南京航空航天大學

Nanjing University of Aeronautics and Astronautics

# Mining the Benefits of Two-stage and One-stage HOI Detection

**Aixi Zhang**[1]* **Yue Liao**[2]* **Si Liu**[2]†
**Miao Lu**[1] **Yongliang Wang**[1] **Chen Gao**[2] **Xiaobo Li**[1]
[1]Alibaba Group [2]Beihang University

***NIPS 2021***

# background--HOI

The goal of Human-Object Interaction (HOI) detection is to make a machine detailedly understand human activities from a static image.

Human activities in this task are abstracted as a set of <human, object, action> HOI triplets.

# background--HOI

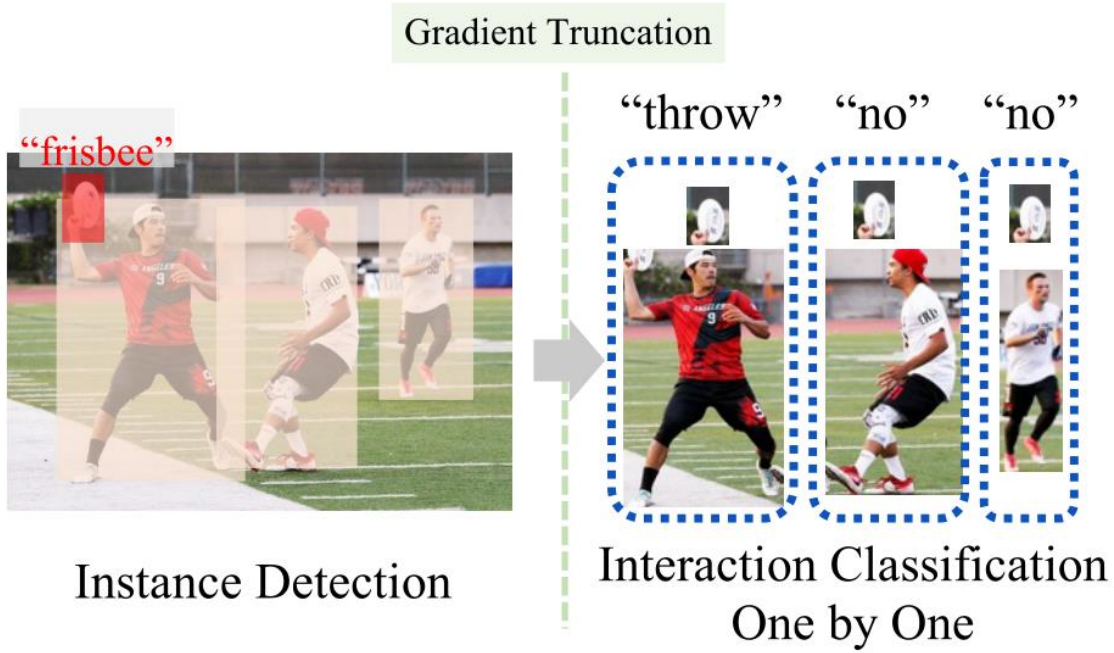


- locate human-object pairs
- classify their corresponding action

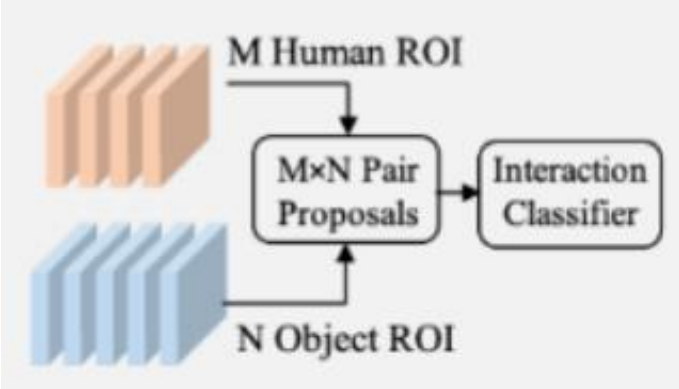
two-stage

one-stage

# two-stage




1. detect humans and objects
2. match humans and objects one by one (human-object pairs)
3. feed the pairs into an interaction classifier



# two-stage



- 缺点

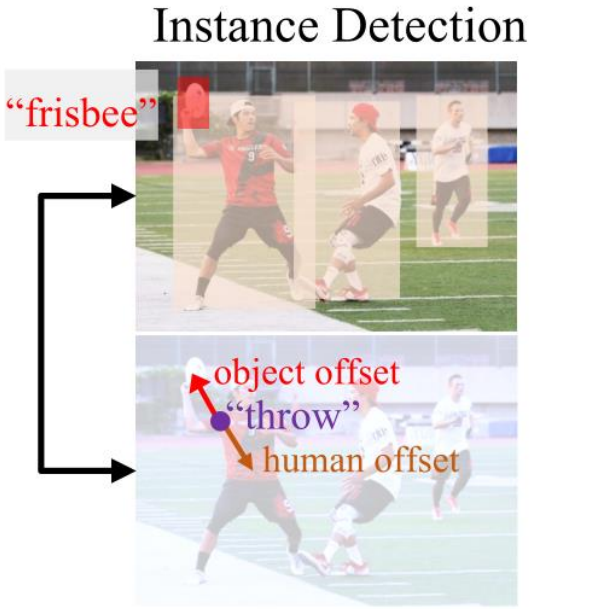
1. produce a more additional computational cost
   O(M*N)>>O(K')
2. the imbalance between positive and negative samples makes the model easily overfit to negative samples  no-interaction
3. the accuracy of interaction classification is influenced by the non-end-to-end pipeline

- 优点

However, it is an excellent property for two-stage methods that disentangling detection and interaction classification makes each stage focus on its task and produce good results in each stage.

# one-stage


Detect HOI Triplets with a Multi-task Learning




Interaction Classification & Association

Detect all HOI triplets S directly and simultaneously with an end-to-end framework

Most one-stage methods are interaction-driven, which directly locate the interaction point or interactive human-object pairs

# one-stage



- 优点

1. easily focus on the interactive human-object pairs
2. the time complexity is reduced to O(K’)

- 缺点

Coupling human-object detection and interaction classification limit their performance because it is hard to generate a unified feature representation for two very different tasks.

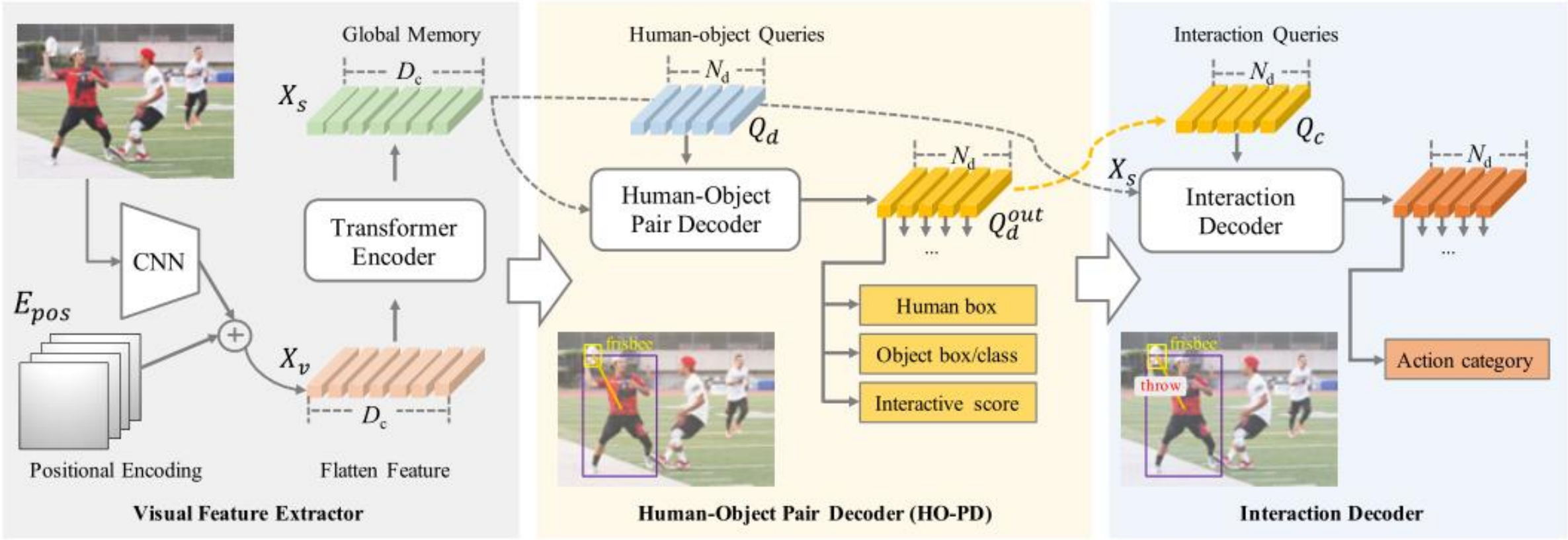
# idea

因此，针对这两种方法各自的优缺点，提出了一种如图所示的新结构，Cascade Disentangling Network (CDN)。它保留了一阶段范式的优点，同时引入了两阶段范式的优势。这是一种级联的一阶段范式。


Human-object Pair Generator

Interaction Classification

# method--CDN framework




**HO-PD:** predict a set of human-object pairs from the sequenced visual features

$$\boldsymbol{P}_{ho} = f_d(\boldsymbol{Q}_d, \boldsymbol{X}_s, \boldsymbol{E}_{pos}) \qquad \boldsymbol{P}_{ho}\text{: } \{(b_i^h, b_i^o), i \in \{1, 2, \cdots, N_d\}\}$$

**Interaction Decoder:** assign one or several action categories for each human-object query

$$\boldsymbol{P}_{cls} = f_{cls}(\boldsymbol{Q}_d^{out}, \boldsymbol{X}_s, \boldsymbol{E}_{pos}) \qquad \boldsymbol{P}_{cls} : \{a_i, i \in \{1, 2, \cdots, N_d\}\}$$

# method—Decoupling Dynamic Re-weighting



- **long-tail class distribution**

$$w_i^{(o,a)}\big|_{i\in\{1,2,\cdots,c(o,a)\}} = \left(\frac{\sum_{i=1}^{c(o,a)} N_i}{N_i}\right)^{p(o,a)}, \quad w_{bg}^{(o,a)} = \left(\frac{\sum_{i=1}^{c(o,a)} N_i}{N_{bg}^{(o,a)}}\right)^{p(o,a)}$$

In detail, we first train the whole model with regular losses. Then, we freeze the parameters of the visual feature extractor and only train the cascade disentangling decoders with a relatively small learning rate and the designed dynamic re-weighted losses.

# Training and Post-processing

The loss of CDN is composed by five parts: the box regression loss $L_b$, the intersection-over-union loss $L_{GIoU}$, the interactive score loss $L_p$, the object class loss $L^o_c$, and the action category loss $L^a_c$. The target loss is the weighted sum of these parts as:

$$L = \sum_{k \in (h,o)} (\lambda_b L_b^k + \lambda_{GIoU} L_{GIoU}^k) + \lambda_p L_p + \lambda_o L_c^o + \lambda_a L_c^a$$

# experiments



| *Strategy* | Full | Rare | Non-Rare |
|---|---|---|---|
| *iCAN** | 14.16 | 12.26 | 14.73 |
| *iCAN* † | 15.37 | 13.23 | 16.01 |
| *HO-PD+iCAN** | 24.05 | 18.32 | 25.76 |
| *QPIC [28]* | 29.07 | 21.85 | 31.23 |
| *CDN-S base* | **30.96** | **27.02** | **32.14** |

Table 1: **Analysis of Two-stage and One-stage Methods**. * denotes our implemented PyTorch version iCAN [6] baseline model. † denotes replacing instance detection boxes given by a HICO-Det fine-tuned DETR detector to extract box features. 'HO-PD+iCAN*' denotes replacing original one-by-one generated human-object pairs with our HO-PD generated. 'CDN-S base' denotes CDN-S w/o re-weighting and PNMS strategies.

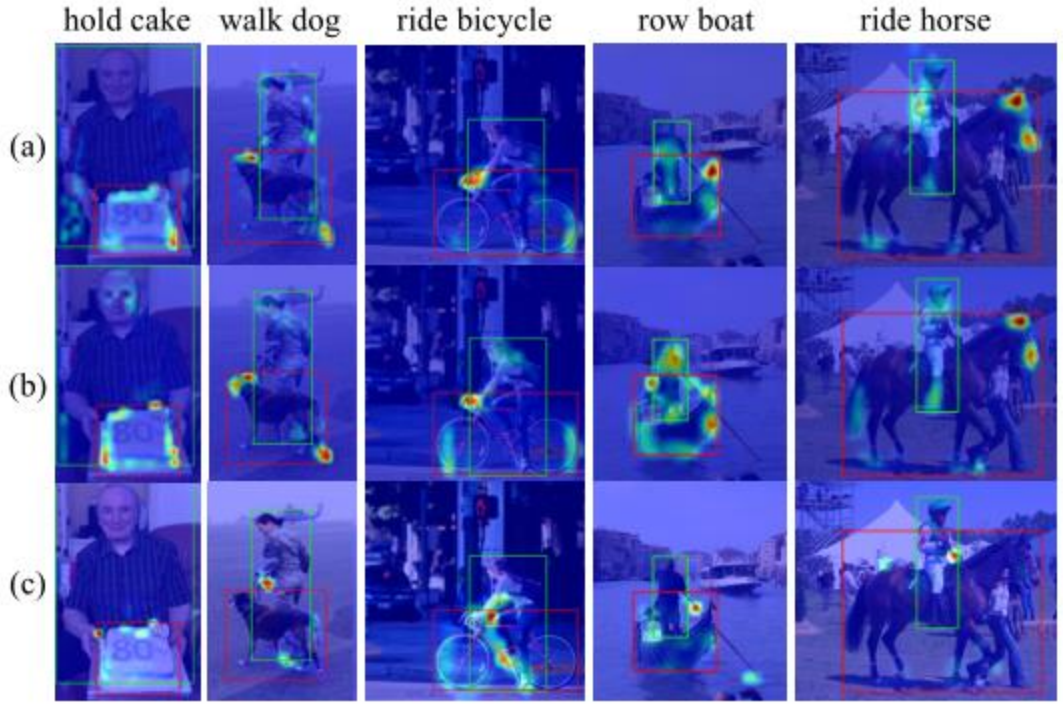



Figure 3: **Visualization of Feature Maps for Queries**. Visual attended features for query with top-1 score extracted from the last layer of the decoder of (a) QPIC, (b) HO-PD in CDN, and (c) interaction decoder in CDN. Zoom in for details.

# experiments

| Method | Detector | Backbone | Extra | Default | | | Know Object | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Full | Rare | Non-Rare | Full | Rare | Non-Rare |
| Two-stage Method: | | | | | | | | | |
| InteractNet [7] | COCO | ResNet-50-FPN | ✗ | 9.94 | 7.16 | 10.77 | - | - | - |
| GPNN [24] | COCO | Res-DCN-152 | ✗ | 13.11 | 9.34 | 14.23 | - | - | - |
| iCAN [6] | COCO | ResNet-50 | ✗ | 14.84 | 10.45 | 16.15 | 16.26 | 11.33 | 17.73 |
| No-Frills [9] | COCO | ResNet-152 | *P* | 17.18 | 12.17 | 18.68 | - | - | - |
| PMFNet [30] | COCO | ResNet-50-FPN | *P* | 17.46 | 15.65 | 18.00 | 20.34 | 17.47 | 21.20 |
| CHGNet [31] | COCO | ResNet-50 | ✗ | 17.57 | 16.85 | 17.78 | 21.00 | 20.74 | 21.08 |
| DRG [5] | COCO | ResNet-50-FPN | *T* | 19.26 | 17.74 | 19.71 | 23.40 | 21.75 | 23.89 |
| VCL [12] | COCO | ResNet-50 | ✗ | 19.43 | 16.55 | 20.29 | 22.00 | 19.09 | 22.87 |
| IP-Net [32] | COCO | Hourglass-104 | ✗ | 19.56 | 12.79 | 21.58 | 22.05 | 15.77 | 23.92 |
| VSGNet [29] | COCO | ResNet-152 | ✗ | 19.80 | 16.05 | 20.91 | - | - | - |
| FCMNet [22] | COCO | ResNet-50 | ✗ | 20.41 | 17.34 | 21.56 | 22.04 | 18.97 | 23.12 |
| ACP [15] | COCO | ResNet-152 | *T* | 20.59 | 15.92 | 21.98 | - | - | - |
| PD-Net [35] | COCO | ResNet-152-FPN | *T* | 20.81 | 15.90 | 22.28 | 24.78 | 18.88 | 26.54 |
| DJ-RN [16] | COCO | ResNet-50 | *P* | 21.34 | 18.53 | 22.18 | 23.69 | 20.64 | 24.60 |
| IDN [17] | COCO | ResNet-50 | ✗ | 23.36 | 22.47 | 23.63 | 26.43 | 25.01 | 26.85 |
| One-stage Method: | | | | | | | | | |
| UnionDet [13] | COCO | ResNet-50-FPN | ✗ | 17.58 | 11.72 | 19.33 | 19.76 | 14.68 | 21.27 |
| DIRV [4] | COCO | EfficientDet-d3 | ✗ | 21.78 | 16.38 | 23.39 | 25.52 | 20.84 | 26.92 |
| PPDM-Hourglass [20] | HICO-DET | Hourglass-104 | ✗ | 21.94 | 13.97 | 24.32 | 24.81 | 17.09 | 27.12 |
| HOI-Trans [39] | HICO-DET | ResNet-50 | ✗ | 23.46 | 16.91 | 25.41 | 26.15 | 19.24 | 28.22 |
| GG-Net [37] | HICO-DET | Hourglass-104 | ✗ | 23.47 | 16.48 | 25.60 | 27.36 | 20.23 | 29.48 |
| ATL [11] | HICO-DET | ResNet-50 | ✗ | 23.81 | 17.43 | 25.72 | 27.38 | 22.09 | 28.96 |
| HOTR [14] | HICO-DET | ResNet-50 | ✗ | 25.10 | 17.34 | 27.42 | - | - | - |
| AS-Net [3] | HICO-DET | ResNet-50 | ✗ | 28.87 | 24.25 | 30.25 | 31.74 | 27.07 | 33.14 |
| QPIC [28] | HICO-DET | ResNet-50 | ✗ | 29.07 | 21.85 | 31.23 | 31.68 | 24.14 | 33.93 |
| CDN-S | HICO-DET | ResNet-50 | ✗ | 31.44 | 27.39 | 32.64 | 34.09 | 29.63 | 35.42 |
| CDN-B | HICO-DET | ResNet-50 | ✗ | 31.78 | **27.55** | 33.05 | 34.53 | **29.73** | 35.96 |
| CDN-L | HICO-DET | ResNet-101 | ✗ | **32.07** | 27.19 | **33.53** | **34.79** | 29.48 | **36.38** |

Table 2: **Performance comparison on the HICO-Det test set.** The 'P', 'T' represent human pose information and the language feature, respectively.

# experiments

| Method | Extra | $AP^{S1}_{role}$ | $AP^{S2}_{role}$ |
|---|---|---|---|
| Two-stage Method: | | | |
| InteractNet [7] | ✗ | 40.0 | - |
| GPNN [24] | ✗ | 44.0 | - |
| iCAN [6] | ✗ | 45.3 | 52.4 |
| TIN [18] | ✗ | 47.8 | 54.2 |
| VCL [12] | ✗ | 48.3 | - |
| DRG [5] | *T* | 51.0 | - |
| IP-Net [32] | ✗ | 51.0 | - |
| VSGNet [29] | ✗ | 51.8 | 57.0 |
| PMFNet [30] | *P* | 52.0 | - |
| PD-Net [35] | *T* | 52.6 | - |
| CHGNet [31] | ✗ | 52.7 | - |
| FCMNet [22] | ✗ | 53.1 | - |
| ACP [15] | *T* | 53.23 | - |
| IDN [17] | ✗ | 53.3 | 60.3 |
| One-stage Method: | | | |
| UnionDet [13] | ✗ | 47.5 | 56.2 |
| HOI-Trans [39] | ✗ | 52.9 | - |
| AS-Net [3] | ✗ | 53.9 | - |
| GG-Net [37] | ✗ | 54.7 | - |
| HOTR [14] | ✗ | 55.2 | 64.4 |
| DIRV [4] | ✗ | 56.1 | - |
| QPIC [28] | ✗ | 58.8 | 61.0 |
| CDN-S | ✗ | 61.68 | 63.77 |
| CDN-B | ✗ | 62.29 | 64.42 |
| CDN-L | ✗ | **63.91** | **65.89** |

Table 3: **Performance comparison on the V-COCO test set**. The 'P', 'T' represent the human pose information and the language feature, respectively.

# Ablation Study

| *Strategy* | Full | Rare | Non-Rare |
|---|---|---|---|
| QPIC [28] | 29.07 | 21.85 | 31.23 |
| *base* | 31.06 | 26.68 | 32.36 |
| + *re-weighting* | 31.38 | 27.36 | 32.58 |
| + *PNMS* | **31.78** | **27.55** | **33.05** |

(a) **Strategies:** Analysis of improvements by various training strategies.

| *Strategy* | $L_Q$ | $p$ | Full | Rare | Non-Rare |
|---|---|---|---|---|---|
| *base* | - | - | 31.06 | 26.68 | 32.36 |
| *decouple* | - | - | 30.90 | 26.09 | 32.33 |
| *static* | - | 0.7 | 31.25 | 27.12 | 32.49 |
| *dynamic* | $2 \times N_s$ | 0.8 | 31.33 | 27.45 | 32.49 |
| *dynamic* | $1 \times N_s$ | 0.7 | 31.34 | **27.48** | 32.49 |
| *dynamic* | $2 \times N_s$ | 0.7 | **31.38** | 27.36 | **32.58** |

(b) **Dynamic re-weighting:** Analysis of decouple training with dynamic re-weighted losses, *i.e.*, different queue length $L_Q$, coefficient $p$ and dynamic or static.

| $\alpha$ | $\beta$ | *thres* | Full | Rare | Non-Rare |
|---|---|---|---|---|---|
| - | - | - | 31.38 | 27.36 | 32.58 |
| *1* | *1* | *0.8* | 31.66 | 27.46 | 32.91 |
| *1* | *1* | *0.7* | 31.75 | 27.50 | 33.03 |
| *1* | *0.7* | *0.7* | 31.77 | 27.54 | 33.03 |
| *1* | *0.5* | *0.8* | 31.75 | 27.51 | 33.02 |
| *1* | *0.5* | *0.7* | **31.78** | **27.55** | **33.05** |

(c) **PNMS:** The effects of different settings of PNMS coefficients, *i.e.*, $\alpha$, $\beta$, and *thres* denotes threshold.

Table 4: **Ablation studies of our proposed method on the HICO-Det test set.** We carry out all experiments based on the base model (CDN-B).

# Thanks