Human Object Interaction Detection via Multi-level Conditioned Network

Xu Sun[1,2], Xinwen Hu[1], Tongwei Ren[1,2,*], and Gangshan Wu[1]
[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2] Shenzhen Research Institute of Nanjing University, Shenzhen, China
{sunx,hu_xinwen}@smail.nju.edu.cn,{rentw,gswu}@nju.edu.cn

ABSTRACT

As one of the essential problems in scene understanding, human object interaction detection (HOID) aims to recognize fine-grained object-specific human actions, which demands the capabilities of both visual perception and reasoning. Existing methods based on convolutional neural network (CNN) utilize diverse visual features for HOID, which are insufficient for complex human object interaction understanding. To enhance the reasoning capablity of CNN, we propose a novel multi-level conditioned network that fuses extra spatial-semantic knowledge with visual features. Specifically, we construct a multi-branch CNN as backbone for multi-level visual representation. We then encode extra knowledge including human body structure and object context as condition to dynamically influence the feature extraction of CNN by affine transformation and attention mechanism. Finally, we fuse the modulated multimodal features to distinguish the interactions. The proposed method is evaluated on two most frequently-used benchmarks, HICO-DET and V-COCO. The experiment results show that our method is superior to the state-of-the-arts.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Human object interaction detection, conditioned network, multi-level visual representation, multimodal feature fusion, feature transformation.

ACM Reference Format:
Xu Sun[1,2], Xinwen Hu[1], Tongwei Ren[1,2,*], and Gangshan Wu[1]. 2020. Human Object Interaction Detection via Multi-level Conditioned Network. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3372278.3390671

1 INTRODUCTION

Human object interaction detection (HOID) aims to localize and classify human-object pairs and their interactions [1], which can be utilized in numerous multimedia applications such as image captioning [14, 22, 28] and retrieval [19, 30, 36]. In some cases, action recognition [42] and human-centric visual relation detection (HCVRD) [55] are considered similar to HOID, but they have substantial differences. Action recognition mainly concentrates on classifying the actions of individual human instances in images [13] or video clips [23, 42, 47] without considering the interacted objects, which is insufficient to describe complex visual scenes in real world. Compared with it, HOID provides more specific and comprehensive description of human activity with object context. As shown in Figure 1, action recognition only captures the human action—jump, while HOID describes how the human interacts with the skateboard in detail. HCVRD concentrates on holistic visual scene, including both interactions and geometrical relations between humans and all the objects in image [55]. Compared with HCVRD, for one thing, HOID focuses on comprehensive and fine-grained interactions, which requires deep understanding of human body structure; for another, HOID ignores the uninformative relations involving objects in background, which can be distinguished with some straightforward visual cues like relative location. As illustrated in Figure 1, HCVRD attempts to capture all the relations between the human and the objects including wall, ground, plant and skateboard. However, HOID only concentrates on the fine-grained and salient interactions between the individual and the skateboard, ignoring the uninformative relation instances.

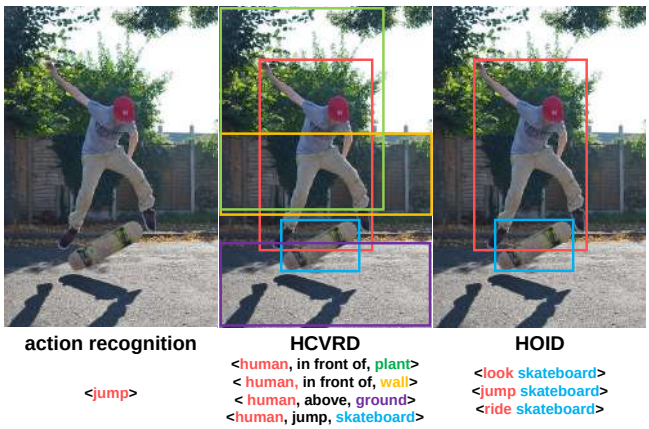

Figure 1: Comparison of action recognition, HCVRD and HOID. The bounding boxes indicate the locations of the human and interested objects. The labels under the images are the recognition results required by corresponding tasks. The colors indicate the consistence between the labels and bounding boxes.

* Corresponding author.

ICMR '20, June 8–11, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7087-5/20/06…$15.00
https://doi.org/10.1145/3372278.3390671

As a challenging task, HOID aims to capture high level semantic information beyond individual entities from complex visual scene. To be specific, the visual patterns within the same human object interaction (HOI) category can be quite distinct because of different human object instances and context. Moreover, since many interactions involve subtle motions of certain body parts, the appearance deviation among different categories can be minor. Following the strategy of object detection framework, early solutions [1, 39] intuitively combined the entity-level visual features [21] of human-object pairs extracted by convolutional neural network (CNN) for interaction classification. To make CNN focus on more informative regions of image, some methods apply visual attention mechanism [10] or supplement CNN features extracted from the regions around human body joints [43]. Although several existing works have made some progress in HOID, they may remain some defects. First of all, pure CNN feature can be insufficient to bridge the gap between low-level visual information of pixels and high-level semantic information of HOI. Recently-proposed methods, RPNN [54] and PMFNet [43], crop multi-level CNN features according to the bounding boxes of detected entities and human body parts to capture detailed visual cues. Although the prior location information is utilized, the CNN features are still sourced from image only. Besides, most of the existing HOID methods [16, 43, 54] use frozen CNN backbone pretrained on an object detection dataset to extract visual features for HOI reasoning. The appearance distributions of interaction phrase (union region of human-object pair) and single object are significantly biased, which are supposed to be learned independently.

Based on these observations, we propose a novel HOID method, a multi-level conditioned network (MLCNet) which aims to fuse extra explicit knowledge with multi-level visual feature. Specifically, we construct a multi-branch CNN structure as backbone to generate multi-level visual representation. To extract features of diverse visual content including global scene, interaction phrase, entities and human body parts, different branches are optimized independently. In this way, the appearance bias of different visual content can be effectively learned. However, the pure visual features are insufficient to understand complex semantics of HOI. Inspired by [45], we utilize extra spatial-semantic information of human body structure and object context as guidance to enhance the reasoning capability of CNN by dynamically influencing the feature extraction procedure. To obtain the comprehensive information of human body structure, we apply human parsing and pose estimation models to localize the body parts and joints respectively. The estimated body part segmentation map and body-object spatial configuration map are encoded with condition network and fed into feature transform layers to generate modulate parameters, which alternate the visual features at different levels by affine transformation.

Another informative cue we exploit is object context [4]. Intuitively, certain object category relates to certain body parts. For example, "bike" is often associated with "leg" and "hip", while "book" is often related to "head" and "arm". Moreover, different objects with similar function probably involve the same interactions, such as ⟨ride, bicycle⟩ and ⟨ride, motorcycle⟩. To explore these correlations, we use word vectors pretrained on large-scale linguistic dataset as object context features to represent object categories and generate attention weights for different body parts, which implicitly encode the functional similarity among different objects, thereby facilitating the transfer of interaction knowledge. We also add an context branch taking object category vectors as input to supplement the visual branches.

We evaluated the proposed MLCNet on two most frequently-used benchmarks, HICO-DET [1] and V-COCO [15]. The experiment results show that our method outperforms the state-of-the-art methods and component analysis confirms the effectiveness of the combination of multi-level CNN features and explicit knowledge. Compared with pure visual models, our method achieves better performance and interpretability.

2 RELATED WORK

2.1 Human Object Interaction Detection

HOID task was firstly formulated by Chao *et al.* [1]. They proposed evaluation metrics and the first large-scale HOID dataset, HICO-DET, which is currently used as a public benchmark. They also developed a multi-stream architecture named HO-RCNN, to aggregate entity appearance and spatial configuration information for HOI reasoning. Gao *et al.* improved the performance of HO-RCNN model with instance-centric attention module [10]. Gkioxari *et al.* proposed a novel multi-task model to simultaneously localize interacted object and recognize HOI under the guidance of human appearance [12]. Recently, Gupta *et al.* proposed a light-weight model and achieved impressive performance with proper feature factorization as well as several training techniques [16]. Wang *et al.* proposed a contextual attention framework, which adaptively selected relevant instance-centric context information to highlight informative regions of image [44]. Zhao *et al.* developed a graph-based network named RPNN to reason HOI by passing message through an object-body graph and a human-body graph [54]. Wan *et al.* designed a pose-aware multi-feature network, PMFNet, combining ROI-aligned CNN features in different levels, which achieved state-of-the-art performance [43].

Different from the existing methods that employ standard CNN features, MLCNet dynamically alternates the feature extraction with extra spatial-semantic knowledge as guidance and outperforms the state-of-the-art methods.

2.2 Human-Centric Visual Relation Detection

Visual relation detection (VRD) aims at detecting object pairs and their relations in terms of space, comparison, interaction and possession for image [29, 41, 50, 51] and video [35, 38, 40]. Lu *et al.* formulated VRD task on still image for the first time and proposed a multi-modal VRD method, combining language priors and deep visual feature [29]. Shang *et al.* proposed the first video VRD framework with the capability of temporally localizing and recognizing dynamic relations [38]. HCVRD [49, 55] concentrates on capturing the human-centric relations of which the subjects are restricted to human instances and is more related to real world application scenarios such as social media analysis [5, 11]. Zhuang *et al.* constructed a large-scale HCVRD dataset for still image and proposed a web-supervised baseline method [55]. Yu *et al.* developed a image HCVRD model using Mask-RCNN and VTransE [49], which obtained desirable results. Recently, Shang *et*

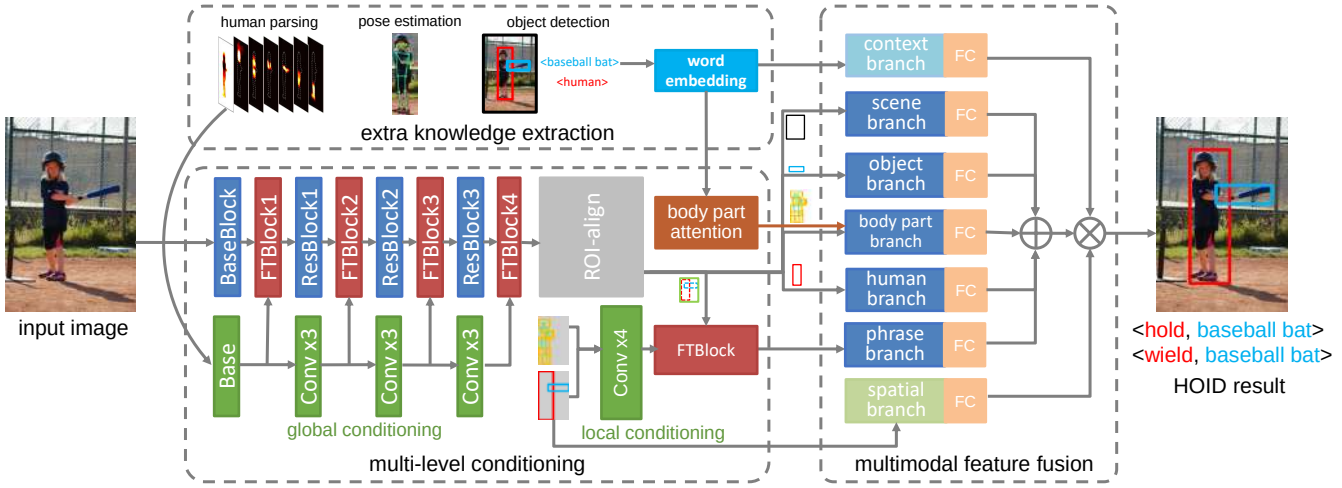

Figure 2: An overview of the proposed MLCNet. Here, FTBlock, ResBlock, Conv and FC are abbreviations of feature transform block, residual block, convolution layer and fully-connected classifier. ⊕ and ⊗ denote addition and element-wise multiplication. Zoom in for better view.

al. constructed a large-scale human-centric dataset, VidOR, for video VRD evaluation [37], on which Sun *et al.* proposed a video relation model with multi-model feature fusion and achieved state-of-the-art performance [40].

Instead of predicting general human-centric relations, HOID focuses on dynamic human object interactions that are more informative and fine-grained. It requires deep understanding of human action and body structure.

2.3 Conditioned Network

Multimodal information fusion is a critical problem of high-level semantic understanding tasks such as visual dialog [6] and visual question answering (VQA) [32]. To dynamically influence the focus of CNN with questions in VQA, Perez *et al.* proposed a feature-wise linear modulation layer, which took language feature as condition and transformed visual feature [33]. This structure effectively enhanced the reasoning capability of standard CNN. Besides, some image synthesis tasks such as style transfer [48, 52] and super resolution [45] also rely on external semantic guidance. In general, these works made an attempt to render different regions of image based on a segmentation map that contains both semantic and spatial information. Wang *et al.* proposed a spatial feature transform (SFT) layer for semantic super resolution [45]. This model encoded the probability map generated by semantic segmentation model as condition and modulated the visual feature of original image with the SFT layers. In this way, the realistic texture could be recovered under the guidance of spatial-semantic condition. Inspired by these works, the proposed HOID method exploits extra knowledge provided by visual perceptual models to improve the reasoning capability of CNN, attempting to fill the gap between visual feature and high-level semantics.

3 METHOD

Given an image I, we apply some off-the-shelf visual perceptual models to extract extra spatial-semantic knowledge $\mathcal{K}$. It is fed into the proposed MCLNet $\mathcal{D}(\cdot)$ together with I to enhance the HOI reasoning capability of CNN:

$$\Psi = \mathcal{D}(I|\mathcal{K}), \tag{1}$$

where Ψ is referred to the detected HOI instances $\{(b^h, b^o, \sigma)\}$, in which b^h and b^o are the bounding boxes of detected human and object, and σ belongs to the HOI category set. An HOI category σ involves an action ω_σ and an object α_σ, which belong to corresponding action and object category sets respectively. In the following sections, we start with the preparation for the extra knowledge we exploit. Then we introduce how to fuse multi-level visual features and extra spatial-semantic information by network conditioning in detail.

3.1 Extra Knowledge Extraction

Object detection. For an image I, we apply a state-of-the-art object detection model , FPN [26], to obtain the locations and categories of humans and objects. Detected human and object instances are referred to as b^h and (b^o, α) respectively. The human and object instances are paired as HOI candidates, $\Theta=\{(b^h, b^o, \alpha)\}$. The object categories are represented with a set of high-dimensional word vectors $v \in \mathbb{R}^{L_v}$ pretrained on large-scale linguistic dataset.

Pose estimation. To obtain the structure information of human body, we adopt an off-the-shelf multi-person pose estimation method, RMPE [9], which estimates N_k body joints for each human instance. Each body joint is expressed as a coordinate with confidence value.

Human parsing. We utilize a pretrained human parsing method, WSHP [8], to generate body part segmentation map Φ, a multi-channel probability map with the same width and height as the original image, each channel of which corresponds to a certain type of body parts. Compared with body joints, this semantic segmentation map provides denser structure information in pixel level including the shapes and edges of human body parts.

Although recent attempts on HOID also exploit object detection and pose estimation, most of them only use the obtained bounding boxes of entities and human body joints to crop CNN feature. Different from the existing methods, MLCNet comprehensively exploits the semantic information, global spatial distribution and relations among body parts and object for HOI reasoning. Serving as a bridge between pure visual feature and complicated semantics, these explicit knowledge contributes to improving both the reasoning capability and interpretability of deep network.

3.2 Multi-level Visual Features

Multi-level visual representation aims to encode both coarse and fine-grained visual information, which is essential for HOID. However, most existing methods utilize a shared CNN backbone to extract different visual features for HOI reasoning [16, 43, 54], which cannot capture the appearance distribution bias of different visual content.

To solve this problem, we construct a multi-branch CNN as the backbone of MLCNet. The proposed network structure encodes comprehensive visual information of global context, interaction phrase, entities and body parts with corresponding branches, which are independently optimized and can efficiently learn different appearance distributions in the training stage. Specifically, we first generate globally conditioned feature f_g for the entire image with a sequence of shared residual blocks and feature transform blocks, from BaseBlock to FTBlock4, as shown in Figure 2. The BaseBlock and ResBlocks are standard modules of ResNet. The feature transform blocks fuse human body structure information with global visual features, which is introduced in Section 3.3 in detail. Based on this, we extract multi-level visual features including f_u, f_h, f_o and f_p by cropping f_g according to the regions of interaction phrase, human, object and body parts respectively and passing them into corresponding branches. These branches share the same structure with the last residual block of ResNet but are optimized independently. The shape of ROI-aligned features f_u, f_h and f_o is $w \times w \times c$. Here w and c represent the width and number of channels. However, compared with entities and body parts, interaction phrase contains more complicated semantic information that pure CNN feature is unable to capture effectively. Therefore, we exploit explicit knowledge of human-object pair to improve f_u by local network conditioning, which is also introduced in Section 3.3.

To extract fine-grained visual features of human body structure, we construct body parts by dividing N_k body joints into N_p groups for each detected human instance following RPNN [54]. We apply ROI-align on f_g for all N_p body parts. The aforementioned f_p is generated by concatenating all the cropped body part features in channel-wise, the shape of which is $w \times w \times (c \times N_p)$. To highlight the informative body parts related to a certain object, we apply body part attention on f_p, which is introduced in the following section. In addition to the visual features f_u, f_h, f_o and f_p, we further supplement a holistic context feature f_s to encode the global scene, which is generated by pooling the CNN feature of the entire image, *i.e.* f_g. The pooled feature is fed into a scene branch, whose structure is the same as aforementioned branches. With global context feature f_s, the multi-level visual representation can be more comprehensive.

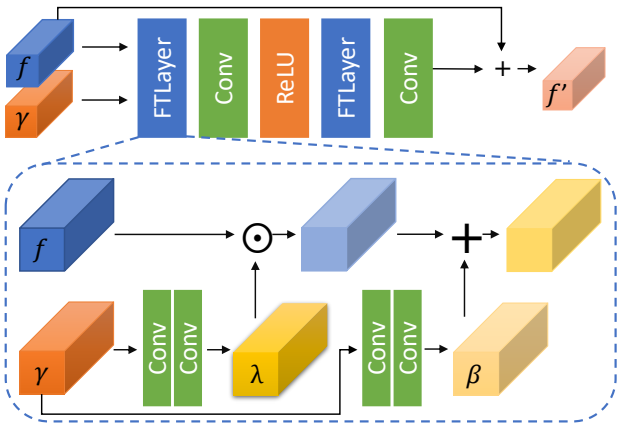

Figure 3: Feature transform block consists of feature transform, convolution and ReLU layers.

In the training stage, the whole network is optimized in an end-to-end way. The bias of features in different levels and large variation of appearance can be learned with limited number of parameters. At last, we apply global average pooling [25] on all visual features to generate feature vectors as inputs of classifiers.

3.3 Multi-level Conditioning

Standard CNN is insufficient to handle complex HOI reasoning because of the gap between low-level visual feature and high-level semantic information. To this end, we adopt a multi-level conditioning mechanism to further improve the reasoning capability of the aforementioned multi-branch CNN. Specifically, the proposed method dynamically alternate the features of global image, interaction phrase and body parts with explicit spatial-semantic information of human body structure and object context.

Global conditioning. We utilize body part segmentation map introduced in Section 3.1 as global condition to enhance the global visual feature of the entire image. The segmentation map is fed into a condition network to generate multi-level condition features $\{\gamma\}$, which encode the relative locations and shapes of human body parts on different scales simultaneously. Figure 2 indicates the overview of the global condition network. It consists of four consecutive convolution blocks, the same number as the blocks of CNN backbone. The first condition block has the same structure as the BaseBlock of CNN backbone and the following ones contain three convolution layers using 1×1 kernel, among which exist two LeakyReLU activation layers [17]. It is worth noting that the condition features are spatially aligned with corresponding visual features all the time. After each block of CNN backbone, global conditioning is implemented via a feature transform block, FTBlock shown in Figure 2, which combines visual and condition features of the same scale. Specifically, as shown in Figure 3, the feature transform layers $\mathcal{T}(f|\lambda, \beta)$ of FTBlock apply affine transformation to dynamically alternate the input visual feature f with modulate parameters (λ, β) following [45]. The parameters are generated by a mapping function $\mathcal{M}(\cdot)$ taking condition feature γ of human body

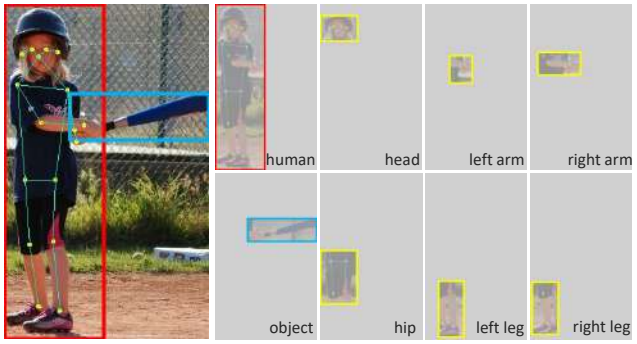

Figure 4: Body-object configuration map.

as input:

$$(\lambda, \beta) = \mathcal{M}(\gamma), \tag{2}$$

$$\mathcal{T}(f|\lambda, \beta) = \lambda \odot f + \beta, \tag{3}$$

where $\mathcal{M}(\cdot)$ is implemented with double convolution layers using 1×1 kernel and LeakyReLU activation, and $\odot$ denotes element-wise multiplication. The feature transform block (FTBlock) indicated in Figure 3 is referred to as $\mathcal{G}^g(f|\gamma)$. Visual and condition features are fused with residual function [18]:

$$f' = \mathcal{G}^g(f|\gamma) + f. \tag{4}$$

The globally conditioned feature f_g is generated by consecutive residual and transform blocks for extracting the multi-level visual features mentioned in Section 3.2. The effectiveness of global conditioning is evaluated in experiments.

Local conditioning. Local interaction phrase is a relatively informative region of image tightly surrounding an HOI candidate. However, because of the diversity of human object instances and context, the appearance variety of interaction phrase can be tremendous, which pure visual features are inadequate to capture. Therefore, we construct a body-object configuration map to guide the feature extraction of interaction phrase by network conditioning. Specifically, a human instance is separated into N_p body parts by grouping the neighboring body joints introduced in Section 3.1. The body parts, $\{p_1, p_2, ..., p_{N_p}\}$, are represented with a set of bounding boxes around the corresponding body joint groups with proper margins. Since the local conditioning is applied on local visual feature f_u cropped from the global feature f_g, fine-grained spatial details like shape and edge cannot be reserved. As shown in Figure 4, we generate a box-level body-object spatial configuration map as local condition, encoding the relative locations of human-object pair (h, o) and all human body parts with two and N_p channels respectively. Each channel is a two-dimensional binary matrix with the same size of interaction phrase. The digits inside bounding box is set to 1, otherwise 0. The configuration map is fed into a local condition network to generate the local condition feature π. The local condition network consists of four convolution layers using 1×1 kernel, among which exist three LeakyReLU activation layers. Local conditioning is implemented with local feature transform module $\mathcal{G}^l$, whose structure is identical to $\mathcal{G}^g$:

$$f'_u = \mathcal{G}^l(f_u|\pi) + f_u. \tag{5}$$

Compared with global conditioning, local conditioning provides more specific spatial-semantic guidance for certain HOI candidates.

Body part attention. The function of an object decides how a human interacts with it and this means close relation exists between an object and certain body parts. To this end, we assign different attention to the visual features extracted from different body parts. We generate attention weights $w \in \mathbb{R}^{N_p}$ by feeding the word vector v of target object category into a fully-connected network:

$$w = \kappa(\kappa(vX_1 + b_1)X_2 + b_2), \tag{6}$$

where $\kappa(\cdot)$ is LeakyReLU activation function, (X_1, X_2) are project parameters and (b_1, b_2) are bias items, $X_1 \in \mathbb{R}^{L_v \times L_h}$, $b_1 \in \mathbb{R}^{1\times L_h}$, $X_2 \in \mathbb{R}^{L_h \times N_p}$, $b_2 \in \mathbb{R}^{1\times N_p}$. Since the word vectors pretrained on large-scale linguistic dataset encode the function of objects in a manner, the knowledge can be transferred among objects in different categories with similar function. The obtained attention weights applied on the visual features of body parts:

$$f'_{p_i} = f_{p_i} \cdot w_i, \tag{7}$$

where $i \in \{1, ..., N_p\}$ and f_{p_i} is the cropped feature from f_g according to the bounding box of i-th body part. The weighted and original features of all body parts are fused as follows:

$$f'_p = \mathcal{E}(\{f'_{p_1}, ..., f'_{p_{N_p}}\}) + \mathcal{E}(\{f_{p_1}, ..., f_{p_{N_p}}\}), \tag{8}$$

where $\mathcal{E}(\cdot)$ referred to channel-wise concatenation. The experiment results in Section 4.3 confirm that object context attention efficiently improves the visual features of human body parts.

3.4 Multimodal Feature Fusion

In addition to multi-level visual features, we further augment a relative location feature f_{loc} and an object context feature f_{ctx} for better performance. f_{loc} is generated by two convolution layers with max pooling, taking human-object configuration map as input following HO-RCNN. f_{loc} is frequently used by HOID methods [1, 10, 24] to encode the relative locations of bounding boxes which surround the human and object instances in interaction phrase. It is also proved effective in visual relation detection [53]. With f_{ctx}, the word vector of detected object category, the functional similarity among different objects can be captured and the interaction knowledge of these functional similar objects can be transferred [41]. So far we have obtained seven types of feature: f_h, f_o, f'_u, f_s, f'_p, f_{loc} and f_{ctx}. All these features are fed into independent fully-connected classifiers, whose output is normalized with sigmoid function to estimate probabilities for all object-independent actions. Then, we adopt a late fusion strategy following iCAN [10] to fuse the confidence values $\{\delta\}$ of actions from all branches as well as ρ_h and ρ_α, the confidence values of detected human and object in HOI candidate as follows:

$$\begin{aligned} \hat{\delta} &= (\delta^h + \delta^o + \delta^u + \delta^p + \delta^s) \odot \delta^{loc} \odot \delta^{ctx}, \\ \rho_\sigma &= \hat{\delta}_{\omega_\sigma} \cdot \rho_h \cdot \rho_{\alpha_\sigma}, \end{aligned} \tag{9}$$

where $\hat{\delta}$ refers to the fused confidence vector of actions and the superscripts of δ denote the corresponding types of feature. As for ρ_σ, it is the confidence value of HOI category σ. It is worth noting that it is impractical to obtain adequate and balanced training data considering the possibility that the category space of HOI can be quite large. We factorize HOI categories into actions and objects following [39], and recognize them independently. In this way, the

proposed method can handle large-scale category space and long-tailed data distribution. Besides, the interaction knowledge can be transferred among different objects, which makes zero-shot HOID possible [20, 39].

3.5 Model Training

In the training stage, we feed a mini-batch $B=\{(b^h, b^o, Y)\}$ into the model for each step, where Y denotes object-independent action labels $Y=\{(y_1, y_2, ..., y_{|\Omega|})\}$, Ω is action category set, $y \in \{0, 1\}$, b^h and b^o are defined in Equation (1). Since a human instance can exert multiple types of action on a target object instance, we formulate HOI recognition as a multi-label classification problem. In the training stage, we calculate independent loss values for all seven branches with binary cross entropy loss function $BCE(,)$:

$$\begin{aligned} \mathcal{L} &= \sum_{p=1}^{|B|} \sum_{q=1}^{|\Omega|} BCE(y_{p,q}, \delta_{p,q}), \\ \tilde{\mathcal{L}} &= \mathcal{L}_h + \mathcal{L}_o + \mathcal{L}_u + \mathcal{L}_s + \mathcal{L}_p + \mathcal{L}_{spa} + \mathcal{L}_{ctx}, \end{aligned} \tag{10}$$

where the subscripts of $\mathcal{L}$ indicate the corresponding branches. Here, the mini-batch loss is a sum instead of an average. It effectively avoids a situation where the samples in rare categories are overlooked and can prevent the model from being partial to the frequently-appearing categories.

4 EXPERIMENTS

4.1 Datasets and Experiment Settings

Datasets. We evaluated the proposed method on two frequently-used benchmarks, HICO-DET [1] and V-COCO [15]. HICO-DET is constructed by augmenting HICO dataset [2] with instance annotations. It includes 47,776 images (38,118 for training and 9,658 for testing) with 600 HOI categories involving 80 object and 117 action categories. Over 150K HOI instances are provided by HICO-DET. V-COCO is constructed by augmenting a subset of MS-COCO dataset [27] with interaction category annotations. It includes 10,346 images (2,533 for training and 2,867 for validation and 4,946 for testing) with 26 HOI categories. Over 16K HOI instances are provided by V-COCO.

Evaluation criteria. The official evaluation metric of both HICO-DET and V-COCO is mean average precision (mAP) on all HOI categories. For a detected HOI instance $\langle b_d^h, b_d^o, \sigma_d \rangle$, it is considered correct if there exists a groundtruth HOI instance $\langle b_g^h, b_g^o, \sigma_g \rangle$, $\min(IoU(b_d^h, b_g^h), IoU(b_d^o, b_g^o)) > \zeta$ and $\sigma_g = \sigma_d$. $IoU(,)$ is the area of bounding box intersection over the area of bounding box union. ζ denotes IoU threshold, which is equal to 0.5 in official evaluation settings.

For HICO-DET, there are two different evaluation modes: *known-object* and *default*. In *known-object* setting, average precision (AP) for a certain HOI category σ is only calculated over the predicted HOI instances of images containing objects categorized as α_σ. This setting mainly focuses on the performance of human-object localization and object-independent action recognition. Meanwhile, the influence of object classification is reduced. In *default* setting, AP for σ is calculated over the HOI detections of all images. For V-COCO, the locations of human-object pair and action category are considered while the specific object category is ignored. We follow the official metrics of both datasets in our experiments.

4.2 Implementation Details

The object detection model we used is FPN [26] with ResNet-50 [18], which is trained on MS-COCO dataset, following [24]. The word vectors for encoding object categories are trained on GoogleNews dataset. The dimension number L_v of the word vector is 300. The multi-person pose estimation model we utilized is RMPE [9], which is trained on MSCOCO-Keypoints dataset [27]. Each human instance in this dataset is represented with N_k=17 body joints, which are grouped into N_p=6 body parts, including "left arm", "right arm", "left leg", "right leg", "head" and "hip" following RPNN [54]. The human parsing method we utilized is WSHP [8] trained on PASCAL-Person-Part dataset [3], in which six different body parts are annotated at pixel level: "head", "left/right upper arms", "left/right lower arms", "left/right upper legs", "left/right lower legs" and "torso".

The proposed method takes ResNet-101 [18] pretrained on ImageNet [7] as backbone and is implemented with Pytorch [31]. The shape of ROI-aligned visual feature (w, w, c) is (7, 7, 2048). In the training stage, we adopt SGD optimizer with initial learning rate 1e−5, which is reduced to 1e−10 evenly, momentum 0.9 and weight decay 5e−4. Dropout with 50% connections is applied on all full-connected classifiers while training. The model is trained for 6 and 18 epochs on HICO-DET and V-COCO, respectively.

In the test stage, we use object detection results provided by iCAN [10] for fair comparison. The HOI candidates are generated by pairing all detected human and object instances whose confidences exceed 0.4.

4.3 Component Analysis

We evaluate the effectiveness of all the proposed components and designs on HICO-DET dataset. We construct a baseline model referred to as *Base* by eliminating all the proposed components in our method. Specifically, *Base* consists of the human branch, object branch, body part branch, phrase branch and spatial branch. The proposed components, including the scene branch (*SB*), context branch (*CB*), global conditioning (*GC*), local conditioning (*LC*) and body part attention (*BPA*), are incrementally added to *Base*.

Evaluation of multi-branch network. Recent works such as PMFNet [43] and No-Frills [16] use frozen CNN to extract visual features for HOI reasoning. To learn the distributions of different types of visual feature, all the layers of the multi-branch network (*Base*) are optimized in the training stage. To confirm the effectiveness of the multi-branch design and training strategy, we construct a comparison baseline, namely *Base-frozen*, by freezing the CNN backbone of *Base* and initializing the backbone with weights pretrained on MS-COCO dataset. The experiment results shown in the first two rows of Table 1 indicate that the CNN based HOI recognition can be greatly improved by independently learning the visual patterns at different levels.

Evaluation of new features. We argument two types of feature in different modalities, *i.e.* the global scene feature f_s and object context feature f_{ctx}, mentioned in Section 3.2 and 3.4, respectively. The corresponding scene branch (*SB*) and context branch (*CB*)

Figure 5: Qualitative results of the proposed MLCNet on V-COCO dataset. The locations of detected human and object instances are indicated with the bounding boxes and the ⟨action, object⟩ labels below the images are predicted HOI categories. The colors indicate the consistence between the labels and bounding boxes. Each label contains k actions split by "/" with the highest confidences, where k is equal to the number of annotated interactions of the human-object pair.

are added to the *Base*, generating two comparisons, *Base+CB* and *Base+CB+SB*. According to the evaluation results in Table 1, *Base+CB* improves the full mAP of *Base* by 0.59%, and *Base+CB+SB* achieves the higher full mAP in comparison with *Base* and *Base+CB*. The experiment results confirm the effectiveness of the proposed scene branch and context branch.

Evaluation of network conditioning. We construct three new comparison methods by incrementally adding body part attention (*BPA*), local conditioning (*LC*) and global conditioning (*GC*) to *Base+CB+SB*, namely *Base+CB+SB+BPA*, *Base+CB+SB+BPA+LC*, and *Ours* (*Base+CB+SB+BPA+LC+GC*). The row 5 to 7 of Table 1 indicate that *BPA*, *LC* and *GC* gradually improve the full mAP by 0.07%, 0.24% and 0.17%, respectively. Furthermore, the complete model *Ours* achieves the best performance in this experiment. The component analysis results confirm that the reasoning capability of CNN is improved with extra spatial-semantic knowledge.

4.4 Comparison with State-of-the-arts

We compare the proposed MLCNet with the existing methods on both HICO-DET and V-COCO datasets, following the official evaluation settings. Table 2 demonstrates the comparison results of our method and current state-of-the-arts.

On HICO-DET dataset, the proposed MLCNet outperforms the state-of-the-art methods under both *default* and *known-object* settings with full mAP=17.95% and 22.28% over all 600 HOI categories. In particular, the mAPs calculated on full, rare, non-rare

Table 1: Component analysis results on HICO-DET in *default* setting.

Method	full	rare	non-rare
Base-Frozen	13.85	10.99	14.71
Base	16.20	13.76	16.93
Base+CB	16.79	14.54	17.46
Base+CB+SB	17.47	15.50	18.06
Base+CB+SB+BPA	17.54	15.72	18.08
Base+CB+SB+BPA+LC	17.78	16.32	18.21
Ours	**17.95**	**16.62**	**18.35**

categories under *known-object* scenario of the proposed method significantly exceed those of the best existing method PMFNet [43], by 1.94%, 3.26% and 1.54%, respectively. It confirms that using extra knowledge as condition can effectively enhance the reasoning capability of CNN for fine-grained human-object interaction.

Owing to the sparse nature of HOI data, few-shot learning is another problem requiring special attention. The proposed MLCNet exploits object context and object-independent action for knowledge transfer, and achieves the highest rare mAPs, 16.62% and 20.73%, in both *default* and *known-object* settings, respectively. The gap between mAPs over rare categories and non-rare categories, for our method is 1.73%, which is dramatically lower than that of RPNN, 5.93%. The experiment results confirm that the adopted knowledge transfer strategies and loss function can effectively limit the negative effects from the long tailed training data and

Table 2: Experiment results of comparison with the state-of-the-art methods on HICO-DET and V-COCO.

Method	HICO-DET						V-COCO
	default			*known-object*			role mAP
	full	rare	non-rare	full	rare	non-rare	
Gupta *et al.* [15]	-	-	-	-	-	-	31.8
Shen *et al.* [39]	6.46	4.24	7.12	-	-	-	-
HO-RCNN [1]	7.81	5.37	8.54	10.41	8.94	10.85	-
Interact-Net [12]	9.94	7.16	10.77	-	-	-	40.0
GPNN [34]	13.11	9.34	14.23	-	-	-	44.0
iCAN [10]	14.84	10.45	16.15	16.26	11.33	17.73	45.3
Xu *et al.* [46]	14.70	13.26	15.13	-	-	-	45.9
Wang *et al.* [44]	16.24	11.16	17.75	17.73	12.78	19.21	47.3
Li *et al.* [24]	17.22	13.51	18.32	19.38	15.38	20.57	48.7
No-frills [16]	17.18	12.17	18.68	-	-	-	-
RPNN [54]	17.35	12.78	**18.71**	-	-	-	47.5
PMFNet [43]	17.46	15.65	18.00	20.34	17.47	21.20	53.0
Ours	**17.95**	**16.62**	18.35	**22.28**	**20.73**	**22.74**	**55.2**

lead to similar capabilities for distinguishing the HOI categories with different frequencies. Our method trades the precision over non-rare categories for better generalization capability and overall performance, which results in a non-rare mAP=18.35, slightly lower than the highest value 18.71% of RPNN [54] in turn.

According to the experiment results on V-COCO dataset, our method achieves the best performance among all the comparison methods. We improve the current state-of-the-art role mAP by 2.2%, thereby demonstrating the effectiveness of the proposed method. The advanced performance on V-COCO dataset is obtained without adjusting the hyper-parameters from the experiments on HICO-DET dataset except for the number of training epochs, showing the robustness of our model. Figure 5 shows some qualitative results. For better demonstration, we supplement object categories that are ignored in V-COCO dataset to the samples in Figure 5.

4.5 Discussion

In the experiments, we also find some limitations of the proposed method. To encode the extra spatial-semantic knowledge, we construct condition networks with multiple convolution layers, which increases the number of parameters. To learn different appearance distributions of diverse visual content, the multi-branch structure further enlarge the size of the network, which requires around 6GB GPU memory for training.

Moreover, in some complex cases involving multiple individuals, some of the detected HOI instances are less informative than the others, and this may bring negative effects to the applications that aim to describe the dominant visual content of image. As shown by the second and fifth samples in row 3 of Figure 5, the HOI instances recognized as ⟨look, person⟩ convey less semantic information than the sport instances. However, since semantically-interest HOI annotation was not taken into consideration in the previous work, existing methods cannot obtain the capability of selecting the most informative HOI instances from complex scene by supervised learning. This problem deserves more attention and future research is needed for further exploration of data collection and model design.

5 CONCLUSION

To bridge the gap between low-level visual feature of image and high-level semantic information of human object interaction, we proposed a multi-level conditioned network, which exploits extra spatial-semantic information as condition to dynamically influence the behavior of CNN. In this way, the explicit prior knowledge and implicit visual features are fused for complicated and fine-grained visual content understanding. We applied off-the-shelf human parsing and pose estimation models to obtain the body structure information of human instances in image. We also utilized object detection model to obtain the locations and categories of the entities in image. The extra knowledge is encoded by condition network and used to guide the visual feature extraction. To evaluate the effectiveness of the proposed method, we conducted the experiments on two public benchmarks, HICO-DET and V-COCO. The experiment results demonstrated that our method significantly outperforms the state-of-the-art methods and confirmed the effectiveness of the proposed multi-level condition mechanism and multimodel feature fusion.

6 ACKNOWLEDGEMENTS

This work is supported by Natural Science Foundation of Jiangsu Province (BK20191248), Science, Technology and Innovation Commission of Shenzhen Municipality (JCYJ20180307151516166), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. Learning to Detect Human-Object Interactions. In *IEEE Winter Conference on Applications of Computer Vision*. 381–389.

[2] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. HICO: A Benchmark for Recognizing Human-Object Interactions in Images. In *IEEE International Conference on Computer Vision*. 1017–1025.

[3] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. Detect What You Can: Detecting and Representing Objects using Holistic Models and Body Parts. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1971–1978.

[4] Zhen Cui, Chunyan Xu, Wenming Zheng, and Jian Yang. 2018. Context-Dependent Diffusion Network for Visual Relationship Detection. In *ACM International Conference on Multimedia*. 1475–1482.

[5] Minh-Son Dao, Pham Quang Nhat Minh, Asem Kasem, and Mohamed Saleem Haja Nazmudeen. 2018. A context-aware late-fusion approach for disaster image retrieval from social media. In *ACM on International Conference on Multimedia Retrieval*. 266–273.

[6] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *IEEE Conference on Computer Vision and Pattern Recognition*. 326–335.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: a Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[8] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. 2018. Weakly and Semi Supervised Human Body Part Parsing via Pose-Guided Knowledge Transfer. *arXiv preprint arXiv:1805.04310* (2018).

[9] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-Person Pose Estimation. In *IEEE International Conference on Computer Vision*. 2334–2343.

[10] Chen Gao, Yuliang Zou, and Jia-Bin Huang. 2018. iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. *arXiv preprint arXiv:1808.10437* (2018).

[11] Yuqi Gao, Jitao Sang, Tongwei Ren, and Changsheng Xu. 2017. Hashtag-centric immersive search on social media. In *ACM International Conference on Multimedia*. 1924–1932.

[12] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. 2018. Detecting and Recognizing Human-Object Interactions. In *IEEE Conference on Computer Vision and Pattern Recognition*. 8359–8367.

[13] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. 2015. Contextual Action Recognition with R*CNN. In *IEEE International Conference on Computer Vision*. 1080–1088.

[14] Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. 2019. Aligning Linguistic Words and Visual Semantic Units for Image Captioning. In *ACM International Conference on Multimedia*. 765–773.

[15] Saurabh Gupta and Jitendra Malik. 2015. Visual Semantic Role Labeling. *arXiv preprint arXiv:1505.04474* (2015).

[16] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. 2019. No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques. In *IEEE International Conference on Computer Vision*. 9677–9685.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *IEEE International Conference on Computer Vision*. 1026–1034.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[19] Svebor Karaman, Xudong Lin, Xuefeng Hu, and Shih-Fu Chang. 2019. Unsupervised Rank-Preserving Hashing for Large-Scale Image Retrieval. In *ACM International Conference on Multimedia Retrieval*. 192–196.

[20] Keizo Kato, Yin Li, and Abhinav Gupta. 2018. Compositional Learning for Human Object Interaction. In *European Conference on Computer Vision*. 234–251.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[22] Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-guided cross-lingual image captioning. In *ACM International Conference on Multimedia*. 1549–1557.

[23] Qing Li, Zhaofan Qiu, Ting Yao, Tao Mei, Yong Rui, and Jiebo Luo. 2016. Action recognition by learning deep multi-granular spatio-temporal video representation. In *ACM International Conference on Multimedia Retrieval*. 159–166.

[24] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. 2019. Transferable Interactiveness Knowledge for Human-Object Interaction Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3585–3594.

[25] Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in Network. *arXiv preprint arXiv:1312.4400* (2013).

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature Pyramid Networks for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2117–2125.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*. 740–755.

[28] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2018. Context-Aware Visual Policy Network for Sequence-Level Image Captioning. In *ACM International Conference on Multimedia*. 1416–1424.

[29] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual Relationship Detection with Language Priors. In *European Conference on Computer Vision*. 852–869.

[30] Xinyao Nie, Hong Lu, Zijian Wang, Jingyuan Liu, and Zehua Guo. 2019. Weakly Supervised Image Retrieval via Coarse-scale Feature Fusion and Multi-level Attention Blocks. In *ACM International Conference on Multimedia Retrieval*. 48–52.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*. 8024–8035.

[32] Liang Peng, Yang Yang, Zheng Wang, Xiao Wu, and Zi Huang. 2019. CRA-Net: Composed Relation Attention Network for Visual Question Answering. In *ACM International Conference on Multimedia*. 1202–1210.

[33] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI Conference on Artificial Intelligence*. 3943–3951.

[34] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. 2018. Learning Human-Object Interactions by Graph Parsing Neural Networks. In *European Conference on Computer Vision*. 401–417.

[35] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. 2019. Video Relation Detection with Spatio-Temporal Graph. In *ACM International Conference on Multimedia*. 84–93.

[36] Omar Seddati, Stéphane Dupont, and Saïd Mahmoudi. 2017. Quadruplet networks for sketch-based image retrieval. In *ACM International Conference on Multimedia Retrieval*. 184–191.

[37] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating Objects and Relations in User-Generated Videos. In *ACM International Conference on Multimedia Retrieval*. 279–287.

[38] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video Visual Relation Detection. In *ACM International Conference on Multimedia*. 1300–1308.

[39] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. 2018. Scaling Human-Object Interaction Recognition Through Zero-Shot Learning. In *IEEE Winter Conference on Applications of Computer Vision*. 1568–1576.

[40] Xu Sun, Tongwei Ren, Yuan Zi, and Gangshan Wu. 2019. Video Visual Relation Detection via Multi-modal Feature Fusion. In *ACM International Conference on Multimedia*. 2657–2661.

[41] Xu Sun, Yuan Zi, Tongwei Ren, Jinhui Tang, and Gangshan Wu. 2019. Hierarchical Visual Relationship Detection. In *ACM International Conference on Multimedia*. 94–102.

[42] Yi Tian, Qiuqi Ruan, Gaoyun An, and Yun Fu. 2016. Action recognition using local consistent group sparse coding with spatio-temporal structure. In *ACM International Conference on Multimedia*. 317–321.

[43] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. 2019. Pose-aware Multi-level Feature Network for Human Object Interaction Detection. In *IEEE International Conference on Computer Vision*. 9469–9478.

[44] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. 2019. Deep Contextual Attention for Human-Object Interaction Detection. In *IEEE International Conference on Computer Vision*. 5694–5702.

[45] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2018. Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform. In *IEEE Conference on Computer Vision and Pattern Recognition*. 606–615.

[46] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. 2019. Learning to Detect Human-Object Interactions With Knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2019–2028.

[47] Baohan Xu, Hao Ye, Yingbin Zheng, Heng Wang, Tianyu Luwang, and Yu-Gang Jiang. 2018. Dense dilated network for few shot action recognition. In *ACM International Conference on Multimedia Retrieval*. 379–387.

[48] Keiji Yanai and Ryosuke Tanno. 2017. Conditional fast style transfer network. In *ACM International Conference on Multimedia Retrieval*. 434–437.

[49] Fan Yu, Xin Tan, Tongwei Ren, and Gangshan Wu. 2018. Human-centric Visual Relation Segmentation Using Mask R-CNN and VTransE. In *European Conference on Computer Vision*. 582–589.

[50] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual Translation Embedding Network for Visual Relation Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5532–5540.

[51] Sipeng Zheng, Shizhe Chen, and Qin Jin. 2019. Visual Relation Detection with Multi-Level Attention. In *ACM International Conference on Multimedia*. 121–129.

[52] Yuheng Zhi, Huawei Wei, and Bingbing Ni. 2018. Structure Guided Photorealistic Style Transfer. In *ACM International Conference on Multimedia Conference*. 365–373.

[53] Hao Zhou, Chongyang Zhang, and Chuanping Hu. 2019. Visual Relationship Detection with Relative Location Mining. In *ACM International Conference on Multimedia*. 30–38.

[54] Penghao Zhou and Mingmin Chi. 2019. Relation Parsing Neural Network for Human-Object Interaction Detection. In *IEEE International Conference on Computer Vision*. 843–851.

[55] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. 2018. HCVRD: A Benchmark for Large-Scale Human-Centered Visual Relationship Detection. In *AAAI Conference on Artificial Intelligence*. 7632–7638.