Progressive Instance-Aware Feature Learning for Compositional Action Recognition

Rui Yan, Lingxi Xie, Xiangbo Shu, *Senior Member, IEEE*, Liyan Zhang, and Jinhui Tang, *Senior Member, IEEE*

***Abstract*—In order to enable the model to generalize to unseen "action-objects" (compositional action), previous methods encode multiple pieces of information (i.e., the appearance, position, and identity of visual instances) independently and concatenate them for classification. However, these methods ignore the potential supervisory role of instance information (i.e., position and identity) in the process of visual perception. To this end, we present a novel framework, namely Progressive Instance-aware Feature Learning (PIFL), to progressively extract, reason, and predict dynamic cues of moving instances from videos for compositional action recognition. Specifically, this framework extracts features from foreground instances that are likely to be relevant to human actions (Position-aware Appearance Feature Extraction in Section III-B1), performs identity-aware reasoning among instance-centric features with semantic-specific interactions (Identity-aware Feature Interaction in Section III-B2), and finally predicts instances' position from observed states to force the model into perceiving their movement (Semantic-aware Position Prediction in Section III-B3). We evaluate our approach on two compositional action recognition benchmarks, namely, Something-Else and IKEA-Assembly. Our approach achieves consistent accuracy gain beyond off-the-shelf action recognition algorithms in terms of both ground truth and detected position of instances.**

***Index Terms*—Compositional action recognition, compositional generalization, human action recognition.**

I. Introduction

HUMAN action recognition aims at understanding the behaviors of humans from given video sequences. In the past decade, inspired by the success of deep learning in image representations, many 2D [2], [3], [4] and 3D [5], [6], [7], [8] neural networks have been designed to extract representations from videos. Existing video backbones work well for basic human actions [9], [10] involving only body motion and posture (such as walking and jumping), but there is still a long way to go in understanding complex activities involving spatio-temporal interactions between humans and objects [11], [12].

Having said that, why do humans easily understand a complex action (e.g., "*Moving sth. and sth. closer to each other*" as shown in Fig. 1(c)) even when it is performed with different objects in various environments? Their ability to perform compositional reasoning between entities in the natural world is probably the most plausible explanation [13]. For example, humans often focus on various visual instances (*i.e.*, two boxes and two hands) in the scene of Fig. 1 (c) and then recognize the whole activity by observing the relative change of the distance between instances. This ability helps humans to learn some knowledge that is easy to generalize to the novel environment with unseen combinations. Inspired by this, we hope to make machines show a similar capability in action recognition, namely, compositional action recognition [11], [12], which requires the "action-object" pairs in training and testing sets do not overlap.

According to the above definition, the main challenge of compositional action recognition is the out-of-distribution [11] issue of the testing set. However, it is well known that deep models inevitably produce inductive biases [14] between the visual input and labels on the training set when the samples are insufficient. For example, the confusion of actions in Fig. 1 (c) and (d) may be caused by the similar appearance. To alleviate this issue, an intuitive solution is to introduce additional modal information, such as position information (e.g., instance positions that can be used to easily distinguish Fig. 1 (c) from (d)). However, the position information does not always work on some actions involving the property of objects, e.g., the paper can be torn by hands in Fig. 1 (b) but the spoon can not be in Fig. 1 (a). Therefore, integrating multiple information has become a promising solution for compositional action recognition, but it is also extremely challenging.

Previous work (STIN [11]) directly encodes instance information (e.g., position and identity) independently into features for interaction and then concatenates them with global appearance feature for classification (as shown in Fig. 3), which is feasible but rough. We hold that instance information is critical for motion feature extraction at different stages to understand compositional actions. For example, when humans understand complex actions, they often need to focus on moving objects

Manuscript received 9 November 2021; revised 7 January 2023; accepted 13 March 2023. Date of publication 27 March 2023; date of current version 30 June 2023. This work was supported by the National Key Research and Development Program of China under Grant 2018AAA0102002, in part by the National Natural Science Foundation of China under Grants 61925204, 62222207, 62072245, and 62172212, and in part by the the Natural Science Foundation of Jiangsu Province under Grant BK20211520. Recommended for acceptance by A. Kovashka. *(Corresponding authors: Liyan Zhang.)*

Rui Yan, Xiangbo Shu, and Jinhui Tang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu, China (e-mail: ruiyan@njust.edu.cn; shuxb@njust.edu.cn; tangjh1981@acm.org).

Lingxi Xie is with the Huawei Inc., Shenzhen, Guangdong 518100, China (e-mail: 198808xc@gmail.com).

Liyan Zhang is with the College of Computer Science and Technology, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu 211106, China (e-mail: zhangliyan@nuaa.edu.cn).

Digital Object Identifier 10.1109/TPAMI.2023.3261659

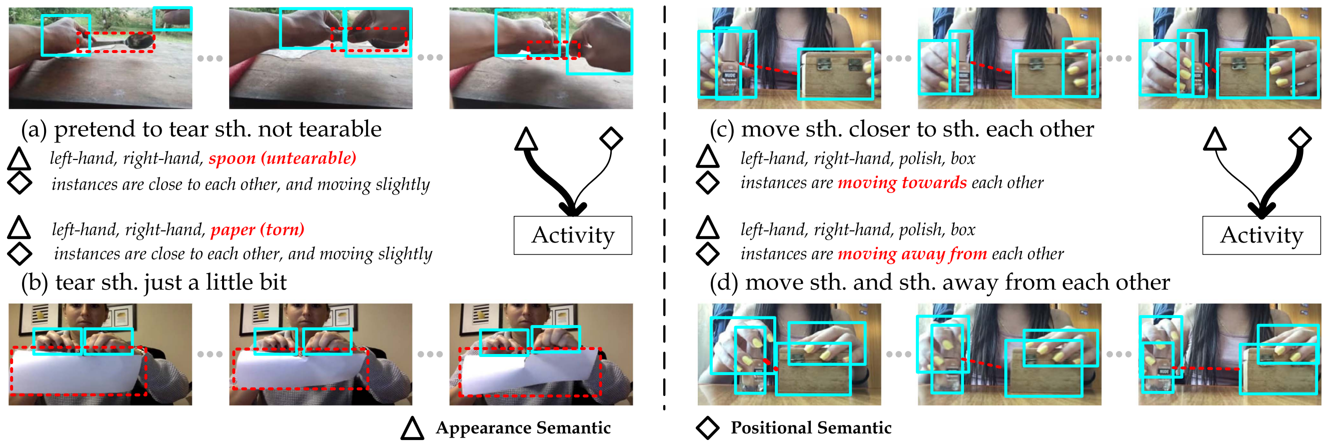

(a) pretend to tear sth. not tearable
△ left-hand, right-hand, spoon (untearable)
◇ instances are close to each other, and moving slightly
△ left-hand, right-hand, paper (torn)
◇ instances are close to each other, and moving slightly
(b) tear sth. just a little bit
△ ◇ Activity
(c) move sth. closer to sth. each other
△ left-hand, right-hand, polish, box
◇ instances are moving towards each other
△ left-hand, right-hand, polish, box
◇ instances are moving away from each other
(d) move sth. and sth. away from each other
△ ◇ Activity
△ Appearance Semantic
◇ Positional Semantic

Fig. 1. Compositional examples in Something-Something [1]. We annotate all instances (*i.e.*, hands and objects) by the cyan boxes and highlight the major difference by the red dashed boxes or lines in each group of comparisons. By observing the appearance changes of objects (e.g., *untearable spoon* and *torn paper*), humans can easily distinguish between (a) and (b). In contrast, humans understand "closer to" in (c) and "away from" in (d) by observing the relative displacement (shown as the red dashed line) among instances rather than by the objects' appearances. Motivated by this, we aim to fuse different types of information to understand compositional actions. Best viewed in color.

and reason about their relationships with the help of identities. Furthermore, if people can accurately predict the future motion of instances (related to the action), they must already have a good understanding of the observed action. Inspired by these points, this work aims at injecting instance information (position and identity) into the process of video feature learning progressively.

Inspired by this, we propose a Progressive Instance-aware Feature Learning framework for compositional action recognition. Progressive Instance-aware Feature Learning achieves compositional generalization by endowing the model with the ability to extract, reason, and predict dynamic cues of moving instances, with the help of instance information. It extracts features from moving instances that are most likely to be related to action semantics; performs identity-aware reasoning among them to earn semantic-specific relational structures, and finally predicts their future position from observed states to enhance the model's capability to motion perception. Specifically, the framework is implemented via the following three steps. i) Position-aware Appearance Feature Extraction: builds instance-centric appearance representation from images according to instance positions and combines them with non-appearance features (from position and identity information [11]) into instance-centric hybrid features; ii) Identity-aware Feature Interaction: builds identity-aware pairwise relationships among these hybrid features in the latent space to generate semantic features for each instance; iii) Semantic-aware Position Prediction: projects the semantic features back to position space by an auxiliary task of instance position prediction, to enhance the model's perception of objects' movement.

The proposed approach is evaluated on two challenging datasets: Something-Else [11] and IKEA-Assembly [15]. Experimental results show that our approach significantly outperforms state-of-the-art methods [6], [11], [16] on compositional action recognition, regardless of whether the object positions are labeled or detected. In particular, on the Something-Else dataset, our approach achieves 2.7% and 3.5% improvements on top-1 and top-5 accuracy with detections, respectively, compared with [11]. In addition, our approach shows good generalization ability in few-shot setting.

Our contributions can be summarized as:

- We propose a novel Progressively Instance-aware Feature Learning framework which progressively injects instance information (position and identity) into the video feature extraction at different stages.
- Instance-aware feature learning is progressively implemented at different stages as i) extracting instance-centric features with the help of position; ii) guiding the feature interaction through identity information; iii) predicting the position from semantic video features.
- Comprehensive experiments on two datasets demonstrate the effectiveness of our approach and some diagnostic studies show the interpretability of our approach.

The rest of this paper is organized as follows. Section II briefly discusses related work on activity recognition, the compositionality of activity, instance-centric video representation, and video prediction. In Section III, we present the formulation of compositional action recognition and our approach as also the implementation details. We then show the results of a number of experiments on two datasets, *i.e.*, Something-Else and IKEA-Assembly and conduct a comprehensive analysis, in Section IV. Finally, our work is concluded in Section V.

II. Related Work

In this section, we briefly review the traditional action recognition task, point out its bottleneck and then introduce the compositionality in the activity we are concerned with. Moreover, two related technologies (*i.e.*, instance-centric video representation and prediction in the video) used in the proposed approach are reviewed for better understanding.

A. Human Action Recognition

Recognizing the behaviors of humans in videos has been a fundamental problem in the computer vision [3], [7], [16], [17]. In recent years, researchers have been scaling up the video databases (e.g., HMDB51 [10], UCF101 [9], Charades [18], Something-Something [1], ActivityNet [19], and Kinetics [20]) and building numerous powerful backbones (e.g., TSN [4], C3D [7], I3D [5], TSM [21], SlowFast [22] and TEA [23]) on them. However, in real world scenarios, human behavior is usually instantiated by objects or environments. These proposed algorithms are still not robust to novel combinations of known actions and objects [11], [24], [25], which hinders the application action recognition technology. Here, we believe that endowing existing models with compositional generalization [26], [27] may be an urgent and promising direction of action understanding.

B. Compositionality in Activity Recognition

Human activities are composed of a set of subactions in the temporal domain [28], [29] and various subjects and objects in the spatial domain [12], [30]. Recent works proposed compositional annotations or settings based on some popular video-based datasets [1], [18]. For example, Materzynska et al. [11] provided object bounding box annotations for the Something-Something dataset [1] and presented a compositional setting in which there is no overlap between the verb-noun combinations in the training and testing sets. In addition, built upon Charades [18], Action Genome [12] decomposed activities into spatiotemporal scene graphs as the intermediate representation for understanding them. To control the scene and object bias, Girdhar et al. [24] created a synthetic video dataset, CATER, in which the events are broken up into several atomic actions in the spatial and temporal domains. In this work, we focus on generalizing compositional action to novel environments by progressively injecting instance information (position and identity) into video feature extraction.

C. Instance-Centric Video Representation

For video understanding, previous works [5], [16], [21], [22], [31] focused on designing deep and powerful backbones to extract appearance features from each frame. However, it is difficult for these approaches to mine rich relationships [6], [12], [32], [33] between different instances (objects or hands or persons, etc.) in the spatial and temporal domains. To this end, some recent studies [6], [17], [30], [33], [34] focused on extracting instance-centric representations from videos and building the spatial and temporal relationships among them. For example, Wang et al. [6] represented videos as space-time region graphs and modeled the long-range relationships among regions for action recognition. Similarly, Baradel et al. [30] and Ma et al. [33] designed more sophisticated neural networks with the ability to reason about the spatio-temporal interactions among the semantic instances/objects in videos. In this paper, we construct instance-centric representations from both multiple information for compositional action recognition.

D. Predictability of Motion

Predicting small image patches [35], [36], [37] or full frames [38], [39], [40] in videos has received increasing attention in recent years. A line of recent works [41], [42], [43], [44], [45] focused on disentangling each frame into several instances and predicting their state in terms of mass, location, velocity, *etc*. The sequence prediction mechanism has also been used in proxy tasks [46], [47], [48] to learn self-supervised representations for video. For example, Oord et al. [48] proposed Contrastive Predictive Coding to learn effective video representations by predicting the future frame-level feature in latent space. Han et al. [46] designed a Dense Predictive Coding framework to learn dense spatiotemporal video embeddings by recurrently predicting the future. In this work, we predict the location and offset of instances from instance-centric semantic representations to promote the fusion of multiple information.

III. Approach

A. Problem Statement

Formally, given a video with T frames, N instances (e.g., objects and hands/persons), and the associated tracklets, we denote the RGB input of this video as $\boldsymbol{V} \in \mathbb{R}^{T\times H\times W\times 3}$ (H is the height and W is the width), the tracklets of instances as $\boldsymbol{B} \in \mathbb{R}^{T\times N\times 4}$, the object identity as $\boldsymbol{C} \in \mathbb{R}^{T\times N\times 1}$ (*indicates only "hand/person" and "object"*), and the activity label of this video as l. We extract the visual appearance features $A \in \mathbb{R}^{\frac{T}{2}\times H\times W\times d_{\text{fea}}}$ from the video V via I3D [5]. $\boldsymbol{B}$ and $\boldsymbol{C}$ can be ground-truth as well as predictions. Compositional activity recognition [11] aims at understanding unseen combinations of action (performed by hands or persons) and objects in each video, which brings up the problem of out-of-distribution generalization [26], [27].

Previous methods [11], [49] often integrate multiple features from different sources of information for this task. For example, appearance features extracted from video $\boldsymbol{V}$ contain the attributes of instances or the environment. Beyond that, we can also obtain position features that describe the trajectories (*i.e.*, motions) of instances from tracklets $\boldsymbol{B}$ and the object identity feature used to identify each instance from $\boldsymbol{C}$. In general, appearance features take on thousands of dimensions, but non-appearance features (from position and object identity) can be represented by vectors of several tens of dimensions.

In [11], the authors showed that both appearance and non-appearance (position and identity) information are useful for understanding a complex compositional activity. However, the features independently extracted from $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{V}$ are simply concatenated for fusion (as shown on the left of Fig. 3). In our view, different information for the same video may complement each other instead of being independent during feature extraction.

B. Progressive Instance-Aware Feature Learning

To this end, we propose a simple yet effective framework that progressively fuses multiple information at different stages of video feature extraction (Fig. 2). The general formulation of

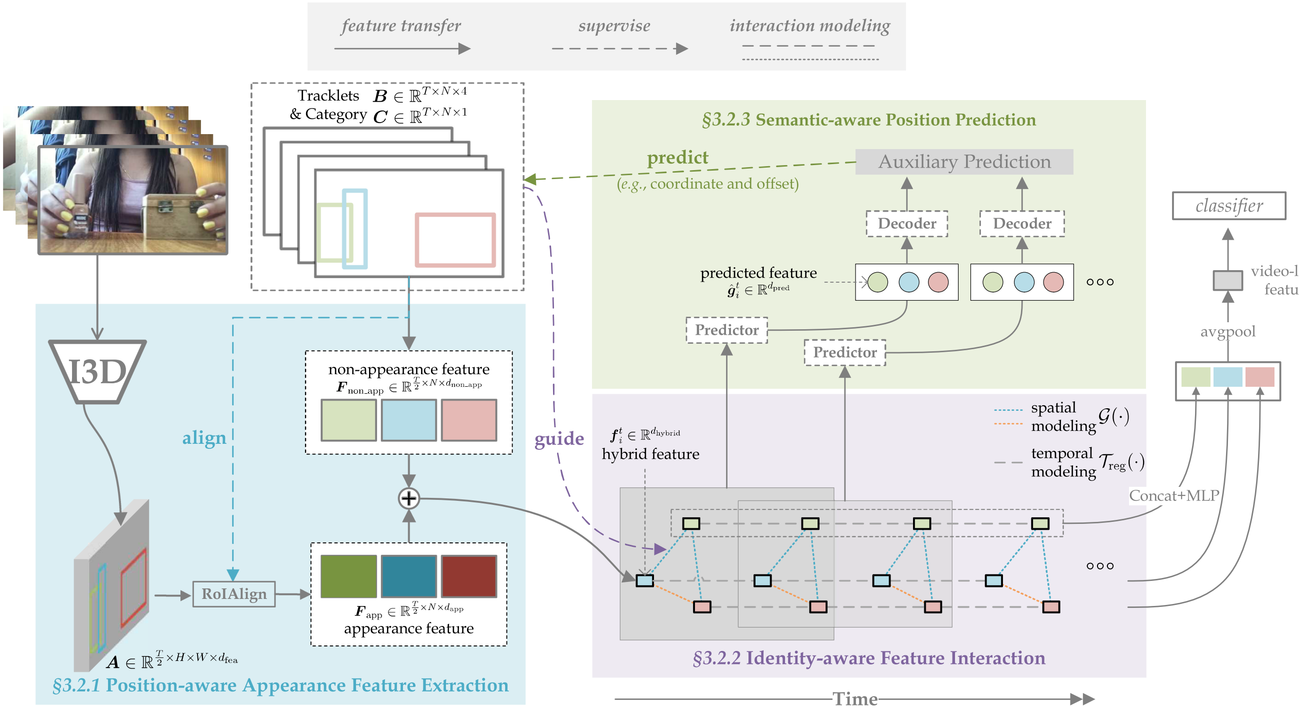

feature transfer
supervise
interaction modeling
Tracklets $B \in \mathbb{R}^{T\times N\times 4}$
& Category $C \in \mathbb{R}^{T\times N\times 1}$
§3.2.3 Semantic-aware Position Prediction
predict
(e.g., coordinate and offset)
Auxiliary Prediction
Decoder
Decoder
predicted feature $\hat{g}_i^t \in \mathbb{R}^{d_{\text{pred}}}$
Predictor
Predictor
classifier
video-level feature
avgpool
I3D
align
non-appearance feature $F_{\text{non\_app}} \in \mathbb{R}^{\frac{T}{2}\times N\times d_{\text{non\_app}}}$
guide
$f_i^t \in \mathbb{R}^{d_{\text{hybrid}}}$ hybrid feature
spatial modeling $\mathcal{G}(\cdot)$
temporal modeling $\mathcal{T}_{\text{reg}}(\cdot)$
Concat+MLP
RoIAlign
$F_{\text{app}} \in \mathbb{R}^{\frac{T}{2}\times N\times d_{\text{app}}}$ appearance feature
$A \in \mathbb{R}^{\frac{T}{2}\times H\times W\times d_{\text{fea}}}$
§3.2.1 Position-aware Appearance Feature Extraction
§3.2.2 Identity-aware Feature Interaction
Time

Fig. 2. Overview of the proposed framework. It takes T frames sampled from a video, the associated position, and identity information (distinguishing only "hand" or "object") of each instance as inputs. This framework is composed of three steps. Position-aware Appearance Feature Extraction (PAFE) (in Section III-B1): pastes a set of boxes on the video to extract instance-centric appearance features and concatenate them with position features into hybrid features; Identity-aware Feature Interaction (IFI) (in Section III-B2): builds pairwise relationships between instance-centric hybrid features with the consideration of object identity (only identifying the hand or object). Semantic-aware Position Prediction (SPP) (in Section III-B3): recovers part of the high-level information from semantic features (output from IFI) by predicting the future position information (e.g., coordinate and offset) of each instance. Best viewed in color.

this framework can be abstracted as follows:

$$\begin{aligned} \boldsymbol{G} &= \mathcal{G}(\mathcal{F}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}),\boldsymbol{C}), \\ \hat{\boldsymbol{B}} &= \mathcal{T}_{\text{aux}}(\boldsymbol{G}), \boldsymbol{Z} = \mathcal{T}_{\text{reg}}(\boldsymbol{G}). \end{aligned} \tag{1}$$

The framework takes as inputs $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$, which denote the appearance features extracted from video $\boldsymbol{V}$, the position (tracklets) and object identity, respectively. $\hat{\boldsymbol{B}}$ and $\boldsymbol{Z}$ represent the predicted position information and instance-centric spatio-temporal features, respectively.

Different from the most related work STIN [11] (more detail can be found in Section III-C), we aim at injecting instances information (position and identity) to the video feature learning process progressively via the following three functions. First, $\mathcal{F}(\cdot)$ aims at making full use of instance position information ($\boldsymbol{B}$) to enhance instances' motion cues when extracting appearance features ($\boldsymbol{A}$) from videos and output hybrid features from three types of information based on instances; Second, $\mathcal{G}(\cdot)$ aims at further exploring the spatio-temporal dependencies among instance-centric hybrid features with the consideration of object identity $\boldsymbol{C}$ and generating high-level semantic representations $\boldsymbol{G}$; Third, $\mathcal{T}_{\text{aux}}(\cdot)$ is an auxiliary function designed to project high-level semantic representations $\boldsymbol{G}$ back to the low-dimensional space by estimating $\hat{\boldsymbol{B}}$.

In addition, $\mathcal{T}_{\text{reg}}(\cdot)$ aggregates instance-centric features in the temporal domain for final recognition. Note that $\mathcal{F}(\cdot)$, $\mathcal{G}(\cdot)$, and $\mathcal{T}_{\text{aux}}(\cdot)$ can be designed in different forms with the requirement that $\mathcal{F}(\cdot)$ is built on instances rather than frames, $\mathcal{G}(\cdot)$ needs to further leverage more existing supervision to promote feature fusion in the latent space, and $\mathcal{T}_{\text{aux}}(\cdot)$ can be free from the additional annotation.

1) Position-Aware Appearance Feature Extraction (PAFE):

As mentioned above, function $\mathcal{F}(\cdot)$ aims at enhancing instances' motion cues with the help of instance position information while extracting appearance features from videos. In this work, we adopt a simple method: extracting instance-centric appearance representations from videos according to the position information, which highlights the dynamic clues from local areas (instances) and relieves the inductive bias from backgrounds.

Specifically, we sample T frames from each video and employ I3D [5] to extract the spatiotemporal appearance representation from the video $\boldsymbol{V}$. The output of the last convolutional layer is a feature map $\boldsymbol{A}$ with dimensions of $\frac{T}{2} \times H \times W \times d_{\text{fea}}$. To obtain N instance-centric features, we apply a RoIAlign [50] on the feature map $\boldsymbol{A}$ to crop and rescale the appearance feature for each instance according to the associated tracklets $\boldsymbol{B}$. Instance-centric appearance features are denoted as $\boldsymbol{F}_{\text{app}} \in \mathbb{R}^{\frac{T}{2}\times N\times d_{\text{app}}}$, where N is the number of instances.

Furthermore, we combine the appearance features with non-appearance features used in [11], *i.e.*, the position features $\boldsymbol{F}_{\text{bb}} \in \mathbb{R}^{\frac{T}{2}\times N\times d_{\text{bb}}}$ from tracklets $\boldsymbol{B}$ and object identity features $\boldsymbol{F}_{\text{idt}} \in \mathbb{R}^{\frac{T}{2}\times N\times d_{\text{idt}}}$ from object identity $\boldsymbol{C}$ by word embedding. $\boldsymbol{F}_{\text{bb}}$ and $\boldsymbol{F}_{\text{idt}}$ are concatenated and embedded into non-appearance representation $\boldsymbol{F}_{\text{non\_app}} \in \mathbb{R}^{\frac{T}{2}\times N\times d_{\text{non\_app}}}$ via an MLP. Finally, we concatenate $\boldsymbol{F}_{\text{app}}$ and $\boldsymbol{F}_{\text{non\_app}}$ at the last dimension as instance-centric hybrid representations $\boldsymbol{F} \in \mathbb{R}^{\frac{T}{2} \times N \times d_{\text{hybrid}}}$ for each video, where $d_{\text{hybrid}} = d_{\text{app}} + d_{\text{non\_app}}$.

2) Identity-Aware Feature Interaction (IFI): Instance-centric hybrid representation $\boldsymbol{F}$ extracted from each instance is isolated now and lacks the spatio-temporal semantics among multiple source information. Consequently, it is natural to model latent interactions among these instances via many choices, including Non-Local [16], STIN [11], STRG [6], and Transformer [51]. However, we know that the relationship between different visual instances is different. For example, the relationship between objects generally reflects position-related semantics (e.g., "on the left," "away from"), while the relationship between humans (hands) and objects often reflects motion-related semantics. Therefore, the differentiated instance-centric feature interaction helps to enhance the generalization of visual features. Here, we simply instantiate function $\mathcal{G}(\cdot)$ by constructing pairwise spatial relationships between each instance with the consideration of their identity and exploring the temporal dependencies via sequential models.

Spatial Modeling. Different from previous relational modules [11], [16], [17] that construct relationships between a specific feature node and its neighbors indistinguishably, we model the different types of pairwise relationships between each instance with the consideration of their identities. Formally, we represent each video by a set of instance-centric hybrid features as $\boldsymbol{F} \in \mathbb{R}^{\frac{T}{2} \times N \times d_{\text{hybrid}}}$ and perform spatial reasoning [16], [52], [53] among them as follows:

$$\boldsymbol{g}_i^t = \mathcal{G}(\boldsymbol{F}^t, \boldsymbol{C}^t) = \psi^{\text{S}} \left(\boldsymbol{f}_i^t + \sum_{\forall (i,j) \in \mathcal{E}_{\text{ss}}} \phi_{\text{ss}} \left([\boldsymbol{f}_i^t, \boldsymbol{f}_j^t] \right) + \sum_{\forall (i,j) \in \mathcal{E}_{\text{so}}} \phi_{\text{so}} \left([\boldsymbol{f}_i^t, \boldsymbol{f}_j^t] \right) + \sum_{\forall (i,j) \in \mathcal{E}_{\text{oo}}} \phi_{\text{oo}} \left([\boldsymbol{f}_i^t, \boldsymbol{f}_j^t] \right) \right), \tag{2}$$

where $\boldsymbol{f}_i^t$ and $\boldsymbol{g}_i^t$ denote the hybrid and spatial relational features for the i-th instance at the t-th time step, respectively. $[\cdot, \cdot]$ represents a concatenation operation used to compose a pair of instance-centric features, *i.e.*, $\boldsymbol{f}_i^t$ and $\boldsymbol{f}_j^t$, where $j \neq i$. $\mathcal{E}_{\text{ss}}$, $\mathcal{E}_{\text{so}}$, and $\mathcal{E}_{\text{oo}}$ represent different instance-pair sets (*i.e.*, subject-subject, subject-object, and object-object interactions, where the subject indicates the hand or person) that can be defined according to instance identity $\boldsymbol{C}$. Functions $\phi_{\text{ss}}(\cdot)$, $\phi_{\text{so}}(\cdot)$, and $\phi_{\text{oo}}(\cdot)$ are designed to encode different pairwise interactions, and $\psi^{\text{S}}(\cdot)$ is used to fuse them with original hybrid features $\boldsymbol{f}_i^t$. All above four functions can be implemented by MLPs. This module outputs instance-centric spatial features $\boldsymbol{G} \in \mathbb{R}^{\frac{T}{2} \times N \times d_{\text{sem}}}$.

Temporal Modeling. Given i-th instance' spatial features, $\boldsymbol{G}_i \in \mathbb{R}^{\frac{T}{2} \times d_{\text{sem}}}$, we further fuse them along the temporal domain as:

$$\boldsymbol{Z}_i = \mathcal{T}_{\text{reg}}(\boldsymbol{G}_i) = \psi^{\text{T}} \left(\texttt{Cat} \left(\texttt{M\_Seq} \left(\left[\boldsymbol{g}_i^1, \ldots, \boldsymbol{g}_i^{\frac{T}{2}} \right]; \boldsymbol{\theta}^{\tau} \right) \right) \right), \tag{3}$$

Notably, `M_Seq` represents a sequential model that is used to capture temporal dependencies and is optional. Without `M_Seq`, this step degrades into a naive temporal concatenation used in [11], [54]. It can be implemented via LSTM [55], [56] or Transformer [57]. $\texttt{Cat}(\cdot)$ is used to concatenate all $\frac{T}{2}$ outputs from the sequential model `M_Seq` along the last dimension, and $\psi^{\text{T}}(\cdot)$ is an MLP used to encode the concatenated features for each instance. $\boldsymbol{\theta}^{\tau}$ denotes the learnable parameters used in `M_Seq`. For each video, N instances' spatio-temporal features $\boldsymbol{Z} = \{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N\}$ are used to recognize the final activity; refer to Section III-B4.

3) Semantic-Aware Position Prediction (SPP): To enhance the model's ability to perceive instance motion, we further implement $\mathcal{T}_{\text{aux}}(\cdot)$ by predicting the future position information of each instance from the observed hybrid features. Inspired by recent self-supervised action recognition methods [35], [58], not only the absolute coordinates of instances but also their relative offsets are predicted. Predicting "coordinate" is similar to locating objects, but predicting "offset" focuses more on the movement of instances. When the position information of each instance is not inaccurate, predicting the coordinates is difficult for training. Therefore, we combine these two objectives and the function defined in (1) can be rewritten as:

$$[\hat{\boldsymbol{B}}, \hat{\boldsymbol{O}}] = \mathcal{T}_{\text{aux}}(\boldsymbol{G}) = \texttt{Decoder}(\texttt{Predictor}(\boldsymbol{G})), \tag{4}$$

where $\boldsymbol{G}$ denotes instance-centric spatial features and $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{O}}$ are predicted coordinates and offsets, respectively.

Predictor. We first estimate the future state of each instance by the observed features. Formally, given the previous t observed spatial features of the i-th instance as $\{\boldsymbol{g}_i^1, \boldsymbol{g}_i^2, \ldots, \boldsymbol{g}_i^t\}$, the $(t+1)$-th state is predicted as:

$$\hat{\boldsymbol{g}}_i^{t+1} = \texttt{M\_Seq} \left(\{\boldsymbol{g}_i^1, \boldsymbol{g}_i^2, \ldots, \boldsymbol{g}_i^t\}; \boldsymbol{\theta}^{\tau} \right). \tag{5}$$

Here, we denote T_{obs} as the number of frames observed by the predictor in each step and $t \in [T_{\text{obs}} : \frac{T}{2})$. `M_Seq` is used to model the temporal structure hidden in sequence data, and the last output is viewed as the predicted state $\hat{\boldsymbol{g}}_i^{t+1} \in \mathbb{R}^{d_{\text{pred}}}$. *It shares parameters with the sequential model used in (3), and they are trained synchronously.*

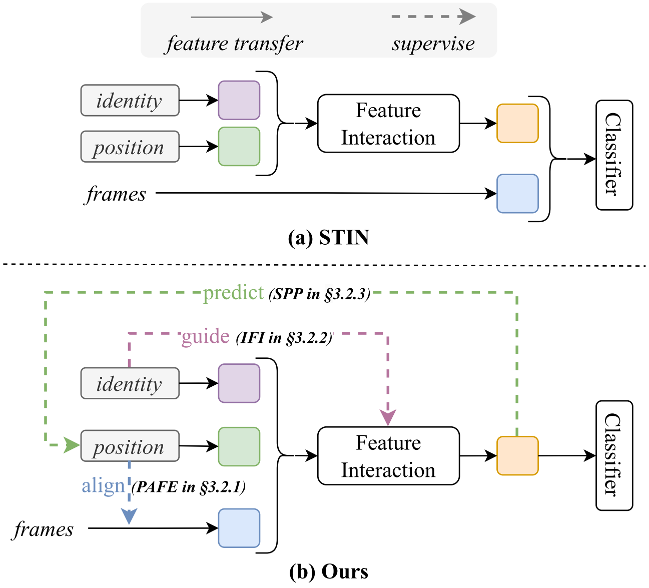
Decoder. Given the predicted state $\hat{\boldsymbol{g}}_i^{t+1}$ for the i-th instance in time $t+1$, we estimate its absolute position and relative offsets (*i.e.*, the difference between the centers of an instance in two consecutive frames) via two simple linear layers as $[\hat{\boldsymbol{b}}_i^{t+1}, \hat{\boldsymbol{o}}_i^{t+1}] = [\boldsymbol{W}_b^{\top} \hat{\boldsymbol{g}}_i^{t+1}, \boldsymbol{W}_o^{\top} \hat{\boldsymbol{g}}_i^{t+1}]$, where $\boldsymbol{W}_b \in \mathbb{R}^{d_{\text{pred}} \times 4}$ and $\boldsymbol{W}_o \in \mathbb{R}^{d_{\text{pred}} \times 2}$. $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{O}}$ are padded with observed values when $t < T_{\text{obs}}$.

4) Training and Inference: During the training phase, our approach not only needs to predict human activities but also needs to estimate the position of the involved visual instances. Therefore, the total objective function of our approach consisting of recognition and auxiliary loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{reg}}(\hat{\boldsymbol{y}}, \boldsymbol{y}) + \mathcal{L}_{\text{aux}}(\hat{\boldsymbol{B}}, \hat{\boldsymbol{O}}, \boldsymbol{B}, \boldsymbol{O}), \tag{6}$$

where $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{B}}$, and $\hat{\boldsymbol{O}}$ denote video-level prediction of activity, predicted coordinates and offsets of tracklets, respectively. $\boldsymbol{y}$, $\boldsymbol{B}$, and $\boldsymbol{O}$ are the corresponding ground-truths.

Recognition loss $\mathcal{L}_{reg}(\cdot)$ is a simple cross-entropy loss that is used to recognize the final compositional activities. It is defined



(a) STIN

(b) Ours

Fig. 3. Differences between STIN [11] and our approach.

as

$$\mathcal{L}_{\text{reg}} = -\frac{1}{N_{\text{Y}}}\sum_{i=1}^{N_{\text{Y}}}\left(\boldsymbol{y}_i\left(\log\frac{\exp(\hat{\boldsymbol{y}}_i)}{\sum_{i=1}^{C}\exp(\hat{\boldsymbol{y}}_i)}\right)\right), \tag{7}$$

where $\boldsymbol{y}$ is the one-hot-encoded ground-truth of activity and N_{Y} is the number of activity classes. We average-pool over N instances' spatio-temporal features $\{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N\}$ output via IFI to the video-level features Z^*, and feed it to a linear layer with the output dimension of N_{Y} for calculating the prediction scores $\hat{\boldsymbol{y}}$.

Auxiliary loss $\mathcal{L}_{aux}(\cdot)$ designed to measure the error between the predicted and ground-truth position information. It is simply defined as the Euclidean distance between the prediction from (4) and the ground truth and sums them over space and time as:

$$\mathcal{L}_{\text{aux}} = \sum_{\forall t}\sum_{i=1}^{N}\left(\left\|\hat{\boldsymbol{b}}_i^{t+1} - \boldsymbol{b}_i^{t+1}\right\|_2^2 + \left\|\hat{\boldsymbol{o}}_i^{t+1} - \boldsymbol{o}_i^{t+1}\right\|_2^2\right), \tag{8}$$

where $t \in [T_{\text{obs}}, \frac{T}{2})$ and N is the number of instances in the video. $\boldsymbol{b}_i^{t+1}$ and $\boldsymbol{o}_i^{t+1}$ are the ground-truth position and offset converted from tracklets $\boldsymbol{B}$, respectively. $\boldsymbol{o}_i^{t+1} = (\boldsymbol{b}_i^{t+1} - \boldsymbol{b}_i^{t})[:2]$, where $[:2]$ takes only the previous two elements of vectors.

Inference. To obtain the final activity prediction $\hat{\boldsymbol{l}}$, we average pool N instances' spatio-temporal features $\boldsymbol{Z}$ as the video-level representation followed by a `softmax` layer, as shown on the right of Fig. 2.

C. Discussion

We compare our approach with the most related work, STIN [11] in Fig. 3 to highlight our contributions. Overall, STIN extracts features from instance information as a supplement to conventional visual/appearance features for compositional action recognition. Different from STIN, our approach treats instance information as a "supervisor" to guide the visual feature extraction for enhancing its compositional generalization.

Specifically, as shown in Fig. 3(a), STIN directly extracts features from instance information (position and identity) independently for interaction and concatenates them with global appearance features (from whole frames) for classification, which is feasible but ignores the dependence between the two types of information. However, our approach aims at extracting, reasoning, and predicting dynamic cues of moving instances from videos progressively for improving the compositional generalization of existing action recognition algorithms, as shown in the colored dotted arrows. This can be summarized as follows,

- To highlight dynamic cues of regions (visual instances) most likely to be action-related, we extract instance-centric appearance features according to position information. (Section III-B1)
- To avoid undifferentiated associations between multiple instances, we apply identity information to guide the process of instance-centric feature interaction. (Section III-B2)
- To enhance the model's ability to perceive instance motion, we predict the position of instances from semantic features output from the interaction module. (Section III-B3)

The designs of the above modules follow the unified motivation, rather than independently or incrementally.

D. Implementation Details

1) Input: Our approach takes as input T frames [1] sampled from each video, and the associated tracklets and object identity for N instances (hands/persons and objects, no more than four) in the scene. Each frame is resized to a resolution of 224 × 224. The tracklet of each object can be ground-truth boxes annotated by humans or detections generated by various off-the-shelf detection algorithms.

2) Network Architecture: For a fair comparison, we employ the I3D model [5] built on ResNet-50 as the backbone unless stated otherwise and more details are introduced in [6]. The RoIAlign [50] extracts region-based features for each instance with a size of 3×3 on top of the last convolutional layer. Thus, the RoIAlign layer generates $N \times 3 \times 3 \times d_{\text{fea}}$ output features for each video, which are then flattened and embedded into $N \times d_{\text{app}}$ via a linear transformation, where $N = 4$ and $d_{\text{fea}} = d_{\text{app}} = 512$.

In addition, the MLP used to fuse non-appearance features in Section III-B1 and $\psi^{\text{T}}(\cdot)$ defined in (3) are composed of two linear layers with output dimensions of 512; thus, $d_{\text{non\_app}} = 512$ and $d_{\text{hybrid}} = 1024$. $\phi_{\text{ss}}(\cdot)$, $\phi_{\text{so}}(\cdot)$, $\phi_{\text{oo}}(\cdot)$ and $\psi^{\text{S}}(\cdot)$ defined in (2) are implemented by MLPs, each of which is a single linear layer with an output dimension of 1024. For comparison, the sequential model `M_Seq` used in (3) and (5) is implemented

[1] Following the standard evaluation protocol used in the previous works [4], [11], [49], [59], we first divide each video into T segments of equal durations. Then we randomly sample one frame from each segmentation for training and sample the middle frame from each segmentation for testing. Previous methods [11], [49], [59] for this task have used different numbers of frames (where more frames are usually better). In our experiments, we set T to 8 for all ablation studies, but also set T to different numbers for fair comparisons with the previous methods.

by an LSTM [56] or transformer [57], respectively, with two layers, and the dimension of the hidden states is 1024. T_{obs} used in prediction is set to $\frac{T}{4}$.

3) Training: We train the proposed approach with PyTorch [60] on two NVIDIA TITAN RTX graphics cards. Overall, we employ SGD with fixed hyper-parameters (0.9 momentum and 10^{-4} weight decay) to optimize all our models with an initial learning rate of 10^{-2}. The I3D model used in this work is pre-trained on Kinetics-400 [20]. For a fair comparison, we train our models with the batch size of 72 for the backbone of I3D in 50 epochs and reduce the learning rate to $1/10$ of the previous one at epoch 35 and 45 following [11] unless specified otherwise. For other different numbers of frames or backbones used in experiments, we adopt the largest batch size within the total memory of available GPUs. Some training details depend on the case and will be described in Sections IV-B and IV-C.

IV. Experiments

In this work, we evaluate our proposed approach on two datasets, *i.e.*, Something-Else [11] and IKEA-Assembly [15], with the compositional setting. Additionally, the few-shot setting and some diagnostic studies are applied to the experiments of Something-Else. The details are as follows.

A. Dataset and Metrics

1) Something-Else: Something-Else [11] is built on SSV2 [1] with the compositional setting forcing the combinations of action and objects cannot overlap between training and testing sets. Something-Else contains 174 categories of activities but only 112,795 videos (54,919 for training and 57,876 for testing) selected from original SSV2 [1] with the consideration of compositional setting. Specifically, the objects and actions from SSV2 are divided into two disjoint sets $\{\mathcal{S}_{\mathrm{A}}, \mathcal{S}_{\mathrm{B}}\}$ and $\{\mathcal{S}_1, \mathcal{S}_2\}$, respectively. With the novel setting, only the `action-object` combinations from the set $\{\mathcal{S}_1\mathcal{S}_{\mathrm{A}} + \mathcal{S}_2\mathcal{S}_{\mathrm{B}}\}$ are used to train the model, but the model is tested on $\{\mathcal{S}_1\mathcal{S}_{\mathrm{B}} + \mathcal{S}_2\mathcal{S}_{\mathrm{A}}\}$. Following [11], [59], we evaluate our approach on this dataset with the standard classification protocol and measure the top-1 and top-5 accuracy.

2) Ikea-Assembly: We further test the compositional generalization ability of our model on IKEA-Assembly [15], although it does not have the above compositional setting. IKEA-Assembly contains 16,764 video samples of furniture assemblies, each of which is annotated with one of 33 compositional action classes, *i.e.*, `verb-object` pairs. Each activity defined in IKEA-Assembly is also composed of an action and an object similar to Something-Else [11] but on a more granular scale. In total, there are 12 actions and 7 objects in the original version dataset. We split these samples in a compositional way in which the combination of action and objects do not overlap over the training and testing sets following [11], [59], leading to 6 compositional activities. More details can be found in [59]. Following [59], we measure our approach via both the mean of per class recall (macro) and micro averaged accuracy (micro) in experiments due to the serious issue of class imbalance.

TABLE I
Results of the Proposed Methods Built on Different Video Backbones. The Baseline Method Simply Average-Pools the Feature Maps Extracted From Visual Backbones Into Video Features. "# Fr" Denotes the Number of Frames Used in Models

Backbone	# fr	Method	Accuracy (%) top-1	Accuracy (%) top-5
I3D-ResNet50 [5], [6]	8	baseline	40.0	69.3
		Ours	**60.4**	**86.0**
	16	baseline	46.8	72.2
		Ours	**64.2**	**87.6**
TSM-ResNet50 [21]	8	baseline	48.4	76.8
		Ours	**62.3**	**87.3**
TEA-ResNet50 [23]	8	baseline	48.2	76.7
		Ours	**62.6**	**87.7**

B. Results on Something-Else

We conduct some ablation studies of our approach and compare it with the state-of-the-art methods on this dataset with the standard compositional setting mentioned above. Beyond that, the few-shot setting is introduced to test the proposed method, following [11].

1) Ablation Study: **Effect of Different Feature Extractors.** To verify that our proposed method is complementary to existing video representations, our approach applies different video backbones (e.g., TSM [21] and TEA [23]). TSM and TEA are initialized with weights pretrained on ImageNet [61] but I3D-ResNet50 is initialized with weights pretrained on kinetics-400 [20] (following [11]). The baseline in Table I removes all proposed modules from our approach and directly pools (*i.e.*, global-avg-pool) the feature map into a single vector for classification. As shown in Table I, our approach steadily improves the robustness of existing video representations on compositional generalization problems. I3D relies on dense frames, thus it shows low performance with only 8 frames compared with TSM and TEA. The better the video representation is, the more obvious the gain of our method is, indicating that our proposed framework is orthogonal to existing video feature encoders.

Effect of Different Feature Interactions. In this work, our IFI aims to introduce high-level information (*i.e.*, category of instance[2]) into relational reasoning among instance-centric representations in the latent space. To verify the effectiveness of such motivation, we compare IFI with several commonly used category-agnostic spatiotemporal interaction modules (i.e., NL, STRG, STIN, and Transformer), and the results are reported in Table II. Without considering the identity of instances, our IFI performs on par with existing methods and is even lower than STRG [6]. However, by introducing instance identity information, IFI is significantly higher than existing interaction modules and nearly 1% higher than the best method (*i.e.*, STRG) with detection. This shows the importance of identity for instance-centric feature interaction.

Effect of Different Temporal Fusions. In this work, temporal fusion is an important step for feature extraction and position

[2] We categorize instances that appeared in videos into only two classes, *i.e.*, hand or object.

TABLE II
Results of the Proposed Methods Built With Different Interaction Methods. NL* [16], STRG* [6], STIN* [11], and Transformer* [57] Are Built on the Proposed Basic Instance-Centric Joint representations (extracted by PAFE) for Fair Comparisons

Interaction Module	Identity	Accuracy (%) top-1	top-5
NL* [16]	✗	54.2	81.4
STRG* [6]		56.4	83.6
STIN* [11]		56.5	83.4
Transformer* [57]		55.7	82.9
IFI (Ours)	✗	57.3	84.2
	✓	**58.3**	**84.8**

(a) with Ground-truth

Interaction Module	Identity	Accuracy (%) top-1	top-5
NL* [16]	✗	44.2	71.3
STRG* [6]		45.9	73.2
STIN* [11]		44.1	71.2
Transformer* [57]		45.2	72.2
IFI (Ours)	✗	45.2	71.9
	✓	**46.8**	**73.9**

(b) with Detection

TABLE III
Results of the Proposed Methods Built With Different Temporal Fusion Methods

Temporal Fusion	Box Type GT	DET	Accuracy (%) top-1	top-5
		✓	47.5	74.4
LSTM [56]	✓		60.4	86.0
		✓	47.5	74.4
Transformer [57]	✓		60.6	86.6
		✓	47.8	75.0

TABLE IV
Effect of Different Position Predictions. "GT" and "DET" Are Short for Ground-Truth and Detection, Respectively

Box Type	Position Pred coord	offset	Pred Acc [3]	Accuracy (%) top-1	top-5
GT	✓		80.4	59.9	85.9
		✓	74.5	59.7	85.5
	✓	✓	**84.0**	**60.4**	**86.0**
DET	✓		74.9	48.6	76.1
		✓	69.0	49.5	76.5
	✓	✓	**79.0**	**50.4**	**76.9**

TABLE V
Ablation Study on Something-Else With the Compositional Setting. "PAFE," "IFI" and "SPP" Are Short for Position-Aware Appearance Feature Extraction, Identity-Aware Feature Interaction and Semantic-Aware Position Prediction, Respectively. "global fea." Represents the Appearance Features Extracted From the Whole Frames. "GT" and "DET" Are Short for Ground-Truth and Detection, Respectively

	Method	Accuracy (%) top-1	top-5
GT	Global fea. + STIN [11]	51.7	80.5
	B1: PAFE + STIN [11]	56.5	83.4
	B2: PAFE + IFI	58.3	84.8
	Ours: PAFE + IFI + SPP	**60.4**	**86.0**
DET	Global fea. + STIN [11]	42.7	69.7
	B1: + PAFE + STIN [11]	44.1	71.2
	B2: + PAFE + IFI	46.8	73.9
	Ours: + PAFE + IFI + SPP	**47.5**	**74.4**

prediction. Here, we explore the effect of different temporal modeling methods on the efficiency and performance of the proposed framework. We try two different temporal fusion methods, namely, LSTM [56] and Transformer [57]. Notably, in this work, LSTM has 2 layers, and Transformer has 2 layers with 8 heads. As shown in Table III, the performance of the two modules is almost the same. Furthermore, in our approach, there is little difference between these two modules in terms of efficiency with limited video frames and layers. Specifically, our approach with 2 layers of LSTM and Transformer requires 0.018 s and 0.017 s for the inference of one video with 16 frames, respectively. However, we recommend using the more advantageous Transformer in practical applications that may involve more frames. Because LSTM is difficult to optimize in parallel, its computational efficiency is significantly lower than that of Transformer when the sequence length is too large or the model has too many layers.

Effect of Different Position Predictions. To determine the effect of the prediction content of Semantic-aware Position Prediction, we also tried different forms of position prediction with ground-truth and detected boxes, such as "coord," "offset," and "coord+offset". As shown in Table IV, in the case of the ground truth, the difference in these three forms is not significant. However, when the bounding boxes (detected) are not inaccurate, the absolute position prediction alone is not conducive to improving the generalization of our framework, but additionally predicting the "offset" of these boxes helps to alleviate this phenomenon.

Effectiveness of SPP. As shown in the third column of Table IV, SPP with "coord" or "offset" prediction shows satisfactory position accuracy using "GT" or "DET" boxes, indicating its ability to perceive the motion tendency of instances to a certain extent. Combining both of them significantly enhances this ability.

We also visualize the prediction results of SPP with "coord + offset" using "GT" boxes. For example, as shown in Fig. 4, we sample 8 frames from each video as the input of our approach. Thus only the instances in the last 4 frames are predicted by SPP and the predicted position of objects or hands is near to the ground truth. These results demonstrate the ability of the proposed SPP to perceive instances' motion, which further improves the compositional generalization of our approach.

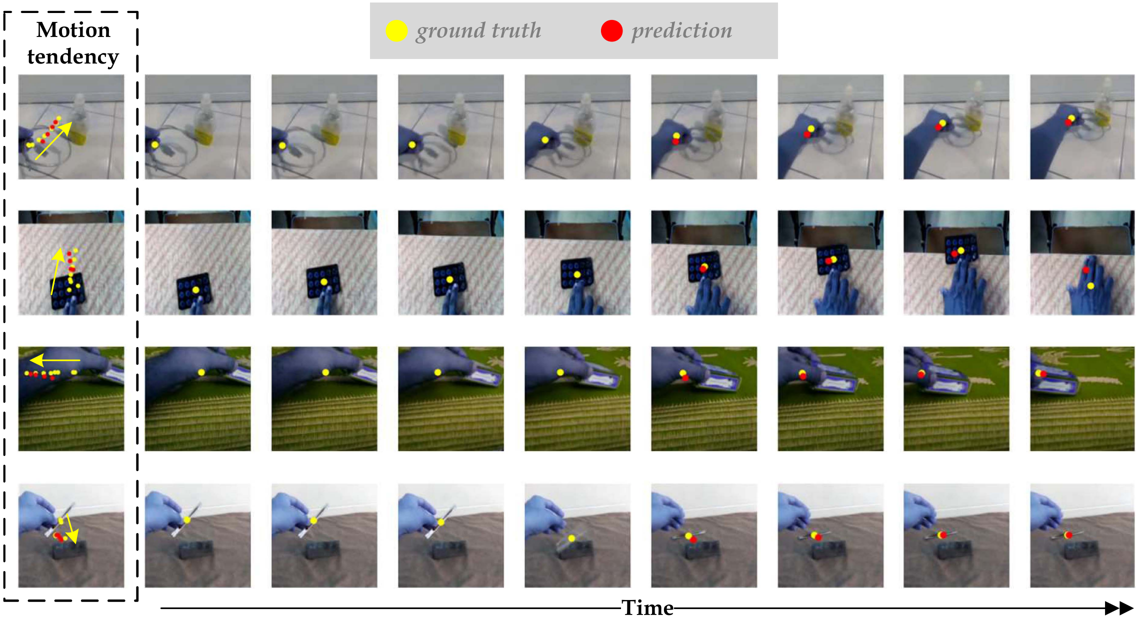


Fig. 4. Visualization results of Semantic-aware Position Prediction (SPP) on Something-Else [11]. Each row contains a sequence of frames uniformly sampled from the raw video. The first image in each row shows the overall motion tendency of the object. The colored points are the center coordinates of instances. For ease of viewing, only one instance's position information is drawn in each sample.

TABLE VI

Compositional Action Recognition on the Something-Else Dataset. For a Fair Comparison, the Numbers of Frames Fed Into the Video Backbone Are Specified. "ensemble" Represents Weighted Fusion of the Prediction Scores From Two Different Models Trained Separately. All Ensemble Methods Fuse the Scores From Different Models Via "naive Sum" Unless Stated Otherwise

Method	# frames	Accuracy (%)	
		top-1	top-5
I3D [5], [6]	8	40.0	69.3
STIN [11]		51.7	80.5
Ours		60.4	86.0
I3D [5], [6]	16	46.8	72.2
STIN [11]		54.6	79.4
Ours		64.2	87.6
Ensemble			
I3D, STIN [11]	16	58.1	83.2
MGAF (SlowFast [22]) [59]	8, 32	68.0	88.7
CDN [49]	16	62.8	87.2
CDN (Log-sigmoid Sum [62]) [49]	16	64.5	88.2
I3D, Ours	16	64.1	87.6
I3D, Ours (Log-sigmoid Sum)		66.1	88.2
TSM [21], Ours		68.5	**90.7**
TEA [23], Ours		**68.8**	90.3

(a) with Ground-truth

Method	# frames	Accuracy (%)	
		top-1	top-5
I3D [5], [6]	8	40.0	69.3
STIN [11]		42.7	69.7
Ours		47.6	74.5
I3D [5], [6]	16	46.8	72.2
STIN [11]		48.2	72.6
STRG [6], [11]		52.3	78.3
Ours		53.1	78.4
Ensemble			
I3D, STIN [11]	16	51.5	77.1
STRG, STIN [11]		56.2	81.3
I3D, Ours		**57.0**	**82.1**

(b) with Detection

Effectiveness of Each Component. We further evaluate each component of our approach using the ground-truth bounding boxes and detections. The results are reported in Table V. The state-of-the-art baseline is "Global fea. + STIN [11]," which builds relational representations among instance-centric hybrid features and concatenates them with the RGB appearance of the whole frames (as shown on the left of Fig. 3). Notably, all methods reported in Table V use the same setting and optimization strategy.

With ground-truth bounding boxes, B1 achieves significant gains in accuracy benefited by the instance-centric appearance feature extracted via PAFE with the help of instance position. Because PAFE can highlight the motion of instances and suppress noisy backgrounds, which is critical for compositional generalization. Compared with B1, our approach with PAFE + IFI (B2) brings further significant 1.8% and 1.4% improvements on top-1 and top-5, respectively, suggesting that identity information is helpful for semantic interactions among instance-centric

TABLE VII

Few-Shot Compositional Action Recognition accuracy (%) on the Something-Else Dataset. "# Fr" Denotes the Number of Frames Used in Models. "base," "5-S," and "10-S" Denote the Base Set, 5-Shot Set and 10-Shot Set, Respectively. "ensemble" Indicates That the Model is Combined With I3D in an Ensemble Way. Only top-1 Accuracy of Models is Reported Here Due to the Limit Space

Method	# fr	*base*	*few shot*	
			5-S	10-S
I3D [5], [6]	8	66.0	18.2	20.1
NL* [16]		76.6	25.3	30.9
STRG* [6]		77.9	26.6	32.4
STIN* [11]		76.4	24.6	29.4
Ours		79.5	30.7	36.2
STIN [11]	16	80.6	28.1	33.6
Ours		82.5	**33.2**	**39.4**
Ensemble				
I3D, STIN [11]	16	81.1	34.0	40.6
I3D, Ours		84.1	**35.3**	**41.7**

(a) with Ground-truths

Method	# fr	*base*	*few shot*	
			5-S	10-S
I3D [5], [6]	8	66.0	18.2	20.1
NL* [16]		69.4	18.3	22.4
STRG* [6]		71.0	20.2	25.3
STIN* [11]		69.3	18.8	23.7
Ours		72.2	20.7	27.2
STIN [11]	16	76.8	23.7	27.0
STRG [11]		75.4	24.8	29.9
Ours		75.4	**26.0**	**31.7**
Ensemble				
I3D, STIN [11]	16	76.1	27.3	32.6
STRG, STIN [11]		78.1	29.1	34.6
I3D, Ours		79.8	**31.6**	**37.5**

(b) with Detections

TABLE VIII

Compositional Action Recognition accuracy (%) on the IKEA-Assembly dataset [15]. Macro is Short for Macro-Recall That Averages the Recall of Each Class; Micro is Short for Micro-Accuracy Computes the Accuracy Globally Without Distinguishing Between Different Classes."ours" Denotes "PAFE + IFI + SPP"

Method	Mixed		Compositional	
	Macro	Micro	Macro	Micro
I3D [5], [6], [15]	51.6	76.0	30.0	44.0
STIN [11]	51.4	73.4	36.3	51.5
PAFE + STIN [11]	57.2	82.1	33.1	51.2
PAFE + IFI	59.6	81.8	35.8	54.3
PAFE + IFI + SPP	62.0	84.6	39.0	56.8
MGAF [59]	49.1	72.4	37.6	55.6
Ours	**62.0**	**84.6**	**39.0**	**56.8**

features. After that, our approach with PAFE + IFI + SPP (Ours) improves top-1 and top-5 again by 2.1% and 1.2% again by introducing auxiliary position prediction to facilitate the fusion of information from different sources.

With detections, each component shows similar improvements, although all variants drop sharply due to the instance-centric features from inexact tracklets. In particular, the improvement brought by SPP is indeed less obvious. To avoid the effect of randomness, we run our final model five times and report the averaged results, 47.5% and 74.4%. It also motivates future works to overcome the inaccuracy of extra modal information (used for enhancing the compositional generalization).

2) Comparisons With the State-of-The-Art Method: We compare our approach with recent works [11], [59] on compositional action recognition with ground-truth and detected boxes, and the results are reported in Table VI. Overall, our approach establishes a new state-of-the-art methods in terms of ground truth and detection.

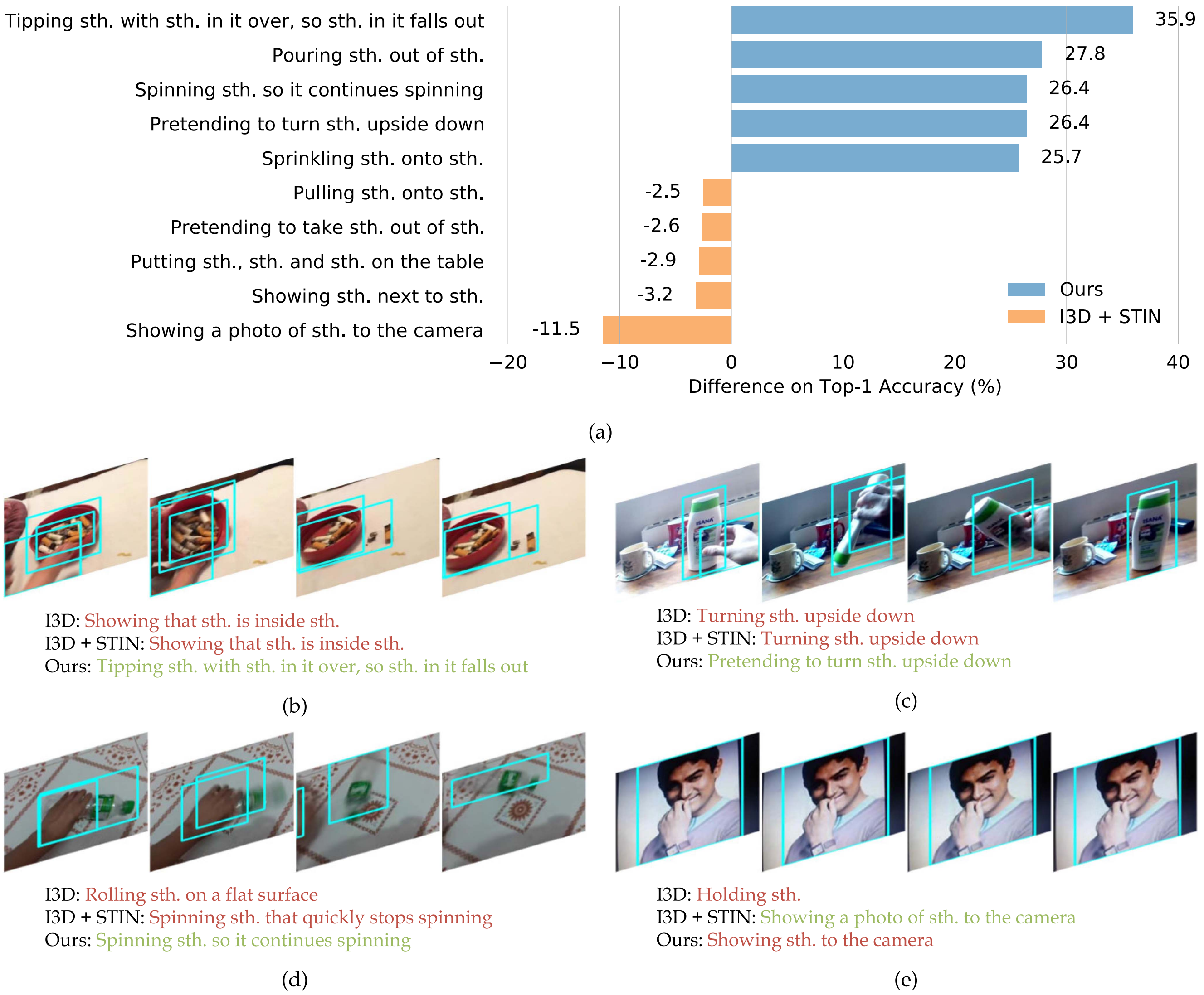
Ground Truth. We first compare our approach with recent state-of-the-art methods using ground-truth boxes based on different frames and training methods, as reported in Table VI a. With an end-to-end training fashion, our approach (*Ours*) easily outperforms the base method I3D [5], [6] and the most related STIN [11], based on whether 8 or 16 frames. In an ensemble way, *I3D, Ours.* still gains 6.0% and 4.4% compared with the baseline method STIN [11] and exceeds the existing best results from [59] with a more powerful video backbone, i.e., TSM [21]. Limited by the available computing resource, we do not use SLOWFAST [22] to extract features from 32 frames similar to [59].

Moreover, CDN [49] is an ensemble method that fuses scores from different pretrained models via two different strategies (i.e., naive sum and log-sigmoid sum [62]). Therefore, it is fairer to compare CDN and our ensemble model, both of which apply the late fusion of predicted scores. Whether using naive sum or log-sigmoid sum, our ensemble model easily exceeds CDN. Beyond that, our end-to-end model can also significantly outperform CDN [49] (Naive Sum).

Detection. Using inaccurate detections, our approach still achieves state-of-the-art results with both end-to-end training and ensemble fashion. Similarly, *Ours* significantly outperforms I3D and STIN with whether 8 or 16 frames. Moreover, *Ours* with 16 frames is superior to the state-of-the-art method STRG [6] by 0.8% in terms of top-1 accuracy. We combine our approach with I3D in an ensemble way, and it becomes the new state-of-the-art result of this benchmark without ground truth.

3) Few-Shot Setting: To evaluate the generalization capability of our approach, we conduct experiments on the few-shot setting proposed in [11].

Problem Formulation. In the above setting, the authors [11] randomly divide the original 174 classes into the base set with 88 categories and the novel set with 86 categories for few-shot compositional action recognition. Models are pretrained on all training samples (total 112,397 videos) from the base set and then finetuned on few-shot samples from the novel set to recognize the rest of the samples. Following [11], we directly fine-tune models on all categories from the novel set instead of the traditional "n-way, k-shot" in few-shot learning [63].



I3D: Showing that sth. is inside sth.
I3D + STIN: Showing that sth. is inside sth.
Ours: Tipping sth. with sth. in it over, so sth. in it falls out

(b)

I3D: Turning sth. upside down
I3D + STIN: Turning sth. upside down
Ours: Pretending to turn sth. upside down

(c)

I3D: Rolling sth. on a flat surface
I3D + STIN: Spinning sth. that quickly stops spinning
Ours: Spinning sth. so it continues spinning

(d)

I3D: Holding sth.
I3D + STIN: Showing a photo of sth. to the camera
Ours: Showing sth. to the camera

(e)

Fig. 5. (a): Top-5 categories that our approach exceeds (blue bar) or lags behind (orange bar) I3D + STIN [11]. The numbers represent the difference between the two models in terms of top-1 accuracy. (b-e): Predictions of I3D [5], [6], I3D + STIN [11], and our approach on some examples from the Something-Else dataset. All instances are annotated by cyan boxes, and the correct and incorrect predictions are highlighted in green and red, respectively. Best viewed in color.

For example, 86 × 10 training examples are used for 10-shot fine-tuning. All the models are trained in 50 epochs with a fixed learning rate of 0.01. We froze the parameters of all layers in the network except the last classifier at the stage of fine-tuning. More details can be found in [11]. In this work, only 5-shot and 10-shot results are reported in Table VII.

Results. As shown in Table VII a, we first report the results based on ground-truth bounding boxes. With only 8 frames, our approach (*Ours*) is clearly superior to previous reasoning methods (NL*, STRG*, and STIN*) that plugged in our instance-based framework and the frame-based method I3D. Using 16 frames, *Ours* outperforms STIN [11] by 5.1% and 5.8%, demonstrating the advantage of the proposed progressive fusion compared with the simple concatenation used in [11]. Finally, our approach can be combined with any existing frame-level video backbone, such as I3D, in an ensemble way and then achieve the best results.

Surprisingly, our approach works well even with detected boxes, suggesting that our model does not rely on strong position supervisions, as reported in Table VII b. In particular, *Ours* achieves 2.5% and 7.1% gains on 5-shot and 10-shot settings, respectively, compared with I3D with 8 frames. With 16 frames, our approach still shows a clear advantage, and "*I3D, Ours.*" outperforms the existing state-of-the-art results from [6] by a remarkable margin. Overall, our approach maintains a strong generalization on the few-shot setting against other methods.

C. *Results on IKEA-Assembly*

The original IKEA-Assembly randomly splits video samples from 33 actions into train/test sets that are joined in the term of verb-noun combination, lacking the ability to evaluate the compositional generalization of methods. For instance, `flip table` will appear in both the training and testing phases. To this end, [59] introduces the compositional task on this dataset by forcing the same verb into combining with different objects between train and test splits. Thus, it becomes a classification on 6 compositional actions, i.e., `align sth.`, `attach sth.`, `flip sth.`, `lay down sth.`, `pick up sth.`, and `push sth.`. More details are described in the supplementary material of [59]. Notably, we use the tracklets released by [15] rather than the ground truth and train all models with 16 frames. The results on both original and compositional settings are reported in Table VIII to demonstrate the effectiveness of the proposed methods.

With the novel compositional setting, STIN [11] shows poor results here due to the rough fusion of appearance and motion on this benchmark, compared with I3D. "PAFE + STIN" surpasses the base method I3D on the metrics of Macro and Micro by extracting instance-centric representations from RGB and position input. As expected, "PAFE + IFI" builds identity-aware semantic interactions among these representations that further boosts our model ("PAPE+STIN"). "PAFE + IFI + SPP" brings significant improvements on this benchmark thanks to the relatively complete and clear motion trajectory of instances (in the third view). Overall, our approach significantly outperforms the state-of-the-art method MGAF [59] on the compositional IKEA-Assembly and shows strong generalization on this benchmark again by aggregating multiple cues progressively.

Moreover, each component of our proposed approach shows similar improvement on the original ("Mixed") split of this dataset. We find that it is easy for the above models to overfit the mixed split with only 8 frames. Therefore, we conduct all our experiments with 16 frames except MGAF. MGAF, including slow (8 frames) and fast (32 frames) paths, shows poor results due to the poor representations from the slow path. In addition, when we train STIN [11], the additional non-local module is dropped to prevent overfitting.

D. Diagnostic Studies

In this section, we show some results to diagnose our model.

1) Category Analysis: There are a total of 174 actions defined in Something-Else, and different models excel in different categories. Overall, our approach surpasses STIN [11] on $\mathbf{84.5\%}$ categories and the top categories are shown in Fig. 5(a). In particular, our approach is far superior [11] on some categories, such as "pouring sth. out of sth.," "spinning sth. so it continues spinning," etc., which involve not only relative motions but also object properties. Understandably, our method is worse than STIN [11] on "showing a photo of sth. to the camera" (in Fig. 5(e)) in which there is only a static photo in the video. Our model tends to capture significant dynamic cues of objects that will not appear in this "static" action.

2) Visualization: We also visualize some predictions on Something-Else in Fig. 5. We can see that STIN [11] easily learns the bad appearance or position bias due to the rough concatenation of different features. For instance, in Fig. 5 (b), STIN [11] prefers to capture the concept of "sth. is inside sth." when seeing a container containing certain objects similar to the I3D model but ignores the negligible yet very important position self-change of objects. In addition, STIN [11] ignores the concept of "pretend" in Fig. 5 (c) due to the worse bias from position features. However, our approach generates predictions by fusing different levels of features progressively, rather than preferring one of them.

V. Conclusion

Integrating multiple information that often differ significantly in modality and dimensionality, is the promising solution to compositional action recognition. Previous methods encode instance information into non-appearance feature independently and fuse it with appearance feature for classification, which ignores the importance of instance information in the process of video feature learning. To this end, we propose a novel framework to inject instance information into the video feature learning progressively. Our framework is composed of three steps, namely, position-aware appearance feature extraction, identity-aware feature interaction, and semantic-aware position prediction. Experimental results on two action recognition datasets show the robust generalization ability of our framework on compositional action recognition. This work demonstrates that the understanding of human action is actually a process of gradually fusing features from multiple sources with diverse strategies. We hope that this work can inspire future research on the compositional generalization of activity understanding.

References

[1] R. Goyal et al., "The "something something" video database for learning and evaluating visual common sense," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 5842–5850.

[2] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[3] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2625–2634.

[4] L. Wang et al., "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.

[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[6] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 399–417.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[8] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 203–213.

[9] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2556–2563.

[11] J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang, and T. Darrell, "Something-else: Compositional action recognition with spatial-temporal interaction networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1049–1059.

[12] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action genome: Actions as compositions of spatio-temporal scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10 236–10 247.

[13] E. S. Spelke and K. D. Kinzler, "Core knowledge," *Devlop. Sci.*, vol. 10, no. 1, pp. 89–96, 2007.

[14] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1521–1528.

[15] Y. Ben-Shabat et al., "The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose," in *Proc. Winter Conf. Appl. Comput. Vis.*, 2021, pp. 847–859.
[16] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[17] R. Yan, L. Xie, J. Tang, X. Shu, and Q. Tian, "HiGCIN: Hierarchical graph-based cross inference network for group activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 27, 2020, doi: 10.1109/TPAMI.2020.3034233.
[18] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 510–526.
[19] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 961–970.
[20] W. Kay et al., "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.
[21] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 7083–7093.
[22] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6202–6211.
[23] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: Temporal excitation and aggregation for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 909–918.
[24] R. Girdhar and D. Ramanan, "CATER: A diagnostic dataset for compositional actions and temporal reasoning," in *Proc. Int. Conf. Learn. Representations*, 2019.
[25] B. Jia, Y. Chen, S. Huang, Y. Zhu, and S.-C. Zhu, "Lemma: A multi-view dataset for learning multi-agent multi-task activities," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 767–786.
[26] D. Keysers et al., "Measuring compositional generalization: A comprehensive method on realistic data," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–13.
[27] B. Lake and M. Baroni, "Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2873–2882.
[28] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Intra-and inter-action understanding via temporal action parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 730–739.
[29] D. Shao, Y. Zhao, B. Dai, and D. Lin, "FineGym: A hierarchical video dataset for fine-grained action understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2616–2625.
[30] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori, "Object level visual reasoning in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 105–121.
[31] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
[32] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
[33] C.-Y. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, and H. P. Graf, "Attend and interact: Higher-order object interactions for video understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6790–6800.
[34] C. Sun, A. Shrivastava, C. Vondrick, K. Murphy, R. Sukthankar, and C. Schmid, "Actor-centric relation network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 318–334.
[35] H. Qi, X. Wang, D. Pathak, Y. Ma, and J. Malik, "Learning long-term visual dynamics with region proposal interaction networks," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–13.
[36] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, "Video (language) modeling: A baseline for generative models of natural videos," 2014, *arXiv:1412.6604*.
[37] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 843–852.
[38] E. L. Denton et al., "Unsupervised learning of disentangled representations from video," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4414–4423.
[39] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–11.
[40] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, "Action-conditional video prediction using deep networks in atari games," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2863–2871.
[41] J.-T. Hsieh, B. Liu, D.-A. Huang, L. F. Fei-Fei, and J. C. Niebles, "Learning to decompose and disentangle representations for video prediction," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2018, pp. 515–524.
[42] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 201–214.
[43] D.-A. Huang, A.-M. Farahmand, K. M. Kitani, and J. A. Bagnell, "Approximate maxent inverse optimal control and its application for mental simulation of human interactions," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2673–2679.
[44] T. Ye, X. Wang, J. Davidson, and A. Gupta, "Interpretable intuitive physics model," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 87–102.
[45] J. Wu, J. J. Lim, H. Zhang, J. B. Tenenbaum, and W. T. Freeman, "Physics 101: Learning physical object properties from unlabeled videos," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1483–1492.
[46] T. Han, W. Xie, and A. Zisserman, "Video representation learning by dense predictive coding," in *Int. Conf. Comput. Vis. (Workshop)*, 2019.
[47] T. Han, W. Xie, and A. Zisserman, "Memory-augmented dense predictive coding for video representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 312–329.
[48] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[49] P. Sun, B. Wu, X. Li, W. Li, L. Duan, and C. Gan, "Counterfactual debiasing inference for compositional action recognition," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 3220–3228.
[50] R. Girshick, "Fast R-CNN," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
[51] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 244–253.
[52] A. Santoro et al., "A simple neural network module for relational reasoning," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4967–4976.
[53] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3588–3597.
[54] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 803–818.
[55] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989.
[56] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[57] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[58] Y. Zhang et al., "Exploiting motion information from unlabeled videos for static image action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12 918–12 925.
[59] T. S. Kim, J. Jones, and G. D. Hager, "Motion guided attention fusion to recognize interactions from videos," 2021, *arXiv:2104.00646*.
[60] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[62] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, "Counterfactual VQA: A cause-effect look at language bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12 700–12 710.
[63] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

Rui Yan received the PhD degree from Intelligent Media Analysis Group (IMAG), Nanjing University of Science and Technology, China. He is currently an assistant researcher with the Department of Computer Science and Technology, Nanjing University, China. He was a visiting researcher with the National University of Singapore (NUS) from 2021 to 2022. He was a research intern with HUAWEI NOAH's ARK LAB from 2018 to 2019. His research interests include mainly focuses on complex human behavior understanding and video-language understanding. He has authored more than 20 journal and conference papers in these areas.

Lingxi Xie received the BE and PhD degrees from the Department of Computer Science and Technology, Tsinghua University, in 2010 and 2015, respectively. He is currently a senior researcher with Huawei Cloud. He was a visiting researcher with the Department of Computer Science, University of Texas at San Antonio (UTSA) from February to July 2014. He was a postdoctoral researcher with the Department of Statistics, University of California, Los Angeles, from 2015 to 2016, and at the Center for Imaging Science, the Johns Hopkins University, from 2016 to 2019. His research interests include computer vision, deep learning, and medical image analysis.

Xiangbo Shu (Senior Member, IEEE) received the PhD degree from the Nanjing University of Science and Technology, in 2016. He is currently a professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. From 2014 to 2015, he worked as a visiting scholar with the National University of Singapore, Singapore. His current research interests include Computer Vision and Multimedia. He has authored more than 80 journal and conference papers in these areas. He has received the Best Student Paper Award in MMM 2016, and the Best Paper Runner-up in ACM MM 2015. He has served as an editorial board of *IEEE Transactions on Circuits and Systems for Video Technology*. He is also the Member of ACM, the senior member of CCF.

Liyan Zhang received the PhD degree in computer science from the University of California, Irvine, CA, USA, in 2014. She is currently a professor with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Her research interests include multimedia analysis, computer vision, and deep learning. She received the Best Paper Awards from the International Conference on Multimedia Retrieval (ICMR) 2013 and the ACM Conference on Multimedia Asia (MMAsia) 2020, and the Best Student Paper Award from the International Conference on Multimedia Modeling (MMM) 2016.

Jinhui Tang (Senior Member, IEEE) received the BE and PhD degrees from the University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a professor with the Nanjing University of Science and Technology, Nanjing, China. He has authored more than 200 articles in toptier journals and conferences. His research interests include multimedia analysis and computer vision. He was a recipient of the Best Paper Awards in ACM MM 2007 and ACM MM Asia 2020, the Best Paper Runner-Up in ACM MM 2015. He has served as an Associate Editor for *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Multimedia*, and *IEEE Transactions on Circuits and Systems for Video Technology*. He is a fellow of IAPR.