DRG: Dual Relation Graph for Human-Object Interaction Detection Supplementary Material

Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang

Virginia Tech
{chengao,jiaruixu,ylzou,jbhuang}@vt.edu

Overview

In this supplementary document, we provide additional experiments and sample visual results to complement the main manuscript.

1. We evaluate our method on HICO-DET under the zero-shot setting.
2. We provide the detailed per-class mAP on the V-COCO *test* set.
3. We conduct an additional evaluation with ResNet-50 backbone on the V-COCO dataset and the HICO-DET dataset.
4. We further validate our proposed method on a recently proposed dataset, Human-Centric Visual Relationship Detection dataset (HCVRD) [10].
5. We provide an error diagnose of the proposed DRG model using the V-COCO dataset.
6. We show an additional ablation study to further explore the model design space.
7. We present additional visual results of HOI detections on the V-COCO, HICO-DET, and HCVRD datasets.

1 Zero-shot setting evaluation

We evaluate our method on HICO-DET under the zero-shot setting to validate the effectiveness of the proposed spatial-semantic representation and dual relation graph for transferring knowledge among object classes. Several methods attempt to detect rare or unseen visual relations [1, 4, 8]. Such a task is crucial because it is challenging to collect sufficient training data for every possible visual relation. For example, during training we may see many examples where a person is riding an elephant, but we may have never seen or only seen few examples where a person is riding a horse. Thus, the model needs to understand the attribute similarity between elephant and horse, i.e., both elephant and horse are animals and can be ridden (with a similar spatial human-object relationship). At testing time, the model can still predict the *riding* action for *riding horse* even though no or few examples of a person riding a horse are available during training.

Existing methods address the zero-shot visual relationship learning via arithmetic operations between the word embeddings [8], leveraging a pre-defined graph

Table 1: **Zero-shot evaluation on HICO-DET *test* set.**

	Default (Full)	
Method	Unseen (120)	Seen (480)
Bansal et al. [1]	11.31 ± 1.03	12.74 ± 0.34
Ours	16.72 ± 1.87	16.84 ± 0.58

according to the external knowledge base [4], or performing data augmentation of semantically similar objects [1]. While our method does not explicitly address the zero-shot setting, the proposed abstract spatial-semantic representation allows the network to effectively transfer knowledge among different interactions and recognize unseen interaction categories.

To evaluate the performance under zero-shot setting, we use the five random splits of 120 unseen and 480 seen relation triplets (with a total of 600 HOI categories) provided by the authors in [1]. Every object within the 120 categories is ensured to occur at least once in the remaining 480 relationships. We report the average mAP and variance over the five splits on the HICO-DET dataset in Table 1.

Our proposed model leads to a sizable absolute gain of 5.41 mAP compared to Bansal et al. [1].

2 Per-class AP_{role} on VCOCO

We show the detailed per-class AP_{role} for individual action classes in Table 2. The proposed DRG model performs particularly well on actions that require a distinctive object such as *skateboard* (85.9%), *surf* (80.6%), *read* (39.5%) and *talk on phone* (55.9%). We attribute this improvement to the use of the proposed spatial-semantic representation. We also observe a sizable performance gain for action classes that are often associated with distinctive scenes, e.g., *cut obj* (42.4%), *throw* (52.1%) and *ride* (69.9%). We believe that the improvement for these interaction classes comes from our DRG, which leverages the contextual information of the scene.

3 Performance under different feature backbones

In this section, we report the additional quantitative results in terms of AP_{role} with the ResNet-50 backbone on V-COCO in Table 3 and HICO-DET in Table 4. We show that adding Feature Pyramid Network (FPN) does not improve the performance by a large margin. Comparing our results using ResNet-50 with other methods using ResNet-50 in Table 1 and Table 2 in the main paper, our method still compares favorably against state-of-the-arts.

Table 2: **Detailed results on V-COCO *test* set.** The best performance is in **bold** and the second best is underscored.

	InteractNet [3] ResNet-50-FPN	iCAN [2] ResNet-50	$RP_{T2}C_D{}^{\star}$ [5] ResNet-50	Ours ResNet-50-FPN
carry	33.1	32.0	40.8	**41.3**
catch	42.5	47.6	48.4	**49.3**
drink	33.8	32.2	34.4	**35.5**
hold	26.4	29.1	37.3	**40.7**
jump	45.1	51.5	53.8	**54.2**
kick	69.4	66.9	66.3	**69.8**
lay	21.0	22.4	**29.6**	26.1
look	20.2	26.5	32.3	**35.5**
read	23.9	30.7	29.9	**39.5**
ride	55.2	61.9	66.3	**69.9**
sit	19.9	26.0	**31.6**	31.4
skateboard	75.5	79.4	83.4	**85.9**
ski	36.5	41.7	50.0	**52.9**
snowboard	63.9	74.4	71.6	**75.6**
surf	65.7	77.2	79.7	**80.6**
talk-on-phone	31.8	52.8	53.6	**55.9**
throw	40.4	40.6	43.3	**52.1**
work-on-computer	57.3	56.3	**65.5**	64.3
cut (object)	23.0	34.8	40.1	**42.4**
cut (instrument)	36.4	37.2	**41.6**	39.9
eat (object)	32.4	37.7	39.1	**41.9**
eat (instrument)	2.0	8.3	**9.4**	5.9
hit (object)	**62.3**	46.1	52.6	54.7
hit (instrument)	43.3	74.1	74.4	**77.6**
mean AP role	40.0	45.3	49.0	**51.0**

Table 3: **Results with different feature backbone on the V-COCO *test* set.**

Method	Use human pose	Feature backbone	AP_{role}
Ours	-	ResNet-50	50.7
Ours	-	ResNet-50-FPN	51.0

Table 4: **Results with different feature backbone on the HICO-DET *test* set.** For the object detector, "COCO" means that the detector is trained on COCO, while "HICO-DET" means that the detector is first pre-trained on COCO and then further fine-tuned on HICO-DET.

				Default			Known Object		
Method	Detector	Use human pose	Feature backbone	Full	Rare	Non Rare	Full	Rare	Non Rare
Ours	COCO	-	ResNet-50	18.87	16.90	19.45	22.83	20.61	23.49
Ours	COCO	-	ResNet-50-FPN	19.26	17.74	19.71	23.40	21.75	23.89
Ours	HICO-DET	-	ResNet-50	24.05	18.72	25.65	27.39	22.01	28.99
Ours	HICO-DET	-	ResNet-50-FPN	24.53	19.47	26.04	27.98	23.11	29.43

4 Experimental results on the HCVRD dataset.

HCVRD [10] is a recently proposed dataset for evaluating human-centric relationship detection. The dataset contains 52,855 images with 1,824 object categories and 927 predicates. We adopt the Recall@50 and Recall@100 as the evaluation metrics for the HCVRD dataset, following the setting of the original paper [10]. As the interactions in the HCVRD dataset are *not* exhaustively annotated, we are not able to use mean Average Precision (mAP) as the metric. For example, a correct interaction prediction may be penalized if this particular ground truth annotation is not given.

Unlike the V-COCO and HICO-DET dataset, which shares the same object categories as the MS-COCO dataset, the HCVRD dataset contains 1,824 object categories and has a severe long-tail distribution. In order to exclude the factor of object detection, here we focus on the particular task called *predicate detection*. In this task, the human bounding boxes, object bounding boxes, and the object category are given a priori. The given object bounding boxes allow us to focus on evaluating the accuracy of recognizing an action/interaction of our model and having a fair comparison with other methods.

We report the quantitative results on the HCVRD dataset in Table 5 under two settings: top-1 and top-3 accuracy. Under the top-3 setting, we choose the predictions with the top 3 scores. For each H-O pair, if the ground truth prediction belongs to any of the top 3 predictions, then we consider the prediction as correct. Compared to the best competing approach [9], our method leads to an absolute gain of 6.5 R@50 (a relative improvement of 17.3%) under the top-1 setting. We achieve an absolute gain of 14.2 R@50 under the top-3 setting.

Table 5: **Performance comparison with the state-of-the-art on the HCVRD *test* set.** The performance is measured by Recall@50 (R@50) and Recall@100 (R@100) under two different settings, namely top-1 and top-3. The best performance is in **bold** and the second best is underscored. Character ⋆ indicates that the results are reported by [9].

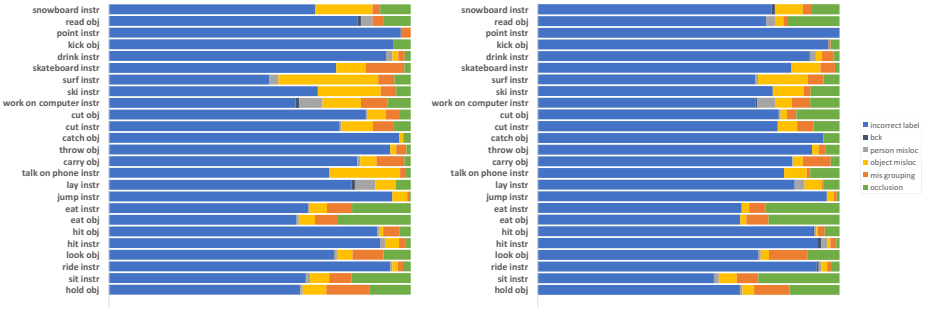
Method	Feature backbone	Predicate Det.			
		R@50		R@100	
		top-1	top-3	top-1	top-3
Multilabel [10]	VGG-16	0.9	2.8	0.9	2.8
JointCNN [10]	VGG-16	2.7	7.4	2.7	7.4
SeparateCNN [10]	VGG-16	29.0	44.4	29.0	45.9
Webly-Supervised [10]	VGG-16	31.1	47.7	31.1	49.0
iCAN⋆ [2]	ResNet-50	33.8	48.9	33.8	49.4
Wang et al. [9]	ResNet-50	37.1	51.3	37.1	51.9
Ours (S-S only)	-	38.1	61.8	38.1	62.2
Ours	ResNet-50-FPN	**43.6**	**65.5**	**43.6**	**65.9**

Using the proposed spatial-semantic stream only (i.e., without using appearance features from human/object stream), we obtain a competitive performance of 38.1 R@50 under the top-1 setting. These results emphasize the contribution of contextual reasoning using abstract spatial-semantic representation.

5 Error analysis

V-COCO. We diagnose the detection errors made by our network using the diagnostic tool provided by iCAN [2]. Figure 1 shows the distribution of the incorrect detections for each action class. We compare the error categories distribution between iCAN and our proposed DRG. The following six types of error are considered:

1. **incorrect label**: the detected person is correctly localized around a ground truth person box, i.e., IoU greater than 0.5, but is incorrectly predicted to perform an action.
2. **bck**: the detected person is incorrectly localized, i.e., IoU less than 0.1 with any of the ground truth persons.
3. **person misloc**: the detected person is mislocalized, i.e., IoU between 0.1 and 0.5 with a ground truth person.
4. **object misloc**: the detected person is correctly localized, and the detected object is mislocalized, i.e., IoU between 0.1 and 0.5 with a ground truth object.
5. **mis grouping**: the detected person is correctly localized, and the detected object is not associated with the ground truth person (i.e., IoU less than 0.1).



(a) iCAN (b) DRG

Fig. 1: **Comparison of the distribution of the error categories between iCAN and our proposed DRG.** We show the distribution of the incorrect detections for each action class. **'incorrect label'** refers to the detected person is predicted to perform an action which he does not. **'bck'** indicates the object detector fails to localize the person (IoU with any ground truth person is less than 0.1). **'person misloc'** means the detected person is not localized well (IoU is between 0.1 and 0.5 with a ground truth person) regardless of the related object. **'object misloc'** refers to the detected object is not localized well while the related person is successfully localized. **'mis-grouping'** indicates the person is successfully localized, but the network fails to match the person to the correct related object (IoU between the ground truth associated object and the predicted object is less than 0.1). **'occlusion'** means we incorrectly associate an object with a correct detected person, while the object is not annotated in the ground truth due to occlusions. **Discussion:** our proposed DRG helps most with the mis-grouping. By leveraging the proposed human-centric subgraph and object-centric subgraph, we can enhance the correct HOIs and suppress the unrelated pairs. As a result, the mis-grouping error is reduced, especially for the action classes that humans and objects have a typical spatial relationship, e.g., *hit obj*, *hit instr*, and *skateboard instr*.

6. **occlusion**: the detected person is correctly localized, and an object is incorrectly associated with the person while there is no ground truth object associated with this person (e.g., due to occlusion).

Our proposed DRG helps most with the mis-grouping. By leveraging the proposed human-centric subgraph and object-centric subgraph, we can enhance the correct HOIs and suppress the unrelated HOIs. As a result, the mis-grouping error is significantly reduced, especially for the action classes that humans and objects have a typical spatial relationship, e.g., *hit obj*, *hit instr*, and *skateboard instr*.

HICO-DET. We notice that the human-object interaction labels in the HICO-DET dataset are not fully annotated. In Figure 2, we show an example that our model correctly detects the HOIs, however, these correct detections were not annotated as ground truth in HICO-DET. In the first example, three persons are

sitting on the bench. However, only one of them is annotated as *human sitting on the bench*. While our method correctly detects the interactions between all the three persons and the bench, our method is penalized because the dataset is not fully annotated. More examples are shown in Figure 2.

6 Additional ablation study

In this section, we examine additional design choices using the V-COCO *val* split.

Exploring different object category representation. We compare different ways to represent object categories in our spatial-semantic representation. Note that here we do *not* apply the proposed DRG in this experiment. In Figure 3 we show the t-SNE [6] visualization of each object category representation. We take the 1024D feature before the final classification layer in the spatial stream and show the t-SNE [6] visualization. Comparing the "spatial only" and the other three, we observe that including the information of the object of interaction help improve the separation. The proposed approach using fastText [7] shows a better separation between action classes than that using object appearance features or one-hot vector encoding of object categories.

More iteration of feature aggregation. In Figure 4, we show the t-SNE [6] visualization of the proposed spatial-semantic representation on both human-centric subgraph and object-centric subgraph with a different iteration of feature aggregation. As we use more iterations of feature aggregation, the action classes become increasingly disentangled.

7 Additional visual results

In this section we show additional HOI detection results on V-COCO in Figure 5, HICO-DET in Figure 6, HCRVD in Figure 7. We also show more examples of detecting persons simultaneously performing multiple actions and associating the detected object instances to their semantic roles of the action in Figure 8.

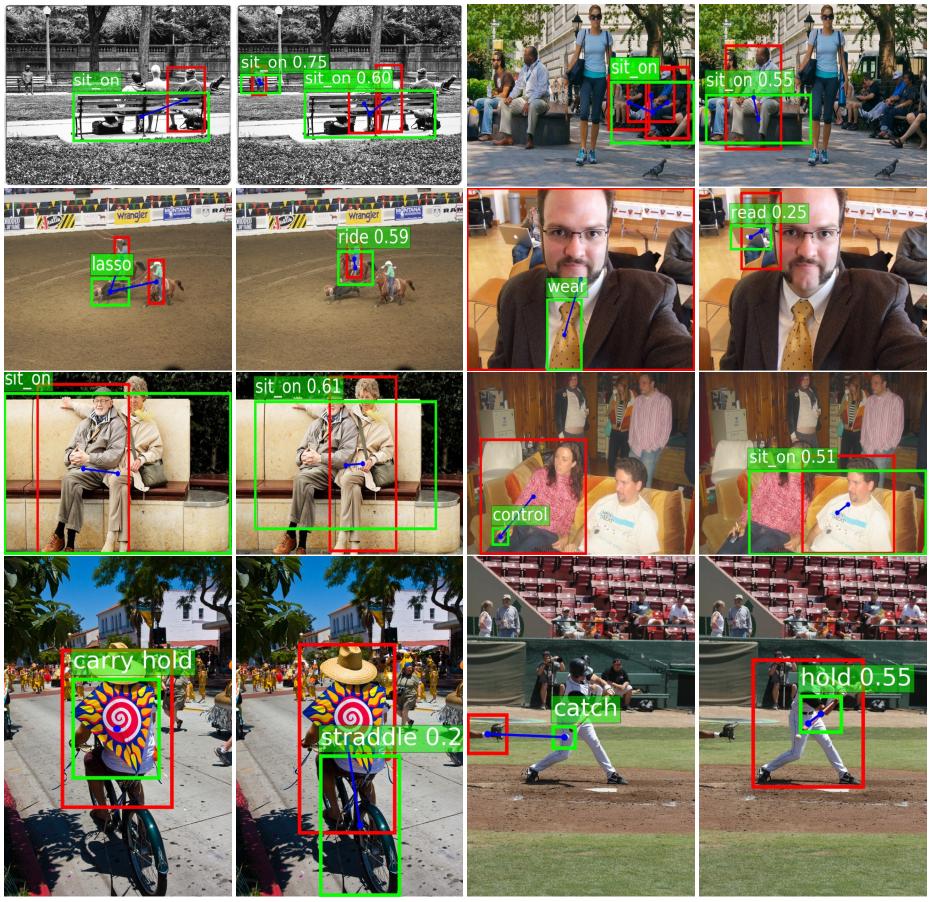


Fig. 2: **HICO-DET is not fully annotated.** The first and third column shows *all* the ground truth annotations. The second and last column shows the HOI detection from our model. While our method correctly detects the HOIs, we are penalized because the dataset is not fully annotated.

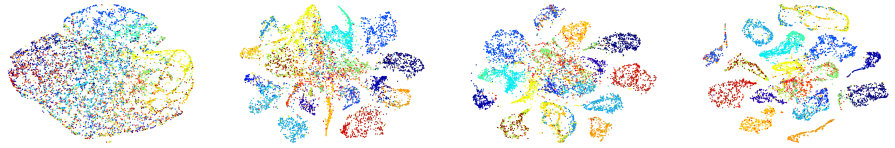


Fig. 3: **t-SNE visualization of different object category representation.**

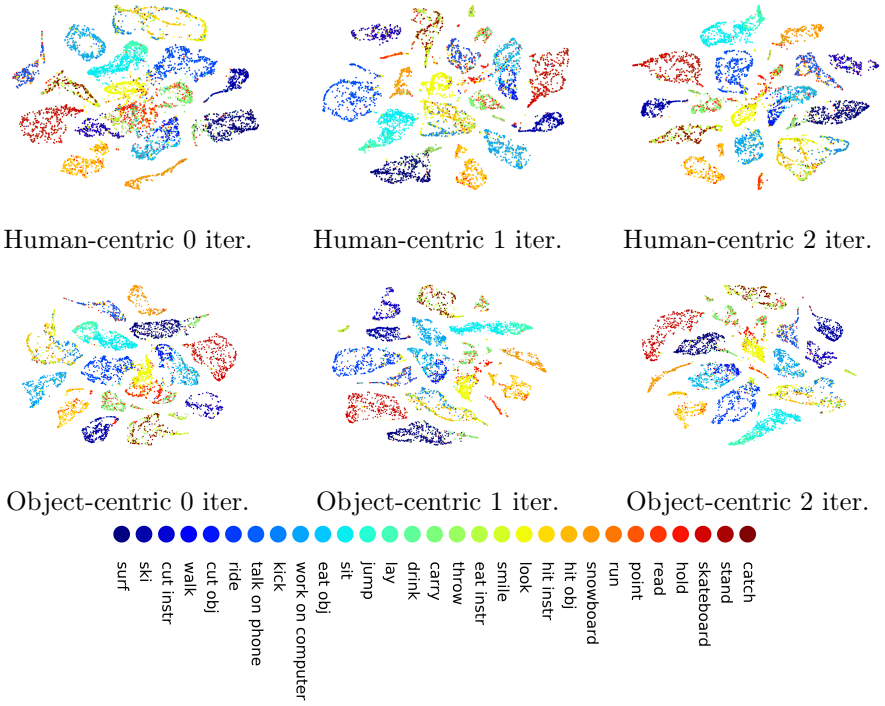


Fig. 4: **t-SNE visualization of spatial-semantic representation with different iterations of feature aggregation.**

Fig. 5: **HOI detection results on V-COCO *test* set.**

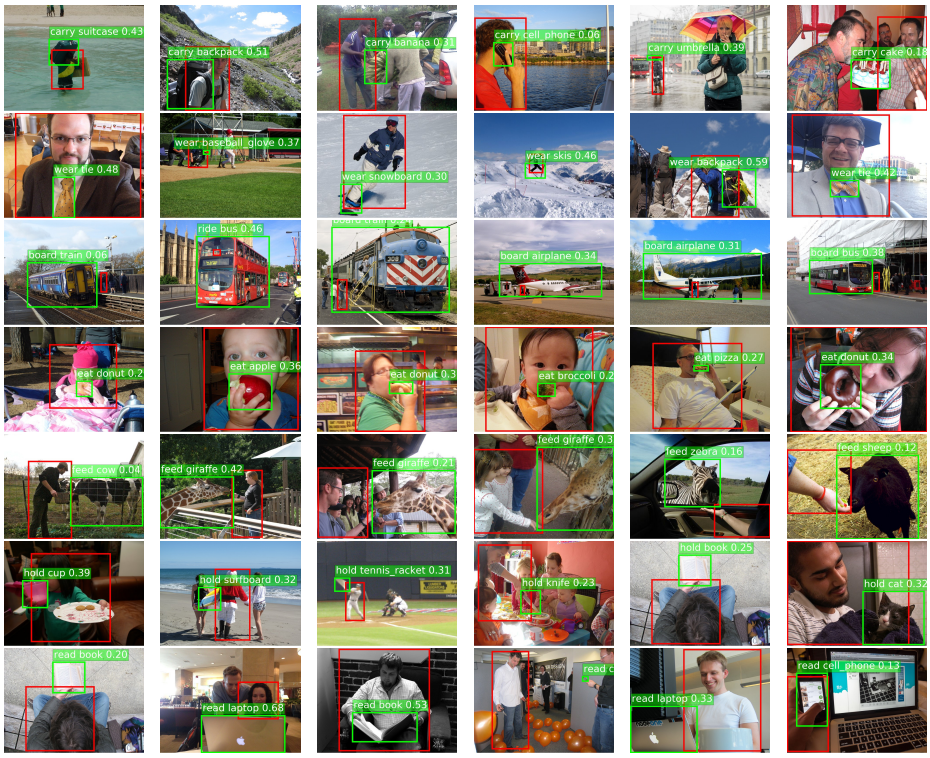


Fig. 6: **HOI detection results on HICO-DET *test* set.**

Fig. 7: **HOI detection results on HCVRD *test* set.**

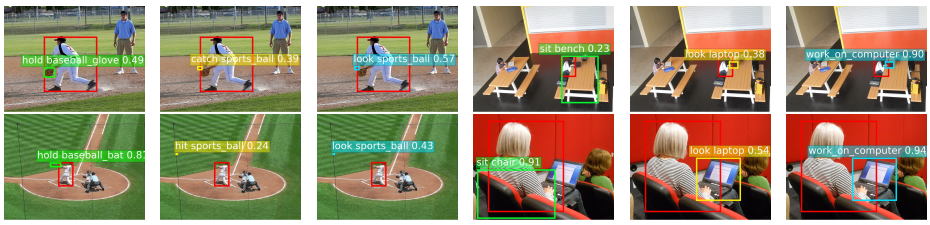


Fig. 8: **Detecting multiple actions.** Our model can detect a person doing different actions with different objects.

References

1. Bansal, A., Rambhatla, S.S., Shrivastava, A., Chellappa, R.: Detecting human-object interactions via functional generalization. In: AAAI (2020)
2. Gao, C., Zou, Y., Huang, J.B.: iCAN: Instance-centric attention network for human-object interaction detection. In: BMVC (2018)
3. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: CVPR (2018)
4. Kato, K., Li, Y., Gupta, A.: Compositional learning for human object interaction. In: ECCV (2018)
5. Li, Y.L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.S., Wang, Y.F., Lu, C.: Transferable interactiveness prior for human-object interaction detection. In: CVPR (2019)
6. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. JMLR (2008)
7. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. In: LREC (2018)
8. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Detecting rare visual relations using analogies. In: ICCV (2019)
9. Wang, T., Anwer, R.M., Khan, M.H., Khan, F.S., Pang, Y., Shao, L., Laaksonen, J.: Deep contextual attention for human-object interaction detection. In: ICCV (2019)
10. Zhuang, B., Wu, Q., Shen, C., Reid, I.D., van den Hengel, A.: Hcvrd: A benchmark for large-scale human-centered visual relationship detection. In: AAAI (2018)