Amplifying Key Cues for Human-Object-Interaction Detection

Yang Liu[1][0000−0002−4259−3882], Qingchao Chen[2][0000−0002−1216−5609], and Andrew Zisserman[1][0000−0002−8945−8573]

[1] Visual Geometry Group, Department of Engineering Science, Oxford, UK
[2] Department of Engineering Science, Oxford, UK

Abstract. Human-object interaction (HOI) detection aims to detect and recognise how people interact with the objects that surround them. This is challenging as different interaction categories are often distinguished only by very subtle visual differences in the scene. In this paper we introduce two methods to amplify key cues in the image, and also a method to combine these and other cues when considering the interaction between a human and an object. First, we introduce an encoding mechanism for representing the fine-grained spatial layout of the human and object (a subtle cue) and also semantic context (a cue, represented by text embeddings of surrounding objects). Second, we use plausible future movements of humans and objects as a cue to constrain the space of possible interactions. Third, we use a gate and memory architecture as a fusion module to combine the cues. We demonstrate that these three improvements lead to a performance which exceeds prior HOI methods across standard benchmarks by a considerable margin.

1 Introduction

Human-Object Interaction (HOI) detection—which focuses specifically on relations involving humans—requires not only to retrieve human and object locations but also to infer the relations between them. Thus, for a given image, the objective of HOI is to identify all triplets of the form $<human, verb, object>$. The ability to predict such triplets robustly is central to enabling applications in robotic manipulations [15] and surveillance event detection [1].

Driven by impressive progress on instance detection and recognition, there has been growing interest in the HOI detection problem. However, the majority of existing methods [9, 30, 38] first detect all human and object instances and then infer their pairwise relations using the appearance feature of the detected instances and their coarse layout (position of human and object boxes). Despite their general efficacy, the performance of prior work may still be limited by some particular design choices, which we discuss next.

First, although recent works have sought to introduce some fine-grained spatial configuration descriptors and context cues from the whole image into the HOI detection, the encoding mechanisms have limitations. Specifically, (1) Some

Fig. 1: (Left): Interactions with similar spatial layouts can be resolved through *fine-grained spatial information*. (Centre): *Global and local context* encode the scene and other local objects to provide strong clues for the interaction taking place; (Right): *Plausible motion* estimation distinguishes between interactions for which dynamics play an important role.

approaches [24, 14, 38, 45] use human pose to distinguish the fine-grained relation (going beyond the standard human and object coarse boxes). However, these approaches encode human pose via key-point estimation (plus post-processing), which is problematic as it loses boundary and shape information. For example in Figure 1(left), encoding the fine-grained spatial information as key points exhibits ambiguity when distinguishing the differences between 'riding a bicycle' and 'walking with a bicycle'. We argue that the boundaries of both human parts and objects should be encoded explicitly, due to their critical ability to reveal interaction boundaries and support inference of relations. Thus we improve the encoding mechanism for the fine-grained information by leveraging the fine-grained human parsing and object semantic segmentation masks, to better capture the geometric relations between them. (2) Some approaches use the visual appearance feature from other image regions or the whole image as the auxiliary context information. However, there are a limited number of triplets in existing HOI datasets—this is insufficient to capture the full intra-class appearance variations of relationships (making it harder to generalise). We draw inspiration from classical recognition techniques (e.g. the use of context for detection [7]) and argue that the semantic categories of other objects present in the surrounding neighbourhood of the candidate instance pair (local context) and the scene category (global context) provide valuable cues for distinguishing between different interactions, but the detailed visual appearance of them is often not crucial for HOI detection. For instance, as shown in Fig.1(middle), the surrounding neighbourhood in the 'eating a cake' category will likely comprise a spoon-like tool, whereas for the 'cutting a cake' category, it is a knife-like tool. But the colour and design of the spoon/knife do not provide useful cues when inferring its relation with the human. Instead of using visual appearance features directly, we encode categories via a semantic embedding (word2vec). This enables the model to leverage language priors to capture possible co-occurrence

and affordance relations between objects and predicates. As we show through careful ablation studies in Sec. 4, these mechanisms for encoding fine-grained spatial configuration and contextual cues bring consistent improvements to HOI detection performance, highlighting the importance of studying these choices.

Second, *plausible motion*—the set of probable movements most likely to follow a static image—is not currently accounted for when detecting human object interactions. Nevertheless, humans can trivially enumerate plausible future movements (what may happen next in an image) and characterise their relative likelihood.

The inference of plausible motions brings two benefits: the first is saliency—it provides a natural attention over the key object and human body parts present in an image; the second is that it constrains the space of relations to the subset that is consistent with these dynamics. For example in Fig. 1(right), it is clear that the object and arm highlighted by motion are concerned with throwing or catching the frisbee, not concerned with eating or writing on it. Furthermore, if the estimation of the direction is also correct then that would distinguish whether the person is throwing or catching the frisbee. Note that while the plausible motion can elucidate a salient signal for human object interaction detection, it can be difficult to learn directly from the image alone. We benefit here from recent work on motion hallucination [10], that has learnt to predict local optical flow from a static image to reveal plausible future movement by identifying which regions of pixels will move (together with their velocity) in the instant following image capture. To the best of our knowledge, this work represents the first study of the utility of plausible motion as an additional cue for HOI detection.

In this paper, we aim to tackle the challenges described above with a unified framework that amplifies important cues for HOI detection (as shown in Fig. 2). We design a novel multi-expert fusion module, where different features (i.e., plausible motion, enhanced fine-grained spatial configuration and context cues) are viewed as cooperative experts to infer the human object interaction. As different cues and their relationships will have different contribution for detecting the human object interaction, we use the gate and memory mechanism to fuse the available cues sequentially, select the discriminative information and gradually generate the representation for the whole scene step by step. By doing so, the final representation is more discriminate than those from existing methods that lack a reasoning mechanism, and this leads to better HOI detection performance.

The contributions of this work are summarised as follows:

(1) We propose a mechanism for amplifying fine-grained spatial layout and contextual cues, to better capture the geometric relations and distinguish the subtle difference between relation categories.

(2) We are the first to explore the utility of the plausible motion estimation (which regions of pixels will move) as an additional cue for HOI detection.

(3) We propose a gate and memory mechanism to perform sequential fusion on these available cues to attain a more discriminative representation.

(4) Our approach achieves state-of-the-art performance on two popular HOI detection benchmarks: V-COCO and HICO-DET.

2 Related Work

Visual Relationship Detection. Visual Relationship Detection (VRD) aims to detect objects and simultaneously predict the relationship between them. This topic has attracted considerable attention, supported by the recent development of large-scale relationship datasets such as VRD [26], Visual Genome [21] and Open Images [22]. However, detecting subtle differences between visual relationship remains difficult and the task is made yet more challenging by the distribution of visual relations, which is extremely long-tailed. Several recent works have proposed various mechanisms to address this problem [26, 43, 20, 5, 42, 23]. Our work focuses on one particular class of visual relationship detection problem: detecting human object interaction (HOI). HOI detection poses additional challenges over VRD: a human can also perform multiple actions with one or more objects simultaneously and the range of human actions we are interested in are typically more fine-grained and diverse than for other generic objects.

Human Object Interaction (HOI) Detection. HOI detection aims to detect and recognise how each person interacts with the objects that surround them—it provides the fundamental basis for understanding human behaviour in a complex scene. Recently, driven by the release of relevant benchmarks, HOI detection has attracted significant attention [9, 12, 24, 14, 38, 45, 40, 33, 41].

The earlier works typically focused on tackling HOI by utilizing human and object visual appearance features or by capturing their spatial relationship through their coarse layout (box locations) [9, 12]. Recently, several methods [24, 14] have been developed that use human pose configuration maps to distinguish fine-grained relations. Wan et al. [38] and Fang et al. [6] use human pose cues to zoom into the relevant regions of the image via attention mechanism. Zhou et al. [45] encode human pose through graph neural networks and message passing. Note however, that each of these approaches encode human pose via keypoint estimation (either draw rectangle around the key point or link the keypoint into skeleton), which removes detailed information about the boundary and shape of the human, in particular, about human parts. By contrast, we argue that the boundaries of both human parts and objects are crucial. We are the first to leverage fine-grained human parsing and object semantic segmentation masks to better capture the geometric relations between them—such cues enable discrimination between subtly different relation categories.

More recently, although some approaches [28, 12, 31, 9, 39] have sought to use contextual cues from the whole image (global context), they do so by learning a spatial attention map in pixel space based on instance visual appearance to highlight image regions, making optimisation challenging when training data is limited. By contrast, we draw inspiration from classical recognition techniques [34, 29] and argue that the semantic categories of other objects present in the surrounding neighbourhood of the candidate instance pair (local context) and the scene information (global context) provide valuable cues for resolving ambiguity between different interactions, but the detailed visual appearance of them is often not crucial. Instead of using visual appearance features, we are the first to encode context information via a semantic embedding, i.e., word2vec, that

enables us to leverage the language priors to capture which objects might afford or co-occur with particular predicates.

The prediction of plausible motion has received limited attention for HOI detection. Nevertheless, estimation of current movement—when coupled with an understanding of the dynamics of an interaction—provides a cue for assessing the degree to which the configuration of humans and objects is probable for that interaction. We are the first to leverage flow prediction from a static image to infer the motion most plausible for a given image (i.e., which regions of pixels will move, together with their velocity, in the instant following image capture) as an auxiliary cue for HOI detection. Our approach is related to a wide body of work on visual future prediction [36, 8, 35] and draws on techniques for flow prediction from static images [37, 10]. Differently from prior work, our objective is to infer plausible motion as an auxiliary cue for HOI detection in static images. In this sense, our method bridges static image HOI detection with video-level action understanding by transferring a motion prior learned from videos to images.

Current approaches predominately use either late or early fusion strategy to combine multi-stream features when inferring relations, while the relationship between different streams is often overlooked. Specifically, Late fusion are adopted by [12, 19, 33, 41], where interaction predictions are performed on each stream independently and then summed at inference time. A wide body of work, i.e., [9, 14, 24, 30, 40] use the early fusion strategy, where multi-stream features are concatenated first and then use it to predict the score (sometimes with attention mechanism as in [38, 39, 31, 45]). In this work, we are the first to use the gate and memory mechanism to fuse the available cues for HOI detection, i.e., select the discriminative information and gradually generate the representation for the whole scene step by step.

3 Method

We now introduce our proposed Fine-grained layout-Context-Motion Network (FCMNet), for localising and recognising all human-object interaction instances in an image. We first provide a high-level overview of FCMNet in Sec. 3.1, followed by a detailed description of the model architecture in Sec. 3.2. Finally, Sec. 3.3 describes the training and inference procedures.

3.1 Overview

Our approach to human-object interaction detection comprises two main stages: (1) human and object detection and (2) interaction prediction. First, given an image I, we use Faster R-CNN [32] (Detectron implementation [11]) to detect all person instances $p = (p^1, p^2, ..., p^m)$ and object instances $o = (o^1, o^2, ..., o^n)$, generating a set of bounding boxes $b = (b^1, b^2, ..., b^{m+n})$ where m denotes the total number of detected person and n denotes the number of detected objects. We use $b_H \in \mathbb{R}^4$ and $b_O \in \mathbb{R}^4$ to denote the human and object bounding boxes. HOI proposals are then generated by enumerating all pairs of candidate human

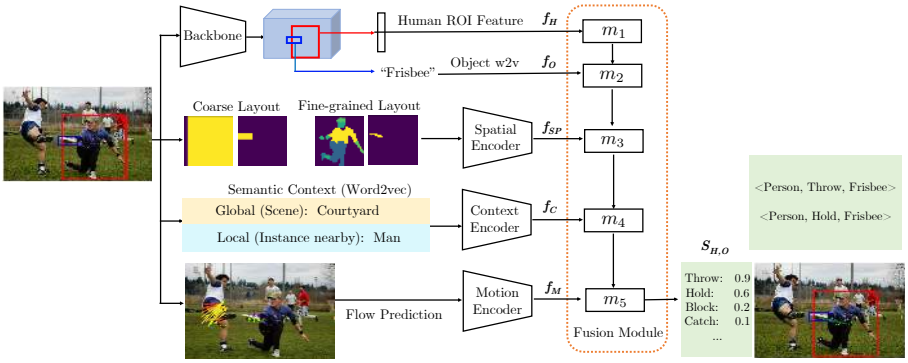

Fig. 2: The proposed **FCMNet** framework. The *backbone module* first detects all human and object boxes and encodes their representations; For each candidate pair of human-object boxes $<b_H, b_O>$, the *spatial encoder* encodes coarse and fine-grained layouts to capture the spatial configuration between them; the *context encoder* accumulates other readily available auxiliary information in the other region of the image, including local and global context; the *motion encoder* infer the likely movement of humans and objects (i.e. which regions of pixels will move together with their velocity in the instant following image capture) via flow prediction from the static image as an approximation to plausible motion. Finally, the *fusion module* combines all available knowledge about the candidate pair $<b_H, b_O>$ to predict the final interaction score for the candidate pair and outputs the detected triplet $<human, verb, object>$.

and object bounding boxes. Next, we process each human-object bounding box pair $<b_H, b_O>$ with FCMNet to predict an *interaction action score* $S_{H,O} \in \mathbb{R}^A$, where A represents the number of interaction classes. FCMNet encodes three image cues independently via the spatial encoder, context encoder, and motion encoder. Finally, a fusion module combines the outputs of these encoders and generates a robust HOI detection. Fig. 2 shows an overview of the model.

3.2 Model Architecture

FCMNet contains the following five modules: (1) a backbone module that detects human and object boxes and encodes their representations; (2) a spatial encoder that leverages both coarse and fine-grained layout information; (3) a context encoder that accumulates auxiliary information in other regions of the image; (4) a motion encoder that predicts which regions of pixels would move in the instant following image capture via flow prediction; (5) a fusion module that combines all cues to predict the final interaction score and outputs the detected triplet $<human, verb, object>$. We next provide details of each component.

(1) Backbone module: We adopt Faster R-CNN (with a ResNet-50 trunk) to detect all human and object instances. To encode a detected human box b_H, we

extract instance level visual appearance feature f_H using standard techniques: we apply ROI pooling, pass the result through a residual block and then perform global average pooling. For the object box b_O, we do not keep the visual appearance feature but instead use the word2vec instead to encode the semantic knowledge of the object category, denoted by f_O. The motivation for this design choice is that the intra-class variation for each object category can be considerable, but detailed visual appearance of the object is often not crucial for the interaction category: for example, the colour and design of the "frisbee" do not provide useful cues when inferring its relation with the human. Using the word2vec semantic embedding for the object enables the model to leverage language priors to capture which objects might afford similar predicates and therefore generalise to previously unseen or rare HOI instances [41].

(2) Spatial Encoder: The *spatial encoder* is designed to encode the spatial relationship between the human and object instance at both a coarse and fine-grained level. The encoding methods for each granularity are described next.

Coarse Layout representation: We use the two-channel binary image representation advocated by [9] to describe the interaction pattern at a coarse level. Specifically, we take the union of the human and object instance boxes as the reference box and construct two binary images (as separate channels) within it: the first channel has value 1 within the human bounding box and 0 elsewhere; the second channel has value 1 within the object box and 0 elsewhere.

Fine-grained Layout representation: Instead of encoding the fine-grained information via key point estimation, we compute a human parse (segmentation masks for the human parts) of the detected human and a segmentation mask for the detected object instance to provide fine-grained layout information. The primary hypothesis underpinning our design is that *the shape and boundary of both the human and object can greatly help disambiguate different actions with a similar coarse spatial layout.* The reason is that these fine-grained cues can reveal the interaction boundary explicitly when inferring the relations. Moreover, fine-grained human parsing both reflects the human's pose and keeps much of the valuable information of their visual appearance. In our work, we use Mask-RCNN [16] to extract the segmentation mask of the object, and use MMAN [27] to extract the segmentation mask of all visible human parts. These two masks are stacked to form the fine-grained layout representation. Specifically, the first channel contains a set of discrete intensity values uniformly ranging from 0.05 to 0.95 to indicate different human parts; the second channel is a binary map that has a value of 1 within the object mask area. Examples of the information conveyed by each channel information can be found in Fig. 2.

We show in our ablation experiments (in Sec. 4.2) that each of the channels forming the coarse and fine-grained layout representations yield improvements in HOI detection performance. Changing the encoding mechanism of the fine-grained information outperforms the one encoding via key point estimation in the literature. In all other experiments, we stack coarse (two channels) and fine-grained (two channels) layout representations together as the input to the spatial encoder unless otherwise specified. The spatial encoder extracts the instance

spatial relationship f_{SP} via a small CNN. The CNN comprises two convolutional layers (the first of which uses an SE block [17] to learn more powerful features) and is trained end-to-end.

(3) Context Encoder: The *context encoder* is designed to uncover complementary cues conveyed in other regions of the image. Instead of using visual appearance to encode the contexts information directly, we change the encoding mechanism by using semantic categories. It comprises a global context encoding, i.e., the estimated scene, and a local context encoding, namely the semantic categories of other objects present in the surrounding neighbourhood of the candidate instance pair.

Global Context: For the global context representation f_G, we use a DenseNet-161 model [18] pretrained on Places365 [44] to extract scene features. After that we encode the scene class (with largest likelihood) using word2vec. The global context embedding f_G (scene feature) is therefore a 300-dimensional vector.

Local Context: During inference of the relationship between the candidate pair containing object o^i with human h, all the other detected objects o^j where $j \neq i$ in the neighbourhood can be considered to provide local context. In particular, their semantic category and position relative to the candidate object o^i are both valuable cues for distinguishing between different interactions. For example, the objects present in the neighbourhood of an 'eating a cake' interaction will likely comprise a spoon-like tool, whereas for the 'cutting a cake' interaction, it is a knife-like tool. The distance between the knife and the cake is also important (if it is far away from the cake, it is very unlikely that the knife is being used to cut the cake).

Motivated by this observation, we first use word2vec to encode the semantic knowledge of each object o^j in the neighbourhood, and then concatenate the resulting embedding with its spatial relationship f_R (computed with respect to the candidate object o^i). Following prior work [46] on visual relationship detection, the spatial geometric relationship feature f_R between candidate object o^i and its neighbour o^j is encoded as follows:

$$f_R = [(\frac{x_1^i - x_1^j}{x_2^j - x_1^j}), (\frac{y_1^i - y_1^j}{y_2^j - y_1^j}), log(\frac{x_2^i - x_1^i}{x_2^j - x_1^j}), log(\frac{y_2^i - y_1^i}{y_2^j - y_1^j})], \tag{1}$$

where $(x_1^i, x_2^i, y_1^i, y_2^i)$ and $(x_1^j, x_2^j, y_1^j, y_2^j)$ are the normalised box coordinates of the candidate object o_i and its neighbour o^j . This geometric feature f_R is a measure of the scales and relative positioning of the two object entities.

To account for the variable number of objects present in the neighbourhood of an interaction, we use NetVLAD [2] to aggregate all object representations when forming the local context embedding f_L. During the end-to-end training, the NetVLAD aggregation module can learn to discriminatively select which information should be promoted (or demoted). Finally, the output of the Context Encoder, f_C, is generated by concatenating the global context f_G and local context f_L embeddings.

(4) Motion Encoder: The *Motion Encoder* aims to infer the likely movements of humans and objects in a given image, and provide cues to detect and recognise

their interactions. We draw inspiration from the work of [36] which sought to learn models from video that were capable of synthesising plausible futures from a singe static image and in particular, the more recent work of [10] for static image flow prediction. In our work, we focus on the latter task of predicting local optical flow as a proxy for plausible future movements of objects. We adopt the Im2Flow model [10] to predict flow from the static image to encode plausible scene motion. The flow describes which pixel regions will move (together with their direction and velocity) in the instant following image capture. Concretely, we first predict the flow information of the image and then use the plausible motion encoder (a CNN with learnable parameters) to extract plausible motion features f_M. Qualitative examples of the predicted flow can be seen in Fig. 1 (right) and Fig. 2.

(5) Fusion Module: The *fusion module* combines the outputs of the backbone, spatial, context and motion encoders into a single feature embedding and uses it to predict a score for the interaction $s_{H,O}$ of the candidate instance pair $<b_H, b_O>$. The design of the fusion module is shown in Fig. 2. Specifically, we perform fusion by putting the sequence of available features $f^* = \{f_1, ..., f_k\} = \{f_H, f_O, f_{SP}, f_C, f_M\}$, one by one into GRUs [4][3]. The description of the whole scene gradually accumulate and update in the memory cell m_k (hidden state), where the lower index k is the number of reasoning steps. At each step k, we start with calculate the update gate z_k as:

$$z_k = \sigma_z(W_z f_k + U_z m_{k-1}) + b_z, \tag{2}$$

where W_z, U_z and b_z are weights and bias and σ_z is a sigmoid activation function. The update gate z_k analyzes the description of the whole scene at last step m_{k-1} and the current input feature f_k to decide how much the current step updates its memory cell. The new added information f_k at step k helping grow the description of the whole scene is computed as follows:

$$\hat{m_i} = \sigma_m(W_m f_k + U_m(r_k \circ m_{k-1}) + b_z), \tag{3}$$

where W_m, U_m and b_m are weights and bias and σ_m is a tanh activation function. $\circ$ is an element-wise multiplication. r_k is the reset gate that decides what content to forget based on the reasoning between the m_{k-1} and f_k, can be computed as

$$r_k = \sigma_r(W_r f_k + U_r m_{k-1}) + b_r, \tag{4}$$

where W_r, U_r and b_r are weights and bias and σ_r is a sigmoid activation function. Then the description of the whole scene m_k at the current step is a linear interpolation using the update gate z_k between the previous description m_{k-1} and the new added content $\hat{m_k}$:

$$m_k = (1 - z_k) \circ m_{k-1} + z_k \circ \hat{m_k}, \tag{5}$$

[3] Empirically, we observe the feature ordering to GRU is not sensitive to the HOI detection performance. So we use this order in all experiments.

where $\circ$ is an element wise multiplication. Lastly, we take the memory cell m_k at the end of sequence f^* as the final representation to predict the relation category: the output of the fusion module is the interaction score $S_{H,O}$ for the candidate pair $<b_H, b_O>$. The proposed gate and memory mechanism allows the model to dynamically select which information should be highlighted or suppressed in the final representation, rendering it more discriminative for the final objective.

3.3 Training and Inference

Since a human can concurrently perform different actions and interact with one or more objects, e.g. "eating a cake and reading a book while sitting on the coach", HOI detection represents a muli-label classification problem in which each predicate is independent and not mutually exclusive. During training, a binary sigmoid classifier is used for each predicate that minimises the cross-entropy between the prediction score $s_{H,O}$ and the ground truth label.

During inference, all human and object instances are first detected in each image. Each human and object box pair $<b_H, b_O>$ is then passed through the network to produce a score $s^a_{H,O}$ for each predicate class $a \in 1, ..., A$, where A denotes the total number of predicate classes. The final relation score is then combined with the detection scores of the human and object instances s_H and s_O that represent the detection quality of the instance boxes b_H and b_O. The final HOI score $S_{H,O}$ for the candidate box pair $<b_H, b_O>$ is then calculated as:

$$S_{H,O} = s_{H,O} \cdot s_H \cdot s_O \tag{6}$$

4 Experiments

We first introduce the dataset used, evaluation metric and implementation details in Sec. 4.1. Next, we conduct a detailed ablation study in Sec. 4.2 to verify the effectiveness of each proposed component and present some qualitative results to demonstrate the strengths and failure cases of our approach. Finally we report our HOI detection results quantitatively and compare with state-of-the-art approaches on two benchmarks: V-COCO [13] and HICO-DET[3].

4.1 Experimental Setup

Datasets: We evaluate our proposed approach on two HOI detection benchmarks: V-COCO [13] and HICO-DET [3]. **V-COCO** contains 5400 images in the training and validation split and 4946 images in the test set. It is annotated with 26 common action category labels and the bounding boxes for human and object instances. **HICO-DET** comprises 47,776 images (38,118 images in the training set and 9,658 in the test set) with more than 150K $<human, verb, object>$ triplets. It is annotated with 600 HOI categories over 80 object classes and 117 unique verbs. HICO-DET is both larger and more diverse than V-COCO.
Evaluation Metric: We follow the standard evaluation setting in [9] and use mean average precision to measure HOI detection performance. Formally, a

triplet of $<human, verb, object>$ is considered as true positive if and only if: (1) the predicted bounding box of both human and object instance overlap with the ground truth bounding box with IoU greater than 0.5, and (2) the predicted HOI matches the ground truth HOI category. For the V-COCO dataset, we evaluate mAP_{role} following [9]. For the HICO-DET dataset, mAP performance is reported for three different HOI category sets: (1) all 600 HOI categories (*Full*), (2) 138 categories with less than 10 training samples (*Rare*) and (3) the remaining 462 categories with more than 10 training samples (*Non-Rare*).

Implementation details: Following [9], we build our approach on the Faster-RCNN [32] object detector (without FPN) with the ResNet-50 backbone to enable a fair comparison with other prior approaches. The human parse masks are obtained with a Macro-Micro Adversarial Net [27], trained on the LIP dataset [27]. Object masks are obtained with Mask-RCNN [16] pretrained on MS-COCO [25]. Global scene features are extracted using a DenseNet-161 model [18] pretrained on Places365 [44]. The object and scene semantic embeddings are obtained from Google News trained word2vec embeddings. We adopt the Im2Flow model [10] to predict flow from the static image to encode plausible motion. All pretrained models described above are kept frozen during the training in this paper. More implementation details can be found in the extended arXiv version of this paper.

4.2 Ablation Studies

In this section, we empirically investigate the sensitivity of the proposed method to different design choices. As the HICO-DET dataset is both larger and more diverse than V-COCO, we perform all ablation studies on HICO-DET. We study four aspects of FCMNet: the contribution of the network components, fine-grained spatial encodings, context features and fusion strategies. Finally, we present some qualitative examples to illustrate the challenging and failure cases.

Architectural variants: We conduct an ablation study by examining the effectiveness of each proposed component in our network structure. We use combinations of human,object embeddings and coarse layout spatial configuration (instance boxes) as our baseline (Base). As shown in Table 1, each proposed module yields improved performance under all three HICO-DET settings. By looking at the first four rows in Table 1, we can observe that the fine-grained spatial configuration information contributes the most significant performance gain as an individual module, suggesting that fine-grained shape and boundary of the human parts and object can greatly help disambiguate different actions with a similar coarse spatial layout. The **FCMNet** which includes all proposed modules (human and object features, spatial encoder with coarse and fine-grained layout, context encoder, motion encoder and fusion) at the same time achieves the best performance, which verifying the effectiveness of the proposed components. More ablation studies can be found in the extended arXiv version of this paper.

Fine-grained spatial configuration encoding: We compare the performance of using different forms of fine-grained spatial information in Table 2. To

Table 1: Ablation Study on Different Network Components

Methods	Full	Rare	Non-Rare
Baseline (Base)	14.8	12.3	15.7
Base+ Fine	17.6	15.4	18.3
Base + Context	16.2	14.1	16.9
Base + Motion	16.5	14.4	17.2
FCMNet (ours)	**20.4**	**17.3**	**21.6**

Table 2: Ablation Study on Different Fine-grained Information

Methods	Full	Rare	Non-Rare
Baseline (Base)	14.8	12.3	15.7
Base +HS	15.8	12.9	16.8
Base +HP	17.1	14.7	18.0
Base +OM	16.2	14.0	17.1
Base +HP+OM	**17.6**	**15.4**	**18.3**

Table 3: Ablation Study on Different Context Information

Methods	Full	Rare	Non-Rare
Baseline (Base)	14.8	12.3	15.7
Base+Local(w2v)	15.7	13.7	16.4
Base+Global(w2v)	15.1	13.0	16.2
Base+Both(visual)	15.2	12.8	15.9
Base+Both(w2v)	**16.2**	**14.1**	**16.9**

Table 4: Ablation Study on Different Fusion Strategy

Methods	Full	Rare	Non-Rare
Baseline	14.8	12.3	15.7
Late Fusion	18.1	15.3	19.0
Concatenation	17.9	14.9	20.0
Fusion (Attention)	19.7	16.6	20.1
FCMNet (ours)	**20.4**	**17.3**	**21.6**

enable a fair comparison, we use the same model architecture, i.e., with only human and object features and a spatial encoder. We observe that each of the investigated fine-grained spatial encoding, i.e., Human Parts (HP) and Object Mask (OM), improves performance. Since prior work ([38, 14]) has shown that encoding fine-grained information via human key-point estimation is also useful to convey pose, we also compare with the Human skeleton (HS) configuration (following [38, 14]) in this table. It can be seen that using human parts information outperforms human skeletons. Using both proposed fine-grained spatial encoding HP and OM concurrently outperforming the baseline by 2.8 mAP in the *Full* setting, which demonstrates the effectiveness of the proposed encoding mechanism for fine-grained spatial configuration.

Context features: We compare the performance of using different contextual information in Table 3. It can be seen that both local context (features of the objects in the neighbourhood) and global context (the scene) contribute to improved performance. Encoding the context via word2vec outperforms the one using the visual appearance directly.

Fusion Mechanism: We compare the performance of using different fusion mechanisms in Table 4. The proposed fusion design strongly outperforms late fusion, simple concatenation and fusion with attention mechanism, demonstrating the utility of providing the model with a gate and memory mechanism for filtering its representation of the candidate pair using all available information gradually. Nevertheless, both late fusion, concatenation, fusion with attention and proposed fusion module boost performance, verifying that the different encoders capture valuable complementary information for HOI detection.

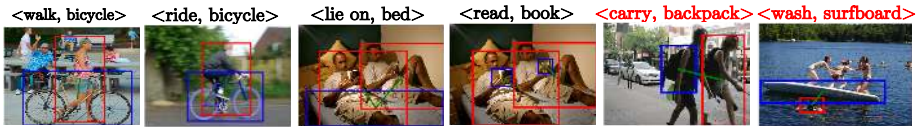

Fig. 3: Samples of human object interactions detected by FCMNet. The first four illustrate correct detections; the last two shows failure cases.

Table 5: Performance comparison on the V-COCO dataset. The scores are reported in mAP(role) as in the standard evaluation metric and the best score is marked in bold. Our approach sets a new state-of-the-art on this dataset.

Method	Feature Backbone	AP_{role}
InteractNet [12]	ResNet-50-FPN	40.0
BAR-CNN [19]	Inception-ResNet	41.1
GPNN [31]	ResNet-152	44.0
iHOI [40]	ResNet-50-FPN	45.8
Xu et al. [41]	ResNet-50	45.9
iCAN [9]	ResNet-50	44.7
Wang et al. [39]	ResNet-50	47.3
RPNN [45]	ResNet-50	47.5
Li et al. [24]	ResNet-50	47.8
PMFNet [38]	ResNet-50-FPN	52.0
Baseline (Ours)	ResNet-50	45.3
FCMNet (Ours)	**ResNet-50**	**53.1**

Qualitative Visual Examples: In Fig. 3, we present qualitative examples to illustrate strengths and failure cases on the HICO-DET dataset. We highlight the detected human and object with red and blue bounding boxes respectively. The first four samples are some challenging cases where our proposed approach produce correct detection. It indicates our model can distinguish subtle visual differences between interactions (first two) and be able to detect co-occurrence relations (third and fourth). The last two show some failure cases.

4.3 Results and Comparisons

In this section, we compare our FCMNet with several existing approaches for evaluation. We use combinations of humans, objects embeddings and coarse layout spatial configuration (instance boxes) as our baseline—the final FCMNet model integrates all the proposed modules.

For the **VCOCO** dataset, we present the quantitative results in terms of AP_{role} in Table 5. Our baseline achieves competitive performance with an AP_{role} of 45.3 (placing it above approximately half of the listed prior work). Different from those approaches, we use word2vec embeddings to represent the object rather than the visual embedding from ROI pooling, which turns out to be very effective for HOI detection in small datasets like V-COCO. Using the word2vec

Table 6: Performance comparison on the HICO-DET dataset. Mean average precision (mAP) is reported for the default and Known object setting. The best score is marked in bold. Our approach sets a new state-of-the-art on this dataset.

Method	Feature Backbone	Default			Known Object		
		Full	Rare	Non-Rare	Full	Rare	Non-Rare
InteractNet [12]	ResNet-50-FPN	9.94	7.16	10.77	-	-	-
GPNN [31]	ResNet-152	13.11	9.34	14.23	-	-	-
iHOI [40]	ResNet-50-FPN	13.39	9.51	14.55	-	-	-
Xu et al. [41]	ResNet-50	14.70	13.26	15.13	-	-	-
iCAN [9]	ResNet-50	14.84	10.45	16.15	16.43	12.01	17.75
Wang et al. [39]	ResNet-50	16.24	11.16	17.75			
Li et al. [24]	ResNet-50	17.03	13.42	18.11	19.17	15.51	20.26
Gupta et al [14]	ResNet-152	17.18	12.17	18.68	-	-	-
RPNN [45]	ResNet-50	17.35	12.78	18.71	-	-	-
PMFNet [38]	ResNet-50-FPN	17.46	15.65	18.00	20.34	17.47	21.20
Baseline (Ours)	ResNet-50	14.77	12.27	15.65	16.07	13.97	16.82
FCMNet (Ours)	**ResNet-50**	**20.41**	**17.34**	**21.56**	**22.04**	**18.97**	**23.12**

semantic embedding for the object representation enables us to leverage language priors to capture which objects might afford similar actions when the training data is limited. Our full FCMNet model (with all components proposed in Section 3) achieves 53.1 mAP, which outperforms prior approaches by a considerable margin.

For the **HICO-DET** dataset, we present quantitative results in terms of mAP in Table 6. We report results on two different settings of Default and Known Objects. Note that our baseline still performs well and surpasses nearly half of existing approaches. FCMNet improves upon our baseline by 5.64 mAP on the default setting (full split) and sets a new state-of-the-art on this dataset for both the default and Known object setting.

5 Conclusions

We have presented FCMNet, a novel framework for human object interaction detection. We illustrated the importance of the encoding mechanism for the fine-grained spatial layouts and semantic contexts, which enables to distinguish the subtle differences among interactions. We also show that the prediction of plausible motion greatly help to constrain the space of candidate interactions by considering their motion and boost performance. By combining these cues via a gate and memory mechanism, FCMNet outperforms state-of-the-art methods on standard human object interaction benchmarks by a considerable margin.

Acknowledgements. The authors gratefully acknowledge the support of the EPSRC Programme Grant Seebibyte EP/M013774/1 and EPSRC Programme Grant CALOPUS EP/R013853/1. The authors would also like to thank Samuel Albanie and Sophia Koepke for helpful suggestions.

References

1. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE transactions on pattern analysis and machine intelligence **30**(3), 555–560 (2008)
2. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307 (2016)
3. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: Hico: A benchmark for recognizing human-object interactions in images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1017–1025 (2015)
4. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
5. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3076–3086 (2017)
6. Fang, H.S., Cao, J., Tai, Y.W., Lu, C.: Pairwise body-part attention for recognizing human-object interactions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 51–67 (2018)
7. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9), 1627–1645 (Sep 2010)
8. Fouhey, D.F., Zitnick, C.L.: Predicting object dynamics in scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2019–2026 (2014)
9. Gao, C., Zou, Y., Huang, J.B.: ican: Instance-centric attention network for human-object interaction detection. In: British Machine Vision Conference (2018)
10. Gao, R., Xiong, B., Grauman, K.: Im2flow: Motion hallucination from static images for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5937–5947 (2018)
11. Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., He, K.: Detectron. https://github.com/facebookresearch/detectron (2018)
12. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8359–8367 (2018)
13. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)
14. Gupta, T., Schwing, A., Hoiem, D.: No-frills human-object interaction detection: Factorization, appearance and layout encodings, and training techniques. arXiv preprint arXiv:1811.05967 (2018)
15. Hayes, B., Shah, J.A.: Interpretable models for fast activity recognition and anomaly explanation during collaborative robotics tasks. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 6586–6593. IEEE (2017)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
17. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE transactions on pattern analysis and machine intelligence (2019)
18. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)

19. Kolesnikov, A., Kuznetsova, A., Lampert, C., Ferrari, V.: Detecting visual relationships using box attention. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2019)
20. Krishna, R., Chami, I., Bernstein, M., Fei-Fei, L.: Referring relationships. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6867–6876 (2018)
21. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
22. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982 (2018)
23. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1261–1270 (2017)
24. Li, Y.L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.S., Wang, Y., Lu, C.: Transferable interactiveness knowledge for human-object interaction detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3585–3594 (2019)
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
26. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European Conference on Computer Vision. pp. 852–869. Springer (2016)
27. Luo, Y., Zheng, Z., Zheng, L., Guan, T., Yu, J., Yang, Y.: Macro-micro adversarial network for human parsing. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 418–434 (2018)
28. Mallya, A., Lazebnik, S.: Learning models for actions and person-object interactions with transfer to question answering. In: European Conference on Computer Vision. pp. 414–428. Springer (2016)
29. Murphy, K.P., Torralba, A., Freeman, W.T.: Using the forest to see the trees: A graphical model relating features, objects, and scenes. In: Advances in neural information processing systems. pp. 1499–1506 (2004)
30. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Detecting rare visual relations using analogies. arXiv preprint arXiv:1812.05736 (2018)
31. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 401–417 (2018)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
33. Shen, L., Yeung, S., Hoffman, J., Mori, G., Fei-Fei, L.: Scaling human-object interaction recognition through zero-shot learning. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1568–1576. IEEE (2018)
34. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition (2003)
35. Villegas, R., Yang, J., Hong, S., Lin, X., Lee, H.: Decomposing motion and content for natural video sequence prediction. arXiv preprint arXiv:1706.08033 (2017)

36. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Advances In Neural Information Processing Systems. pp. 613–621 (2016)
37. Walker, J., Gupta, A., Hebert, M.: Dense optical flow prediction from a static image. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2443–2451 (2015)
38. Wan, B., Zhou, D., Liu, Y., Li, R., He, X.: Pose-aware multi-level feature network for human object interaction detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9469–9478 (2019)
39. 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. arXiv preprint arXiv:1910.07721 (2019)
40. Xu, B., Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S.: Interact as you intend: Intention-driven human-object interaction detection. IEEE Transactions on Multimedia (2019)
41. Xu, B., Wong, Y., Li, J., Zhao, Q., Kankanhalli, M.S.: Learning to detect human-object interactions with knowledge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
42. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5410–5419 (2017)
43. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5532–5540 (2017)
44. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)
45. Zhou, P., Chi, M.: Relation parsing neural network for human-object interaction detection. In: Proceedings of the IEEE international conference on computer vision (2019)
46. Zhuang, B., Liu, L., Shen, C., Reid, I.: Towards context-aware interaction recognition for visual relationship detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 589–598 (2017)