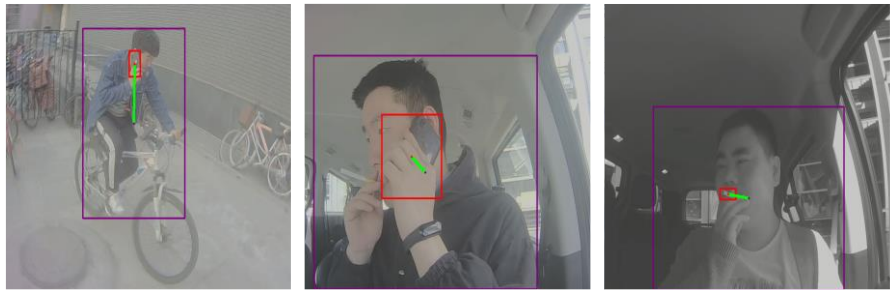
Yue Liao （学生）; Si Liu* et al, PPDM : Parallel Point Detection and Matching for Real-time Human-Object Interaction Detection, CVPR 2020.
Yue Liao （学生）; Si Liu* et al, A Real-Time Cross-modality Correlation Filtering Method for Referring Expression Comprehension, CVPR 2020.

首个自顶而下、实时人-物实例关系匹配框架
并应用于商汤智能车舱和互联网教育产品线

重定义人-物实例关系匹配问题为并行点检测和点匹配问题



危险动作检测结果可视化



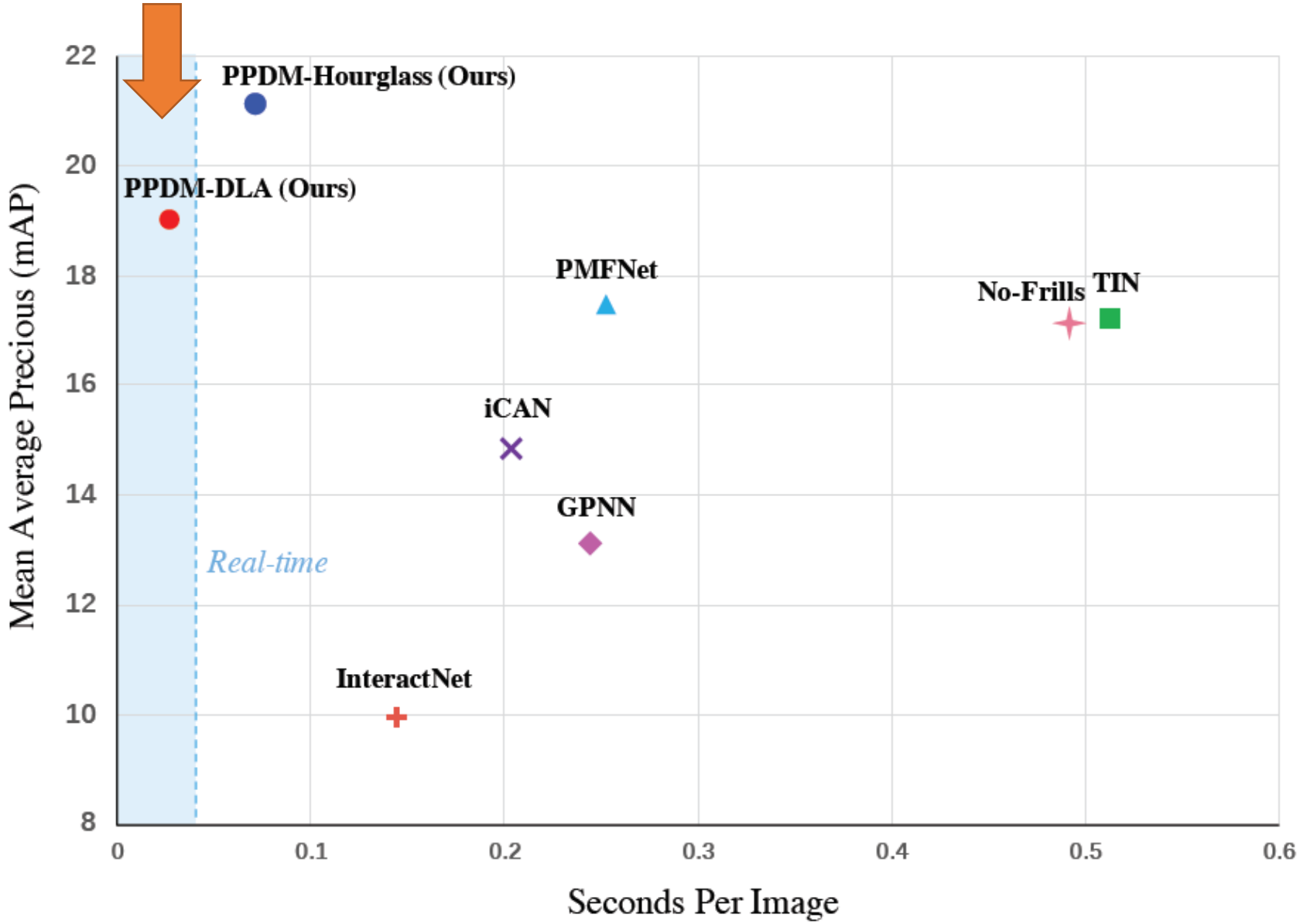
<人，用…通话，手机><人，用…通话，手机> <人，抽(烟)，香烟>

定量实验结果

Method	mAP (%)	Time (ms)
Faster Interaction Net [1]	56.93	-
GMVM [1]	60.26	-
C-HOI [34]	66.04	-
iCAN [7]	44.23	194
TIN [16]	48.64	501
PPDM-DLA	67.45	**27**
PPDM-Hourglass	**71.23**	71

提速近20倍
精度提升38.4%

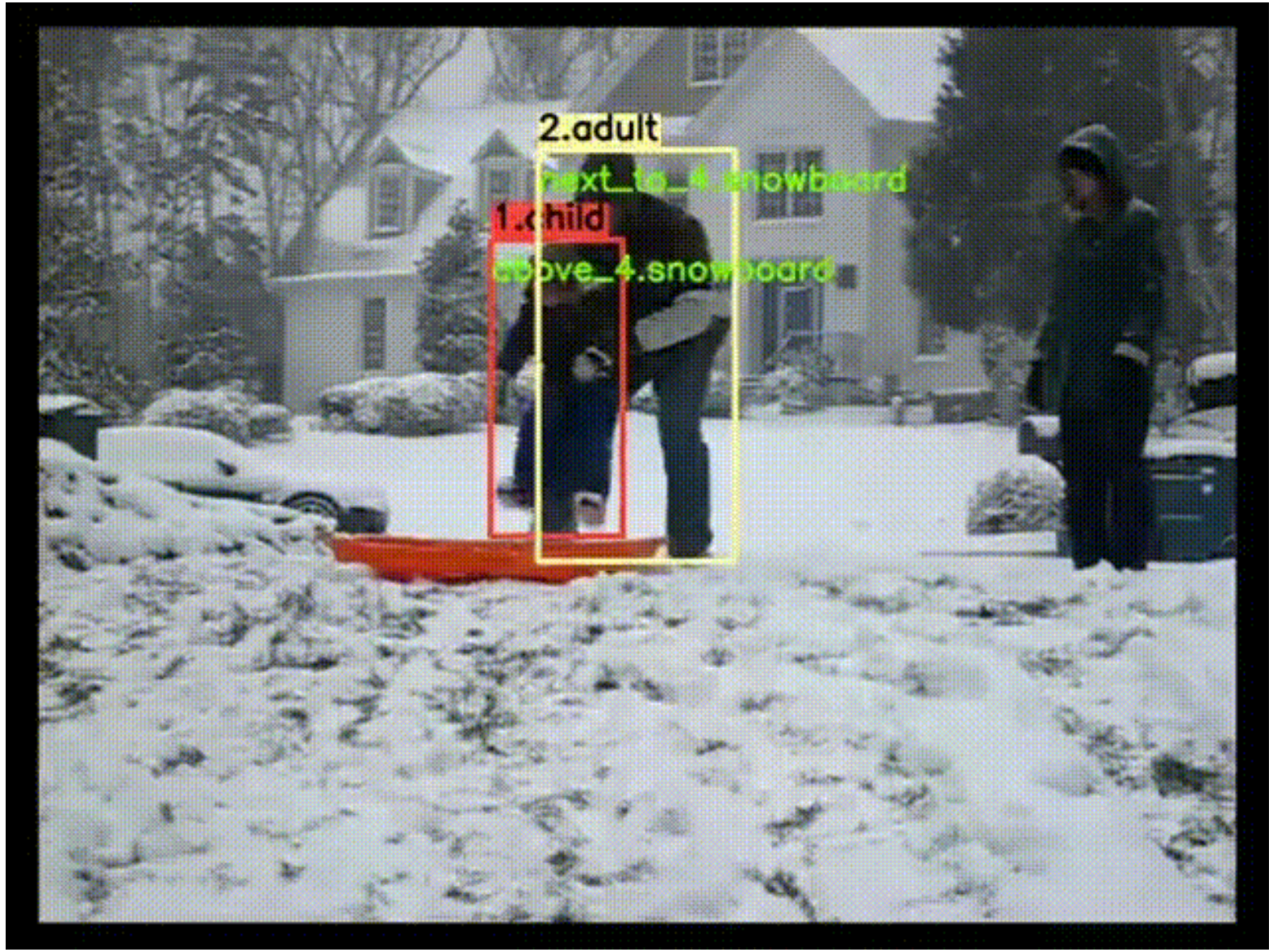
Yue Liao （学生）; Si Liu* et al, PPDM : Parallel Point Detection and Matching for Real-time Human-Object Interaction Detection, CVPR 2020.
Yue Liao （学生）; Si Liu* et al, A Real-Time Cross-modality Correlation Filtering Method for Referring Expression Comprehension, CVPR 2020.



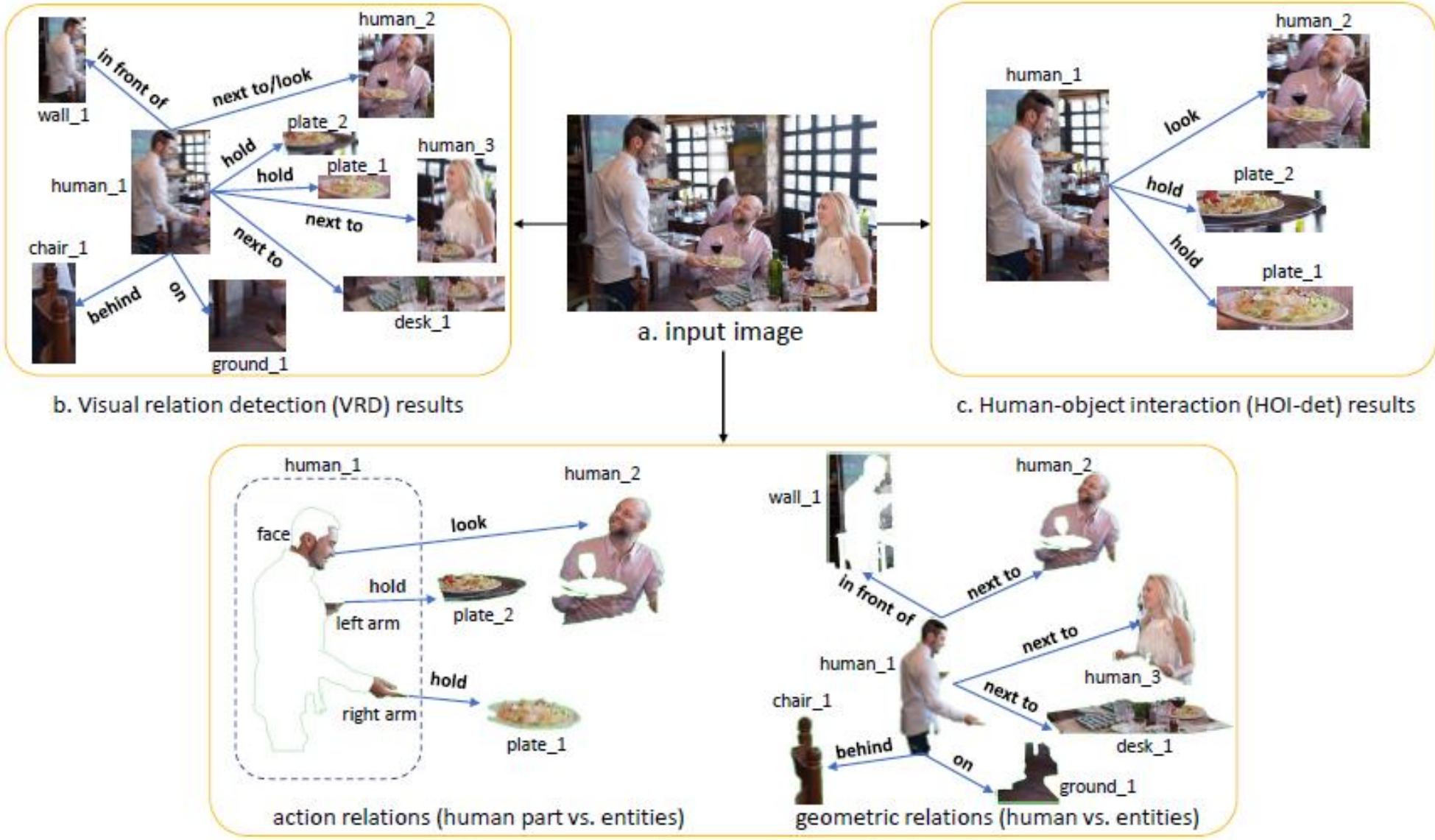
Yue Liao, **Si Liu** et al, PPDM: Parallel Point Detection and Matching for Real-time Human-Object Interaction Detection, CVPR 2020

https://github.com/YueLiao/PPDM

Method	Feature	Full(mAP %) ↑	Rare(mAP %) ↑	Non-Rare(mAP %) ↑	Inference Time (ms) ↓	FPS ↑
Shen *et. al* [23]	A + P	6.46	4.24	7.12	-	-
HO-RCNN [2]	A + S	7.81	5.37	8.54	-	-
VSRL [10]	A	9.09	7.02	9.71	-	-
InteractNet [8]	A	9.94	7.16	10.77	145	6.90
GPNN [21]	A	13.11	9.34	14.23	197 + 48 = 245	4.08
Xu *et. al* [26]	A + L	14.70	13.26	15.13	-	-
iCAN [6]	A + S	14.84	10.45	16.15	92 + 112 = 204	4.90
PMFNet-Base [24]	A + S	14.92	11.42	15.96	-	-
Wang *et. al* [25]	A	16.24	11.16	17.75	-	-
No-Frills [11]	A + S + P	17.18	12.17	18.68	197 + 230 + 67 = 494	2.02
TIN [15]	A + S + P	17.22	13.51	18.32	92 + 98 + 323 = 513	1.95
RPNN [31]	A + P	17.35	12.78	18.71	-	-
PMFNet [24]	A + S + P	17.46	**15.65**	18.00	92 + 98 + 63 = 253	3.95
PPDM-DLA	A	19.02	12.65	20.92	**27**	**37.03**
PPDM-Hourglass	A	**21.10**	14.46	**23.09**	71	14.08


2.adult
next_to_4.snowboard
1.child
above_4.snowboard

2.Human centric relation segmentation

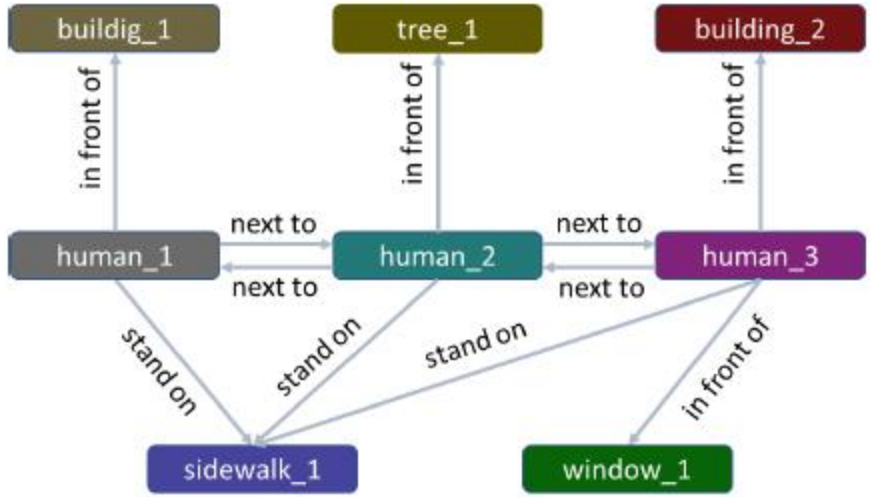


http://picdataset.com/challenge/index/

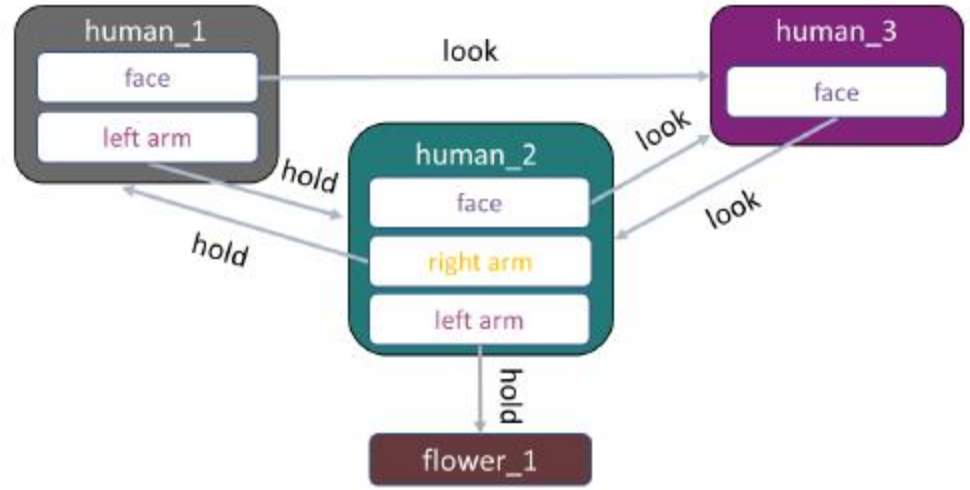
(a) Original image

(b) Entity segmentation

(c) Human parsing



(d) Geometric relations

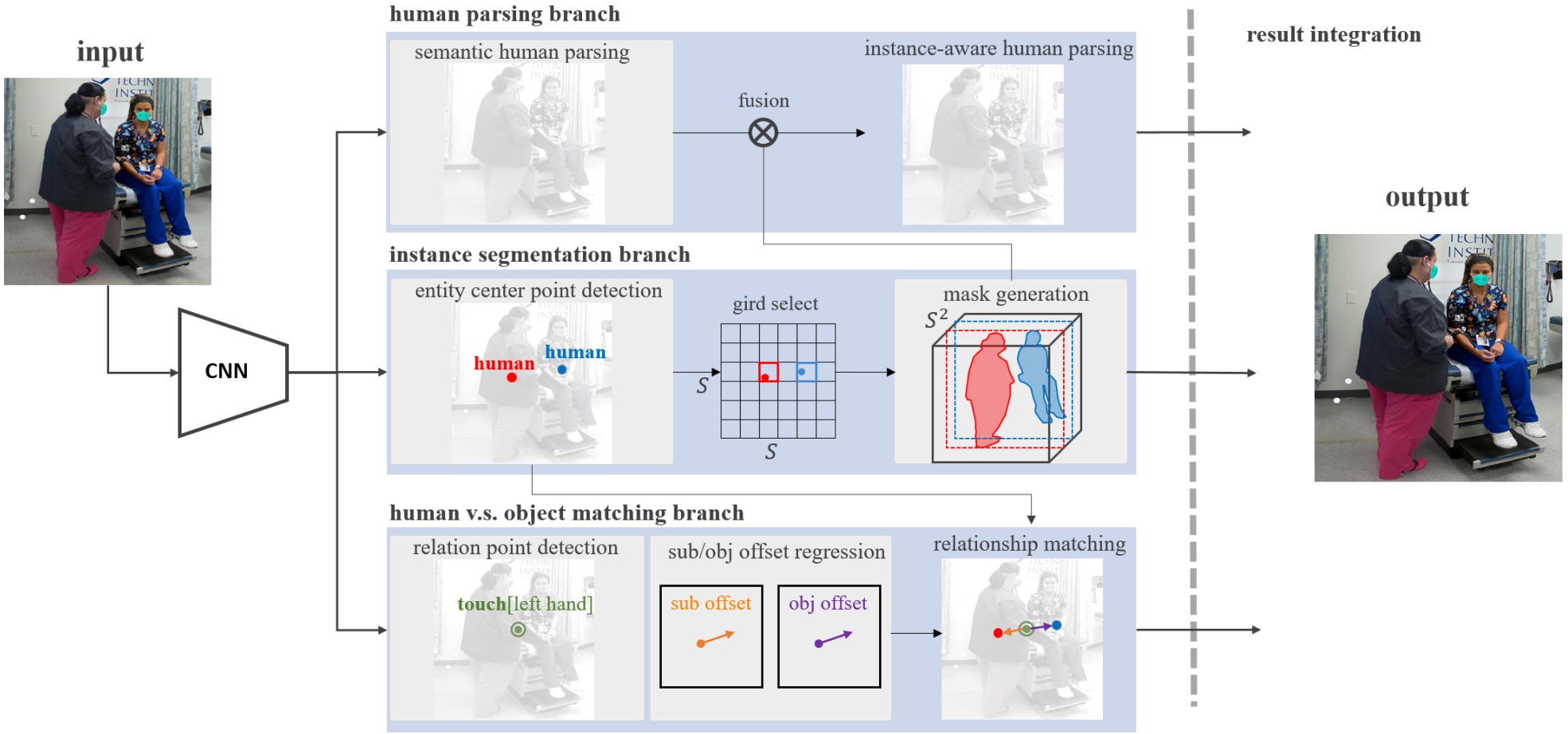


(e) Action relations

Fig. 4: (a) Distribution of action triplets and (b) distribution of geometric triplets.

Dataset	Stuff	Segmentation	Relation	Human part	Human Pose	Average resolution
Visual Gennome [11]	✓	✓	✓	×	✓	413*500
VRD [18]	✓	×	✓	×	×	764*950
VCOCO [16]	×	✓	✓	×	✓	481*578
HICO-DET [17]	×	×	✓	×	×	497*593
PIC	✓	✓	✓	✓	✓	1427*1882

并行匹配分割网络



1. 单阶段并行预测
2. 可以适用于不同backbone，不用head。

(a)

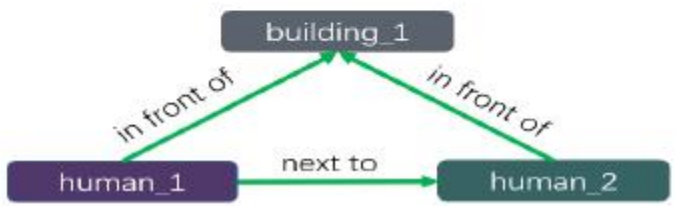

human_1
human_2
vegetation_1
human_3

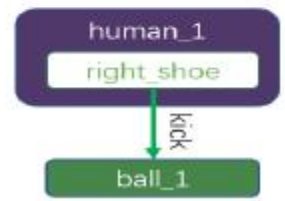

vegetation_1
in front of
human_1
in front of
human_2
in front of
human_3

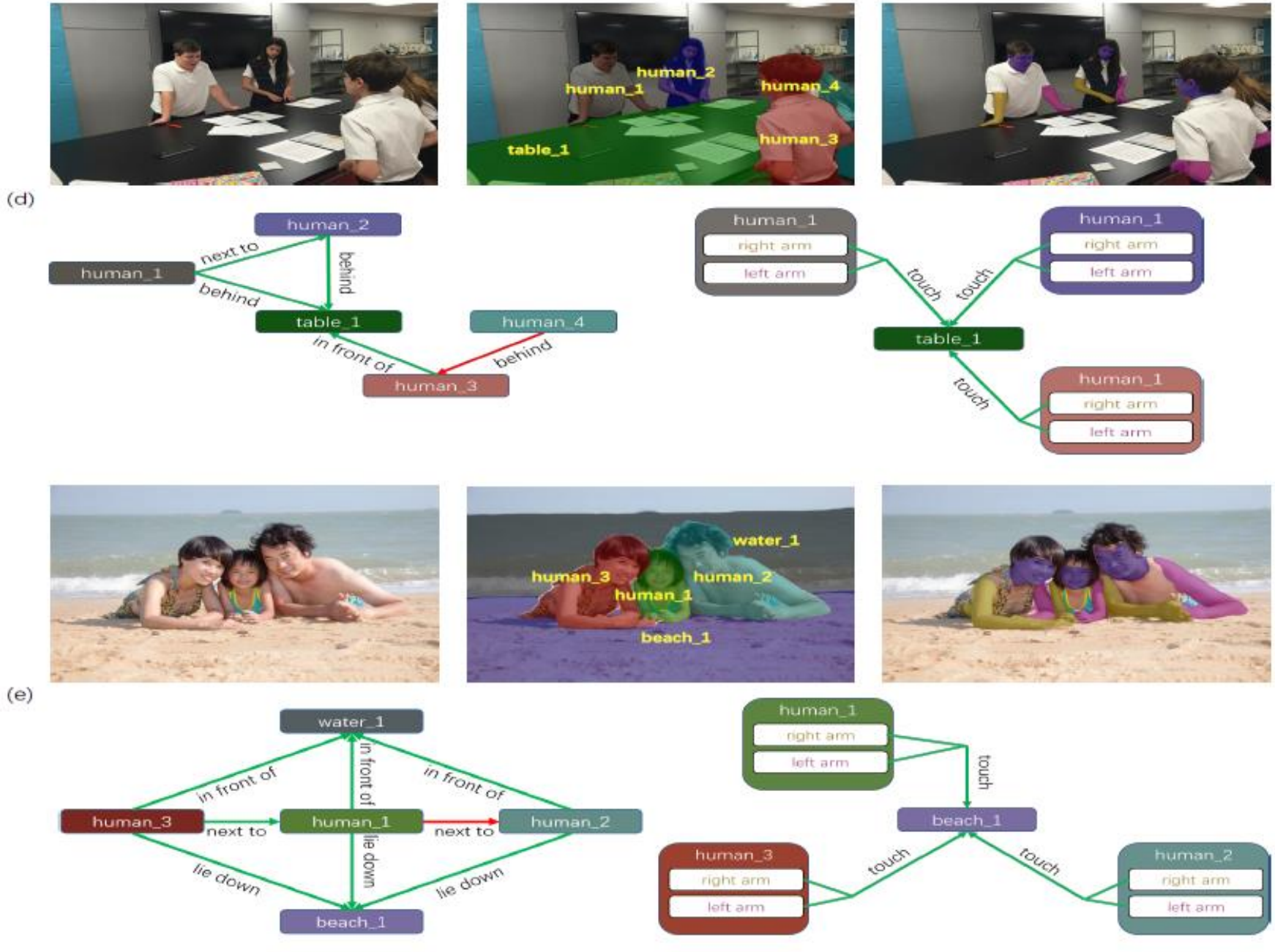

human_1
left arm
pull
pull
human_2
right arm
left arm
pull
pull
human_3
right arm

(b)


building_1
human_2
human_1
ball_1


building_1
in front of
in front of
human_1
next to
human_2


human_1
right_shoe
kick
ball_1


(d)
human_1
human_2
human_3
human_4
table_1
beach_1
water_1
next to
behind
in front of
lie down
touch
right arm
left arm
(e)

两次主办CV顶会竞赛

ECCV 2018
European Conference on Computer Vision
8 – 14 September 2018 | Munich, Germany

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

NUS National University of Singapore

360

ICCV 2019
Seoul, Korea

商汤 sensetime

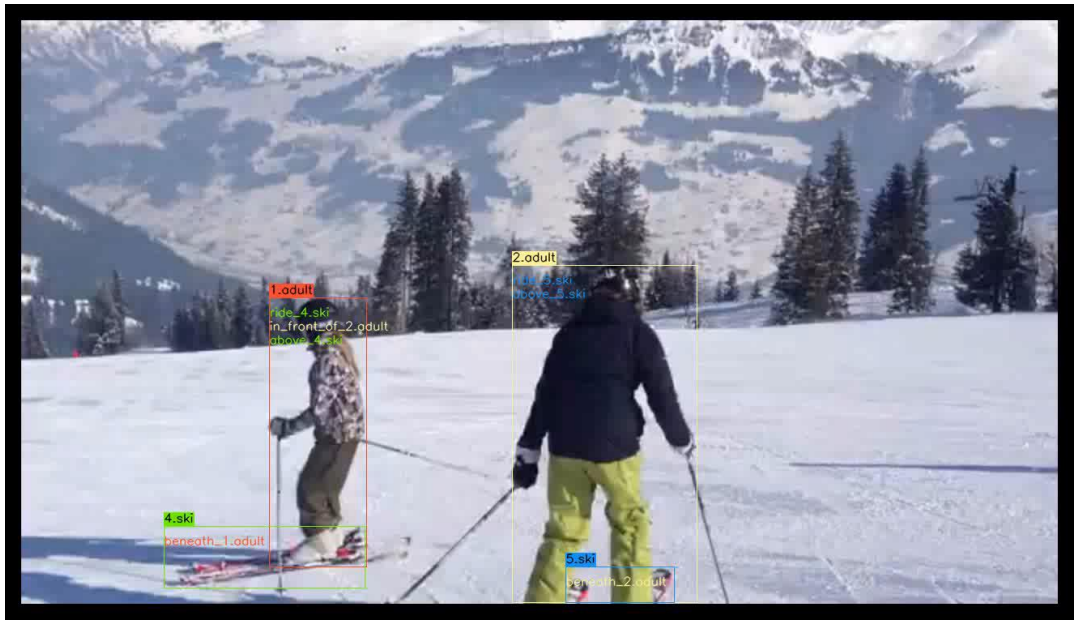
北京航空航天大學
BEIHANG UNIVERSITY

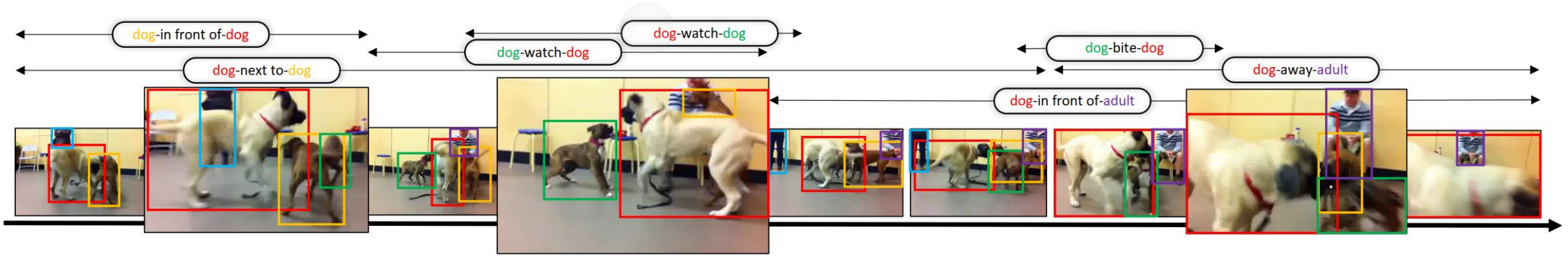
http://picdataset.com/challenge/index/

3. Video relation detection

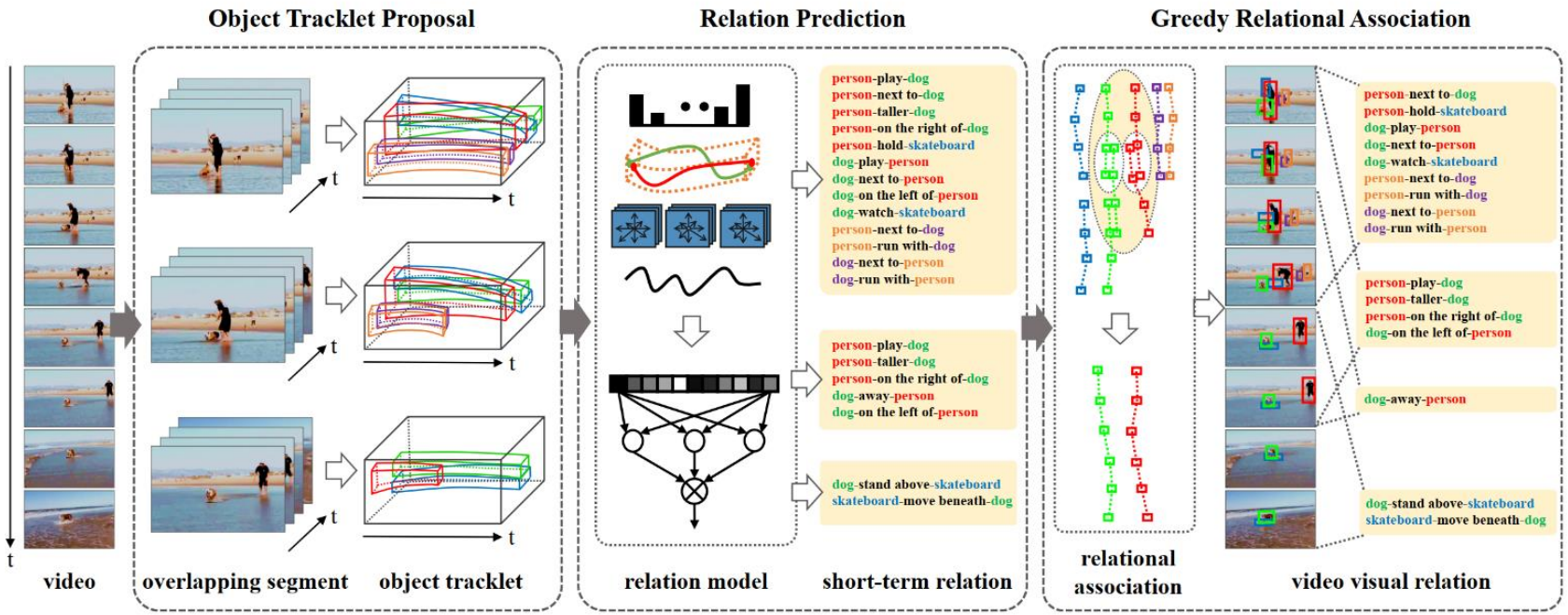
- **在视频中定位物体**：找出视频中出现的所有物体，并得到不同帧中每个相同物体的位置，得到物体轨迹和物体类别

- **识别不同物体之间的交互关系**：包括空间关系和动作关系



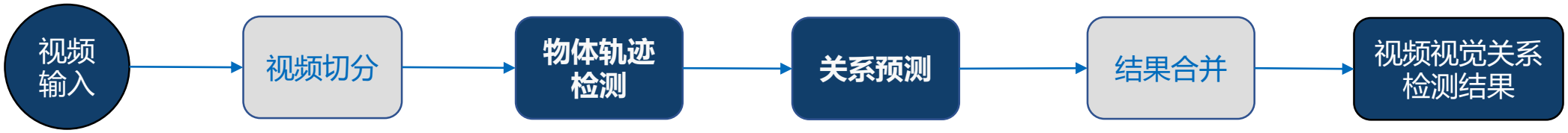


视频关系检测



Xindi Shang, Video Visual Relation Detection.

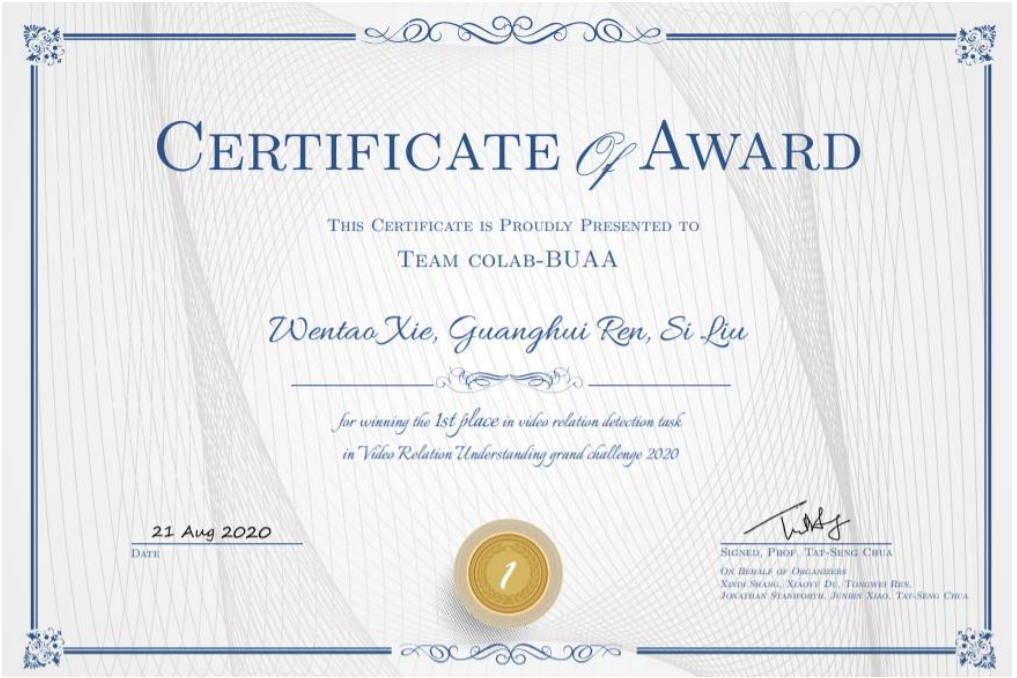
In Proceedings of the 2017 ACM on Multimedia Conference,MM2017



视频关系检测竞赛

Video Object Relation(VidOR) dataset

- **数据集规模**
  1. 10000段互联网短视频
  2. 7000训练集，850验证集，2165测试集
  3. 平均长度30秒
- **标注类别**
  1. 80类物体标注
  2. 50类关系标注（其中8类空间关系，42类动作标注）



Main Task: Video Relation Detection

Rank	Team Name	Performance: mean AP*	Team Members
1	colab-BUAA	0.1174	Beihang University
2	ETRI_DGRC	0.0665	Electronics and Telecommunications Research Institute
3	Zixuan Su	0.0599	Fudan University
4	GKBU	0.0328	Renmin University of China
5	DeepBlueAI	0.0024	DeepBlue Technology (Shanghai) Co., Ltd

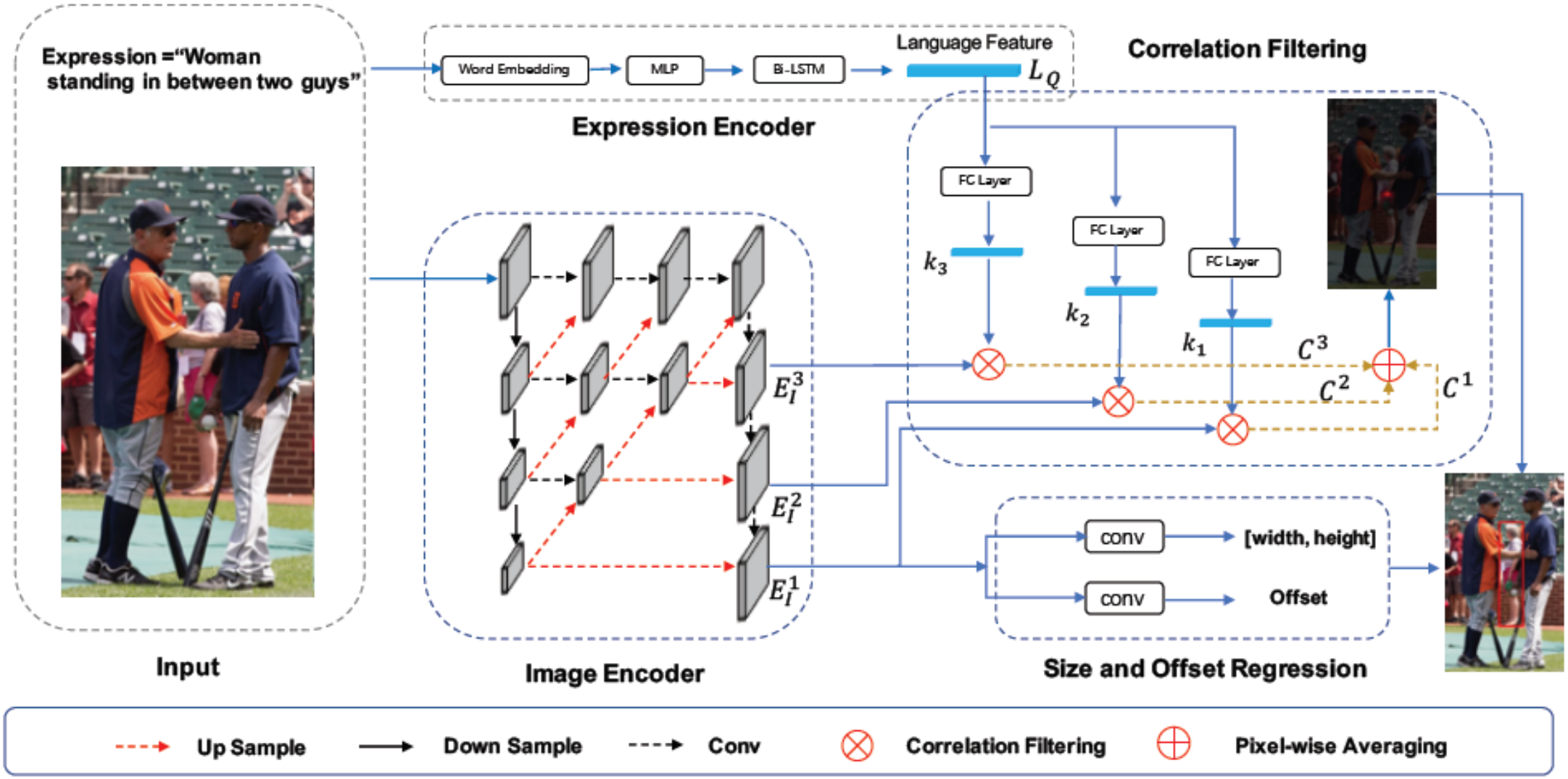
- Visual relation detection
- Referring expression

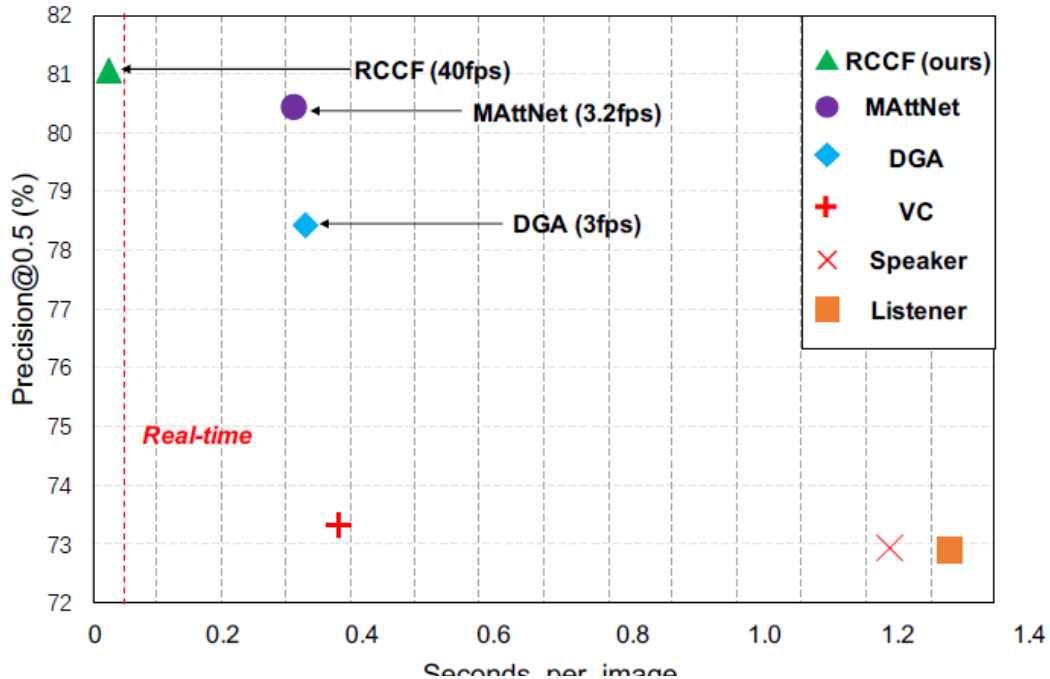
1. Referring expression

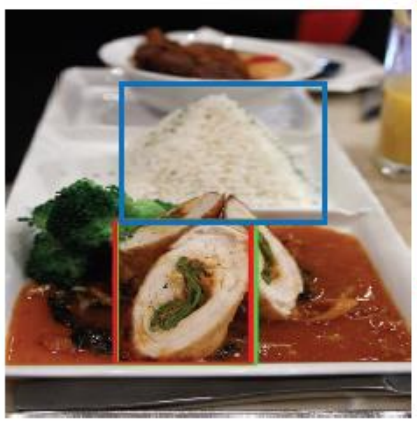
文本: *拿着白色飞盘的红衣男子*

挑战：算法复杂度高，无法实时部署

Yue Liao, **Si Liu** et al, <A Real-Time Cross-modality Correlation Filtering Method for Referring Expression Comprehension > CVPR 2020


Expression ="Woman standing in between two guys"
Word Embedding
MLP
Bi-LSTM
Language Feature
L_Q
Expression Encoder
Correlation Filtering
FC Layer
FC Layer
FC Layer
k_3
k_2
k_1
C^3
C^2
C^1
E_I^3
E_I^2
E_I^1
conv
[width, height]
conv
Offset
Input
Image Encoder
Size and Offset Regression
Up Sample
Down Sample
Conv
Correlation Filtering
Pixel-wise Averaging

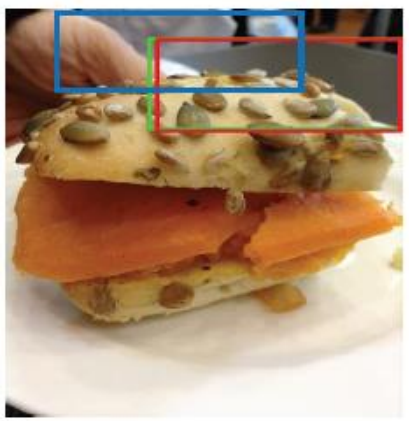




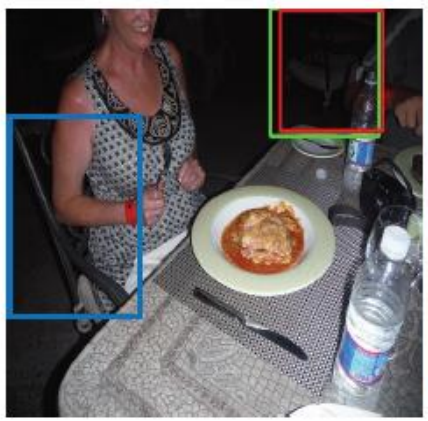
(a)“the middle piece of the chicken rollup”

(b)“man's hand with ring on it”

(c)“ table behind pizza box”



(d)“ The corner of the gray table visible to the right of the hand”



(e)“A steel chair near a lady and back of the man”

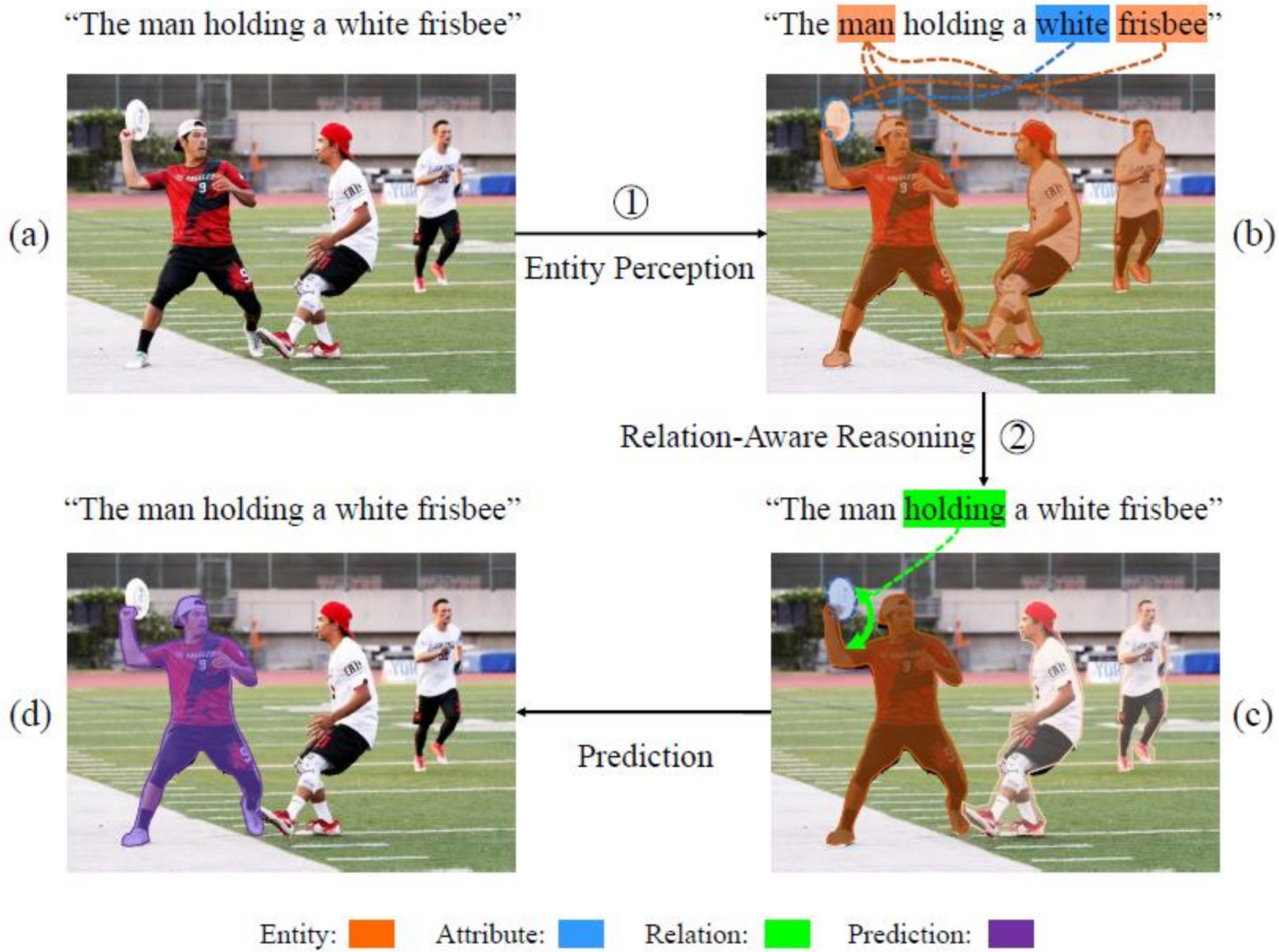
(f)“space between two train cars”

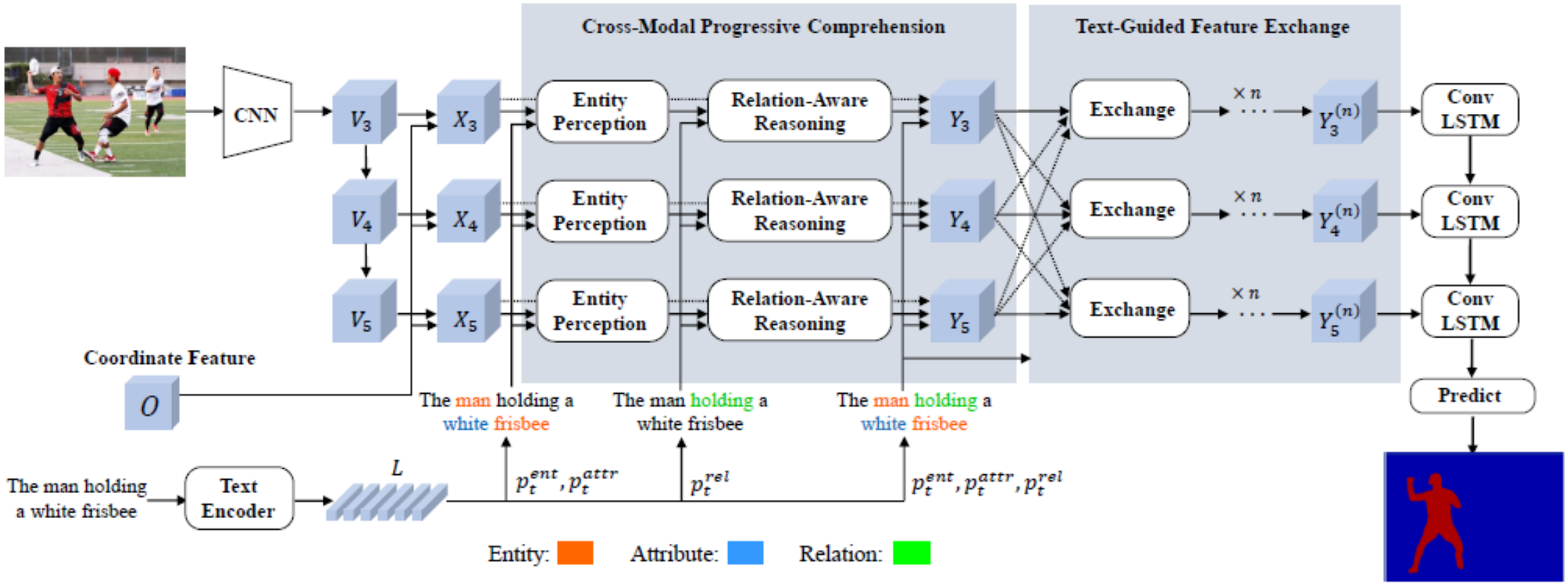
2. Referring segmentation

文本: *拿着白色飞盘的红衣男子*

挑战：无法显式建模实体之间的关系

- Shaofei Huang*, Tianrui Hui*, **Si Liu** et al, <Referring Image Segmentation via Cross-Modal Progressive Comprehension> CVPR 2020 (* equal contribution)
- Tianrui Hui, Si Liu et a, Linguistic Structure Guided Context Modeling for Referring Image Segmentation, ECCV 2020

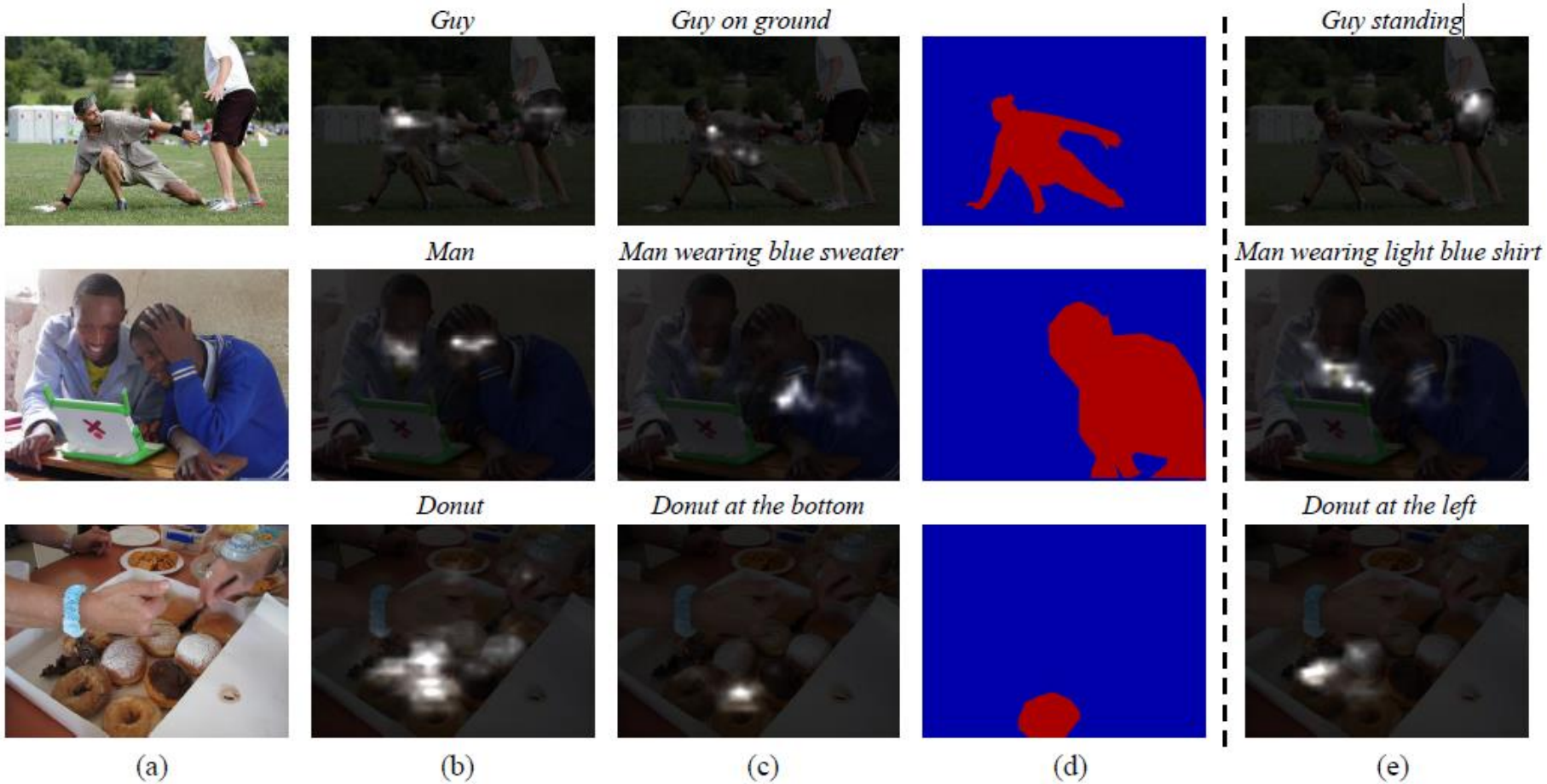

"The man holding a white frisbee"
"The man holding a white frisbee"
(a)
①
Entity Perception
(b)
Relation-Aware Reasoning
②
"The man holding a white frisbee"
"The man holding a white frisbee"
(d)
Prediction
(c)
Entity:
Attribute:
Relation:
Prediction:



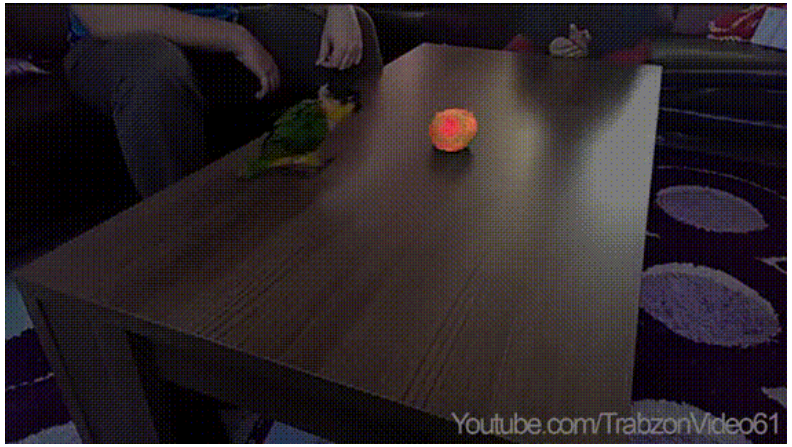
https://github.com/spyflying/CMPC-Refseg

Method	UNC			UNC+			G-Ref	ReferIt
	val	testA	testB	val	testA	testB	val	test
LSTM-CNN [17]	-	-	-	-	-	-	28.14	48.03
RMI [30]	45.18	45.69	45.57	29.86	30.48	29.50	34.52	58.73
DMN [34]	49.78	54.83	45.13	38.88	44.22	32.29	36.76	52.81
KWA [38]	-	-	-	-	-	-	36.92	59.09
ASGN [36]	50.46	51.20	49.27	38.41	39.79	35.97	41.36	60.31
RRN [23]	55.33	57.26	53.95	39.75	42.15	36.11	36.45	63.63
MAttNet [45]	56.51	62.37	51.70	46.67	52.39	40.08	n/a	-
CMSA [44]	58.32	60.61	55.09	43.76	47.60	37.89	39.98	63.80
CAC [8]	58.90	61.77	53.81	-	-	-	44.32	-
STEP [3]	60.04	63.46	57.97	48.19	52.33	40.41	46.40	64.13
Ours	**61.36**	**64.53**	**59.64**	**49.56**	**53.44**	**43.23**	**49.05**	**65.53**

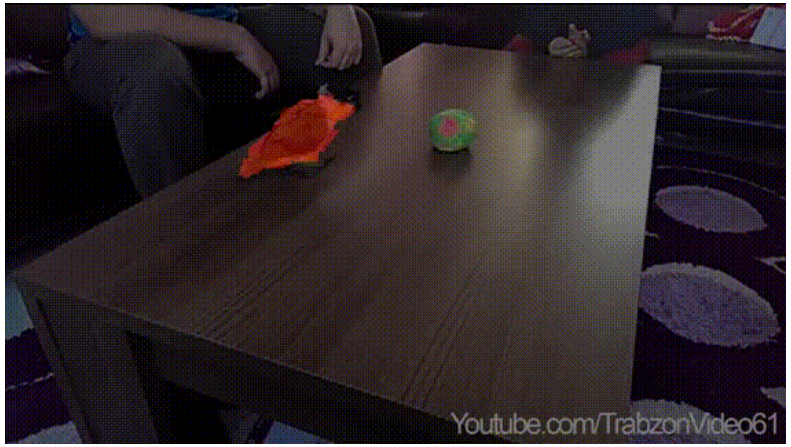
结果展示


Guy
Guy on ground
Guy standing
Man
Man wearing blue sweater
Man wearing light blue shirt
Donut
Donut at the bottom
Donut at the left
(a)
(b)
(c)
(d)
(e)

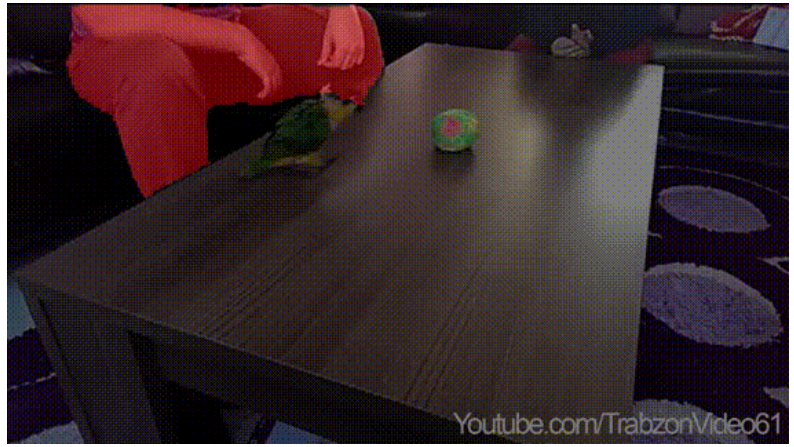
文本：桌上的球



文本：一只玩球的小鸟

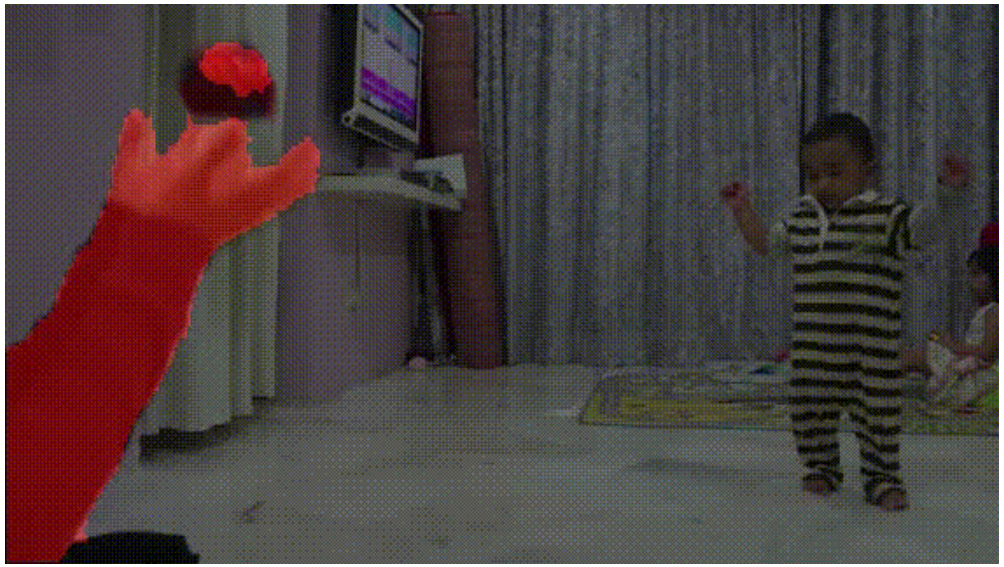


文本：坐在沙发上的人



文本：蹒跚学步的婴儿

文本：用玩具逗婴儿玩的母亲

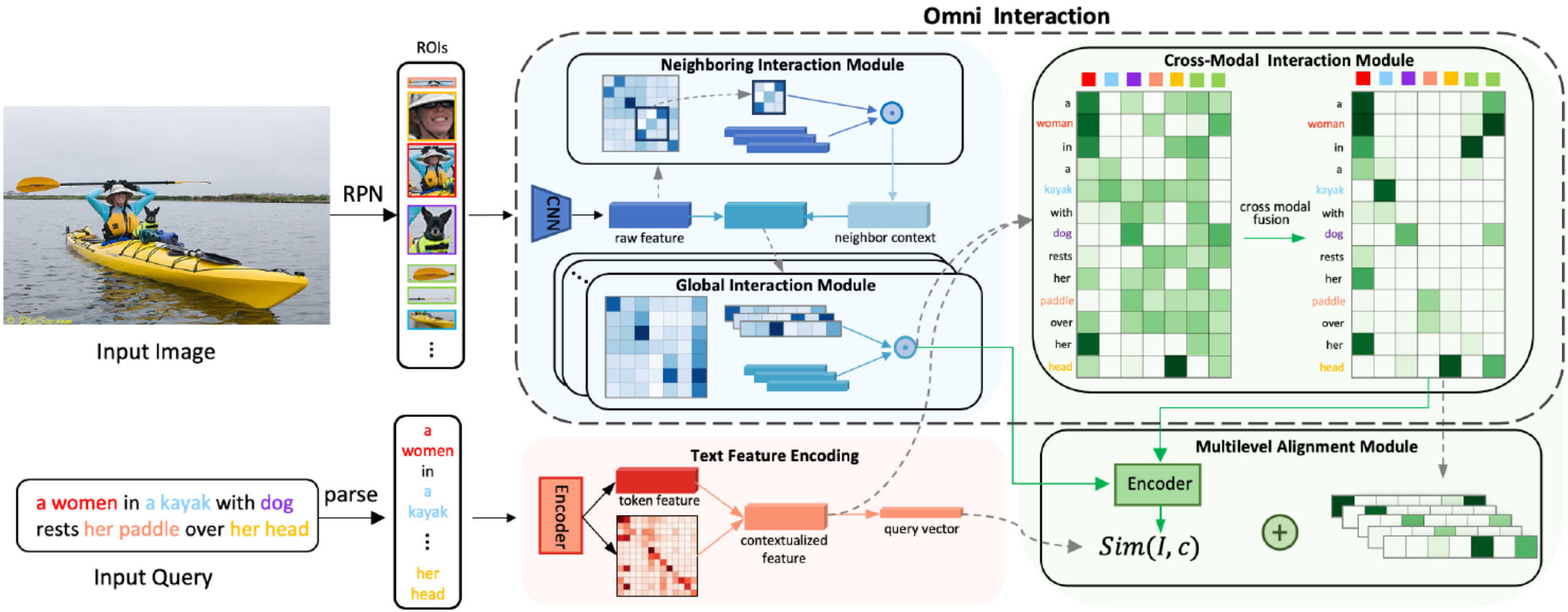


3. Phrase grounding

- Task: Given a natural language specification and an image, find all objects in the image.

Figure 1: Example image from Flickr30K Entities annotated with bounding boxes corresponding to entities in the caption "A man wearing a tan coat signs papers for another man wearing a blue coat."

Tianyu Yu, Tianrui Hui, Zhihao Yu, Yue Liao, Sansi Yu, Faxi Zhang, si liu* , Cross-Modal Omni Interaction Modeling for Phrase Grounding, ACM MM 2020


Omni Interaction
ROIs
RPN
Input Image
a women in a kayak with dog rests her paddle over her head
Input Query
parse
a
women
in
a
kayak
her
head
Neighboring Interaction Module
CNN
raw feature
neighbor context
Global Interaction Module
Cross-Modal Interaction Module
cross modal fusion
a
woman
in
a
kayak
with
dog
rests
her
paddle
over
her
head
Text Feature Encoding
Encoder
token feature
contextualized feature
query vector
Multilevel Alignment Module
Encoder
$Sim(I, c)$

Method	Accu@0.5
Similarity Network [40]	51.05
RPN + QRN[6]	53.48
IGOP [45]	53.97
SPC+PPC [33]	55.49
SS+QRN[6]	55.99
CITE[32]	59.27
SeqGROUND [10]	61.60
G3RAPHGROUND++ [3]	66.93
Visual-BERT [26]	71.33
Contextual Grounding [25]	71.36
COI Net (Ours)	**77.51**

Method	Accu@0.5
SCRC [18]	17.93
MCB + Reg + Spatial [5]	26.54
GroundeR + Spatial [36]	26.93
Similarity Network + Spatial [40]	31.26
CGRE [31]	31.85
MNN + Reg + Spatial [5]	32.21
EB + QRN (VGG_{cls}-SPAT) [6]	32.21
CITE [32]	34.13
IGOP [45]	34.70
QRC Net [6]	44.07
G3RAPHGROUND++ [3]	44.91
COI Net (Ours)	**66.16**

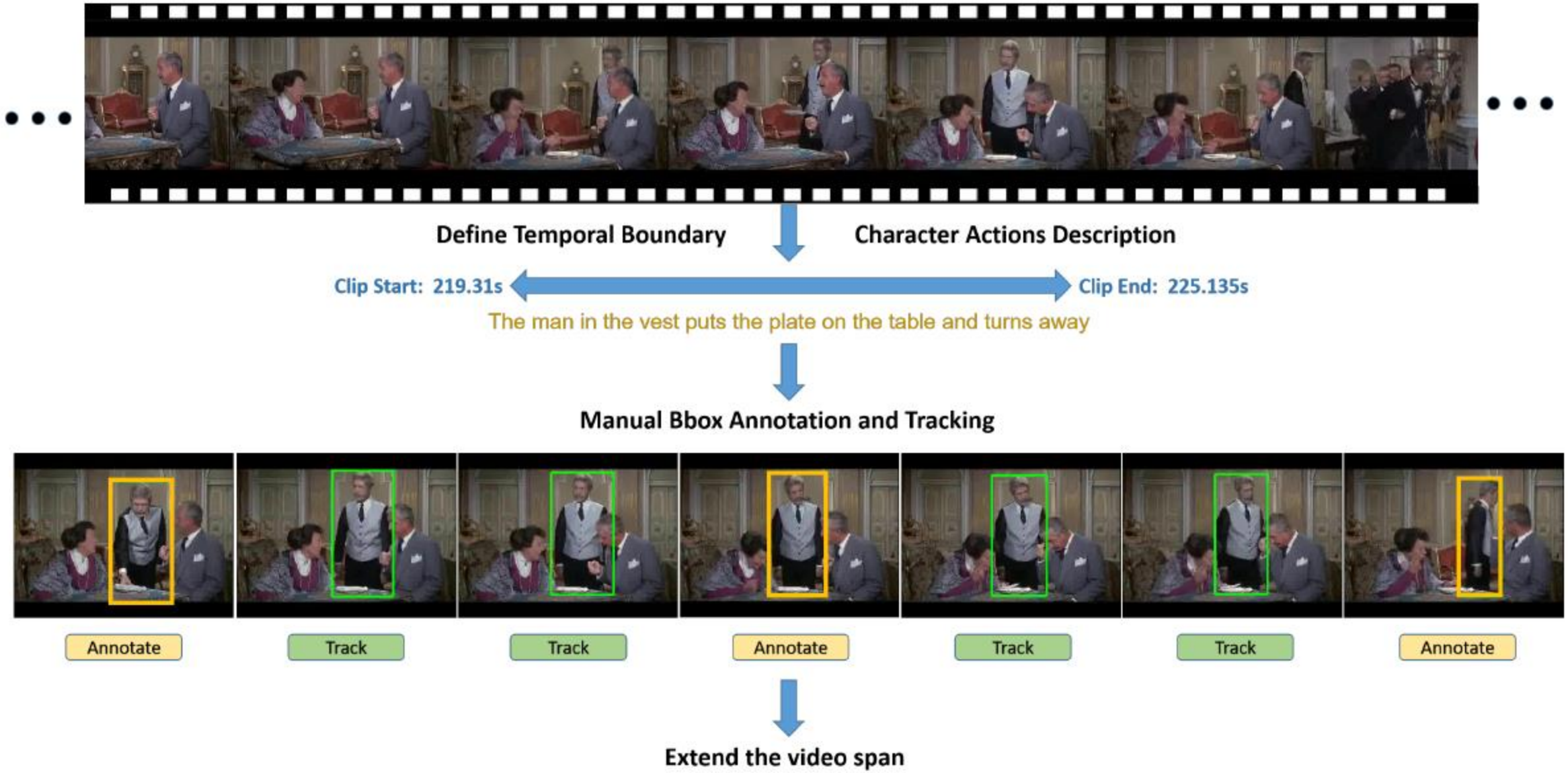
4. Video grounding

Input: untrimmed Video + natural language description

Output: target person bounding boxes + temporal duration

Characteristic: spatial-temporal referring; human centric; complicate expression

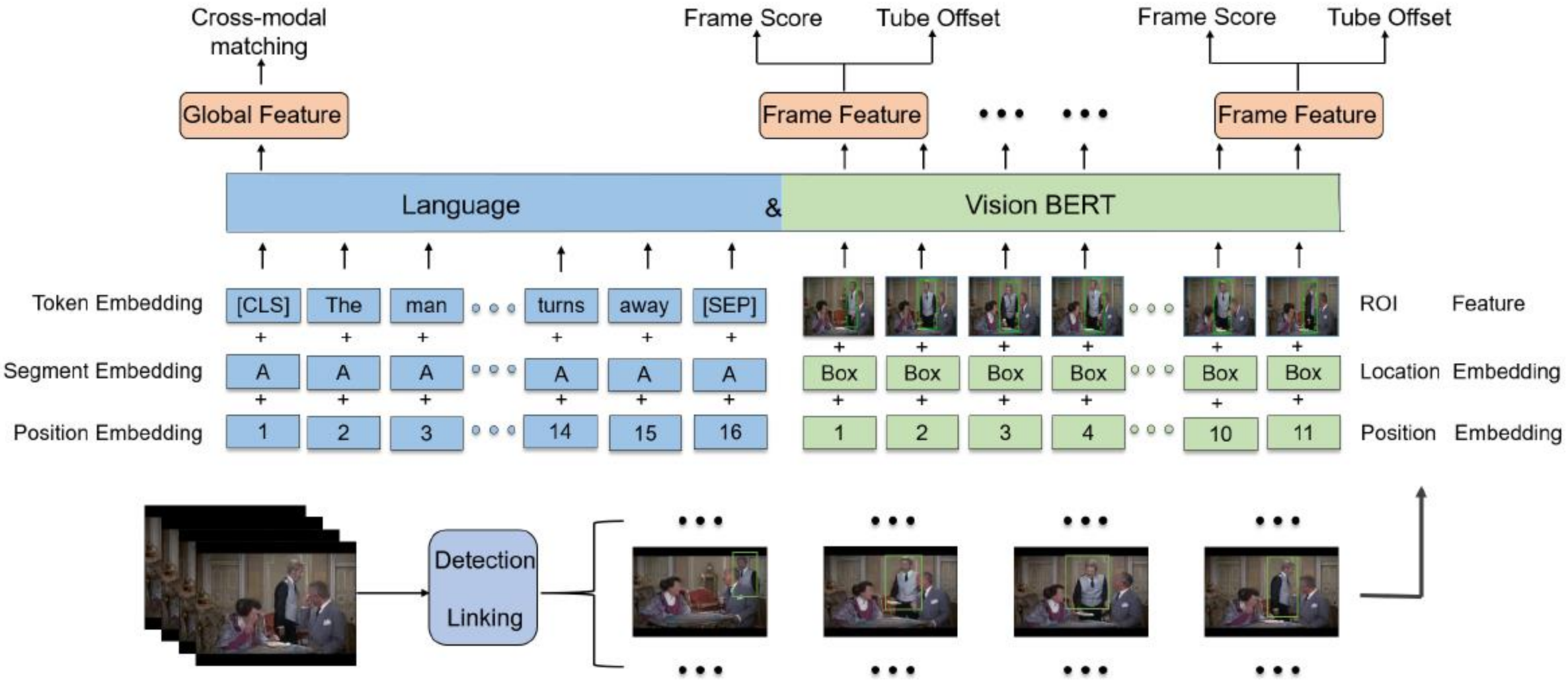
labeling process



语言指导的视频人物检测数据库

Dataset	Queries	Average Words	Spatial Annotation	Temporal Annotation	Human Centric
TACoS	16158	10.5	×	√	√
DiDeMo	41206	8	×	√	×
ActivityNet-C	71492	14.8	×	√	√
CharadesSTA	16124	7.2	×	√	√
VID-sentence	7654	13.2	√	×	×
YouCook2-B	2000	8.8	√	√	×
Our Dataset	16685	**17.25**	**√**	**√**	**√**

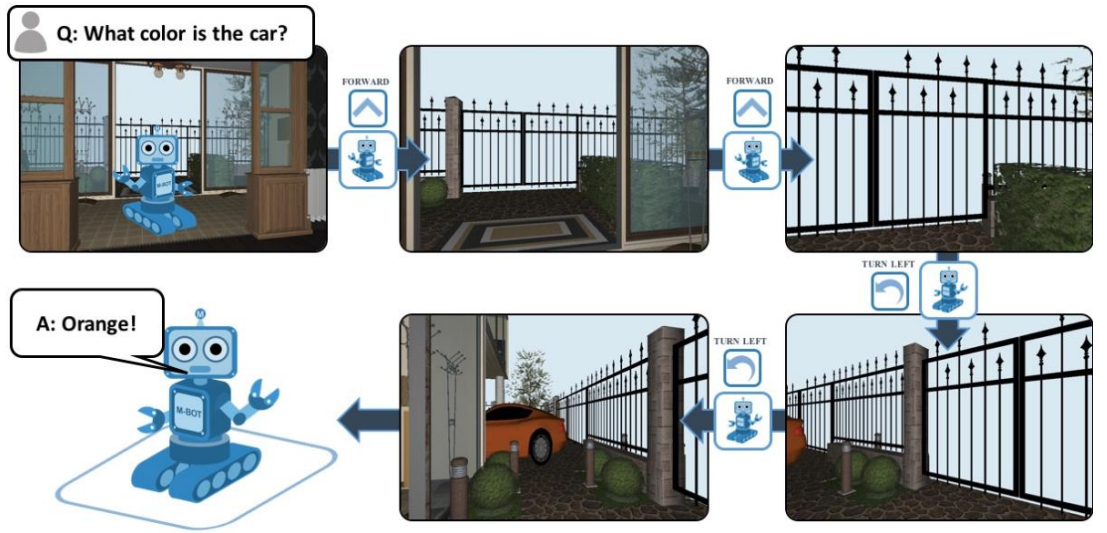
Spatio-Temporal Referring with Visual Transformers



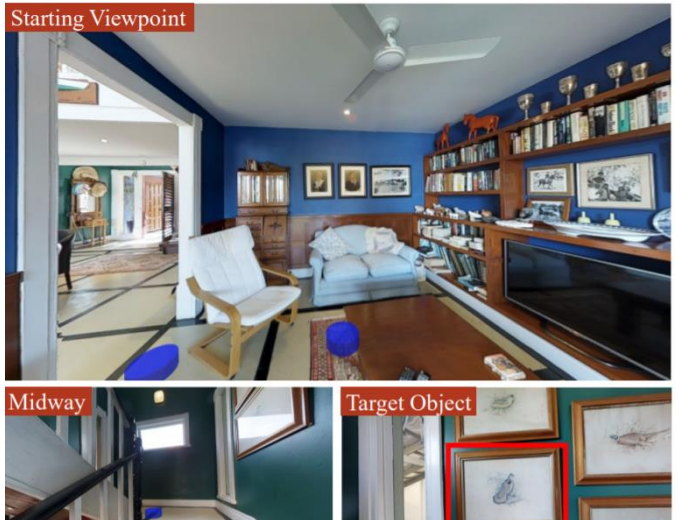
应用场景：短视频、电影搜索； 监控视频搜索

5. REVERIE: Remote Embodied Visual Referring Expression

- 1.通过探索环境，寻找问题答案（embodied QA）



2.通过探索环境，找到指令指代物体的具体位置（remote referring expression）



REVERIE Challenge

ResNet 特征提取

R-CNN
物体检测

匹配度评分 $r_{t,k,i}$ | 匹配度评分 $r_{t,k,i+1}$

多层感知机 | 多层感知机

融合图像特征 $v'^{o}_{t,k,i}$ | 融合图像特征 $v'^{o}_{t,k,i+1}$

跨模态渐进式理解

图像特征 $v^{o}_{t,k,i}$ | 图像特征 $v^{o}_{t,k,i}$ | 语言特征 | 语言特征 | 语言特征 | 语言特征

图像特征提取 | 图像特征提取 | 语言特征提取 | 语言特征提取 | 语言特征提取 | 语言特征提取

二号 病床边 站着的 护士

Top 3

Bi_LSTM

取平均

交互模块

$< v'^{o}_{t,k}, x^{o}_{t,k}, v^{f}_{t,k} >= v^{f,}_{t,k}$

跨模态定位模块

行为预测模块

多层感知机

软性注意力机制 $\hat{v}'_t$

h_{t-1} LSTM h_t

软性注意力机制 $\hat{L}'_t$

行为预测

L

辅助推理监督