Supplementary Material

TABLE I
COMPARISON OF THE GROUP ACTIVITY RECOGNITION ACCURACY (%) ON THE VOLLEYBALL DATASET.

	Student	Teacher	Teacher-Student
GCN	90.4	92.3	91.2
GCN$_2$	90.7	92.6	91.1

I. VISUALIZATION OF ACTION SEGMENTATION RESULTS ON THE UNTRIMMED VOLLEYBALL DATASET

We present several action segmentation results in Fig. 1. We observe that combining our proposed model with TCN [2] can obtain better performance than other approaches.

II. EXPLORATION ON TRANSFERRING KNOWLEDGE BETWEEN GRAPHS

In our original GCN formulation, the adjacency matrix A is obtained based on the locations of different people. In order to explore transferring knowledge between graphs, we first tried to built a new GCN$_1$ model, where the adjacency matrix A$_1$ is "self-learned" but not constructed based on the location information. Specifically, we fed the features of the n-th nodes into a fully-connected layer and obtained the n-th row of A$_1$ (n=1, 2, ..., N). However, we found the performance was decreased compared with the original GCN model for Teacher network (GCN$_1$: 87.8% vs GCN: 92.4%) due to the lack of location information. Then, we tried to feed the input into the GCN and GCN$_1$ simultaneously, and concatenated their outputs before the next layer. We denote this approach as GCN$_2$, and applied it to the Teacher and Student network respectively. Furthermore, we transferred the knowledge of GCN$_2^{teacher}$ to GCN$_2^{student}$ by adding an MSE based on the learned adjacency matrix A$_1$, which is a similar scheme to the J_{SPA} in the original paper.

Table I presents the experiments results. We find that the performance of new GCN$_2$ model for Student and Teacher networks have been improved slightly, while the final accuracy of Student-Teacher network is comparable with our original model. This may be because the "self-learned" adjacency matrix A$_1$ is similar to the "self-attention" scores in our original model, so the improvement is not significant. We will further explore this interesting direction in the future.

III. ANALYSIS ON DIFFERENT TYPES OF LOSS FUNCTIONS

In the original paper, we employ an MSE based loss function, which was adopted in a related work [3]. We further conducted experiments on the KL loss, which is defined as follow:

$$KL(att_i^{teacher}||att_i^{student}) = \sum_i att_i^{teacher} log \frac{att_i^{teacher}}{att_i^{student}}, \quad (1)$$

TABLE II
COMPARISON OF THE GROUP ACTIVITY RECOGNITION ACCURACY (%) OF DIFFERENT LOSS FUNCTIONS ON THE VOLLEYBALL DATASET.

J_{KD} J_{SPA}	MSE	KL
MSE	**91.2**	90.9
KL	90.7	90.8

TABLE III
COMPARISON OF THE GROUP ACTIVITY RECOGNITION ACCURACY (%) WITH λ_1 AND λ_2 ON THE VOLLEYBALL DATASET.

λ_1 λ_2	0.25	0.50	1	2
0.25	91.1	90.7	91.0	90.7
0.50	91.0	90.9	90.8	90.8
1	91.0	90.8	**91.2**	90.5
2	90.7	90.7	90.6	90.4

where $att_i^{teacher}$ and $att_i^{student}$ denote the i-th element in the attention scores of Teacher network and Student network respectively. Table II presents the experimental results on the volleyball dataset, which indicates the MSE based loss is better for the volleyball dataset in practise.

IV. ANALYSIS ON THE HYPER-PARAMETERS

Our loss function is defined as:

$$J \quad = \quad J_{CLS} + \lambda_1 J_{SPA} + \lambda_2 J_{KD}.$$

In order to study the effect of attention transfer and knowledge distillation, we conducted experiments on different λ_1 and λ_2. Table III presents the comparing results of λ_1 and λ_2 on the Volleyball dataset. We observe that when $\lambda_1 \leq 1$ and $\lambda_2 \leq 1$, the results vary slightly and the peak is achieved when $\lambda_1 = \lambda_2 = 1$. Besides, we find the minimum is 90.4% when $\lambda_1 = \lambda_2 = 2$, which indicates that the J_{SPA} and J_{KD} should not be over emphasized compared with J_{CLS}.

V. CONFUSION MATRICES ON THE VOLLEYBALL DATASET

Fig. 2 presents four confusion matrices of the baseline method and three variants of our model. As it shown, Ours†+GCN$_{-\ SPA\ +\ KD}$ clearly improves the performance of the baseline method [1], especially for the classes of "left pass", "right set", "left win" and "right win". And combining optical flows can increase the accuracy of "left spike", "left win" and "right win".

REFERENCES

[1] Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: Hierarchical deep temporal models for group activity recognition. CoRR **abs/1607.02643** (2016)

[2] Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: CVPR. (2017) 1003–1012

[3] Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: ICLR. (2017)

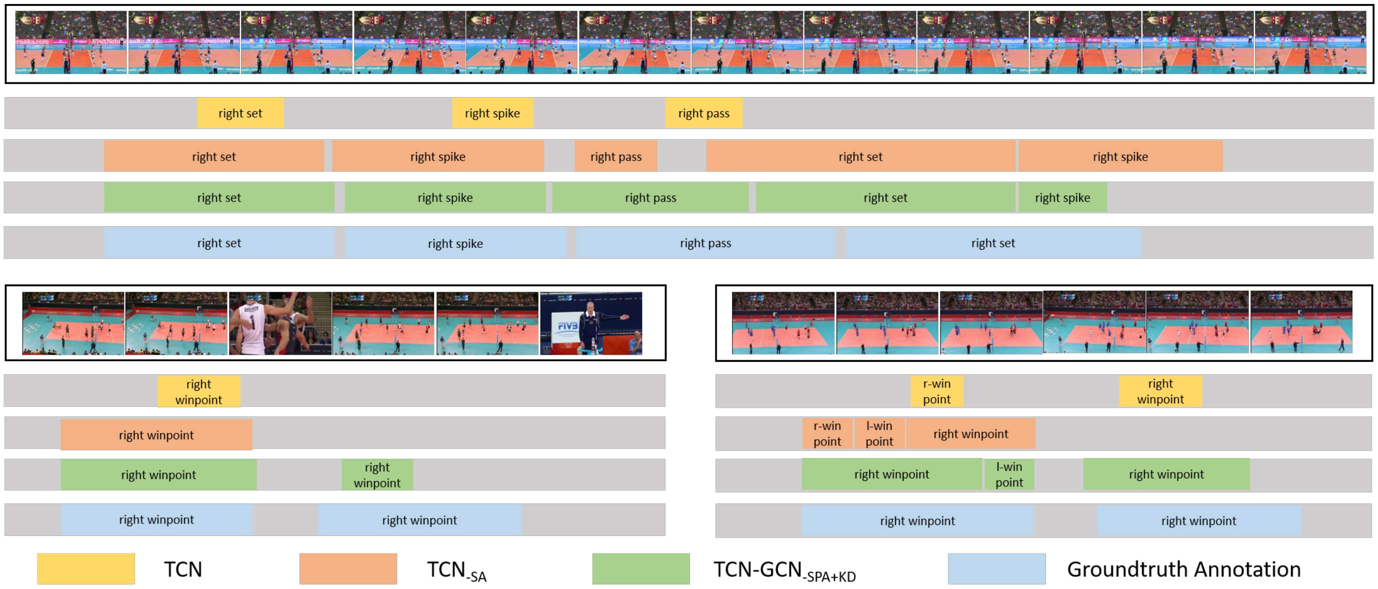
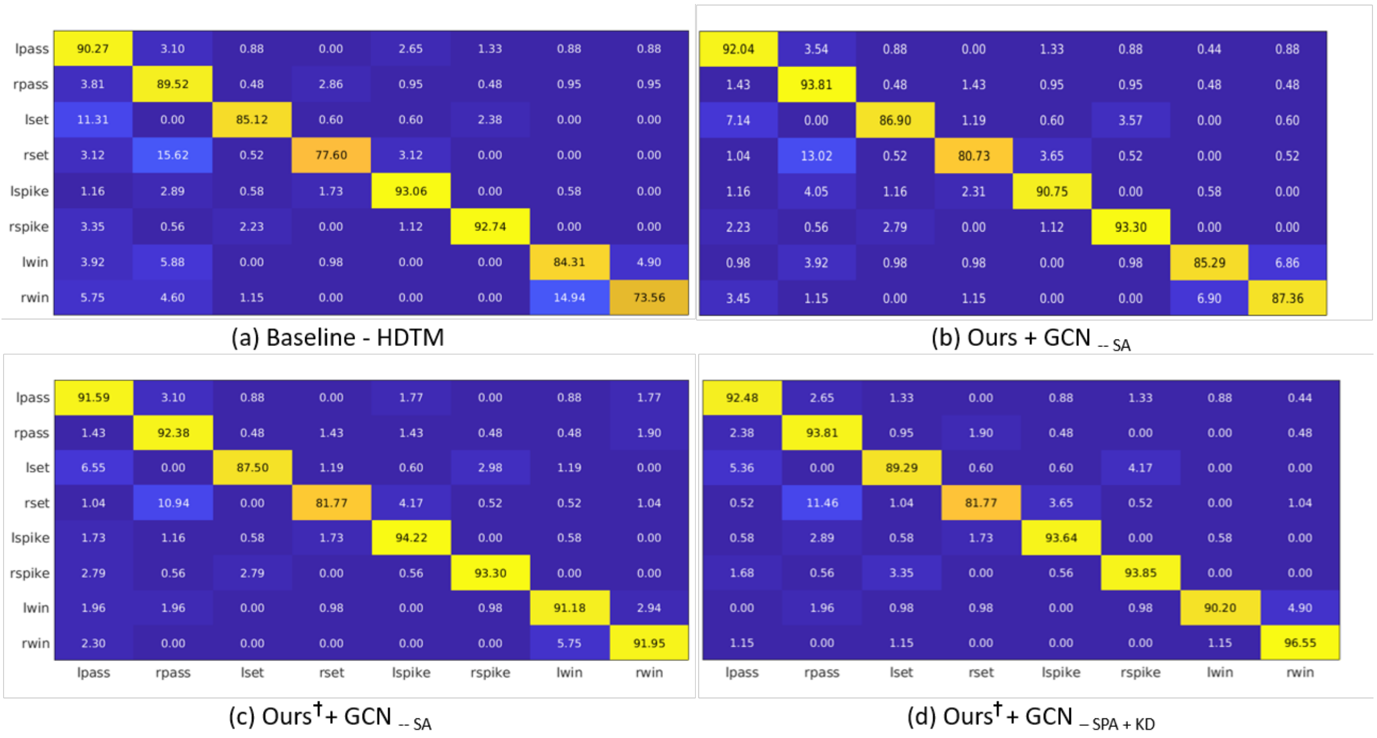


Fig. 1. Visualization of action segmentation results on the untrimmed Volleyball dataset.



	lpass	rpass	lset	rset	lspike	rspike	lwin	rwin
lpass	90.27	3.10	0.88	0.00	2.65	1.33	0.88	0.88
rpass	3.81	89.52	0.48	2.86	0.95	0.48	0.95	0.95
lset	11.31	0.00	85.12	0.60	0.60	2.38	0.00	0.00
rset	3.12	15.62	0.52	77.60	3.12	0.00	0.00	0.00
lspike	1.16	2.89	0.58	1.73	93.06	0.00	0.58	0.00
rspike	3.35	0.56	2.23	0.00	1.12	92.74	0.00	0.00
lwin	3.92	5.88	0.00	0.98	0.00	0.00	84.31	4.90
rwin	5.75	4.60	1.15	0.00	0.00	0.00	14.94	73.56

(a) Baseline - HDTM

	lpass	rpass	lset	rset	lspike	rspike	lwin	rwin
lpass	92.04	3.54	0.88	0.00	1.33	0.88	0.44	0.88
rpass	1.43	93.81	0.48	1.43	0.95	0.95	0.48	0.48
lset	7.14	0.00	86.90	1.19	0.60	3.57	0.00	0.60
rset	1.04	13.02	0.52	80.73	3.65	0.52	0.00	0.52
lspike	1.16	4.05	1.16	2.31	90.75	0.00	0.58	0.00
rspike	2.23	0.56	2.79	0.00	1.12	93.30	0.00	0.00
lwin	0.98	3.92	0.98	0.98	0.00	0.98	85.29	6.86
rwin	3.45	1.15	0.00	1.15	0.00	0.00	6.90	87.36

(b) Ours + GCN $_{\text{-- SA}}$

	lpass	rpass	lset	rset	lspike	rspike	lwin	rwin
lpass	91.59	3.10	0.88	0.00	1.77	0.00	0.88	1.77
rpass	1.43	92.38	0.48	1.43	1.43	0.48	0.48	1.90
lset	6.55	0.00	87.50	1.19	0.60	2.98	1.19	0.00
rset	1.04	10.94	0.00	81.77	4.17	0.52	0.52	1.04
lspike	1.73	1.16	0.58	1.73	94.22	0.00	0.58	0.00
rspike	2.79	0.56	2.79	0.00	0.56	93.30	0.00	0.00
lwin	1.96	1.96	0.00	0.98	0.00	0.98	91.18	2.94
rwin	2.30	0.00	0.00	0.00	0.00	0.00	5.75	91.95

(c) Ours†+ GCN $_{\text{-- SA}}$

	lpass	rpass	lset	rset	lspike	rspike	lwin	rwin
lpass	92.48	2.65	1.33	0.00	0.88	1.33	0.88	0.44
rpass	2.38	93.81	0.95	1.90	0.48	0.00	0.00	0.48
lset	5.36	0.00	89.29	0.60	0.60	4.17	0.00	0.00
rset	0.52	11.46	1.04	81.77	3.65	0.52	0.00	1.04
lspike	0.58	2.89	0.58	1.73	93.64	0.00	0.58	0.00
rspike	1.68	0.56	3.35	0.00	0.56	93.85	0.00	0.00
lwin	0.00	1.96	0.98	0.98	0.00	0.98	90.20	4.90
rwin	1.15	0.00	1.15	0.00	0.00	0.00	1.15	96.55

(d) Ours†+ GCN $_{\text{– SPA + KD}}$

Fig. 2. Comparison of Confusion Matrices on the volleyball dataset [1]. † denotes that the model takes both RGB images and optical flows as inputs.