Learning Label Semantics for Weakly Supervised Group Activity Recognition

Lifang Wu, *Member, IEEE*, Meng Tian, Ye Xiang, Ke Gu, *Senior Member, IEEE*, and Ge Shi

***Abstract*—Weakly supervised group activity recognition deals with the dependence on individual-level annotations during understanding scenes involving multiple individuals, which is a challenging task. Existing methods either take the trained detectors to extract individual features or utilize the attention mechanisms for partial context encoding, followed by integration to form the final group-level representations. However, the detectors require individual-level annotations during the training phase and have a mis-detection issue, and the partial contexts extracted immediately from the whole complex scene are too ambiguous without the guidance of concrete semantics. In this article, we investigate the hierarchical structure inherent in group-level labels to extract the fine-grained semantics without using detectors for weakly supervised group activity recognition. A multi-hot encoding strategy combined with a semantic encoder is first adopted to get the label semantics embeddings. The semantic and visual scene information are then fused through a semantic decoder to obtain activity-specific features. Lastly, we employ the multi-label classification and integrate the scores of hierarchical activity labels. Experimental results show that our proposed method achieves the state-of-the-art performance on three benchmarks, and the accuracy on the Volleyball dataset exceeds the second-best method by 2%.**

***Index Terms*—Weakly Supervised Group Activity Recognition, Label Semantics, Multi-Label Classification.**

I. Introduction

GROUP activity recognition [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11] aims to infer the type of group activity of participants based on important group cues in a scene. It has a wide range of promising applications in intelligent video surveillance, social public safety and sports video analysis. Its potential applications and economic value have made group activity recognition gradually become a research hotspot in the field of computer vision.

The core of group activity recognition lies in understanding contextual interaction among individuals within a scene. Existing methods [4], [12], [13], [14], [15], [16], [17] explore the spatio-temporal interactions between actors with the help of individual labels and bounding box annotations. Specifically, individual features are extracted by using off-the-shelf tools such as RoI-Align [18], their spatio-temporal relationships are then modelled by transformer or graph convolutional networks, generating the actor-level features that are subsequently aggregated to form the group-level representations. Although these works have achieved remarkable results, their implementations rely on individual labels and bounding box annotations, which are therefore labour intensive and unrealistic for inference in practical applications, severely limiting their applicability.

To address the above issues, a few works have been dedicated to the study of weakly supervised group activity recognition. Some researchers [12], [19], [20] proposed to make the actor detection and group activity reasoning collaborate in a unified framework by sharing convolutional layers between them, thus avoiding the use of bounding box annotations during the inference phase. Furthermore, Yan et al. [21] adopted the detector trained on an external dataset combined with pruning irrelevant suggestions, eliminating the need for actor-level annotations during both training and inference phases. However, these detector based methods easily suffer from mis-detection issue, leading to a reduction in recognition accuracy. Recently, a detector-free method [22] was proposed to utilize attention mechanism in prevailing Transformer model to locate and encode partial context, which achieves impressive performance without requiring any actor-level annotations. The shortcomings of this algorithm are reflected in two aspects: 1) adjusting to obtain the optimal number of tokens can be time-consuming; 2) extracting partial context directly from a whole complex scene is difficult due to the lack of explicit semantic information.

In fact, the scene patterns of the video sequences are directly related to their activity labels and there is no need to extract the scene elements separately and explore their combined relationships. As shown in Fig. 1. The group activities $2p\text{-}fail\text{-}off$ and $2p\text{-}fail\text{-}def$ show the same scene pattern in the first half. The second half of the field has visually distinct differences between the offense and the defense, and thus should be represented differently in the feature space. Intuitively, the different group activities should have different semantic representations and correspond directly to their specific scene features.

In this paper, inspired by [23] we propose a new weakly supervised group activity recognition model, which generates activity-specific features for different group activities through the guidance of labeled semantic features and solves the problem of group activity recognition from a multi-label

Manuscript received 9 May 2023; revised 8 November 2023; accepted 24 December 2023. Date of publication 4 January 2024; date of current version 10 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 62106011, 62236010, 61976010, 62322302, 62273011, 62021003, and 62076013, and in part by Beijing Natural Science Foundation under Grant JQ21014. The Associate Editor coordinating the review of this manuscript and approving it for publication was Mr. Y. Fang. *(Corresponding author: Ye Xiang.)*

The authors are with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: lfwu@bjut.edu.cn; tianmeng@emails.bjut.edu.cn; xiangye@bjut.edu.cn; guke@bjut.edu.cn; tinkersxy@gmail.com).

Digital Object Identifier 10.1109/TMM.2024.3349923

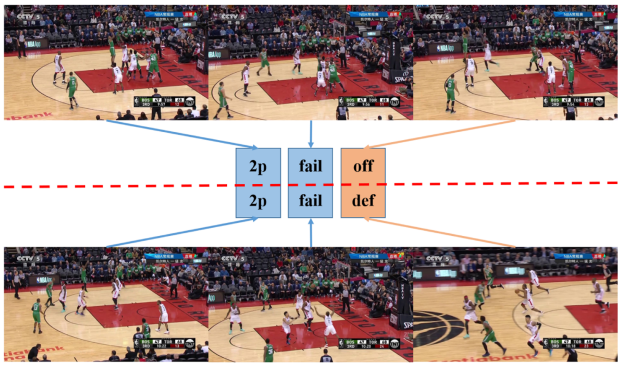

Fig. 1. Comparison of the similarities and differences between the different group activities in the NBA dataset in terms of scenarios and labels. It can be observed that the group activities $2p$-$fail$-off and $2p$-$fail$-def have the same sub-labels $2p$ and $fail$ in the first half of the game, the difference being whether the second half performs offensively or defensively. The variation in group sub-labels is a direct reflection of scene differences, and the different label semantics are directly related to scene information. Knowing where the specific differences between group activities are manifested helps in the classification task.

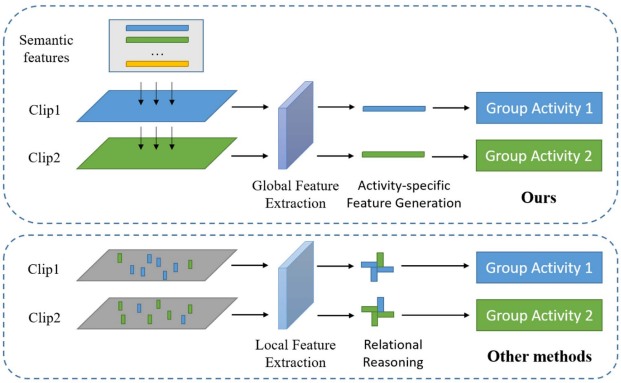

Fig. 2. Comparison of our method with other methods.

classification perspective for the first time. Different from most previous approaches, which generate shared feature vectors by extracting local features within the scene and represent group activities through the interactive combination of multiple feature vectors, our network generates specific feature representations for each activity based on semantics. As shown in Fig. 2. Specifically, we assign multiple labels to group activities by analysing the intrinsic relationships between group labels. Given a set of group labels, we use multi-hot encoding to generate a semantic representation relative to the group activity based on the predetermined correspondence between the labels. The semantic encoder we design encodes the semantic representation as a semantic token specific to the group tokens. The semantic decoder is designed to decode information from scene features that are semantically relevant for labeling purposes. Specifically, we use two structurally symmetric transformers to do the work of backbone network output feature enhancement and activity-specific feature extraction respectively, as the same pattern facilitates the matching of semantic features with visual features. A regular feature aggregation module outputs the group representation by aggregating activity-specific features. Finally we use a multi-label classifier to complete the outcome prediction and group activity transformation. Experimental results show that the proposed network outperforms state-of-the-art methods on two benchmarks on the widely adopted volleyball datasets [15], NBA datasets [21] and collective activity datasets [24]. The contributions of this paper are as follows:

- We embed the specific label semantics into group activity recognition framework to extract corresponding fine-grained information, facilitating the representation learning under weak supervision.
- For the first time, we address group activity recognition with multi-label classification, achieving the full utilization of label information under weak supervision setting.
- We conduct experiments on three benchmarks and obtain the state-of-the-art performance. The accuracy on Volleyball dataset is even 2% higher than the second best method.

II. Related Work

A. Group Activity Recognition

Distinct from traditional action recognition, group activity recognition focuses on understanding the activity of multiple individuals in a scene. Most of the existing algorithms are dedicated to mining contextual information for group activity recognition. Earlier work used graph structure models [2], [3], [25] to construct relationships between different actions based on hand-crafted features. However, these methods require a large amount of a priori knowledge and have the disadvantages of poor performance and insufficient generalization of the model, so are limited in their specific application.

With the increasing development of deep learning, group activity recognition algorithms based on deep learning have become the mainstream of research. [4], [5], [12], [13], [14], [15], [16], [17], [26] achieved satisfactory results for hierarchical temporal modelling of group context based on RNN. Ibrahim et al. [15] proposed a combined CNN and LSTM framework for activity recognition, first extracting the appearance features of individual players through CNN, and then realising the modelling of temporal sequences with the help of LSTM, by achieving the encoding of individual players in spatial and temporal features, and then obtaining the representation of group activity. Another attempt was made to model the spatio-temporal relationships between participants based on a graphical approach [6], [7], [8], [27], [28]. [6] building relationship graphs based on the appearance characteristics of individuals as nodes. Interpersonal interactions are modelled by stacking multiple person-relationship graphs, using graph convolution to achieve feature update augmentation, which in turn completes the group activity recognition task. Recognition is further enhanced by a method based on transformers [9], [10], [29], [30], [31] for relational inference on individual features. [10] extract individual features and group feature representations from feature maps at different scales, and then optimise individual features by clustering spatio-temporal Transformer joint spatio-temporally to finalise the group recognition task. The reliance of these methods on individual annotation strictly limits their applicability.

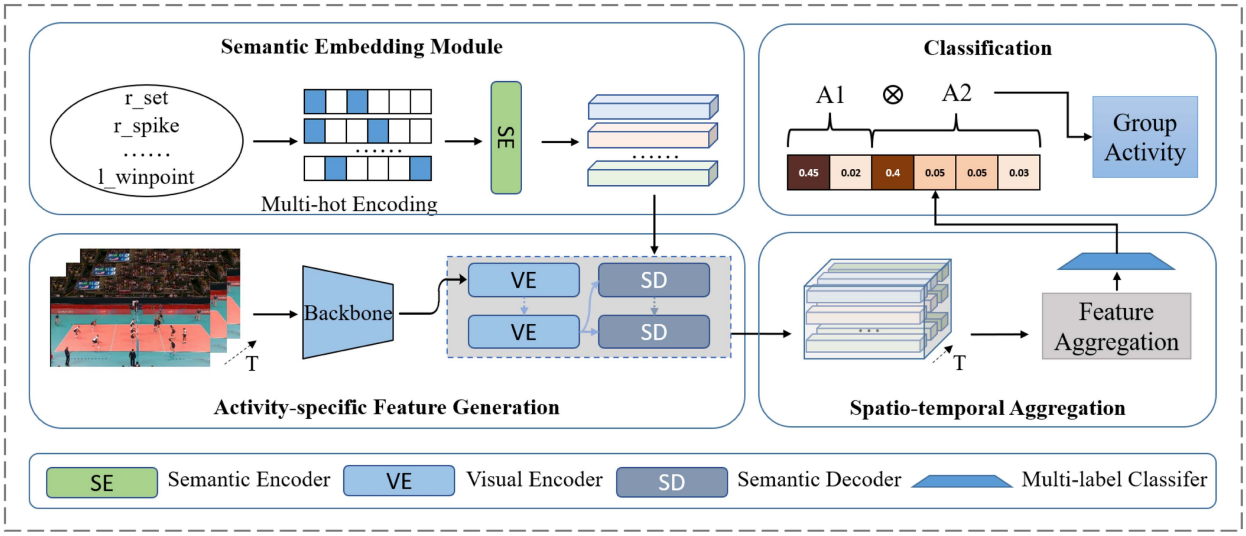

Fig. 3. Framework of our proposed algorithm. It consists of four parts, label semantics embedding, activity-specific feature extraction, spatio-temporal aggregation and classification. Firstly, the semantic encoder of the label semantic embedding module encodes multi-label embeddings to generate semantic representations of group activities. A set of visual encoders are used to enhance the scene features extracted from the backbone network. The enhanced features are then fed into the semantic decoder together with the semantic features to complete the extraction of activity-specific features. After the activity-specific features are obtained, the feature aggregation module completes the refinement of the features, a process that includes both spatio-temporal aggregation. It is worth noting that in the final classification stage, we use a multi-label classifier to obtain the classification results and then complete the transformation to group activity labels based on the sub-label correspondence.

As research progressed, some algorithms [20], [21], [22], [23] began to try to escape the constraints of individual labels and explore group activity recognition algorithms in a weakly supervised setting. wu [20] et al. obtained masks to locate the spatial location of scene activities and eliminate background information through an attention, and used them as visual markers to construct spatio-temporal relationships at different scales. However, he still needed the assistance of individual annotation in the training phase. [23] extracts activity-specific features based on an attention mechanism, and then generates co-occurrence matrices to augment the features based on label co-occurrence probabilities to finalise the task of behaviour recognition. However, it is clear from the title that it is primarily designed for multi-labelled behaviour recognition, and although experiments are set up for group behaviour recognition, the performance difference from the baseline in the WSGAR setting, i.e. after removing individual labels, is small. Detector-Free [22] The Transformer-based model encodes part of the context of the group activity as a set of visual embeddings, eventually aggregating the embeddings and capturing the temporal evolution of the embedding vectors, thus bypassing explicit target detection, with surprising results. However, the lack of semantic guidance and the lack of clear physical meaning of the individual embedding vectors may lead to an inadequate exploration of group context. To fully explore the population context, we introduce semantics to guide the network recognition process.

B. Multi-Label Image Classification

Multi-label image recognition [32], [33], [34], [35], [36], [37] is a fundamental task in the field of computer vision and plays an important role in applications such as human attribute recognition, recommender systems and medical image recognition. Unlike single-label classification, multi-label image recognition requires the assignment of multiple labels to a single image. Current multi-label image recognition [38], [39], [40], [41], [42], [43], [44] is mostly carried out based on CNN approaches, either by roughly localising multiple regions and then using neural networks to identify each region [39], [40], [41], [42], or by using label correlations or region relationships [38], [43], [44], [45], [46], [47] to guide learning.

III. Proposed Method

A. Overview

We aim to explore the semantic information of hierarchical structure inherent in the fine-grained activity labels for weakly supervised group activity recognition. The overall architecture is illustrated in Fig. 3, which contains four modules: label semantics embedding, activity-specific feature extraction, spatio-temporal aggregation and classification. Firstly, we adopt a multi-hot encoding strategy combined with a semantic encoder to get the semantic embeddings for group labels. The semantic information is then fused with the visual information extracted from RGB frames by using a semantic decoder, obtaining the activity-specific features. Different activity-specific features of different frames are aggregated across spatial and temporal domains. Lastly, a multi-label classifier is used to predict the group activity labels.

B. Label Semantic Embedding

Multi-hot codes are often used for recommender systems [48] and click-through rate prediction [49], [50], since its ability to represent the multi-valued discrete features. For group activity

TABLE I
FEATURE DOMAINS AND MULTI-HOT CODING FOR SUB-LABEL OF VOLLEYBALL DATASETS

Activity labels	right	left	spike	set	pass	win-point
r-set	1	0	0	1	0	0
r-spike	1	0	1	0	0	0
r-pass	1	0	0	0	1	0
r-winpoint	1	0	0	0	0	1
l-set	0	1	0	1	0	0
l-spike	0	1	1	0	0	0
l-pass	0	1	0	0	1	0
l-winpoint	0	1	0	0	0	1

TABLE II
FEATURE DOMAINS AND MULTI-HOT CODING FOR SUB-LABEL OF NBA DATASETS

Activity labels	2p	3p	layup	w/o layup	succ.	def.	off.
2p-succ.	1	0	0	1	1	0	0
2p-fail.-off.	1	0	0	1	0	0	1
2p-fail.-def.	1	0	0	1	0	1	0
2p-layup-succ.	1	0	1	0	1	0	0
2p-layup-fail.-off.	1	0	1	0	0	0	1
2p-layup-fail.-def.	1	0	1	0	0	1	0
3p-succ.	0	1	0	1	1	0	0
3p-fail.-off.	0	1	0	1	0	0	1
3p-fail.-def.	0	1	0	1	0	1	0

recognition, we note that many labels are formed by rigidly combining the words describing different aspects of information, for example, *left-spike* and *2p-succ*. Inspired by this, we propose to use the multi-hot codes to initially encode group activity labels before generating semantic embeddings. The main benefits of multi-hot encoding strategy include: 1) It explicitly models both commonalities and differences between group activity labels, thereby helping to learn clearer and hierarchical semantic information; 2) Compared with common one-hot encoding strategy, the distances (*e.g.*, Hamming distance) between the codes of labels appear to become larger, similarly, the separation between representations incorporating the semantics across different labels can also become clearer, as verified by experiments in Section IV-E.

We divide the group labels into multiple sub-labels, according to different feature domains. The division process needs to be done only once for the entire dataset, and will not incur additional training samples labeling work. Before dividing group labels, we observe the datasets and identify the appropriate feature domains. The sub-labels within the same feature domain should be mutually exclusive, and we expect it to be convenient for the subsequent label transformation. For example, in the volleyball dataset, right and left belong to one same feature domain, and the remaining four labels belong to another feature domain. In the NBA dataset, offensive and defensive only occur after a failed shot and are mutually exclusive, thus belonging to the same domain. As a result, we divide the volleyball dataset into 2 feature domains containing 6 sub-labels of right, left, spike, set, pass, and win-point, and divide the NBA dataset into 3 feature domains containing 7 sub-feature labels of 2p, 3p, layup, w/o layup, succ., def., and off. The specific division is shown in Tables I and II.

After obtaining the multi-hot codes for group activity labels, we can further encode them into the semantic embeddings. Let $Z \in \mathbb{R}^{K\times L}$ denote the multi-hot codes for group labels, where K is the number of group activity labels and L is the number of sub-labels. We first adopt a linear layer to get the sub-label embeddings of dimension D. The embeddings for group labels can then be represented as $\mathbf{Z_1} \in \mathbb{R}^{K\times L\times D}$. To update $\mathbf{Z_1}$, we apply a semantic encoder based on self-attention mechanism, as illustrated in Fig. 4(a), to capture the interactions between sub-labels. Afterwards, we concatenate the L sub-label embeddings for each group label, and feed them into a fully connected layer followed by a LayerNorm layer to reduce the dimension to D. The obtained $\tilde{\mathbf{Z}} \in \mathbb{R}^{K\times D}$ is just the final semantic embeddings of group labels.

C. Activity-Specific Feature Extraction

We use ResNet-18 as the backbone network to extract visual features. A visual encoder and a semantic decoder, taking the output of the backbone network and the semantic embedding module as input, are applied to encode the visual features and fuse the semantic and visual information.

In the case of sports videos, different elements within a scene are highly interrelated. For example, the actor behaviour, actor position and ball position interact with each other, and are all directly related to the group activity. Modelling the relationships of different elements is hence necessary. We employ a visual encoder based on the self-attention mechanism to capture the relationships between elements within the scene. A semantic decoder based on the cross-attention mechanism is used to generate the activity-specific features, by combining the semantic and visual information. The visual encoder and the semantic decoder work in conjunction with each other, as shown in Fig. 4(b).

For the visual encoder, we use a transformer [51] composed of a multi-head self-attention layer and a feed-forward network (FFN), whose implementation follows DETR [52]. Specifically, given the features $\mathbf{F} \in \mathbb{R}^{T\times H\times W\times C}$ output by the backbone for T frames, we first reduce the channel dimension to D by 1×1 convolution. The spatial correlation enhancement is then performed by using self-attention mechanism, generating the features $\mathbf{F}'$. For the semantic decoder, we use a similar structure to extract the activity-specific features. The difference is that the query(Q) used for cross-attention of the semantic decoder is generated from the semantic embeddings $\tilde{\mathbf{Z}} \in \mathbb{R}^{K\times D}$, key($K$) and value($V$) are generated from the enhancement features $\mathbf{F}'$. Through the attention mechanism, the label-related group cues in the scene can be located, and the semantic and visual information are integrated together. For the input T-frame images and K semantic embeddings of group activity labels, the module outputs the fused features $\mathbf{W} \in \mathbb{R}^{T\times K\times D}$.

In order to search for a better way to fuse the semantic and visual information, we design three more schemes, as displayed in Fig. 5. The two schemes in the first row both take semantic embeddings as queries and visual features as keys and values, while the two schemes in the second row additionally take visual features as queries and updated semantic embeddings as keys and values. These schemes can be combined with multi-hot

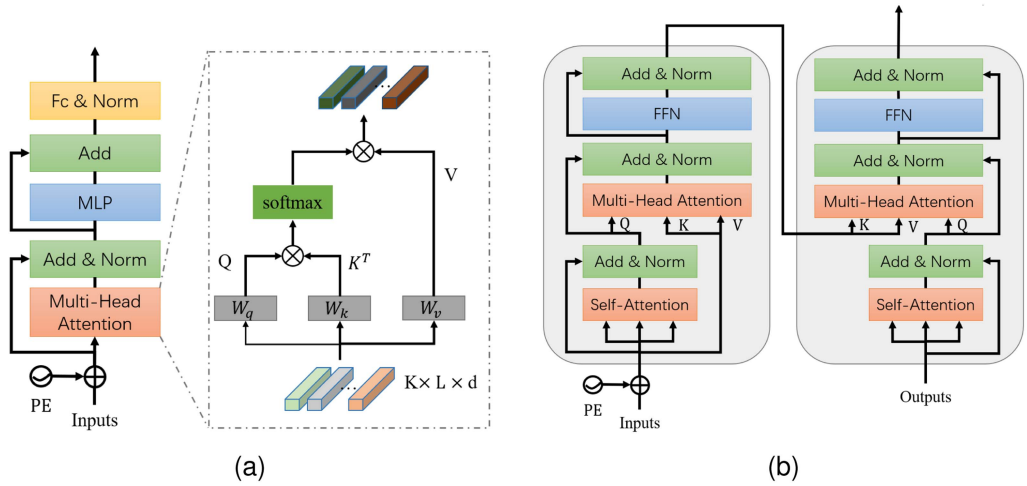

Fig. 4. Detailed architectures of the Semantic Encoder and Visual Encoder and Semantic Decoder modules. (a) Semantic Encoder. (b) Visual Encoder and Semantic Decoder.

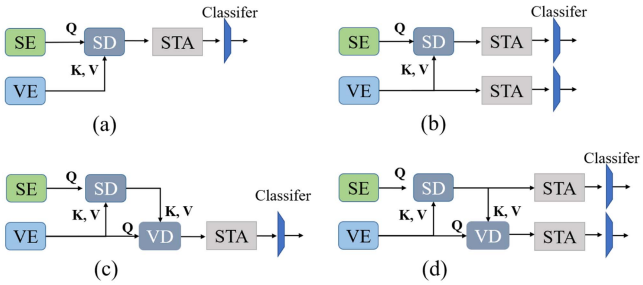

Fig. 5. Framework of the different schemes. (a) SD + 1 Classifier, (b) SD + 2 Classifiers, (c) SD + VD + 1 Classifier, (d) SD + VD + 2 Classifiers.

encoding and multi-label classification respectively, forming four different methods.

D. Spatio-Temporal Aggregation

In this section, we will perform the aggregation of the activity-specific features. For the token embeddings $\mathbf{W} \in \mathbb{R}^{T\times K\times D}$ obtained from activity-specific feature generation module, we first aggregate the embeddings expressing the same semantics by stacking one-dimensional temporal convolutional blocks. An average pooling layer is taken to capture the temporal dynamics and output $\tilde{\mathbf{W}} \in \mathbb{R}^{K\times D}$. We then use the multi-head self-attention layer to capture the dynamic relationships among the K group token embeddings. The role of self-attention here is to find the similarities and differences between labels, in order to obtain the most relevant group representation to the scene. By applying another average pooling operation, the group representation $\mathbf{g} \in \mathbb{R}^{D}$ can be obtained.

E. Classification

Traditional approaches for group activity recognition usually perform the multi-class classification, using the standard cross-entropy loss. In contrast to these methods, we design the network to optimise the sub-label classification results. This ensures that the extracted activity-specific features are more consistent with semantics, through the ability of distinguishing the labels and sub-labels. Concretely, after obtaining the group representation $\mathbf{g}$, we predict the classification scores for sub-labels by using a classifier, generating the vector $\mathbf{s} = [\mathbf{s}^1, \mathbf{s}^2, ..., \mathbf{s}^C]$, where C is the number of sub-labels, $\mathbf{s}$ is classification score. We use a multi-label loss function as a constraint, in which the ground truth for multi-label classification can be obtained from the correspondence between group activity labels and sub-labels.

We compute the final scores and train the entire network using a multi-label classification loss as follows:

$$\mathcal{L}(\mathbf{y}, \mathbf{s}) = \sum_{c=1}^{C} \mathbf{y}^c \log\left(\sigma(\mathbf{s}^c)\right) + (1-\mathbf{y}^c)\log\left(1-\sigma(\mathbf{s}^c)\right) \quad (1)$$

where $\sigma(\cdot)$ is the Sigmoid function, $\mathbf{y}$ is the ground truth of sub-labels.

For the volleyball dataset, we define the number of sub-labels to be 6. We divide the multi-label classification scores into two groups according to the feature domains (i.e., one for left and right, another for pass, set, spike and win-point). The two groups are combined by the Kronecker product, generating the 8 group label classification results. For the NBA dataset, we define the number of sub-labels as 7 and evaluate the scores in three feature domains (i.e., 3p and 2p as one group, layup and w/o layup as one group, and success, off and def as one group). We select the highest-scoring labels in each of the feature domains as the final results and translate them into group activity based on the label correspondence in Table II. We consider the predictions valid when all three groups are predicted correctly.

IV. Experiments

In this section, we first describe three datasets, namely the volleyball, NBA and collective activity datasets, and show the implementation details for each. Afterwards, we provide extensive ablation experimental results to demonstrate the role of the various modules of our approach. Finally, we provide feature

TABLE III
COMPARISON WITH STATE-OF-THE-ART SCHEMES ON VD AND CAD

scheme	Backbone	Training		Testing	MCA(VD)	Merged MCA(VD)	MCA(CAD)
		AL	Bbox	Bbox			
HDTM [15]	AlexNet	✓	✓	✓	81.9	-	81.5
HANs+HCNs [53]	GoogLeNet	✓	✓	✓	85.1	-	84.3
CCGL [54]	AlexNet	✓	✓	✓	87.6	-	90.0
CERN [5]	Vgg16	✓	✓	✓	87.6	-	87.2
stagNet [4]	Vgg16	✓	✓	✓	89.3	-	89.1
PCTDM [17]	ResNet-18	✓	✓	✓	90.3	94.3	-
ARG [6]	Vgg16	✓	✓	✓	91.9	-	90.1
ARG [6]	Inception-v3	✓	✓	✓	92.5	-	91.0
ARG [6]	ResNet-18	✓	✓	✓	91.1	95.1	-
SACRF [29]	ResNet-18	✓	✓	✓	90.7	92.7	-
PRL [28]	Vgg16	✓	✓	✓	91.4	-	-
GAIM [55]	Inception-v3	✓	✓	✓	91.9	-	90.6
Xu et al [56]	Inception-v3	✓	✓	✓	92.8	-	-
Ehsanpour et al [27]	I3D	✓	✓	✓	93.0	-	89.4
STBiP [31]	Inception-v3	✓	✓	✓	93.3	-	-
AT [9]	I3D	✓	✓	✓	91.4	-	90.8
AT [9]	ResNet-18	✓	✓	✓	90.0	94.0	-
GINs [57]	Vgg16	✓	✓	✓	91.7	-	-
GroupFormer [10]	Inception-v3	✓	✓	✓	94.1	-	93.6
Dual-AI [57]	Inception-v3	✓	✓	✓	94.4	-	-
HIGCIN [7]	Resnet-18		✓	✓	91.4	-	92.5
DIN [8]	Vgg16		✓	✓	93.6	-	-
DIN [8]	Resnet-18		✓	✓	93.1	**95.6**	-
Zappardino *et al* [58]	OpenPose		✓	✓	89.4	-	-
PoseConv3D [59]	3D-CNN		✓	✓	91.3	-	-
Kong *et al* [60]	Inception-v3		✓	✓	92.0	-	-
SSU [12]	Inception-v3	✓	✓		87.1	-	-
CRM [61]	I3D	✓	✓		92.1	-	83.4
SBGAR [16]	Inception-v3	✓			38.7	-	83.7
Zhang *et al* [19]	ZFNet		✓		86.0	-	83.8
ASPHRI [20]	Inception-v3		✓		92.4	94.9	85.0
Detector-free [22]	ResNet-18				90.5	94.4	85.9
Ours	ResNet18				**92.5**	95.3	**86.7**

Only RGB images are used as input."AL" and "-" denote action labels, the results are not provided."✓"denotes the information is used.
Numbers in bold indicate the best performance.

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS IN A WEAKLY SUPERVISED ENVIRONMENT WITHOUT ACTION ANNOTATION ON VD

Method	Backbone	MCA	Merged MCA
PCTDN [17]	ResNet-18	80.5	90.0
ARG [6]	ResNet-18	87.4	92.9
AT [9]	ResNet-18	84.3	89.6
SACRF [29]	ResNet-18	83.3	86.1
DIN [8]	ResNet-18	86.5	93.1
SAM [21]	ResNet-18	86.3	93.1
Dual-AI [57]	Inception v3	-	**95.8**
Ours	ResNet-18	**92.5**	95.3

"-" denote the results are not provided.
Numbers in bold indicate the best performance.

visualisation results with classification confusion matrices to validate the effectiveness of our method by comparing it with state-of-the-art schemes.

A. Dataset

The Volleyball Dataset is edited from 55 volleyball match videos, including 3493 training clips and 1337 test clips. For each video clip, there are eight group activity labels including right set, right spike, right pass, right win-point, left set, left spike, left pass and left win-point. And each actor in the scene has nine individual action labels and bounding boxes including waiting, setting, digging, falling, spiking, blocking, jumping, moving and standing. In order to make a fair comparison, we follow the weak supervision setting of [22], and only consider the use of group activity labels in the experiment. We use two indicators, MCA and merged MCA, to evaluate the accuracy, where merging MCA is to combine the right set and right pass of labels into the right set, and merging the left set and left transfer of labels into the left set, as shown in SAM [21]

The NBA Dataset comes from 181 NBA game videos, including 7624 training clips and 1548 test clips. For each video clip, there are only nine group activity annotations including '2p-succ.', '2p-fail.-off.', '2p-fail.-def.','2p-layup-succ.', '2p-layup-fail.-off.', '2p-layup-fail.-def.', '3p-succ.', '3p-fail.-off.', '3p-fail.-def.', and the annotation cost is low. In addition, it is difficult to identify since the NBA dataset has the characteristics of a large time span, changes in the number of actors, changes in camera perspective, and so on. In the experiment, we followed the experimental setup of DF, using multi-class classification accuracy (MCA) and average accuracy per category (MPCA) measurements.

The Collective Activity Dataset consists of 44 video sequences, with 31 videos as the training set and 13 videos as the test set, for a total of 2481 activity clips. Each video clip has five group activity labels, including crossing, walking, waiting, talking and queuing. And each actor in the scene was annotated with one of 6 individual action labels and bounding boxes, including NA, Crossing, Walking, Waiting, Talking and Queuing.

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE NBA DATASET

Method	Backbone	# Params	FLOPs	MCA(NBA)	MPCA(NBA)
ARG [6]	ResNet-18	49.5M	307G	59.0	56.8
AT [9]	ResNet-18	29.6M	305G	47.1	41.4
SACRF [29]	ResNet-18	53.7M	339G	56.3	52.8
DIN [8]	ResNet-18	26.0M	304G	61.6	56.0
SAM [21]	ResNet-18	25.5M	304G	54.3	51.5
Dual-AI [57]	Inception-v3	-	-	51.5	44.8
Detector-free [22]	ResNet-18	17.5M	313G	75.8	71.2
Ours-Lite	ResNet-18	**17.0M**	**303G**	75.8	71.5
Ours	ResNet-18	17.8M	309G	**77.1**	**72.7**

Numbers in bold indicate the best performance.

TABLE VI
ABLATION STUDY ON DIFFERENT SEMANTIC ENCODING AND CLASSIFICATION MANNERS

	Multi-class	Multi-label(5)	Multi-label(6)
Gaussian	91.3	90.8	91.8
One-hot + SE	91.8	91.5	92.2
Multi-hot (5) + SE	91.8	92.1	92.0
Multi-hot (6) + SE	92.1	91.7	92.5

One-hot and multi-hot stand for One-hot encoding and multi-hot encoding respectively. The number in () represents the number of sub-labels or the length of the encoding.

TABLE VII
RESULTS OF DIFFERENT OPTIONS FOR VISUAL ENCODER AND SEMANTIC ENCODER

Visual Encoder		Semantic Encoder			MCA
Self-Attention	w/o VE	Self-Attention	MLP	w/o SE	
✓		✓			92.5
✓			✓		91.3
✓				✓	91.8
	✓	✓			91.3
	✓			✓	90.7

TABLE VIII
RESULTS OF DIFFERENT FUSION METHODS

Model	Group Activity (VD)
(a) SD +1 Classifier	92.5
(b) SD +2 Classifiers	91.9
(c) SD + VD + 1 Classifier	89.4
(d) SD + VD + 2 Classifiers	90.7

TABLE IX
COMPARISON ON VOLLEYBALL DATASET / NBA DATASET UNDER LIMITED TRAINING DATA

Scheme	Data Ratio			
	10%	25%	50%	100%
Detector-free [22]	67.9 / -	78.0 / 39.2	82.6 / 60.3	90.5 / 75.8
Ours	80.1 / -	83.7 / 49.8	86.2 / 64.1	92.5 / 75.8

B. Implementation Details

We segmented two datasets, NBA dataset T=18 and Volleyball dataset T=5, following the Detector-free [22] strategy and adjusted the sampled image pixels to 720 × 1280 for fair comparison following the Groupformer [10] processing strategy. We chose the ResNet-18 model as the CNN backbone. We set the same settings on the volleyball and NBA datasets for the semantic encoder, stacking 4 Transformer encoder layers, with 2 attention heads and 256 channels. For NBA dataset, we stack visual encoder and semantic decoder layer of 2, with 2 attention heads and 256 channels; For volleyball dataset, two layers are stacked, with 4 attention heads and 256 channels.

We used ADAM [62] to optimise the network parameters of the model on both datasets with β_1=0.9, β_2=0.999, $\epsilon = 10^{-8}$. For the volleyball dataset, the learning rate was initially set to 1e-6, with a linear warm-up to 1e-4 for 5 epochs and a linear decay after the 6th epoch. The weight decay was set to 1e-3, using a small batch of size 4 and trained for 30 epochs. For the NBA dataset, the learning rate was initially set to 5e-7, with a linear warm-up to 5e-5 for 5 epochs and a linear decay after the 6th epoch. The weight decay was set to 1e-4, using a small batch of size 2 and trained for 50 epochs.

C. Comparison With the State-of-The-Arts

As shown in Table III, we have compared the proposed scheme with a number of state-of-the-art group activity recognition schemes. For a fair comparison, we only report the results of the comparative schemes with RGB input. In addition, we compare these schemes in two parts. The first part is a fully supervised setup scheme [1], [4], [5], [6], [9], [10], [15], [17], [27], [28], [29], [31], [53], [54], [55], [56], [57] that uses bounding boxes and individual action labels during the training and testing phases. The second part for [7], [8], [12], [16], [19], [20], [22], [58], [59], [60], [61] used only bounding boxes or action labels or neither.

From the experimental results, it can be seen that on the volleyball dataset, our method performs much better with MCA 92.5 and Merged MCA 95.3 than detector-free, which is currently the only recognition algorithm that does not use individual annotations. Our scheme outperforms almost all other non-fully supervised algorithms and is only slightly below DIN [8], which uses Vgg16 as the backbone and uses individual bounding box annotations during the testing and training phases. In addition, our scheme outperforms several fully supervised setup schemes.

We also provide the results of our experiments on the Collective dataset, as shown in the last column of Table III. We use one-hot encoding instead of multi-label embedding and remove the label conversion module. In the case of using only group labels, our scheme is 86.7%, which is better than the results of Detector-free and outperforms some of the more strongly supervised algorithms. It is worth noting that Detector-free does not use motion information in our experiments, since the results are better without the motion information.

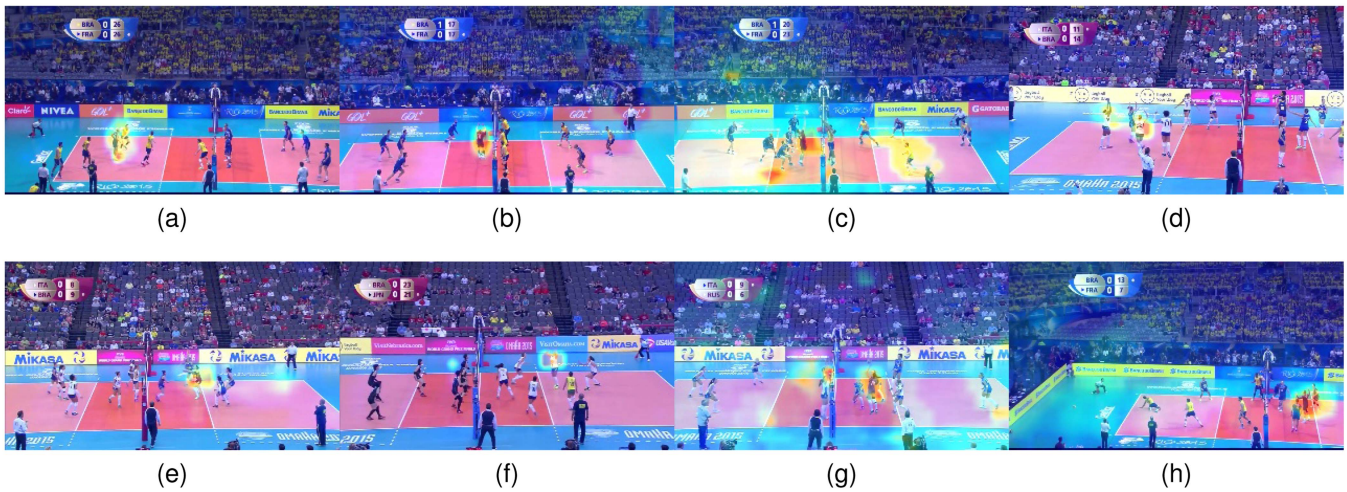

Fig. 6. Visualization of the Semantic Decoder attention maps on the volleyball dataset. (a) l-pass. (b) l-set. (c) l-spike. (d) l-win-point. (e) r-pass. (f) r-set. (g) r-spike. (h) r-win-point.

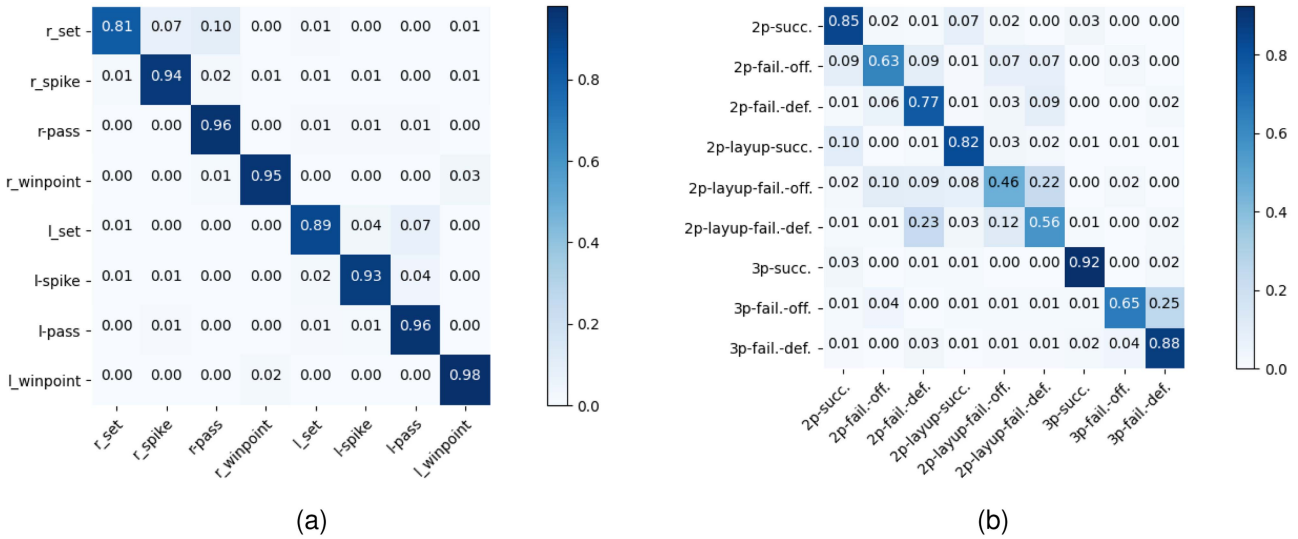

Fig. 7. Confusion matrix on VD and NBA. (a) The Volleyball dataset, (b) The NBA dataset. Each row denotes the predicted class and each column denotes an instance of the real class. "l-" and "r-" are abbreviations for "left" and "right" of the group activity labels in the volleyball dataset.

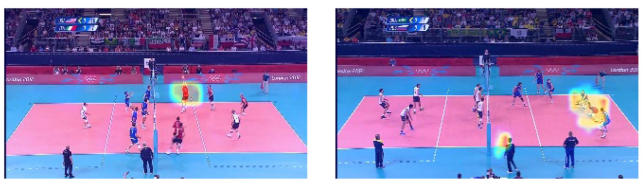

Fig. 8. Visualisation of prediction error cases. The dataset labels are in green and the predictions are in red.

In addition, we provide an experimental comparison of existing fully supervised advanced algorithms in the weakly supervised setting for the volleyball dataset, as shown in Table IV. Data were obtained from Detector-free [22], ground truth bounding boxes are replaced using object detectors preprocessed on an external dataset and individual action classification heads are removed. The results show that our approach outperforms all group activity recognition models in the weakly supervised setting.

For the NBA dataset, we compared our method with state-of-the-art group activity recognition methods, and these experimental results are from Detector-free [22]. We also compared the computational complexity of the model in the weakly supervised setting of the NBA dataset, and the results are shown in Table V. We provide experimental results for two variants, using two layers of VE and SD, and one layer of VE and SD (Lite). In the Lite case, our method achieved an MCA of 75.8, which is comparable to the results of Detector-free, and an MPCA of 71.5, which is slightly higher than Detector-free's by 0.3%. However, our method requires fewer parameters and FLOPs than other group activity recognition methods, even without computing the computational complexity of their object detectors. In the case of stacking 2 layers of VE and SD, our method achieves the optimal performance of MCA 77.1 and MPCA 72.7.

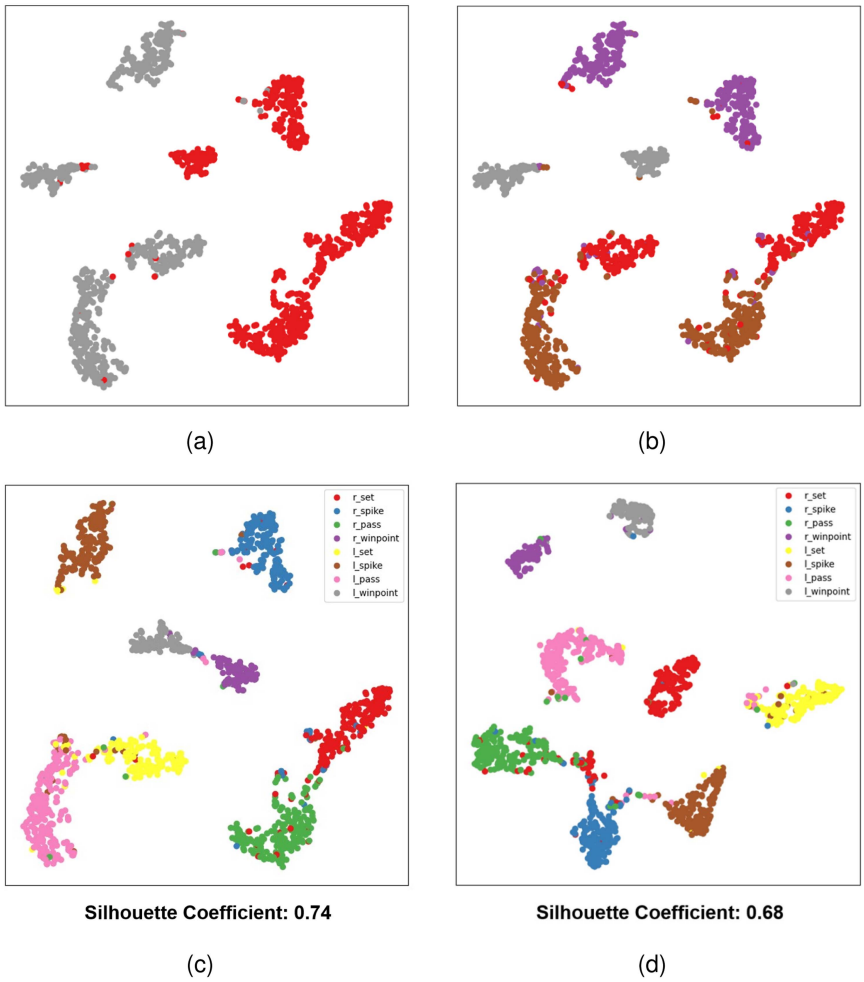

Fig. 9. t-SNE [63] visualisation of feature embedding learned by our model on the volleyball dataset. (a) Visualisation of the left and right sub-labels in our model. (b) Visualisation of the pass, set, spike and win-point sub-labels in our model. (c) Visualisation of group labels for our model in a multi-hot encoding manner. (d) Visualisation of group labels for our model in a one-hot encoding manner.

D. Ablation Studies

Effects of Multi-Hot Encoding and Multi-Label Classification: We investigate the impact of both multi-hot encoding and multi-label classification on Volleyball dataset. We implement a total of 12 variant models for comparison, in which there are four different semantic encoding manners and three different classification losses. The semantic encoding manners include the Gaussian function with mean 0 and standard deviation of 1, one-hot encoding, multi-hot encoding with five sub-labels by merging the left and right sub-labels, and multi-hot encoding with six sub-labels. The classification losses include common multi-class classification loss, multi-label classification loss with five sub-labels, and multi-label classification loss with six sub-labels. As shown in Table VI, the multi-hot encoding with six sub-labels combined with the multi-label classification loss with six sub-labels achieves the best performance. Moreover, under distinct semantic encoding manners, the multi-label classification loss using six sub-labels can always lead to the performance improvement, compared with common multi-class classification loss.

Effects of Semantic Encoder and Visual Encoder: The semantic encoder and visual encoder are two sub-modules to build semantic embeddings and extract visual features, respectively. We investigate different ways to implement the two functions. For semantic encoder, we experiment with three options: Self-Attention, MLP and w/o SE. For visual encoder, we experiment with two options: Self-Attention and w/o VE. The comparison results on volleyball dataset are shown in Table VII. We can observe that the recognition performance drops significantly without SE and VE, and the self-attention mechanism works well for both SE and VE.

Variants of Semantic and Visual Information Fusion: We compare the four different schemes proposed in Section III-C to search for the best way to integrate the semantic and visual information. The comparison results on Volleyball dataset are shown in Table VIII. We can see that the first scheme using the semantic decoder and one classifier is optimal. This may be because the visual feature maps contain more noises that can easily affect the recognition under weakly supervised setting. We conjecture

that it might be helpful to study the sparse visual tokens in future work.

Experiment under Limited Training Data: Following the experimental setup of Dual-AI [57], we report the experimental results for the volleyball dataset/NBA dataset under 10%, 25%, 50%, and 100% scale training set conditions, as shown in Table IX. The results show that on the volleyball dataset, our scheme maintains more than 80 recognition accuracies with 10% of the training data. For the NBA dataset, we only compare the experimental results for training data above 25% scale, and the results show that our method performs significantly better than Detector-free [22]. The advantage of our scheme is more obvious with less training data compared to detector-free, which is attributed to our multi-label semantic encoding approach. Under the same conditions, our scheme is able to locate more useful scene information for training.

E. Visualization and Analysis

In Fig. 6 we illustrate the attention visualisation obtained from the semantic decoder on the volleyball dataset. The results show that semantic embeddings are learned to focus on the parts of the scene that are most relevant to the group activity. Activities such as set and pass focus on the key actor in which the action takes place, win-point focuses on the actor celebrating a win, and spike focuses not only on the spiking actor but also on the defending actor of the opponent.

Fig. 7 shows the confusion matrices on the volleyball dataset and NBA dataset. For the volleyball dataset, the accuracies of all categories are above 90%, except for l-set and r-set which are easily misclassified as l-pass and r-pass. The confusion between set and pass activities has been a common issue of existing methods. Figs. 6(f) and 8(a) are respectively a success case and a typical failure case to illustrate the difficulty of distinguishing between pass and set, even when the proper attentive regions involving key actors are captured. Besides, the volleyball dataset contains quite a few wrong labels. For example, in Fig. 8(b), the r-pass activity is incorrectly labelled as r-set, in which case our method can correct it, demonstrating the superiority of our approach. For the NBA dataset, the recognition accuracy gradually decreases with the number of basic words in the compound words denoting group activity labels. The reason is video samples in this dataset for the compound words containing more basic words are more likely to have nontrivial temporal structure.

Fig. 9 shows a visualisation of the feature t-SNE [63] clustering results learned by our model on the volleyball dataset. our scheme distinguishes well between left and right activities as shown in Fig. 9(a), which is attributed to the classification of our left and right sub-labels. We not only show the classification results for the group labels, but also provide classification visualisations for the sub-labelled features. In order to show more clearly the differences between the multi-hot encoding and one-hot encoding approaches, we provide the clustered silhouette coefficients of the features for the two models. We can find that the multi-hot encoding approach helps to clearly distinguish each category.

V. Conclusion and Future Work

We propose a novel approach for weakly supervised group activity recognition. The method extracts fine-grained semantics based on the hierarchy inherent in group-level labels to help encode group visual contexts. In addition, we devise a unique resultant transformation based on the correspondence between the labels to achieve the final recognition. We have experimented on three datasets and established new state-of-the-art results in the weakly supervised setting. Experimental results fully verify the significance of the hierarchical information of labels for weakly supervised group activity recognition. The similar conclusion also appeared in the advanced works [64], [65] that organized classes in a hierarchical structure instead of a flat list for large-scale image classification. In the future, we will further consider the automatic learning of label hierarchy for the applications in complex scenes with many labels or sub-labels.

References

[1] L. Lu et al., "GAIM: Graph attention interaction model for collective activity recognition," *IEEE Trans. Multimedia*, vol. 22, pp. 524–539, 2020.

[2] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1549–1562, Aug. 2011.

[3] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S. C. Zhu, "Joint inference of groups, events and human roles in aerial videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4576–4584.

[4] M. Qi et al., "stagNet: An attentive semantic RNN for group activity recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 101–117.

[5] T. Shu, S. Todorovic, and S.-C. Zhu, "CERN: Confidence-energy recurrent network for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5523–5531.

[6] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9964–9974.

[7] 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.*, vol. 45, no. 6, pp. 6955–6968, Jun. 2023.

[8] H. Yuan, D. Ni, and M. Wang, "Spatio-temporal dynamic inference network for group activity recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 7476–7485.

[9] K. Gavrilyuk, R. Sanford, M. Javan, and C. G. Snoek, "Actor-transformers for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 839–848.

[10] S. Li et al., "Groupformer: Group activity recognition with clustered spatial-temporal transformer," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 13668–13677.

[11] W. Zhou, L. Kong, Y. Han, J. Qin, and Z. Sun, "Contextualized relation predictive model for self-supervised group activity representation learning," *IEEE Trans. Multimedia*, vol. 26, pp. 353–366, 2023.

[12] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4315–4324.

[13] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4772–4781.

[14] M. S. Ibrahim and G. Mori, "Hierarchical relational networks for group activity recognition and retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 721–736.

[15] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1971–1980.

[16] X. Li and M. C. Chuah, "SBGAR: Semantics based group activity recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2876–2885.

[17] R. Yan, J. Tang, X. Shu, Z. Li, and Q. Tian, "Participation-contributed temporal dynamic model for group activity recognition," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1292–1300.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[19] P. Zhang, Y. Tang, J.-F. Hu, and W.-S. Zheng, "Fast collective activity recognition under weak supervision," *IEEE Trans. Image Process.*, vol. 29, pp. 29–43, 2020.
[20] L. Wu et al., "Active spatial positions based hierarchical relation inference for group activity recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2839–2851, Jun. 2023.
[21] R. Yan, L. Xie, J. Tang, X. Shu, and Q. Tian, "Social adaptive module for weakly-supervised group activity recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 208–224.
[22] D. Kim, J. Lee, M. Cho, and S. Kwak, "Detector-free weakly supervised group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20083–20093.
[23] Y. Zhang, X. Li, and I. Marsic, "Multi-label activity recognition using activity-specific features and activity correlations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14625–14635.
[24] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2009, pp. 1282–1289.
[25] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2012, pp. 215–230.
[26] M. Wang, B. Ni, and X. Yang, "Recurrent modeling of interaction context for collective activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3048–3056.
[27] M. Ehsanpour et al., "Joint learning of social groups, individuals action and sub-group activities in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 177–195.
[28] G. Hu, B. Cui, Y. He, and S. Yu, "Progressive relation learning for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 980–989.
[29] R. R. A. Pramono, Y. T. Chen, and W. H. Fang, "Empowering relational network by self-attention augmented conditional random fields for group activity recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 71–90.
[30] R. R. A. Pramono, W.-H. Fang, and Y.-T. Chen, "Relational reasoning for group activity recognition via self-attention augmented conditional random field," *IEEE Trans. Image Process.*, vol. 30, pp. 8184–8199, 2021.
[31] H. Yuan and D. Ni, "Learning visual context for group activity recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3261–3269.
[32] D. Zhao, Q. Gao, Y. Lu, and D. Sun, "Non-aligned multi-view multi-label classification via learning view-specific labels," *IEEE Trans. Multimedia*, vol. 25, pp. 7235–7247, 2023.
[33] Z.-M. Chen, Q. Cui, X.-S. Wei, X. Jin, and Y. Guo, "Disentangling, embedding and ranking label cues for multi-label image recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 1827–1840, 2021.
[34] X. Deng, S. Feng, G. Lyu, T. Wang, and C. Lang, "Beyond word embeddings: Heterogeneous prior knowledge driven multi-label image classification," *IEEE Trans. Multimedia*, vol. 25, pp. 4013–4025, 2022.
[35] Z. Ge, D. Mahapatra, S. Sedai, R. Garnavi, and R. Chakravorty, "Chest X-rays classification: A multi-label and fine-grained problem," 2018, *arXiv:1807.07247*.
[36] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2016, pp. 684–700.
[37] H. Jain, Y. Prabhu, and M. Varma, "Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications," in *Proc. ACM Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 935–944.
[38] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 522–531.
[39] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3286–3293.
[40] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 28, pp. 91–99.
[42] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.
[43] J. Wang et al., "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2285–2294.
[44] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 464–472.
[45] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5177–5186.
[46] Q. Li, M. Qiao, W. Bian, and D. Tao, "Conditional graphical lasso for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2977–2986.
[47] Y. Wang et al., "Multi-label classification with label graph superimposing," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 12265–12272.
[48] M. Tkalcic, A. Odic, A. Kosir, and J. Tasic, "Affective labeling in a content-based recommender system for images," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 391–400, Feb. 2013.
[49] G. Zhou et al., "Deep interest network for click-through rate prediction," in *Proc. ACM Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1059–1068.
[50] Y. Qu et al., "Product-based neural networks for user response prediction," in *Proc. IEEE Int. Conf. Data Mining*, 2016, pp. 1149–1154.
[51] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 5998–6008.
[52] N. Carion et al., "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
[53] L. Kong, J. Qin, D. Huang, Y. Wang, and L. V. Gool, "Hierarchical attention and context modeling for group activity recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1328–1332.
[54] J. Tang, X. Shu, R. Yan, and L. Zhang, "Coherence constrained graph LSTM for group activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 636–647, Feb. 2022.
[55] D. Xu et al., "Group activity recognition by using effective multiple modality relation representation with temporal-spatial attention," *IEEE Access*, vol. 8, pp. 65689–65698, 2020.
[56] Y. Tang, Y. Wei, X. Yu, J. Lu, and J. Zhou, "Graph interaction networks for relation transfer in human activity videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 2872–2886, Sep. 2020.
[57] M. Han et al., "Dual-AI: Dual-path actor interaction learning for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2990–2999.
[58] F. Zappardino, T. Uricchio, L. Seidenari, and A. D. Bimbo, "Learning group activities from skeletons without individual action labels," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2021, pp. 10412–10417.
[59] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2969–2978.
[60] L. Kong, D. Pei, R. He, D. Huang, and Y. Wang, "Spatio-temporal player relation modeling for tactic recognition in sports videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6086–6099, Sep. 2022.
[61] S. M. Azar, M. G. Atigh, A. Nickabadi, and A. Alahi, "Convolutional relational machine for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7892–7901.
[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[63] L. V. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[64] S. Zhang, R. Xu, C. Xiong, and C. Ramaiah, "Use all the labels: A hierarchical multi-label contrastive learning framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16660–16669.
[65] S. Zeng, R. T. d. Combes, and H. Zhao, "Learning structured representations by embedding class hierarchy," in *Proc. 11th Int. Conf. Learn. Representations*, 2022.

Lifang Wu (Member, IEEE) received the B.S., M.S. and Ph.D. degrees from the Beijing University of Technology, Beijing, China, in 1991, 1994, and 2003, respectively. She is currently a Professor with the Beijing University of Technology. Her research interests include image/video understanding, group activity recognition, face anti-spoofing, multi-model sentiment analysis, and intelligent 3D printing. She is a CCF outstanding member. She was the recipient of five provincial and ministerial science and technology awards. She was also the recipient of the Best Paper Award of ICCV 2021 Workshop on Human-centric Trustworth Computer Vision and PRCV 2021 Best Paper Hornorable Mentions. She participated in organizing the CCCV 2017, PRCV 2019, PRCV 2022, ChinaMM 2021, ChinaMM 2020, and CCIG 2022.

Meng Tian received the bachelor's degree in communication engineering in 2021 from the Beijing University of Technology, Beijing, China, where he is currently working toward the master's degree in information and communication engineering. His research interests include video analysis and group activity recognition.

Ke Gu (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2009 and 2015, respectively. He is currently a Professor with the Beijing University of Technology, Beijing, China. He is also with the Faculty of Information Technology, the Engineering Research Center of Intelligent Perception and Autonomous Control of Ministry of Education, Beijing Laboratory of Smart Environmental Protection, the Beijing Key Laboratory of Computational Intelligence and Intelligent System, and the Beijing Artificial Intelligence Institute, Beijing University of Technology, Beijing. His research interests include industrial vision, environmental perception, image processing, and machine learning.

Ye Xiang received the B.S. degree from the University of Science and Technology, Beijing, China, in 2012, and the Ph.D. degree from the Beijing Institute of Technology, Beijing, in 2020. She is currently a Lecturer with the School of Information and Communication Engineering, Beijing University of Technology, Beijing. Her main research interests include image processing, video analysis, and machine learning.

Ge Shi received the Ph.D. degree in computer science from the Beijing Institute of Technology, Beijing, China, in 2020. He is currently a Lecturer with the Faculty of Information Technology, Beijing University of Technology, Beijing. His main research interests include information extraction, text generation, and cross-modal learning.