Optimised ARG based Group Activity Recognition for Video Understanding

Pranjal Kumar*

Abstract In this paper, we propose a robust video understanding model for activity recognition by learning the actor's pair-wise correlations and relational reasoning, exploiting spatial and temporal information . In order to measure the similarity between the pair appearances and construct an actor relations map, the Zero Mean Normalized Cross-Correlation (ZNCC) and the Zero Mean Sum of Absolute Differences(ZSAD) are proposed to allow the Graph Convolution Network (GCN) to learn how to distinguish group actions. We recommend that MNASNet be used as the backbone to retrieve features. Experiments shows 38.50% and 23.7% reduction in training time in 2-stage training process along with 1.52% improvement in accuracy against traditional methods.

Keywords group activity recognition · graph convolution network · video understanding · video analytics · activity recognition

1 Introduction

Precise video understanding involves consideration of links between actors, objects and their surroundings, often over long time periods. One reason why video comprehension is so difficult is because it requires an understanding of the interactions between actors, objects and other contexts on the scene. Moreover, these interactions cannot always be seen from a single frame, and therefore require reasoning over long periods of time. As such, some of them only model spatial relationships between actors and objects, but not the evolution of those interactions with time. Alternative approaches model long-range time interactions[1], but do not capture and do not train spatial relations. Although certain methods model spatio-temporary interactions between objects[2, 3], further supervision is required for their explicit representations of objects. Early works in this field included modelling human-object interaction[4, 5], various objects[6], and human actions/scenes context relations[7, 8]. Moreover, human vision has also proven to be context dependent[9].

A major problem in video understanding is recognition of human action and recognition of group activities[10]. The techniques of action and activity recognition have been widely used, for example in the fields of social behavior understanding, sport video analysis and video monitoring. It is important to better understand a video scene with several people and to understand the action and collective activity of all individuals. The group activity recognition based on the Actor Relation Graph (ARG) is the state-of-the-art model that aims to capture the appearance and position of the actors in the scene, and to identify the action and the group activity[10]. Z Kuang et al.[11](2020) further attempted to improve the functioning of the ARG module. However, real-time aspect of video understanding was not considered (e.g., noise, abrupt illumination changes, large pixel differences.)

In this paper, we suggest several ways to improve the functionality and efficiency of the Actor Relation Graph model in order to improve video understanding not only to be tested on available dataset but also considering the real world aspect of it. We use MNASNet[12] in the

Pranjal Kumar
NIT Hamirpur, H.P, India-177005
Tel.: +918637511985
Fax: +91-1972-223834
E-mail: pranjal@nith.ac.in

CNN layer to increase the performance of human action and group activity recognition. With 78ms of latency on a pixel-phone, MnasNet achieves 75.2% top- 1 precision, 1.8 times faster than MobileNetV2[13] with an increase of 0.5% and 2.3 times faster than NASNet[14] with a higher accuracy of 1.2%. Our experimental results show that optimised ARG is able to achieve superior performance compared to all previous approaches. Main contribution of the proposed methodology is summarized below:

- We propose to incorporate MNASNet for significant reduction in training time while maintaining similar accuracy .
- Our method leverages ZNCC and ZSAD template matching algorithms in order to improve the accuracy for group activity recognition.

2 Related Work

2.1 Action Recognition

In earlier works, hand-crafted attributes for encoding information from motion were used[15,16]. Advances in deep learning saw first the repurposes of video ”two stream” networks of 2D image-convolutionary neural networks (CNNs)[17,18], and then the space-time 3D CNN[19–22]. These architectures, however, concentrate on extracting broad features, video-based features and are not suitable for studying fine grain relationships. Graph neural network (GNN), by modelling them as nodes in a directed, undirected graph, explicitly models the interaction between entities[23–25] through a neighbourhood defined in each node. Each feature maps element in each function is a node, and all nodes are fully connected. The self-attention [26] and non-local operators [27] are also considered GNNs. Such models have been outstanding in several processing tasks for the natural language and informatics, inspiring numerous follow-up methods[28–32].

2.2 Human Object Interaction

The objective of Human Object Interaction (HOI) detection is to locate humans and objects and to recognise their interactions. Previous studies[33–38] show promising results of HOI sensing by decoupling it into the detection and classification of objects. In particular, the results of human and object detection first come from an object detector pre-trained, and then a pair of combined proposals for human objects interaction classification. In recent approaches[39, 38, 37], a substitute detection problem was introduced, which would indirectly optimise the HOI detection. Firstly, the proposal of interaction was predefined on the basis of human priors. UnionDet[38], for example, defines the proposal for interactions as a union box for human and object boxes. As an interaction point, the central point from the human to the object is used by PPDM[37].

2.3 Group Activity Recognition

Group recognition of activity is a major problem for the understanding of video[40, 18, 41, 42] and has a wide variety of practical applications, such as monitoring, sports video and social competencies. The model must not only outline the individual ongoing actions in the scene, but also takes into account for their collective activity to understand the scene of multiple individuals. For understanding multiple persons' group activities the ability to capture correctly the corresponding relationships among actors and conduct relation reasoning is crucial[43–47]. The relationship between actors from other aspects, including looks and relative location, is expected to be deduced. Therefore, when designing effective profound models for group activity understanding, these two important indications are necessary.

Recent methods of deep learning have demonstrated promising results on the recognition of group activities in videos[48, 49]. These methods usually follow a two-stage pipeline of recognition. Zijian et al. has proposed to use MobileNet to extract features from each video frame[11]. First, a convolutionary neural network is used to extract personal features (CNN). A global module is then designed to add these representations to provide a scene level functionality. The relationship between these actors is modelled by existing methods with a rigid graphical model [46] whose structure is predetermined manually, or by using a complex but unintuitive message passing mechanism[50,49].

2.4 Actor Relation Graphs for Group Activity Recognition

A major problem in video understanding is recognition of human action and recognition of group activities[10]. J Wu et al. suggested that Actor Relational Graph (ARG) be used to model actor-to-actor relationships and so that group activity learning be recognized with multiple participants[10]. The relationship between actors from a similarity to the appearance and the relative location is determined and captured using the ARG in a multi-person scene. When compared to the use of CNN to extract person-level features and then add them into a scene-level feature, or the application of RNN to collect temporal information from densely sampled frames, the computationally costly and flexible way of learning with ARG is less while addressing variations in group activity. With a video sequence with bounds for actors in the scene, the trained network can recognize individual actions and group activity in a multiperson scene. The network can also recognize action. ARG efficiency is improved for long range video clips by forcing a relational connection in a local neighborhood alone and by dropping several frames alone while maintaining the variations in training sample data and minimizing the risk of overfitting. Initially, the features of the actors will be extracted from the provided bounding boxes by model CNN and RoIAlign[51]. After the feature vectors of actors in the scene have been obtained, several graphs are created to represent the diverse information of the same set of actors. Finally, the GCN is designed to perform learning in order to identify the actions and activities of individual groups based on ARG. The pooled ARG applies respectively to two classifiers used for each action and group recognition activity. The representation in scene-level is generated by maxpooling individual actors, that later use them for classification of groups of activities.

3 Methodology

J. Wu et al. has shown that in each frame, the ARG can represent the graphical structure of the information in pairs between the pairs of actors, and use the related information to understand group activity[10]. Both features of appearance and position information are used to construct the ARG to better understand the relation between two actors. The value of the relationship is defined as a composite function f_a which indicates the relationship of appearance, and f_s indicates position relationship. x_i^a and x_j^a refers to the appearance features of the actor i and actor j while x_i^s and x_j^s refers to the location features (the center of the bounding box of each actor) of the actor i and actor j. Function h combines appearance along with position to scalar weight[10].

$$G_{ij} = h\left(f_a\left(x_i^a, x_j^a\right), f_s\left(x_i^s, x_j^s\right)\right) \tag{1}$$

The normalisation of every actor node with SoftMax function is further adopted in order to ensure that the sum of all values corresponding to each actor node is always one[10]:

$$\mathbf{G}_{ij} = \frac{f_s\left(\mathbf{x}_i^s, \mathbf{x}_j^s\right)\exp\left(f_a\left(\mathbf{x}_i^a, \mathbf{x}_j^a\right)\right)}{\sum_{j=1}^{N} f_s\left(\mathbf{x}_i^s, \mathbf{x}_j^s\right)\exp\left(f_a\left(\mathbf{x}_i^a, \mathbf{x}_j^a\right)\right)} \tag{2}$$

3.1 Appearance Relation

The Embedded Dot-Product is used in J Wu's paper to calculate the similitude between the appearance features of both actors (the image within each actor's boundary box) in the space where they are embedded[10]: The following is the corresponding function:

$$f_a\left(\mathbf{x}_i^a, \mathbf{x}_i^a\right) = \frac{\theta\left(\mathbf{x}_i^a\right)^{\mathsf{T}} \varphi\left(\mathbf{x}_j^a\right)}{\sqrt{d_k}} \tag{3}$$

Both θ and φ are functions utilizing $Wx + b$, where W nad b are learnable weights. The learned changes with respect to the original features can better comprise the correlation among two actors in a subspatial environment. Two other methods for appearance calculation are: Zero Mean Normalized Cross-Correlation (ZNCC) and Zero Mean Sum of Absolute Differences (ZSAD).

The NCC and ZNCC are easy to assess because they range from 1 to 1 (respectively the worst and best matching). As such, different attempts of correspondence can be compared. But the calculation can be dominated by large pixel differences. In the lighting changes, it is expected that ZNCC will perform better than NCC[52]. The advantage of the normalized cross-relation is that it is less sensitive to linear changes in the illumination amplitude of the two comparative images and can be written in the following way[51]:

$$\phi^I_{x_i^a x_j^a}(t) = \frac{\phi_{x_i^a x_j^a}(t)}{\sqrt{\phi_{x_i^a x_i^a}(0)\phi_{x_j^a x_j^a}(0)}} \tag{4}$$

The quantity $\phi^I_{x_i^a x_j^a}(t)$ vary between -1 to 1. The value of the NCC can help us better understand the relationship of appearance between the two actors. Another method we evaluate in order to calculate the relationship of appearance of ARG is the sum of absolute differences (SAD). By calculating the amount of absolute difference between the matrix components as the formula, the SAD computes the distance between two matrices:

$$SAD\left(x_i^a, x_j^a\right) = \sum_{k}^{n} \left|x_{ik}^a - x_{jk}^a\right| \tag{5}$$

SAD is stronger against extreme data values, making it robust when comparing appearance characteristics and better capturing the relationship between appearances. ZSAD algorithm uses about 30 % more hardware than SAD, but it is preferred to prevent the random distortion of non-ideal stereo cameras[53].

doi:10.20944/preprints202106.0313.v1

3.2 Position Relation

Furthermore, spatial structural information is considered to better capture the relationship between actors. To obtain signals from entities not remotely distant, a distancing mask has been applied. As relation is crucial to the understanding of group activities in a local context in comparison to the global relationship, an evaluation of Euclidean distance G_{ij} between two actors is taken as follows[10]:

$$f_s \ \ \mathbf{x}^s_i, \mathbf{x}^s_j \ \ = I \ \ d \ \ \mathbf{x}^s_i, \mathbf{x}^s_j \ \ \leq \mu \tag{6}$$

Where $I(\cdot)$is the indicator function and $d\ \mathbf{x}^s, \mathbf{x}^s{}_i \ \ {}_j$ evaluates the Euclidean distance between the center coordinates of two actor's bounding boxes, μ is a distance threshold.

4 Proposed Method

We propose to use an optimised model based on the Actor relations graph, which focuses on recognition of group activity.

Firstly, the model extracts the actor from sampled video frames with CNN and RoIAlign bounding boxes[54]. It then builds a N by d feature matrix with a d dimensional vector to represent each bounding box of the actor and N to present the number of video frame bounding boxes. The graphs of the actor relation are then constructed to capture the appearance and position of each actor in the scene. The model will then analyze from the ARG the relations of each actor using Graph Convolutional Networks (GCN). Finally, two distinct classifiers aggregate and utilize the original and relational features to carry out actions and recognition of group activities. As the study focuses mainly on the recognition of group activities.

Although high accuracy predictions of group activities can be made using the ARG model. Some areas for improvement still exist. We proposed to use MNASNet for extracting image characteristic maps in CNN and to calculate the pair-wise appearance similarity for the Actor Relation Graph by using Zero mean normalized cross-correlation (ZNCC) and the Zero mean sum of absolute differences (ZSAD). ZNCC enables linear brightness differences to be tolerated. In addition, the ZNCC is more robust than the NCC because of the removal of the local mean[55]. It is easy for NCC and ZNCC to be evaluated as they vary between 1 and 1 (the worst and best matches, respectively); as such, different correlation efforts can also be compared. However, the calculation can be dominated by large pixel differences (e.g., sudden scene change in real-time feed from surveillance). In terms of light change, ZNCC is expected to perform better than NCC[52]. ZSAD algorithm uses about 30 % more hardware than SAD, but it is preferred to prevent the random distortion of nonideal stereo cameras[53]. In order to give our model a more visualized result we also use a visualization model to display every video frame with bounding boxes[11] on each human object. Fig 1 shows the output examples.

5 Results & Discussion

5.1 Implementation Details

The minibatch size is 16, the learning rate is 0.0001 and we train our network through 100 epochs. The single action weight loss $\lambda = 1$ is used. The GCN parameters are determined as $d_k = 256$, $d_s = 32$, and the distance mask threshold μ is adopted as 1/2 of the image width. The default CNN network backbone for function extraction is set to Inception-v3, and an embedded dot-product is the default appearances relating function. The PyTorch framework and individual instance of Tesla K80 GPU are the basis for our implementation.

5.2 Experiments

- *Experiment 1: Evaluation with different backbone networks*
  In this section we conduct extensive studies on the dataset of the collective activities, in order to understand the modeling relationship of the proposed backbone network using group activity prediction precision as the assessment metric. Table 1 presents the results of the experiments.

Table 1 Group Activity Prediction from Two Backbone Networks

Backbone Network	Stage	Time(seconds)	Accuracy
Inception-v3	1	25035	91.03%
Inception-v3	2	25360	92.82%
MNASNet	1	15395	88.63%
MNASNet	2	19332	91.37%

During our two-stage training, the ImageNet pretrained backbone network is first finalized with each of the samples randomly selected from the training. Then, in stage 2 the weights of the backbone network feature extraction are fixed. We continue training the GCN network and use the embedded dot-product to calculate the appearance relationship.

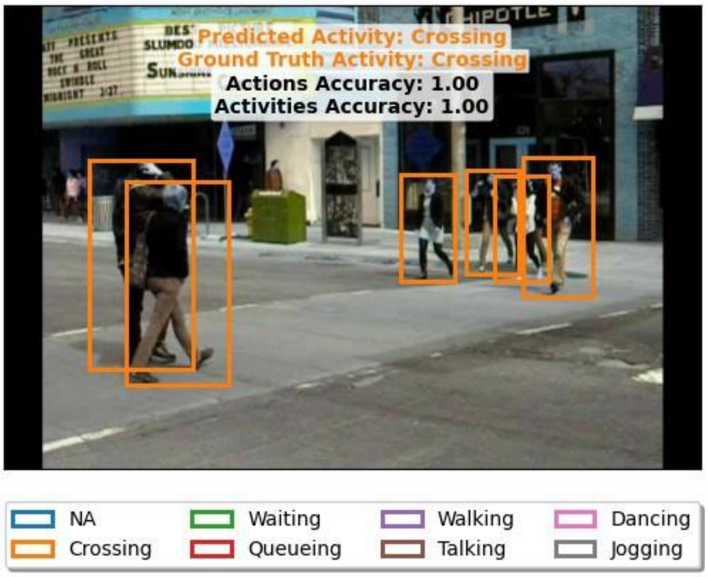
Predicted Activity: Crossing
Ground Truth Activity: Crossing
Actions Accuracy: 1.00
Activities Accuracy: 1.00
NA
Waiting
Walking
Dancing
Crossing
Queueing
Talking
Jogging

Fig. 1

Optimized ARG-based human actions and group activity recognition

Inception v3 as a backbone network is the start of our experiments. It takes about 6.95 hours to complete the first training stage, while about 7.04 hours to complete the second stage. Inception v3 achieves the group activity accuracy of 91.03% after stage 1 training. Our model delivers a higher recognition accuracy of 92.82% with additional training of GCN in stage 2.
We also use MNASNet as the backbone network to increase our model speed. MNASNet is a low-weight but efficient deep convolution neural network. The training time for stage 1, which reduces the time used at stage 1 by 38.50%, is 4.27 hours with MNASNet. The training time of stage 2 is 5.37 hours, 23.7% below the training time of Inceptionv3 at the same time. However, the accuracy of activity recognition decreases slightly from 92.82% to 91.37%.

- *Experiment 2: Evaluation with different appearance relation functions*
  In this experiment, we analyze the performance of the group activity recognition with different functions. The group activity recognition performance is first trained and validated on the default Inceptionv3 backbone and the embedded dot-product for calculating appearance-relation. We have 92.06 % as the best result. The appearance relation function is then updated to the zero mean normalized cross correlation (ZNCC), and the best outcome is 93.58 %. The zero mean sum of the absolute distance (ZSAD) function is further evaluated in order to calculate appearance similarity and 94.37 % is the best result we can achieve. Similar experimentation is then conducted with MNASNet as backbone. In this Experiment 2, we prove to be more accurate in our proposed model with the ZNCC or ZSAD as an apparence-related calculative function. Table 2 shows the results of the experiments. Fig 1 shows the output from visualization model.

6 Ablation Study

In this section, we conduct a ablation study to understand the contributions to the relationship modelling by using the accuracy of group activity and optimal number of graphs as an evaluation metric.

6.1 Appearance Relational Function

For the computation of appearance relation value, we conduct the experimentation by evaluating the ramifications of modelling the relationship between actors in the scene and different functions. We build a single ARG without position relationships on the basis of a single frame. Table 2 shows the results. First, we note that modelling the actors' relationship explicitly improves performance substantially. All MNASnet models exceed the basic model.

Table 2 Group Activity Prediction on Backbone Network Inception-v3(I) and MNASNet(M) with Different Appearance Relation Functions with time cost(in seconds).

Method	Time(I)	Time(M)	Accuracy
Embedded Dot-product	26357	17874	92.06%
ZNCC	25560	16386	93.58%
ZSAD	27704	23564	93.37%

6.2 Number of Graphs

As shown in Table 3, the construction of several graphs activate agreeing and notable progress over only building one graph and further improves its accuracy from 90.3% to 91.8%.

Table 3 Variation with total number of graphs

Number	1	4	8	16	32
Accuracy	90.3%	91.2%	91.4%	91.8%	91.7%

7 Comparative Study with State of The Art

In this section, we provide the results our best models with the most advanced methods. We examined our results with both Inception-v3 and MNASnet backbone networks for a fair comparison with previous methods.We could not conduct the experimentation on the volleyball dataset[56] due to memory constraints of the system. We conduct a proposition-based experiment in the meantime.

Table 4 The Collective Activity dataset comparison with the state of the art[10]

Method	Backbone Net.	Accuracy
SIM[57]	AlexNet	81.2%
stagNet (PRO)[49]	VGG-16	87.9%
HDTM [46]	AlexNet	81.5%
stagNet (GT)[49]	VGG-16	89.1%
Cardinality Kernel[58]	n/a	83.4%
CERN [59]	VGG-16	87.2%
SBGAR[60]	Inception-v3	86.1%
OURS	MNASNet	91.3%
OURS	Inception-v3	92.8%

Proposed method shows encouraging results over all existing methods. Our approach with MNASNet uses the same feature extraction strategy as[48], which is about 2% superior in group activity recognition precision, as the relationship information between actors can be captured and exploited by our model. This is mainly because we explicitly model and adopt a more efficient ZNCC and ZSAD similarity matching functions, which is not using any hierarchical relationship networks[46] or semantian RNN[49] as in ARG to build relationship between actors.

The proposed model for the collective activity dataset is further evaluated. Table 4 lists the results and comparisons with previous methods[10]. Again, with 91.3 % group activity recognition accuracy our model achieves state of the art performance. The efficiency and generality of the proposed method are demonstrated by these outstanding results.

8 Conclusion & Future Work

This paper uses the model based on the actor relation graph (ARG) for the recognition of group activity. To enhance the performance of our model, we learn ARG to use zero-mean normalised cross-correlation (ZNCC) and the zero mean sum of absolute difference in the graphs (ZSAD). We also have MNASNet as the backbone network in our proposed model to improve computational speed. In addition, extensive experiments show that the suggested methods for improving precision and speed on the Collective Activity dataset are robust and effective.

In future, we would further attempt to enhance the accuracy for activity recognition by applying skeleton extraction framework (incorporating adaptive sampling and consideration of neighbouring points) as current ARG module use randomised sampling i.e, *randomized* ARG and fine tuning MNASNet. After GCN there remains an open question as to how to fuse a group of graphs together for better group activity recognition.

References

1. Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.
2. Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018.
3. Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–121, 2018.
4. Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(10):1775–1789, 2009.

5. Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24. IEEE, 2010.
6. Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.
7. Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936. IEEE, 2009.
8. Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015.
9. Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007.
10. Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9964–9974, 2019.
11. Zijian Kuang and Xinran Tie. Improved actor relation graph based group activity recognition, 2020.
12. Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile, 2019.
13. Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
14. Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.
15. Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013.
16. Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64(2-3):107–123, 2005.
17. Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.
18. Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.
19. 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.
20. R Christoph and Feichtenhofer Axel Pinz. Spatiotemporal residual networks for video action recognition. *Advances in Neural Information Processing Systems*, pages 3468–3476, 2016.
21. Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.
22. Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.
23. Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
24. Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.
25. Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
26. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
27. 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.
28. Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
29. Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. $a\hat{}2$-nets: Double attention networks. *arXiv preprint arXiv:1810.11579*, 2018.
30. Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2019.
31. Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*, 2019.
32. Li Zhang, Dan Xu, Anurag Arnab, and Philip HS Torr. Dynamic graph message passing networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3726–3735, 2020.
33. Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018.
34. Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020.
35. Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018.
36. Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020.
37. Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 482–490, 2020.

38. Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, pages 498–514. Springer, 2020.
39. Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020.
40. Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1430–1439, 2018.
41. 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.
42. Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alex G Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2568–2577, 2015.
43. Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1354–1361. IEEE, 2012.
44. Mohamed Rabie Amer, Peng Lei, and Sinisa Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *European Conference on Computer Vision*, pages 572–585. Springer, 2014.
45. Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pages 1282–1289. IEEE, 2009.
46. Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 721–736, 2018.
47. Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023. PMLR, 2016.
48. Timur Bagautdinov, Alexandre Alahi, François 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.
49. 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 (ECCV)*, pages 101–117, 2018.
50. Sovan Biswas and Juergen Gall. Structural recurrent neural network (srnn) for group activity analysis. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1625–1632. IEEE, 2018.
51. Y Raghavender Rao, Nikhil Prathapani, and E Nagabhooshanam. Application of normalized cross correlation to image registration. *International Journal of Research in Engineering and Technology*, 3(5):12–16, 2014.
52. Niccolò Dematteis and Daniele Giordan. Comparison of digital image correlation methods and the impact of noise in geoscience applications. *Remote Sensing*, 13(2):327, 2021.
53. Vaddi Chandra Sekhar, Satyajit Bora, Monalisa Das, Pavan Kumar Manchi, S Josephine, and Roy Paily. Design and implementation of blind assistance system using real time stereo vision algorithms. In *2016 29th International Conference on VLSI Design and 2016 15th International Conference on Embedded Systems (VLSID)*, pages 421–426. IEEE, 2016.
54. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
55. Luigi Di Stefano, Stefano Mattoccia, and Federico Tombari. Zncc-based template matching using bounded partial correlation. *Pattern recognition letters*, 26(14):2129–2134, 2005.
56. 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.
57. 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.
58. Hossein Hajimirsadeghi and Greg Mori. Multi-instance classification by max-margin training of cardinality-based markov networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1839–1852, 2016.
59. 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.
60. 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.