Effects of Motion-Relevant Knowledge From Unlabeled Video to Human–Object Interaction Detection

Xue Lin, Qi Zou, Xixia Xu, Yaping Huang, and Ding Ding

***Abstract*— The existing works on human–object interaction (HOI) detection usually rely on expensive large-scale labeled image datasets. However, in real scenes, labeled data may be insufficient, and some rare HOI categories have few samples. This poses great challenges for deep-learning-based HOI detection models. Existing works tackle it by introducing compositional learning or word embedding but still need large-scale labeled data or extremely rely on the well-learned knowledge. In contrast, the freely available unlabeled videos contain rich motion-relevant information that can help infer rare HOIs. In this article, we creatively propose a multitask learning (MTL) perspective to assist in HOI detection with the aid of motion-relevant knowledge learning on unlabeled videos. Specifically, we design the appearance reconstruction loss (ARL) and sequential motion mining module in a self-supervised manner to learn more generalizable motion representations for promoting the detection of rare HOIs. Moreover, to better transfer motion-related knowledge from unlabeled videos to HOI images, a domain discriminator is introduced to decrease the domain gap between two domains. Extensive experiments on the HICO-DET dataset with rare categories and the V-COCO dataset with minimum supervision demonstrate the effectiveness of motion-aware knowledge implied in unlabeled videos for HOI detection.**

***Index Terms*— Domain adaptation, human–object interaction (HOI), multitask learning (MTL), self-supervised learning (SSL).**

I. Introduction

HUMAN–OBJECT interaction detection task, as a subtask of visual relationship detection, aims to localize all humans and objects, and infer the interactions between them, i.e., ⟨human, verb, object⟩ triplets, from an input image. HOI detection is critical for many multimedia tasks, such as activity analysis [1], visual question answering [2], [3], robotic manipulation [4], [5], and weakly supervised object detection [6], [7].

While object detection and action classification, the most relevant areas of HOI detection, have achieved great progress with the development of deep learning, HOI detection is still challenging. One main challenge is the overreliance on a large number of labeled samples since even small sets of verbs and objects create a huge verb-object combination. Although some HOI detection datasets, such as V-COCO and HICO-DET, are available, it is highly unlikely that the training dataset can uniformly cover all these combinations. For example, it is frequent that a person drives or gets on the bus, but it is rare that the people wave the bus. The lack of labeled training samples will lead to inferior HOI detection performance. Thus, it is a challenge to subtly recognize the rare interaction combinations, such as human wave bus, according to the general human and object, as shown in Fig. 1(b).

Existing works detect the human–object interactions (HOIs) with fully supervised learning [8]–[13], but the performance still has much room to improve (e.g., the best mAP on the rare set of HICO-DET dataset is still lower than 20%). One way to improve the performance is to exploit external knowledge from other modalities, such as linguistics [8]. However, it excessively relies on off-the-shelf well-learned knowledge. Another way is to design more complicated modules [9] but with increasing costs of training and complexity. Some works [10]–[13] explore compositional learning to tackle the problem of limited training samples but still need manually annotating large-scale data of actions and objects. Thus, we aim to address the problem with unlabeled videos that are easy to obtain with the rapid development of multimedia and contain motion-relevant knowledge. Taking human learning as an example, people can build up motion-aware visual representations through constant observation in the open world, as illustrated in Fig. 1(a), which will facilitate the detection of rare HOI categories, e.g., people wave bus, in static images, as shown in Fig. 1(b). In light of this, we aim to learn motion-relevant knowledge from unlabeled videos in a self-supervised manner to guide the detection of rare interactions for common objects.

Furthermore, it is noticeable that there exists a domain gap, including various scenes, illuminations, styles, and scales, between the unlabeled videos and HOI images, as shown in Fig. 1. That will greatly degrade the transfer performance of motion-relevant knowledge from auxiliary unlabeled videos to HOI images. Considering this, we propose to diminish the domain gap via adversarial learning. More specifically, we formulate our task as a multitask learning (MTL) problem with the help of motion-relevant knowledge learning on unlabeled

Manuscript received February 10, 2021; revised August 6, 2021 and November 9, 2021; accepted November 22, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61906013 and Grant 62106017 and in part by the Fundamental Research Funds for the Central Universities under Grant 2019JBM019. *(Corresponding author: Qi Zou.)*

The authors are with the Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China (e-mail: 18112028@bjtu.edu.cn; qzou@bjtu.edu.cn; 19112036@bjtu.edu.cn; yphuang@bjtu.edu.cn; dding@bjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3131154.

Digital Object Identifier 10.1109/TNNLS.2021.3131154

Fig. 1. Illustration of our proposed model for HOI detection. (a) Auxiliary unlabeled video dataset contains rich spatiotemporal information complementary to HOI images. (b) Motion-relevant knowledge implied in videos benefits the **rare** HOI category detection, e.g., people wave bus.

videos and domain adaptation in a mid-level feature space similar to [14].

To the best of our knowledge, this is the *first* time that the motion-related knowledge obtained from unlabeled videos is exploited to promote the detection of HOI categories. Moreover, the learned motion-relevant knowledge can greatly facilitate the HOI detection with minimum supervision, in which only few or rare labeled training samples of most categories are available. Specifically, for appearance representations, we propose a novel appearance reconstruction loss (ARL) to enforce the model to focus on the obvious dynamics. For motion pattern, we introduce a sequential motion mining (SMM) module on the basics of temporal coherence in the video. All of these will help learn the general motion-relevant knowledge and further compensate for the rare samples in the HOI detection task. A common backbone, e.g., ResNet-50, is shared to extract features of two domains covering unlabeled videos and HOI images. The proposed domain discriminator tries to distinguish which domain the features come from, while the backbone network aims to confuse the discriminator. Thus, the domain-invariant features in regard to videos and HOI images will be obtained and further facilitate the motion-related knowledge transfer from the unlabeled videos to HOI images, leading to the advancement of HOI detection performance.

We summarize our contributions as follows.

1) We creatively propose to transfer motion-aware knowledge from unlabeled videos to HOI images in an MTL manner, which can effectively compensate for insufficient training samples in HOI detection.
2) We propose a novel motion-relevant knowledge learning strategy, including ARL and SMM, to exploit the spatiotemporal coherence of video in a self-supervised manner.
3) We introduce a domain discriminator to decrease the domain shift between unlabeled videos and HOI images to obtain domain-invariant features, which will further benefit the motion-related knowledge transfer of two domains.
4) We perform extensive experiments on the HICO-DET dataset with rare categories and V-COCO dataset with minimum supervision, i.e., only 7% of labeled HOI images, to validate the effectiveness of our approach. Exploiting unlabeled videos with domain adaptation, our approach achieves competitive performance compared with the state-of-the-art methods.

The remainder of this article is organized as follows. Section II describes some works related to our work. In Section III, we describe our proposed approach. In Section IV, we demonstrate the performance of our approach by comparing it with the state-of-the-art algorithms. This article concludes with a discussion of our work in Section V.

II. Related Works

A. Human–Object Interaction Detection

Different from general visual relationships, which focus on two arbitrary objects in the images [12], [15], human interaction is human-centric with fine-grained labels that can be roughly classified into human–human interaction and HOI. Human–human interaction involves at least two individual actions from multiple persons, which contains two-person interaction [16] and multiple-person interaction (i.e., group activity) [16], [17]. HOI is labeled with fine-grained verb-object labels. With humans as a subject, the interactions with objects are more fine-grained and diverse than other general objects. In this work, we mainly focus on HOI detection. HOI detection is essential for understanding human activity in a complex scene. In recent years, several human–object interaction datasets, such as V-COCO [18] and HICO-DET [19], are developed by some researchers for the exploration of HOI detection. Early studies tackle HOIs recognition by utilizing multistream information, including human, object appearance, and spatial information, which can be further divided into two categories: with well-designed complicated architectures and with well-trained extra knowledge.

1) With Well-Designed Complicated Architectures: Chao *et al.* [19] propose a multistream model to aggregate human, object, and spatial configuration information to achieve HOIs' detection. Gkioxari *et al.* [20] introduce an action-specific density map estimation method based on detected human appearance to locate objects that interacted with human. Gao *et al.* [21] introduce an instance-centric attention network (iCAN) to highlight the information from the interest region for detecting HOIs. Wang *et al.* [22] propose a contextual attention framework for HOI detection, which can adaptively select relevant instance-centric context information to highlight image regions that likely contained HOIs. Ulutan *et al.* [23] fully utilize relative spatial reasoning and structural connections between objects to detect HOIs. Qi *et al.* [24] propose graph parsing neural network (GPNN) to model the structured scene into a graph and propagate messages between each human and object node, offering a generic HOI representation that applies to both static images and dynamic videos. A similar human-centric work focusing on dynamic videos is human–human interaction recognition.

For example, Shu *et al.* [17] propose a novel hierarchical long short-term concurrent memory (H-LSTCM) to model the long-term interrelated dynamics among a group of persons for recognizing the human interactions, which may be helpful to HOI detection in videos.

2) With Well-Trained Extra Knowledge: There have been several attempts that use extra knowledge, such as word embedding, human pose, and human body part states [25], for detecting human–object interaction. Fang *et al.* [26] exploit the pairwise human parts' correlation to help solve HOIs' detection problem. Xu *et al.* [27] propose a human intention-driven HOI detection (iHOI) framework to model human pose with the relative distances from body joints to the object instances. Li *et al.* [28] explore interactiveness prior existed in multiple datasets and combine human pose and spatial configuration to form a pose configuration map. Zhou and Chi [29] propose a novel model named relation parsing neural network (RPNN) to detect HOIs, which addresses the HOI detection problem by modeling two attention-based graphs according to human pose. Wan *et al.* [30] propose a multilevel relation detection strategy that utilizes human pose cues to capture global spatial configurations of relations and as an attention mechanism to dynamically zoom into relevant regions at the human part level. Xu *et al.* [8] introduce a knowledge graph based on the ground-truth annotations of training dataset and external source, which leads to an enhanced semantic embedding space for HOI detection by multimodal learning. There also exist some works [31], [32] using language priors to handle the verb/action polysemy problem in HOI detection. Li *et al.* [25] introduce a large-scale knowledge-based PaStaNet based on human body part states (PaSta) and reason out the activities based on the inferring human part states. Except for the successful application of extra knowledge on HOI detection, it is also widely used in visual relationships' detection [15] and human–human interaction recognition [16]. Zhuang *et al.* [15] use word2vec to encode context into a semantic space and combine with the interaction to explicitly construct an interaction classifier, which is also popular in HOI detection. Kong *et al.* [16] present high-level semantic knowledge and interactive phrases to build primitives for representing human interactions. Although the approaches using extra knowledge get good performance in HOI detection, they excessively rely on a well-trained model or exhaustive labeled annotations.

In addition to the two-stage multistream approaches, which are sequential and separated, another emerging direction is regarding HOI detection as an interaction point detection [33], [34] and matching problem [34]. Although they get good performance on HOI detection, they require a more complicated model, e.g., Hourglass-104, as a feature extractor, which is hard to train. Although the existing works on HOI detection get good performance by means of extra knowledge or more complicated architecture, most of them do not consider the problem of insufficient training samples in reality. In contrast, we propose a novel MTL perspective to settle the above problem with the aid of free unlabeled videos.

B. Multitask Learning

MTL aims to improve the performance of each task, when compared to training a separate model for each task [35], [36]. It can be considered an approach to inductive knowledge transfer that improves generalization by sharing the domain information between complimentary tasks. This can be achieved by using a shared representation to learn multiple tasks to help other tasks' learning [37]. MTL has been used for a variety of vision problems, including semantic segmentation [38], [39], facial landmark detection [40], pose estimation [41], robot manipulation [42], [43], and surface normal and depth prediction [44]. Inspired by these works, we propose a novel motion-relevant knowledge learning method and jointly train it with HOI detection in an MTL manner for obtaining generalizable visual features. We demonstrate that our MTL approach learns better representations compared to the single-task learning on human–object interaction detection.

C. Self-Supervised Learning

Since it takes a lot of human efforts to create high-quality annotations for supervised learning, self-supervised learning (SSL) constructs pretext tasks by discovering supervisory signals directly from the input data itself without additional human effort [45], [46]. In computer vision studies, different types of information have been adopted as a signal for SSL, including colorization [47], inpainting [48], and spatial patches [49], [50]. In addition to SSL on images, another line of research is the visual feature representation learning on unlabeled videos by exploring temporal clues [51], [52], which has been applied in object detection [51], image classification [51], action recognition [53]–[56], and video prediction [57], [58]. Inspired by the successful applications of SSL, we propose a novel motion-relevant knowledge learning strategy, including ARL and SMM, to learn more generalized action features for boosting the performance of HOI detection. The work similar to our ARL is EPVA [57]. Both of them use LSTM-like tools to minimize the L1 or L2 loss between the prediction and the ground truth. The differences between EPVA and ours can be concluded as follows. First, different from EPVA whose predictor LSTM does not consider the spatial structure of video data, the ConvLSTM adopted in our article can capture the spatial–temporal features of videos. Second, EPVA has no auxiliary task, so the encoder and predictor are directly supervised by the task-relevant goal to produce reliable frame prediction. However, in our work, the predictor is optimized by an indirect task-irrelevant goal to better express the video dynamics, which can help obtain more general representations. Third, EPVA aims to discover high-level features for video prediction in a single-task manner, whereas our ARL learns general motion-relevant knowledge from videos for HOI detection in a multitask manner. In addition, our proposed SMM is inspired by order prediction network (OPN) [51] and odd-one-out network (O3N) [55]. All of them leverage the spatiotemporal coherence in videos as a supervisory signal for representation learning. The OPN compares all pairs of frames to reason the chronological order

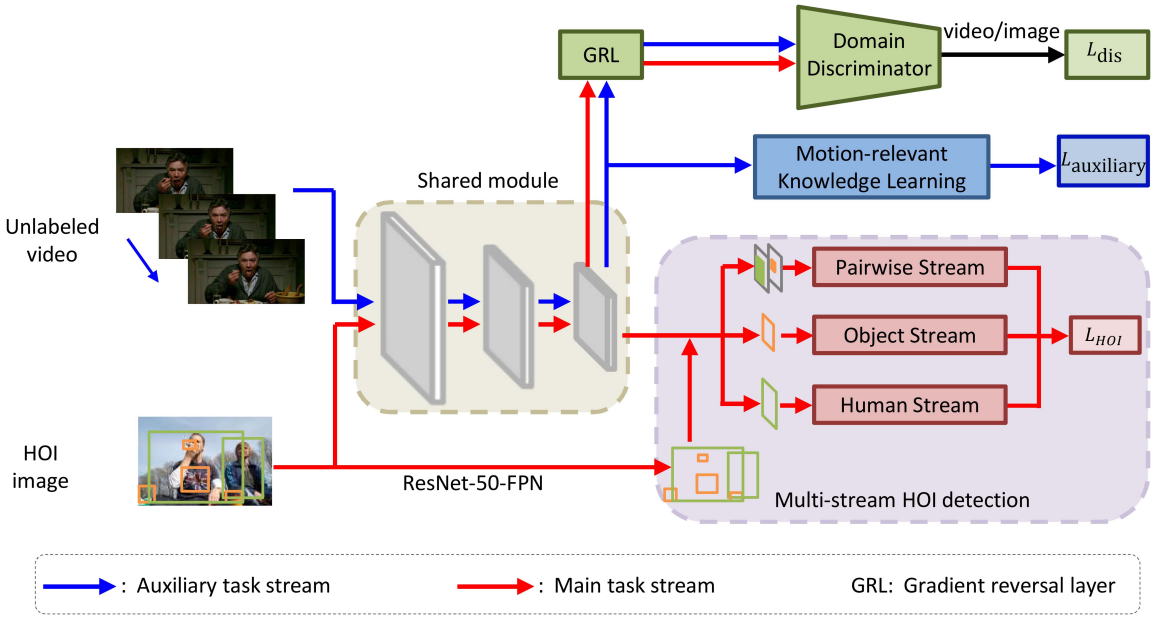


Fig. 2. Overview of our framework for HOI detection. The main task, HOI detection (III-A1), fed with static HOI images, is a common multistream architecture, such as previous work [21]. The auxiliary task stream (III-A2) takes the unlabeled videos as input and learns the motion-relevant knowledge in an SSL manner. Note that the backbone module is shared by the above two tasks. In addition, the feature representations of videos and HOI images, obtained from the shared backbone, are fed into a domain discriminator to reduce the domain gap (III-B).

of the sequence. The O3N compares all video clips to find regularities and picks the one with irregularities, whereas our proposed SMM can be regarded as a simple two-way classification problem, distinguishing if the video clips are consecutive or not.

D. Domain Adaptation

Domain adaptation was first introduced in [59] to eliminate the dataset bias and promote the performance of the target domain by leveraging common knowledge from the source domain. With the advancement of CNNs, many approaches reduce domain shift by learning domain-invariant features, including criterion- [60], [61] and adversarial learning-based [62]–[65] methods. The former aligns the domain distributions by minimizing some statistical distances across the domains, and the latter introduces the domain classifier to construct minimax optimization with the feature extractor. Different from the general domain adaptation that has exactly the same categories, our source domain, i.e., unlabeled videos, has no labels and is even irrelevant to the target domain, i.e., the HOI dataset. The difficulty of our task inspires us to design a specific domain discriminator to learn the domain-invariant features for better knowledge transfer from unlabeled videos to HOI images in an adversarial learning manner.

III. Proposed Method

Our proposed approach contains auxiliary task stream, motion-relevant knowledge learning, main task stream, and HOI detection, as shown in Fig. 2. For the main task HOI detection, following the setting of [21], we use Faster R-CNN [66] from Detectron [67] with ResNet-50-FPN [68] to obtain all the human/object instances and their corresponding detection scores. Then, a multistream architecture, including human, object, and pair, and late fusion strategy are used to achieve interaction prediction. Specifically, we propose a multitask visual learning perspective to compensate for insufficient training samples in HOI detection. Concretely, the motion-relevant knowledge learning by exploiting the spatiotemporal coherence on unlabeled videos, as the auxiliary task, is jointly trained with HOI detection in an end-to-end way, achieving the action-related knowledge transfer from videos to HOI images. Moreover, a domain discriminator is introduced to decrease the domain shift between videos and HOI images, which can greatly promote knowledge transfer across two domains.

A. Multitask Visual Representation Learning

Due to the fact that the number of training samples per class is limited in reality, the learned feature representations are nondiscriminative for HOI detection. Fortunately, the unlabeled videos relevant to action have some similarities with HOI and are freely available. Therefore, we propose a multitask feature learning perspective to compensate for the lack of training samples. In particular, a novel motion-relevant knowledge learning strategy, including ARL and SMM, is designed to exploit the spatiotemporal coherence of videos in a self-supervised manner and jointly trained with the main task, i.e., HOI detection, to improve the performance of the main task, instead of the averaged performance of all main and auxiliary tasks.

1) HOI Detection as Main Task: For our main task HOI detection, following [21], we employ ResNet-50 as the feature extraction backbone. Specifically, we denote ResNet-50 as a visual feature extractor with five blocks, where $F(\cdot)$ means the intermediate features of the third block. That is to say,

the blocks b_1, b_2, and b_3 are used to extract mid-level features of input HOI images $\mathbb{I} = \{I_1, \ldots, I_m, \ldots, I_M\}$ and output the feature representations $F(\mathbb{I})$, where M represents the number of HOI images in the main task. Then, the blocks b_4 and b_5 are used to further extract the high-level features of the human and object bounding boxes.

Then, a multistream architecture is built to explore the influence of human, object, and pairwise spatial information. Let s_h and s_o denote the confidence for the individual object detections of human and object, respectively. The interaction prediction based on the appearance of the person is represented as s_h^a and the object s_o^a. The score prediction based on the spatial relationship between the person and the object is denoted as s_{sp}^a. The input of spatial relationship is a two-channel tensor consisting of a human map and an object map. Human and object maps are all 64×64 and obtained from the human–object union box. In the human channel, the value is 1 in the human bounding box and 0 in other areas. The object channel is similar, which has value of 1 in the object bounding box and 0 elsewhere. Since a person can concurrently perform different actions to one or multiple target objects, HOI detection is, thus, a multilabel classification problem. We apply binary sigmoid classifiers for each action category and minimize the binary cross entropy losses between action scores s_h^a, s_o^a, and s_{sp}^a and the ground-truth action labels for each action category, denoted as L_H, L_O, and L_{HO}, respectively. Thus, the HOI detection loss is commonly formulated as follows:

$$L_{\text{HOI}} = \alpha * L_H + \beta * L_O + \gamma * L_{\text{HO}} \tag{1}$$

where α, β, and γ are hyperparameters that control the influence of human, object, and pairwise spatial information. Following [21], [28], we set $\alpha = 2$ and $\beta = \gamma = 1$, respectively, in all experiments.

2) Motion-Relevant Knowledge Learning as Auxiliary Task: Unlabeled videos, related to human action, are freely available and have rich spatiotemporal information that shares similar human action characteristics with HOI samples. We propose a novel motion-relevant knowledge learning strategy in a self-supervised manner, containing ARL for spatial–visual representations and SMM for temporal motion patterns. There are several ways to capture the temporal structure of an unlabeled video, such as 3-D convolutions, recurrent encoders, gated recurrent unit (GRU), and LSTM, any of which can also be used in our SMM module. Unlike the other task, such as action recognition whose goal is only to predict the action categories of the input images, HOI detection not only needs to localize all humans and objects but also infers the interactions between them. Therefore, we design ARL based on ConvLSTM to learn the more fine-grained appearance information, which benefits distinguishing whether the objects are interactive or not in HOI detection. Assuming that a video $\mathbb{V}$ is comprised of N successive frames $\{V_1, \ldots, V_n, \ldots, V_N\}$, the feature backbone, shared with HOI images, takes the frame V_n as input and outputs the features $F(V_n)$ frame by frame, as shown in Fig. 3. Then, the ConvLSTM is introduced to capture the action dynamics and predict the future. Concretely, the ConvLSTM is fed with the frame features $F(V_n)$ and

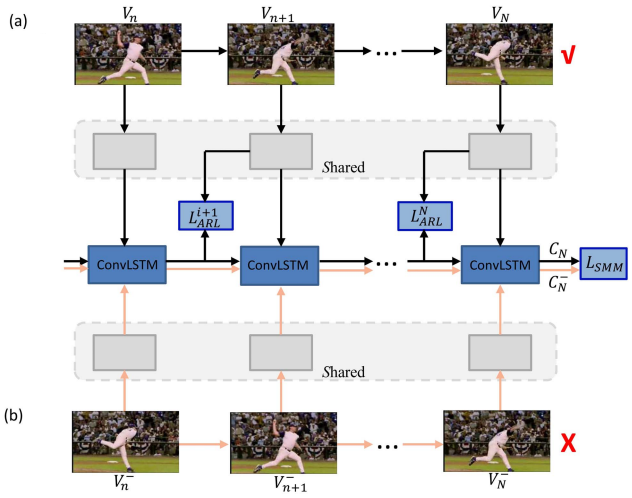


Fig. 3. Diagram of motion-relevant knowledge learning on unlabeled video. (a) Consecutive video frames $\mathbb{V}$ are fed into the shared backbone to generate mid-level features frame by frame. Then, the ConvLSTM takes the obtained representations as input and outputs the feature of the next frame. The ARL L_{ARL} is introduced to satisfy that the predicted feature and ground-truth feature should be similar as possible. (b) Muddled sequence (out-of-order) $\mathbb{V}^-$ is fed into the ConvLSTM to obtain the final memory cell that contains the history information until the last time. Trying to distinguish whether the sequences are correct or not, i.e., the proposed SMM loss L_{SMM}, can help the model learn the general motion-related features.

generates the predictions as follows:

$$[C_n, H_n] = \text{ConvLSTM}(F(V_n), C_{n-1}, H_{n-1}) \tag{2}$$

where C_n is the memory cell that retains information from the history of the inputs until time n and H_n represents the hidden state of the ConvLSTM at time step n. Although the ConvLSTM can predict the next and save the history information, it does not further discriminate typical dynamics of various actions. What is more, to learn the motion-aware knowledge in the aspects of visual appearances and motion patterns, we creatively design ARL and SMM modules based on ConvLSTM to exploit the spatiotemporal coherence and temporal order of videos in a self-supervised manner detailed as follows.

a) Appearance reconstruction loss: For appearance representation learning, an unsupervised video generation task is designed based on the mid-level features, where the stability of consecutive frames in unlabeled videos can be exploited. Considering the characteristic of ConvLSTM in video generation, the hidden state H_n can approximately express the frame prediction $\hat{F}(V_{n+1})$ at time $n+1$, that is,

$$\hat{F}(V_{n+1}) = H_n. \tag{3}$$

It is obvious that the feature of frame V_{n+1} can be exactly obtained by the backbone as $F(V_{n+1})$. Thus, we introduce the ARL to minimize the distance between the predicted feature $\hat{F}(V_{n+1})$ and the relatively accurate ones $F(V_{n+1})$, not only making the ConvLSTM capture motion-related representations but also enforcing the feature backbone $F(\cdot)$ to extract the discriminative features, as described in the following:

$$L_{\text{ARL}}^i = \left\| \hat{F}(V_{n+1}) - F(V_{n+1}) \right\|_1 \tag{4}$$

where $\|\cdot\|_1^1$ means ℓ_1 loss. L_{ARL}^n denotes the ARL of frame V_{n+1}. The final loss of all frames is formulated as follows:

$$L_{\text{ARL}} = \sum_{i=2}^{N} L_{\text{ARL}}^i. \tag{5}$$

Note that the subscript i starts from 2 since the first predicted feature corresponds to the second frame. The ARL can enforce the model to learn the obvious motion-related visual dynamics on unlabeled videos, which further facilitates the acquisition of discriminative features for HOI detection.

b) Sequential motion mining: The spatiotemporal coherence in videos provides strong supervisory signals, so one can achieve good representation learning by understanding the sequential motion patterns. In light of this, we propose an SMM model to perceive the reasonable motion-related knowledge of the sequence, as shown in Fig. 3. The correct sequence [see Fig. 3(a)] contains general spatiotemporal information, while the wrong sequence [out of order; see Fig. 3(b)] is ruleless. Trying to distinguish the correct and wrong sequences forces the model to learn the valuable consecutive motion pattern.

Therefore, except for the sampled correct sequence $\mathbb{V}$, we randomly shuffle the sampled frames to form a wrong sequence $\mathbb{V}^-$. The ConvLSTM takes the frame features $F(V_n^-)$ obtained from the shared extractor one by one as input and outputs the memory cell C_n^- until time n. It can be formulated as follows:

$$\left[C_n^-, H_n^-\right] = \text{ConvLSTM}\left(F\left(V_n^-\right), C_{n-1}^-, H_{n-1}^-\right) \tag{6}$$

where $V_n^- \in \mathbb{V}^-$ is a frame of the wrong sequence. Since the input sequence $\mathbb{V}^-$ is wrong, the history information of sequence kept in the final memory cell C_N^- of ConvLSTM is irregular. Thus, we design an SMM loss L_{SMM} to classify the correct and wrong sequences as follows:

$$L_{\text{SMM}} = \frac{1}{N}\sum_{v} -\left[y_v \cdot \log(p_v) + (1-y_v)\cdot\log(1-p_v)\right] \tag{7}$$

where y_v represents the label of sample $v \in \{C_N, C_N^-\}$. If v belongs to C_N (the history information of $\mathbb{V}$), then $y_v = 1$; otherwise (C_N^-, the history information of $\mathbb{V}^-$) $y_v = 0$. p_v is the corresponding probability prediction defined as follows:

$$p_v = \sigma(\Phi(v)), \quad v \in \left\{C_N, C_N^-\right\} \tag{8}$$

where σ is the sigmoid function and $\Phi(\cdot)$ denotes the prediction module consisting of multiple fully connected layers.

B. Feature Space Domain Adaptation

While the multitask feature learning can achieve the transfer of action-relevant information from the unlabeled videos to the static HOI images, the performance of the main task is suboptimal due to the domain gap between the irrelevant videos and HOI images. Thus, it is of significant importance to eliminate the domain shift for better knowledge transfer between two domains, which will help learn discriminative features and further promote the performance of main HOI detection. Next, we will elaborate on the details of feature adaptation between the unlabeled videos and HOI samples.

Following [14], performing domain adaptation between the very low and very high layers can greatly contribute to the improvements of transfer learning performance. Therefore, we integrate a domain discriminator D after the third block $F(\cdot)$ in the backbone network to align the feature distributions across domains, where a player minimax game is constructed in an adversarial learning manner. That is to say, the domain discriminator D tries to distinguish which domain the features come from, while the backbone module $F(\cdot)$ aims to confuse the classifier. Concretely, $F(\cdot)$ and D are connected by the special gradient reverse layer (GRL) [69], as shown in Fig. 2. There are no other parameters associated with it apart from the metaparameter λ. During the forward process, the GRL does not do any operations on the passing features. When backpropagating, the gradient propagated to the GRL would be multiplied by $-\lambda$, i.e., $((\partial L_{\text{dis}})/(\partial \Theta_F))$ is effectively replaced with $-\lambda((\partial L_{\text{dis}})/(\partial \Theta_F))$ and is then passed to the preceding layer. Mathematically, the gradient reversal layer $R_\lambda(\mathbf{x})$ can be defined by two equations describing its forward propagation and backpropagation behavior, as shown in the following, respectively,

$$R_\lambda(\mathbf{x}) = \mathbf{x} \tag{9}$$

$$\frac{dR_\lambda}{d\mathbf{x}} = -\lambda\mathbf{I} \tag{10}$$

where $\mathbf{I}$ is an identity matrix, λ is not updated by backpropagation, and $\mathbf{x} \in \{F(\mathbb{V}), F(\mathbb{I})\}$ denotes the mid-level feature representations of unlabeled videos and HOI images.

When the training process converges, $F(\cdot)$ tends to extract domain-invariant feature representations. Formally, the objective of adversarial learning can be written as follows:

$$L_{\text{dis}} = \min_{\Theta_F}\max_{\Theta_D} \mathbb{E}_{V_n\sim\mathbb{V}}\log D(F(V_n)) + \mathbb{E}_{I_m\sim\mathbb{I}}\log(1-D(F(I_m))) \tag{11}$$

where Θ_F and Θ_D are the parameters of the shared backbone F and domain classifier D, respectively. Different from the general GAN-based adversarial methods that alternately update the generator and discriminator in a two-stage manner, our domain classifier is jointly trained with the auxiliary and main tasks in an end-to-end way due to the usage of GRL. With the help of domain discrimination, the shared feature extractor can obtain domain-invariant features so that the knowledge transfer between two domains is achieved in the same feature distribution space, leading to the improvement of the main task.

C. Training and Inference

1) Inference: Since our goal is to boost the performance of HOI detection, only the main task stream is considered in the inference process. For each human–object bounding box pair ho_i, we predict the score s_h^a for each action $a \in \{1, \ldots, A\}$, where A denotes the number of possible actions. The score $s_{h,o}^a$ depends on: 1) the confidence for the individual object detections s_h and s_o; 2) the interaction prediction based on the appearance of the person s_h^a and the object s_o^a; and (3) the score prediction based on the spatial relationship between the person

and the object s^a_{sp}. Specifically, our HOI score $s^a_{h,o}$ can be formulated as follows:

$$s^a_{h,o} = s_h \cdot s_o \cdot \left(s^a_h + s^a_o\right) \cdot s^a_{\text{sp}}. \tag{12}$$

2) Adversarial Multitask Training: In order to obtain the generalized features for HOI detection, our framework jointly trains the auxiliary task, motion-relevant knowledge learning, and main task, HOI detection, in an MTL manner. In addition, a domain discriminator is included to decrease the domain difference to extract domain-invariant features for achieving better information transfer between domains. Our final loss is summed over all losses as follows:

$$L = L_{\text{HOI}} + \zeta * L_{\text{auxiliary}} + \eta * L_{\text{dis}} \tag{13}$$

$$L_{\text{auxiliary}} = L_{\text{ARL}} + L_{\text{SMM}} \tag{14}$$

where ζ and η are hyperparameters that control the importance of auxiliary task and adversarial learning. We use $\zeta = 0.1$ and $\eta = 0.01$ in our experiments.

IV. Experimental Results

A. Datasets and Metrics

We use two HOI image datasets (i.e., V-COCO and HICO-DET) and two video datasets (i.e., UCF101 and HMDB51) to construct three pairs of video→image datasets corresponding to three experimental settings, i.e., UCF101→V-COCO, UCF101→HICO-DET, and HMDB51→HICO-DET in order to demonstrate the effectiveness of our proposed approach.

1) HOI Image Datasets: V-COCO is a subset of MS-COCO [70], including 10 346 images (2533 for training, 2867 for validation, and 4946 for test) and 16 199 human instances. Each person is annotated with binary labels for 26 action categories. Note that three action classes (i.e., cut, hit, and eat) are annotated with two types of targets (i.e., instrument and direct object). Since this work focuses on minimum supervision, we use only a few labeled training images per category. In detail, the category number of most HOIs is less than 30, and only four HOIs have more than 30 training samples. Therefore, a total of 182 images, corresponding to a 7% ratio of minimum samples to full samples, are obtained for training. The HICO-DET [19] dataset consists of 47 776 images with more than 150k human–object pairs (38 118 images in the training set and 9658 in the test set). It has 600 HOI categories over 80 object categories (as in MS-COCO [70]) and 117 unique action verbs. Note that there exist 138 HOI categories with only less than ten training instances (Rare) in the full categories. It is obvious that the above two settings have the same problem of limited training samples for some categories.

2) Unlabeled Video Datasets: HMDB51 contains 6849 short video clips distributed in 51 actions and UCF101 comprises 13 320 videos from 101 action categories. Note that we do not use the provided action labels of videos. We randomly sample K video clips from both datasets. Empirically, we extract the first $N = 5$ key frames from each video at the interval of six and three frames on the UCF101 and HMDB51 datasets considering the difference of video lengths between these two datasets.

TABLE I

Performance Comparison With the State-of-the-Art Methods on the V-COCO Dataset. Note That Only About 7% (182/2533) Labeled Training Samples of V-COCO Dataset on the Main Task Is Available. The Best Results Are Marked in Red

Ratio	Methods	Extra Knowledge	Feature Backbone	mAP$_{role}$ Scenario1	mAP$_{role}$ Scenario2
100%	Xu *et al.* [8]	Word embedding	ResNet-50	45.9	-
	RP_DC_D [28]	Pose	ResNet-50	47.8	-
	iCAN [21]	None	ResNet-50	45.3	-
7%	Xu *et al.* [8]	Word embedding	ResNet-50	27.42	32.80
	RP_DC_D [28]	Pose	ResNet-50	39.48	45.86
	iCAN [21]	None	ResNet-50	38.74	45.20
	Ours	None	ResNet-50	40.71	47.37

3) Evaluation Metrics: Since we finally focus on the performance of the main task rather than the auxiliary task in the MTL, following the standard evaluation setting in [21], we use role mean average precision (mAP) to measure the HOI detection performance. The goal is to detect the agents and the objects in the various roles for the action, denoted as the $<$ human, action, object $>$ triplet. The HOI detection is considered as a true positive if it has the correct action label, and the intersection-over union (IoU) between the human and object bounding-box predictions and the respective ground-truth boxes is greater than the threshold of 0.5.

B. Implementation Details

We deploy Detectron [67] with a ResNet-50-FPN [68] backbone to obtain human and object bounding-box predictions. To select a predicted bounding box as a training sample, we set the confidence threshold to be higher than 0.8 for humans and 0.4 for objects. For a fair comparison, we adopt the object detection results and the pretrained weights for MS-COCO [70] from [21]. For interaction prediction, we implement our network based on Faster R-CNN with a ResNet-50 feature backbone. The image-centric training strategy [21] is also applied. The ConvLSTM is fed with mid-level features of the backbone, e.g., the third block. The domain discriminator consists of one gradient reversal layer and two fully connected layers, outputting which domain the input features belong to. We use the SGD optimizer for training with an initial learning rate of 5e-6, a weight decay of 5e-4, and a momentum of 0.9 for all datasets. In training, the ratio of positive and negative samples is 1:3. All experiments are conducted on a single Nvidia Titan XP GPU.

C. Quantitative Results

We compare our proposed model with several existing approaches under conditions of limited training samples for some categories to manifest the effects of motion-relevant knowledge implied in unlabeled videos to HOI detection.

For the V-COCO dataset only with minimum supervised labels, we evaluate mAP$_{\text{role}}$ of 24 actions with the label ratio 100% and 7%, respectively. As shown in Table I, for 7% label ratio, our method achieves 40.71 mAP$_{\text{role}}$ and

TABLE II

PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE HICO-DET DATASET. THE BEST RESULTS ARE MARKED IN RED

Methods	Extra knowledge	Feature Backbone	Default			Known Object		
			Full	Rare	Non-Rare	Full	Rare	Non-Rare
Xu *et al.* [8]	Word embedding	ResNet-50	14.70	13.26	15.13	-	-	-
iHOI [27]	Pose	ResNet-50	13.39	9.51	14.55	-	-	-
RPNN [29]	Pose	ResNet-50	17.35	12.78	18.71	-	-	-
Li *et al.* (RP_DC_D) [28]	Pose	ResNet-50	17.03	13.42	18.11	19.17	15.51	20.26
PMFNet [30]	Pose	ResNet-50-FPN	17.46	15.65	18.00	20.34	17.47	21.20
InteractNet [20]	None	ResNet-50-FPN	9.94	7.16	10.77	-	-	-
GPNN [24]	None	ResNet-152	13.11	9.34	14.23	-	-	-
iCAN [21]	None	ResNet-50	14.84	10.45	16.15	16.26	11.33	17.73
Wang *et al.* [22]	None	ResNet-50	16.24	11.16	17.75	17.73	12.78	19.21
AGRR [71]	None	ResNet-50	16.63	11.30	18.22	19.22	14.56	20.61
Ours (HMDB51)	None	ResNet-50	16.58	11.76	18.01	19.14	14.93	20.40
Ours (UCF101)	None	ResNet-50	16.55	11.81	17.96	19.13	14.97	20.37

47.37 mAP_{role}, with improvements of 13.29(14.57) compared with approach [8] relied on extra word embedding and 1.23(1.51) compared with ones [28] based on well-trained pose estimation model on Scenario1 (Scenario2), respectively. Note that the method [8] trained with word embedding is extremely worse in the case of minimum supervised labels. This is maybe due to the fact that the lack of training samples makes the method overly rely on other extra knowledge irrelevant to HOI images, leading to the inferior performance of HOI detection. Although the approach with the help of pose information gets good performance, the well-trained pose estimation model needs more fine-grained pose annotations. Moreover, compared with the case of full supervision (with 100% label ratio), the minimum supervision (with label ratio of 7%) can achieve a competitive performance for HOI detection.

For the HICO-DET dataset, we report results on three different HOI category sets: full, rare, and nonrare with two different settings of default and known objects. As shown in Table II, our proposed method achieves 1.88 mAP improvements compared to [8] trained with word embedding in full category sets under default settings. Although the approaches [28]–[30] with the assistance of pose information are superior to us, they extremely rely on the well-trained pose estimation model that also needs a large number of labeled samples to train. Furthermore, our method outperforms most of the well-designed complicated approaches [20]–[22], [24] on all settings of default and known objects, while it is inferior to [71] on the full category but superior on the hard rare category. In particular, the percentage gains reach 4.5% and 2.8% for mAP on rare HOI category with two different settings of default and known objects compared with [71]. Those obviously prove that our proposed multitask feature learning with the aid of unlabeled videos is effective for action-related feature learning on HOI detection, especially for rare categories, even without the well-trained pose estimation model or elaborated architecture.

D. Ablation Study

1) Effectiveness of Appearance Reconstruction Loss: To demonstrate the effectiveness of ARL on HOI detection with insufficient training samples, we jointly train the whole model based on the main task HOI detection loss and ARL to learn the discriminative action-relevant features. As shown in Table III for the V-COCO dataset, the proposed ARL improves the mAP performance of baseline by 0.74 and 0.67 with the UCF101 as an auxiliary dataset on Scenario1 and Scenario2, respectively. Moreover, it exceeds the baselines 1 and 3.03 at metric mAP, with the assistance of unlabeled UCF101 dataset, on rare category with two different settings of default and known objects, as shown in Table IV. All of these certify that our proposed ARL is beneficial to discriminative feature learning and further compensates for the limited training samples in the HOI detection task.

TABLE III

ABLATION STUDY ON MINIMUM LABELED V-COCO DATASET ABOUT EACH PROPOSED MODULE. "ARL" DENOTES APPEARANCE RECONSTRUCTION LOSS. "SMM" MEANS SEQUENTIAL MOTION MINING. "DA" REPRESENTS DOMAIN ADAPTATION BETWEEN TWO DOMAINS. THE BEST RESULTS ARE MARKED IN RED

w/ ARL	w/ DA	w/ SMM	mAP_{role} Scenario1	mAP_{role} Scenario2
-	-	-	38.74	45.20
✓	-	-	39.48	45.87
✓	✓	-	40.15	46.89
-	-	✓	40.20	46.96
✓	✓	✓	40.71	47.37

2) Effectiveness of Sequential Motion Mining: Furthermore, we train the model with the supervision of HOI detection loss and SMM loss to show the effectiveness of the proposed SMM on learning the motion-relevant features. One can observe from Table III that the proposed SMM improves the mAP performance of baseline by 1.46 and 1.76 with the help of UCF101 videos on Scenario1 and Scenario2 of V-COCO, respectively. For the HICO-DET dataset shown in Table IV, it exceeds the baseline 2.24 at metric mAP with the assistance of UCF101 videos on rare category of known objects' setting. These prove that our proposed SMM can help the model learn general motion-relevant knowledge that will cover the shortage of rare training samples in HOI detection. Moreover, when combined with ARL and DA, the model gets improvements of 0.56 and 0.48 mAP on Scenario1 and Scenario2, respectively, shown in Table III. It demonstrates that the proposed SMM may be complementary for ARL in terms of learning motion-aware knowledge to some extent. In addition, the model with SMM outperforms the one with both ARL and DA on the

TABLE IV

ABLATION STUDY ON THE HICO-DET DATASET ABOUT EACH PROPOSED MODULE. SIMILAR TO TABLE III, "ARL" DENOTES APPEARANCE RECONSTRUCTION LOSS. "SMM" MEANS SEQUENTIAL MOTION MINING. "DA" REPRESENTS DOMAIN ADAPTATION BETWEEN TWO DOMAINS

w/ ARL	w/ SMM	w/ DA	Unlabeled Video Dataset	Default Full	Default Rare	Default Non-Rare	Known Object Full	Known Object Rare	Known Object Non-Rare
-	-	-	-	14.84	10.45	16.15	16.26	11.33	17.73
✓	-	-	UCF101	16.35	11.45	17.81	18.79	14.36	20.12
-	✓	-	UCF101	16.24	10.54	17.94	18.77	13.57	20.33
✓	-	✓	UCF101	16.53	11.74	17.96	19.04	14.73	20.33
✓	✓	✓	UCF101	16.55	11.81	17.96	19.13	14.97	20.37

TABLE V

COMPARISON OF DIFFERENT STRATEGIES ON THE HICO-DET DATASET. THE SELF-SUPERVISED PRETRAINING REPRESENTS PRETRAINING WITH OUR PROPOSED SSL ON UNLABELED VIDEO AND THEN TRAINING WITH THE MAIN HOI DETECTION TASK. THE SSL ON STATIC IMAGE DENOTES CONSTRUCTING SSL ON IMAGES AND TRAINING WITH HOI DETECTION IN A MULTITASK MANNER

Models	Default Full	Default Rare	Default Non-Rare	Known Object Full	Known Object Rare	Known Object Non-Rare
baseline	14.84	10.45	16.15	16.26	11.33	17.73
self-supervised pre-training	12.23	7.43	13.66	14.46	9.60	15.91
self-supervised learning on static image	16.22	10.46	17.95	18.79	13.65	20.32
Ours (HMDB51)	16.58	11.76	18.01	19.14	14.93	20.40
Ours (UCF101)	16.55	11.81	17.96	19.13	14.97	20.37

V-COCO dataset and vice versa on the HICO-DET dataset. Compared with V-COCO, HICOD-DET has more diverse interaction classes and is, thus, more difficult to recognize. Since interactions in V-COCO are relatively simple and regular, exploring the temporal order features (motion patterns) rather than visual appearance may benefit further learning the discriminative representations. In contrast, the learning of HICO-DET is essentially hard for the model; therefore, the temporal order classification auxiliary task (SMM) may be not enough for capturing subtle differences of numerous various interactions and, thus, cannot provide more help compared with the simple task (ARL).

3) Comparison With Self-Supervised Pretraining: For the model with self-supervised pretraining, it is pretrained with our proposed SSL on unlabeled video and then trained with the main HOI detection task. From Table V, one can observe that the performance of it falls from 11.81 to 7.43 and from 14.97 to 9.60 on rare HOI category with two different settings of default and known objects compared with ours (UCF101). Our intuition for the weak performance of self-supervised pretraining is that pretraining is not aware of the main task of interest and can fail to adapt, which is consistent with the observation in [74]. It demonstrates that MTL of jointly training main and auxiliary tasks is better than separately training of them because it can address the mismatch between the two different tasks.

4) Comparison With Self-Supervised Learning on Static Images in a Multitask Manner: For the auxiliary training dataset, we randomly choose 1k videos from the HMDB51 and sample one frame from each video. The sampled static images are rotated 0° and 180°, respectively. The SSL on the static image is constructed based on predicting the rotation angles of the input images, as shown in Table V. It is observed that, although the SSL on static images has a certain improvement, it is still worse than ours (HMDB51), especially for rare categories. Concretely, the mAPs of a rare category with two different settings of default and known objects decrease by 11.05% and 8.57%, respectively. This may be due to the fact that the SSL on images only helps learn the robust feature representations by enriching the training dataset but does not obtain the motion-relevant knowledge that is beneficial to the main task.

TABLE VI

COMPARISON WITH OTHER VARIOUS STRATEGIES UNDER CONDITION OF LIMITED TRAINING SAMPLES ON THE HICO-DET DATASET. THE BASELINE IS A COMMON MULTISTREAM ARCHITECTURE WITHOUT ANY STRATEGIES TO DEAL WITH RARE SAMPLES

Models	Default Full	Default Rare	Default Non-Rare	Known Object Full	Known Object Rare	Known Object Non-Rare
baseline	14.84	10.45	16.15	16.26	11.33	17.73
re-sampling (RFS) [72]	15.87	10.28	17.54	18.41	13.26	19.95
re-weighting [28]	15.67	10.37	17.25	18.14	13.28	19.59
Kang *et al.* [73]	15.77	10.43	17.37	18.19	13.26	19.66
Ours (UCF101)	16.55	11.81	17.96	19.13	14.97	20.37

5) Effectiveness of Domain Discrimination (DA): To decrease the domain shift between unlabeled videos and HOI images, we use a domain discriminator to distinguish which domain the features come from and simultaneously train the backbone module to confuse the classifier in an adversarial manner. It can be observed from Table III that the proposed DA achieves 0.67 and 1.02 mAP improvements compared to the model without DA on Scenario1 and Scenario2 of the V-COCO dataset, respectively. For the HICO-DET dataset, as shown in Table IV, the percentage gains of domain discrimination reach 2.53% and 2.57% for mAP, with the aid of the UCF101 dataset, on rare HOI category compared to the model without it. This strongly indicates that the domain discrimination can decrease the domain gap to obtain the domain-invariant features, which implicitly contributes to the motion-relevant features' learning for rare HOI detection.

6) Comparison With Other Various Strategies on Dealing With the Rare Categories: We further explore the influence on detecting rare categories with different strategies, including resampling (repeat factor sampling, RFS) [72], reweighting [28], and decoupling representation and classification [73], as shown in Table VI. One can observe that our proposed method achieves the best performance on all settings compared to the other approaches. Concretely, ours (UCF101) exceeds the RFS 1.53 and 1.71 at metric mAP on *rare* HOI category with two different settings of default and known objects. It reveals that the perspective of adversarial MTL can effectively compensate for the limitation of insufficient training samples per class and significantly boost the performance of human–object interaction detection, especially for the rare HOI category.

TABLE VII
INFLUENCE OF VIDEO FRAME INTERVAL ON RARE HOI DETECTION OF THE HICO-DET DATASET

Frame Interval	Default			Known Object		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
2	16.43	11.70	17.84	18.88	14.57	20.17
6	16.55	11.81	17.96	19.13	14.97	20.37
8	16.46	11.75	17.86	18.91	14.61	20.19

TABLE VIII
ABLATION STUDY ON THE HICO-DET DATASET ABOUT THE INFLUENCE OF DIFFERENT AUXILIARY VIDEOS. Λ_v REPRESENTS THE NUMBER OF TRAINING VIDEOS. Λ_c DENOTES THE NUMBER OF VIDEO CATEGORIES

Video Dataset	Λ_v	Λ_c	Default			Known Object		
			Full	Rare	Non-Rare	Full	Rare	Non-Rare
-	-	-	14.84	10.45	16.15	16.26	11.33	17.73
UCF101	1K	20	16.28	11.40	17.74	18.69	14.15	20.05
UCF101	1K	51	16.53	11.76	17.95	19.01	14.66	20.31
UCF101	1K	101	16.55	11.81	17.96	19.13	14.97	20.37
UCF101	5K	101	16.32	11.42	17.78	18.80	14.38	20.12
HMDB51	1K	51	16.58	11.76	18.01	19.14	14.93	20.40

E. Further Analysis

1) Influence of Video Frame Interval on Rare HOI Detection: We change the video frame interval from 2 to 8 of the UCF101 video, as shown in Table VII, to explore its influence on detecting rare HOI categories. One can observe that, when the frame interval is 6, the best performance on the rare category is achieved. This may be due to that the motion range with the interval of 6 is suitable for motion-relevant knowledge learning. Too big interval may result in missing the discriminative information, and too small interval makes it difficult to learn obvious motion dynamics. Similarly, we experimentally set the frame interval of HMDB51 to 3 in this work.

2) Do More Training Samples Benefit HOI Detection: We try to change the number of video clips when performing motion-relevant knowledge learning on unlabeled videos. As shown in Table VIII, when increasing the number of training samples from 1k to 5k, the mAPs of a rare category with two different settings of default and known objects decrease by 3.3% and 3.9%, respectively. This may be due to the fact that a large number of random video clips lead to the diversity and complexity of auxiliary datasets, and deciding which valuable video clips can transfer to HOI images is hard. Moreover, the large-scale unlabeled videos in the auxiliary task may lead to excessively focusing on video representations, thus neglecting the specific characteristic of HOI to some extent.

3) How Does the Diversity of Video Category Influence the HOI Detection: Fixed the auxiliary videos and number of samples, changing the number of categories from 101 to 20 results in the decline of mAP on all categories with two different settings of default and known objects, as shown in Table VIII. In addition, the HOI detection achieves inferior performance with the help of 20 video categories from UCF101. Those show that too few categories cannot achieve the optimum transfer from videos to HOI images, which may be due to the fact that too few categories hardly contain useful knowledge for HOI detection. It reveals that the scarcity of categories is less helpful to motion-relevant knowledge learning, but a too large auxiliary dataset may be unnecessarily difficult to learn. Facing the increasing number of available unlabeled videos, the categories of videos that are diverse and relevant to the classes of the main task have rather simple relations and, with a reasonable number close to the one in the main task, maybe benefit to obtain greater performance gains in the main task. After meeting the above conditions, the scale of selected videos should not be too small, but too large-scale videos are not suggested since they may involve more complex dynamics to learn. This needs to design more complicated architectures and more sophisticated auxiliary tasks but with higher costs of training complexity. Note that the video label information is only used to explain the influence of category diversity and is not available in our whole experiments.

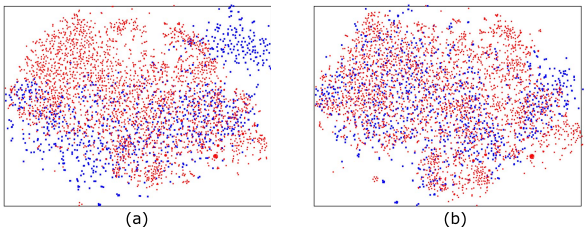


Fig. 4. Visualization of features using t-SNE: (a) without domain discrimination and (b) with domain discrimination. The red and blue points are samples from the HOI images and UCF101 unlabeled videos, respectively.

4) Different Auxiliary Unlabeled Video Datasets: In order to explore how different unlabeled video datasets influence the main task performance, we, respectively, select 1k video clips covering 51 categories from UCF101 and HMDB51 datasets, as shown in Table VIII. One can notice that using the HMDB51 as an auxiliary dataset gets a little better performance on all categories with two different settings of default and known objects compared with using the UCF101 dataset. It may reveal that the HMDB51 dataset has similar motion-relevant features shared with the HOI images so that it can greatly boost the HOI detection performance.

Apart from our multitask joint learning manner, pretraining on unlabeled videos is an alternative strategy for learning generalizable visual features. For the pretraining methods, in order to obtain comprehensive representations, various and abundant videos are indispensable. After that, the pretrained model is separately applied to the downstream application tasks. Different from the pretraining ways that do not consider the specificity of the main tasks, the MTL methods need to take the main task of interest into account and achieve the adaptation between proxy and main tasks. Thus, placing additional constraints on the selection of auxiliary video sets is more helpful for MTL compared with the pretraining strategy.

F. Qualitative Visualization Results

1) Feature Visualization: We visualize the feature distribution of the adaptation layer, corresponding to the output of $F(\cdot)$, using t-SNE [75] projection, as shown in Fig. 4.

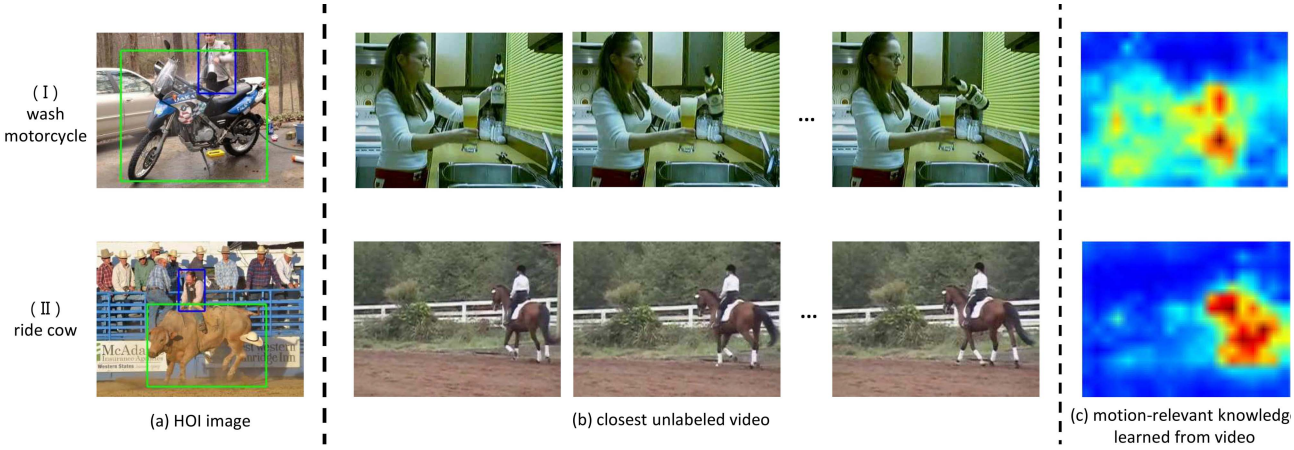


Fig. 5. Implicit influence of motion-relevant knowledge learned from video on rare categories of the HICO-DET dataset [19]. (a) Rare HOI categories of the main task. (b) Closest unlabeled video that may be helpful for rare categories. (c) Motion-related knowledge learned from the unlabeled video.

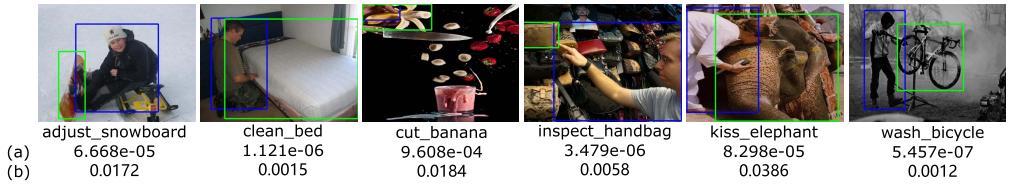


Fig. 6. Influence of our proposed motion-relevant knowledge learning on the predicted HOI interaction scores for **rare** categories of the HICO-DET dataset [19]. (a) Model without motion-relevant knowledge learning. (b) Model with it.

Concretely, the red points represent HOI image features, while the blue points are UCF101 video features. Fig. 4(a) denotes the feature visualization without domain discrimination, while Fig. 4(b) is with domain discrimination. One can observe that the features aligned with domain discrimination are more indistinguishable, indicating that it is effective to reduce domain shift between the unlabeled videos and the HOI images by confusing the domain classifier in an adversarial manner.

2) Implicit Influence of Motion-Relevant Knowledge Transferred From Videos on Rare Categories: To further explain how the unlabeled videos promote the rare HOI categories detection, we show the intermediate motion-relevant knowledge obtained from the closest unlabeled videos, as shown in Fig. 5. Concretely, we visualize the heatmap of the middle frame in the video. One can observe that our proposed approach can learn the generalized motion-aware features [e.g., Fig. 5(c)] through the motion-relevant knowledge learning on unlabeled videos [e.g., Fig. 5(b)], which will guide the detection of rare interaction categories [e.g., Fig. 5(a)] of the main task. Moreover, the proposed motion-relevant knowledge learning on unlabeled videos can pick up the dominant mover in the presence of camera motion, as shown in Fig. 5(II) (b)–(c).

3) Effectiveness of Motion-Relevant Knowledge Learning for Rare HOI Detection: To further manifest the effectiveness of motion-relevant knowledge learning for rare interaction detection, we show the predicted scores of the model without and with motion-relevant knowledge learning, as illustrated in Fig. 6. It is obvious that the action score with its help outperforms the ones without it by a large margin. That is maybe due to the fact that the knowledge learned from auxiliary unlabeled videos can compensate for the lacking of rare categories to some extent.

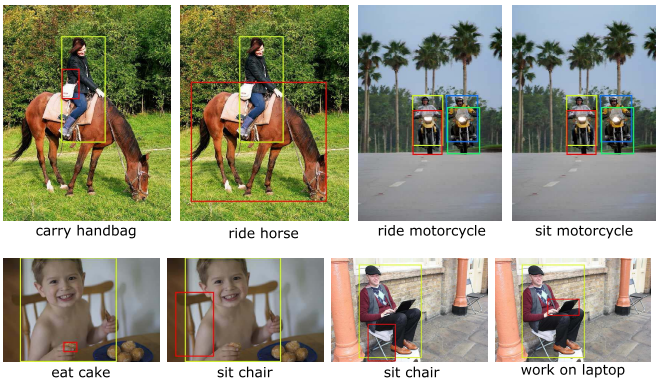


Fig. 7. Multiple interaction detections on the V-COCO dataset [18]. Our model detects human instances doing multiple actions and interacting with different objects, e.g., the person sitting on the chair is working on a laptop.

4) HOI Detection Visualization: Fig. 7 shows interaction detection examples that the detected human has different interactions with various objects simultaneously. Fig. 8 displays the HOI detections on the V-COCO test set [18], which demonstrates that our model can detect various objects that the human instances are interacting with in different situations. Fig. 9 represents the sample HOI detections on the HICO-DET test set [19]. It reveals that our approach can detect multiple interactions with the same object.

5) Failure Cases: Although we demonstrate improved performance, our model is far from perfect, as shown in Fig. 10.

Fig. 8. Sample HOI detections on the V-COCO dataset [18]. For actions, "ride" and "eat," our model detects various objects that the human instances are interacting with in different situations.

Fig. 9. Sample HOI detections on the HICO-DET test set [19]. Our model detects different types of interactions with objects from the same category.

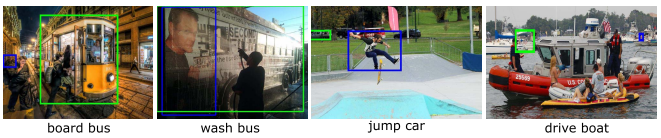


Fig. 10. Failure cases of our method where irrelevant humans and objects are considered interactive due to the neglect of relations between objects.

Generally, we leverage the off-the-shelf object detector to detect various object instances in an image. When the scene is complex, our model will be confused about whether the detected human and object are interactive or not. Our approach mainly captures the discriminative representations of HOI with the help of unlabelled videos. Based on the rich contextual cues, further taking the relations of objects into consideration may reduce such mistakes to some extent. Also, this is what we have been working on, i.e., distinguishing whether the human and object are interactive or not.

V. Conclusion

In this article, we propose to deal with the problem of limited training samples in HOI detection from a novel multitask feature learning perspective. An auxiliary task based on motion-relevant knowledge learning on unlabeled videos is proposed and jointly trained with the HOI detection in an MTL manner to learn more generalizable representations for boosting the performance of HOI detection. In addition, a domain classifier is introduced to decrease the domain shift between videos and HOI images for better knowledge transfer between two domains. The experiments conducted on the HICO-DET dataset with rare categories and the V-COCO dataset with minimum supervision confirm the effectiveness of the proposed approach, especially for the limited labeled categories. We believe that our work not only contributes to the learning of rare HOI categories but offers inspiration for general knowledge transfer in the future.

References

[1] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 961–970.

[2] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6904–6913.

[3] A. Mallya and S. Lazebnik, "Learning models for actions and person-object interactions with transfer to question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 414–428.

[4] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robot. Auton. Syst.*, vol. 57, no. 5, pp. 469–483, 2009.

[5] S. Ekvall and D. Kragic, "Interactive grasp learning based on human demonstration," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2004, pp. 3519–3524.

[6] Z. Yang, D. Mahajan, D. Ghadiyaram, R. Nevatia, and V. Ramanathan, "Activity driven weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2917–2926.

[7] D. Kim, G. Lee, J. Jeong, and N. Kwak, "Tell me what they're holding: Weakly-supervised object detection with transferable knowledge from human-object interaction," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11246–11253.

[8] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, "Learning to detect human-object interactions with knowledge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2019–2028.

[9] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "DRG: Dual relation graph for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 696–712.

[10] K. Kato, Y. Li, and A. Gupta, "Compositional learning for human object interaction," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 247–264.

[11] L. Shen, S. Yeung, J. Hoffman, G. Mori, and L. Fei-Fei, "Scaling human-object interaction recognition through zero-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2018, pp. 1568–1576.

[12] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3107–3115.

[13] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 584–600.

[14] Z. Ren and Y. J. Lee, "Cross-domain self-supervised multi-task feature learning using synthetic imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 762–771.

[15] B. Zhuang, L. Liu, C. Shen, and I. Reid, "Towards context-aware interaction recognition for visual relationship detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 589–598.

[16] Y. Kong, Y. Jia, and Y. Fu, "Interactive phrases: Semantic descriptionsfor human interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 1775–1788, Sep. 2014.

[17] X. Shu, J. Tang, G.-J. Qi, W. Liu, and J. Yang, "Hierarchical long short-term concurrent memory for human interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1110–1118, Mar. 2021.

[18] S. Gupta and J. Malik, "Visual semantic role labeling," 2015, *arXiv:1505.04474*.

[19] Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2018, pp. 381–389.

[20] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8359–8367.

[21] C. Gao, Y. Zou, and J. Huang, "iCAN: Instance-centric attention network for human-object interaction detection," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–13.

[22] T. Wang *et al.*, "Deep contextual attention for human-object interaction detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5694–5702.

[23] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, "VSGNet: Spatial attention network for detecting human object interactions using graph convolutions," 2020, *arXiv:2003.05541*.

[24] S. Qi, W. Wang, B. Jia, J. Shen, and S. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 407–423.

[25] Y.-L. Li *et al.*, "PaStaNet: Toward human activity knowledge engine," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 382–391.

[26] H. Fang, J. Cao, Y. Tai, and C. Lu, "Pairwise body-part attention for recognizing human-object interactions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 52–68.

[27] B. Xu, J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Interact as you intend: Intention-driven human-object interaction detection," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1423–1432, Jun. 2020.

[28] Y.-L. Li *et al.*, "Transferable interactiveness knowledge for human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3585–3594.

[29] P. Zhou and M. Chi, "Relation parsing neural network for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 843–851.

[30] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9469–9478.

[31] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for robust human–object interaction detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–22.

[32] Y. Liu, J. Yuan, and C. W. Chen, "ConsNet: Learning consistency graph for zero-shot human-object interaction detection," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 4235–4243.

[33] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4116–4125.

[34] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "PPDM: Parallel point detection and matching for real-time human-object interaction detection," 2019, *arXiv:1912.12898*.

[35] J. A. Baxter, "A model of inductive bias learning," *J. Artif. Intell. Res.*, vol. 12, no. 1, pp. 149–198, 2000.

[36] S. Thrun, "Is learning the n-th thing any easier than learning the first?" in *Proc. Neural Inf. Process. Syst.*, 1995, pp. 640–646.

[37] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[38] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3994–4003.

[39] Y. Liao, S. Kodagoda, Y. Wang, L. Shi, and Y. Liu, "Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2016, pp. 2318–2325.

[40] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 94–108.

[41] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "R-CNNs for pose estimation and action detection," 2014, *arXiv:1406.5212*.

[42] L. Pinto, D. Gandhi, Y. Han, Y. Park, and A. Gupta, "The curious robot: Learning visual representations via physical interactions," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 3–18.

[43] L. Pinto and A. Gupta, "Learning to push by grasping: Using multiple tasks for effective learning," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2017, pp. 2161–2168.

[44] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.

[45] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. Bradski, "Self-supervised monocular road detection in desert terrain," in *Proc. Robot., Sci. Syst. Conf.*, Aug. 2006, pp. 1–7.

[46] W. Lee, J. Na, and G. Kim, "Multi-task self-supervised object detection via recycling of bounding box annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4984–4993.

[47] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.

[48] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2536–2544.

[49] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1422–1430.

[50] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–84.

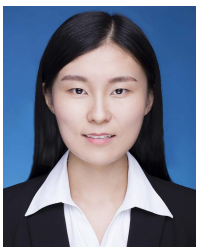
[51] H. Lee, J. Huang, M. Singh, and M. Yang, "Unsupervised representation learning by sorting sequences," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 667–676.

[52] J. J. Walker, C. Doersch, A. Gupta, and M. Hebert, "An uncertain future: Forecasting from static images using variational autoencoders," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 835–851.

[53] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating visual representations from unlabeled video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 98–106.

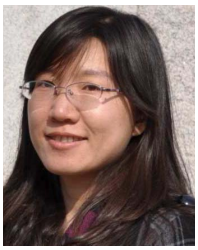
[54] D. Jayaraman and K. Grauman, "Slow and steady feature analysis: Higher order temporal coherence in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3852–3861.

[55] B. Fernando, H. Bilen, E. Gavves, and S. Gould, "Self-supervised video representation learning with odd-one-out networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5729–5738.
[56] Y. Zhang *et al.*, "Exploiting motion information from unlabeled videos for static image action recognition," 2019, *arXiv:1912.00308*.
[57] N. Wichers, R. Villegas, D. Erhan, and H. Lee, "Hierarchical long-term video prediction without supervision," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 6033–6041.
[58] X. Lin, Q. Zou, X. Xu, Y. Huang, and Y. Tian, "Motion-aware feature enhancement network for video prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 688–700, Feb. 2021.
[59] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 213–226.
[60] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.
[61] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–450.
[62] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3764–3773.
[63] W. Cong *et al.*, "DoveNet: Deep image harmonization via domain verification," 2019, *arXiv:1911.13239*.
[64] Y. Zheng, D. Huang, S. Liu, and Y. Wang, "Cross-domain object detection through coarse-to-fine feature adaptation," 2020, *arXiv:2003.10275*.
[65] X. Ma, T. Zhang, and C. Xu, "Deep multi-modality adversarial networks for unsupervised domain adaptation," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2419–2431, Sep. 2019.
[66] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[67] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He. (2018). *Detectron*. [Online]. Available: https://github.com/facebookresearch/detectron
[68] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[69] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by back-propagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[70] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[71] X. Lin, Q. Zou, and X. Xu, "Action-guided attention mining and relation reasoning network for human-object interaction detection," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1104–1110.
[72] A. Gupta, P. Dollár, and R. Girshick, "LVIS: A dataset for large vocabulary instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5356–5364.
[73] B. Kang *et al.*, "Decoupling representation and classifier for long-tailed recognition," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–16.
[74] B. Zoph *et al.*, "Rethinking pre-training and self-training," 2020, *arXiv:2006.06882*.
[75] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.



Xue Lin received the B.S. and M.E. degrees from the School of Information Science and Engineering, University of Jinan, Jinan, China, in 2015 and 2018, respectively. She is currently pursuing the Ph.D. degree with the Department of Computer and Information Technology, Beijing Jiaotong University, Beijing, China, in 2018.

Her research interests are computer vision, image processing, and machine learning, and their applications on human-centric activity understanding.



Qi Zou received the Ph.D. degree in computer science from Beijing Jiaotong University, Beijing, China, in 2006.

She is currently a Professor and a Doctoral Supervisor with the School of Computer and Information Technology, Beijing Jiaotong University. Her research interests include computer vision and intelligent transportation systems.

Xixia Xu received the B.S. degree in software engineering from Lanzhou Jiaotong University, Lan Zhou, China, in 2018. She is currently pursuing the Ph.D. degree with the Department of Computer and Information Technology, Beijing Jiaotong University, Beijing, China.

Her current research interests include computer vision, image processing, and machine learning with applications on 2-D multiperson pose estimation analysis and human-centric behavior analysis.

Yaping Huang received the B.S., M.S., and Ph.D. degrees from the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China, in 1995, 1998, and 2004, respectively.

She is currently a Professor with the School of Computer and Information Technology, Beijing Jiaotong University. Her research interests are computer vision, machine learning, and pattern recognition.

Ding Ding received the Ph.D. degree in computer application technology from Beijing Jiaotong University, Beijing, China, in 2011.

She is currently an Associate Professor with the School of Computer and Information Technology, Beijing Jiaotong University. Her current research focuses on building a highly efficient and accurate service recommendation mechanism.