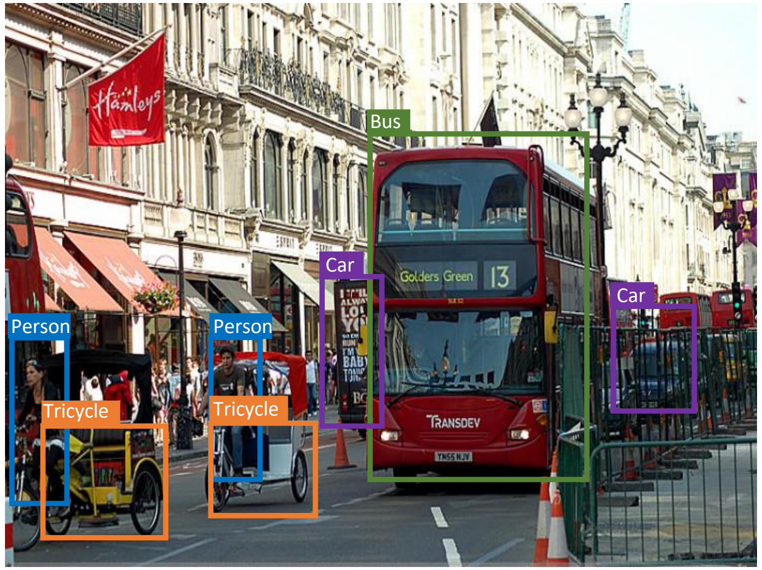
Object Detection Using Dual Graph Network

Shengjia Chen[1], Zhixin Li[1,*], Feicheng Huang[1], Canlong Zhang[1], Huifang Ma[1,2]

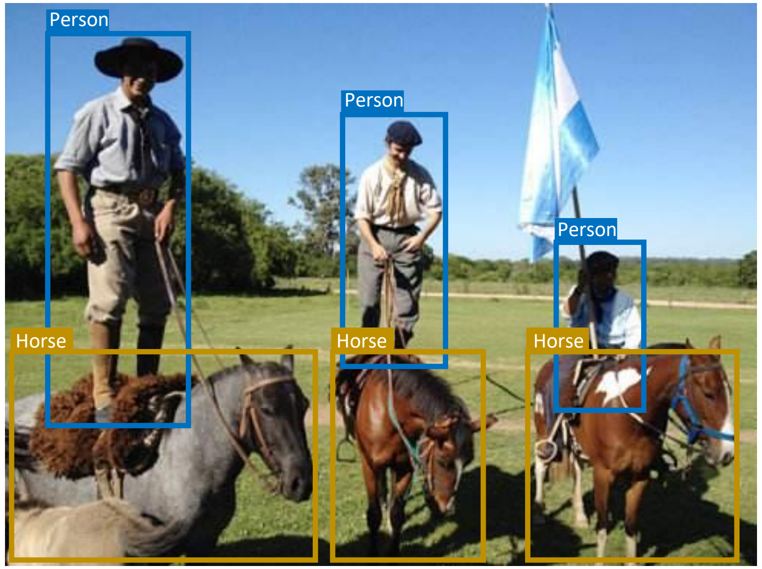
[1]Guangxi Key Lab of Multi-source Information Mining and Security,
Guangxi Normal University, Guilin 541004, China
[2]College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China

Motivation

- Deteriorated quality of feature in the propagation process of the neural network
- Traditional detectors utilize information within one region proposal
- Hard for traditional detectors to identify a small object



(a)



(b)

Motivation

- Prevalent detectors only focus on local information near an object's region of interest within the image. Usually an image contains rich **spatial relation** information including ***context*** and ***object relationships***.

- Previous detectors ignore the **semantic relation** information including ***global correlations*** and ***important dependencies*** between labels which require to be inferred from knowledge beyond a single image.

Ignoring these information inevitably places constraints on the accuracy of objects detected. Therefore, we study the following problem:

*How to capture more **semantic relation** and **spatial relation** information during training?*

Motivation

- Prevalent detectors only focus on local information near an object's region of interest within the image. Usually an image contains rich **spatial relation** information including ***context*** and ***object relationships***.

- Previous detectors ignore the **semantic relation** information including ***global correlations*** and ***important dependencies*** between labels which require to be inferred from knowledge beyond a single image.

Ignoring these information inevitably places constraints on the accuracy of objects detected. Therefore, we study the following problem:

*How to capture more **semantic relation** and **spatial relation** information during training?*

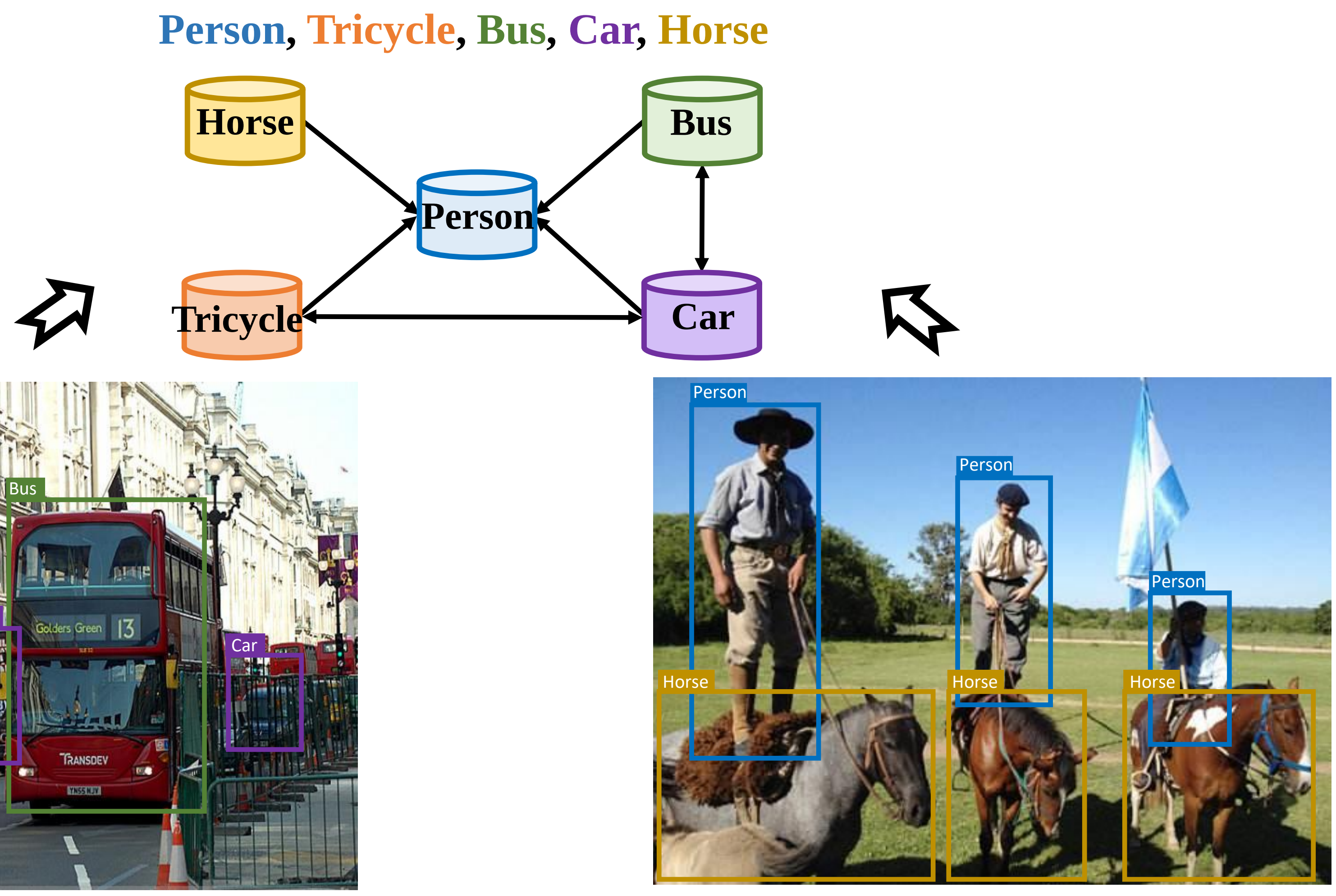
Motivation

- Prevalent detectors only focus on local information near an object's region of interest within the image. Usually an image contains rich **spatial relation** information including ***context*** and ***object relationships***.
- Previous detectors ignore the **semantic relation** information including ***global correlations*** and ***important dependencies*** between labels which require to be inferred from knowledge beyond a single image.

Ignoring these information inevitably places constraints on the accuracy of objects detected. Therefore, we study the following problem:

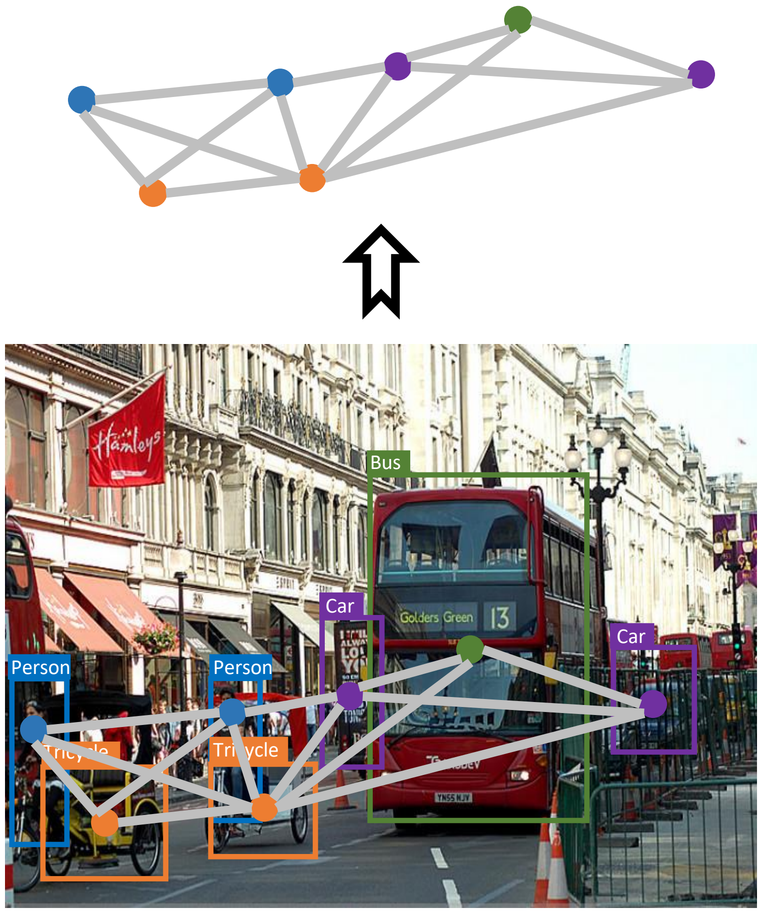
*How to capture **semantic relation** and **spatial relation** information during training?*

Global Semantic Relation

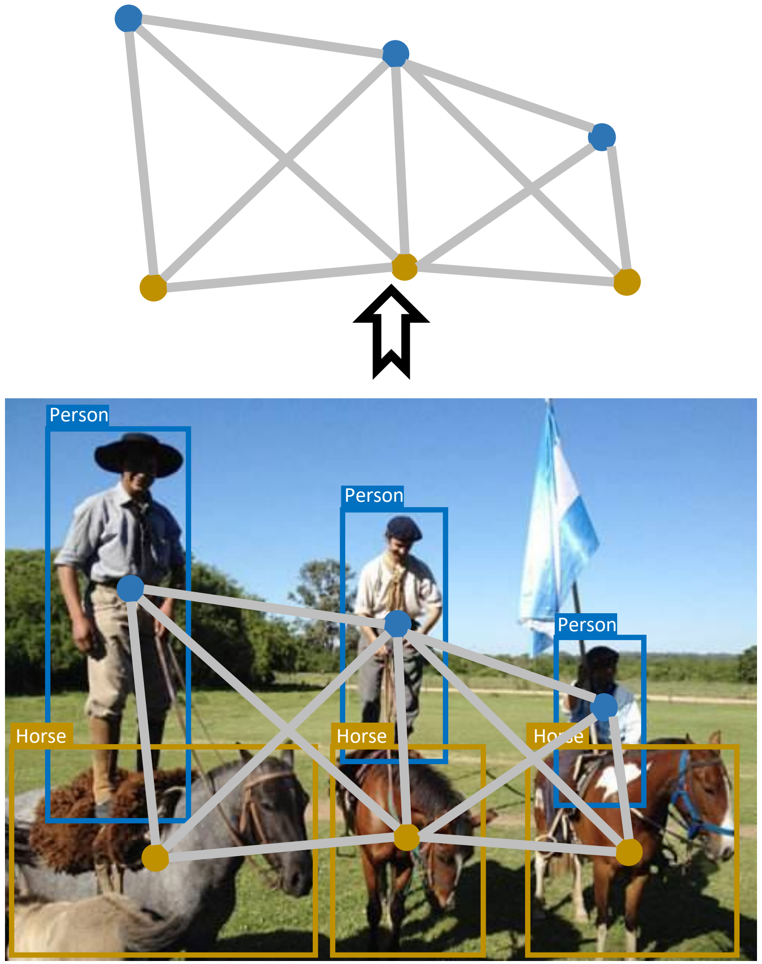


(a) (b)

Local Spatial Relation



(a)

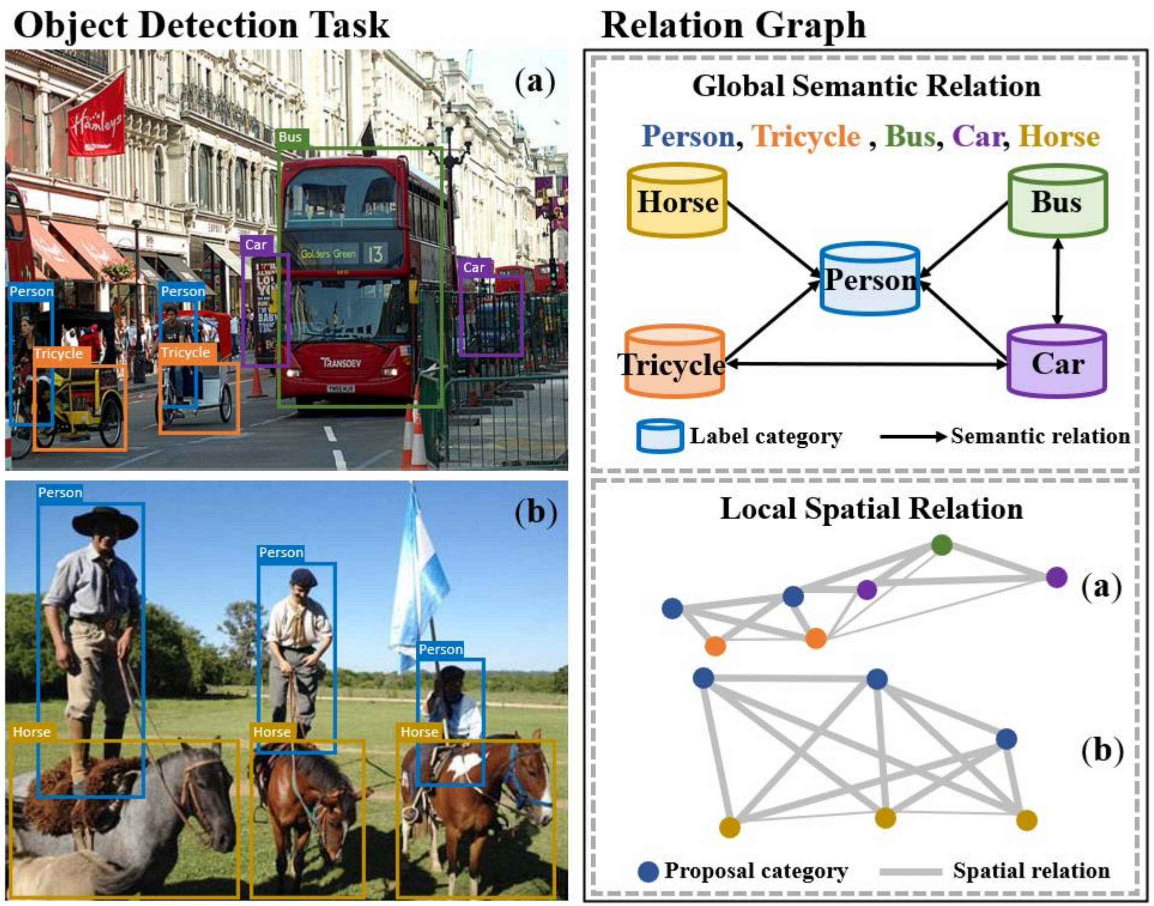


(b)

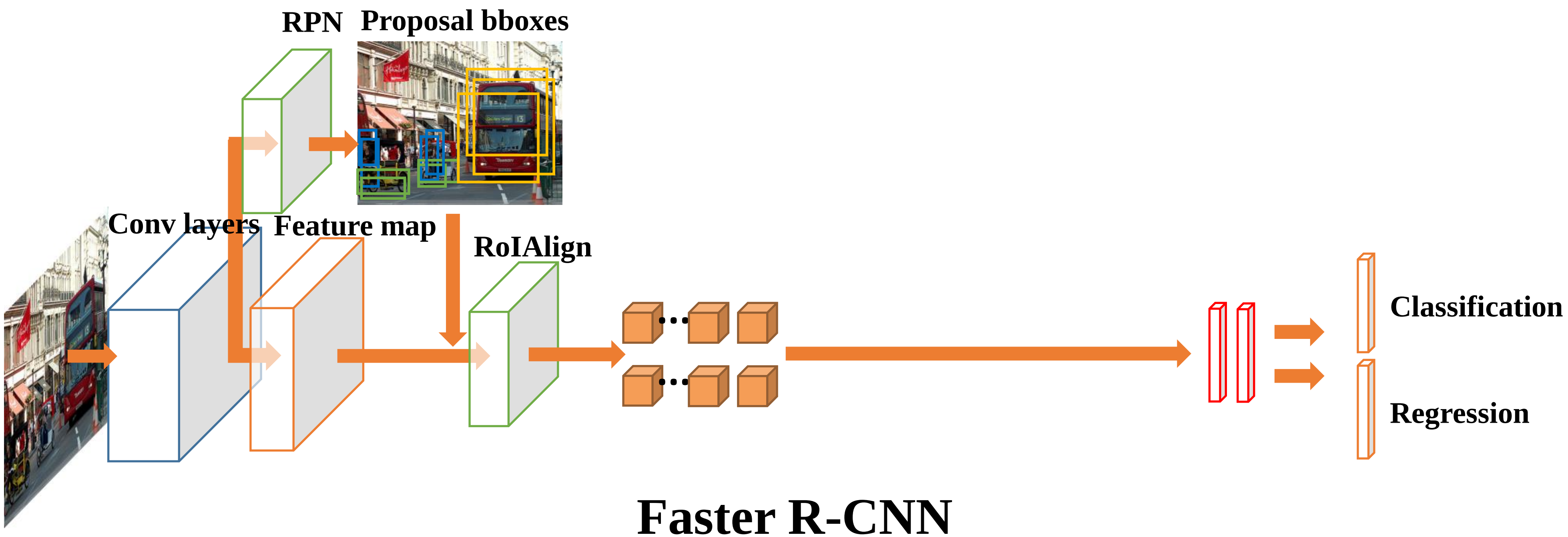
Contributions

- *Causes of Constraints on the Accuracy*
- ◻ Ignoring **glocal semantic relation** information
- ◻ Ignoring **local spatial relation** information
- ◻ Hard for traditional detector to identify a small object

- *Our Solution: Dual Graph Network*
- ✓ capture **global semantic relation** information
- ✓ capture **local spatial relation** information
- ✓ The ability to detect small objects can be significantly improved



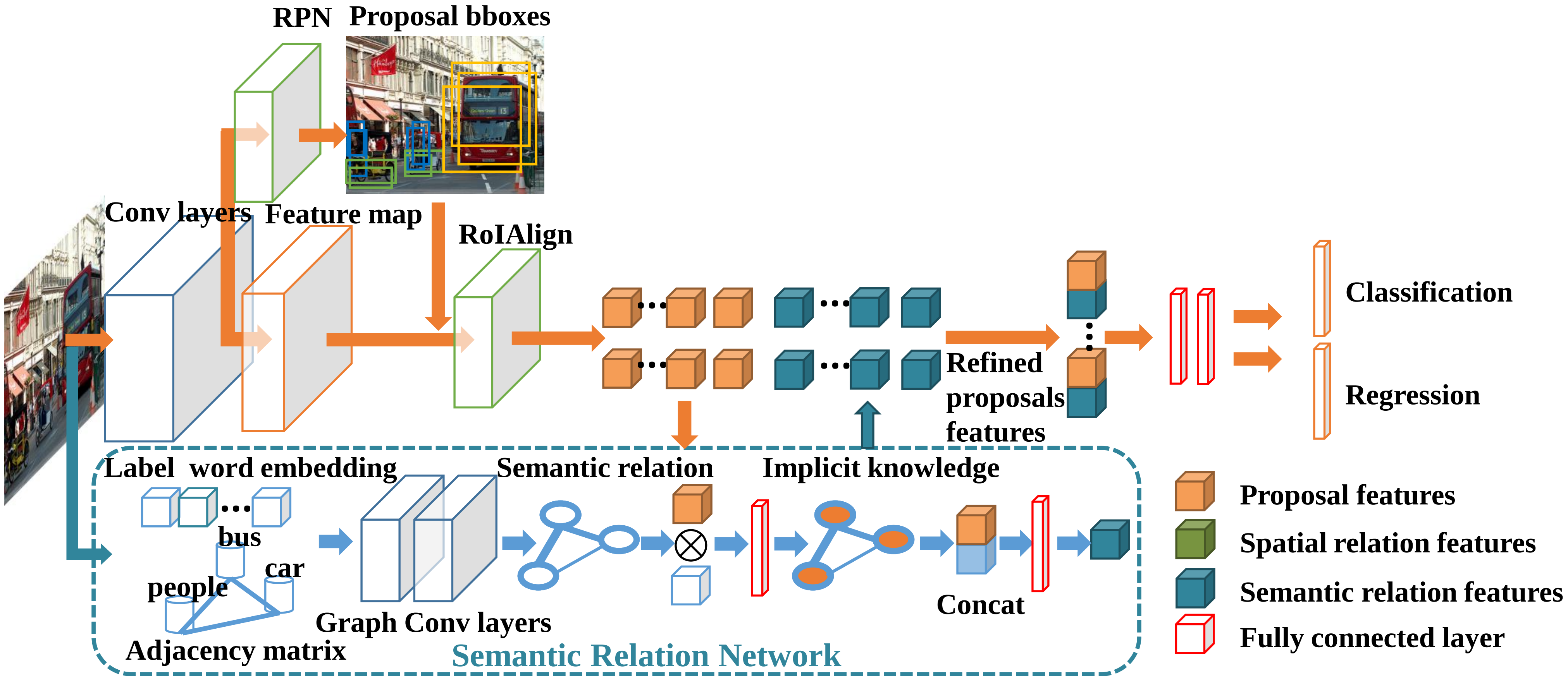
Baseline



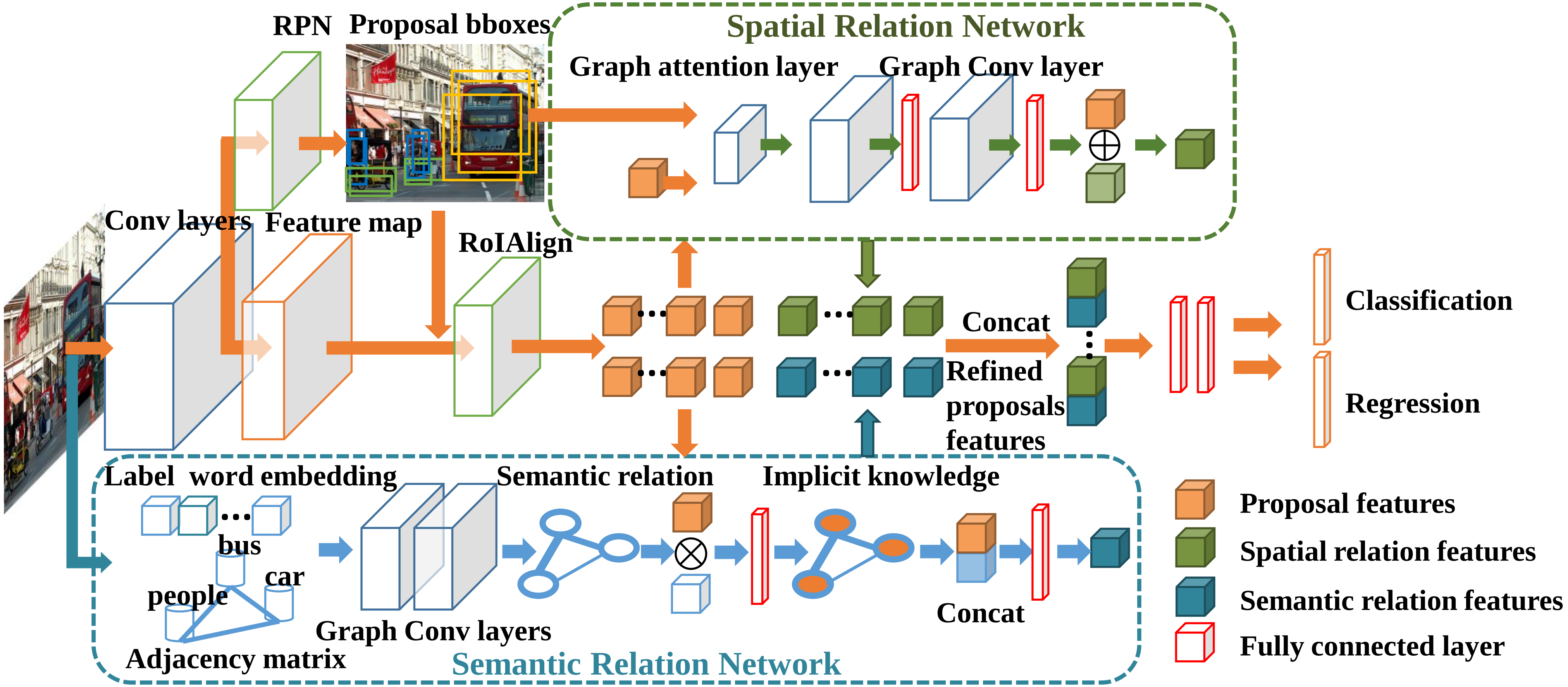
Faster R-CNN

- Traditional detectors focus only on the information around **one region proposal**
- They only propagate the **visual features** of the objects in the network
- Ignoring the key relation in **labels** and **images**
- Hard for these detectors to identify a **small object**

Relation R-CNN



Relation R-CNN



Quantitative Results on VOC

Method	Backbone	Data	Input resolution	mAP
General Detector				
Faster R-CNN [3](Baseline)	VGG16	07+12	600×1000	73.2
Fast R-CNN [2]	VGG16	07+12	600×1000	70.0
NOC [26]	VGG16	07+12	600×1000	73.3
SSD [4]	VGG16	07+12	321×321	75.1
RON384 [17]	VGG16	07+12	384×384	75.4
Relation Information				
KG-CNet [21]	VGG16	07	600×1000	66.6
SMN [10]	VGG16	07	600×1000	70.0
ACCNN [19]	VGG16	07+12	600×1000	72.0
ION [6]	VGG16	07+12	600×1000	75.6
SIN [11]	VGG16	07+12	600×1000	76.0
Relation R-CNN(Ours)	VGG16	07+12	600×1000	**76.6**

Method	Backbone	Data	Input resolution	mAP
General Detector				
Faster R-CNN [3](Baseline)	ResNet101	07+12	600×1000	76.4
SSD321 [4]	ResNet101	07+12	321×321	77.1
DSOD300 [27]	DenseNet	07+12	300×300	77.7
YOLOv2 [5]	DarkNet	07+12	544×544	78.6
CenterNet [18]	ResNet101	07+12	512×512	78.7
Relation Information				
GBDNet [20]	Inception v2	07+12	600×1000	77.2
HKRM [22]	ResNet101	07+12	600×1000	78.8
Relation R-CNN(Ours)	ResNet101	07+12	600×1000	**78.9**

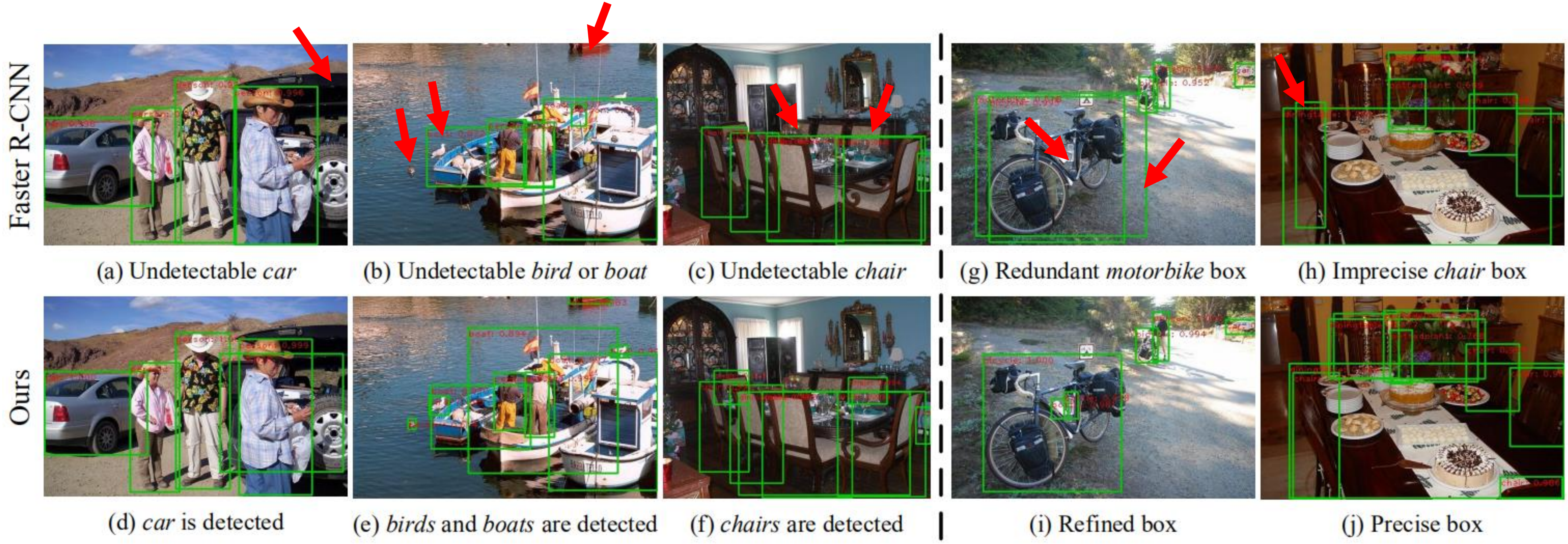
Quantitative Results on MS COCO

Method	Backbone	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}
General Detector							
Faster R-CNN [3](Baseline)	ResNet101	34.7	54.7	37.2	14.8	39.4	**51.8**
YOLOv2 [5])	DarkNet	33.0	**57.9**	34.4	18.3	35.4	41.9
TripleNet [7])	ResNet50	35.9	57.8	38.0	17.7	37.2	50.7
Relation Information							
ION [6])	VGG16	23.0	42.0	23.0	6.0	23.8	37.3
SIN [11]	VGG16	23.2	44.5	22.0	7.3	24.5	36.3
KG-CNet [21]	VGG16	24.4	-	-	-	-	-
GBDNet [20]	Inception v2	27.0	45.8	-	-	-	-
SMN [10]	ResNet101	31.6	52.2	33.2	14.4	35.7	45.8
Relation Network [9]	ResNet101	35.4	56.1	38.5	-	-	-
Relation R-CNN(Ours)	ResNet101	**36.2**	56.9	**39.3**	**19.5**	**41.2**	49.1

The ability to detect small objects can be significantly improved !

Qualitative results

More objects are detected: small, occluded, and indistinct !



(a) Undetectable *car* (b) Undetectable *bird* or *boat* (c) Undetectable *chair* (g) Redundant *motorbike* box (h) Imprecise *chair* box

(d) *car* is detected (e) *birds* and *boats* are detected (f) *chairs* are detected (i) Refined box (j) Precise box

(a) Semantic Relation Network (b) Spatial Relation Network

More precise bounding box !

Conclusion

- **Relation R-CNN**
  - ✓ The semantic relation network is proposed to capture the **global semantic relations** in labels. the detector can find more objects, and the ability to detect **small objects** and **occluded objects** can be significantly improved
  - ✓ The spatial relation network is proposed to capture the **local spatial relations** between objects in images. It can make the **detection box** more accurate and reasonable
  - ✓ Relation R-CNN has more advantages, better robustness, and better generalization ability than other advanced methods. This makes the detector more consistent with **human visual perception**

Thanks for watching !