Supplementary Material: 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

1 Network Structure

Detailed network architecture can be found in Table. 1. The input of the spatial encoder is a 4 channel layout representation, i.e., coarse (two channels) and fine-grained (two channels) are stacked together, more detail is in the main paper. The input of the Context Encoder is the object representation (w2v) of the instances detected in the neighbourhood f_{others} and the spatial relationship feature f_R between each neighbour instance and the candidate object. The input of the Motion Encoder is a 3 channel input, i.e., the flow prediction from the static image. The first two channels are the motion angle $sin\theta$ and $cos\theta$, and the third channel is the optical flow magnitude. We provide an example in Figure 1. We plot the flow as arrows (only the subset with large magnitude is plotted) on the left, the corresponding 3-channel (input to motion encoder) is shown on the right. The fusion module combines the outputs of the backbone, spatial, context and motion encoders into a single compact feature embedding and predicts the interaction score. Specifically, we perform a reasoning by putting the sequence of available features $f^* = \{f_H, f_O, f_{SP}, f_C, f_M\}$, one by one into GRUs (with hidden dimension 512).

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. Besides four aspects of FCMNet we discussed in section 4.2, we also study: the feature ordering to fusion module, object encoding mechanisms and local context aggregation approaches in the supplementary. Finally, we present some qualitative example.

Feature Ordering to Fusion Module: Since our fusion module sequentially processes available cues, select the discriminative information and gradually generate the representation for the whole scene step by step, we compare different feature ordering to the fusion module in Table 2. One possible setting (specific $\rightarrow$ general) is encouraging the model to focus on the specific information to the current candidate pair f_H, f_O and f_{SP} first and then accumulate more

Table 1: Network Structure for HOI detection

Spatial Encoder Input: Layout ($64 \times 64 \times 4$)			Motion Encoder Input: Flow Prediction ($128 \times 128 \times 3$)		
Layer	Filter/Stride	Output Size	Layer	Filter/Stride	Output Size
$Conv1$	$3 \times 3/1$	$64 \times 64 \times 32$	$Conv1$	$3 \times 3/1$	$128 \times 128 \times 64$
SE	$1 \times 1/1$	$64 \times 64 \times 32$	$Pool1$	$2 \times 2/2$	$64 \times 64 \times 64$
$Pool1$	$2 \times 2/2$	$32 \times 32 \times 32$	$Conv2$	$3 \times 3/1$	$64 \times 64 \times 64$
$Conv2$	$3 \times 3/1$	$32 \times 32 \times 16$	$Pool2$	$2 \times 2/2$	$32 \times 32 \times 64$
$Pool2$	$2 \times 2/2$	$16 \times 16 \times 16$	$Conv3$	$3 \times 3/2$	$16 \times 16 \times 16$
$Flatten$	-	4096	$Flatten$	-	4096
$FC1$	-	512	$FC1$	-	512
$Output$	-	$f_{SP} \in R^{512}$	$Output$	-	$f_M \in R^{512}$
Context Encoder Input: $f_{others} \in R^{n \times 300}$, $f_R \in R^{n \times 4}$,$f_G \in R^{300}$			Fusion Module Input: $f_H \in R^{2048}$, $f_O \in R^{300}$, f_{SP}, f_C, f_M		
Layer	Input	Output Size	Layer	Input	Output Size
$Concat1$	f_{others}, f_R	$n \times 304$	$FC1$	f_H	$f_H \in R^{512}$
$NetVlad(K=3)$	-	$f_L \in R^{912}$	$FC2$	f_O	$f_O \in R^{512}$
$Concat2$	f_L, f_G	1212	GRU	f_H, f_O, f_{SP}, f_C, f_M	$m_k \in R^{512}$
$FC1$	-	512	$FC3$	-	n_{class}
$Output$	-	$f_C \in R^{512}$	$Output$	-	$S_{H,O} \in R^{n_{class}}$

(a) Visualization of the flow prediction (only the subset with large magnitude is plotted)

(b) 3-channel flow prediction (input to the motion encoder)

Fig. 1: Flow prediction from static image. (Left:) Visualisation of the flow prediction (only the subset with large magnitude is plotted). (Right): 3-channel flow prediction, which used as the input to the motion encoder.

general knowledge (semantic contexts and plausible motion) about the whole image f_C and f_M. Another option is accumulate from general to specific (general $\rightarrow$ specific), as this lets the model to obtain global scene information first and then zoom in to the candidate pair. We also test the model with random ordering of the cues (Random). Results in Table 2 shows that reasoning from specific to general performs the best. Furthermore, we also find the variance of these different settings is tiny, which indicates that the feature ordering to GRU is not sensitive to the HOI detection performance.

Object Encoding Mechanism: We compare the performance of using different forms of object encoding mechanisms in Table 3, i.e, ROI pooling fea-

Table 2: Ablation Study on Feature ordering to Fusion Module

Methods	Full	Rare	Non-Rare
Random	19.9	16.9	21.1
specific → general	**20.4**	**17.3**	**21.6**
general → specific	20.2	17.2	21.3

Table 3: Ablation Study on Different Object Encoding Mechanisms

Methods	Full	Rare	Non-Rare
Baseline (visual)	13.9	9.8	14.8
Baseline (w2v)	14.8	12.3	15.7
Base (w2v)+Local(visual)	14.4	11.0	14.9
Base (w2v)+Local(w2v)	15.7	13.7	16.4
Base (w2v)+Global(visual)	13.5	10.3	14.2
Base (w2v)+Global(w2v)	15.1	13.0	16.2
Base (w2v)+Both(visual)	15.2	12.8	15.9
Base (w2v)+Both(w2v)	**16.2**	**14.1**	**16.9**

ture (visual appearance) and word2vec (semantic category information). We use combinations of human, object embeddings f_O and coarse layout spatial configuration (instance boxes) as our baseline. By looking at the first two rows in Table 3, we can observe that using the word2vec for object embedding f_O consistently outperforms the one using the ROI pooling feature, especially for the Rare setting where we have a limited number of training samples per category. We also compare the performance of these two encoding mechanisms (visual appearance and w2v) for the local and global context information in the last 6 rows of Table 3. It can be seen that using w2v consistently outperforms the one using the visual appearance feature. This is probably due to (1) using w2v enables the model to leverage language priors to capture possible co-occurrence between objects and predicates; (2) The limited number of triplets in existing datasets is insufficient to capture the full intra-class visual variations of relationships. (3) The lower dimension of w2v and therefore a reduced risk of overfitting.

Local Context Encoding Mechanism: We compare the performance of using different forms of local context encoding mechanisms in Table 4. It can be seen that both the semantic categories (w2v) of other object present in the surrounding neighbourhood and their spatial relationship f_R contribute to improved performance. Using NetVLAD [3] to aggregate variable number of objects representation outperforms the one using simple average directly.

Visual Examples: We highlight the importance of the encoding mechanism (for representing the fine-grained spatial layout and semantic context) and the

[3] The choice of NetVLAD was inspired by its empirical effectiveness for aggregating variables numbers of objects in other tasks (retrieval) and its computationally-lightweight structure.

Table 4: Ablation Study on Different Local Context Encoding Mechanisms

Methods	Full	Rare	Non-Rare
Baseline	14.8	12.3	15.7
NetVLAD (w2v)	15.3	13.1	15.7
NetVLAD (w2v + F_R)	**15.7**	**13.7**	**16.4**
Average (w2v + F_R)	14.6	12.9	15.3

utility of plausible future movement in tackling the challenges of HOI detection in this paper. In this section, we provide more visual examples in Figure 2. From the left to right, we present input, fine-grained spatial layout, plausible motion estimation (optical flow, i.e., which regions of pixels will move), and the semantic context (both global and local). It can be seen that the fine-grained spatial layouts can greatly help disambiguate different actions with a similar coarse layout (i.e. only the bounding boxes), which renders each relation triplet more discriminative. It is worth noting that if the segmentation mask of one object is not available (segmentation mask with a low confidence score), we can use the coarse box instead. The plausible future movements of humans and objects provide information to constrain the space of candidate interaction s by considering their motion. The semantic context enables the resolution of ambiguity amongst the representations of humans and objects.

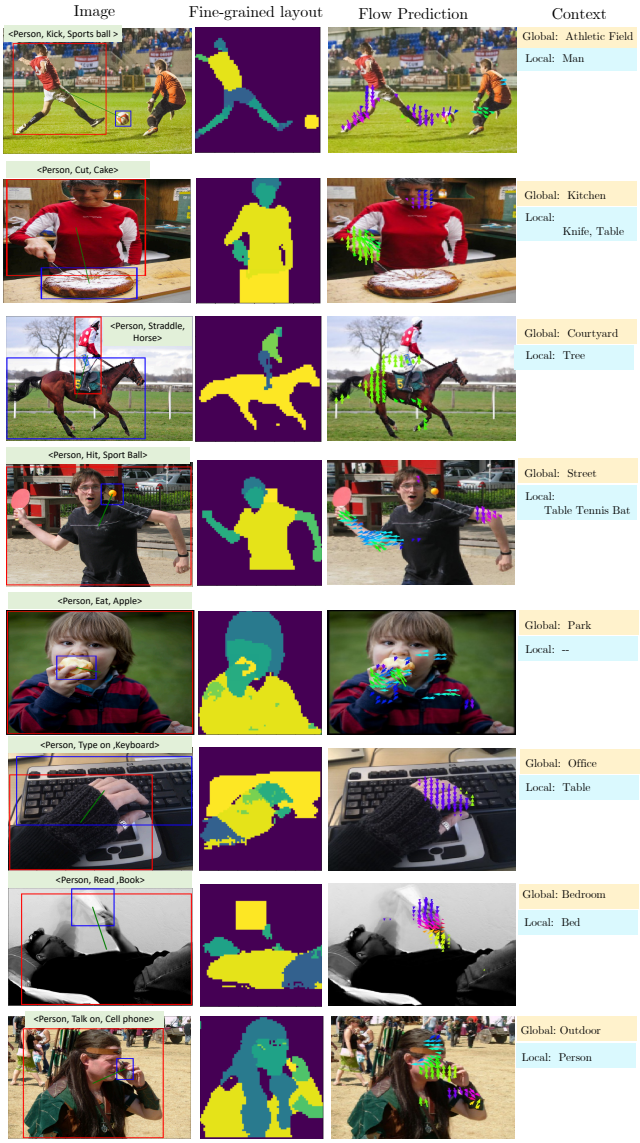

Fig. 2: We highlight the importance of the encoding mechanism (for representing the fine-grained spatial layout and semantic context) and the utility of plausible future movement in tackling the challenges of HOI detection. (First Column): The input images; (Second Column): *fine-grained spatial layouts* can greatly help disambiguate different actions with a similar coarse layout (i.e. only the bounding boxes); (Third Column): *Plausible motion* estimation distinguishes between interactions for which dynamics play an important role; (Last Column): *Global and local context* encode the scene and other local objects to provide strong clues for the interaction taking place.