Empowering Relational Network by Self-Attention Augmented Conditional Random Fields for Group Activity Recognition

Rizard Renanda Adhi Pramono[0000−0002−6668−5167], Yie Tarng Chen[0000−0002−7221−1603], and Wen Hsien Fang[0000−0001−6402−2688]

National Taiwan University of Science and Technology
{d10702801,ytchen,whf}@mail.ntust.edu.tw

Abstract. This paper presents a novel relational network for group activity recognition. The core of our network is to augment the conditional random fields (CRF), amenable to learning inter-dependency of correlated observations, with the newly devised temporal and spatial self-attention to learn the temporal evolution and spatial relational contexts of every actor in videos. Such a combination utilizes the global receptive fields of self-attention to construct a spatio-temporal graph topology to address the temporal dependency and non-local relationships of the actors. The network first uses the temporal self-attention along with the spatial self-attention, which considers multiple cliques with different scales of locality to account for the diversity of the actors' relationships in group activities, to model the pairwise energy of CRF. Afterward, to accommodate the distinct characteristics of each video, a new mean-field inference algorithm with dynamic halting is also addressed. Finally, a bidirectional universal transformer encoder (UTE), which combines both of the forward and backward temporal context information, is used to aggregate the relational contexts and scene information for group activity recognition. Simulations show that the proposed approach surpasses the state-of-the-art methods on the widespread Volleyball and Collective Activity datasets.

Keywords: bidirectional universal transformer encoder, self-attention mechanism, conditional random field, graph cliques, group activity.

1 Introduction

Group activity recognition has received much attention in view of numerous applications in abnormal event detection [1], sport tactical analysis [2], social behaviours [3], and *etc.* Understanding group activities requires reasoning on how interactions of every actor with different individual actions can lead to a collective activity. This is a challenging issue as the relations among the actors are dynamic [4] and, in addition, some individual actions may not be directly related to the group activity [1, 5]. Therefore, it is of great importance to effectively learn the spatial relational contexts and temporal evolution of the actors in the group activity, as illustrated in Fig. 1.

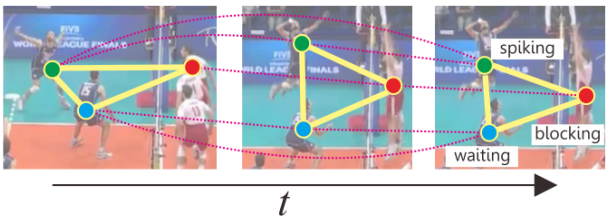


Fig. 1: A spatial-temporal graph learnt by the self-attention augmented CRF to model spatial relations and temporal evolution of actors in 'Left Spiking' activity.

A number of methods have been proposed to deal with group activity recognition. Earlier approaches [6–8] reasoned interactions of actors without a deep network architecture. However, these approaches do not properly address the temporal relationship of the actors and leverage complex semantic information from deep networks. To tackle this setback, [3, 5, 9] utilized recurrent neural network (RNN) to learn the dynamics of the individual actions. These methods, however, do not consider the spatial relational structure of the actors that is important in understanding complex group activities, where the actors' appearances and movements are dynamically changing. To learn the interactions of the actors, Deng *et al.* [10] introduced a structure inference machine to construct a graphical RNN model using a gating function. Shu *et al.* [11] proposed a graphical long short-term memory (LSTM), composed of an energy-based layer that can be optimized with a relatively small scale of data. Wang *et al.* [12] developed an efficient interaction model that combines person-level, group, and scene information. Biswas *et al.* [13] designed a grid pooling layer to aggregate the interaction information from the graphical RNN with a varying number of nodes and edges. Ibrahim *et al.* [2] proposed an autoencoder network comprising of multiple relational layers to learn multi-person interactions. Qi *et al.* [1] developed a soft attentive mechanism with message passing to model the interaction of relevant actors. However, the aforementioned approaches [1, 2, 10, 11, 13] are based on either RNN or LSTM, which generally requires a large variety of training data and may encounter the vanishing gradient problem [14]. Azar *et al.* [15] proposed a specific convolutional neural network (CNN) to learn group activities without explicitly detecting individual actions. Wu *et al.* [4] constructed relational graphs based on self-similarity of the actors. However, the relational graphs in [4] are limited to a few frames of observation without considering the diversity of actors' relationships at various spatial distances.

In this paper, we propose a novel relational network for group activity recognition. The core of our network is to augment the mean-field conditional random fields (CRF) [16] with the newly devised temporal and spatial self-attention to learn the temporal evolution and spatial relational contexts of every actor in videos. In contrast to the convolutional or recurrent architectures in [16, 17], self-attention, which calculates the response of every position in a sequence by relating it to all other positions, has global receptive fields across the whole data [18, 19]. Such a combination thereby allows CRF, amenable to learning inter-dependency of correlated observations, to infer individual actions based on a spatio-temporal graph topology that considers temporal dependency of ev-

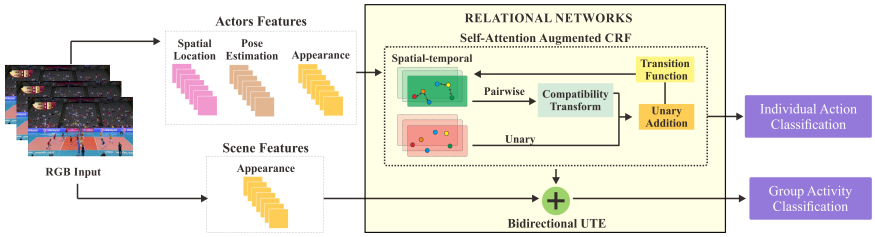


Fig. 2: Overview of the proposed network, which first uses a self-attention augmented CRF to produce the spatial relational contexts and temporal evolution of every actor. A bidirectional UTE is then used to aggregate the relational contexts and scene information.

ery actor across the frames and their non-local relationships. The network first employs the temporal self-attention along with the spatial self-attention, which considers multiple fully-connected sub-graphs, cliques [20], with different scales of locality to address the diversity of the actors' relationships in the group activities, in which their interactions can be local or non-local. As an illustration in Fig. 1, the interaction between 'blocking' and 'spiking' in the last frame of the 'Left Spiking' activity is local as those actions are close to each other while in the first frame it is non-local. Thereafter, to accommodate the distinct characteristics of each video, a new mean-field inference algorithm with dynamic halting is also addressed. Finally, a bidirectional universal transformer encoder (UTE) [21], which combines both of the forward and backward temporal context information to deal with videos with similar patterns in the first few frames, is utilized to aggregate the relational contexts with scene information. Simulations show that our network can achieve state-of-the-art performance on the widely adopted Volleyball and Collective Activity datasets.

The contributions of this paper include: (i) the mean-field CRF inference is reinforced by the temporal and spatial self-attention to facilitate the learning of the spatial relations and temporal evolution of the actors. To the best of the authors' knowledge, this is the first time CRF and self-attention are combined together to jointly model the spatial-temporal relations of multiple actors in action recognition; (ii) the CRF considers the pairwise energy with multi-scale cliques to deal with the diverse relationships of multiple actors in every frame; (iii) the proposed mean-field inference algorithm can adaptively decide an appropriate number of iterations; (iv) a bidirectional UTE is devised to aggregate the relational contexts and scene information for group activity recognition.

2 Related Works

CNN Based Action Recognition. CNN has become an important milestone in video context understanding because of its effectiveness to extract meaningful image features. In action recognition, the majority of CNN architectures can be classified into two categories: two-stream networks [22–24], which are trained to capture appearance and motion information from RGB frames and optical flow images, and 3D networks [25, 26], which are composed of spatial and temporal convolutional layers to process a number of consecutive video frames. Some

recent methods [27–29] combined both architectures to attain a good trade-off between the accuracy and efficiency of the network in training and inference.

Attention Mechanism. Attention mechanism has been extensively used to improve the capability of CNN to extract fine-grained image features [30–33]. Li *et al.* [30] made use of an attentional masking scheme to filter noisy background information for more precise video object segmentation. Zhang *et al.* [31] proposed a generative adversarial network with a spatial attention mechanism that helps localize attribute-specific regions for face attribute editing. Zhao *et al.* [32] developed a pyramid feature attention network that can capture multi-level visual contexts for saliency detection. Fu *et al.* [33] introduced a dual attention network to learn correlation among channel and spatial feature maps.

Learning of Temporal Dependency. Temporal dependency is a core issue in video understanding as the past information can help infer the present and future behavior. For instance, a time-delayed graphical model was developed in [34] to detect global anomaly from multiple disjoint cameras. Swears *et al.* [35] took advantage of the Granger Causality to measure the temporal dependency between two time sequences. Meanwhile, the majority of the approaches [36–38] relied on RNN to learn temporal dependency from sequences of video frames. However, RNN and its variants such as LSTM have difficulty to generalize when the volume of training data is small and highly aperiodic [14]. Inspired by the impressive success of transformer based methods [18, 21, 19] in natural language processing, several recent works [39–43] made use of similar self-attention mechanisms in action recognition and detection.

Graphical Models. Graphical models [44–46] have been remarkably successful in various image and video analysis tasks. Intille *et al.* [45] modeled correlation of object trajectories to recognize complex activities. Morariu *et al.* [44] designed markov logic networks to recognize events in structured scenarios. Xu *et al.* [46] developed causal and-or graphs to learn the multiple person-object interaction for tracking humans in videos. However, [44–46] are based on non-deep networks, which can not capture spatial semantic information and not be integrated end-to-end with deeper networks. Several approaches [47, 48] have addressed this setback by using CNN for graph models to learn complex semantic information. Li *et al.* [47] devised a deep relational network based on self-similarity to detect important persons in images. However, it is limited to low dimensional features from still images. Wang *et al.* [48] made use of graphical convolutional network to model local and long-term dependency for single human action recognition, but it is not devised for multiple-action scenario.

3 Feature Extraction Network

A feature extraction network, composed of a faster R-CNN on feature pyramid network (FPN) with ResNet-50 [49] and a two-stream inflated 3D network (I3D) [27] fine-tuned for individual action recognition, is employed to generate multi-scale and 3D features, respectively. The final fully-connected layer and the temporally-averaged last convolutional layer of FPN and I3D, respectively, are used to extract appearance features from actors. Also, scene features from the whole frame are generated using the same networks. The spatial location

cues of each actor are obtained by concatenating its spatial position and the spatial distances with the other actors. In addition, the pose information of every actor, characterized by 17 keypoints, is extracted using AlphaPose [50]. The appearance, spatial location, and pose information are aggregated by a linear feed-forward layer to provide the final feature representation for every actor.

4 CRF for Individual Action Recognition

The graph representation for individual action recognition is illustrated in Fig. 1, where actors' interactions in every frame and temporal evolution of every actor are jointly modelled to infer individual action categories. Here, we define an actor's features in a particular frame obtained from the feature extraction network in Sec. 3 as a node. A node in each frame is fully connected to the other nodes in the same frame and a set of nodes from the same actor across a temporal sliding window of M frames is temporally interconnected. Denote a set of individual action labels $\mathcal{X} = \{x_1, \cdots, x_K\}$ and a set of random variables $\mathbf{z} = \{z_{1,1}, \cdots, z_{M,N}\}$, where $z_{i,j} \in \mathcal{X}$ is a random variable assigned to node i in frame j and N is the maximum number of actor in all frames.

The graph can be learnt by a conditional random fields (CRF) that abides by the Markov random fields conditioned on all actors' features $\mathbf{B}$. A CRF graph can be characterized by the Gibbs distribution of the form $P(\mathbf{z}|\mathbf{B}) = \frac{1}{n(\mathbf{B})}\exp(-E(\mathbf{z}|\mathbf{B}))$, where $E(\mathbf{z}|\mathbf{B})$ is the energy of the label assignment and $n(\mathbf{B})$ is the partition function [51]. In a fully-connected CRF model, the total energy can be expressed as a summation of the unary and pairwise energies [52]:

$$E(\mathbf{z}|\mathbf{B}) = \sum_{i,j} \phi_u(z_{i,j}|\mathbf{B}) + \sum_{(i',j')\neq(i,j)} \phi_p(z_{i,j}, z_{i',j'}|\mathbf{B}), \tag{1}$$

where the first term is the unary energy used to compute the cost of assigning a label $z_{i,j} \in \mathbf{z}$ to node i in frame j and the second term is the pairwise energy utilized to determine the cost of assigning labels to the same actor across the frames and to different actors in the same frame.

Minimizing the total energy can provide the most probable label assignment. However, in a dense pairwise graph, the exact minimization is intractable. So, a mean-field algorithm [52] can be adopted to approximate $P(\mathbf{z}|\mathbf{B})$ by a product of independent marginal distributions. However, the mean-field algorithm is not suitable for end-to-end training with deep CNN. Moreover, it is not designed for a spatio-temporal graph, in which the temporal dependency and non-local relationships of the actors are also considered.

5 Proposed Method

This section first introduces temporal and spatial self-attention in Sec. 5.1, followed by a self-attention augmented CRF to generate the relational context and an individual action label for every actor in Sec. 5.2. Next, reformulation of the mean-field inference as a self-attention network is described in Sec. 5.3, followed by the bidirectional UTE for group activity recognition in Sec. 5.4. For easy reference, the overall architecture of the proposed method is depicted in Fig. 2.

5.1 Temporal and Spatial Self-Attention

Inspired by the success of self-attention to encode the structural information of a sequence of data, we use it to model the link between every pair of the nodes of a graph by their feature similarity. By self-attention, non-local edges can be constructed as it has global receptive fields that can simultaneously relate every node to the other nodes in the graph. Given an input sequence of nodes $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_{T_s}] \in \mathbb{R}^{T_s \times F}$ with a feature length of F and a sequence length of T_s, the self-attention function $S(\mathbf{v}_m)$ can be defined as [18]:

$$S(\mathbf{v}_m) = \mathbf{v}_m + \sum_{\forall m' \in \{1,\cdots,T_s\}} p(\mathbf{v}_m, \mathbf{v}_{m'}) e_3(\mathbf{v}_m), \tag{2}$$

$$p(\mathbf{v}_m, \mathbf{v}_{m'}) = \frac{e_1(\mathbf{v}_m) e_2(\mathbf{v}_{m'})^T}{\sqrt{F}}, \tag{3}$$

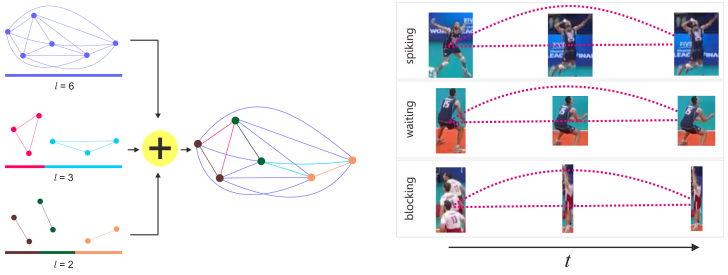
where $e_1(\cdot)$, $e_2(\cdot)$, and $e_3(\cdot)$ are linear transformations implemented by matrix multiplication with trainable weights. Eqn. (3) computes the pairwise similarity between two different nodes. For our problem, we consider two types of self-attention, *i.e.* temporal self-attention and spatial self-attention.

The essence of **temporal self-attention** is to learn the temporal evolution of an actor by feature similarity of the same actor across the frames. The temporal evolution of actor i can be obtained by applying (2) to every node representation of the actor i over M frames, $\{\mathbf{b}_{i,1}, \cdots, \mathbf{b}_{i,M}\}$. Suppose that the temporal self-attention output for node i in frame j is $\mathbf{r}^t_{i,j} = S(\mathbf{b}_{i,j})$. Thereby, the temporal self-attention applied to all actors in all frames, $\mathbf{B} = [\mathbf{b}_{1,1}, \cdots, \mathbf{b}_{N,M}]$, can be expressed as $S_T(\mathbf{B}) = [\mathbf{r}^t_{1,1}, \cdots, \mathbf{r}^t_{N,M}] \in \mathbb{R}^{(M \times N) \times F}$, where N is the maximum number of actors in all frames.

A new **spatial self-attention** mechanism is devised here to learn the statistical correlations of all actors in every frame by their pairwise similarity. It is designed to resolve the natural diversification of relationships among the actors as their interactions can be local or non-local, thereby making it difficult to learn the statistical correlations of all actors. The major idea behind the spatial self-attention is to prioritize the pairwise relationships of nodes with different scales of locality. A pair of local nodes is assigned a high priority edge while a pair of non-local nodes is assigned a low-priority edge. To compute the spatial self-attention, the actors in every frame are divided according to their horizontal positions into a set of cliques defined as fully-connected sub-graphs with l nodes, $2 \leq l \leq N$, where l represents the scale of locality of the cliques. For every clique, (3) is used to compute the edges based on the pairwise similarity of the nodes within that clique. The priority assignment for node i in frame j, $\mathbf{r}^s_{i,j}$, can thus be expressed as a linear combination of the self-attention outputs from all cliques enclosing node i:

$$\mathbf{r}^s_{i,j} = w_2 \cdot \mathbf{p}^{s,2}_{i,j} + \cdots + w_k \cdot \mathbf{p}^{s,k}_{i,j} + \cdots + w_N \cdot \mathbf{p}^{s,N}_{i,j}, \tag{4}$$

where $\mathbf{p}^{s,k}_{i,j} \in \mathbb{R}^{1 \times F}$ is the result of applying (2) to relate node i to the other k nodes in frame j and w_k is the corresponding trainable scalar weight. Eqn. (4)



(a) (b)

Fig. 3: (a) a pairwise fully connected graph constructed from three complete sub-graphs with the number of nodes $l = 2, 3, 6$; (b) temporal self-attention connecting the same actor across the frames.

facilitates the learning of actors' diverse relationships by combining the edges from all cliques with different localities, as illustrated in Fig. 3. Likewise, the spatial self-attention applied to all nodes in all frames can be expressed as $S_S(\mathbf{B}) = [\mathbf{r}^s_{1,1}, \cdots, \mathbf{r}^s_{N,M}] \in \mathbb{R}^{(M\times N)\times F}$.

5.2 Self-Attention Augmented Conditional Random Fields

To resolve the setback of the conventional CRF, this section considers a self-attention augmented CRF to learn the temporal evolution and spatial relational contexts of every actor. Different from convolutional or recurrent architectures, self-attention can generate global receptive fields to facilitate CRF inference having non-local edges while addressing temporal dependency of the actors.

The **unary energy** of assigning an action label to an actor is now obtained by applying a linear feed-forward classifier, $f_u(\cdot)$, to the feature of each node, obtained from the feature extraction network described in Sec. 3:

$$\sum_{(i,j)} \phi_u(z_{i,j}|\mathbf{B}) = -\sum_{i,j} f_u(\mathbf{b}_{i,j}), \tag{5}$$

The **pairwise energy** of assigning different individual action labels to the same actor across the frames and to different actors in the same frame can be modelled by the spatial and temporal self-attention as follows:

$$\sum_{(i',j')\neq(i,j)} \phi_p(z_{i,j}, z_{i',j'}|\mathbf{B}) = -\sum_{i,j} f_p(\mathbf{r}^s_{i,j} + \mathbf{r}^t_{i,j}), \tag{6}$$

where $f_p(\cdot)$ is a linear transformation applied to the outputs of the spatial and temporal self-attention for each node, $\mathbf{r}^s_{i,j}$ and $\mathbf{r}^t_{i,j}$, respectively, discussed in Sec. 5.1. Such a definition of pairwise energy can provide a measure of the cost for assigning the relevant labels to a pair of nodes based on their feature similarity. The new pairwise energy can be viewed as a degenerated version of the higher-order CRF [53–55], where several graph cliques are used to enforce consistency in labelling in image segmentation. However, learning the compatibility of several actors in a clique simultaneously is more difficult as the actors' interactions are dynamic. Therefore, in contrast to the higher order potentials, we restrict

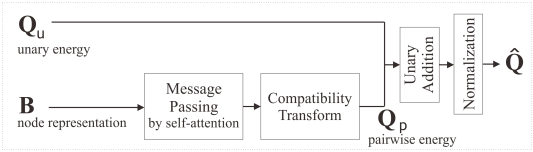


Fig. 4: A single mean-field iteration modelled as a self-attention network.

the problem to the minimization of the pairwise energy based on multi-scale cliques. As illustrated in Fig. 3, the edges of the pairwise graph are obtained by a combination of the spatial self-attention outputs at different scales of cliques.

The total energy formed by the unary and pairwise energy can be minimized by a new mean-field inference to be described in the next section.

5.3 Reformulation of Mean-Field Inference

The mean-field inference can be used to approximate Gibbs distribution of action labels by a product of independent marginal distribution of all actors, $Q(\mathbf{z}) = \prod_{i,j} Q_{z_{i,j}}$, each of which is obtained from the unary and pairwise energy [52]:

$$Q_{z_{i,j}} = \frac{1}{Z_{i,j}} \exp\big(- \mathbf{q}_u(i,j) - \mathbf{q}_p(i,j)\big), \tag{7}$$

in which $Z_{i,j}$ is the normalization constant [52], and $\mathbf{q}_u(i,j) = -f_u(\mathbf{b}_{i,j})$ and $\mathbf{q}_p(i,j) = -f_p(\mathbf{r}^s_{i,j} + \mathbf{r}^t_{i,j})$ are respectively the unary and pairwise energies of assigning a label $z_{i,j}$ to a node i in frame j and two labels at once to both the node and the other node connected to it. Such an inference can be implemented by iteratively stacking CNN kernels to refine the marginal distribution [16, 53, 17], which, however, do not consider temporal dependency and non-local actors' interactions. To resolve the setback, we reformulate the mean-field inference algorithm as a self-attention network as summarized in Algorithm 1, described in details below.

Multiple mean-field iterations can be implemented by refining the marginal distribution $\hat{\mathbf{Q}} = [-\mathbf{q}_p(1,1), \cdots, -\mathbf{q}_p(M,N)]$ using self-attention networks, as depicted in Fig. 4. The marginal distribution is initialized by the unary energy $\mathbf{Q}_u = [-\mathbf{q}_u(1,1), \cdots, -\mathbf{q}_u(M,N)]$, as given in Step 5 of Algorithm 1. In each iteration, the pairwise energy, $\mathbf{Q}_p = [-\mathbf{q}_p(1,1), \cdots, -\mathbf{q}_p(M,N)]$, is calculated using the spatial and temporal self-attention, as given in Steps 7 to 8, followed by compatibility transform, in Step 9, to learn the penalty of assigning labels to a pair of nodes based on their correlation. Subsequently, in Steps 10 to 11, the marginal distribution $\hat{\mathbf{Q}}$ is refined by an addition of the pairwise energy and then normalized using a softmax layer [16]. Meanwhile, the node representation is updated by combining the outputs of the temporal and spatial self-attention, as given in Step 13. Finally, in Step 14, the probability of halting the iteration is computed using (8) based on the current node representation.

Message Passing. We employ the temporal and spatial self-attention in Sec. 5.1 to connect every node by their feature similarity and address the non-local interactions of the nodes. Implementing the message passing within mean-field iterations is similar to increasing the depth of the self-attention network in [18,

Algorithm 1 Mean-field inference of the self-attention augmented CRF

Input: $\mathbf{B}$, max_iter ▷ actors' features, maximum number of iterations
Output: $\hat{\mathbf{Q}}, \bar{\mathbf{C}}$ ▷ individual action scores, final context representation
halt $= 0$ ▷ initialization of the halting probability
$v = 0$
$\mathbf{C}_v = \mathbf{B}$ ▷ initialization of the node representation
$\mathbf{Q}_u = \mathrm{f_u}(\mathbf{B})$ ▷ feed forward classifier to obtain unary energy
$\hat{\mathbf{Q}} = \mathrm{softmax}(\mathbf{Q}_u)$ ▷ initialize the marginal distribution by the unary energy
while halt ≤ 1 and $v \leq$ max_iter **do**
  Compute $S_T(\mathbf{C}_v)$ and $S_S(\mathbf{C}_v)$ ▷ message passing by temporal and spatial self-attention
  Feed-forward layer applied to $S_T(\mathbf{B})$ and $S_S(\mathbf{B})$ to obtain $\mathbf{Q}_p^t, \mathbf{Q}_p^s$
  $\mathbf{Q}_p = \mathbf{Q}_p^t \mathbf{U}^t + \mathbf{Q}_p^s \mathbf{U}^s$ ▷ compatibility transform
  $\bar{\mathbf{Q}} = \mathbf{Q}_u + \mathbf{Q}_p$ ▷ unary addition
  $\hat{\mathbf{Q}} = \mathrm{softmax}(\bar{\mathbf{Q}})$ ▷ update the marginal distribution of action labels
  $v = v + 1$
  $\mathbf{C}_v = S_T(\mathbf{C}_{v-1}) + S_S(\mathbf{C}_{v-1})$ ▷ update the node representation
  halt $= \mathrm{f_s}(\mathbf{C}_v)$ ▷ adaptive halting probability by sigmoidal function in Eqn. (8)
end while
$\bar{\mathbf{C}} = \mathbf{C}_v$ ▷ final relational context information

21]. First, denote the node feature representation at the v^{th} mean-field iteration, $\mathbf{C}_v$, as the combination of the outputs of the temporal and spatial self-attention in the previous iteration, *i.e.*, $\mathbf{C}_v = S_T(\mathbf{C}_{v-1}) + S_S(\mathbf{C}_{v-1})$, where $\mathbf{C}_0 = \mathbf{B}$. The pairwise energy by the temporal and spatial self-attention $\mathbf{Q}_p^t, \mathbf{Q}_p^s \in \mathbb{R}^{(M\times N)\times K}$, where K is the number of individual action labels, is thus obtained by propagating the temporal and spatial self-attention outputs of the current node representation, $S_T(\mathbf{C}_v)$ and $S_S(\mathbf{C}_v)$, to linear feed forward layers. Thereafter, compatibility transform is performed by multiplying $\mathbf{U}^s$ and $\mathbf{U}^t \in \mathbb{R}^{K\times K}$ with $\mathbf{Q}_p^s$ and $\mathbf{Q}_p^t$, respectively, to learn the penalty of assigning labels to a pair of nodes based on their correlation. Instead of using fixed penalty, the compatibility matrices are trained to provide data-dependent penalty, provided that different labels are assigned to nodes with high correlation.

Adaptive Mean-Field Iterations. Due to the diverse nature of actor interactions in group activities, the mean-field inference in some videos may require more iterations to converge. Consequently, instead of using a fixed number of iterations for all videos, we resort to a dynamic halting scheme, which computes the halting probability based on the current node representation. At the v^{th} iteration, the halting probability given the current node representation, $\mathbf{C}_v$, is obtained by using the sigmoid function with the corresponding weight matrix $\mathbf{W}_h$ and bias B_h [21, 56]:

$$f_s(\mathbf{C}_v) = f_s(\mathbf{C}_{v-1}) + \sigma(\mathbf{W}_h \mathbf{C}_v + B_h) \quad v \geq 1, \tag{8}$$

where $f_s(\mathbf{C}_0) = 0$. Eqn. (8) computes the accumulated halting probability up to the current iteration. Once the probability reaches one, the iteration stops and the node representation at the last iteration is considered as the final relational context information $\bar{\mathbf{C}}$. All inference parameters, including the trainable weights in (2)-(8), are updated by back propagation.

5.4 Bidirectional UTE for Group Activity Recognition

To recognize group activity in each frame, we aggregate the scene information and the relational context representation, $\bar{\mathbf{C}}$, obtained in Secs. 3 and 5.2, respectively, by UTE [21]. UTE combines the advantages of recurrent neural networks and self-attention in modelling the temporal correlation at distant positions with more flexible network depth. To this end, the relational representation in each frame is first summarized as a weighted sum of feature vectors. Suppose $\bar{\mathbf{C}}_j \in \mathbb{R}^{N\times F}$ is the relational representation in frame j, it can be aggregated as:

$$\mathbf{g}_j = \mathbf{1}^T(\bar{\mathbf{C}}_j\mathbf{q}_j)^T\bar{\mathbf{C}}_j, \quad j = 1,\ldots,M. \tag{9}$$

where $\mathbf{1}$ is an all-one vector and $\mathbf{q}_j \in \mathbb{R}^{F\times N}$ is a trainable weight. Subsequently, for positional direction consideration, we employ the bidirectional self-attention encoding to combine both of the forward and backward temporal contexts to deal with videos with similar patterns in the first few frames. More specifically, denote a concatenation of scene and context information for frame j as $\mathbf{n}_j = [\mathbf{g}_j, \mathbf{f}_j] \in \mathbb{R}^{1\times 2F}$. We can modify (2) to include positional masks as follows:

$$S_f(\mathbf{n}_j) = \mathbf{n}_j + \sum_{\forall j'}\Big(p(\mathbf{n}_j,\mathbf{n}_{j'}) \odot M^f_{j,j'}\Big)e_3(\mathbf{n}_j), \tag{10}$$

$$S_b(\mathbf{n}_j) = \mathbf{n}_j + \sum_{\forall j'}\Big(p(\mathbf{n}_j,\mathbf{n}_{j'}) \odot M^b_{j,j'}\Big)e_3(\mathbf{n}_j), \tag{11}$$

where $S_f(\cdot)$ and $S_b(\cdot)$ are the self-attention functions in the forward and backward directions, respectively, and

$$\begin{cases}\mathcal{M}^f_{j,j'} = 1, & j < j' \\ 0, & \text{otherwise}\end{cases} \quad \text{and} \quad \begin{cases}\mathcal{M}^b_{j,j'} = 1, & j > j' \\ 0, & \text{otherwise.}\end{cases} \tag{12}$$

The bidirectional representation of self-attention in UTE can be expressed as

$$S_{fb}(\mathbf{n}_j) = S_f(\mathbf{n}_j) + S_b(\mathbf{n}_j), \tag{13}$$

which is then propagated to a fully connected layer and a softmax layer to obtain the group activity distribution for every frame.

6 Experimental Results

We evaluate the performance of our proposed method on two popular group activity recognition datasets, Volleyball [5] and Collective Activity [57, 6], which provide bounding boxes and tracking annotations.

Volleyball Dataset [5]. This dataset is a collection of 55 volleyball matches, which are trimmed into 4830 short videos. Every video is composed of 41 frames and categorized into one of eight group activity classes: 'right set', 'right spike', 'right pass', 'right winpoint', 'left winpoint', 'left pass', 'left spike' and 'left set' with nine possible individual action classes for every actor: 'waiting', 'setting', 'digging', 'falling', 'spiking', 'blocking', 'jumping', 'moving' and 'standing'.

Table 1: Impact of input features. The best results are bold-faced.

Appearance	Spatial Location	Pose	Accuracy		
			Volleyball		Collective Activity
			Group Activity	Individual Action	Group Activity
✓	-	-	94.1	81.9	93.9
✓	✓	-	94.5	82.3	94.6
✓	✓	✓	**95.0**	**83.1**	**95.2**

Table 2: Performance comparison with various combination of modules. The best results are bold-faced.

Spatial Max Pooling [5]	Self-Attention Augmented CRF	UTE [21]	Bidirectional UTE	Accuracy		
				Volleyball		Collective Activity
				Group Activity	Individual Action	Group Activity
✓	-	-	-	87.1	78.4	84.6
✓	✓	-	-	94.0	82.5	93.9
-	✓	✓	-	94.6	82.8	94.8
-	✓	-	✓	**95.0**	**83.1**	**95.2**

Collective Activity Dataset [57, 6]. This dataset contains 44 untrimmed video sequences captured by hand-held cameras from crowded daily environment. There are five collective activities: 'crossing', 'waiting', 'queueing', 'walking', and talking' and six individual action labels: 'NA', 'crossing', 'waiting', 'queueing', 'walking', and 'talking'. The group activity label in one frame is decided by the largest number of the existing individual actions.

6.1 Experimental Settings

Faster R-CNN is fine-tuned with RGB images using stochastic gradient descent (SGD) and a pre-trained COCO model with a learning rate of 0.0001, a momentum of 0.9, and a decay rate of 0.1 after 160K iterations for a total of 240K iterations. The two-stream I3D is fine-tuned with a volume of 64 RGB and optical flow images [58] using Adam and a pre-trained Kinetics with a learning rate of 0.0001, a batch of 6, and a decay rate of 0.1 in every 5K iterations for a total of 20K iterations. AlphaPose [50] with a pre-trained COCO model is used to generate the keypoints. The features of the faster R-CNN, two-stream I3D, spatial location, and pose information for each actor enclosed by a bounding box from the annotation are aggregated into a feature dimension of $F = 1024$. The self-attention augmented CRF and the bidirectional UTE are jointly trained for both individual action and group activity recognition using Adam with a learning rate of 0.0005, a decay rate of 0.1 after 60 epochs, for 100 epochs. The multi-task loss function for our relational network consists of the cross entropy for group activity and individual action recognition, L_2 regularization loss [59], and the pondering time penalty for the dynamic halting [21]. Same as [18], the number of the parallel self-attention heads is 8 and the drop out rate is 0.1. To simplify the pairwise graph complexity, we set $l = \{2, 4, 6, N\}$, where $N = 12$ and 13 for Volleyball and Collective Activity, respectively. The length of temporal sliding window, M, and the maximum number of iterations of the self-attention CRF are set as 10 for both datasets. The compatibility matrices are initialized with the Potts model [16]. The experiments mainly follow the protocols and evaluation metrics provided by Volleyball [5] and Collective Activity [57, 6].

6.2 Ablation Studies

Input Features. We first scrutinize the performance with a different combination of input features as shown in Table 1, from which we can see that with the

Table 3: Impact of the number of iterations on our self-attention augmented CRF. The best results are bold-faced.

Number of Iterations	Accuracy		
	Volleyball		Collective Activity
	Group Activity	Individual Action	Group Activity
1	92.1	81.0	92.3
5	93.4	82.1	94.3
8	94.2	82.5	93.5
10	93.1	81.8	92.7
Adaptive stopping (5 methods)	**95.0**	**83.1**	**95.2**

Table 4: Comparison with the state-of-the-art methods on Volleyball and Collective Activity. The best results are bold-faced.

Method	Backbone	Accuracy		
		Volleyball		Collective Activity
		Group Activity	Individual Action	Group Activity
Discriminative Latent Models (RGB) [60]	-	-	-	79.1
Iterative Belief Propagation (RGB) [6]	-	-	-	79.6
Structure Inference Machine (RGB) [10]	AlexNet	-	-	81.2
HDTM (RGB) [5]	AlexNet	81.9	-	81.5
HiRF (RGB) [7]	-	-	-	83.1
Cardinality Potential Kernel (RGB) [8]	-	-	-	83.4
SBGAR (RGB + Flow) [9]	Inception-v3	67.6	-	86.1
SRNN (RGB) [13]	AlexNet	83.5	-	-
CERN (RGB) [11]	VGG16	83.3	69.1	87.2
StagNet w/o Attention (RGB) [1]	VGG16	87.9	81.9	87.7
StagNet w/ Attention (RGB) [1]	VGG16	89.3	-	89.1
SSU (RGB) [3]	Inception-v3	90.6	81.8	-
Recurrent Modelling (RGB + Flow) [12]	AlexNet + GoogleNet	-	-	89.4
RCRG (RGB) [2]	VGG19	89.5	-	-
ARG (RGB) [4]	Inception-v3	92.5	83.0	91.0
CRM (RGB + Flow) [15]	I3D	93.0	-	85.8
Ours (RGB + Flow)	I3D + FPN	**95.0**	**83.1**	**95.2**

addition of spatial location information, the group activity recognition performance can be improved by 0.4% and 0.7% on Volleyball and Collective Activity, respectively. This is because some volleyball activities usually have distinct actors' positions as part of the game strategies. Also, for Collective Activity, the majority of the action classes, which determines the overall group activity, are located within a close distance to each other. The performance of individual action classification on Volleyball can also be boosted by 0.4%, as some individual action classes like 'spiking' and 'waiting' have different spatial positions throughout the video. Adding pose information helps bolster both group activity and individual action recognition performance on Volleyball by 0.5% and 0.8%, respectively. Also, the group activity classification on Collective Activity is improved by 0.6%. This is because our CRF can learn the dynamic changes of the actors' poses and their relation with the other actors that are crucial in determining the group activity and the individual action categories. Consequently, we employ all three features in the following simulations.

Functions of Modules. We assess the effect of the modules in our relational network, as shown in Table 2, from which we can see that compared with using only the node features by the feature extraction network, the proposed self-attention augmented CRF can improve the individual action recognition performance by 4.1% on Volleyball. The accuracy of group activity classification is also boosted by 6.9% and 9.3% on Volleyball and Collective Activity, respectively, compared to applying max pooling directly to aggregate the actors' features in every frame. This is because it can leverage the spatial relational context and temporal evolution of every actor to precisely infer the group activity category. Lastly, we assess the impact of bidirectional UTE. We can see that using the uni-

directional UTE [21], the group activity recognition performance can be slightly improved by 0.6% and 0.9% on Volleyball and Collective Activity, respectively. The performance gain can be further enhanced respectively by 0.4% on both of the datasets with a replacement of the bidirectional UTE.

Number of Iterations. We inspect the significance of using the adaptive number of iterations by using the dynamic halting scheme instead of a fixed one. As shown in Table 3, the performance of the individual action and group activity recognition on Volleyball is improved with the number of iterations, but it begins to drop after it reaches 8 iterations. Similarly, on Collective Activity, the accuracy of group activity classification stops to improve after 5 iterations. This is because different videos require different numbers of iterations and increasing the number of iterations can lead to overfitting. Meanwhile, the proposed adaptive inference, which in average converges in 5 iterations and 3 iterations for Volleyball and Collective Activity, respectively, achieves the best performance.

6.3 Comparison with the State-of-the-Art Works

We first compare the proposed method with state-of-the-art works, including HDTM [5], SBGAR [9], SRNN [13], CERN [11], StagNet [1], Recurrent Modelling [12], SSU, [3], RCRG [2], ARG [4], and CRM [15], on Volleyball in terms of the group activity and individual action recognition accuracy, as shown in Table 4, from which we can see that SBGAR [9] is the worst as it relies on the high-level semantic caption data. This approach [9] is inferior to HDTM [5], CERN [11], and SRNN [13], which aggregate individual actions using two-level of RNN. Considerable improvement is attained by StagNet [1] as it models the relationships among the actors using semantic attentive graphs. RCRG [2] achieves similar performance by utilizing stacks of relational encoders. SSU [3] has even better performance by multi-task learning of action detection and group activity recognition. ARG [4] outperforms [3] by constructing an actor relational graph based on self similarity. Without explicitly learning individual actions, CRM [15] yields slightly better results by incorporating multi-stage refinement. Our work outperforms all of the aforementioned methods by learning the spatial relation and the temporal evolution of actors using the self-attention endowed CRF. Also, our approach excels the state-of-the-art works that reported their performance on the individual action recognition. This is because our self-attention augmented CRF can model the non-local relationships and temporal dependency of the actors. Some visualization results are shown in Fig. 5 (a), from which we can observe that the the error comes from differentiating between 'setting' and 'passing' that have similar actors' movements.

Next, we compare the group activity recognition with twelve baselines, Discriminative Latent Models [60], Iterative Belief Propagation (RGB) [6], Structure Inference Machine [10], HiRF [7], Cardinality Potential Kernel[8], HDTM [5], SBGAR [9], CERN [11], StagNet [1], Recurrent Modelling [12], ARG [4], and CRM [15], on Collective Activity. As shown in Table 4, CRM [15] has superior performance compared with non-deep networks [60, 6–8] and relatively shallow networks [10, 5] by learning the group activity from multi-stage convolutional maps. SBGAR [9] outperforms [15] by capturing the dynamics of the semantic

Fig. 5: Some group activity and individual action recognition results by our method, where the individual action classification is in white while the incorrect classification is marked by a red cross.

caption data. CERN [11] achieves even better performance by using an energy based optimization scheme for training. StagNet outperforms [11] by incorporating an attention mechanism to determine important actor features for more precise group activity recognition. Recurrent Modelling [12] yields even better results as it uses multi-context information. ARG [4] attains higher accuracy by constructing an actor relational graph based on self-similarity. Our method outperforms all other works by modelling the spatial relation and temporal evolution of the actors using the self-attention strengthened CRF. Some visualization results are depicted in Fig. 5 (b), from which we can see that 'waiting' can be easily confused with 'crossing' and 'walking' as it has similar appearances.

7 Conclusions

This paper has developed an efficacious relational network for group activity recognition. Our network first utilizes a self-attention augmented CRF to learn the spatial relational context and temporal evolution of every actor. Such a combination explores actors' interactions from a graph topology with different localities. Next, an adaptive mean-field inference is addressed. Finally, a bidirectional UTE is used to amass the actors' relational context and scene information. Simulations show the effectiveness of the new approach on two common datasets.

Acknowledgement This work was supported by the Ministry of Science and Technology and by ITRI, R.O.C., under contracts MOST 109-2221-E-011-131 and 109-2221-E-011-116, and CCL/ITRI B5-10903-HQ-07.

References

1. Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. StagNet: An attentive semantic RNN for group activity recognition. In *Proceedings of the European Conference on Computer Vision*, pages 101–117, 2018.
2. Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 721–736, 2018.
3. Timur Bagautdinov, Alexandre Alahi, Francois Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4315–4324, 2017.
4. Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9964–9974, 2019.
5. Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1980, 2016.
6. Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Proceedings of the European Conference on Computer Vision*, pages 215–230, 2012.
7. Mohamed Rabie Amer, Peng Lei, and Sinisa Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *Proceedings of the European Conference on Computer Vision*, pages 572–585, 2014.
8. Hossein Hajimirsadeghi, Wang Yan, Arash Vahdat, and Greg Mori. Visual recognition by counting instances: A multi-instance cardinality potential kernel. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2596–2605, 2015.
9. Xin Li and Mooi Choo Chuah. Sbgar: Semantics based group activity recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2876–2885, 2017.
10. Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4772–4781, 2016.
11. Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. CERN: confidence-energy recurrent network for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5523–5531, 2017.
12. Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3048–3056, 2017.
13. Sovan Biswas and Juergen Gall. Structural recurrent neural network (SRNN) for group activity analysis. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1625–1632, 2018.
14. Judith Butepage, Michael J. Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6158–6166, 2017.

15. Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2019.
16. Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015.
17. Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2017.
18. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
19. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
20. Richard D Alba. A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, 3(1):113–126, 1973.
21. Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. In *Proceedings of the International Conference on Learning Representations*, 2019.
22. Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Proceedings of the Neural Information Processing Systems*, pages 568–576, 2014.
23. Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.
24. Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. Spatiotemporal residual networks for video action recognition. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 3468–3476, 2016.
25. Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international Conference on Computer Vision*, pages 4489–4497, 2015.
26. Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.
27. Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
28. Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019.
29. Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision*, pages 305–321, 2018.

30. Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *Proceedings of the European Conference on Computer Vision*, pages 90–105, 2018.
31. Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European Conference on Computer Vision*, pages 417–432, 2018.
32. Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3085–3094, 2019.
33. Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.
34. Chen Change Loy, Tao Xiang, and Shaogang Gong. Modelling activity global temporal dependencies using time delayed probabilistic graphical model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 120–127, 2009.
35. Eran Swears, Anthony Hoogs, Qiang Ji, and Kim Boyer. Complex activity recognition using granger constrained DBN (GCDBN) in sports and surveillance video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 788–795, 2014.
36. Yongyi Lu, Cewu Lu, and Chi-Keung Tang. Online video object detection using association lstm. In *Proceedings of The IEEE International Conference on Computer Vision*, pages 2344–2352, 2017.
37. Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. LSTM pose machines. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pages 7852–7861, 2018.
38. Toby Perrett and Dima Damen. DDLSTM: Dual-domain LSTM for cross-dataset action recognition. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pages 5207–5215, 2019.
39. Didik Purwanto, Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. Three-stream network with bidirectional self-attention for action recognition in extreme low resolution videos. *IEEE Signal Processing Letters*, 2019.
40. Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.
41. Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*, 2019.
42. Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.
43. Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. Hierarchical self-attention network for action localization in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 61–70, 2019.
44. Vlad I Morariu and Larry S Davis. Multi-agent event recognition in structured scenarios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3289–3296, 2011.
45. Stephen S Intille and Aaron F Bobick. Recognizing planned, multiperson action. *Computer Vision and Image Understanding*, 81(3):414–445, 2001.
46. Yuanlu Xu, Lei Qin, Xiaobai Liu, Jianwen Xie, and Song-Chun Zhu. A causal and-or graph model for visibility fluent reasoning in tracking interacting objects. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2178–2187, 2018.
47. Wei-Hong Li, Fa-Ting Hong, and Wei-Shi Zheng. Learning to learn relation for important people detection in still images. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 5003–5011, 2019.
48. Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision*, pages 399–417, 2018.
49. Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.
50. Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. pages 2334–2343, 2017.
51. John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*, 2001.
52. Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 109–117, 2011.
53. Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr. Higher order conditional random fields in deep neural networks. In *European Conference on Computer Vision*, pages 524–540, 2016.
54. Buyu Liu and Xuming He. Learning dynamic hierarchical models for anytime scene labeling. In *European Conference on Computer Vision*, pages 650–666, 2016.
55. Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019.
56. Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.
57. Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Proceedings of the International Conference on Computer Vision Workshops*, pages 1282–1289, 2009.
58. Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Proceedings of the Joint Pattern Recognition Symposium*, pages 214–223, 2007.
59. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.
60. Tian Lan, Yang Wang, Weilong Yang, Stephen N Robinovitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1549–1562, 2011.