# Detecting Human-Object Interaction via Fabricated Compositional Learning

Zhi Hou[1], Baosheng Yu[1], Yu Qiao[2,3], Xiaojiang Peng[4], Dacheng Tao[1]
[1] School of Computer Science, Faculty of Engineering, The University of Sydney, Australia
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[3] Shanghai AI Laboratory
[4] Shenzhen Technology University
zhou9878@uni.sydney.edu.au, baosheng.yu@sydney.edu.au, yu.qiao@siat.ac.cn,
pengxiaojiang@sztu.edu.cn, dacheng.tao@sydney.edu.au

## 1. Overview of Appendixes

In this supplementary file, we provide additional details of the proposed method in Section B. Section C demonstrates more quantitative analysis (*e.g*. Object identity embedding, VRD, Semantic verb regularization, comparison of detector, additional ablation study and so on). In the last section, we illustrate the qualitative results (*e.g*. analysis of fabricated features).

## 2. Additional Details of the Proposed Method

### 2.1. More examples of Open Long-tailed HOI Detection

Figure 1 provides more clear illustration of open long-tailed HOI detection. Open long-tailed HOI detection aims to detect head, tail and unseen classes in one integrated way from long-tailed HOI examples.

### 2.2. Factorized model

We implement the factorized model under our framework. In details, we replace the HOI branch in Figure 3 in the paper with verb and object stream. The two streams predict the verb and object respectively. During inference, we merge the score of verb and object to obtain HOI score as follows,

$$\mathbf{S}_{hoi} = (\mathbf{S}_o\mathbf{A}_o) + (\mathbf{S}_v\mathbf{A}_v), \tag{1}$$

where $\mathbf{A}_v$ ($\mathbf{A}_o$) is the co-occurrence matrix between verbs (objects) and HOIs, $\mathbf{S}_o$ is the score from object stream and $\mathbf{S}_v$ is the score from verb stream.

### 2.3. The Effect of Objects on HOI Detection

In the nature, different types of objects form a long-tail distribution. Then, all those actions that people perform on those objects are inevitably long-tailed. As a result, those HOIs that we observed are long-tailed. This motivates us to fabricate balanced objects for composing HOI samples with visual verbs. We have demonstrated the long-tailed distribution of objects in Figure 2 in the paper and the effect of different object detector on HOI detection in Table 7 in paper. We further illustrate HOI detection has roughly similar performance to object detection among most object categories in Figure 2, which also illustrates the importance of object detector for HOI detection at the same time. Meanwhile, it is necessary to balance the the distribution of objects.

### 2.4. The Number of Primitives in two Zero-Shot Setting

We have count the number of unseen HOI primitives (*i.e*. verb and object) in the remaining data of two zero-shot setting. Unseen HOIs of rare first zero-shot has 40 verbs, 5 of which have less than 10 instances in the remaining data, while Unseen HOIs of non-rare first zero-shot have only 30 verbs and all have more 10 instances. We think this partly explains why Factorized method has worse result on unseen category in rare first setting. When the primitives of unseen HOI are few in the training data. Factorized method possibly achieves worse result on unseen category.

### 2.5. Fusion of HOI prediction and Generic Object Detector

In our experiment, we directly predict 600 HOI classes in HICO-DET. The predictions of HOI (verb-object pair) also contain object information. We think the object information in HOI prediction and the generic object detector might be complementary. Thus, we convert HOI scores $S_{hoi}$ to object scores and fuse it with $s_o$ as follow,

$$\hat{s}_o = \beta_1 \frac{(S_{sp} \cdot S_{hoi})\mathbf{A}_o^{\mathsf{T}}}{\mathbf{B}} + \beta_2 s_o, \tag{2}$$

Where $\beta_1$ and $\beta_2$ are 0.3 and 0.7 respectively, $\mathbf{B} \in R^{N_o}$ and $\mathbf{B}_i = \sum_{j=0}^{C} \mathbf{A}_{o_{i,j}}$. Then, we use the new object score

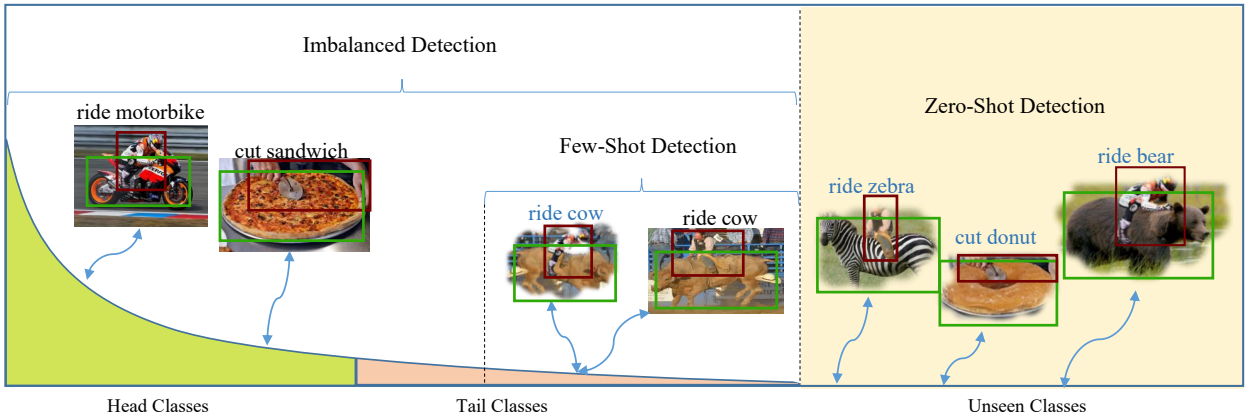


Figure 1. Open long-tailed HOI detection addresses the problem of imbalanced learning and zero-shot learning in a unified way. We propose to compose new HOIs for open long-tailed HOI detection. Specifically, the blurred HOIs, *e.g.*, "ride bear", are composite, while the black HOIs are real.

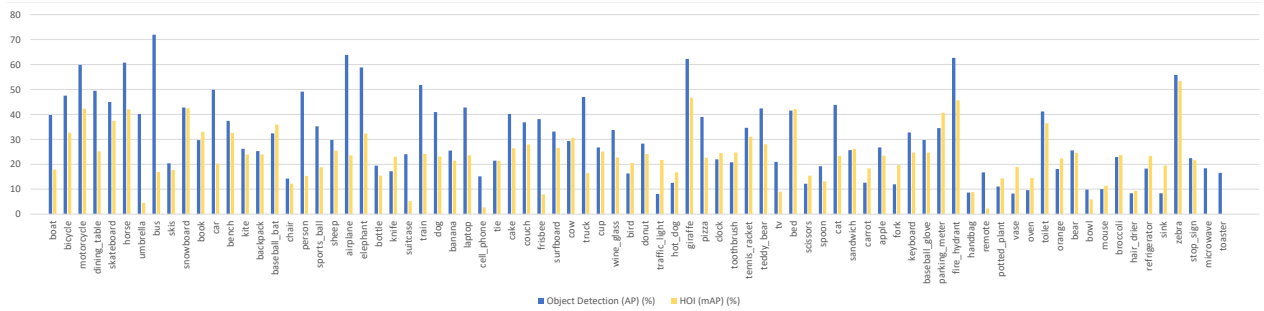


Figure 2. Illustration of Object detection result and HOI detection result in HICO-DET dataset. Blue is Object result. Yellow is HOI result. We average HOI detection AP according to the object categories for a direct comparison.

Table 1. Illustration of the effect of fusing HOI predicition to object score. This experiment is based on word-embedding object identity FCL model.

| Method | Full | Rare | NonRare |
|---|---|---|---|
| FCL$^{VCL}$ w/o Fusion | 24.42 | 19.68 | 25.84 |
| FCL$^{VCL}$ | 24.68 | 20.03 | 26.07 |

$\hat{s}_o$ in Equation 6. Meanwhile, we can also update the object category according to $\hat{s}_o$. Table 1 shows we can improve the result a bit under VCL detector which provides all scores for each object category. Noticeably, our baseline under VCL detector also uses this strategy and we do not use this in zero-shot settings. For the DRG object detector, we also do not use this strategy. To some extent, this slightly shows *HOI prediction and object detection can be mutually promoted*, and provides some insights for our future work although this strategy is not much useful.

# 3. Additional Quantitative analysis

## 3.1. Object Identity

In Table 2, we compare three kinds of object identity. The object variables are identified after we fine-tune the fabricator in the first step. Meanwhile, in the end-to-end optimization, the object variables can maintain object semantic information. We find word embedding [7] and object variables achieve similar performance ( 24.78% vs 24.68%), while the performance of one-hot representation is a bit worse. Particularly, the HOI model is initialized with a pre-trained object detector model. Thus, one-step optimization can also optimize the Fabricator according to the pre-trained backbone. In the main paper, the result of long-tailed HOI detection is the model using word embedding as identity embedding. For simplicity, we use randomly initialized variables as object identity embedding for other model, *i.e.* randomly initialize identity embedding.

Table 2. Illustration of the effect of different object identity in the proposed fabricator on HICO-DET dataset[1].

| Method | Full | Rare | NonRare |
|---|---|---|---|
| object variables | **24.78** | **20.05** | **26.19** |
| word embedding | 24.68 | 20.03 | 26.07 |
| one-hot | 24.38 | 19.49 | 25.84 |

| Method | Zero-Shot | All |
|---|---|---|
| MFURLN [10] | - | 58.2 |
| MFURLN [10]* | 25.26 | 57.87 |
| Ours | **27.31** | **58.31** |

Table 3. Illustration of Predicate Detection in Visual Relation Detection. Zero-shot means the relation (subject, predicate, object) do not exist in the training data.

| FCL | S | V | Full | Rare | NonRare | Unseen |
|---|---|---|---|---|---|---|
| - | ✓ | - | 18.22 | 15.69 | 20.74 | 12.98 |
| ✓ | ✓ | - | 19.39 | 17.99 | 21.21 | 14.83 |
| ✓ | - | ✓ | 19.61 | 18.69 | 21.13 | 15.86 |
| ✓ | ✓ | ✓ | 19.62 | 18.38 | 21.61 | 14.73 |

Table 4. Illustration of semantic regularization modules based on the ablated setting in paper. FCL Means proposed Compostional Learning. S means semantic regularize loss. V means auxiliary verb loss (verb regularization loss in paper).

### 3.2. Visual Relation Detection

We also present the efficiency of FCL in Predicate Detection on Visual Relation Detection [6] in Table 3. Here, we combine subject, predicate and fabricated object to generate novel relation samples [10]. Table 3 illustrates an important improvement on zero-shot predicate detection compared to the state-of-the-art approach with FCL.

### 3.3. Semantic Verb Regularization

We also experiment with semantic verb regularization similar to [9] with Graph Convolutional Network and verb word embeddings graph. In details, we use the cosine distance loss to regularize the visual verb representation to be similar to the corresponding word embedding. Here, similar to [9], we equally treat same category of verbs among different HOIs as same. Table 4 illustrates FCL is orthogonal to semantic regularization. Meanwhile, auxiliary verb loss achieve similar performance compared to semantic verb regularization [9]. When we incorporate both semantic regularization and auxiliary verb loss, the improvement is limited. This means verb regularization loss in the paper and semantic verb regularization have similar effect on the model.

Table 5. Illustration of auxiliary object loss on HICO-DET dataset[1] based object variables identity. Here, auxiliary object loss aims to regularize visual objects

| Method | Full | Rare | NonRare |
|---|---|---|---|
| w/o object loss | **24.78** | **20.05** | **26.19** |
| auxiliary object loss | 24.54 | 19.93 | 25.92 |

Table 6. Illustration of the box for verb representation on HICO-DET dataset[1].

| Method | Full | Rare | NonRare |
|---|---|---|---|
| baseline(human box) | 22.91 | 16.66 | 24.77 |
| FCL (human box) | 23.83 | 18.62 | 25.39 |

Table 7. The result while filtering out the composite HOIs according to the similarity between the fake objects and original objects. #Neighbors ($K$) means top $K$ neighbors according to similarity. This experiment is based on ablated setting in Table 3 in paper. When the number of neighbors is 80, we do not filter out composite HOIs according to similarity.

| #Neighbors ($K$) | 1 | 5 | 10 | 20 | 40 | 80 |
|---|---|---|---|---|---|---|
| FCL (Full) | 18.70 | 19.15 | 19.19 | 19.48 | 19.60 | 19.61 |

### 3.4. Object Feature Regularization

**visual object feature regularization.** Object features are usually more discriminative. Meanwhile, we initialize our backbone with the faster-rcnn pre-trained in COCO dataset, which largely helps us to obtain discriminative object features. Thus, it is unnecessary to use auxiliary object loss to regularize object features (See Table 5). Meanwhile, we find the object features is more discriminative from the t-SNE graph in Figure 6.

### 3.5. The Effect of Union Box on FCL

We extract verb representation from the union box of human and object. In Table 6, we illustrate with human box verb, FCL still effectively improves the baseline. This shows the proposed method is orthogonal to the verb representation. Noticeably, although the union box contains the object, the HOI model mainly learns the verb representation via compositional learning, and largely ignores the identity information of the object. Thus, the object in the union box do not have much effect on Fabricator. By comparing human box and union box for verb representation in Table 2 in paper and Table 6, we find verb representation from union box largely improves the performance since it provides more context information for verb representation.

Table 8. Comparison between step-wise optimization and one step optimization in unseen object HOI detection.

| Method | Full | Rare | NonRare | Unseen |
|---|---|---|---|---|
| one step | 19.87 | 15.01 | 22.51 | **15.54** |
| step-wise | **20.13** | **16.71** | **22.82** | 13.85 |

Table 9. Illustration of recall of HOI under DRG detector, VCL detector and GT boxes.

| Detector | Full (mAP) | Recall (mRec) |
|---|---|---|
| FCL$^{VCL}$ | 24.68 | 62.07 |
| FCL$^{DRG}$ | 29.12 | 82.81 |
| FCL$^{GT}$ | 44.26 | 86.08 |

Table 10. Illustration of FCL without re-weighting on long-tailed HOI detection.

| FCL | Full | Rare | NonRare |
|---|---|---|---|
| - | 20.79 | 13.19 | 23.06 |
| ✓ | 21.20 | 15.48 | 22.90 |

Table 11. Illustration of proposed modules on long-tailed HOI detection. FCL Means proposed Fabricated Compostional Learning. V means verb regularization loss.

| FCL | V | Full | Rare | NonRare |
|---|---|---|---|---|
| - | - | 23.35 | 17.08 | 25.22 |
| ✓ | - | 23.86 | 18.16 | 25.56 |
| - | ✓ | 23.94 | 17.48 | 25.87 |
| ✓ | ✓ | **24.78** | **20.05** | **26.19** |

### 3.6. Additional Object Detector Analysis

We notice there is a large gap between VCL [4] detector and DRG [2]. VCL provides the detection result (*i.e.* 30.79% mAP), while we do not know the detection result of DRG detector. We do not achieve the similar object detection performance to DRG [2] when we fine-tune Faster R-CNN on HICO-DET training set. However, we think we can compare the two detector by the recall of HOI detection as illustrated in Table 9. Recall can also be used to compare the object detection performance between one-stage HOI detection and two-stage HOI detection. Table 9 shows FCL$^{DRG}$ nearly achieves similar result to FCL$^{GT}$ on Recall. FCL$^{GT}$ still requires the network to discriminate which pair of human and object boxes has interaction.

### 3.7. Verb Analysis

The same verb might has different meanings in different HOIs. However, the verb in HOI dataset (e.g. HICO-DET) mainly represents action. Thus, the verb in HOI dataset is usually not ambiguous. Meanwhile, the deep convolutional network (*e.g.* Resnet) is able to fit some ambiguous and even random data [11]. Therefore, we can use factorized method [9] for HOI detection and the ambiguous verbs do not affect the compositional learning on HICO-DET [4], even if there are still some ambiguous verbs (e.g. hold) who can be related to multiple objects.

Besides, we further demonstrates the improvement of FCL among different categories of verbs in Figure 3. We find the ambiguity does not affect the performance of those verbs in fact. For example, although the verb "hold" is related to 61 kinds of objects in HICO-DET, the correpsonding HOIs of "hold" still achieves considerable improvement.

Inspired by that people interact similar objects in a similar manner. we also design an approach to select composite HOIs according to the similarity between different object of objects, *i.e.* we only keep those composite HOIs whose object is in the top $K$ neighbors of the verb's original object. The original object of the verb is the visual object paired with the verb in the HOI annotation. This helps us to filter out those ambiguous composite HOIs. Specifically, we calculate the similarity between different classes of objects by its word embedding [7]. Then we can obtain the top $K$ neighbors for each class of objects. Table 7 shows with more similar objects, the performance steadily improves. Particularly, there are only one verb relating to more than 40 HOIs, and 4 verbs with more than 20 HOIs in HICO-DET. When $K = 1$, we only keep composite HOIs whose objects have the same label to the original object.

### 3.8. Orthogonality to previous methods

**Orthogonal to spatial pattern**. Table 13 illustrates that the spatial pattern strategy [3, 5, 8] largely improves the performance, and the proposed compositional learning is orthogonal to spatial pattern.

**Orthogonal to re-weighting**. In our baseline, we utilize the re-weighting strategy that is used in [5, 4] to compare directly with [4]. We demonstrate FCL is orthogonal to re-weighting in Table 10. Without the useful re-weighting strategy, FCL still achieves similar improvement than baseline.

### 3.9. Complementary Analysis of fabricator

In this section, we conduct analysis of fabricator on HOI detection without unseen data (the full long-tailed HOI detection). We witness the similar trend compared to the ablation study in the paper.

**Verb and Noise for fabricating objects**. Table 3.9 demonstrates the efficiency of verb and noise. Particularly, the performance in the full HOI detection drops larger than that in zero-shot study in the paper. We think it is because the improvement on unseen category is large, while there are no unseen category in the full HOI detection.

**Verb Fabricator**. Table 3.9 illustrates if we fabricate verb features to augment HOI samples, the performance ap-

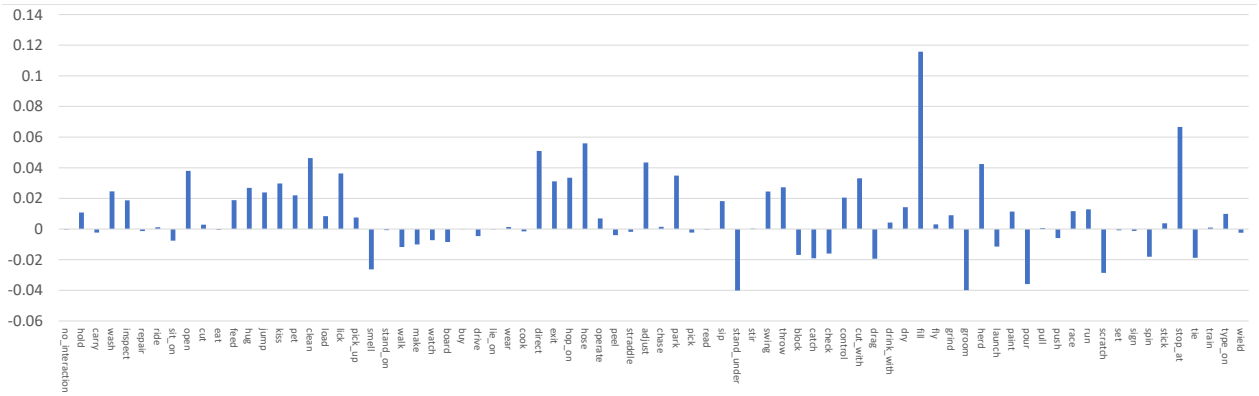


Figure 3. The improvement among the classes of verbs on HICO-DET. The verbs are sorted by the number of HOIs that the particular verb is related. The clear figure is in the directory of Compressed package.

Table 12. Ablation study of fabricator. Verb fabricator means we fabricate verb features.

| Method | Full | Rare | NonRare |
|---|---|---|---|
| FCL | **24.78** | **20.05** | **26.19** |
| FCL w/o noise | 24.22 | 19.23 | 25.72 |
| FCL w/o verb | 24.29 | 18.98 | 25.87 |
| verb fabricator | 23.93 | 17.10 | 25.97 |

| FCL | SP | ZS | Full | Rare | NonRare | Unseen |
|---|---|---|---|---|---|---|
| - | - | - | 21.07 | 14.11 | 23.15 | - |
| ✓ | - | - | 21.68 | 16.92 | 23.11 | - |
| ✓ | ✓ | - | **24.78** | **20.05** | **26.19** | - |
| - | - | ✓ | 15.29 | 14.45 | 17.85 | 8.27 |
| ✓ | - | ✓ | 16.82 | 16.57 | 18.17 | 12.94 |
| ✓ | ✓ | ✓ | **19.61** | **18.69** | **21.13** | **15.86** |

Table 13. Illustration of spatial pattern. SP means we use spatial pattern. ZS means zero-shot setting.

parently decreases to 23.93% in long-tailed HOI detection. This again illustrates that the verb feature is more complex and it is difficult to generate efficient verb features to facilitate HOI detection.

### 3.10. Additional Ablation Study

**Step-wise optimization**. We also provide the comparison between step-wise optimization and one-step optimization in unseen object HOI detection in Table 8.

**Hyper-Parameters**. We follow the hyper-parameters in [4] for $\lambda_1$ and $\lambda_2$. For $\lambda_3$, we provide the ablated experiment in Table 14 based on 0.5 because we think $L_{reg}$ is less important than $L_{CL}$.

**Fine-tune the network**. In the step-wise optimization, we fine-tune the whole FCL network in the last step. For a fair comparison, we also fine-tune our baseline after we train our network. Table 15 shows fine-tuning the network improves effectively the baseline. This is the reason why our baseline is strong. It might be because the initial learning 0.01 in our optimization is high.

Table 14. Illustration of ablated study on $\lambda_3$ in HICO-DET based on open long-tailed HOI detection (corresponding to Table 3 in paper).

| $\lambda_3$ | 0.1 | 0.3 | 0.5 |
|---|---|---|---|
| FCL | 19.30 | 19.61 | 19.10 |

Table 15. Ablation study of fine-tuning the network.

| Method | Full | Rare | NonRare |
|---|---|---|---|
| Baseline (w/o fine-tune) | 22.83 | 16.32 | 24.77 |
| Baseline | **23.35** | **17.08** | **25.22** |

## 4. Additional Qualitative Analysis

### 4.1. Object Representations

We analyze the real object features and fabricated object features in detail in Figure 4, 5 by selecting top 10 frequent classes in HICO-DET. 1) In Figure 4 (a) and Figure 5 (a), we find the fake object features of the same class are close to each other, while the features from different classes are separable although they might share the same verb. 2) Figure 4 (b) and Figure 5 (c) show features of different verbs slightly cluster together within each object class. **We can find there are outliers in some object classes because those outliers have different verbs**. 3) for unseen object ZSL, Figure 5 shows all fake object features of the same class are also closer to each other. Particularly, the unseen objects (red edge in row b) are also separable from others. 4) The Column 3 in Figure 4 and Figure 5 illustrate fake object features are still separable from its real objects of the same class. However, there are still some fabricated features are closer to it's corresponding real features (*e.g.* the dark blue class in Figure 4 and the jade-green class in Figure 5). We think

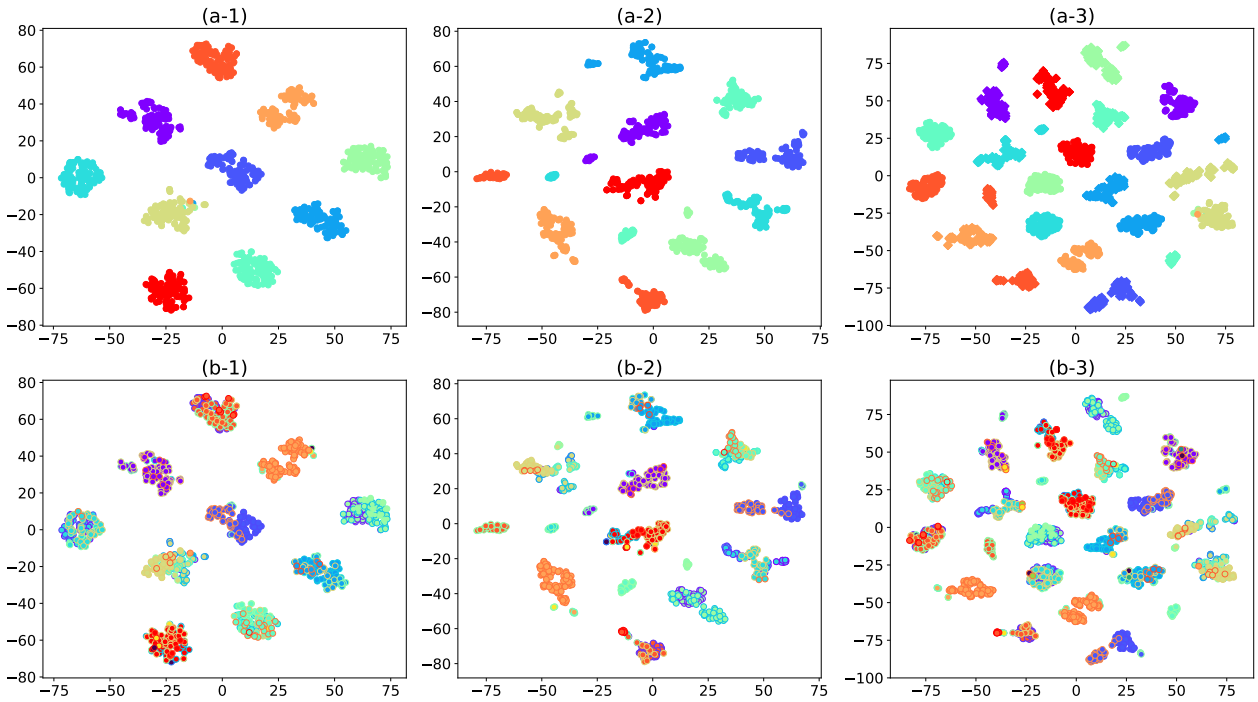


Figure 4. The illustration of real object representations, fabricated object representations and joint representations extracted from long-tailed HOI detection model. We select top 10 frequent object classes from HICO-DET training data. For each classes, we randomly select 100 instances. Column 1 is real object representations, Column 2 is fabricated object representations and Column 3 is the joint representations. In Column 3, diamond point means fabricated object representations. Raw a is the base t-SNE figure. In raw b, we label different verbs with different edges (color) in Raw b.

Column 3 in the two Figures also shows a future direction for fabricating objects, *i.e*. generate more realistic objects.

### 4.2. Primitive Features

Figure 6 illustrates verb features are apparently more difficult to distinguish. The verb representation is abstract and complicated. By contrast, object representations extracted from modern object detector are more discriminative. By comparing Figure 6 with the Figures in VCL [4], we can find the objects of FCL are more discriminative.

### 4.3. Qualitative Comparison

In Figure 7, we compare our baseline with our proposed method. Apparently, our proposed method efficiently detects rare categories, while the corresponding baseline can not. In fact, all the HOIs detected by our method in Figure 7 have less than five samples in training set which is much less than the rare setting (less than 10 samples).

### 4.4. Failure cases analysis

We provide some false positive results on Rare category in Figure 8. All failure cases can be separated into four groups: blurry image, wrong verb, wrong object, wrong match. If the image is blurry or has partial occlusion, it is hard to detection the interaction right. Besides, verb is usually hard to classify. Meanwhile, small objects also cause that the network detect object wrongly (*e.g*. the carrot in Figure 8). Lastly, even though the network can recognize action and object correctly, it also possibly mismatches the interaction. For example, in Figure 8, the women do not interact with the banana on the corner of the table.

## References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, pages 381–389. IEEE, 2018. 3

[2] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 4

[3] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 4

[4] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 4, 5, 6

[5] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. Transferable interactiveness prior for human-object interaction detection. In *CVPR*, 2019. 4

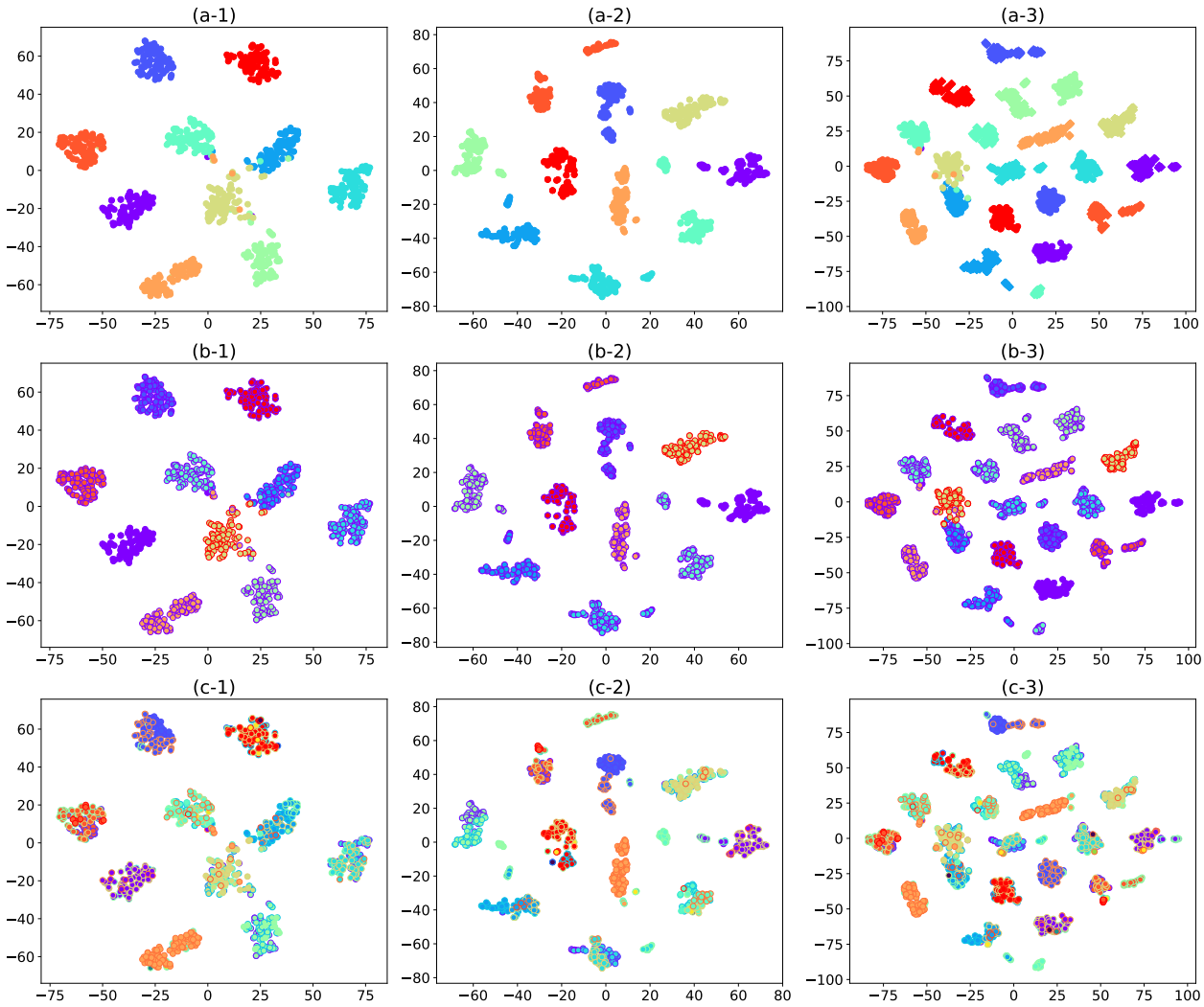


Figure 5. The illustration of real object representations, fabricated object representations and joint representations extracted from unseen object zero-shot model. Column 1 is real object representations, Column 2 is fabricated object representations and Column 3 is the joint representations. In Column 3, diamond point means fabricated object representations. Raw a is the base t-SNE figure. In raw b, we point out the unseen objects with red edge. In Raw c, we label different verbs with different edges (color).

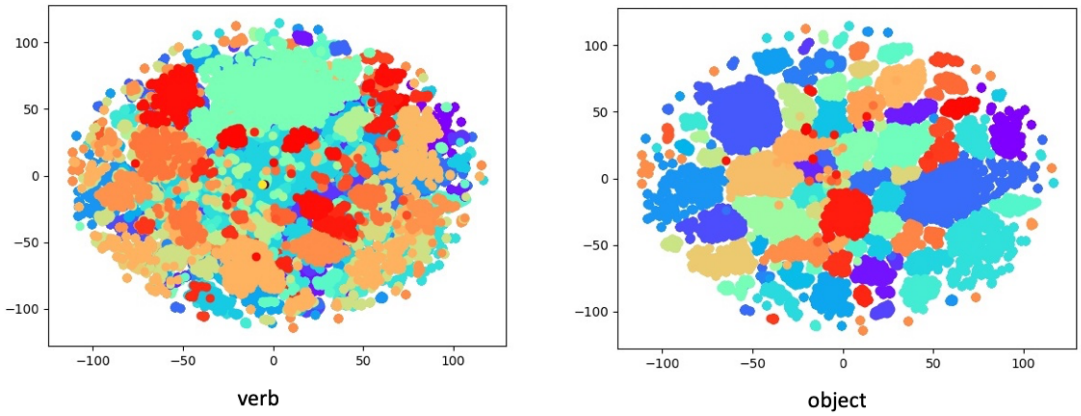


Figure 6. The comparison between verb features and object features.

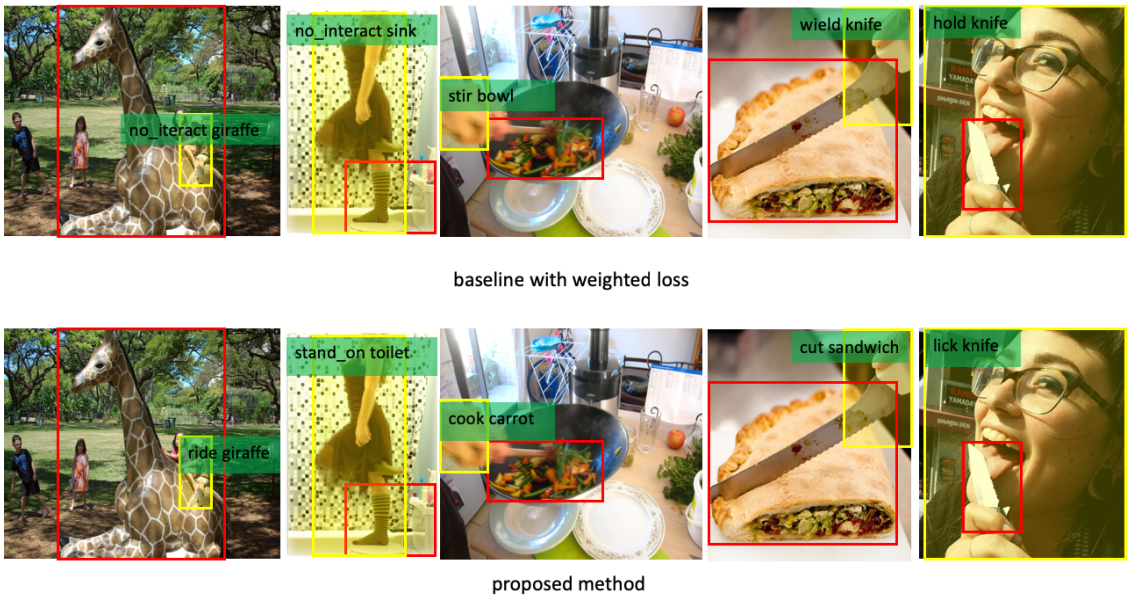


Figure 7. Visual Comparison between FCL and our baseline. The two models use same detector.

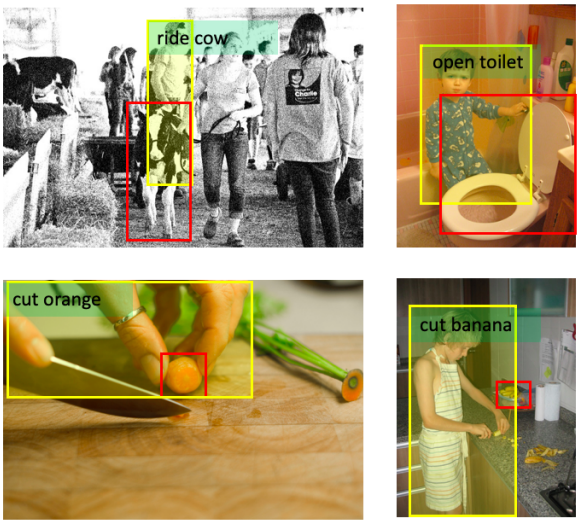


Figure 8. Illustration of failure cases.

[6] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869. Springer, 2016. 3

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 2, 4

[8] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, pages 9469–9478, 2019. 4

[9] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019. 3, 4

[10] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. On exploring undetermined relationships for visual relationship detection. In *CVPR*, pages 5128–5137, 2019. 3

[11] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017. 4