Wang, Tiancai; Anwer, Rao Muhammad; Khan, Muhammad Haris; Khan, Fahad Shahbaz; Pang, Yanwei; Shao, Ling; Laaksonen, Jorma

Deep Contextual Attention for Human-Object Interaction Detection

Published in:
Proceedings of the International Conference on Computer Vision (ICCV2019)

DOI:
10.1109/ICCV.2019.00579

Published: 01/02/2020

Document Version
Peer reviewed version

Please cite the original version:
Wang, T., Anwer, R. M., Khan, M. H., Khan, F. S., Pang, Y., Shao, L., & Laaksonen, J. (2020). Deep Contextual Attention for Human-Object Interaction Detection. In *Proceedings of the International Conference on Computer Vision (ICCV2019)* (pp. 5693-5701). [9008846] (Proceedings of the IEEE International Conference on Computer Vision; Vol. 2019-October). IEEE. https://doi.org/10.1109/ICCV.2019.00579

Deep Contextual Attention for Human-Object Interaction Detection

Tiancai Wang[1*†], Rao Muhammad Anwer[2*], Muhammad Haris Khan[2], Fahad Shahbaz Khan[2],
Yanwei Pang[1], Ling Shao[2], Jorma Laaksonen[3]
[1]School of Electrical and Information Engineering, Tianjin University
[2]Inception Institute of Artificial Intelligence (IIAI), UAE
[3]Department of Computer Science, Aalto University School of Science, Finland
[1]{wangtc, pyw}@tju.edu.cn, [2]{rao.anwer, muhammad.haris, fahad.khan, ling.shao}@inceptioniai.org
[3]{jorma.laaksonen}@aalto.fi

Abstract

Human-object interaction detection is an important and relatively new class of visual relationship detection tasks, essential for deeper scene understanding. Most existing approaches decompose the problem into object localization and interaction recognition. Despite showing progress, these approaches only rely on the appearances of humans and objects and overlook the available context information, crucial for capturing subtle interactions between them. We propose a contextual attention framework for human-object interaction detection. Our approach leverages context by learning contextually-aware appearance features for human and object instances. The proposed attention module then adaptively selects relevant instance-centric context information to highlight image regions likely to contain human-object interactions. Experiments are performed on three benchmarks: V-COCO, HICO-DET and HCVRD. Our approach outperforms the state-of-the-art on all datasets. On the V-COCO dataset, our method achieves a relative gain of 4.4% in terms of role mean average precision (mAP_{role}), compared to the existing best approach.

1. Introduction

Recent years have witnessed tremendous progress in various instance-level recognition tasks, including object detection and segmentation. These instance-level problems have numerous applications in robotics, autonomous driving and surveillance. However, such applications demand a deeper knowledge of scene semantics beyond instance-level recognition, such as the inference of visual relationships between object pairs. Detecting human-object interactions (HOI) is a class of visual relationship detection. Given an image, the task is to not only localize a human and an object, but also recognize the interaction between them. Specifically, it boils down to detecting $\langle$*human, action, object*$\rangle$ triplets. The problem is challenging as it focuses on both human-centric interactions with fine-grained actions (*i.e.*, riding a horse vs. feeding a horse) and involves multiple co-occurring actions (*i.e.*, eating a donut and interacting with a computer while sitting on a chair).

Most existing HOI detection approaches typically tackle the problem by decomposing it into two parts: object localization and interaction recognition [1, 10, 11, 13, 20, 26]. In the first part, off-the-shelf two-stage object detectors [7, 22, 8] localize both human and object instances in an image. In the second part, detected human and object instances and the pairwise interaction between them are treated separately in a multi-stream network architecture. Recent works have attempted to improve HOI detection by integrating, *e.g.*, structural information [20], gaze and pose cues [26]. Despite these recent advances, the HOI detection performance is still far from satisfactory compared to other vision tasks, such as object detection and instance segmentation.

Current HOI detection approaches tend to focus on appearance features of human and object instances (bounding-boxes) that are central to scoring human-object interactions, and thereby identifying triplets. However, the readily available auxiliary information, such as context, at various levels of image granularity is overlooked. Context information is known to play a crucial role in improving the performance of several computer vision tasks [4, 27, 18, 2]. However, it is relatively underexplored for the high-level task of HOI detection, where context around each candidate detection is likely to provide complementary information to standard bounding-box appearance features. Global context provides valuable image-level information by determining the presence or absence of a specific object category. For instance, when detecting *driving a boat* interaction category, person, boat and water are likely to co-occur in an image. However for *drive a car* category, interaction (drive) remains the

*Equal contribution
†Work done at IIAI during Tiancai's internship.

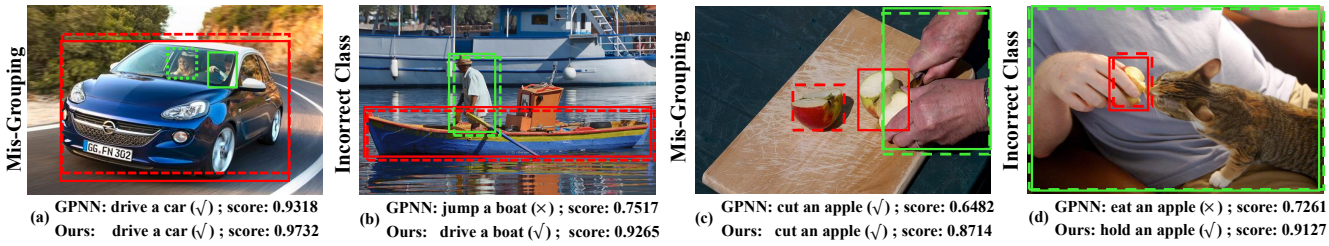

Figure 1. Example of HOI detections using the proposed approach and the recently introduced GPNN method [20]. The four examples depict two HOI detection cases. First in (a) and (b), different object categories (*car* and *boat*) involve the same human-object interaction (*drive*). Second in (c) and (d), different human-object interactions (*cut an apple* and *hold an apple*) involve the same object (*apple*). In case of (a) and (c), GPNN method fails to correctly pair the agent (person) and object, while it miss-classifies the action categories (b) and (d). Our approach accurately groups the agent and the respective object, while correctly classifying the action labels (scores) in all four cases.

same and only context (water) is changed. Besides global context, information in the immediate vicinity of each human/object instance provides additional cues to distinguish different interactions, *e.g.*, various interactions involving the same object. For instance, the surrounding neighborhood in *eating an apple* category is the face of the person whereas for *cutting an apple* category, it is knife and part of the hand (see Fig. 1). In this work, we leverage the context information to the relatively new problem of HOI detection.

Contributions: We first introduce a contextually enriched appearance representation for human and object instances. While providing auxiliary information, global context also introduces background noise which hampers interaction recognition performance. We therefore propose an attention module to suppress the background noise, while preserving the relevant contextual information. Our attention module is conditioned to specific instances of humans and objects to highlight the interaction regions, *i.e.*, *kick a sports ball* versus *throw a sports ball* categories. The resulting human/object attention maps are then used to modulate the global features to highlight image regions that are likely to contain a human-object interaction.

We validate our approach on three HOI detection benchmarks: V-COCO [11], HICO-DET [1] and HCVRD [32]. We perform a thorough ablation study to show the impact of context information for HOI detection. The results clearly demonstrate that the proposed approach provides a significant improvement over its non-contextual baseline counterpart. Further, our contextual attention-based HOI detection framework sets a new state-of-the-art on all datasets. On HICO-DET dataset, our approach yields a relative gain of 9.4% in terms of mean average precision (mAP), compared to the best published method [5]. Fig. 1 shows a comparison of our approach with GPNN [20] on HICO-DET images.

2. Related Work

Object Detection: Significant progress has been made in the field of object detection [7, 23, 22, 8, 29, 15, 21, 17], predominantly due to deep convolutional neural networks (CNNs). Generally, CNN-based object detectors can be divided into two-stage and single-stage approaches. In the two-stage approach, object detection methods [7, 22, 8] first employ an object proposal generator to generate regions of interests, which are then passed through an object classification and bounding-box regression pipeline. In contrast, single-stage detection methods [21, 17] directly learn object category predictions (classification) and bounding-box locations (regression) using anchors to predict the offsets of boxes instead of coordinates. Two-stage object detectors are generally more accurate compared to their single-stage counterparts. As in previous HOI detection works [10, 1], we employ an off-the-shelf two-stage FPN detector [15] to detect both human and object instances.

Human-Object Interaction Detection: Gupta and Malik [11] were the first to introduce the problem of visual semantic role labeling. In this problem, the aim is to detect a human, an object, and label the interaction between them. Gkioxrari *et al.*, [10] proposed a human-centric approach by extending the Faster R-CNN pipeline [22] with an additional branch to classify both actions and action-specific probability density estimation over the target object location. The work of [20] proposed a Graph Parsing Neural Network (GPNN) in which HOI structures are represented with graphs and then optimal graph structures are parsed in an end-to-end fashion. The work of [26] introduced a human intention-driven approach, where both pose and gaze information are exploited in a three-branch framework: object detection, human-object pairwise interaction and gaze-driven stream. Kolesnikov *et al.*, [13] proposed a joint probabilistic model for detecting visual relationships. Chao *et al.*, [1] introduced a human-object region-based CNN approach that extends the region-based object detector (Fast R-CNN) and has three streams: human, object and pairwise. Further, they introduced a new large-scale human-object interaction detection benchmark (HICO-DET).

Contextual Cues in Vision: Context provides an auxiliary cue for several vision problems, such as object detection

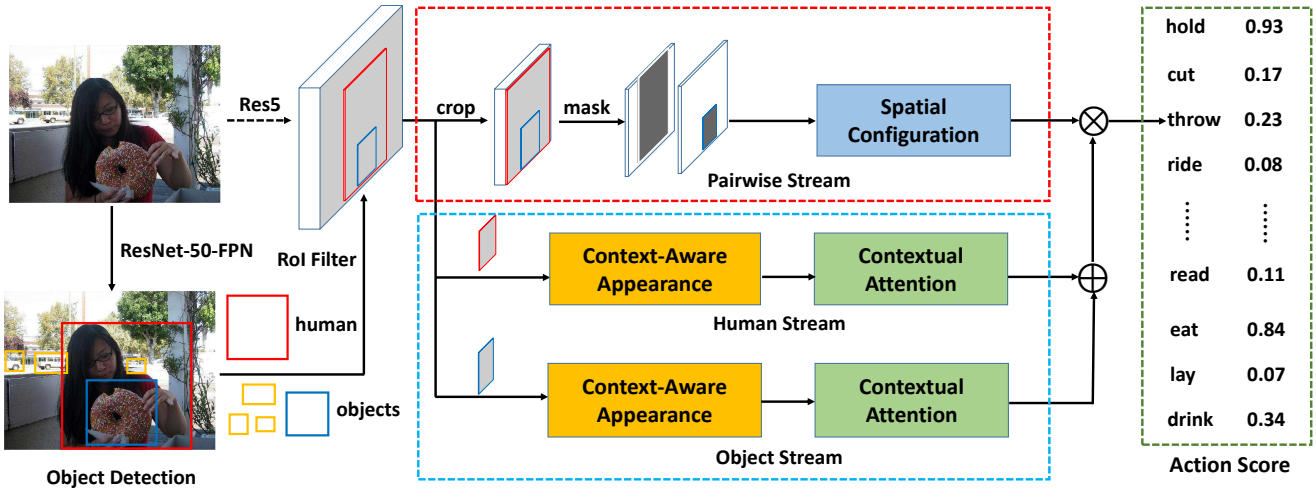

Figure 2. Overall multi-stream architecture of our proposed HOI detection framework comprising a localization and an interaction stage. For localization, we follow the standard object detector [15] to obtain human and object bounding-box predictions. For interaction prediction, we fuse scores from a human, an object, and a pairwise stream. We introduce context-aware appearance and contextual attention modules in the human and object streams. Final predictions are obtained by fusing the scores from human, object and pairwise streams.

[18, 2], action recognition [27], and semantic segmentation [4]. Recently, learnable context has gained popularity with the advent of deep neural networks [6, 18]. Despite its success in several tasks [19, 6, 18, 31, 14, 28], the impact of contextual information to the relatively new task of HOI detection is yet to be fully explored.

3. Overall Framework

The overall framework comprises two stages: localization and interaction prediction (see Fig. 2). For localization, we follow the popular paradigm of FPN [15] as a standard object detector to generate bounding-boxes for all possible human and object instances in the input image. For interaction prediction, following [1], we fuse scores from the three individual streams: a human, an object, and a pairwise. Scores from human and object streams are added. The resulting scores are then multiplied with pairwise stream.

Multi-Stream Pipeline: The inputs to the multi-stream architecture are the bounding-box predictions from FPN [15] and the original image. The output of the multi-stream architecture is a detected $\langle$*human, action, object*$\rangle$ triplet. The overall framework comprises three separate streams: human, object and pairwise interaction. Both the human and object streams are appearance oriented; they employ CNN feature extraction to generate confidence scores on the detected human and object bounding-boxes. The pairwise interaction stream encodes the spatial relationship between the person and object as in [1].

3.1. Proposed Human/Object Stream

The standard multi-stream architecture encodes instance-centric (bounding-box) appearance features in the human and object streams and ignores the associated contextual information. In this work, we argue that the bounding-box appearance alone is insufficient and that the contextual information in the vicinity of a human and object instances provides complementary information useful to distinguish complex human-object interactions. We therefore enrich the human and object streams (see Fig. 3) with contextual information by introducing contextually-aware appearance features f_{app} (sec. 3.1.1). These contextual appearance features f_{app} are then fed into the contextual attention module (sec. 3.1.2), where they are used to modulate the global feature map A to obtain a modulated feature representation F_m. The modulated feature representation F_m is further refined in the attention refinement block to obtain the refined modulated features F_r, which further passes through global average pooling to obtain refined modulated vector f_r. Subsequently, both representations f_{app} and f_r are concatenated to obtain action predictions from the human/object streams. Note that the same architecture is employed for both the human and object streams. Thus, the only difference between the two streams is their inputs, which are human and object bounding-box predictions, respectively. Next, we describe different components of our proposed human/object stream.

3.1.1 Contextually-Aware Appearance Features

Given the CNN features (Res5 block of the ResNet-50 backbone) of the whole image, as well as human/object bounding-box predictions from the detector, standard instance-centric appearance features are extracted by employing region-of-interest (ROI) pooling followed by a residual block and global average pooling. Though theoretically the image-level CNN features used in the construction

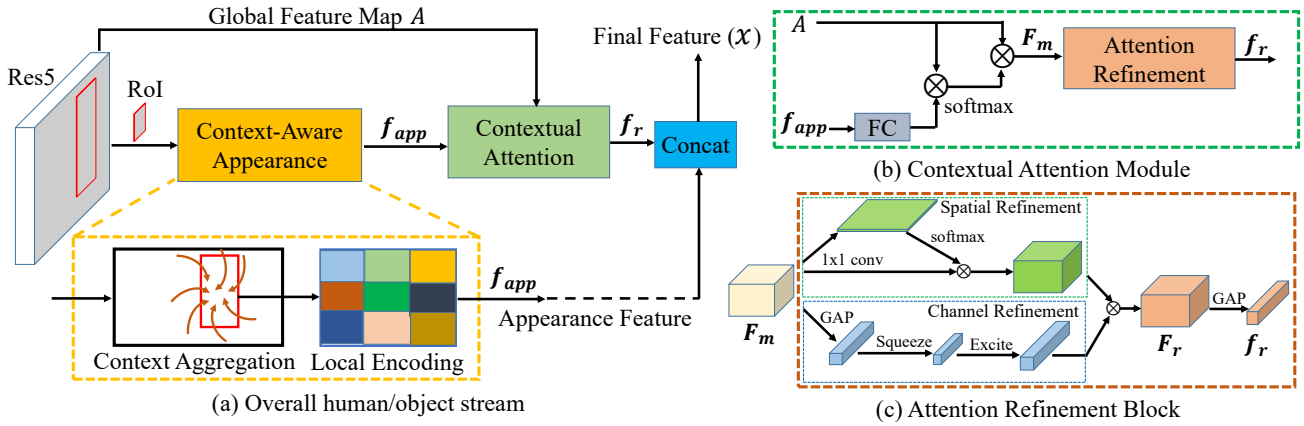

Figure 3. On the left (a), the proposed overall human/object stream. Both the contextual attention module (b) and the attention refinement block (c) are shown on the right. The context-aware appearance module produces contextual appearance features that encode both appearance and context information. The contextual appearance features are then fed into the contextual attention module to suppress the background noise resulting in a modulated feature representation. The modulated feature representation is further enriched in the attention refinement block to obtain refined modulated features. Consequently, both contextual appearance and refined modulated features are concatenated to obtain action predictions from the human/object stream.

of the standard appearance representation are supposed to cover entire spatial image extent, their valid receptive field is much smaller in practice [30]. This implies that the larger global scene context prior is ignored in such a standard appearance feature construction. Our context-aware appearance module is designed to capture additional context information and consists of context aggregation and local encoding blocks (see Fig. 3(a)).

The context aggregation block aims to capture a larger field-of-view (FOV) to integrate context information in instance-centric appearance features, while preserving spatial information. A straightforward way to capture a larger FOV is through a fully connected (FC) layer or cascaded dilated convolutions. However, the former collapses spatial dimensions, while the latter produces sparser features. Therefore, our context aggregation block employs a large convolutional kernel (LK) previously used for semantic segmentation [19]. To the best of our knowledge, we are the first to introduce a large kernel-based context aggregation block to construct contextual appearance features for the problem of HOI detection. The input to the context aggregation block is the CNN features (Res5 block) of the image with size $h \times w \times c_{in}$, where c_{in} denotes the number of channels and h and w denote the input feature dimensions. The output of the context aggregation block is then context-enriched features of size $h \times w \times c_{out}$, obtained after applying a large kernel of size $k \times k$ to the original CNN features. In this work, we utilize the factorized large kernel, which is efficient as its computational complexity and number of parameters are only $O(2/k)$, compared to the trivial $k \times k$ convolution.

Beside context aggregation, our context-aware appearance module contains a local encoding block. Existing HOI detection approaches employ standard ROI warping, which involves a max-pooling operation performed on the cropped ROI region. Our local encoding block aims to preserve locality-sensitive information in each bounding-box ROI region by encoding the position information with respect to a relative spatial position. Such a strategy has been previously investigated to encode spatial information within ROI regions in the context of generic object detection [3]. However, [3] directly employs a 1×1 convolution on the standard CNN feature map (Res5). Instead, we encode locality-sensitive information in each ROI region based on the contextualized CNN feature map obtained from our context aggregation block. Further, [3] utilizes PSRoIpooling with average pooling. Instead, we employ the PSRoIAlign together with max-pooling. PSRoIAlign is employed to reduce the impact of coarse quantization caused by PSRoIpooling through bilinear interpolation. Fig. 4 shows the impact of PSRoIAlign-based local encoding on the input feature maps of an image. Consequently, the output of the local encoding block is flattened and passed through a fully-connected layer to obtain conextual appearance features f_{app}.

3.1.2 Contextual Attention

The contextual appearance features, described above, encode both appearance and global context information. However, not all background information is equally useful for the HOI problem. Further, integrating meaningless background noise can even deteriorate the HOI detection performance. Therefore, a careful identification of useful contextual information is desired to distinguish subtle human-object interactions that are difficult to handle otherwise. Generally, attention mechanisms are used to highlight the discriminative features particularly important for a given

task [25]. The contextual attention module in our human/object stream consists of bottom-up attention and attention refinement components. The bottom-up attention component is based on the recently introduced approach of [6] for action recognition and exploits a scene-level prior to focus on relevant features. Note, [6] computes image-level attention, whereas we aim to generate bounding-box based attention. Further, contrary to standard appearance features, the bottom-up attention maps in our attention module are generated using *contextually-aware appearance* features f_{app} (sec. 3.1.1) that encode both appearance and context. We generate modulated features by first constructing a contextual attention map, which is then deployed to modulate the input CNN feature map (see Fig. 3(b)).

Specifically, we project the input (Res5) feature maps f using a 1×1 convolution onto a 512-dimensional space, denoted as A. Then, we compute the dot product between these projected global features A and contextual-appearance features f_{app} to obtain an attention map, which is then used to modulate A, such that,

$$F_m = \text{softmax}(f_{app} \otimes A) \otimes A \tag{1}$$

Here, F_m are the resulting modulated features. The discriminative ability of F_m is further enhanced in the attention refinement block, which consists of spatial and channel-wise attention refinement. The attention refinement block is simple and light-weight (see Fig. 3(c)). During spatial refinement, we first apply a 1×1 conv on modulated features F_m to generate a single-channel heatmap H, followed by a softmax-operation-based normalization. Then, we perform an element-wise multiplication between the normalized heatmap and the modulated features F_m. The resulting spatial refinement S_{att} learns the most relevant features as:

$$S_{att}(F_m) = H \otimes F_m \tag{2}$$

Beside spatial refinement, we also perform a channel-wise refinement. Inspired by the squeeze-and-excitation network (SENet) of [12], we first apply global average pooling on the modulated features F_m to squeeze global spatial information into a channel descriptor z. Then, the excitation stage is a stack of two FC layers, followed by a sigmoid activation with input z and is described as:

$$C_{att}(F_m) = \sigma(W_1\delta(W_2 z)) \tag{3}$$

Here, z is the output of the squeeze operation, and W_1 and W_2 refer to fully-connected operations. δ and σ are ReLU and sigmoid activations, respectively. Finally, C_{att} modulates the spatially-attended features S_{att} to further highlight regions relevant to human-object interaction to obtain a refined modulated feature representation F_r as:

$$F_r = S_{att}(F_m) \otimes C_{att}(F_m) \tag{4}$$

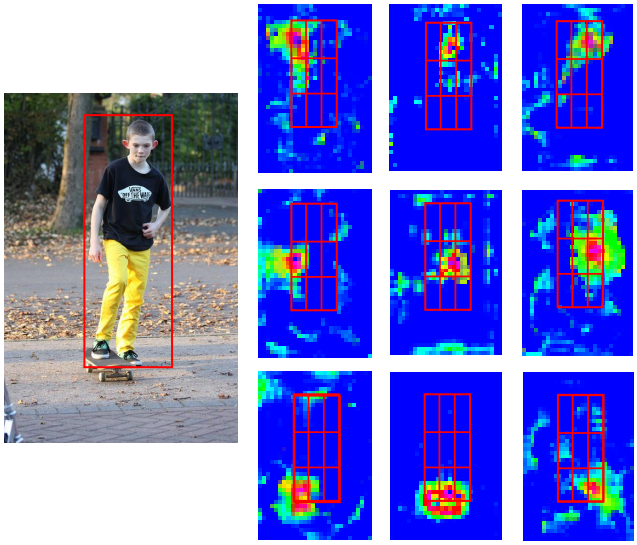

Figure 4. Visual depiction of the local encoding block that preserves locality-sensitive information. For illustration purposes, the detected human bounding-box is divided into 3×3 sub-regions and there are 9 score maps. Each sub-region votes for the presence of a specific object part, relative to the position of the object, based on how good the bounding-box overlaps with the score maps.

Finally, the refined modulated features F_r are passed through global average pooling to obtain the refined modulated vector f_r. We combine contextual appearance features f_{app} and the refined modulated vector f_r to produce the final representation x. This representation x is then passed through two FC layers to estimate action predictions from the human/object stream, respectively. Given an HOI predicted bounding-box, the final prediction is obtained by fusing the scores from the human, object and pairwise streams.

4. Experiments

4.1. Dataset and Evaluation Protocol

V-COCO [11]: is the first HOI detection benchmark and a subset of popular MS-COCO dataset [16]. The V-COCO dataset contains 10,346 images in total, with 16,199 human instances. Each human instance is annotated with 26 binary action labels. Note that three action classes (*i.e.*, cut, hit, eat) are annotated with two types of targets (*i.e.*, instrument and direct object). It includes 2533, 2867, and 4946 images for training, validation and testing, respectively.

HICO-DET [1]: is a challenging dataset and has 47,776 images in total, with 38,118 images for training and 9658 images for testing. There are more than 150k human instances annotated with 600 types of different human-object interactions. The HICO-DET dataset contains same 80 object categories as MS-COCO and 117 action verbs.

HCVRD [32]: is a large-scale dataset and is labeled with both human-centric visual relationships and corresponding human and object bounding boxes. It has 52,855 images

Add-on	Baseline		
Res5-share	✓	✓	✓
Context-aware appearance (sec. 3.1.1)		✓	✓
Contextual attention (sec. 3.1.2)			✓
mAP_{role}	44.5	46.0	**47.3**

Table 1. A baseline comparison when integrating our proposed context-aware appearance and contextual attention modules into the multi-stream architecture. Results are reported in terms of role mean average precision (mAP_{role}) on the V-COCO dataset. For fair comparison, we use the same feature backbone (Res 5 block of ResNet-50) for both our approach and the baseline. Both context-aware appearance and contextual attention modules contribute in the overall improvement in HOI detection performance. Our overall architecture achieves a relative gain of 6.3% over the baseline.

with 1,824 object categories and 927 predicates. It contains 256,550 relationships instances and there are on average 10.63 predicates per object category. We evaluate our method on the predicate detection task, where the goal is to perform predicate recognition given the labels and bounding boxes for both object and human.

Evaluation Protocol: We use the original evaluation protocols for all three datasets, as provided by their respective authors. For the V-COCO dataset, we use role mean Average Precision (mAP_{role}) as an evaluation metric. Here, the aim is to detect the $\langle$*human, action, object*$\rangle$ triplet. The HOI detection is considered correct if the intersection-over-union (IoU) between the human and object bounding-box predictions and the respective ground-truth boxes is greater than the threshold 0.5 together with the correct action label prediction. For HICO-DET, results are reported in terms of mean average precision (mAP). For HCVRD, we report top-1 and top-3 results at 50 and 100 recall.

4.2. Implementation Details

We deploy Detectron [9] with a ResNet-50-FPN [15] backbone to obtain human and object bounding-box predictions. To select a predicted bounding-box as a training sample, we set the confidence threshold to be higher than 0.8 for humans and 0.4 for objects. For interaction prediction, we employ ResNet-50 as the feature extraction backbone pre-trained on ImageNet. The initial learning rate is set to 0.001, weight decay of 0.0001 and a momentum of 0.9 is used for all datasets. The network is trained for 300k on V-COCO and 1800k iterations on HICO-DET and HCVRD, respectively. For input image of size 480×640, our interaction recognition part of the approach takes 130 milliseconds (ms) to process, compared to its baseline counterpart (111ms) on a Titan X GPU.

4.3. Results on V-COCO Dataset

Baseline Comparison: We first evaluate the impact of integrating our proposed context-aware appearance (sec. 3.1.1) and contextual attention (sec. 3.1.2) modules into the human/object stream of the multi-stream architecture. Tab. 1 shows the results on the V-COCO dataset. The baseline multi-stream architecture contains standard appearance features from the *Res5* block of the ResNet-50 backbone, which have a size of $h \times w \times 2048$. These standard appearance features are directly passed through the classifier to obtain the final action scores in the human/object stream, achieving a mAP_{role} of 44.5. The introduction of the proposed contextual appearance features improves the HOI detection performance from 44.5 to 46.0 in terms of mAP_{role}. The performance is further improved by 1.3%, in terms of mAP_{role} when integrating our proposed contextual attention module. Our final framework achieves an absolute gain of 2.8% in terms of mAP_{role}, compared to the baseline.

We further evaluate the impact of contextual information on improving the classification capabilities of the multi-stream architecture. This is done by selecting different IoU thresholds in the range [0.1-0.9] used in the test evaluation of interaction recognition performance. Tab. 2 shows the results on the V-COCO dataset. At a high threshold value (0.9), few ground-truth bounding-boxes are matched, whereas at a low threshold (0.1) most them are matched. Therefore, comparison at lower thresholds mainly focuses

Overlap thresh	0.1	0.3	0.5	0.7	0.9
Baseline	50.1	47.8	44.5	35.9	2.5
Our Approach	**53.5**	**50.8**	**47.3**	**37.0**	**2.8**

Table 2. Performance (in terms of mAP_{role}) with different IoU thresholds, used in the testing, to compare the classification capabilities of our approach with the baseline on the V-COCO dataset. The performance gap between our approach and the baseline increases at lower threshold values.

Backbone Architecture	Baseline	Our Approach
VGG-16	42.0	**44.5**
ResNet-50	44.5	**47.3**
ResNet-101	45.0	**47.8**

Table 3. A comparison (in terms of mAP_{role}) of our approach with the baseline when using different backbone network architectures on the V-COCO dataset. Our approach always provides consistent improvements over the baseline using different backbones.

Methods	Feature Backbone	$\textbf{mAP}_{role}$
Gupta et .al[11]*	ResNet-50-FPN	31.8
InteractNet [10]	ResNet-50-FPN	40.0
BAR [13]	Inception-ResNet	41.1
GPNN [20]	ResNet-50	44.0
iCAN [5]	ResNet-50	45.3
Our Approach	ResNet-50	**47.3**

Table 4. State-of-the-art comparison on the V-COCO dataset. * refers to implementation of the approach of [11] by [10]. The scores are reported in mAP_{role} and the best result is in bold. Our approach sets a new state-of-the-art on this dataset, achieving an absolute gain of 2.0% over the best existing method.

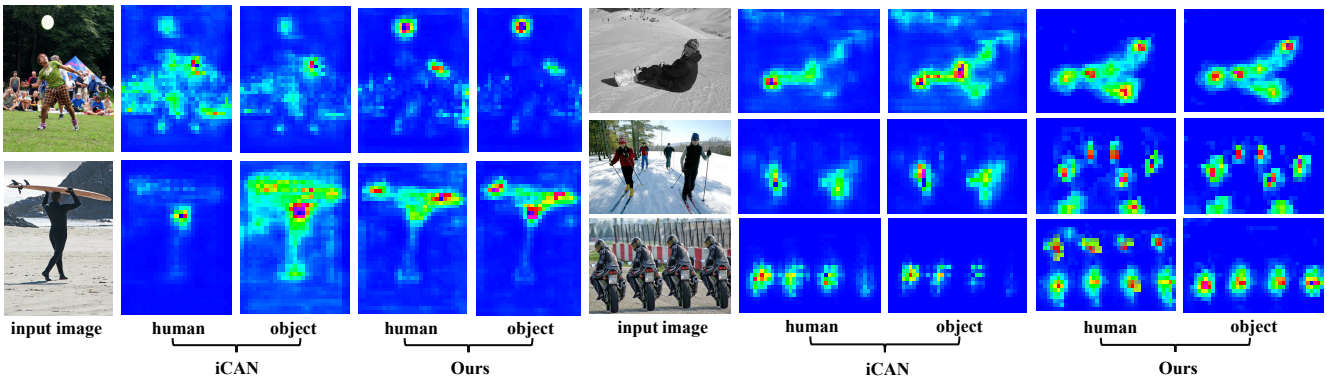

Figure 5. Comparison of attention maps obtained using our approach and iCAN [5] on example images from the V-COCO dataset. Human and object attention maps in iCAN are constructed using standard appearance features. In contrast, human and object attention maps in our approach are constructed using contextual appearance features extracted using the context aggregation and local encoding blocks in our context-aware appearance module. We show examples for both single and multiple human-object interactions.

Figure 6. Example detection results on V-COCO dataset. Each example can involve a single human- object interaction such as *skateboarding* and *eat donut* or multiple humans sharing the same interaction and object - *hold* and *eat pizza*, *throw* and *catch ball*.

on the classification capabilities of our approach. Tab. 2 shows that our approach is superior in terms of classification capabilities, compared to the baseline.

Tab. 3 shows the generalization capabilities of our approach with respect to different network architectures. We perform experiments using VGG-16, ResNet-50 and ResNet-101, each pre-trained on the ImageNet dataset, as the underlying network architectures. In all cases, our approach provides consistent improvements over the baseline.

Comparison with State-of-the-art: In Tab. 4, we compare our approach with state-of-the-art methods in the literature on the V-COCO dataset. Among existing works, InteractNet [10] jointly learns to detect humans, objects and their interactions achieving a mAP$_{role}$ of 40.0. The GPNN approach [20] integrates structural information in a graph neural network architecture and provides a mAP$_{role}$ of 44.0. The iCAN approach [5] combines human, object and their pairwise interaction streams in an early fusion manner using the standard appearance features and bottom-up attention strategy. Our approach sets a new state-of-the-art on this dataset by achieving a mAP$_{role}$ of 47.3.

Qualitative Comparison: Fig. 5 shows comparison between the attention maps obtained using our approach and iCAN [5] on example images from the V-COCO dataset. Note that the attention maps in iCAN [5] are constructed using standard appearance features. In contrast, the attention maps in our approach are constructed using contextual appearance features generated using the context aggrega-

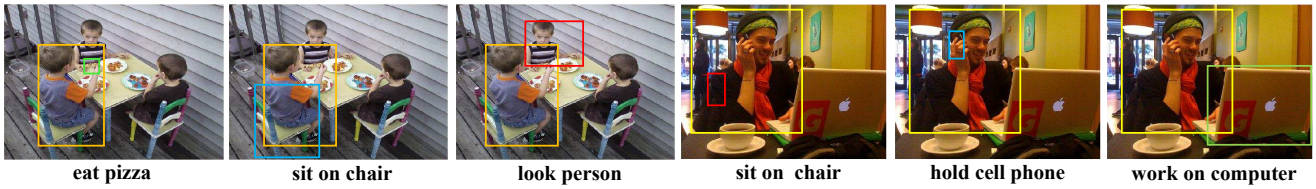

Figure 7. Multiple interaction detection on V-COCO. Our approach detects human instance doing multiple (different) actions and interacting with various objects (represented with different colors). In all cases, the detected agent is represented with the same color.

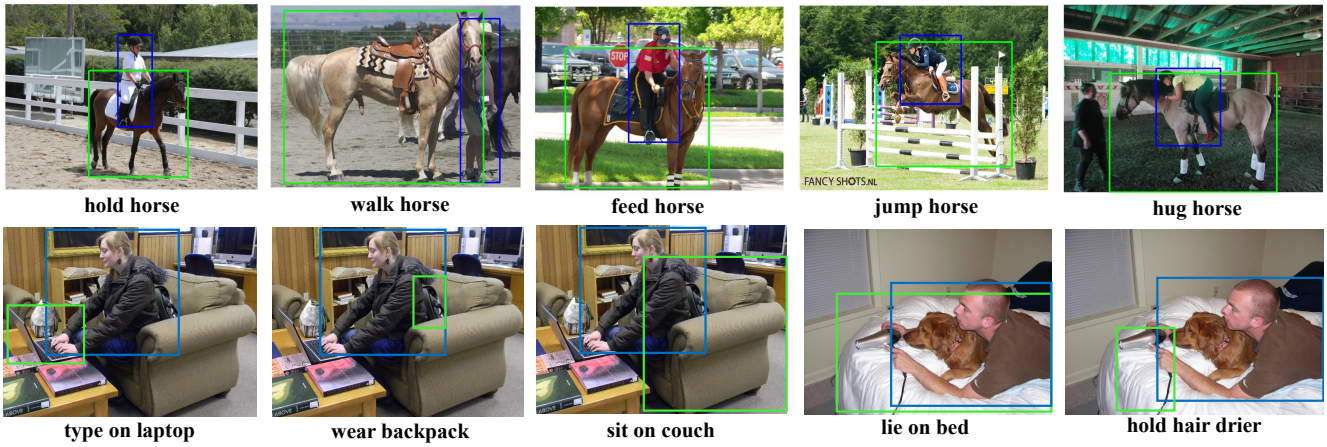

Figure 8. Results on HICO-DET showing one detected triplet. Blue boxes represent a detected human instance, while the green boxes show the detected object of interaction. Our approach detects various fine-grained interactions (top row) and multiple interactions (second row).

tion and local encoding blocks in our context-aware appearance module. Our attention maps focus on relevant regions in the human and object branches that are likely to contain human-object interactions (*e.g.*, in case of *throwing frisbee* and *riding bike*). In addition, for both single and multiple human-object interactions, our approach produces more anchored attention maps compared to the iCAN method.

Fig. 6 shows examples showing both single human-object interactions such as *skateboarding* and *eat a donut*, and multiple humans sharing same interaction and object – *holding* and *eating pizza*, *throw* and *catch ball*. Fig. 7 shows examples of a human performing multiple interactions.

4.4. Results on HICO-DET and HCVRD datasets

On HICO-DET we report results on three different HOI category sets: full, rare, and non-rare with two different settings of Default and Known Objects [1]. Our approach outperforms the state-of-the-art in all three category sets under both Default and Known Object settings (see Tab. 5. The relative gain of 9.4%, 6.7%, and 9.8% is obtained over the best existing method on all three sets in Default settings. Fig. 8 shows results on HICO-DET. On HCVRD dataset, iCAN achieves top-1 and top-3 accuracies at R@50 of 33.8 and 48.9, respectively. Our approach outperforms iCAN with top-1 and top-3 accuracies at R@50 of 37.1 and 51.3, respectively. Similarly, our approach provides superior results at R@100 (top-3 accuracy of iCAN: 49.4 vs. top-3 accuracy of ours: 51.9).

	Default			Known Object		
Methods	full	rare	non-rare	full	rare	non-rare
Shen *et al.*, [24]	6.46	4.24	7.12	-	-	-
Chao *et al.*, [1]	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [10]	9.94	7.16	10.77	-	-	-
GPNN [20]	13.11	9.34	14.23	-	-	-
iCAN [5]	14.84	10.45	16.15	16.43	12.01	17.75
Ours	**16.24**	**11.16**	**17.75**	**17.73**	**12.78**	**19.21**

Table 5. State-of-the-art comparison on the HICO-DET using two different settings: Default and Known Object on all three sets (full, rare, non-rare). Note that Shen *et al.* [24], InteractNet [10] and GPNN [20] only report results on the Default settings. Our approach achieves a relative gain of 9.4%, 6.7%, and 9.9% over the best existing method on all three HOI sets in Default settings.

5. Conclusion

We propose a deep contextual attention framework for HOI detection. Our approach learns contextually-aware appearance features for human and object instances. To suppress the background noise, our attention module adaptively selects relevant instance-centric context information crucial for capturing human-object interactions. Experiments are performed on three HOI detection benchmarks: V-COCO, HICO-DET and HCVRD. Our approach has been shown to outperform state-of-the-art methods on all datasets.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (Grant # 61632018), Academy of Finland project number 313988 and the European Unions' Horizon 2020 (Grant # 780069).

References

[1] Yuwei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1, 2, 3, 5, 8

[2] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. In *CVPR*, 2017. 1, 3

[3] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 4

[4] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *CVPR*, 2018. 1, 3

[5] Chen Gao, Yuliang Zou, and Jia-Bin Huang. iCAN: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 2, 6, 7, 8

[6] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *NIPS*, 2017. 3, 5

[7] Ross Girshick. Fast R-CNN. In *ICCV*, 2015. 1, 2

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1, 2

[9] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 6

[10] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object intaractions. In *CVPR*, 2018. 1, 2, 6, 7, 8

[11] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1, 2, 5, 6

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation networks. *arXiv preprint arXiv:1709.01507*, 2017. 5

[13] Alexander Kolesnikov, Christoph H. Lampert, and Vittorio Ferrari. Detecting visual relationships using box attention. In *arXiv preprint arXiv:1807.02136*, 2018. 1, 2, 6

[14] Jianan Li, Yunchao Wei, Xiaodan Liang, Jian Dong, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Attentive contexts for object detection. *TMM*, 2017. 3

[15] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2, 3, 6

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016. 2

[18] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *CVPR*, 2018. 1, 3

[19] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel mattersimprove semantic segmentation by global convolutional network. In *CVPR*, 2017. 3, 4

[20] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 1, 2, 6, 7, 8

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. Look once: Unified, real-time object detection. In *CVPR*, 2016. 2

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 2

[23] Fahad Shahbaz Khan, Jiaolong Xu, Joost van de Weijer, Andrew Bagdanov, Rao Muhammad Anwer, and Antonio Lopez. Recognizing actions through action-specific person detection. *IEEE Transactions on Image Processing*, 24(11):4422–4432, 2015. 2

[24] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, 2018. 8

[25] John Tsotsos, Sean Culhane, Winky Yan, Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 1995. 5

[26] Bingjie Xu, Junnan Li, Yongkang Wong, Mohan S. Kankanhalli, and Qi Zhao. Interact as you intend: Intention-driven human-object interaction detection. *arXiv preprint arXiv:1808.09796*, 2018. 1, 2

[27] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 1, 3

[28] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 3

[29] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection. In *CVPR*, 2018. 2

[30] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014. 4

[31] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *ICCV*, 2017. 3

[32] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. Hcvrd: a benchmark for large-scale human-centered visual relationship detection. In *AAAI*, 2018. 2, 5