Supplementary Material for Glance and Gaze: Inferring Action-aware Points for One-Stage Human-Object Interaction Detection

Xubin Zhong[1] Xian Qu[1] Changxing Ding[1,2] Dacheng Tao[3]
[1] South China University of Technology [2] Pazhou Lab, Guangzhou [3] The University of Sydney
{eexubin, eequxian.scut}@mail.scut.edu.cn, chxding@scut.edu.cn, dacheng.tao@sydney.edu.au

This supplementary material includes five sections. Section A illustrates the structure of GGNet in the inference stage. Section B conducts ablation study on the value of hyper-parameter β. Section C carries out the sensitivity analysis for λ_1. Section D shows the structure of "Variants of Feature Aggregation" in Section 5.2 of the main paper. Section E visualizes the HOI detection results of GGNet and PPDM; some failure cases of GGNet for HOI detection are also presented here.

A. Structure of GGNet in the Inference Stage

Figure 1 illustrates the structure of GGNet in the inference stage. During inference, the glance step and the first gaze step are only utilized to infer ActPoints; therefore, some layers in these two steps are removed.

B. Ablation Study on the Value of β

Experiments are conducted on the V-COCO database. The experimental results are summarized in Table 1. We can observe that the HNA loss achieves the best performance when β is set to 7.

Table 1. Ablation study on the value of β.

β	5	6	7	8
mAP$_{role}$	54.33	54.28	**54.72**	54.45

C. Sensitivity analysis for λ_1

Experiments are conducted on the V-COCO database. The experimental results are listed in Table 2. We can observe that the GGNet achieves the best performance when λ_1 is set to 0.1.

Table 2. Sensitivity analysis for λ_1.

λ_1	0.1	0.5	1
mAP$_{role}$	**54.72**	54.01	53.28

D. Model Structure of Variants for Feature Aggregation

In this section, we show the structure of "Variants of Feature Aggregation" in Table 3 of the main paper. All methods in Table 3 share the same structure of the human-object pair matching module as the baseline model. Besides, they all adopt the ordinary V-dimensional element-wise focal loss [10] for optimization.

The structure of Model "I + H" ("I + O"). The model "I + H" is illustrated in Figure 2. To obtain the human feature for each human-object pair, we first attach one human (H) branch on the backbone model. The H branch runs in parallel with the interaction point detection (I) branch. The H branch is realized using a 3×3 Conv layer with ReLU, followed by a 1×1 Conv layer and a sigmoid layer. To enable the feature maps $\mathbf{H^1}$ to be action-aware, we apply a V-dimensional element-wise focal loss to the H branch as supervision.

Next, we utilize the offset predicted by the human-object pair matching module, i.e. the point matching branch in Figure 2, to predict the human center point for each interaction point in $\mathbf{F^1}$. Features of the human center point are extracted on $\mathbf{H^1}$ using the bilinear sampling [26] and are further concatenated with the features of the interaction point. The concatenated features are processed by two successive 1×1 Conv layers for interaction prediction.

The model "I + O" can be constructed in a similar manner by replacing the above H branch with an object (O) branch. Features of the object center are utilized to augment the features of the corresponding interaction point.

The Structure of Model "I + H + O". This model can be constructed by adding both the H and O branches. The features of human center, object center, and the interaction point are concatenated for interaction prediction.

E. Qualitative Visualization Results

Figure 3 presents the qualitative comparisons between GGNet and PPDM [10] in terms of HOI detection results on HICO-DET. We can observe that PPDM fails to pre-

dict interaction categories for some images. This is because the interaction points often locate at the background area or unimportant human body area; therefore, their features are ambiguous in semantics for interaction prediction. In comparison, GGNet infers the interaction categories accurately, as the discriminative interaction areas can be captured by our proposed glance-and-gaze strategy. Qualitative comparisons on V-COCO are shown in Figure 4.

We also present some failure cases of GGNet in terms of HOI detection in Figure 5.

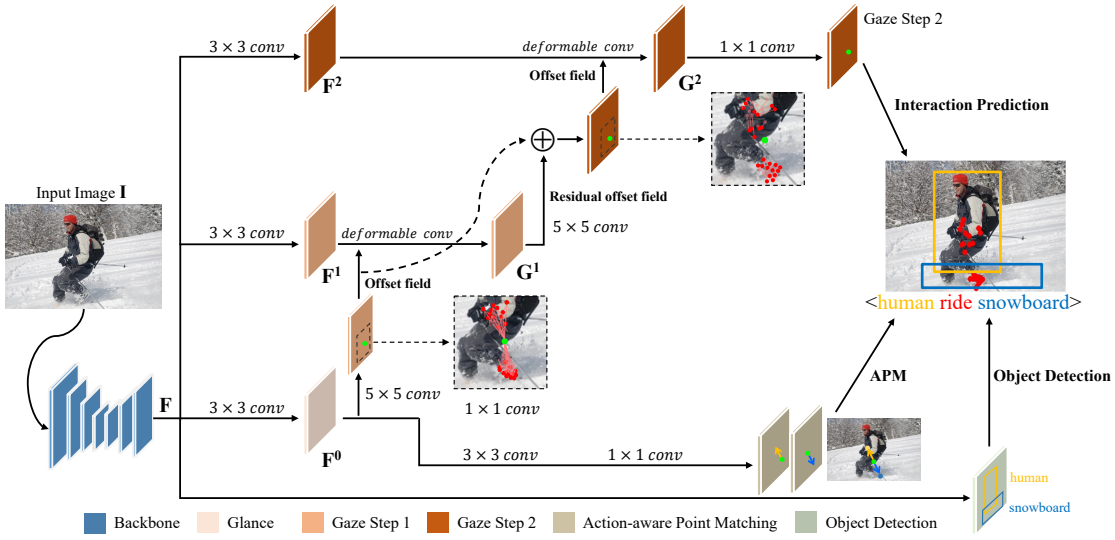

Figure 1. Overview of GGNet in the inference stage. The Glance step and Gaze Step 1 are only used to infer the ActPoints, and the irrelevant layers are discarded.

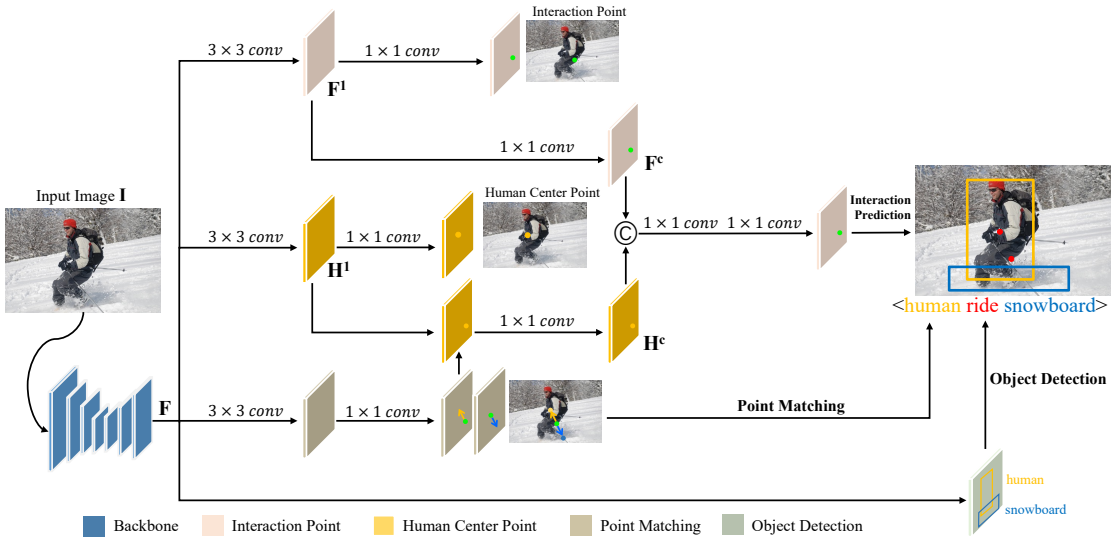

Figure 2. Overview of the model "I + H" in the training stage. The model is composed of four branches, namely the interaction point detection branch, the human branch, the point matching branch, and the object detection branch. The four branches run in parallel. Features of each interaction point and those of the corresponding human center point are concatenated for interaction prediction. Ⓒ denotes the concatenation operation in the channel dimension.

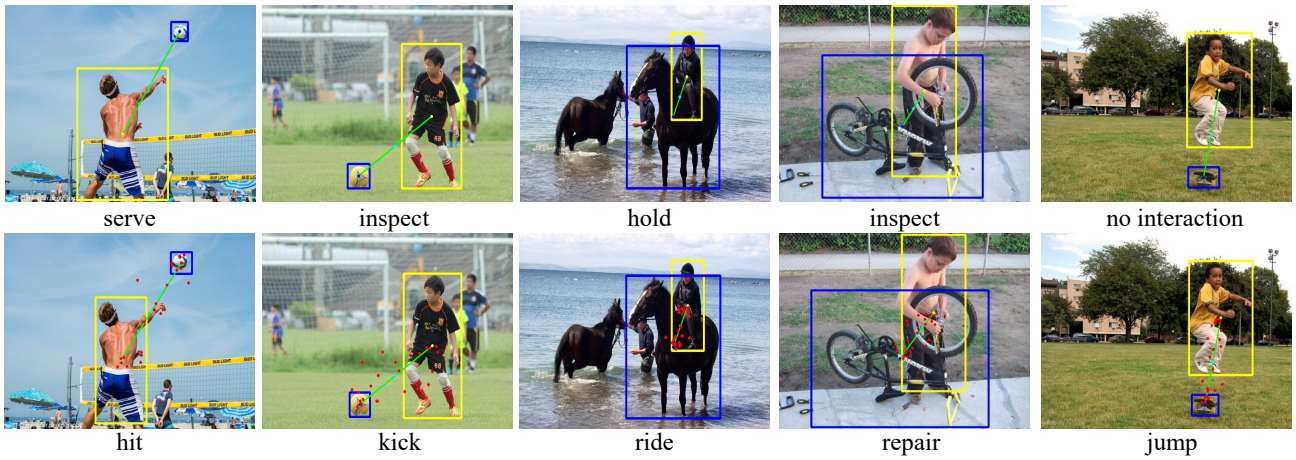

Figure 3. Qualitative comparisons between GGNet and PPDM on HICO-DET. The first and second rows show the predictions by PPDM and GGNet respectively. Cyan denotes the interaction points and red stands for ActPoints. Moreover, the human and objects are represented using yellow and blue, respectively. If a person has interaction with an object, they are linked by a green line. We show the top-1 triplet according to the prediction confidence per image.

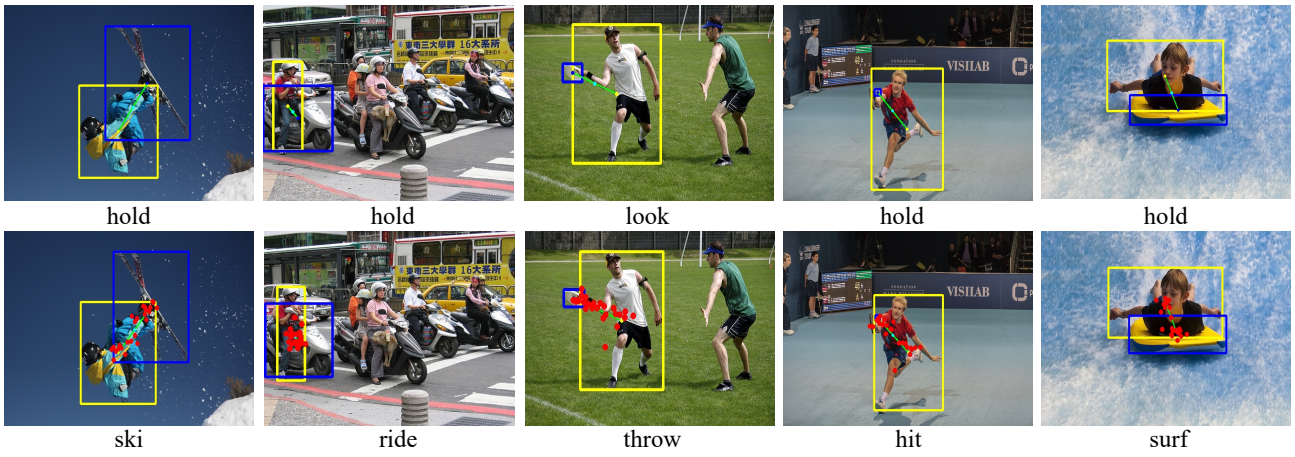

Figure 4. Qualitative comparisons between GGNet and PPDM on V-COCO.

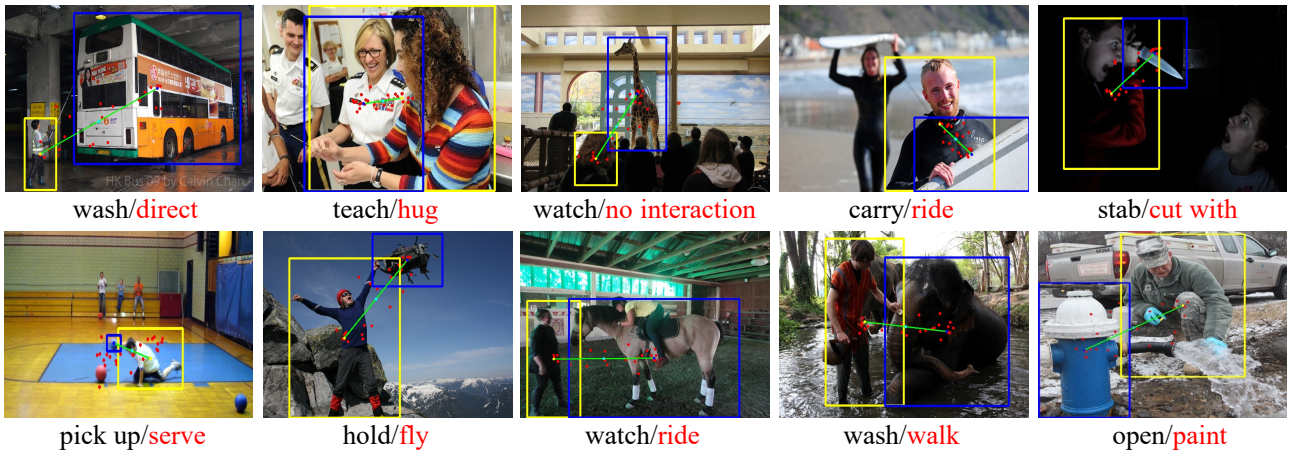

Figure 5. Failure cases of GGNet for HOI detection on HICO-DET. The ground-truth interaction and the predicted one are typed in black and red, respectively.