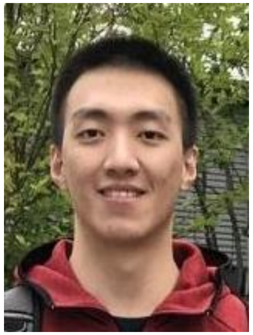
Human-centric Visual Relation Segmentation Using Mask R-CNN and VTransE



Fan Yu Xin Tan Tongwei Ren Gangshan Wu

State Key Laboratory for Novel Software Technology, Nanjing University

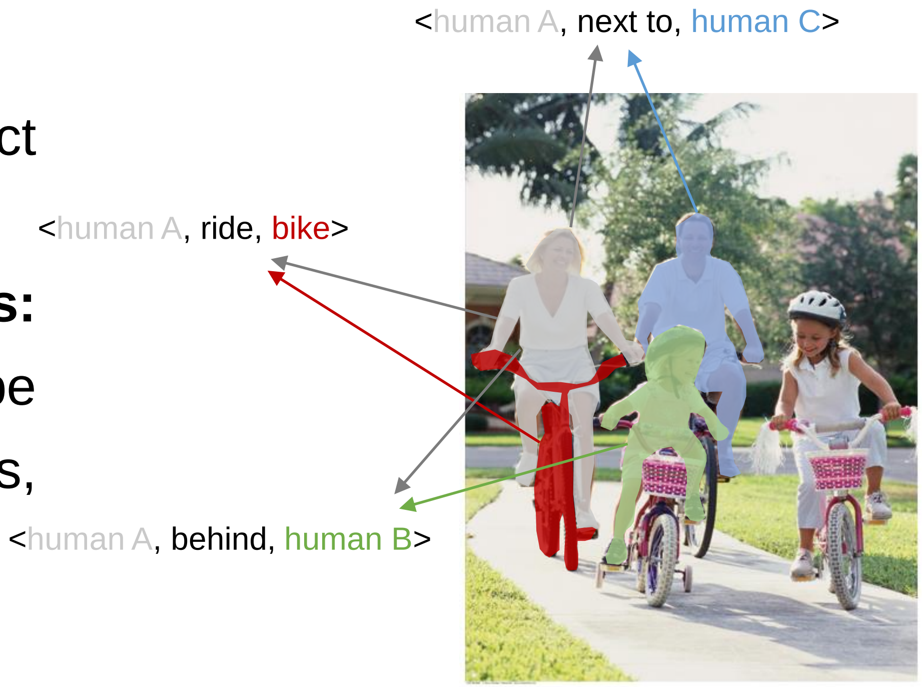
南京大學
NANJING UNIVERSITY

MAGUS
Multimedia AnalyzinG
and UnderStanding

Introduction

- **Human-centric visual relation segmentation:** estimate human-object relations and human-human relations in the form of <human, predicate, object> or <human A, predicate, human B>

- **Focus on human:** the subject in the triple must be human
- **Need to estimate the masks:** subjects and objects can be represented with their shapes, not just the bounding box

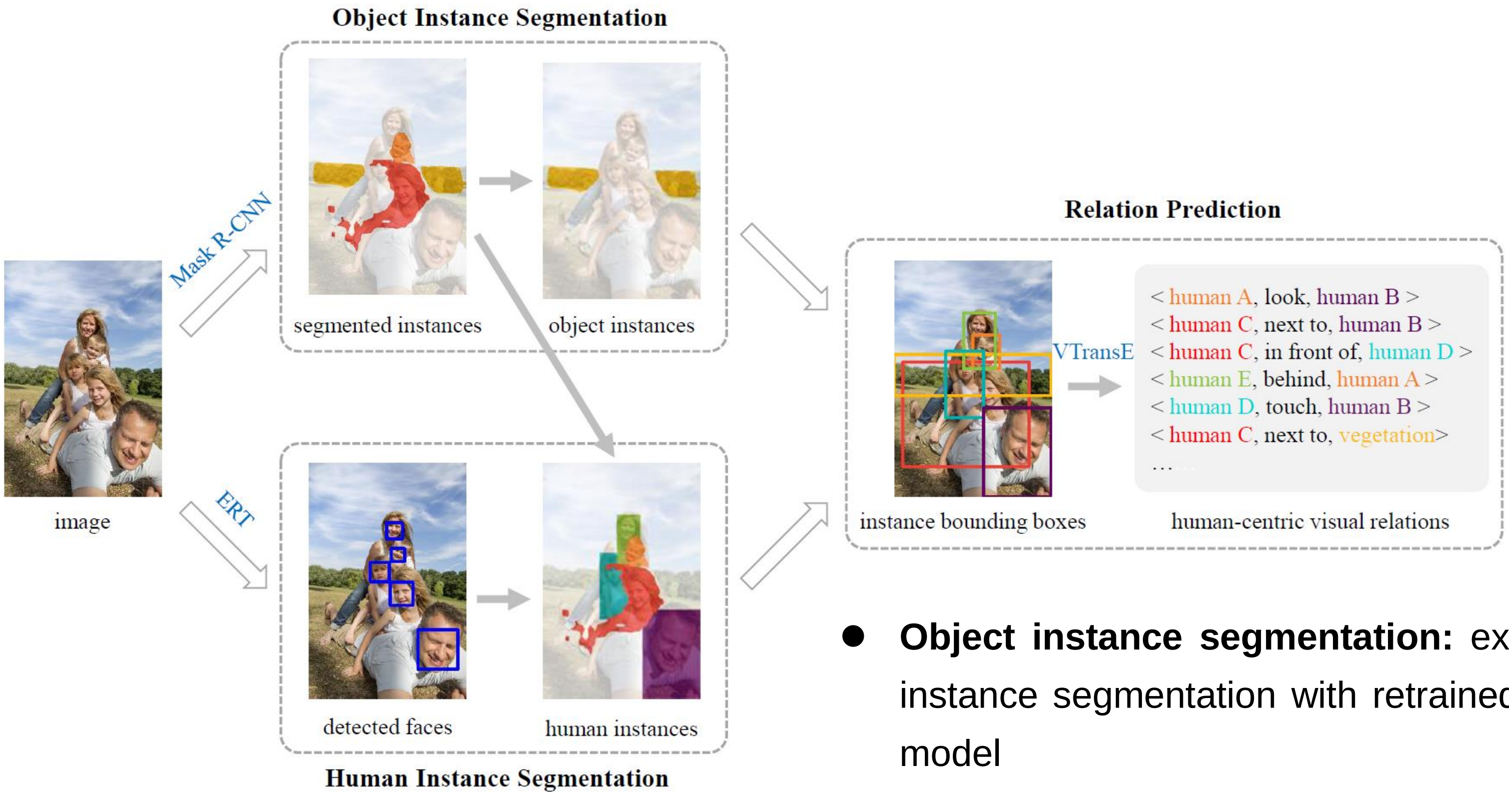


<human A, next to, human C>

<human A, ride, bike>

<human A, behind, human B>

Our Method

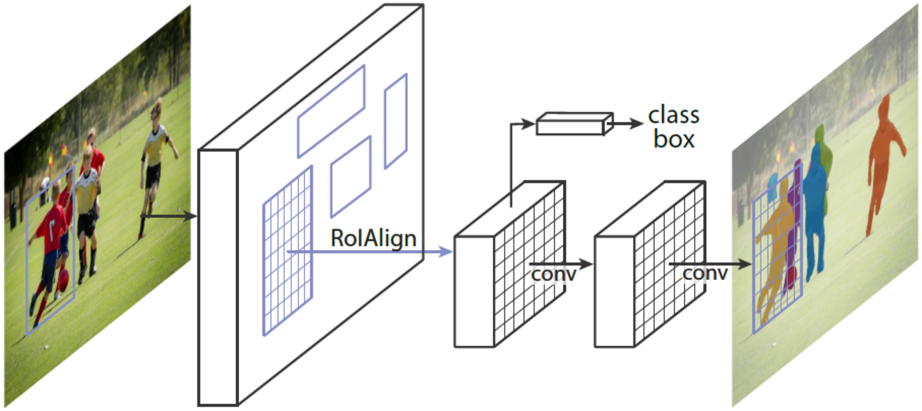


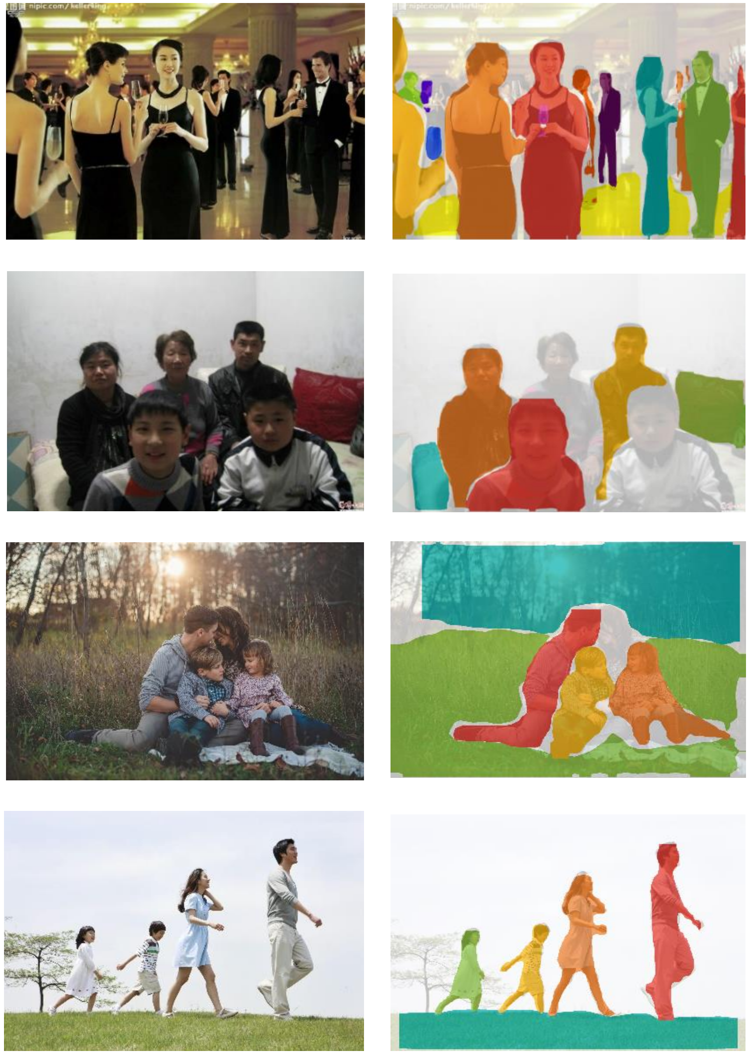
Object Instance Segmentation
Mask R-CNN
segmented instances
object instances
image
ERT
detected faces
human instances
Human Instance Segmentation
Relation Prediction
VTransE
< human A, look, human B >
< human C, next to, human B >
< human C, in front of, human D >
< human E, behind, human A >
< human D, touch, human B >
< human C, next to, vegetation>
… …
instance bounding boxes
human-centric visual relations

- **Object instance segmentation:** export the object instance segmentation with retrained Mask R-CNN model
- **Human instance segmentation:** combine the human instance segmentation exported by Mask R-CNN and face detection result exported by ERT
- **Relation prediction:** import the bounding box of humans and objects into the retrained VTransE model and export the relation triples

Our Method

Object Instance Segmentation





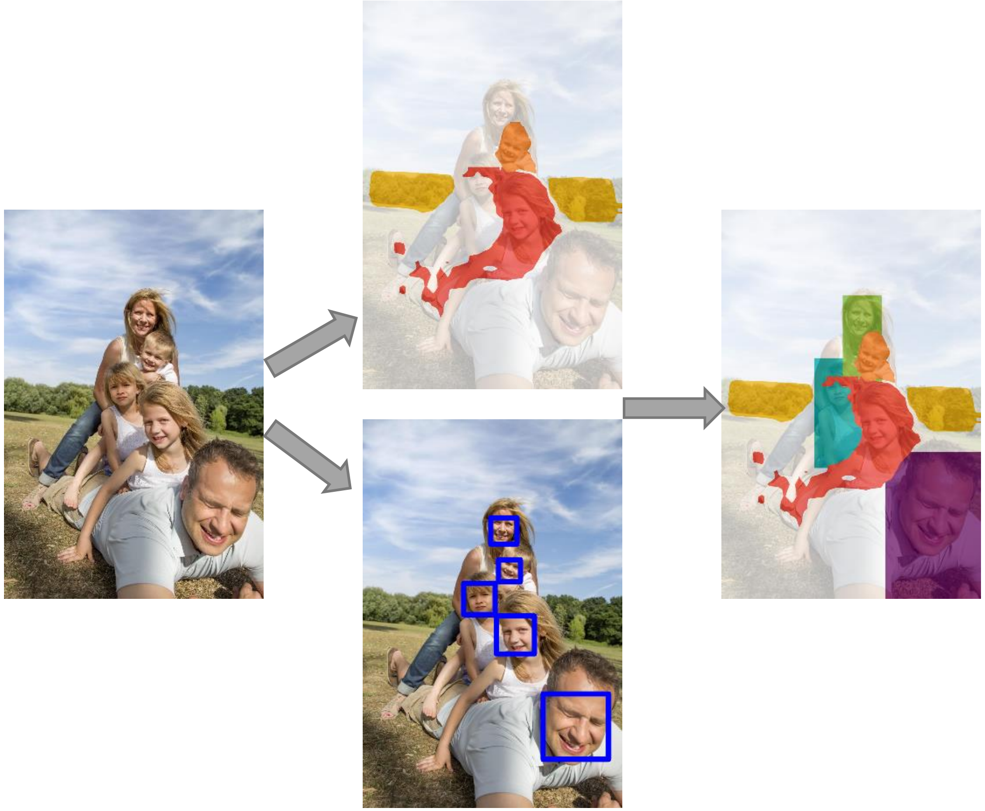
- Mask R-CNN extends Faster R-CNN by adding a branch for **predicting an object mask** in parallel with the existing branch for bounding box recognition
- According to the object categories provided by PIC dataset, we **retrain the last layer** of Mask R-CNN model with this loss function:

$$L = L_{class} + L_{box} + L_{mask}$$

- The result contains **object** instance segmentation and **human** instance segmentation

Our Method

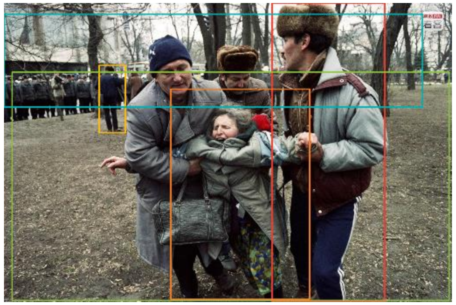
Human Instance Segmentation



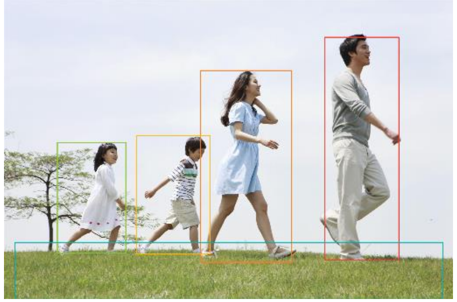
- Object instance segmentation result generated by Mask R-CNN **omit some human** in some images.
- We use ERT method to **detect human faces** and **estimate the location of the whole human body** with face location and common sense.
- Mask exporting from **Mask R-CNN should have a higher priority** and the expanding of **face detection should be a supplementary result**.

Our Method

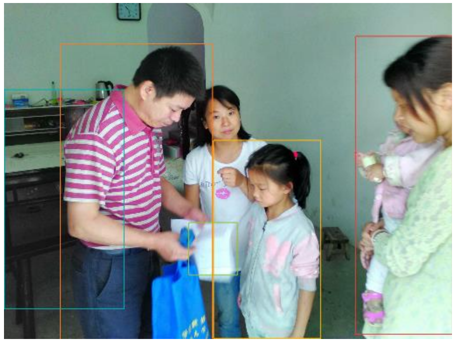
Relation Prediction



human next-to human
human behind human
human on grass
human in-front of human
human next-to human



human next-to human
human look human
human on vegetation
human next-to human



human next-to human
human next-to human
human look bag
human in-front-of table
human next-to human

- VTransE predicts relations from an image **in an end-to-end fashion** and refers to a visual relation as a **subject-predicate-object triple**.
- Inputs are **original images** and **bounding box** exporting from the result of human and object instance segmentation
- We **filter out the triples with little probability** and keep the triples with the same subject and object but higher score
- We **remove some of the result according to language prior**

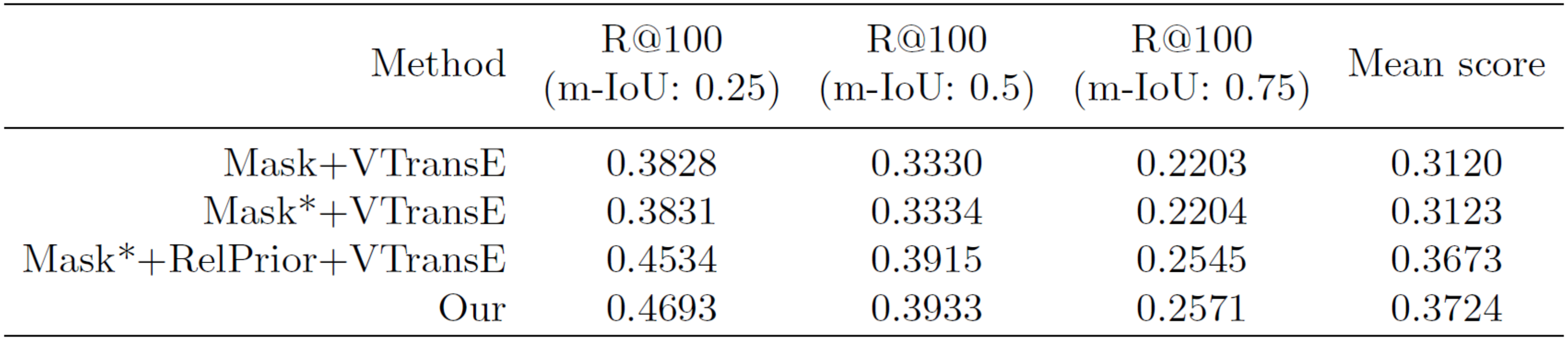
Experiments

	training set	validation set	test set
image	10000	1135	2998
subject/object category	85	85	85
relation category	31	31	31
segment	106959	12061	-
visual relation instance	167916	18729	-

Experiments

I7 3.5GHz CPU, 32GB memory, 1080Ti GPU, 1.9 seconds per image

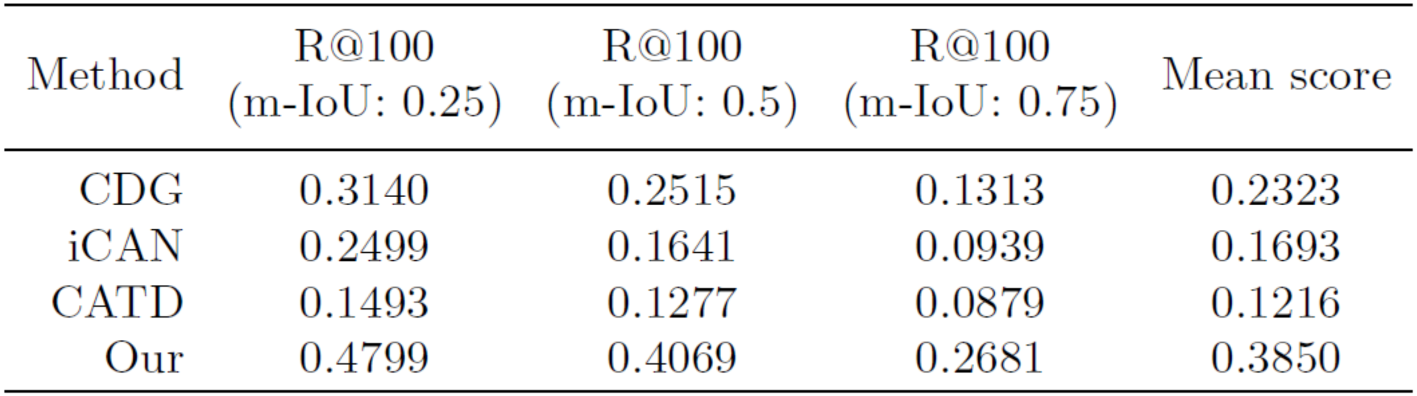
Mask R-CNN + VTransE: retrained Mask R-CNN model and retrained VTransE model
Mask R-CNN * + VTransE: fine-tuned Mask R-CNN model and retrained VTransE model
Mask R-CNN * + relation prior + VTransE: filter infrequent result
Mask R-CNN * + face detection + relation prior + VTransE: additional face detection



Method	R@100 (m-IoU: 0.25)	R@100 (m-IoU: 0.5)	R@100 (m-IoU: 0.75)	Mean score
Mask+VTransE	0.3828	0.3330	0.2203	0.3120
Mask*+VTransE	0.3831	0.3334	0.2204	0.3123
Mask*+RelPrior+VTransE	0.4534	0.3915	0.2545	0.3673
Our	0.4693	0.3933	0.2571	0.3724

- Object instance segmentation cannot be easily improved by global parameter adjustment.
- Face detection based person localization cannot accurately localize the persons.
- Relation prior is effective to visual relation predication.

Experiments



Method	R@100 (m-IoU: 0.25)	R@100 (m-IoU: 0.5)	R@100 (m-IoU: 0.75)	Mean score
CDG	0.3140	0.2515	0.1313	0.2323
iCAN	0.2499	0.1641	0.0939	0.1693
CATD	0.1493	0.1277	0.0879	0.1216
Our	0.4799	0.4069	0.2681	0.3850

- Our method has good generalization ability.
- Our method is better but the performance is far from the requirement in real applications.
- Human-centric visual relation segmentation is still a challenging task.

Thank you

Welcome to contact us!

Email: yf@smail.nju.edu.cn

南京大學
NANJING UNIVERSITY

MAGUS
Multimedia AnalyzinG
and UnderStanding