Hunting Group Clues with Transformers for Social Group Activity Recognition Supplementary Material

Masato Tamura, Rahul Vishwakarma, and Ravigopal Vennelakanti

Hitachi America, Ltd.
masato.tamura.sf@hitachi.com,
{rahul.vishwakarma,ravigopal.vennelakanti}@hal.hitachi.com

A. Individual Recognition

In our method, individuals are recognized by simply adding an action classification head to the detection heads in Deformable DETR [25]. Given a set of feature embedding $\boldsymbol{H} = \{\boldsymbol{h}_i \mid \boldsymbol{h}_i \in \mathbb{R}^{D_p}\}_{i=1}^{N_q}$ from the deformable transformer decoder, the predictions of person class probabilities $\{\hat{c}_i \mid \hat{c}_i \in [0,1]\}_{i=1}^{N_q}$, bounding boxes $\{\hat{\boldsymbol{b}}_i \mid \hat{\boldsymbol{b}}_i \in [0,1]^4\}_{i=1}^{N_q}$, and action class probabilities $\{\hat{\boldsymbol{a}}_i \mid \hat{\boldsymbol{a}}_i \in [0,1]^{N_a}\}_{i=1}^{N_q}$ are obtained as $\hat{c}_i = f_c(\boldsymbol{h}_i)$, $\hat{\boldsymbol{b}}_i = f_b(\boldsymbol{h}_i, \boldsymbol{r}_i)$, and $\hat{\boldsymbol{a}}_i = f_a(\boldsymbol{h}_i)$, where N_q is the number of query embeddings, N_a is the number of action classes, $f_c(\cdot)$, $f_b(\cdot,\cdot)$, and $f_a(\cdot)$ are the detection heads for the predictions, and $\boldsymbol{r}_i \in [0,1]^2$ is a reference point, which is used in the same way as the localization in Deformable DETR. Note that the localization results are denoted in the normalized image coordinates.

We view individual recognition as a direct set prediction problem and match predictions and ground truths with the Hungarian algorithm [10] during training. The optimal assignment of ground truths and predictions is determined by calculating the matching cost with the predicted person class probabilities, bounding boxes, and action class probabilities. Given a ground truth set of individual recognition, the set is first padded with $\phi^{(id)}$ (no person) to change the size of the set to N_q. Using the padded ground truth set, the matching cost of the i-th element in the ground truth set and j-th element in the prediction set for individual recognition is calculated as follows:

$$\mathcal{H}_{i,j}^{(id)} = \mathbb{1}_{\{i \notin \boldsymbol{\Phi}^{(id)}\}} \left[\eta_c \mathcal{H}_{i,j}^{(c)} + \eta_b \mathcal{H}_{i,j}^{(b)} + \eta_o \mathcal{H}_{i,j}^{(o)} + \eta_a \mathcal{H}_{i,j}^{(a)}\right], \tag{1}$$

$$\mathcal{H}_{i,j}^{(c)} = -\hat{c}_j, \tag{2}$$

$$\mathcal{H}_{i,j}^{(b)} = \left\|\boldsymbol{b}_i - \hat{\boldsymbol{b}}_j\right\|_1, \tag{3}$$

$$\mathcal{H}_{i,j}^{(o)} = -f_{GIoU}\left(\boldsymbol{b}_i, \hat{\boldsymbol{b}}_j\right), \tag{4}$$

$$\mathcal{H}_{i,j}^{(a)} = -\left(\frac{\boldsymbol{a}_i^T \hat{\boldsymbol{a}}_j + (\boldsymbol{1} - \boldsymbol{a}_i)^T (\boldsymbol{1} - \hat{\boldsymbol{a}}_j)}{N_a}\right), \tag{5}$$

where $\boldsymbol{\Phi}^{(id)}$ is a set of ground-truth indices that correspond to $\phi^{(id)}$, $\boldsymbol{b}_i \in [0,1]^4$ is a ground truth bounding box normalized with the image size, $\boldsymbol{a}_i \in \{0,1\}^{N_a}$ is a ground truth action label, $f_{GIoU}(\cdot,\cdot)$ is a function that calculates generalized IoU [16], and $\eta_{\{c,b,o,a\}}$ are the hyper-parameters. The Hungarian algorithm is applied to the matching cost to find the optimal assignment $\hat{\omega}^{(id)} = \arg\min_{\omega \in \boldsymbol{\Omega}_{N_q}} \sum_{i=1}^{N_q} \mathcal{H}^{(id)}_{i,\omega(i)}$, where $\boldsymbol{\Omega}_{N_q}$ is the set of all possible permutations of N_q elements.

The training loss for individual recognition $\mathcal{L}_{id}$ is calculated between matched ground truths and predictions as follows:

$$\mathcal{L}_{id} = \lambda_c \mathcal{L}_c + \lambda_b \mathcal{L}_b + \lambda_o \mathcal{L}_o + \lambda_a \mathcal{L}_a, \tag{6}$$

$$\mathcal{L}_c = \frac{1}{|\bar{\boldsymbol{\Phi}}^{(id)}|} \sum_{i=1}^{N_q} \left[\mathbb{1}_{\{i \notin \boldsymbol{\Phi}^{(id)}\}} l_f\left([1], [\hat{c}_{\hat{\omega}^{(id)}(i)}] \right) + \mathbb{1}_{\{i \in \boldsymbol{\Phi}^{(id)}\}} l_f\left([0], [\hat{c}_{\hat{\omega}^{(id)}(i)}] \right) \right], \tag{7}$$

$$\mathcal{L}_b = \frac{1}{|\bar{\boldsymbol{\Phi}}^{(id)}|} \sum_{i=1}^{N_q} \mathbb{1}_{\{i \notin \boldsymbol{\Phi}^{(id)}\}} \left\| \boldsymbol{b}_i - \hat{\boldsymbol{b}}_{\hat{\omega}^{(id)}(i)} \right\|_1, \tag{8}$$

$$\mathcal{L}_o = \frac{1}{|\bar{\boldsymbol{\Phi}}^{(id)}|} \sum_{i=1}^{N_q} \mathbb{1}_{\{i \notin \boldsymbol{\Phi}^{(id)}\}} \left[1 - f_{GIoU}\left(\boldsymbol{b}_i, \hat{\boldsymbol{b}}_{\hat{\omega}^{(id)}(i)} \right) \right], \tag{9}$$

$$\mathcal{L}_a = \frac{1}{|\bar{\boldsymbol{\Phi}}^{(id)}|} \sum_{i=1}^{N_q} \mathbb{1}_{\{i \notin \boldsymbol{\Phi}^{(id)}\}} l_f\left(\boldsymbol{a}_i, \hat{\boldsymbol{a}}_{\hat{\omega}^{(id)}(i)} \right), \tag{10}$$

where $\lambda_{\{c,b,o,a\}}$ are hyper-parameters and $l_f(\cdot,\cdot)$ is the element-wise focal loss function [13] whose hyper-parameters are described in [24].

In our training, the hyper-parameters $\eta_{\{c,b,o,a\}}$ and $\lambda_{\{c,b,o,a\}}$ are set as $\eta_c = \lambda_c = 1$, $\eta_b = \lambda_b = 5$, $\eta_o = \lambda_o = 2$, and $\eta_a = \lambda_a = 2$.

B. Implementation Details of Detection Heads

In our method, all the detection heads are constituted by feed-forward networks with the subsequent sigmoid functions. The details of the detection heads are as follows:

Person class head
This head has 1 linear layer with the subsequent sigmoid function.

Box head
This head has 3 linear layers with the ReLU activation between the layers and the subsequent sigmoid function. A reference point is added to each corresponding box position before applying the sigmoid function.

Action head
This head has 1 linear layer with the subsequent sigmoid function.

Activity head
This head has 1 linear layer with the subsequent sigmoid function.

Group size head
This head has 3 linear layers with the ReLU activation between the layers and the subsequent sigmoid function.

Member point head
This head has 3 linear layers with the ReLU activation between the layers and the subsequent sigmoid function. $2 \times M$ values are output from the last linear layer and then split into M group member points, where M denotes the maximum group size. A reference point is added to each corresponding group member point before applying the sigmoid function.

C. Group Annotations in The Volleyball Dataset

Group annotations are critical components to fully leverage the learning capability of our method. In the evaluation of the Volleyball dataset [8], we use the original annotation set combined with the extra annotation set provided by Sendo and Ukita [17] because the original annotations do not contain group information. The group annotations in the extra set are transferred to the original set by matching bounding boxes from each set with intersection over union (IoU). IoU is first calculated for each pair of a box from the original set and that from the extra set in the same frame. The calculated IoU values are then used as costs for the Hungarian algorithm [10] to match the boxes. If a box from the extra set has a label indicating that the person in the box is involved in an activity, we assign a group member flag to the matched box from the original set.

The players involved in each group activity are defined by Sendo and Ukita [17] as follows:

Pass
Players who are trying an underhand pass independently of whether or not they successfully do it.

Set
A player who is doing an overhand pass and those who will spike the ball whether they are trying or faking.

Spike
Players who are spiking and blocking.

Winpoint
All players in the team scoring a point. This group activity is observed for a few seconds right after scoring.

D. Additional List of Comparison

Table 1 shows the additional list of the comparison against state-of-the-art group activity recognition methods on the Volleyball and Collective Activity datasets. The values without the brackets demonstrate the detection-based performances, while those inside the brackets indicate the performances with ground

Table 1: Comparison of group activity recognition. The values w/ and w/o the brackets show the performances in the ground-truth-based and detection-based settings, respectively.

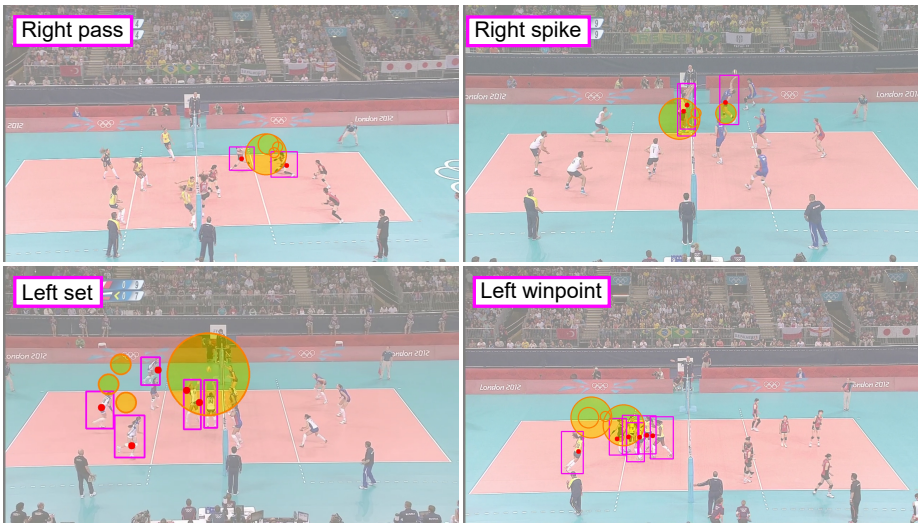
	Volleyball				Collective Activity			
Method	Activity		Action		Activity		Action	
SIM [3]	–	(–)	–	(–)	–	(81.2)	–	(–)
HDTM [8]	–	(81.9)	–	(–)	–	(81.5)	–	(–)
CERN [18]	–	(83.3)	–	(69.1)	–	(87.2)	–	(–)
SSU [2]	86.2	(90.6)	–	(81.8)	–	(–)	–	(–)
SBGAR [12]	–	(66.9)	–	(–)	–	(86.1)	–	(–)
HACN [9]	–	(85.1)	–	(–)	–	(84.3)	–	(–)
HRN [7]	–	(89.5)	–	(–)	–	(–)	–	(–)
stagNet [15]	87.6	(89.3)	–	(–)	87.9	(89.1)	–	(–)
PCTDM [21]	–	(87.7)	–	(–)	–	(92.1)	–	(–)
ARG [19]	91.5	(92.5)	39.8	(83.0)	86.1	(88.1)	49.6	(77.3)
CRM [1]	–	(93.0)	–	(–)	–	(85.8)	–	(–)
HiGCIN [22]	–	(91.4)	–	(–)	93.4	(–)	–	(–)
PRL [6]	–	(91.4)	–	(–)	–	(–)	–	(–)
Actor-Transformers [5]	–	(94.4)	–	(85.9)	–	(92.8)	–	(–)
Ehsanpour *et al.* [4]	93.0	(93.1)	41.8	(83.3)	89.4	(89.4)	55.9	(78.3)
Pramono *et al.* [14]	–	(95.0)	–	(83.1)	–	(95.2)	–	(–)
P^2CTDM [20]	–	(92.7)	–	(–)	–	(96.1)	–	(–)
DIN [23]	–	(93.6)	–	(–)	–	(95.9)	–	(–)
GroupFormer [11]	95.0*	(95.7)	–	(85.6)	85.2*	(87.5†/96.3)	–	(–)
Ours	**96.0**	(–)	**65.0**	(–)	**96.5**	(–)	**64.9**	(–)

* We evaluated the performance with the publicly available source codes.
† We evaluated but were not able to reproduce the reported accuracy because the configuration file for the Collective Activity dataset is not publicly available.

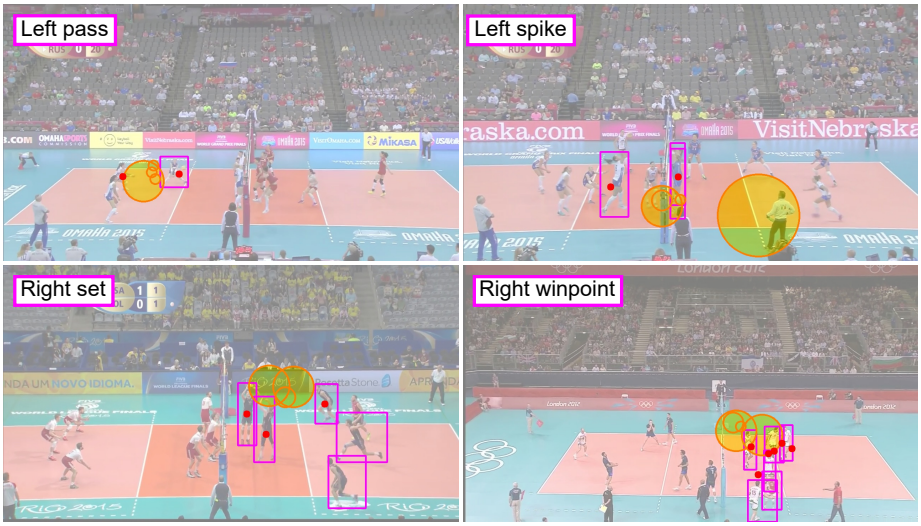
truth bounding boxes. SIM [3], HDTM [8], CERN [18], SBGAR [12], HACN [9], HRN [7], PCTDM [21], HiGCIN [22], and PRL [6] are additionally compared in this table. As shown in the table, our method outperforms state-of-the-art methods on both datasets, indicating the effectiveness of our feature extraction method.

E. Additional Qualitative Analysis

We further analyze the recognition results qualitatively with our method's success and failure cases on the Volleyball dataset [8]. The results of the successful cases and failure cases are shown in Fig. 1a and 1b, respectively. The purple bounding boxes show the ground truth group members, the red circles show the predicted group member points, and the yellow circles show the attention locations. The small and large yellow circles mean that the locations are in the high



(a) Successful cases.



(b) Failure cases.

Fig. 1: Visualization of the social group activity recognition results. The purple bounding boxes, red circles, and yellow circles show the ground truth group members, predicted group member points, and attention locations in the deformable transformer decoder, respectively.

and low-resolution feature maps, respectively, offering a rough range of image areas affecting the features used for the predictions.

As seen from the figures, features are successfully aggregated from the areas around the group members in the successful cases, while those are aggregated from the regions around the non-group members, backgrounds, and part of the group members in the failure cases. It is worth noting that our method successfully recognizes social group activities even when one group member is apart from the other members such as in the cases of "Right spike" and "Left winpoint" in Fig. 1a, demonstrating the effectiveness of our feature aggregation method. In the failure case of "Left pass", a non-group member is falsely recognized as a group member probably because the non-group member has the pose of the underhand pass, which is quite similar to the group member. In the failure case of "Left spike", a group member cannot be identified due to the occlusion. To correctly identify group members in these cases, long-term temporal context should be leveraged effectively. In the failure cases of "Right set" and "Right winpoint", the group members are widely distributed especially in the vertical direction. As discussed in the main manuscript, the group member point prediction is designed on the assumption that group members are seen side by side at the same vertical positions in an image. This design might affect the performance of the failure cases. These observations present opportunities for future work.

References

1. Azar, S.M., Atigh, M.G., Nickabadi, A., Alahi, A.: Convolutional relational machine for group activity recognition. In: CVPR (June 2019)
2. Bagautdinov, T.M., Alahi, A., Fleuret, F., Fua, P.V., Savarese, S.: Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: CVPR (July 2017)
3. Deng, Z., Vahdat, A., Hu, H., Mori, G.: Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: CVPR (June 2016)
4. Ehsanpour, M., Abedin, A., Saleh, F., Shi, J., Reid, I., Rezatofighi, H.: Joint learning of social groups, individuals action and sub-group activities in videos. In: ECCV (August 2020)
5. Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C.G.M.: Actor-transformers for group activity recognition. In: CVPR (June 2020)
6. Hu, G., Cui, B., He, Y., Yu, S.: Progressive relation learning for group activity recognition. In: CVPR (June 2020)
7. Ibrahim, M.S., Mori, G.: Hierarchical relational networks for group activity recognition and retrieval. In: ECCV (September 2018)
8. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: CVPR (June 2016)
9. Kong, L., Qin, J., Huang, D., Wang, Y., Gool, L.V.: Hierarchical attention and context modeling for group activity recognition. In: ICASSP (April 2018)
10. Kuhn, H.W., Yaw, B.: The hungarian method for the assignment problem. Naval Res. Logist. Quart pp. 83–97 (March 1955)

11. Li, S., Cao, Q., Liu, L., Yang, K., Liu, S., Hou, J., Yi, S.: GroupFormer: Group activity recognition with clustered spatial-temporal transformer. In: ICCV (October 2021)
12. Li, X., Chuah, M.C.: SBGAR: Semantics based group activity recognition. In: ICCV (October 2017)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (October 2017)
14. Pramono, R.R.A., Chen, Y.T., Fang, W.H.: Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In: ECCV (August 2020)
15. Qi, M., Qin, J., Li, A., Wang, Y., Luo, J., Gool, L.V.: stagNet: An attentive semantic rnn for group activity recognition. In: ECCV (September 2018)
16. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR (June 2019)
17. Sendo, K., Ukita, N.: Heatmapping of people involved in group activities. In: MVA (May 2019)
18. Shu, T., Todorovic, S., Zhu, S.C.: CERN: Confidence-energy recurrent network for group activity recognition. In: CVPR (July 2017)
19. Wu, J., Wang, L., Wang, L., Guo, J., Wu, G.: Learning actor relation graphs for group activity recognition. In: CVPR (June 2019)
20. Yan, R., Shu, X., Yuan, C., Tian, Q., Tang, J.: Position-aware participation-contributed temporal dynamic model for group activity recognition. IEEE TNNLS (June 2021)
21. Yan, R., Tang, J., Shu, X., Li, Z., Tian, Q.: Participation-contributed temporal dynamic model for group activity recognition. In: ACMMM (October 2018)
22. Yan, R., Xie, L., Tang, J., Shu, X., Tian, Q.: HiGCIN: Hierarchical graph-based cross inference network for group activity recognition. IEEE TPAMI (October 2020)
23. Yuan, H., Ni, D., Wang, M.: Spatio-temporal dynamic inference network for group activity recognition. In: ICCV (October 2021)
24. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points (April 2019), arXiv:1904.07850
25. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: ICLR (May 2021)