Abstract

Human-Object Interaction (HOI) consists of human, object and implicit interaction/verb. Different from previous methods that directly map pixels to HOI semantics, we propose a novel perspective for HOI learning in an analytical manner. In analogy to Harmonic Analysis, whose goal is to study how to represent the signals with the superposition of basic waves, we propose the HOI Analysis. We argue that coherent HOI can be decomposed into isolated human and object. Meanwhile, isolated human and object can also be integrated into coherent HOI again. Moreover, transformations between human-object pairs with the same HOI can also be easier approached with integration and decomposition. As a result, the implicit verb will be represented in the transformation function space. In light of this, we propose an Integration-Decomposition Network (IDN) to implement the above transformations and achieve state-of-the-art performance on widely-used HOI detection benchmarks. Code is available at https://github.com/DirtyHarryLYL/HAKE-Action-Torch/tree/IDN-(Integrating-Decomposing-Network).

1 Introduction

Human-Object Interaction (HOI) takes up most of the human activities. As a composition, HOI consists of three parts: <human, verb, object>. To detect HOI, machines need to simultaneously locate human and object and classify the verb [4]. Except for the direct thinking that maps pixels to semantics, in this work we rethink HOI and explore two questions in a novel perspective (Fig. 1): First, as for the inner structure of HOI, how do isolated human and object compose HOI? Second, what is the relationship between two human-object pairs with the same HOI?

For the first question, we may find some clues from psychology. The view of Gestalt psychology is usually summarized as one simple sentence: “The whole is more than the sum of its parts” [4]. This is also in line with human perception. Baldassano et al. [1] studied the mechanism of how the brain builds HOI representation and concluded that the encoding of HOI is not the simple sum of human and object: a higher-level neural representation exists. Specific brain regions, e.g., posterior superior temporal sulcus (pSTS), are responsible for integrating isolated human and object into coherent HOI [1]. Hence, to encode HOI, we may need complex nonlinear transformation (integration) to combine isolated human and object. Also, we argue that a reverse process is also essential to decompose HOI into isolated human and object (decomposition). Here we use $T_I(·)$ and $T_D(·)$ to indicate integration and decomposition functions. Accordingly to [1], isolated human and object are different from coherent HOI pair. Therefore, $T_I(·)$ should be able to add interactive relationship to isolated elements. On the contrary, $T_D(·)$ should eliminate this interactive information. Through the
Figure 1: Two questions about HOI. First, we want to explore its inner structure. Second, the relationship between two human-object pairs with same HOI is studied.

Before and after transformations, we can reveal the “eigen” structure of HOI carrying the semantics. Considering that verb is hard to represent explicitly in image space, our transformations are conducted in latent space. For the second question, directly transforming one human-object pair to another (inter-pair transformation) is difficult. We need to consider not only isolated element differences but also interaction pattern change. However, with $T_I(\cdot)$ and $T_D(\cdot)$, things are different. We can first decompose HOI pair-$i$ into isolated person-$i$ and object-$i$ and eliminate the interaction semantics. Next, we transform human-$i$ (object-$i$) to human-$j$ (object-$j$) ($j \neq i$). The last step is to integrate human-$j$ and object-$j$ into pair-$j$ and add the interaction simultaneously.

Interestingly, we find the above process is kind of like Harmonic Analysis: to process the signal, we usually use Fourier Transform (FT) to decompose it into the integration of basic exponential functions; then we can modulate the exponential functions via very simple transformations like scalar-multiplication; finally, inverse FT can help us integrate the modulated elements and map them back to the input space. This elegant property brings a lot of convenience for signal processing. Therefore, we mimic this insight and design our methodology, i.e., HOI Analysis. To implement HOI Analysis, we propose an Integration-Decomposition Network (IDN). In detail, after extracting the features from human/object and human-object tight union boxes, we perform the integration $T_I(\cdot)$ to integrate the isolated human and object into the union in latent space. Moreover, decomposition $T_D(\cdot)$ is then performed to decompose the union into isolated human and object instances again. Through the transformations, IDN can learn to represent the interaction/verb with $T_I(\cdot)$ and $T_D(\cdot)$. That said, we first embed verbs in transformation function space, then learn to add and eliminate interaction semantics and classify interactions during transformations. For the inter-pair transformation, we adopt a simple instance exchange policy. For each human/object, we beforehand find its similar instances as candidates and randomly exchange the original instance with candidates in training. This policy can avoid complex transformation like motion transfer. Hence, we can focus on the learning of $T_I(\cdot)$ and $T_D(\cdot)$. Moreover, the lack of samples for rare HOIs can also be alleviated. To train IDN, we adopt the objectives derived from transformation principles, such as integration validity, decomposition validity and interactiveness validity (detailed in Sec. 3.4). With them, IDN can effectively model the interaction/verb in transformation function space. Subsequently, IDN can be applied to the HOI detection task by comparing the above validities and greatly advance it.

Our contributions are threefold: (1) Inspired by Harmonic Analysis, we thereon devise HOI Analysis to model the HOI inner structure. (2) A concise Integration-Decomposition Network (IDN) is proposed to conduct the transformations in HOI Analysis. (3) By learning verb representation in transformation function space, IDN achieves state-of-the-art performance on HOI detection.

2 Related Work

Human-Object Interaction (HOI) detection is crucial for deeper scene understanding and can facilitate behavior and activity learning. Recently, huge progress has been made in this field with the promotion of large-scale datasets and deep learning. HOI has been studied for a long history. Previously, most methods adopted hand-crafted features. With the renaissance of neural networks, recent works start to leverage learning-based features with end-to-end paradigm. HO-RCNN utilized a multi-stream model to leverage human, object and spatial patterns respectively, which is widely followed by subsequent works. Differently, GPNH adopted a graph model to address HOI learning for both images and videos. Instead of directly processing all
Figure 2: HOI Analysis. $T_D(\cdot)$ and $T_I(\cdot)$ indicate the decomposition and integration. First, we decompose the coherent HOI into isolated human and object. Next, human and object can be integrated into HOI again. Through $T_D(\cdot)$ and $T_I(\cdot)$, we can model the verb in transformation function space and conduct the inter-pair transformation (IPT) more easily. Red “X” means it is hard to operate IPT directly. $g_h$, $g_o$ indicate the inter-human/object transformation functions.

human-object pairs generated from detection, TIN [27] utilized interactiveness estimation to filter out non-interactive pairs in advance. In terms of modality, Peyre et al. [41] explored to learn a joint space via aligning the visual and linguistic features and used word analogy to address unseen HOIs. DJ-RN [24] recovered 3D human and object (location and size) and learned a 2D-3D joint representation. Finally, some works also explore to encode HOI with the help of a knowledge base. Based on human part-level semantics, HAKE [25] built a large-scale part state [30] knowledge base and Activity2Vec for finer-grained action encoding. Xu et al. [53] constructed a knowledge graph from HOI annotations and the external source to advance the learning.

Besides the computer vision community, HOI is also studied in human perception and cognition researches. In [1], Baldassano et al. studied how human brain models HOI given HOI images. Interestingly, besides the brain regions responsible for encoding isolated human or object, certain regions can integrate isolated human and object into a higher-level joint representation. For example, pSTS can coherently model the HOI, instead of simply summing isolated human and object information. This phenomenon inspires us to rethink the nature of HOI representation. Thus we propose a novel HOI Analysis method to encode HOI by integration and decomposition.

On the other hand, HOI learning is similar to another compositional problem: attribute-object learning [35,34,26]. Attribute-object compositions have many interesting properties such as contextuality, compositionality [34,35] and symmetry [26]. To learn the attribute-object, attributes are seen as primitives equal with objects [34] or linear/non-linear transformations [35,26]. Different from attributes expressed on object appearance, verbs in HOIs are more implicit and hard to locate in images. They are a kind of holistic representation of composed human and object instances. Thus, we propose several transformation validities to embed and capture the verbs in transformation function space, instead of utilizing an explicit classifier to classify them [34] or using language priors [35,26].

3 Method

3.1 Overview

In an image, human and object can be explicitly seen. However, we can hardly depict which region is the verb. For “hold cup”, “hold” may be obvious and center on the hand and cup. But for “ride bicycle”, most parts of the person and bicycle all represent “ride”. Hence, vision systems may struggle given diverse interactions as it is hard to capture the appropriate visual regions. Though attention mechanism [12] may help, the long-tail distribution of HOI data usually makes it unstable. In this work, instead of directly finding the interaction region and mapping it to semantics [4,12,27,49,19], we propose a novel learning paradigm, i.e., learning the verb representation via HOI Analysis.

Inspired by the perception study [1], we propose the integration $T_I(\cdot)$ and decomposition $T_D(\cdot)$ functions. HOI naturally consists of human, object and implicit verb. Thus, we can decompose HOI
We introduce the feature preparation as follows. First, as shown in Fig. 3, IDN constructs a loop consists of two inverse transformations: integration and decomposition, which construct a loop between the union and human/object features; the second one is the inter-pair transformation like motion transfer [3], which is complex and not our main goal, we implement the human/object transformation functions. Because the strict inter-pair transformation like motion transfer [3] is complex and not our main goal, we implement the same dataset (train set) and are randomly sampled during training. From the experiment (Sec. 4.5), we find that transferring the ideal interaction representation learning.

Second, to encode the box location, we generate location features \( f^b_h, f^b_o \) where \( b \) indicates box location. We follow the box coordinate normalization method [41], getting the normalized box \( \hat{b}_h, \hat{b}_o \). Next, for union box, we concatenate \( \hat{b}_h \) and \( \hat{b}_o \) and feed them to an MLP to get \( \hat{f}_u \). For human/object box, \( \hat{b}_h \) or \( \hat{b}_o \) is also fed to an MLP to get \( \hat{f}_h \) or \( \hat{f}_o \). The size of \( \hat{f}_h \) or \( \hat{f}_o \) is half the size of \( f_h \) or \( f_o \). Third, the location features \( \hat{f}_h, f^b_h, \hat{f}_o, f^b_o \) are concatenated respectively to their corresponding appearance features \( \hat{f}_h, f^a_h, f^a_o \), getting \( \hat{f}_u, \hat{f}_h, \hat{f}_o \). The size of \( \hat{f}_h \) and \( \hat{f}_o \) are also half the size of \( f_u \). For convenience, we concatenate \( \hat{f}_h \) and \( \hat{f}_o \) as \( \hat{f}_h \oplus \hat{f}_o \).

Before transformations, we compress these features to reduce the computational burden via an auto-encoder (AE). This AE is given \( f_u \) as input and pre-trained with an input-output reconstruction loss and a verb classification loss (Sec. 3.4). The classification score is denoted as \( \hat{f}_u \). After pre-training, we use AE to compress \( f_u \) and \( \hat{f}_h \oplus \hat{f}_o \) to 1024 sized \( f_u \) (coherent) and \( f_h \oplus f_o \) (isolated) respectively. Finally, we have \( f_u, f_h \oplus f_o \) for integration and decomposition. The ideal transformations are:

\[
T_D(f_u) = f_h \oplus f_o, \quad T_I(f_h \oplus f_o) = f_u, \quad (1)
\]

where \( T_D(\cdot), T_I(\cdot) \) indicates the decomposition and integration functions, \( \oplus \) denotes the linear operation between isolated human and object features such as element-wise summation or concatenation. In most cases, concatenation performs better. As for the inter-pair transformation, we use

\[
g_h(f^a_h) = f^b_h, \quad g_o(f^a_o) = f^b_o, \quad i \neq j, \quad (2)
\]

where \( f^a_h, f^a_o \) indicate the features of human/object instances, and \( g_h(\cdot), g_o(\cdot) \) are the inter-human/object transformation functions. Because the strict inter-pair transformation like motion transfer [3] is complex and not our main goal, we implement \( g_h(\cdot) \) and \( g_o(\cdot) \) as simple feature replacement for simplicity. For human instances, we find their substitutional persons with the same HOI according to the pose similarity. As to object instances, we use the objects of the same category and similar sizes as the substitutions. All substitutional candidates come from the same dataset (train set) and are randomly sampled during training. From the experiment (Sec. 4.5), we find that this policy performs well and effectively improves the interaction representation learning.

We propose a concise Integration-Decomposition Network (IDN) as shown in Fig. 3. IDN mainly consists of two parts: the first one is the integration and decomposition transformations (Sec. 3.2) which construct a loop between the union and human/object features; the second one is the inter-pair transformation (Sec. 3.4) that exchanges the human/object instances between pairs with same HOI. In Sec 3.2, we introduce the training objectives derived from the transformation principles. With them, IDN would learn more effective interaction representations and advance HOI detection in Sec. 3.5.

### 3.2 Integration and Decomposition

As shown in Fig. 3, IDN constructs a loop consists of two inverse transformations: integration and decomposition implemented with MLPs. That is, we represent the verb/interaction in MLP weight
space or transformation function space. For each verb, we adopt a pair of appropriative MLPs as integration and decomposition functions, e.g., $T^v_i(\cdot)$ and $T^o_i(\cdot)$ for verb $v_i$. For integration, when inputting a pair of isolated $f_h$ and $f_o$, \{$T^v_i(\cdot)$\}_{n=1}^n$ integrates them into $n$ outputs for $n$ verbs:

$$f^{v_i}_{n} = T^v_i(f_h \oplus f_o),$$  \hspace{1cm} (3)

where $i = 1, 2, 3, ..., n$ and $n$ is the number of verbs, $f^{v_i}_{n}$ is the integrated union feature for the $i$-th verb. $\oplus$ indicates concatenation. Through the integration function set \{$T^v_i(\cdot)$\}_{n=1}^n, we get a set of 1024 sized integrated union features \{$f^{v_i}_{n}$\}_{i=1}^n. If the original $f_u$ contains the semantics of the $i$-th verb, it should be close to $f^{v_i}_{n}$ and far away from the other integrated union features. Second, the subsequent decomposition is depicted as follows. Given the integrated union feature set \{$f^{v_i}_{n}$\}_{i=1}^n, we also use $n$ decomposition functions \{$T^v_D(\cdot)$\}_{n=1}^n to decompose them respectively:

$$f^{h_i} + f^{o_i} = T^v_D(f^{v_i}_{n}).$$  \hspace{1cm} (4)

The decomposition output is also a set of features \{$f^{h_i} + f^{o_i}$\}_{i=1}^n, where $f^{h_i}$ and $f^{o_i}$ all have the same size with $f_h$ and $f_o$. Similarly, if this human-object pair is performing the $i$-th interaction, the original input $f_h \oplus f_o$ should be close to the $f^{h_i} + f^{o_i}$ and far from the other \{$f^{h_i} + f^{o_i}$\}_{j\neq i}.

### 3.3 Inter-Pair Transformation

Inter-Pair Transformation (IPT) is proposed to reveal the inherent nature of implicit verb, i.e., the shared information between different pairs with the same HOI. Here, we adopt a simple implementation: instance exchange policy. For humans, we first use pose estimation \[9\] to obtain poses and then operate alignment and normalization. In detail, the pelvis keypoints of all persons are aligned and all the distances between head and pelvis are scaled to one. Hence, we can find similar persons according to the pose similarity, which is calculated as the sum of Euclidean distances between the corresponding keypoints of two persons. To keep the semantics, similar persons should have at least one same HOI. Selecting similar objects is simpler, we directly choose the objects of the same category. An extra criterion is that we choose objects with similar sizes. We use the area ratio between the object box and the paired human box as the criteria. Finally, $m$ similar candidates are selected for each human/object. This whole selection is operated within one dataset. Formally, with instance exchange, Eq. 3 can be rewritten as:

$$f^{v_i}_{n} = T^v_i(g_h(f_h \oplus g_o(f_o)) = T^v_i(f^{k_1} + f^{k_2}),$$  \hspace{1cm} (5)

where $k_1, k_2 = 1, 2, 3, ..., m$ and $m$ is the number of selected similar candidates, here $m = 5$. And \{$T^v_i(\cdot)$\}_{n=1}^n and \{$T^v_D(\cdot)$\}_{n=1}^n should be equally effective before and after instance exchange. During training, we first use Eq. 5 for a certain number of epochs and then replace Eq. 3 with Eq. 5 (Sec. 4.2). When using Eq. 5, we put the original instance and its exchanging candidates together and randomly sample them. Notably, we focus on the transformations between pairs with the same HOI. The transformations between different HOIs which need to manipulate the corresponding human posture, human-object spatial configuration and interactive pattern are beyond the scope of this paper. For IPT, more sophisticated approaches are also possible, e.g., using motion transfer \[3\] to adjust
We further apply IDN to HOI detection, which needs to simultaneously locate human-object and whether an interaction exists within human-object through the integration and decomposition. So while sitting/standing), recovering 3D HOI [39, 24] and adjusting 3D pose [22, 10] to generate new images/features, using language priors to change the classes of interacted objects or HOI compositions [2, 36], etc. But these are beyond the scope of our main insight, so we leave these to the future work.

3.4 Transformation Principles as Objectives

Before training, we first pre-train AE to compress the inputs. We first feed \( f_u \) to the encoder and obtain the compressed \( f_{u\text{recon}} \). Then an MLP takes \( f_u \) as input to classify the verbs with Sigmoids (one pair can have multiple HOIs simultaneously) with cross-entropy loss \( L_{cls}^{AE} \). Meanwhile, \( f_u \) is decoded and generates \( f_{u\text{recon}} \). We construct MSE reconstruction loss \( L_{recon}^{AE} \) between \( f_{u\text{recon}} \) and \( f_u \). The overall loss of AE is \( L_{AE}^{AE} = L_{cls}^{AE} + L_{recon}^{AE} \). After pre-training, AE will be fine-tuned together with transformation modules. Next, we detail the objectives derived from transformation principles.

**Integration Validity.** As aforementioned, we integrate \( f_h \) and \( f_o \) into the union feature set \( \{ f_i^{u} \}_{i=1}^{n} \) for all verbs (Eq. 3 or 5). If integration is able to "add" the verb semantics, the corresponding \( f_i^{u} \) that belongs to the ongoing verbs should be close to the real \( f_u \). For example, if coherent \( f_u \) contains the semantics of verb \( v_p \) and \( v_q \), then \( f_i^{u} \) and \( f_x^{u} \) should be close to \( f_u \). Meanwhile, \( \{ f_i^{u} \}_{i\neq p,q} \) should be far away from \( f_u \). Hence, we can construct the distance:

\[
d_i^{u} = ||f_u - f_i^{u}||_2.
\]

For \( n \) verb classes, we can get distance set \( \{ d_i^{u} \}_{i=1}^{n} \). Considering above principle, if \( f_u \) carries the \( p \)-th verb semantics, \( d_p^{u} \) should be small, and vice versa. Therefore, we can directly use the negative distances as the score of verb classification, i.e. \( S_u^{v} = \{ -d_i^{u} \}_{i=1}^{n} \). Naturally, \( S_u^{v} \) is then used to generate verb classification loss \( L_{cls}^{u} = L_{ent}^{u} + L_{hinge}^{u} \), where \( L_{ent}^{u} \) is cross-entropy loss, \( L_{hinge}^{u} = \sum_{i=1}^{n} [y_i \max(0, d_i^{u} - t_{i)} + (1 - y_i) \max(0, t_{i} - d_i^{u})] \), \( y_i = 1 \) indicates this pair has verb \( v_i \) and otherwise \( y_i = 0 \). \( t_0 \) and \( t_1 \) are chosen following semi-hard mining strategy: \( t_1^{v} = \min_{v \neq v} (d_i^{v}) \) and \( t_0^{v} = \max_{v \neq v} (d_i^{v}) \), where \( B_i^{v} \) denotes all the pairs without verb \( v_i \) in the current mini-batch, and \( B_i^{v} \) denotes all the pairs with verb \( v_i \) in the current mini-batch.

**Decomposition Validity.** This validity is proposed to constrain the decomposed \( \{ f_h^{v} + f_o^{v} \}_{i=1}^{n} \) (Eq. 4). Similar to Eq. 6, we also construct \( n \) distances between \( \{ f_h^{v} + f_o^{v} \}_{i=1}^{n} \) and \( f_h + f_o \) as

\[
d_i^{v} = ||f_h + f_o - f_h^{v} + f_o^{v}||_2
\]

and obtain \( \{ d_i^{v} \}_{i=1}^{n} \). Again \( \{ d_i^{v} \}_{i=1}^{n} \) should obey the same principle according to ongoing verbs. Thus, we get the second verb score \( S_h^{v} = \{ -d_i^{v} \}_{i=1}^{n} \) and verb classification loss \( L_{cls}^{v} \).

**Interactivity Validity.** Interactivity [22] depicts whether a person and an object are interactive. Thus, it is False if and only if human-object do not have any interactions. As the “1+1=2” property [11], the interactivity of isolated human \( f_h \) or object \( f_o \) should be False, so does \( f_h \oplus f_o \). But after we integrate \( f_h \oplus f_o \) into \( \{ f_i^{v} \}_{i=1}^{n} \), its interactivity should be True. Meanwhile, the original union \( f_u \) should have True interactivity. We adopt one shared FC-Sigmoid as the binary classifier for \( f_h \oplus f_o \) and \( \{ f_i^{v} \}_{i=1}^{n} \). The binary label converted from HOI label is zero if and only if a pair does not have any interactions. Notably, we also adopt the interactivity validity upon decomposed \( \{ f_h^{v} + f_o^{v} \}_{i=1}^{n} \) but achieve limited improvement. To keep the model concise, we just adopt the other three effective interactivity validities hereinafter. Thus, we obtain three binary classification cross entropy losses: \( L_{bin}^{u} - L_{bin}^{h} + L_{bin}^{o} \). For clarity, we use a unified \( L_{bin} = L_{bin}^{h} + L_{bin}^{o} \).

The overall loss of IDN is \( L = L_{cls}^{u} + L_{cls}^{h} + L_{bin} \). With the guidance of these principles, IDN can well capture the interaction changes during the transformations. Different from previous methods that aim at encoding the entire HOI representations statically, IDN focuses on dynamically inferring whether an interaction exists within human-object through the integration and decomposition. So IDN can alleviate the learning difficulty of complex and various HOI patterns.

3.5 Application: HOI Detection

We further apply IDN to HOI detection, which needs to simultaneously locate human-object and classify the ongoing interactions. For locations, we adopt the detected boxes from a COCO [29].
pre-trained Faster R-CNN [43], so does the object class probability \( P_o \). Then, verb scores can be obtained from \( S^{\text{u}}_v = \{ -d^{\text{u}}_v \} \) and \( S^{\text{ho}}_v = \{ -d^{\text{ho}}_v \} \) and \( S^{\text{AE}}_v \) obtained from AE are then fed to exponential functions or Sigmoids to generate \( P^u_v = \exp(S^{\text{u}}_v) \), \( P^{\text{ho}}_v = \exp(S^{\text{ho}}_v) \) and \( P^{\text{AE}}_v = \text{Sigmoid}(S^{\text{AE}}_v) \). Since the validity losses would pull the features that meet the labels together and push away the others, thus here we directly use three kinds of distances to classify verbs. For example, if \( f_u \) contains the \( i \)-th verb, \( d^{\text{u}}_v = ||f_u - f_v||_2 \) should be small (probability should be large); if not, \( d^{\text{u}}_v \) should be large (probability should be small). The final verb probabilities is acquired via \( P_v = \alpha (P^u_v + P^{\text{ho}}_v + P^{\text{AE}}_v) \), here \( \alpha = \frac{1}{7} \). For HOI triplets, we get their HOI probabilities using \( P_{\text{HOI}} = P_v \cdot P_o \) for all possible compositions according to the benchmark setting.

### 4 Experiment

In this section, we first introduce the adopted datasets, metrics (Sec. 4.1) and implementation (Sec. 4.2). Next, we compare IDN with the state-of-the-art on HICO-DET [4] and V-COCO [18] in Sec. 4.3. As HOI detection metrics [4, 18] expect both accurate human/object locations and verb classification, the performance strongly relies on object detection. Hence, we conduct experiments to evaluate IDN with different object detectors. At last, ablation studies are conducted (Sec. 4.5).

#### 4.1 Dataset and Metric

We adopt the widely-used HICO-DET [4] and V-COCO [18]. **HICO-DET** [4] consists of 47,776 images (38,118 for training and 9,658 for testing) and 600 HOI categories (80 COCO [29] objects and 117 verbs). **V-COCO** [18] contains 10,346 images (2,533 and 2,867 in train and validation sets, 4,946 in test set). Its annotations include 29 verb categories (25 HOIs and 4 body motions) and same 80 objects with HICO-DET [4]. For HICO-DET, we use mAP following [4]: true positive needs to contain accurate human/object locations (box IoU with reference to GT box is larger than 0.5) and accurate verb classification. The role means average precision [18] is used for V-COCO.

#### 4.2 Implementation Details

The encoder of the adopted AE compresses the input feature dimension from 4608 to 4096, then to 1024. The decoder is structured symmetrical to the encoder. For HICO-DET [4], AE is pre-trained for 4 epochs using SGD with a learning rate of 0.1, momentum of 0.9, while each batch contains 45 positive and 360 negative pairs. The whole IDN is trained without IPT for 45 epochs using SGD with a learning rate of 0.1, momentum of 0.9, while each batch contains 45 positive and 360 negative pairs. The whole IDN (AE and transformation modules) is first trained without inter-pair transformation (IPT) for 20 epochs using SGD with a learning rate of 0.002, momentum of 0.9. Then we finetune IDN with IPT for 30 epochs using SGD, with a learning rate of 1e-3, momentum of 0.9. Each batch for the whole IDN contains 15 positive and 120 negative pairs.

For V-COCO [18], AE is first pre-trained for 60 epochs. The whole IDN is trained without IPT for 45 epochs using SGD with a learning rate of 0.002, momentum of 0.9. Then we finetune IDN with IPT for 20 epochs using SGD with a learning rate of 1e-3, momentum of 0.9. Each batch for the whole IDN contains 15 positive and 120 negative pairs. The whole IDN (AE and transformation modules) is first trained without IPT for 20 epochs using SGD with a learning rate of 0.002, momentum of 0.9. Then we finetune IDN with IPT for 30 epochs using SGD, with a learning rate of 1e-3, momentum of 0.9. Each batch for the whole IDN contains 15 positive and 120 negative pairs. The whole IDN (AE and transformation modules) is first trained without IPT for 20 epochs using SGD with a learning rate of 0.1, momentum of 0.9, while each batch contains 45 positive and 360 negative pairs. The whole IDN is trained without IPT for 45 epochs using SGD with a learning rate of 0.1, momentum of 0.9, while each batch contains 45 positive and 360 negative pairs.

#### 4.3 Results

**Setting.** We compare IDN with state-of-the-art [45, 4, 13, 42, 12, 27, 19, 49, 41, 48, 24, 11, 21, 53, 50, 56, 28, 51, 18, 49] on two benchmarks in Tab. 1 and Tab. 2. For HICO-DET, we follow the settings in [4]. Full (600 HOIs), Rare (138 HOIs), Non-Rare (462 HOIs) in Default and Known Object sets. For V-COCO, we evaluate \( AP_{\text{role}} \) (24 actions with roles) on Scenario 1 (S1) and Scenario 2 (S2). To purely illustrate the HOI recognition ability without the influence of object detection, we conduct evaluations with three kinds of detectors: COCO pre-trained (COCO), pre-trained on COCO and then finetuned on HICO-DET train set (HICO-DET), GT boxes (GT) in Tab. 1.

**Comparison.** With \( T_I(\cdot) \) and \( T_D(\cdot) \), IDN outperforms previous methods significantly and achieves 23.36 mAP on the Default Full set of HICO-DET [4] with COCO detector. Moreover, IDN is the first to achieve more than 20 mAP on all three Default sets without additional information used. Moreover, the improvement on the Rare set proves that the dynamically learned interaction representation can greatly alleviate the data deficiency of rare HOIs. With the HICO-DET finetuned detector, IDN also shows great improvements and achieves more than 26 mAP and further proves...
while the simple linear combination of V-COCO \cite{18}, IDN achieves 4.90 FPS, TIN \cite{27}: 1.95 FPS, PPDM \cite{28}: 14.08 FPS, PMFNet \cite{49}: 3.95 FPS). We also considered one-hot/Word2Vec \cite{33}/Glove \cite{40}. For new verbs, we just change the verb indicator instead of translating.

To verify the effectiveness of transformations, we use t-SNE \cite{31} to visualize the IPT. In detail, we randomly select a pair with verb $v$ and denote its features as $f_h, f_o, f_u$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Detector</th>
<th>Feature</th>
<th>Default</th>
<th>Full</th>
<th>Rare</th>
<th>Non-Rare</th>
<th>Full</th>
<th>Rare</th>
<th>Non-Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shen et al. \cite{45}</td>
<td>COCO VGG-19</td>
<td></td>
<td>6.46</td>
<td>4.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HO-RCNN \cite{4}</td>
<td>COCO CaffeNet</td>
<td></td>
<td>7.81</td>
<td>5.37</td>
<td>8.54</td>
<td></td>
<td>10.41</td>
<td>8.94</td>
<td>10.85</td>
</tr>
<tr>
<td>InteractNet \cite{13}</td>
<td>COCO ResNet50-FPN</td>
<td></td>
<td>9.94</td>
<td>7.16</td>
<td>10.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPNN \cite{42}</td>
<td>COCO ResNet101</td>
<td></td>
<td>13.11</td>
<td>9.34</td>
<td>14.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xu et al. \cite{53}</td>
<td>COCO ResNet50</td>
<td></td>
<td>14.70</td>
<td>13.26</td>
<td>15.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>iCAN \cite{12}</td>
<td>COCO ResNet50</td>
<td></td>
<td>14.84</td>
<td>10.45</td>
<td>16.15</td>
<td></td>
<td>16.26</td>
<td>11.33</td>
<td>17.73</td>
</tr>
<tr>
<td>Wang et al. \cite{50}</td>
<td>COCO ResNet50</td>
<td></td>
<td>16.24</td>
<td>11.16</td>
<td>17.75</td>
<td></td>
<td>17.73</td>
<td>12.78</td>
<td>19.21</td>
</tr>
<tr>
<td>TIN \cite{27}</td>
<td>COCO ResNet50</td>
<td></td>
<td>17.03</td>
<td>13.42</td>
<td>18.11</td>
<td></td>
<td>19.17</td>
<td>15.51</td>
<td>20.26</td>
</tr>
<tr>
<td>No-Frills \cite{19}</td>
<td>COCO ResNet152</td>
<td></td>
<td>17.18</td>
<td>12.17</td>
<td>18.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhou et al. \cite{56}</td>
<td>COCO ResNet50</td>
<td></td>
<td>17.35</td>
<td>12.78</td>
<td>18.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMFNet \cite{49}</td>
<td>COCO ResNet50-FPN</td>
<td></td>
<td>17.46</td>
<td>15.65</td>
<td>18.00</td>
<td></td>
<td>20.34</td>
<td>17.47</td>
<td>21.20</td>
</tr>
<tr>
<td>DRG \cite{11}</td>
<td>COCO ResNet50-FPN</td>
<td></td>
<td>19.26</td>
<td>17.74</td>
<td>19.71</td>
<td></td>
<td>23.40</td>
<td>21.75</td>
<td>23.89</td>
</tr>
<tr>
<td>Peyre et al. \cite{41}</td>
<td>COCO ResNet50-FPN</td>
<td></td>
<td>19.40</td>
<td>14.60</td>
<td>20.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VCL \cite{21}</td>
<td>COCO ResNet50</td>
<td></td>
<td>19.43</td>
<td>16.55</td>
<td>20.29</td>
<td></td>
<td>22.00</td>
<td>19.09</td>
<td>22.87</td>
</tr>
<tr>
<td>VSGNet \cite{48}</td>
<td>COCO ResNet50</td>
<td></td>
<td>19.80</td>
<td>16.05</td>
<td>20.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DJ-RN \cite{24}</td>
<td>COCO ResNet50</td>
<td></td>
<td>21.34</td>
<td>18.53</td>
<td>22.18</td>
<td></td>
<td>23.69</td>
<td>20.64</td>
<td>24.60</td>
</tr>
<tr>
<td>IDN</td>
<td>COCO ResNet50</td>
<td></td>
<td>23.36</td>
<td>22.47</td>
<td>23.63</td>
<td></td>
<td>26.43</td>
<td>25.01</td>
<td>26.85</td>
</tr>
</tbody>
</table>

Table 1: Results on HICO-DET \cite{4}. “COCO” is the COCO pre-trained detector, “HICO-DET” means the “COCO” is further fine-tuned on HICO-DET, “GT” means the ground truth boxes. Superscript $\text{D}_D$ or $\text{V}_C$ indicates that the HICO-DET fine-tuned detector from DRG \cite{11} or VCL \cite{21} is used. Indicate the affect from detections (23.36 to 26.29 mAP). Given GT boxes, the gaps among the other three methods \cite{12, 27, 41} are marginal. But IDN achieves more than 9 mAP improvement on HOI recognition solely. All these greatly verify the efficacy of our integration and decomposition. On V-COCO \cite{18}, IDN achieves 53.3 mAP on S1 and 60.3 mAP on S2, both significantly outperforming previous methods. Moreover, we also apply our IDN to existing HOI methods since its flexibility as a plug-in. In detail, we apply integration and decomposition to iCAN \cite{12} as a proxy task to enhance its feature learning. The performance improves from 14.84 mAP to 18.98 mAP (HICO-DET Full).

Efficiency and Scalability. In IDN, each verb is represented by a pair of MLPs ($T^v_i(\cdot)$ and $T^u_D(\cdot)$). To ensure the efficiency, we carefully designed the data flow to make IDN is able to run on a single GPU. All transformations are operated in parallel and the inference speed is 10.04 FPS (iCAN \cite{12}: 4.90 FPS, TIN \cite{27}: 1.95 FPS, PPDM \cite{28}: 14.08 FPS, PMFNet \cite{49}: 3.95 FPS). We also considered an implementation which utilizes a single MLP for all verbs for scalability, i.e., conditioned MLP functions $f^v_i = T_i(f_h \oplus f_o, f^u_{vi})$ and $f^u_i \oplus f^v_o = T_D(f^v_i, f^{ID}_o)$, where $f^{ID}_o$ is the verb indicator (one-hot/Word2Vec \cite{33}/Glove \cite{40}). For new verbs, we just change the verb indicator instead of increasing MLPs. It works similar to zero-shot learning like TAFE-Net \cite{52} and Nan et al. \cite{56}, but performs worse (20.86 mAP, HICO-DET Full) than the reported version (23.36 mAP).

4.4 Visualization

To verify the effectiveness of transformations, we use t-SNE \cite{31} to visualize $f_h, f_h \oplus f_o, T^v_i(f_h \oplus f_o)$ for different $v$ in Fig. 4. We can find integrated $T^v_i(f_h \oplus f_o)$ obviously closer to the real union $f_u$, while the simple linear combination $f_h \oplus f_o$ cannot represent the interaction information. We also analyze the IPT. In detail, we randomly select a pair with verb $v$ and denote its features as $f_h, f_o, f_u$. 


Assume there are \( m \) other pairs with verb \( v \), whose features are \( \{ f_v^i, f_o^i \}_{i=1}^m \). Then we calculate \( D_1^v = \sum_{i=1}^m \| f_v^i - f_o^i \|^2 \) and \( D_2^v = \sum_{i=1}^m \| T(f_v^i \oplus f_o^i) - f_v^i \|_2 + \| T(f_h^i \oplus f_o^i) - f_o^i \|_2 \). Here, \( D_1^v \) is the mean distance from \( f_v \) to \( \{ f_v^i \}_{i=1}^m \). \( D_2^v \) is the mean distance from \( T(f_h \oplus f_o) \) and \( T(f_h \oplus f_o) \) to \( f_v^i \) \((i = 1, 2, ..., m)\). If IPT can effectively transform one pair to another by exchanging the human/object, there should be \( D_1^v > D_2^v \). We compare \( D_1^v \) and \( D_2^v \) of 20 different verbs in Fig. 5. As shown, in most cases \( D_1^v \) is much larger than \( D_2^v \), indicating the effectiveness of IPT.

### 4.5 Ablation Study

We conduct ablation studies on HICO-DET [4] with COCO detector. The results are shown in Tab. 3. (1) **Modules**: The performance of each module is evaluated. \( T_I, T_D \) and \( AE \) achieve 21.26, 21.05, 17.27 mAP respectively and show complementary property. (2) **Objectives**: During training, we drop one of the three validity objectives respectively. Without anyone of them, IDN shows obvious degradation, especially integration validity. (3) **Inter-Pair Transformation (IPT)**: IDN without IPT achieves 22.63 mAP, showing the importance of instance exchange policy. (4) **AE**: AE is pre-trained with: reconstruction loss \( L_{rec}^{AE} \) and verb classification loss \( L_{cls}^{AE} \). The removal of \( L_{cls}^{AE} \) hurts the performance more severely, especially on the Rare set, while \( L_{rec}^{AE} \) also plays an important auxiliary role in boosting the performance. (5) **Transformation Order**: In practice, we construct a loop \( \{ f_h \oplus f_o \} \subset \{ f_v \}_{i=1}^n \) and \( \{ f_v \}_{i=1}^n \) to train IDN with the consistency. Using \( f_v \) instead of \( f_h \oplus f_o \) to \( \{ f_{h_v} \}_{i=1}^n \) and \( f_v \) to \( \{ f_h \oplus f_{o_v} \}_{i=1}^n \) performs worse (21.77 mAP).

### 5 Conclusion

In this paper, we propose a novel HOI learning paradigm named HOI Analysis, which is inspired by Harmonic Analysis. And an Integration-Decomposition Network (IDN) is introduced to implement it. With the integration and decomposition between the coherent HOI and isolated human and object, IDN can effectively learn the interaction representation in transformation function space and outperform the state-of-the-art on HOI detection with significant improvements.
Broader Impact

In this work, we propose a novel paradigm for Human-Object Interaction detection, which would promote human activity understanding. Our work could be useful for vision applications, such as the health care system in an intelligent hospital. Current activity understanding systems are usually computationally expensive and require high computational resources, and could cost many financial and environmental resources. Considering this, we will release our code and trained models to the community, as part of efforts to alleviate the repeated training of future works.

Acknowledgments and Disclosure of Funding

This work is supported in part by the National Key R&D Program of China, No. 2017YFA0700800, National Natural Science Foundation of China under Grants 61772332 and Shanghai Qi Zhi Institute, SHEITC (2018-RGZN-02046).

References

[1] Christopher Baldassano, Diane M Beck, and Li Fei-Fei. Human-object interactions are more than the sum of their parts. Cerebral Cortex, 2017.


