Response to Submission #591 Reviews

We sincerely thank reviewers and ACs for their time and efforts. Typos will be addressed and **our code will be made publicly available**.

To Reviewer 1: **R1-Q1.** One MLP sub-net for one verb ($\{T_I^{v_i}(\cdot)\}_{i=1}^n$, $\{T_D^{v_i}(\cdot)\}_{i=1}^n$). **A1.** We clarify this in two aspects. (1) **Efficiency**: With carefully designed data flow, IDN can run on a **single** GPU (training: **5.0** GB, inference: **2.5** GB). We operate all transformations **in parallel** and the inference speed is **10.04** FPS (iCAN [9]: **4.90** FPS, TIN [20]: **1.95** FPS, PPDM [21]: **14.08** FPS, PMFNet [30]: **3.95** FPS). (2) **Scalability**: We also considered an implementation which utilizes **a single MLP for all verbs**, *i.e.*, conditioned MLP functions $f_u^{v_i} = T_I(f_h \oplus f_o, f_{v_i}^{ID})$ and $f_h^{v_i} \oplus f_o^{v_i} = T_D(f_u^{v_i}, f_{v_i}^{ID})$, where $f_{v_i}^{ID}$ is the verb indicator (one-hot/Word2Vec/Glove). When facing new verbs, we just change the verb indicator instead of increasing MLPs. It works similar to zero-shot learning methods like <TAFE-Net: Task-Aware Feature Embeddings for Low Shot Learning> (CVPR2019) and <Recognizing Unseen Attribute-Object Pair with Generative Model>(AAAI2019) but performs worse (**20.86** mAP on HICO-DET Full) than reported version (**23.12** mAP). We will discuss them in the final version.

R1-Q2. Threshold and "pull and push" losses. **A2.** (1) We directly use distances in classification (L197-200, L203-204): negative distance $-d_u^{v_i}$ acts as the **binary classification score** of v_i. And the gradient will make the transformation generate reasonable $\{d_u^{v_i}\}_{i=1}^n$. $-d_{ho}^{v_i}$ works similar. Thus, we did not use thresholds to measure the distances. (2) Hence, the classification losses L_{cls}^u and L_{cls}^{ho} can *pull* the features that meet the labels together and *push* away the others.

R1-Q3. Balance between method and experiment. **A3.** Thanks very much! We will revise these sections.

To Reviewer 2: **R2-Q1.** Exponential function and hinge loss. **A1.** Thanks! We adopt the exponential function and hinge loss, and results (HICO-DET Full) are: **Sigmoid-23.12 mAP, Exponential-23.18 mAP, Hinge-23.26 mAP**.

R2-Q2. IDN applied to existing HOI methods. **A2.** We apply HOI integration and decomposition to iCAN as **proxy tasks** to enhance the feature learning. The performance improves from 14.84 mAP to **18.98** mAP (HICO-DET Full). If further combining the IDN result via late fusion, it would be boosted to 23.42 mAP. Thanks for this interesting advice!

R2-Q3. Prior work. **A3.** We will add the results of these works to Tab. 1 and 2 in the final version.

To Reviewer 3: **R3-Q1.** Clarity. **A1.** We will simplify the introduction and highlight the main idea.

R3-Q2. The importance of Inter-Pair Transformation (IPT). **A2.** IPT is essential because it reveals the inherent nature of the implicit verb of HOI: the **shared information** between different H-O pairs with the same verb. Here, we adopt a simplified implementation, *i.e.*, human/object replacement. When replacing the human/object, transformation functions $\{T_I^{v_i}(\cdot)\}_{i=1}^n$ and $\{T_D^{v_i}(\cdot)\}_{i=1}^n$ should be **equally effective** before and after instance replacement. We can also adopt more sophisticate approaches: use motion transfer [2] to adjust the human posture according to another person with same HOIs but different posture (*e.g.*, eating while sitting/standing), change the object pose, use Wor2Vec to change the object class (similar to the method of <Detecting Human-Object Interactions via Functional Generalization, AAAI 2020>), *etc*. But these are beyond the scope of our paper. In this work, we mainly explored to leverage the integration and decomposition to learn the verb representation. We will discuss more IPT implementations in our final version.

R3-Q3. Attribute-as-Operator and Red-Wine. **A3.** Thanks, we will discuss them in our final version.

To Reviewer 4: **R4-Q1.** Method exposition. **A1.** We will revise our paper to improve clarity. The points are explained as follows. (1) *Coherent HOI* carries the interaction semantics and is **more** than the sum of isolated H and O [1], *i.e.*, the *incoherent* ones. (2) *Transformation and Analogy*: we will tune down them. (3) *Eigen*: implicit structure carrying the HOI semantics. (4) *"represent verb with T_I and T_D"*: embed verbs in transformation model space. (5) *Interactive validity*: for f_u, $f_h \oplus f_o$ and $\{f_u^{v_i}\}_{i=1}^n$, we use three binary classifiers to classify them. $L_{bin}^u, L_{bin}^{ho}, L_{bin}^I$ have the same definitions ($-[ylog(p) + (1-y)log(1-p)]$). We will add illustration to the figure. (6) *Fig. 2*:"X" means it is hard to transform between HOI pairs directly. We will add descriptions of "X", g_h and g_o. (7) *Fig. 3*: we will specify the losses. Eq. 1 indicates the **ideal** transformations. In practice, we construct a loop ($f_h \oplus f_o$ to $\{f_u^{v_i}\}_{i=1}^n$ to $\{f_h^{v_i} \oplus f_o^{v_i}\}_{i=1}^n$) to train IDN with the consistency. Using f_u instead of $\{f_u^{v_i}\}_{i=1}^n$, *i.e.*, $f_h \oplus f_o$ to $\{f_u^{v_i}\}_{i=1}^n$ and f_u to $\{f_h^{v_i} \oplus f_o^{v_i}\}_{i=1}^n$, cannot form a cycle and performs worse (**21.77** mAP on HICO-DET Full). (8) *Slice*: the encoded feature $f_h \oplus f_o$ is the **sum** of isolated H and O and thus not yet integrated (t-SNE visualization in Fig. 4). In the implementation, we did not slice $f_h \oplus f_o$ but replace H/O feature before AE compression. We will clarify this in the illustration.

R4-Q2. AAAI 2020 work, arXiv citation. **A2.** More please refer to **R3-Q2**. We will discuss it and fix the arXiv citation.

R4-Q3. Scalability. **A3.** Please refer to **R1-Q1**.

R4-Q4. Ablation study. **A4.** We respectfully explain that the ablation study mainly compares the performances of IDNs in different settings. In model tuning, the test set is **unseen**. Please also refer to the ablation studies of the above AAAI 2020 work, HICO-DET [3], Shen *et al*. [28], GPNN [26], iCAN [9], No-frills [16], Peyre *et al*. [25], PMFNet [30], *etc*.