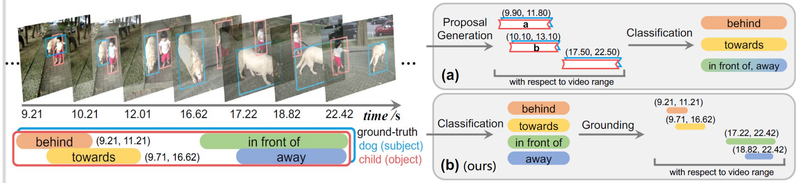
Classification-Then-Grounding: Reformulating Video Scene Graphs as Temporal Bipartite Graphs

Kaifeng Gao[1], Long Chen[2], Yulei Niu[2], Jian Shao[1], Jun Xiao[1]
[1]Zhejiang University [2]Columbia University

arXiv GitHub

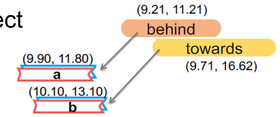
CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Detecting Video Visual Relations



Previous Work *subj-obj pair proposal → relation classification*

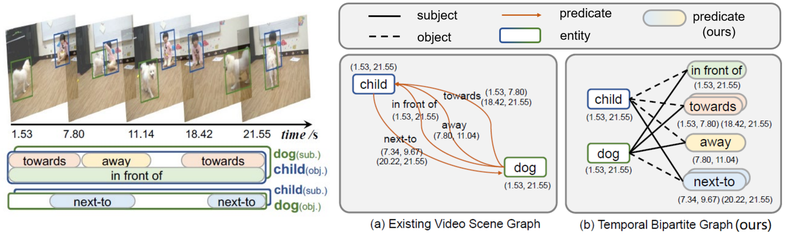
- Label assignment is partially correct
- Discard relation context
- Upper-bounded by proposals



Ours *relation classification → temporal grounding*

- A new classification-then-grounding framework
- A novel **BI**partite **G**raph based model BIG
- Reformulate video scene graphs as temporal bipartite graphs

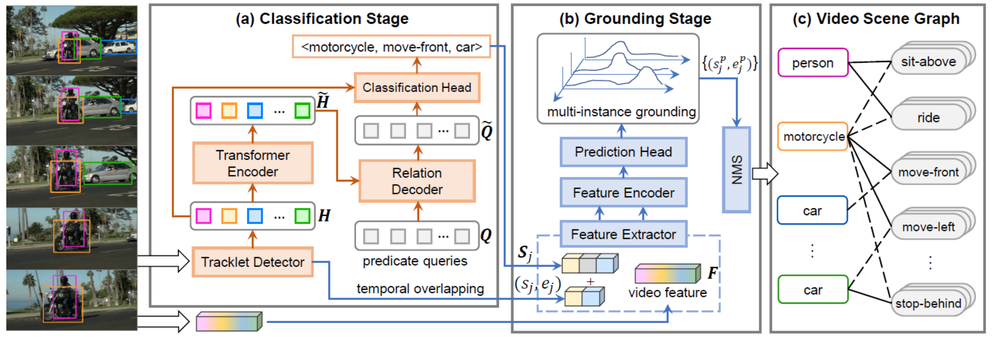
Temporal Bipartite Graphs



Advantages

- it avoids enumerating all entity pairs for predicate (relation) prediction
- it is easier to model entity pairs with multiple predicates
- it can be easily extended to more general relations with more semantic roles

Classification-Then-Grounding



Classification Stage

- A Transformer-based model
- Classifiy the categories of all entity & predicate nodes
- Learn the edges of the bipartite graph

R-norm	F_*	RelDet (%) mAP	RelDet (%) R@50	RelTag (%) P@1	RelTag (%) P@5
	✓	7.98	7.71	61.65	51.10
✓		8.02	7.36	61.65	51.68
✓	✓	8.29	7.92	64.42	51.70

Table 5. Ablations of BIG-C for the R-norm and F_* of RaCA module on VidOR.

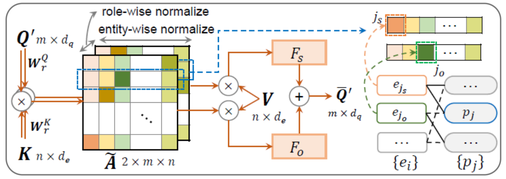
Grounding Stage

- Localize the temporal location of each predicate node
- Take triplet category as language query

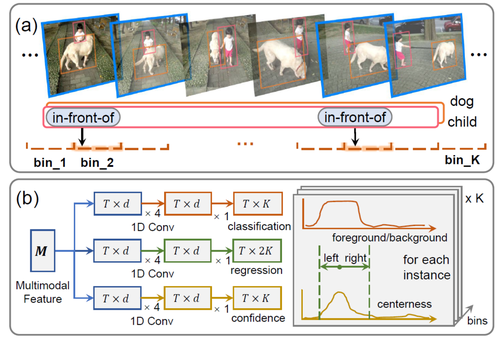
#Bins	fR$_S$@K (%) 50	100	150	fR$_M$@K (%) 50	100	150
1	12.96	15.59	16.76	5.53	6.86	7.46
5	13.07	15.83	17.26	5.75	7.20	8.05
10	13.04	15.89	17.61	5.75	7.30	8.25

Table 6. Ablations for multi-instance grounding with different number of bins on VidOR.

Role-aware Cross-Attention



Multi-instance Grounding



Experiments Results

➢ Compare with SOTA on VidOR dataset

Models	Detector	Features Visual	Lang	Motion	Mask	RelDet mAP	RelDet R@50	RelDet R@100	RelTag P@1	RelTag P@5	RelTag P@10
Liu *et al.* [24]$_{CVPR'20}$	RefineDet	RoI+I3D$_r$		✓		6.85	8.21	9.90	51.20	40.73	—
Chen *et al.* [13]$_{ICCV'21}$	Faster R-CNN	RoI+I3D$_r$	✓	✓		10.04	8.94	10.69	61.52	50.05	38.48
Chen *et al.* [13]$_{ICCV'21}$	Faster R-CNN	RoI+I3D$_r$	✓	✓	✓	11.21	9.99	11.94	68.86	55.16	43.40
IVRD [22]$_{MM'21}$	Faster R-CNN	RoI		✓		7.42	7.36	9.41	53.40	42.70	—
Chen *et al.* [13]$_{ICCV'21}$	Faster R-CNN	RoI		✓		**8.93**	7.38	9.22	56.89	44.76	34.07
VidVRD-II [34]$_{MM'21}$	Faster R-CNN	RoI		✓		8.65	**8.59**	**10.69**	57.40	44.54	33.30
BIG-C (Ours)	MEGA	RoI				8.03	7.60	9.39	**62.25**	50.96	40.30
BIG (Ours)	MEGA	RoI+I3D$_f$				8.28	7.74	9.82	62.13	**51.25**	**40.48**
VRU^5 19-top1 [37]$_{MM'19}$	FGFA	—	✓	✓		6.56	6.89	8.83	51.20	40.73	—
MHA [36]$_{MM'20}$	FGFA	—	✓	✓		6.59	6.35	8.05	50.72	41.56	—
VRU^5 20-top1 [46]$_{MM'20}$	CascadeRCNN	RoI	✓	✓	✓	**9.93**	**9.12**	—	**67.43**	—	—
Chen *et al.* [13]$_{ICCV'21}$	Faster R-CNN	RoI	✓	✓		9.54	8.49	**10.17**	59.24	47.24	35.99
BIG-C (Ours)	MEGA	RoI	✓			8.29	7.92	9.65	64.42	51.70	**41.05**
BIG (Ours)	MEGA	RoI+I3D$_f$	✓			8.54	8.03	10.04	64.42	**51.80**	40.96

Table 3. Performance (%) on VidOR of SOTA models. The **Best** and second best are marked in according formats. **Visual**: I3D$_r$ and I3D$_f$ denote region-level and frame-level I3D features, respectively. **Lang**: The word embeddings of entity categories. **Motion**: It refers to the relative motion feature of entity pairs [34]. **Mask**: It means the localization mask of entities [46].

➢ Compare with SOTA on VidVRD dataset

Models	Features Visual	Features Motion	RelDet mAP	RelDet R@50	RelDet R@100	RelTag P@1	RelTag P@5	RelTag P@10
VidVRD [35]$_{MM'17}$	iDT	✓	8.58	5.54	6.37	43.00	28.90	20.80
GSTEG [40]$_{CVPR'19}$	iDT	✓	9.52	7.05	8.67	51.50	39.50	28.23
VRD-GCN [30]$_{MM'19}$	iDT	✓	16.26	8.07	9.33	57.50	41.00	28.5 0
MHA [36]$_{MM'20}$	iDT	✓	19.03	9.53	10.38	57.50	41.40	29.45
IVRD [22]$_{MM'21}$	RoI	✓	22.97	12.40	14.46	68.83	49.87	35.57
VidVRD-II [34]$_{MM'21}$	RoI	✓	29.37	19.63	22.92	70.40	53.88	40.16
Liu *et al.* [24]$_{CVPR'20}$	RoI+I3D†	✓	18.38	11.21	13.69	60.00	43.10	32.24
Chen *et al.* [13]$_{ICCV'21}$	RoI+I3D	✓	20.08	13.73	16.88	62.50	49.20	38.45
Liu *et al.* [24]$_{CVPR'20}$	RoI†		14.01	8.47	**11.00**	56.50	36.70	26.60
TRACE [39]$_{ICCV'21}$	RoI		15.06	7.67	10.32	—	—	—
BIG-C (Ours)	RoI†		**17.56**	**9.59**	10.92	**56.50**	**44.30**	**32.35**
Liu *et al.* [24]$_{CVPR'20}$	RoI+I3D†		14.81	9.14	**11.39**	55.50	38.90	28.90
TRACE [39]$_{ICCV'21}$	RoI+I3D		17.57	9.08	11.15	**61.00**	**45.30**	**33.50**
BIG-C (Ours)	RoI+I3D†		**17.67**	**9.63**	11.29	56.00	43.80	32.85
BIG-C (Ours)	RoI‡		**26.08**	**14.10**	**16.25**	**73.00**	**55.10**	**40.00**

Bipartite Graph Visualization