

# Three-stream network with context convolution module for human–object interaction detection

Thomhert S. Siadari<sup>1,2</sup>  | Mikyong Han<sup>2</sup> | Hyunjin Yoon<sup>1,2</sup>

<sup>1</sup>ICT Major of ETRI School, University of Science and Technology, Daejeon, Rep. of Korea

<sup>2</sup>City and Transportation ICT Research Department, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea

## Correspondence

Hyunjin Yoon, City and Transportation ICT Research Department, Electronics and Telecommunications Research Institute; and ICT Major of ETRI School, University of Science and Technology, Daejeon, Rep. of Korea.

Email: hjyoon73@etri.re.kr

## Funding information

This research was supported by the Cross-Ministry Giga KOREA Project grant of the Korea government (MSIT), Republic of Korea (No.GK19P0600, Development and Demonstration of Smart City Service over 5G Network).

Human–object interaction (HOI) detection is a popular computer vision task that detects interactions between humans and objects. This task can be useful in many applications that require a deeper understanding of semantic scenes. Current HOI detection networks typically consist of a feature extractor followed by detection layers comprising small filters (eg,  $1 \times 1$  or  $3 \times 3$ ). Although small filters can capture local spatial features with a few parameters, they fail to capture larger context information relevant for recognizing interactions between humans and distant objects owing to their small receptive regions. Hence, we herein propose a three-stream HOI detection network that employs a context convolution module (CCM) in each stream branch. The CCM can capture larger contexts from input feature maps by adopting combinations of large separable convolution layers and residual-based convolution layers without increasing the number of parameters by using fewer large separable filters. We evaluate our HOI detection method using two benchmark datasets, V-COCO and HICO-DET, and demonstrate its state-of-the-art performance.

## KEYWORDS

context convolution module, deep learning, HOI detection, human–object interactions, three-stream network

## 1 | INTRODUCTION

Human–object interaction (HOI) detection aims to detect human and object locations and classify their interactions at the instance level (eg, a person riding a bike, carrying a backpack, and throwing a frisbee), which can be formulated as detecting a triplet (human, action, and object) [1–5]. This task is beneficial to many applications that require a deeper understanding of semantic scenes, such as video surveillance [6–9] and visual question answering [10].

Recently, HOI detection methods based on deep learning have improved considerably. Gkioxari and others [11] proposed a model to predict HOI detection based on an object

location density map from the appearance of a detected human. Chao and others [12] introduced a multistream model combining visual features and spatial locations to detect HOIs from images. Shen and others [13] scaled HOI recognition to the long tail of categories through the zero-shot learning method to predict unknown verbs and objects. Gao and others [14] exploited an instance-centric attention module to selectively aggregate features relevant for recognizing HOIs and to enhance information from the region of interest. Previous studies typically employed a feature extractor followed by detection layers comprising convolution layers with small filters (eg,  $1 \times 1$  or  $3 \times 3$ ) because small filters can capture local spatial features with only a few parameters. However, because of their small

receptive regions, small filters fail to capture larger context information from input feature maps, which can be useful for recognizing interactions between humans and distant objects. Figure 1A shows that the instance-centric attention network (iCAN) [14] method fails to detect the HOIs of a person on the left wearing a yellow shirt, who is throwing a frisbee, and a person on the right wearing a sky blue shirt, who is looking at another person.

Inspired by [15] wherein large separable convolution layers were used for semantic segmentation to improve classification and localization on large-scale objects, we herein propose a three-stream HOI detection network that employs a context convolution module (CCM) in each stream branch. The CCM adopts combinations of large separable convolution layers and residual-based convolution layers to capture larger context information from input feature maps. We used fewer filters in the CCM to reduce the number of network parameters increased by larger filters employed in the CCM. Furthermore, we conducted extensive experiments on two benchmark datasets, V-COCO [5] and HICO-DET [12], and demonstrated that our method outperformed existing HOI detection methods.

In summary, our main contributions are two-fold:

1. Considering the importance of larger spatial features required for detecting interactions of humans and distant objects, we propose using a CCM in each stream branch of our three-stream HOI detection network to capture larger contexts from input feature maps by adopting combinations of large separable convolution layers and residual-based convolution layers without increasing network parameters.

2. We performed extensive experiments on two benchmark datasets and demonstrated that our method outperformed existing HOI detection methods.

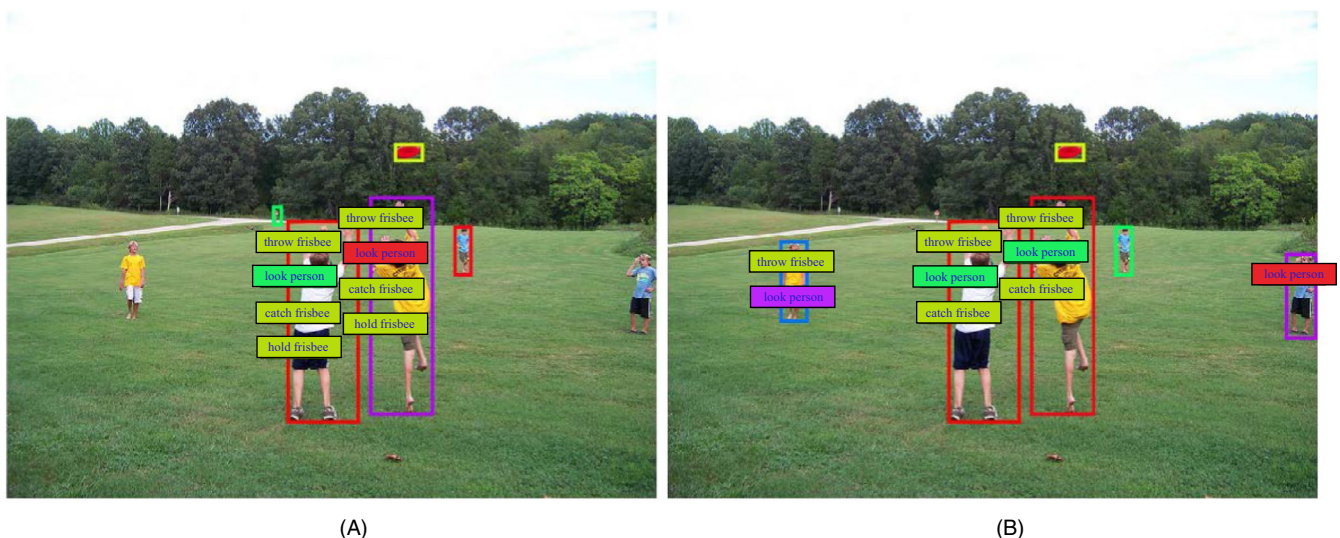
The remainder of the article is organized as follows: Section 2 presents the related work regarding HOI detection. Section 3 describes the details of our proposed method and also elaborates our experiments and discusses our results. Section 4 concludes the paper.

## 2 | RELATED WORK

This section reviews prior works related to visual relationship detection, context embedding, and HOI detection.

### 2.1 | Visual relationship detection

Visual relationship detection involves detecting objects in images and classifying the relationships between them [16–23]. Lu and others [17] proposed a model that learned the visual appearance of objects and predicates; additionally, they combined the model with language priors from semantic embedding. Zhang and others [20] introduced a new model that used end-to-end and fully convolution networks comprising an object detector, differentiable feature extraction layer, and visual translation embedding layer to classify visual relationships. Although these studies are interesting, our focus is on HOI detection, which is a human-centric problem, to detect action interactions between humans and objects.



**FIGURE 1** (A) Results using iCAN [14]. (B) Results using our method. The iCAN method fails to detect the HOIs of a person wearing a yellow shirt, who is throwing a frisbee (left), and a person wearing a sky blue shirt, who is looking at another person (right)

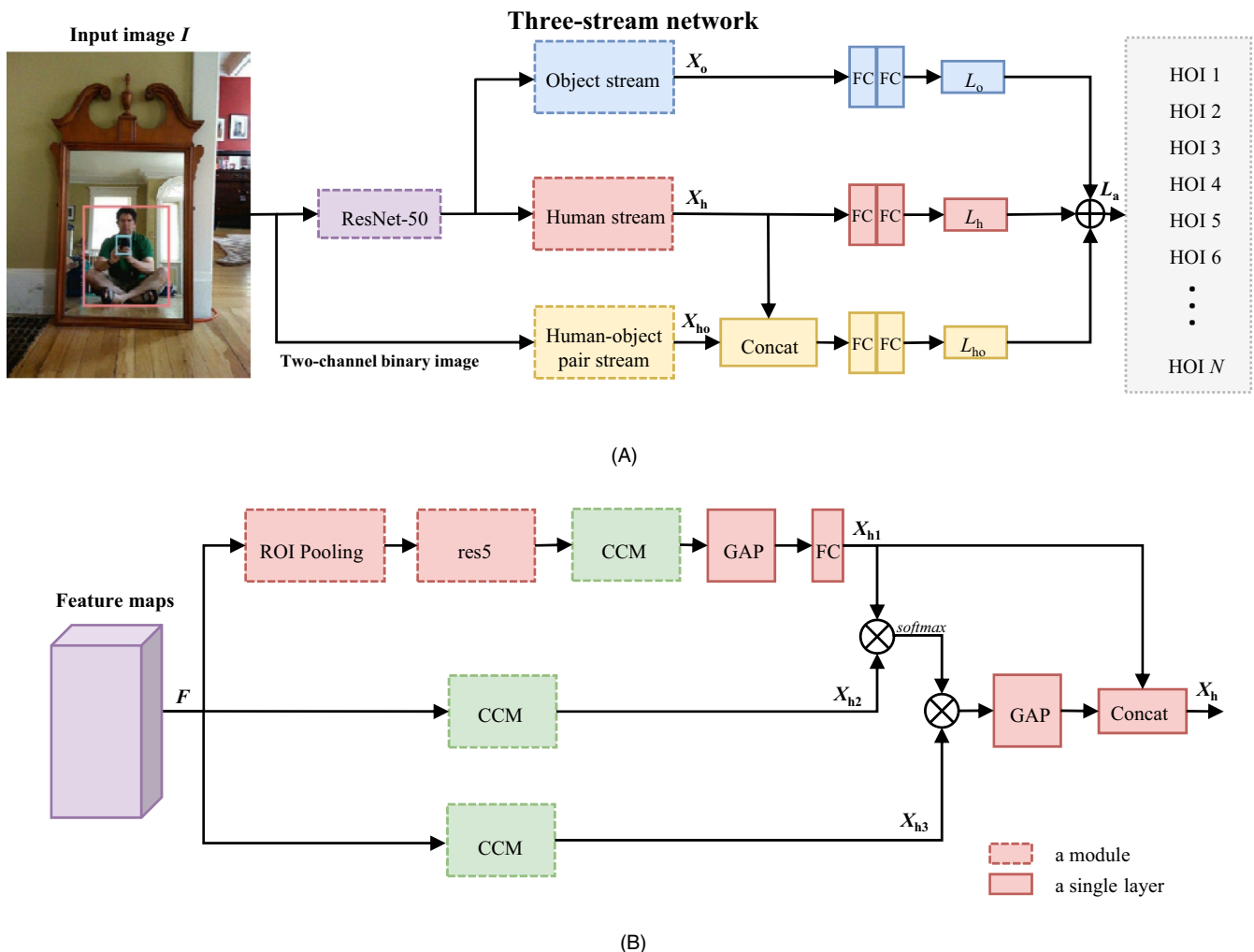
## 2.2 | Context embedding

Context embedding is typically used in semantic segmentation [24–26]. Zoom-out [24] proposed a model that used multilevel and zoom-out feature maps; additionally, they trained multilayer neural networks to conduct pixel-wise classification. Parse-Net [25] introduced a model that adds a global pooling branch into deep convolution networks to extract context information. Dilated-Net [26] presented a model that systematically aggregates multiscale context information using dilated convolutions; this helps to extend the receptive field exponentially without losing coverage.

## 2.3 | HOI detection

HOI detection focuses on detecting humans and surrounding objects and then predicts the interactions between them [1–5]. Recently, several large-scale datasets, such as HICO-DET [12] and V-COCO [5], have been proposed for

the investigation of HOI detection by providing bounding boxes for humans and objects along with the corresponding action interactions. Gkioxari and others [11] proposed a model to predict HOI detection based on an object location density map from the appearance of a detected human. Chao and others [12] introduced a multistream model combining visual features and spatial locations to detect HOIs from images. Shen and others [13] scaled HOI recognition to the long tail of categories through the zero-shot learning method to predict unknown verbs and objects. Gao and others [14] exploited a new attention network by highlighting the multistream instance features for recognizing HOIs and enhanced the information from the region of interest. Unlike our model, all previous studies employed a feature extractor followed by detection layers comprising convolution layers with small filters (eg,  $1 \times 1$  or  $3 \times 3$ ) but failed to capture larger context information relevant for recognizing interactions between humans and distant objects. We herein propose a model that manages to capture larger contexts from input feature maps.



**FIGURE 2** (A) Overview of our HOI detection network. (B) Human stream branch as part of the three-stream network. GAP and FC stand for global average pooling and fully connected layer, respectively

### 3 | PROPOSED METHOD

#### 3.1 | Overview

The architecture of our proposed HOI detection network is shown in Figure 2A. It comprises ResNet-50 [27] as a feature extractor and a three-stream network comprising object stream, human stream, and human–object pair stream branches. The object stream branch exploits instances detected as objects to which humans intend to interact, while the human stream branch focuses on the human, an actor who interacts with objects. The human and object stream branches have the same principal architecture, as shown in Figure 2B. The human–object pair stream branch learns rich feature embeddings to produce a spatial relationship between a human and an object. Each branch of the three-stream network employs a CCM, as shown in Figure 3, to capture larger contexts.

The HOI detection task was conducted as follows: given an input image  $I$ , we first detected human and object candidates from the image using Detectron [28] and retained only the bounding box information for instances where the confidence score was more than a specific threshold. Subsequently, we used ResNet-50 [27] to extract feature maps that will be used for the human and object stream branches only. The human–object pair stream used a two-channel binary image from the input image. Each branch calculated its action score, but the final HOI score was the element-wise addition of scores from all branches.

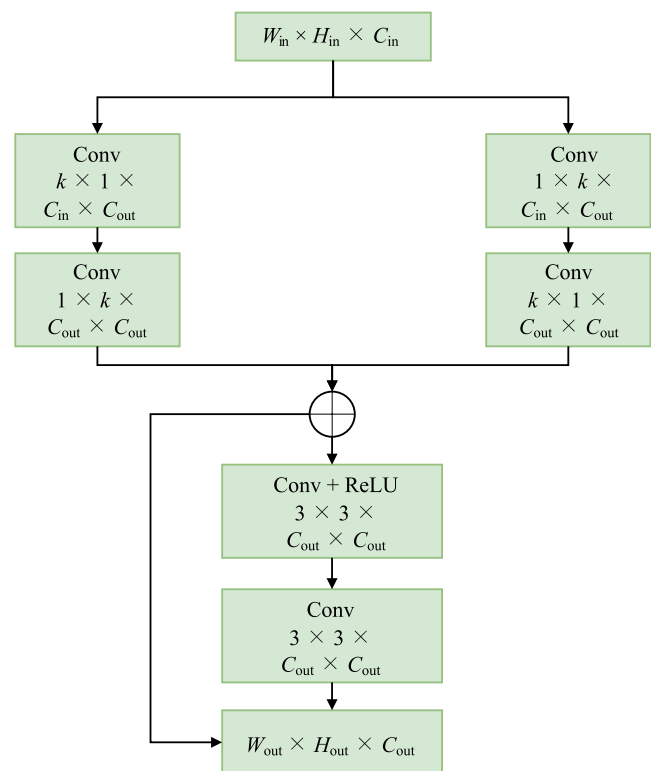
#### 3.2 | Three-stream network

Our three-stream network was adopted from [14]. The three-stream network uses two inputs, feature maps  $F$  extracted by ResNet-50 for the human and object streams, and a two-channel binary image for the human–object pair stream. Within the human or object stream, an ROI pooling, residual block (res5), CCM, average pooling layer (GAP), and fully connected (FC) layer were employed. In the human stream branch, ROI pooling crops the feature maps of human instances, while in the object stream branch, ROI pooling crops the feature maps of object instances.  $X_{h1}$  is the feature map generated from ROI pooling, res5, CCM, GAP, and FC, while  $X_{h2}$  and  $X_{h3}$  are feature maps generated from two CCMs. Intuitively,  $X_{h1}$  is a local feature with small receptive fields focusing on human or object instances, and  $X_{h2}$  and  $X_{h3}$  are global features with larger receptive fields from an input image, as described in Figure 2B. An element-wise multiplication of  $X_{h1}$  and  $X_{h2}$  followed by softmax was used to produce an attention map for target instances. The final feature maps that represent the human  $X_h$  is a concatenation of local and global features.

The human–object pair stream comprises two convolution layers followed by the CCM. To extract the spatial features of the human–object pair from an image, we converted an input image into dual spatial masks by encoding a union of two types of bounding boxes, that is a human and an object. For example, given an input image that contains a pair of human and object bounding boxes, as shown in Figure 2A, we represent the spatial relationship between these two using a two-channel binary image of size  $64 \times 64$ . The first channel represents the human mask, and the second channel the object mask.

#### 3.3 | CCM

Motivated by [15] where large separable convolution layers were used for semantic segmentation to improve classification and localization on large-scale objects, we herein propose a new module called the CCM to capture larger context information relevant to interactions between humans and distant objects from input feature maps. The CCM adopts combinations of large separable and residual-based convolution layers, as shown in Figure 3. The first large separable convolution layers (on the left) comprises convolution layers with  $k \times 1$  and  $1 \times k$ , while the second separable convolution



**FIGURE 3** Context convolution module.  $W_{in}$ ,  $H_{in}$ ,  $C_{in}$  are the width, height, and channels of the input feature maps, respectively, while  $W_{out}$ ,  $H_{out}$ ,  $C_{out}$  are the width, height, and channels of the output feature maps from the CCM

layer (on the right) comprises convolution layers with  $1 \times k$  and  $k \times 1$ . Subsequently, we conducted an element-wise addition for the outputs of these two separable convolution layers followed by the residual-based convolution layers of small kernel size. It is noteworthy that the third layer was employed with a convolution layer containing a rectified linear unit. The other layers were employed with a linear activation function.

### 3.4 | Training

The training of our three-stream HOI detection network is defined as a multitask learning that jointly trains all stream branches with a single loss. We applied a sigmoid binary classifier for each action category and minimized the cross-entropy loss between the predicted score and the ground truth. We considered HOI detection as a multilabel classification problem because a person may be involved in more than an action interaction. Each action was independent and not mutually exclusive. The overall loss is defined as a weighted sum of three branch losses, formulated as follows:

$$L_a = w_h * L_h + w_o * L_o + w_{ho} * L_{ho}, \quad (1)$$

where  $L_h$ ,  $L_o$ , and  $L_{ho}$  are the action classification loss from the human stream, object stream, and human–object pair stream branches, respectively.  $w_h$ ,  $w_o$ , and  $w_{ho}$  are weights applied to each corresponding branch loss and are empirically set as  $w_h = 2$ ,  $w_o = 1$ , and  $w_{ho} = 1$ , respectively. Therefore,  $L_a = 2 * L_h + L_o + L_{ho}$ . Additionally, the human stream was computed over 16 bounding boxes at the most, which were predicted as the human instances. The loss for the human–object pair stream was only computed on positive triplets.

### 3.5 | Inference

During inference, we calculated the triplet of the HOI detection scores for the human  $S_h$ , object  $S_o$ , and human–object pair  $S_{ho}$ . The total score  $S_a$  is computed as

$$S_a = S_h + S_o + S_{ho}. \quad (2)$$

First, we computed the HOI detection scores for the human and object streams. Subsequently, we calculated the HOI detection scores for every candidate human–object pair in the human–object pair stream branch. Calculating every candidate human–object pair is intractable. Therefore, we implemented predefined relevant object categories as prior knowledge (eg, pizza is irrelevant to the action ride but relevant to the action “eat”) and filtered out irrelevant detected objects to particular actions for each human–object

pair, as in [5]. Subsequently, we selected objects that maximized the triplet score  $S_a$  within each relevant category that to be higher than the threshold. With this selection, we obtained a triplet score of the human, action, and object. For the final output, we show the bounding boxes of the human and interacted object, along with the respective HOI actions. For HOI actions that do not associate with any objects (eg, walk, smile), we only used the action score from the human stream network.

## 4 | EXPERIMENTS

### 4.1 | Dataset and metrics

#### 4.1.1 | V-COCO

The popular benchmark V-COCO dataset [5] is a subset of COCO [29] that comprises 10 346 images including 2533 images for training, 2867 for validating, 4946 for testing, and 16 199 person instances. Each person in an image has an average of 2.9 actions categories. V-COCO has 26 action class annotations. Cut, eat, and hit classes are labeled with two different objects: instrument and direct object. We treat these three actions as six different action categories [5].

#### 4.1.2 | HICO-DET

HICO-DET [12] is an extension of the previous dataset called HICO [30], where multiple detection annotations are provided in an image. This dataset comprises 47 774 images annotated with 600 HOI categories (eg, ride-motorcycle, ride-horse). From the HOI categories, it has 117 action labels and 80 objects with an average of 6.5 actions on each object. As each image may have multiple HOI labels, the total annotation for HICO-DET reaches 150K annotations. The training data were set to be 80% of the dataset with 38 116 images, and the testing data were 20% with 9958 images [12].

#### 4.1.3 | Metrics

We evaluate our experiments using the average precision (AP or  $AP_{role}$  as in V-COCO) for each action and mean average precision (mAP) for the overall performances following [11]. The mAP is formulated as

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i, \quad (3)$$

where  $N$  is the total number of HOI categories. The AP role in the HOI detection is the AP of the target of interest in the triplet of human, verb, and object [5]. A true-positive triplet can be obtained if three conditions are satisfied: (a) Intersection-over-union (IoU) of ground truth and the prediction of the human bounding box is more than 0.8, (b) IoU of ground truth and the prediction of the object bounding box is more than 0.4, and (c) the ground truth and predicted human actions labels are matched.

## 4.2 | Implementation details

In this article, we used Detectron [28] with a feature pyramid network [31] built on the ResNet-50 backbone [27] to generate human and object bounding boxes from the images. We only retained human bounding boxes that have a confidence score higher than 0.8 and object bounding boxes that have a confidence score higher than 0.4. For the feature extractor, we employed ResNet-50 [27]. During the training and inference, we froze ResNet-50 and only used it as a feature extractor. In the CCM, we set the kernel size ( $k$ ) as 15,  $C_{mid}$  as 21, and  $C_{out}$  as 21. Therefore, we could incorporate the entire feature map yet used only a few channels. A three-stream network was trained for 300K iterations on the training set with a learning rate of  $10^{-3}$ , weight decay of  $10^{-4}$ , and stochastic gradient descent optimizer of momentum 0.9 on a single NVIDIA GTX1080 GPU.

**TABLE 1** Comparison results on V-COCO. Results are reported in mean average precision (mAP) (%)

Method	mAP (%)
Gupta et al [5]	31.8
InteractNet [11]	40.0
GPNN [32]	44.0
iCAN [14]	44.7
Ours	<b>45.3</b>

Bold values indicate the highest performance scores obtained from the corresponding experiments

**TABLE 2** Comparison results on HICO-DET dataset. Results are reported in mean average precision (mAP) (%)

Method	Default			Known object		
	Full	Rare	Non-rare	Full	Rare	Non-rare
Shen et al [13]	6.46	4.24	7.12	N/A	N/A	N/A
HO-RCNN [12]	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [11]	9.94	7.16	10.77	N/A	N/A	N/A
iHOI [33]	9.97	7.11	10.83	N/A	N/A	N/A
iCAN [14]	12.80	<b>8.53</b>	14.07	14.70	<b>10.79</b>	15.87
Ours	<b>13.67</b>	7.59	<b>15.49</b>	<b>15.85</b>	10.26	<b>17.52</b>

Bold values indicate the highest performance scores obtained from the corresponding experiments

In addition, the input ratio of our implementation was 2:1 for the negative and positive triplets of the HOIs (human, action, and object).

## 4.3 | Results

We compare the results of our proposed method with those of previous HOI detection methods. Tables 1 and 2 show that our method outperforms other competing methods. For experiments on V-COCO, our method achieves 45.3%, gaining 0.6 points compared with the best-performing existing method iCAN [14]. As for the HICO-DET dataset, our method achieves the best performances on Full Default Object, Non-Rare Default Object, Full Known Object, and Non-Rare Known Object with improvements of +0.87, +1.49, +1.15, and +1.65, respectively. However, our method achieved worse results than iCAN on Rare objects for Default and Known objects. We believe that our method requires more samples to learn the larger context information from images, as we empirically demonstrated the highest improvement on Non-Rare objects.

**TABLE 3** Comparison results using CCM on two scenarios: (1) only on human and object streams; (2) on all three streams

Method	mAP (%)	Size (MB)
iCAN [14] + 1 FC	43.8	571.1
Ours w/CCM on human and object stream (1 FC)	44.4	410.0
Ours w/CCM on human and object stream (2 FCs)	44.8	425.0
Ours w/CCM on all three-stream branches (1 FC)	44.8	364.6
Ours w/CCM on all three-stream branches (2 FCs)	<b>45.3</b>	374.4

Bold values indicate the highest performance scores obtained from the corresponding experiments



**FIGURE 4** Visualization of HOI detection using our method. This figure illustrates the detected triplet of (*human, action, and object*)

#### 4.4 | Ablation

In this section, we discuss the effectiveness of the CCM in two scenarios: 1) employing the CCM on only the human and object streams and 2) employing the CCM on all three streams. For each scenario, we compared two conditions by using only a single FC layer (1 FC) and two FC layers (2 FCs) in each branch of the three-stream network. Using 1 FC, we obtained even fewer parameters (lighter model). For a fair comparison, we present the iCAN performance in two scenarios by using only 1 FC (without dropout) and 2 FCs (results in Table 2) in each branch of its three-stream network. The overall results are reported in Table 3. In the first scenario, our method achieved 44.4% and 44.8% when the human and object streams employed 1 FC and 2 FCs, respectively. In the second scenario, we employed the CCM on the human, object, and human–object pair streams. We demonstrated that the performance of our method without the CCM on the human–object pair stream with 2 FCs are the same as that of our method using the CCM on all three streams with 1 FC. However, employing the CCM on all three streams resulted in fewer parameters. Finally, the

best-performing model is when the three-stream network employs the CCM on all its three streams with 45.3%. Therefore, we have empirically proven that the CCM improves HOI detection performance while using fewer network parameters. In addition to these quantitative results, we show the qualitative results from using our method in Figure 4.

## 5 | CONCLUSION

We herein proposed a three-stream HOI detection network that employed a CCM in each of its stream branches. Using our method, we could capture larger context information relevant for recognizing interactions between human and distant objects from input feature maps. The proposed method was evaluated over two public benchmark datasets. Results from extensive experiments demonstrated the effectiveness of our method, which outperformed the state-of-the-art HOI detection methods.

#### ORCID

Thomhert S. Siadari  <https://orcid.org/0000-0001-7811-0953>

## REFERENCES

1. A. Gupta, A. Kembhavi, and L. S. Davis, *Observing human-object interactions: Using spatial and functional compatibility for recognition*, IEEE Trans. Pattern Anal. Mach. Intell. **31** (2009), no. 10, 1775–1789.
2. V. Delaitre, I. Laptev, and J. Sivic, *Recognizing human actions in still images: a study of bag-of-features and part-based representations*, in Proc. BMVC 2010-21st British Mach. Vision Conf., 2010, pp. 97:1–11.
3. B. Yao and L. Fei-Fei, *Modeling mutual context of object and human pose in human-object interaction activities*, in Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recogn., San Francisco, CA, USA, June 2010, pp.17–24.
4. C. Y. Chen and K. Grauman, *Predicting the location of interactees in novel human-object interactions*, Asian conference on computer vision, Springer, Cham, Switzerland, 2014, pp. 351–367.
5. S. Gupta and J. Malik, *Visual semantic role labeling*, arXiv preprint arXiv:1505.04474, 2015.
6. L. Wang and D. Sng, *Deep learning algorithms with applications to video analytics for a smart city: a survey*, arXiv preprint arXiv:1512.03131, 2015.
7. J. W. Choi, D. Moon, and J. H. Yoo, *Robust multi-person tracking for real-time intelligent video surveillance*, ETRI J. **37** (2015), no. 3, 551–561.
8. J. Moon et al., *Extensible hierarchical method of detecting interactive actions for video understanding*, ETRI J. **39** (2017), no. 4, 502–513.
9. K. Yun et al., *Vision-based garbage dumping action detection for real-world surveillance platform*, ETRI J. **41** (2019), no. 4, 494–505.
10. Y. Licheng et al., *Visual madlibs: fill in the blank image generation and question answering*, arXiv preprint arXiv:1506.00278, 2015.
11. G. Gkioxari et al., *Detecting and recognizing human-object interactions*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Salt Lake City, UT, USA, June 2018, pp. 8359–8367.
12. Y. W. Chao et al., *Learning to detect human-object interactions*, in Proc. IEEE Winter Conf. Applicat. Comput. Vision, Lake Tahoe, NV, USA, Mar. 2018, pp. 381–389.
13. L. Shen et al., *Scaling human-object interaction recognition through zero-shot learning*, in Proc. IEEE Winter Conf. Applicat. Comput. Vision, Lake Tahoe, NV, USA, Mar. 2018, pp. 1568–1576.
14. C. Gao, Y. Zou, and J. B. Huang, *iCAN: Instance-centric attention network for human-object interaction detection*, British Machine Vision Conference, 2018.
15. C. Peng et al., *Large kernel matters—improve semantic segmentation by global convolutional network*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Honolulu, HI, USA, 2017, pp. 4353–4361.
16. M. A. Sadeghi and A. Farhadi, *Recognition using visual phrases*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Providence, RI, USA, 2011, pp. 1745–1752.
17. L. Cewu et al., *Visual relationship detection with language priors*, European Conference on Computer Vision, Springer, Cham, Switzerland, 2016, pp. 852–869.
18. M. Yatskar, L. Zettlemoyer, and A. Farhadi, *Situation recognition: Visual semantic role labeling for image understanding*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Las Vegas, NV, USA, 2016, pp. 5534–5542.
19. B. Dai, Y. Zhang, and D. Lin, *Detecting visual relationships with deep relational networks*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Honolulu, HI, USA, 2017, pp. 3076–3086.
20. H. Zhang et al., *Visual translation embedding network for visual relation detection*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Honolulu, HI, USA, 2017, pp. 5532–5540.
21. H. Ronghang et al., *Modeling relationships in referential expressions with compositional modular networks*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Honolulu, HI, USA, 2017, pp. 1115–1124.
22. J. Peyre et al., *Weakly-supervised learning of visual relations*, in Proc. IEEE Int. Conf. Comput. Vision, Venice, Italy, 2017, pp. 5179–5188.
23. A. Kolesnikov, C. H. Lampert, and V. Ferrari, *Detecting visual relationships using box attention*, arXiv preprint arXiv:1807.02136, 2018.
24. M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, *Feedforward semantic segmentation with zoom-out features*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Boston, MA, USA, 2015, pp. 3376–3385.
25. W. Liu, A. Rabinovich, and A. C. Berg, *Parsenet: Looking wider to see better*, arXiv preprint arXiv:1506.04579, 2015.
26. F. Yu and V. Koltun, *Multi-scale context aggregation by dilated convolutions*, arXiv preprint arXiv:1511.07122, 2015.
27. K. He et al., *Deep residual learning for image recognition*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Las Vegas, NV, USA, June 2016, pp. 770–778.
28. R. Girshick et al., *Detectron*, <https://github.com/facebookresearch/detectron>, 2018.
29. T. Y. Lin et al., *Microsoft COCO: Common objects in context*, in Proc. Computer Vision—ECCV, Zurich, Switzerland, Sept. 2014, pp. 740–755.
30. Y. W. Chao et al., *HICO: A benchmark for recognizing human-object interactions in images*, in Proc. IEEE Int. Conf. Comput. Vision, Santiago, Chile, 2015, pp. 1017–1025.
31. T. Y. Lin et al., *Feature pyramid networks for object detection*, in Proc. IEEE Conf. Comput. Vision Pattern Recogn., Honolulu, HI, USA, July 2017, pp. 2117–2125.
32. S. Qi et al., *Learning human-object interactions by graph parsing neural networks*, in Proc. Eur. Conf. Comput. Vision (ECCV), 2018, pp. 401–417.
33. X. Bingjie et al., *Interact as you intend: Intention-driven human-object interaction detection*, CoRR abs/1808.09796, 2018.

## AUTHOR BIOGRAPHIES



**Thomhert S. Siadari** received his BS degree in telecommunication engineering from Telkom University, Bandung, Indonesia, in 2011, and his MEng in IT convergence engineering from Kumoh National Institute of Technology, Gumi, Republic of Korea, in 2013. He is currently pursuing PhD in ICT from the ETRI School, University of Science & Technology, Daejeon, Republic of Korea. His main research interests include machine learning, deep learning, and computer vision.



**Mikyong Han** received her MS degree in computing engineering from the School of Electronics and Information, Kyung Hee University, Seoul, Republic of Korea, in 1993. She joined the Electronics and Telecommunications Research Institute in 1993 and is currently a principal research member. From March 2012 to February 2013, she was a visiting professor at the School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, USA. Her major research interests include multimedia service platforms, immersive media service platforms, and gigamedia service platforms.



**Hyunjin Yoon** received her BS and MS degrees in computer science and engineering from Ewha Womans University, Seoul, Republic of Korea, in 1996 and 1998, respectively, and her PhD degree in computer science from the Viterbi School of Engineering, University of Southern California, Los Angeles, USA, in 2009. Since 2010, she has been with the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. Currently, she is an associate professor in ICT of the ETRI School, University of Science & Technology, Daejeon, Republic of Korea. Her main research interests include machine learning, deep learning, and pattern recognition, with applications in smart cities and autonomous agents.