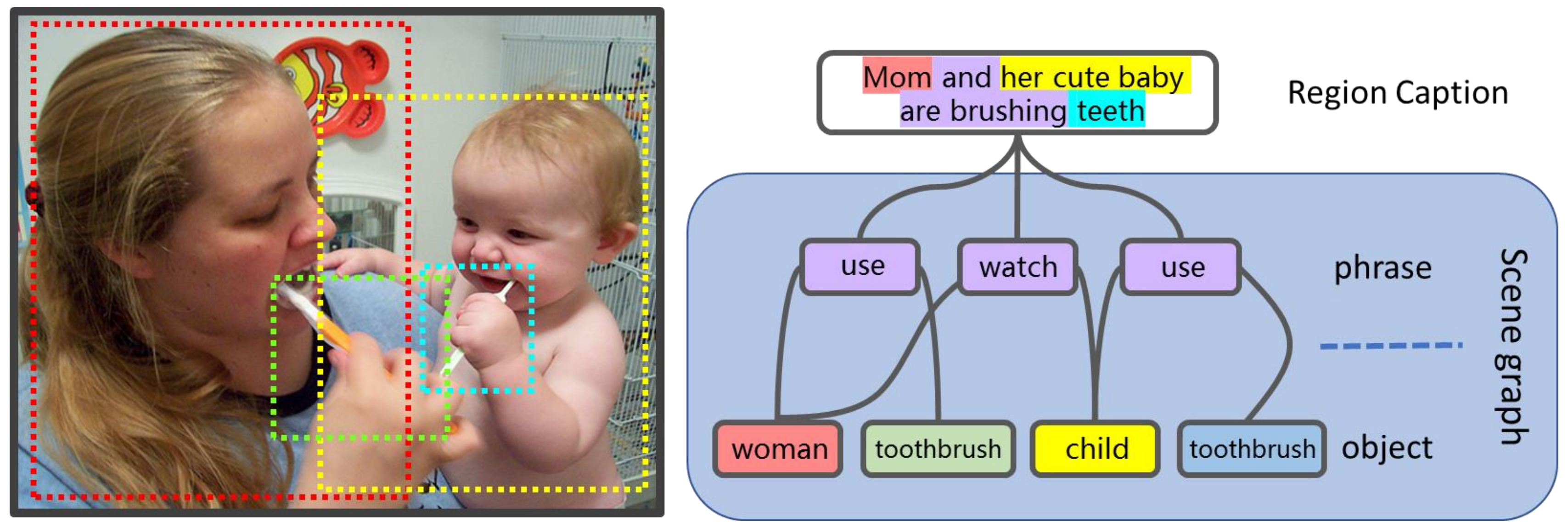
香 港 中 文 大 學
The Chinese University of Hong Kong

Scene Graph Generation from Objects, Phrases and Region Captions

Yikang Li[1], Wanli Ouyang[1, 2], Bolei Zhou[3], Kun Wang[1], Xiaogang Wang[1]
[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]University of Sydney, Australia, [3]Massachusetts Institute of Technology, USA

ICCV 17
International Conference on Computer Vision 2017

Motivation



- Scene understanding includes numerous vision tasks of different semantic levels. These tasks are tightly connected and describe the image from different aspects.
- Object detection localizes the objects and recognize their categories, targeting on the details of the image.
- Scene graph generation involves pair-wise relationships between objects, describing the image with a series of <subject-predicate-object> phrases.
- Image/region captioning generates a free-form sentence with an uncertain number of the objects, their attributes, and their interactions

Contribution

Multi-level Scene Description Network~(MSDN) model is proposed to leverage complementary effects of features at different semantic levels.

- MSDN could simultaneously detects objects, recognizes their relationships and generates captions for salient image regions.
- A graph is dynamically constructed to establish the links among regions with different semantic meaning.
- A feature refining structure is used to pass message from different semantic levels through the graph.

Quantitative Results

Comparison with existing works:

- LP: Visual Relationship detection using word embeddings as language prior (Lu, Cewu, et al., ECCV 2016)
- ISGG: Scene graph generation using iterative message passing (Xu, Danfei, et al. arXiv:1701.02426)

Task		LP [23]	ISGG [33]	Ours
PredCls	R@50	26.67	58.17	**67.03**
	R@100	33.32	62.74	**71.01**
PhrCls	R@50	10.11	18.77	**24.34**
	R@100	12.64	20.23	**26.50**
SGGen	R@50	0.08	7.09	**10.72**
	R@100	0.14	9.91	**14.22**

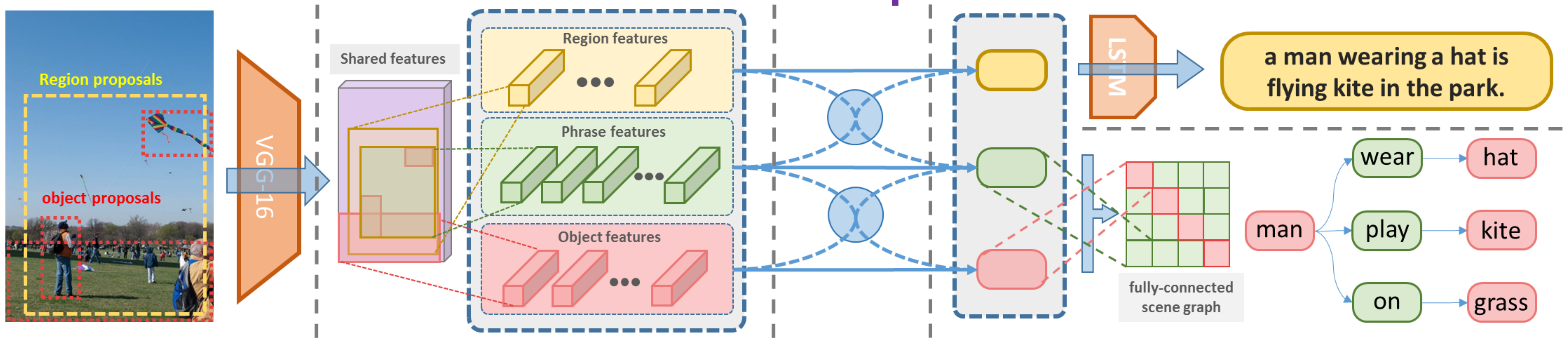
Component analysis:

ID	M.P.	Cap.B.	Cap.S.	FR-iters	PredCls		PhrCls		SGGen	
					Rec@50	Rec@100	Rec@50	Rec@100	Rec@50	Rec@100
1	-	-	-	0	49.28	52.69	7.31	10.48	2.39	3.82
2	✓	-	-	1	63.12	66.41	19.30	21.82	7.73	10.51
3	✓	✓	-	1	63.82	67.23	20.91	23.09	8.20	11.35
4	✓	✓	✓	1	66.70	**71.02**	23.42	25.68	10.23	13.89
5	✓	✓	✓	2	**67.03**	71.01	**24.22**	**26.50**	**10.72**	**14.22**
6	✓	✓	✓	3	66.23	70.43	23.16	25.28	10.01	13.62

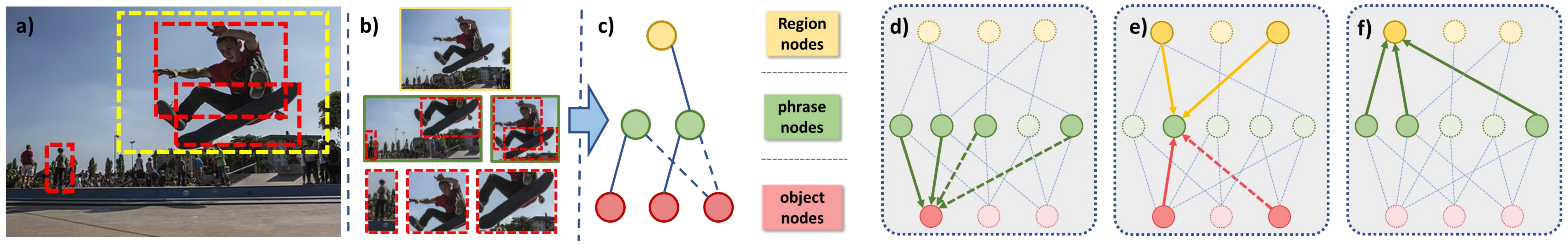
- M.P.: whether to use message passing
- Cap.B.: whether to use caption branch
- Cap.S.: whether to use language model
- FR-iters: feature refining iterations

- PredCls: predicate recognition
- PhrCls: phrase recognition task
- SGGen: scene graph generation

Multi-level Scene Description Network



Overview of MSDN. The two RPNs for objects and captions are omitted for simplicity, which share the convolutional layers with other parts. Phrase regions are generated by grouping object regions. With the region proposals for objects, phrases, and captions, ROI-pooling is used for obtaining their features. These features go through two fully connected layers and then pass messages to each other. After message passing, features for objects are used for object detection, similarly for phrase detection and region captioning. Message passing is guided by the dynamic graph constructed from the object and caption region proposals. Features, bounding boxes and predicted labels for object (red), phrase (green) and region (yellow) are assigned with different colors.



Dynamic Scene Graph Generation

- Connections between of constructing phrase proposals.
- Each phrase proposal will be connected to its corresponding two object proposals as a <subject-predicate-objecteen> phrase and objects are naturally built based on our triplet.
- When the caption region proposal covers 0.7 of the phrase proposal, there is an undirected edge between the two proposals.
- Since phrases are used as the intermediate semantic level connecting objects and caption regions, connections between caption regions and objects are omitted.

Feature Refining

- Feature refining procedure is divided into three parallel steps, object refining, phrase refining and region refining. Without loss of generality, we only show the refining of object features, which can be applied to phrase and caption region features as well.
- **Phrase feature merge:** Since the features from different phrases have different importance factors for refining objects, we use a gate function to determine weights.

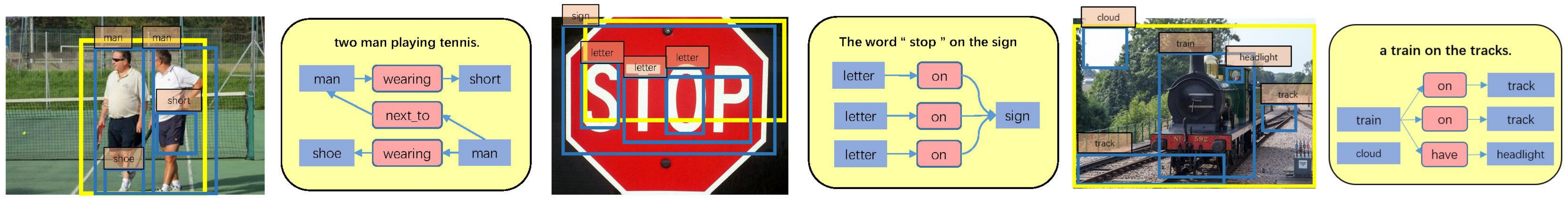
$$\tilde{\boldsymbol{x}}_i^{(p\to s)} = \frac{1}{\|\boldsymbol{E}_{i,p}\|} \sum_{(i,j)\in \boldsymbol{E}_{s,p}} \sigma_{\langle o,p\rangle}\left(\boldsymbol{x}_i^{(o)}, \boldsymbol{x}_j^{(p)}\right) \boldsymbol{x}_j^{(p)}$$

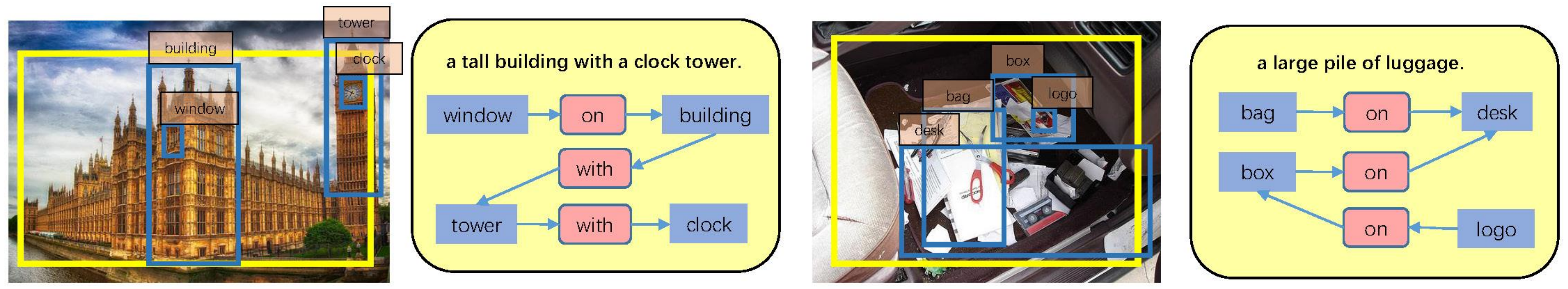
The gate function is defined as:

$$\boldsymbol{x}_{i,t+1}^{(o)} = \boldsymbol{x}_{i,t}^{(o)} + \boldsymbol{F}^{(p\to s)}\left(\tilde{\boldsymbol{x}}_i^{(p\to s)}\right) + \boldsymbol{F}^{(p\to o)}\left(\tilde{\boldsymbol{x}}_i^{(p\to o)}\right)$$

- **Refine object features**: For the *i*-th object, there are two merged features:

$$\sigma_{\langle o,p\rangle}\left(\boldsymbol{x}_i^{(o)}, \boldsymbol{x}_j^{(p)}\right) = \sum_{g=1}^{G} \mathrm{sigmoid}\left(\boldsymbol{w}_{\langle o,p\rangle}^{(g)} \cdot \left[\boldsymbol{x}_i^{(o)}, \boldsymbol{x}_j^{(p)}\right]\right),$$





Github Repo: MSDN

Yikang's homepage