Consistency Learning via Decoding Path Augmentation for Transformers in Human Object Interaction Detection

Jihwan Park\textsuperscript{1,2} SeungJun Lee\textsuperscript{1} Hwan Heo\textsuperscript{1} Hyeong Kyu Choi\textsuperscript{1} Hyunwoo J. Kim\textsuperscript{1,*}

\textsuperscript{1}Department of Computer Science and Engineering, Korea University \textsuperscript{2}Kakao Brain

\{jseven7071, lapal0413, gjghks950, imhgchoi, hyunwoojkim\}@korea.ac.kr

\{jwan.park\}@kakaobrain.com

\section*{Summary}
In this supplement, we provide the implementation details of cross-path consistency (CPC) learning and additional experimental results. This includes (1) the details of CPC loss and hyperparameters, (2) results for other HOI transformers that we did not discuss in the main paper, (3) limitations of our work, (4) negative social impacts, and (5) license information.

\section{Implementation Details of CPC}
Existing HOI transformers \cite{lu2016visual, zeng2017detecting, zeng2017collaborative, yu2016modeling} have some variants in output logits. We will shortly explain the details of each model’s outputs, and specific CPC losses. For simplicity, we discuss the consistency loss between $P_k$ and $P_{k'}$ for further explanation, which is specified in the main paper as:

$$
\mathcal{L}_{P_kP_{k'}} = \lambda_h \cdot \mathcal{L}_h(\hat{y}_k^h, \hat{y}_{k'}^h) + \lambda_o \cdot \mathcal{L}_o(\hat{y}_k^o, \hat{y}_{k'}^o) + \lambda_{act} \cdot \mathcal{L}_{act}(\hat{y}_k^{act}, \hat{y}_{k'}^{act})
$$

(1)

To avoid clutter, query index $\tilde{\sigma}_{k,n}$ of each path that is matched to the same ground truth label is omitted.

\subsection{CPC loss for QPIC}
For each HOI triplet element, the outputs of QPIC on a decoding path $P_k$ are composed of box regression $\hat{b}_k^b$ for human prediction $\hat{y}_k^h$, box regression $\hat{b}_k^o$ and softmax class probabilities $\hat{c}_k^o$ for object prediction $\hat{y}_k^o$, and multi-label class probabilities $\hat{a}_k$ for interaction prediction $\hat{y}_k^{act}$. For (1), we use the following loss:

$$\mathcal{L}_{P_kP_{k'}} = \lambda_h \cdot \text{MSE}(\hat{b}_k^b, \hat{b}_{k'}^b) + \lambda_o \cdot \text{MSE}(\hat{b}_k^o, \hat{b}_{k'}^o) + \lambda_{act} \cdot \text{MSE}(\hat{c}_k^o, \hat{c}_{k'}^o) + \lambda_{act} \cdot \text{MSE}(\hat{a}_k, \hat{a}_{k'})
$$

(2)

where MSE, JSD denotes mean-squared error and Jensen-Shannon divergence. The loss weights, $\lambda_h, \lambda_o, \lambda_{act}$, are set to 2.5, 2.5, 1, and 1, respectively.

\subsection{CPC loss for HOTR}
The outputs of HOTR on a decoding path $P_k$ include softmax class probabilities $\hat{c}_k^o$ for human prediction $\hat{y}_k^h$, softmax class probabilities $\hat{c}_k^o$ for object prediction $\hat{y}_k^o$, and multi-label class probabilities $\hat{a}_k$ for interaction prediction $\hat{y}_k^{act}$. CPC loss for HOTR is

$$\mathcal{L}_{P_kP_{k'}} = \lambda_h \cdot \text{JSD}(\hat{c}_k^h, \hat{c}_{k'}^h) + \lambda_o \cdot \text{JSD}(\hat{c}_k^o, \hat{c}_{k'}^o) + \lambda_{act} \cdot \text{MSE}(\hat{a}_k, \hat{a}_{k'})
$$

(3)

The loss weights, $\lambda_h, \lambda_o, \lambda_{act}$ are set to 1.1, and 10, respectively.

\textit{Remarks.} The loss weights for CPC are hyperparameters. In our experiments, we simply adopted the loss weights of the baseline HOI detectors, i.e., supervision loss weights. For instance, $\lambda_h, \lambda_o$ are the same as the loss weights for bounding box regression in the baseline HOI detector. We believe that this is a good starting point for hyperparameter tuning.

\subsection{Weight scheduler for total CPC loss}
As we mentioned in the main paper, we use a weight scheduler $w(t)$ for the total CPC loss. Weight coefficient increases for the first few epochs ($t_{max}$) along the sigmoid-shaped function $\epsilon^{-0.5(1-x)^2}$, and then the maximum value $\lambda$ is maintained. The ramp-up function for weight coefficient can be written as

$$w(t) = \lambda \cdot \epsilon^{-0.5(1-\min(1,t/t_{max}))^2}.
$$

(4)

Table 1, and 2 show $\lambda$, $t_{max}$ used for HOTR and QPIC.

\section{Comparison with other HOI transformers}
We present the experiment results on additional HOI transformers that we did not discuss in the main paper to
<table>
<thead>
<tr>
<th>Dataset</th>
<th>epoch</th>
<th>$\lambda$</th>
<th>$t_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-COCO</td>
<td>90</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>HICO-DET</td>
<td>50</td>
<td>0.2</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 1. **Weight scheduler settings for HOTR**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>epoch</th>
<th>$\lambda$</th>
<th>$t_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-COCO</td>
<td>90</td>
<td>0.2</td>
<td>30</td>
</tr>
<tr>
<td>HICO-DET</td>
<td>90</td>
<td>0.2</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 2. **Weight scheduler settings for QPIC**

show the effectiveness of our method in Table 3, 4. Our experiments show that our CPC learning consistently improves the performance of other transformer-based HOI detectors (e.g., HoiT, and AS-Net) on V-COCO and HICO-DET.

<table>
<thead>
<tr>
<th>Method</th>
<th>V-COCO</th>
<th>HICO-DET</th>
</tr>
</thead>
<tbody>
<tr>
<td>HoiT</td>
<td>150</td>
<td>-</td>
</tr>
<tr>
<td>HoiT + ours</td>
<td>150 0.2</td>
<td>49.34</td>
</tr>
</tbody>
</table>

Table 3. **Comparison of our training strategy with HoiT on V-COCO.** * signifies our results reproduced with the official implementation codes of [11].

<table>
<thead>
<tr>
<th>Method</th>
<th>V-COCO</th>
<th>HICO-DET</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS-Net</td>
<td>90</td>
<td>28.87</td>
</tr>
<tr>
<td>AS-Net + ours</td>
<td>90 0.5</td>
<td>29.13</td>
</tr>
</tbody>
</table>

Table 4. **Comparison of our training strategy with AS-Net on HICO-DET.**

3. **Limitations**

Apart from no additional computation at inference, the training complexity of our method scales linearly as the number of paths is augmented. In addition, our method mainly targets the HOI detection task; applications on simpler yet more general tasks as image classification or object detection were not covered in our work. Regarding our decoding-path augmentation method, further discussions on how our methods should be applied to tasks with non-separable output (e.g., image classification) are required. This may be one interesting future direction of our work.

4. **Negative Social Impact**

Human-Object Interaction detection is a task that predicts human behavior and localize it. Although we used the popular and public benchmark datasets, the datasets may not sufficiently contain misrepresented population. Potentially the bias of the datasets may lead to the bias of our model predictions. Also, HOI detection can be abused to illegally monitor individual behaviors.

5. **Licence**


**References**