Solution for Homage Scene-graph Generation

Aixi Zhang[1*], Yue Liao[2*], Yongliang Wang[1], Miao Lu[1], Xiaobo Li[1], Si Liu[2]

[1]Alibaba Group, [2]Beihang University

[*]Equal contribution

2021-06

Outline

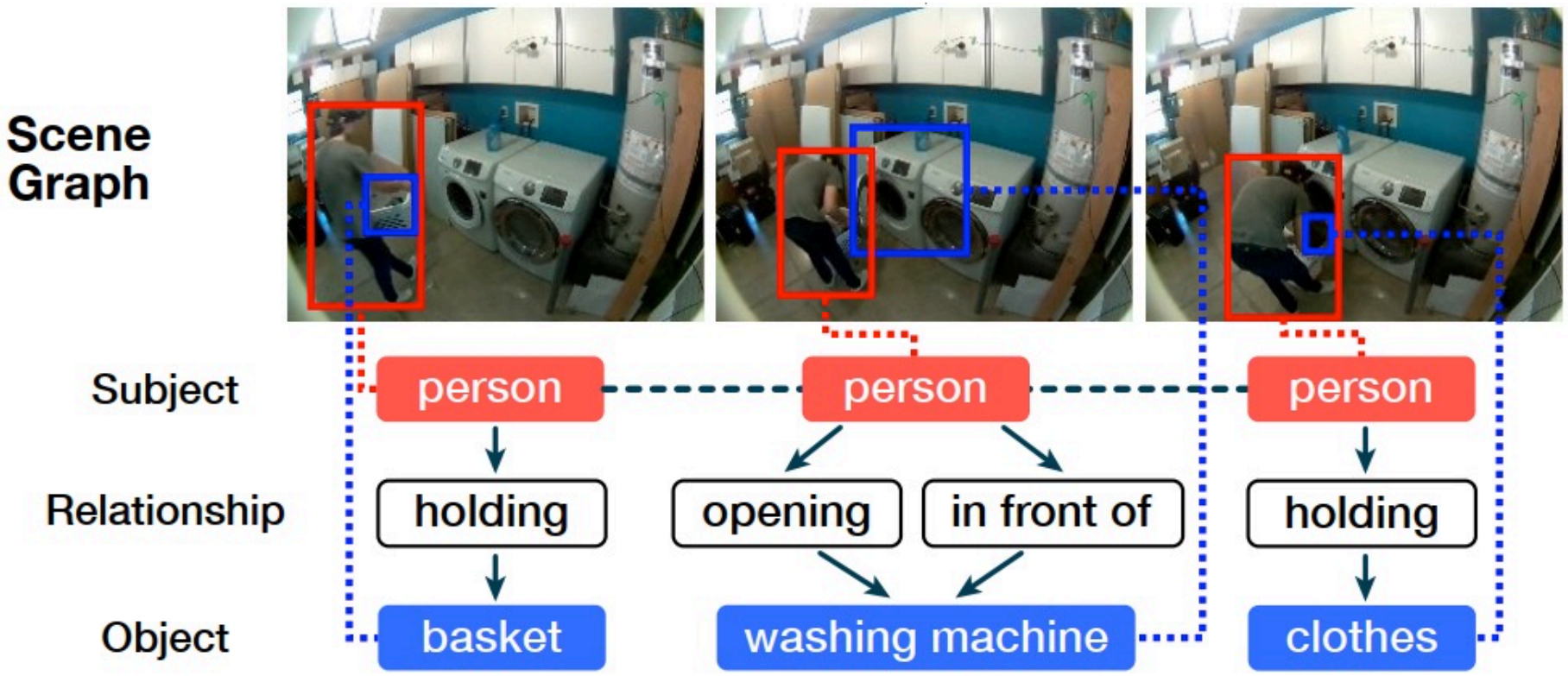
Alibaba Group

- Task Analysis
- Our Solution
- Experimental Results
- Summary

Homage SGG Task Analysis

Task Description：

Predict per-frame scene graphs to describe the relationship between a person and the object used during the execution of an action, as well as how they change as the video progresses



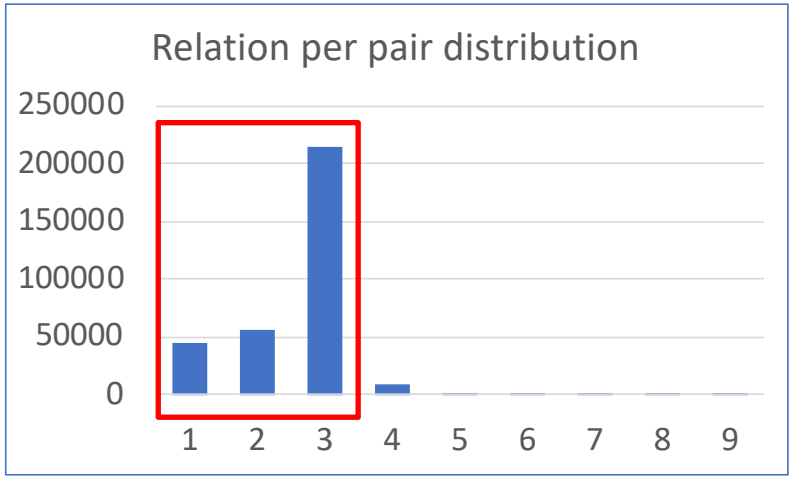
Home Action Genome: Cooperative Compositional Action Understanding, CVPR 2021

Homage Database Analysis

Database Description：

A large-scale multi-view video dataset of daily activities at home

Valid video num: 2,560
Valid frame num: 116,200
Pair labels num: 396,152, Pair per frame: 3.4
Relation labels num: 992,698, relation per pair: 2.5
Proportion of 1-3 relations per pair is 96.5%



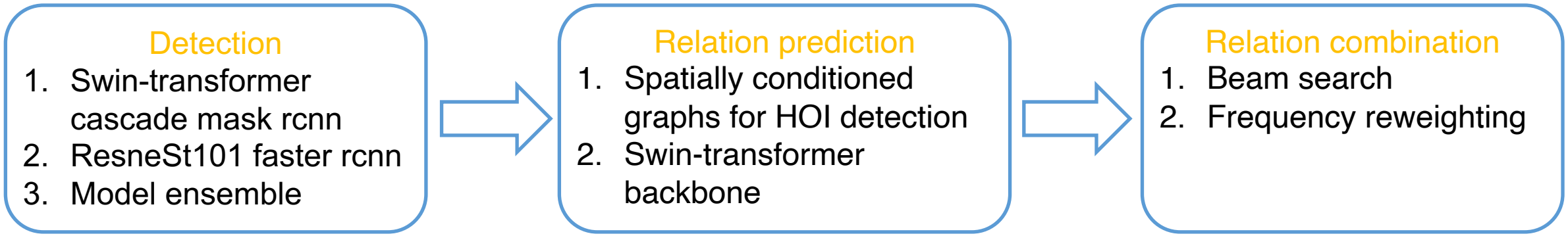
Object class num: 84, imbalance factor: 115.7 (42,819 / 370)
Relation class num: 25, imbalance factor: 20,164.5 (241,974 / 12)

We randomly sample 400 video as our validation subset, for following experiments.

Algorithm Modules

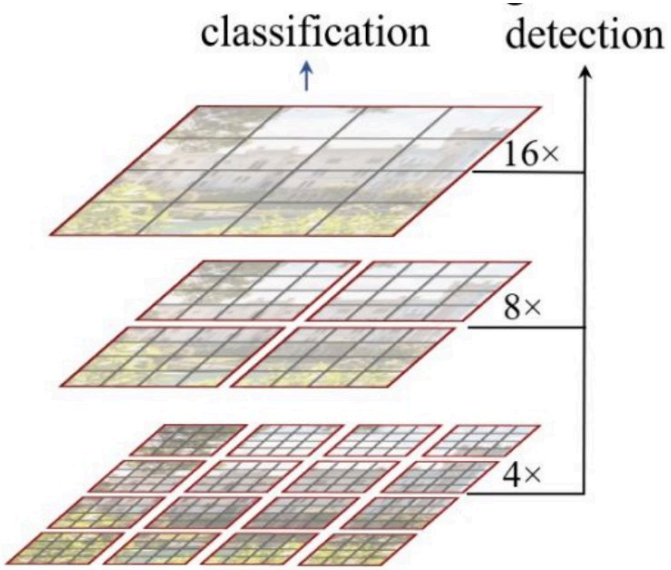
Alibaba Group

Three stage algorithm flow：

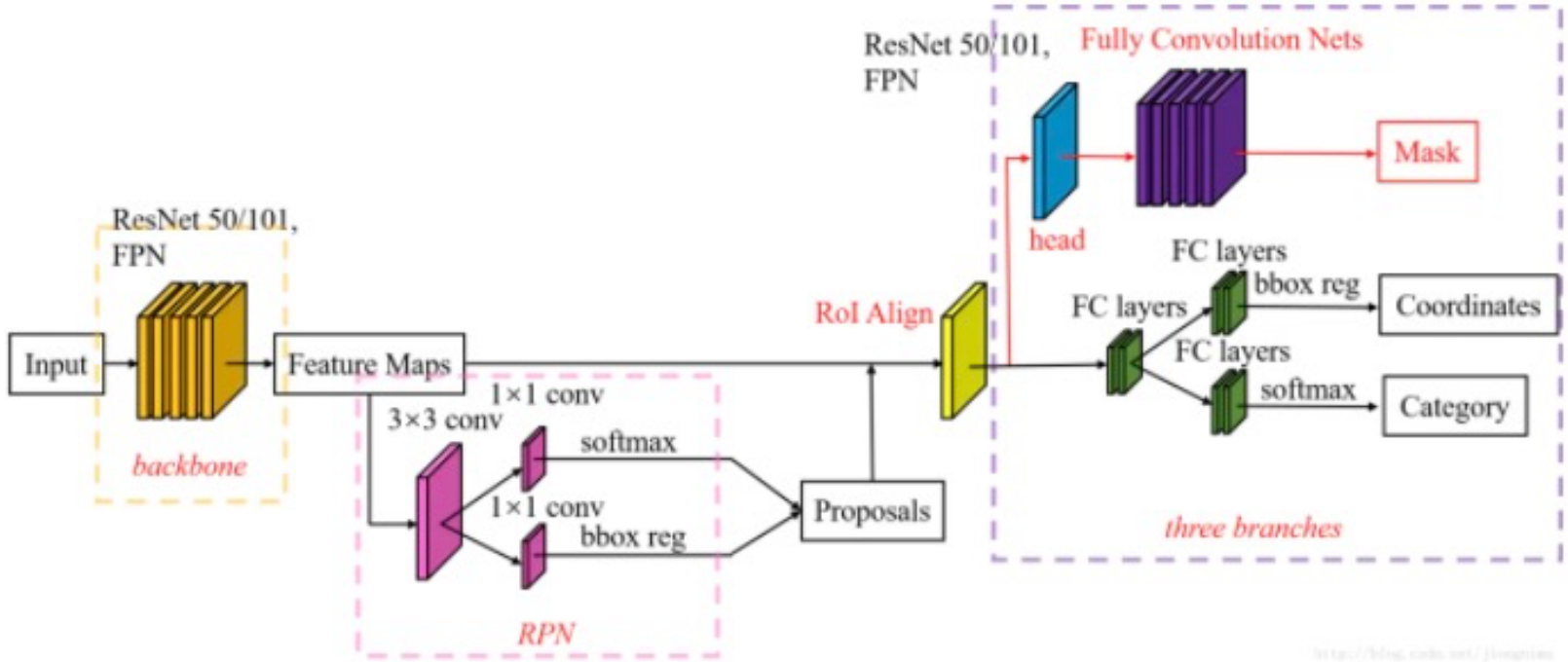


Detection

Swin-transformer backbone



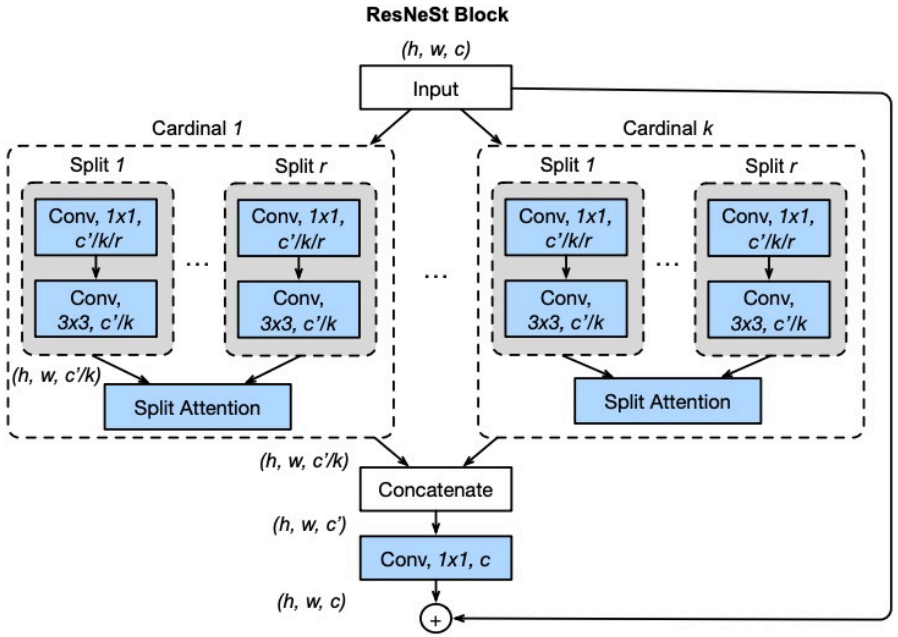
Mask Rcnn pipeline



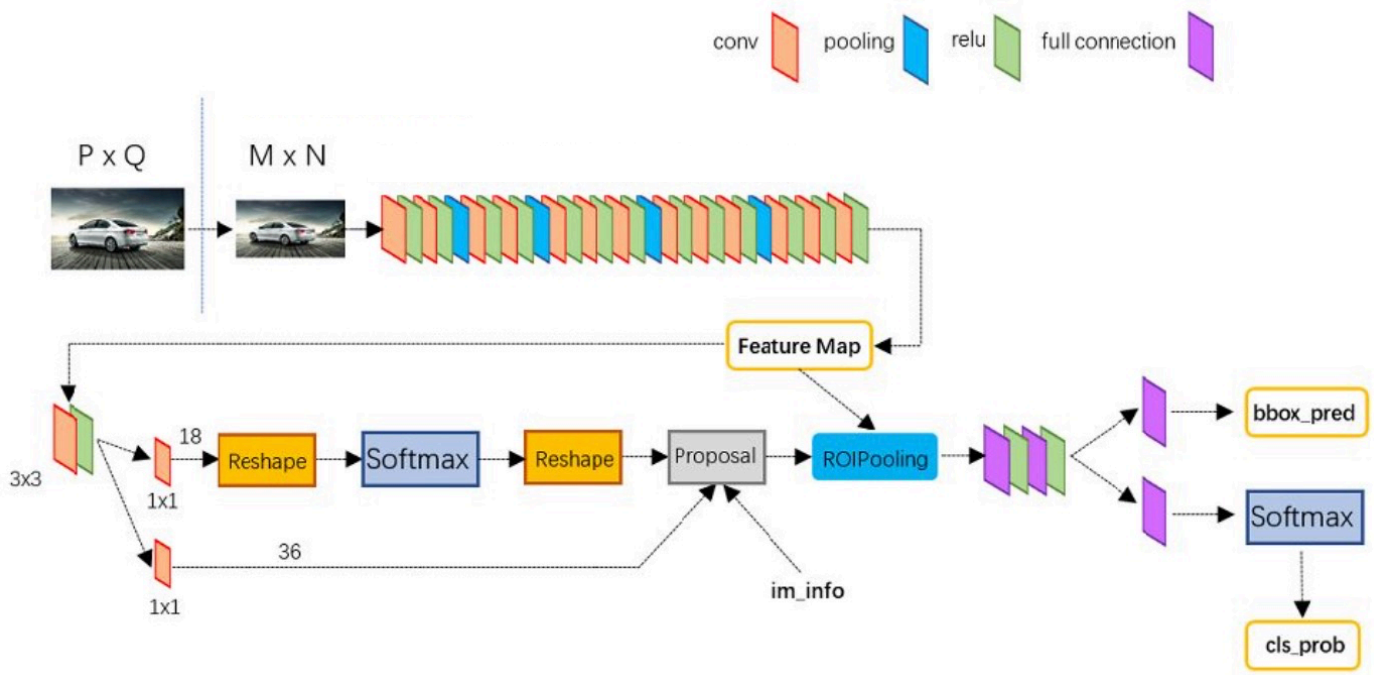
Detection

Alibaba Group

ResNeSt backbone



Faster RCNN pipeline



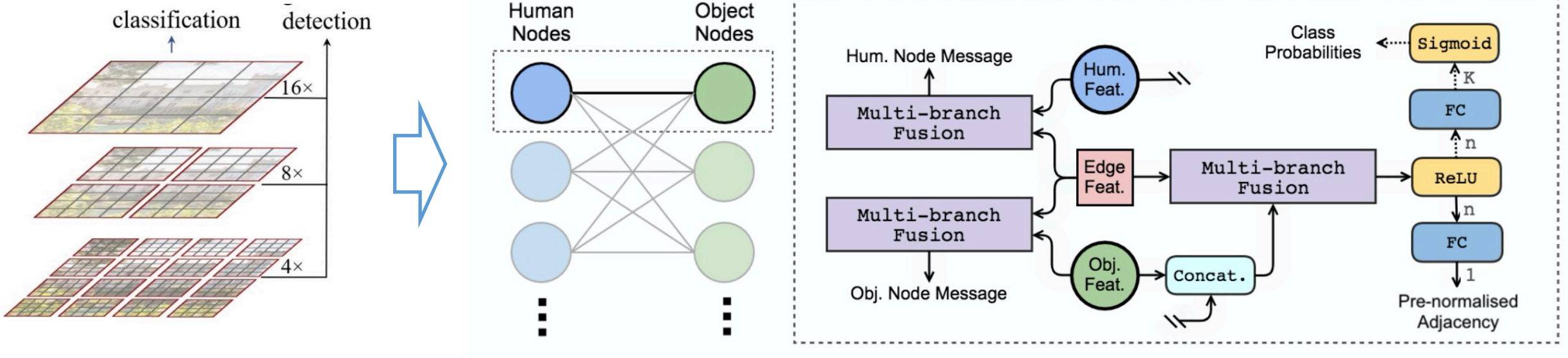
Detection Results

Alibaba Group

Methods	Description	Accuracy (%)
1	Swin-B	81.70
2	Multi-scale Swin-B	83.40
3	ResNeSt101	79.95
4	Multi-scale ResNeSt101	80.64
5	Merge 2 & 4	84.65

Relation Prediction

Swin-Transformer based spatially conditioned graph architecture for HOI detection



Spatially Conditioned Graphs for Detecting Human–Object Interactions, 2021

Relation Prediction Results

Alibaba Group

Evaluation results on our validation subset.

Methods	Description	Score (%)	Recall@10	Recall@20
1	On swin-transformer only detections	69.4	65.8	73.0
2	On merged detections	73.1	69.6	76.7
3	With frequency promotion	78.6	74.9	82.3

Finally, we get 76.57% accuracy and rank the first place.

Summary

Alibaba Group

- Design a three-stage algorithm flow for the Homage scene graph generation task
- Apply Swin-Transformer based SCG model for the task
- Achieve the first place in the Homage scene graph generation challenge

Thanks!