Learning Group Activity Feature Through Person Attribute Prediction

Supplementary Material

Chihiro Nakatani[1] Hiroaki Kawashima[2] Norimichi Ukita[1]
[1] Toyota Technological Institute, Japan [2] University of Hyogo, Japan

1. Implementation Details

This section shows the implementation details that are not mentioned in the main paper.

1.1. Experimental Conditions

We trained all of our models on the Nvidia A100 GPUs with batch size 8. For pertaining the person feature extractor in the GAFL-PAC, we set the learning rates to be 0.0001 and 0.00005 for the volleyball and collective activity datasets, respectively. As mentioned in the main paper, the person feature extractor is fine-tuned through our GAF learning in GAFL-PAC while we freeze the person feature extractor trained on ImageNet in GAFL-PAF following the previous method [14]. As the person feature extractor, we utilize the code implemented for DIN [39] available at https://github.com/JacobYuan7/DIN-Group-Activity-Recognition-Benchmark. In our GAF learning, the learning rates are 0.0001 and 0.00005 for the volleyball and collective activity datasets, respectively.

We utilize the experimental setting of DIN [39] as follows. We use video clips consisting of 10 frames (i.e., $T = 10$) for both the volleyball and collective activity datasets. The feature of each person is embedded into 1024 dimensional vectors (i.e., $C = 1024$) after RoIAlign in the person feature extractor. The RoIAlign in the person feature extractor is applied to each person with the ground-truth full-body bounding boxes as used in [11,14,37,39]. For evaluation, the threshold of action set retrieval (i.e., IoU and AF-IDF) is defined as 0.5 following [14].

1.2. Implementation of Previous Methods

All methods, including our method, are evaluated in the same setting for a fair comparison.

1.2.1 GAFL-PAC

HiGCIN [37], DIN [39] and Dual-AI [11] are used for a comparison in GAFL-PAC. As mentioned in the main paper, we train these models with person action labels for a fair comparison as with our method in GAFL-PAC. These codes are prepared as follows:

- **HiGCIN and DIN**: These codes are available at https://github.com/JacobYuan7/DIN-Group-Activity-Recognition-Benchmark.
- **Dual-AI**: The code is not available. We implement the method based on [11].

1.2.2 GAFL-PAF

As with the GAFL-PAC, we compare our method with the SOTA methods in GAFL-PAF. For evaluation, B1-Compact [14], B2-VGG18 [14], and HRN [14] are used for comparison because these methods are trained without group activity and person action annotations as with our method in GAFL-PAF. These codes are prepared as follows:

- **B1-Compact**: The code is not available. However, we implement the method based on [14] since the details of the model are mentioned in [14]. The dimension of each person feature is 128.
- **B2-VGG19**: As with B1-Compact, we implement B2-VGG19 based on [14]. In this method, the output of fc7 layer in a pretrained VGG19 network is regarded as a person feature of each person. The dimension of each person feature is 4096.
- **HRN**: While the code is available at https://github.com/mostafa-saad/hierarchical-relational-network, the retrieval function is not included in the provided code. Therefore, we implement the network for the retrieval function using person features acquired by HRN. This network is based on [14]. The dimension of each person feature is 128 as with B1-Compact.

2. Additional Experiments

Additional experiments, which are not included in the main paper for the page limitation, are presented in this section. As with the main paper, comparative experiments and detailed analysis are shown in Secs. 2.1 and 2.2 in this supplementary material.

2.1. Comparative Experiments

2.1.1 Retrieval

As noted in [14], previous methods (i.e., B1-Compact, B2-VGG19, and HRN) sort person image features of all people

Table 5. Additional quantitative comparison of retrieval on the VolleyBall Dataset (VBD) in GAFL-PAF. In addition to the results shown in Table 1 of the main paper, variants of these previous methods using $\boldsymbol{F}_{loc}$ are compared with our method in this table. The best result in each column is colored in red. Results obtained by the concatenation of output features (i.e., $\boldsymbol{F}_{grp}^{TS}$ and $\boldsymbol{F}_{grp}^{ST}$) and $\boldsymbol{G}$ are denoted as "Ours-ind" and "Ours-grp", respectively.

Retrieval type	Action set (IoU [14]				Action set (AF-IDF)				Group activity		
Method	Hit@1	Hit@2	Hit@3	mAP	Hit@1	Hit@2	Hit@3	mAP	Hit@1	Hit@2	Hit@3
B1-Compact128	57.9	75.7	84.3	45.8	41.3	60.8	71.4	29.3	30.3	48.0	59.9
B1-Compact128 w/ $\boldsymbol{F}_{loc}$	60.6	80.9	88.0	45.9	46.1	64.1	74.8	29.3	34.8	52.1	63.6
B2-VGG19	63.8	80.6	86.8	46.8	46.7	65.8	75.7	29.4	35.4	53.6	65.0
B2-VGG19 w/ $\boldsymbol{F}_{loc}$	63.6	80.5	88.0	46.1	48.5	66.6	76.3	28.9	51.6	69.1	78.3
HRN	60.9	78.6	86.0	46.9	40.8	60.9	72.9	28.7	31.2	47.0	57.6
HRN w/ $\boldsymbol{F}_{loc}$	60.3	77.9	85.0	46.9	42.3	62.2	73.2	28.7	29.3	44.9	56.8
Ours-ind	64.2	80.8	88.3	45.0	50.4	69.3	77.6	30.1	55.0	72.3	79.2
Ours-grp	64.8	82.7	90.3	46.4	52.3	71.4	81.0	31.4	61.1	75.1	82.4

Table 6. Additional quantitative comparison of retrieval on the Collective Activity Dataset (CAD) in GAFL-PAF. In addition to the results shown in Table 2 of the main paper, variants of these previous methods using $\boldsymbol{F}_{loc}$ are compared with our method in this table.

Retrieval type	Action set (IoU [14]				Action set (AF-IDF)				Group activity		
Method	Hit@1	Hit@2	Hit@3	mAP	Hit@1	Hit@2	Hit@3	mAP	Hit@1	Hit@2	Hit@3
B1-Compact128	48.8	60.3	68.2	38.0	81.8	88.2	89.7	52.6	82.4	88.4	90.1
B1-Compact128 w/ $\boldsymbol{F}_{loc}$	59.1	73.7	78.7	40.1	70.7	79.1	82.2	46.4	70.7	79.6	82.9
B2-VGG19	53.6	61.6	66.1	35.3	71.1	80.3	83.8	46.7	72.2	80.8	84.2
B2-VGG19 w/ $\boldsymbol{F}_{loc}$	43.5	54.8	60.9	37.0	81.2	86.7	89.3	52.2	80.9	85.1	86.9
HRN	37.1	50.1	58.6	22.2	53.2	64.8	72.5	34.2	54.0	64.8	72.4
HRN w/ $\boldsymbol{F}_{loc}$	32.0	45.8	54.2	21.1	49.2	60.7	69.8	32.9	49.9	60.7	69.3
Ours-ind	67.6	81.3	85.9	53.3	83.7	88.9	90.2	57.5	88.5	91.2	91.9
Ours-grp	52.7	70.3	74.1	46.4	74.0	80.5	82.6	60.1	79.2	81.0	82.0

based on their locations to avoid comparison with each possible permutation for retrieval in Sec. 4.4.1 of the main paper. However, the location-based sorting only captures the spatial interaction between people coarsely. Therefore, we also compare variants of these previous methods in which location features (denoted by $\boldsymbol{F}_{loc}$ as with our method) are added to their person image features to capture such spatial interaction precisely. As with our method, the location of each person is encoded in a feature vector by positional encoding.

Volleyball dataset (GAFL-PAF). Table 5 shows the results obtained by the previous methods (i.e., equal to the results shown in Table 1 of the main paper) and results obtained by their variants. In both action set and group activity retrieval, we can see that adding $\boldsymbol{F}_{loc}$ is somewhat effective in these previous methods. Specially, the performance gain between "B2-VGG19" and "B2-VGG19 w/ $\boldsymbol{F}_{loc}$" is large (i.e., 16.2 %). However, our method is still better than these variants in all metrics. These results demonstrate that our GAF learned through location-guided person attribute prediction in the APN (Fig. 3 (c) of the main paper) is better for representing people in a group.

Collective activity dataset (GAFL-PAF). As with the volleyball dataset, the results on the collective activity dataset are shown in Table 6. Different from the results on the volleyball dataset, $\boldsymbol{F}_{loc}$ has a negative impact on the collective activity dataset. These results may come from the fact that $\boldsymbol{F}_{loc}$ makes the training of the person feature extractor complex in these previous methods. For example, in B2-VGG19, we can see the positive impact of $\boldsymbol{F}_{loc}$. This is because the method only employs the frozen model (i.e., VGG19) without any training. Compared with all of these methods, our method is still the best in all metrics. These results reveal that our location-guidance disentangles such location information in the complex GAF, as its effectiveness is also validated in Tables 3 and 7 of the main paper and the supplementary materials.

2.1.2 Group Activity Recognition

In this section, we show additional results of group activity recognition using 1-nearest neighbor classification, which is shown in Sec. 4.4.2 of the main paper.

Confusion matrices in GAFL-PAC. As with the results in GAFL-PAF shown in Fig. 4 of the main paper, confu-

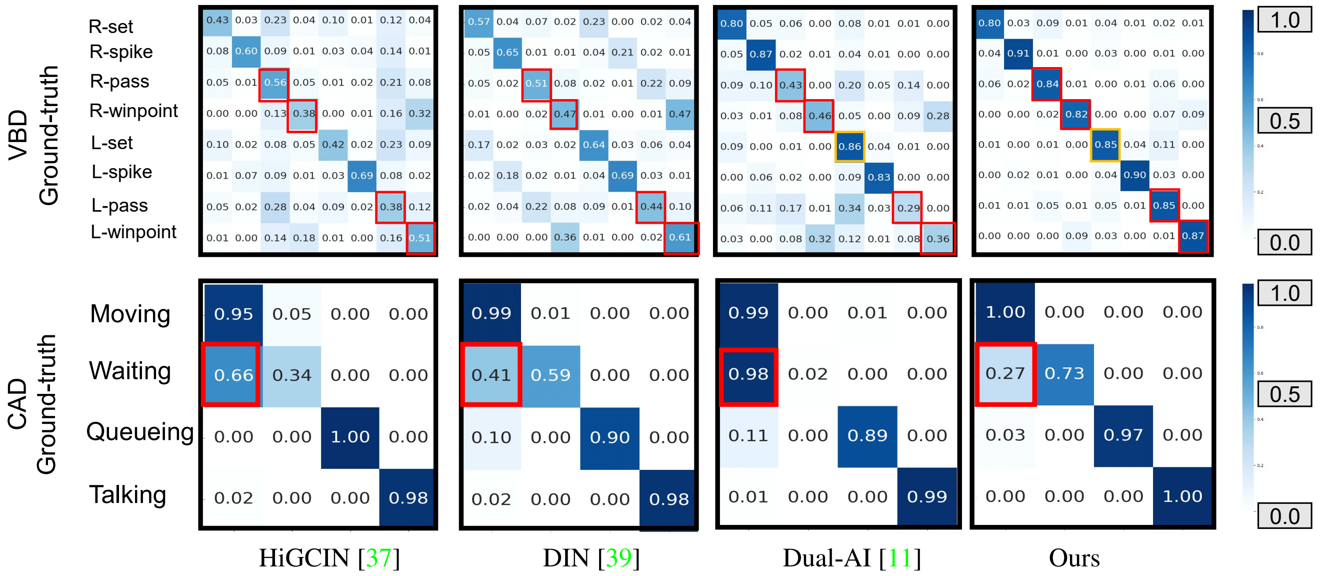


Figure 8. Confusion matrices of GAR by nearest neighbor retrieval on the VolleyBall Dataset (VBD) and Collective Activity Dataset (CAD) in GAFL-PAC. Each row and column show the ground-truth of the GA label and the recognized GA, respectively.

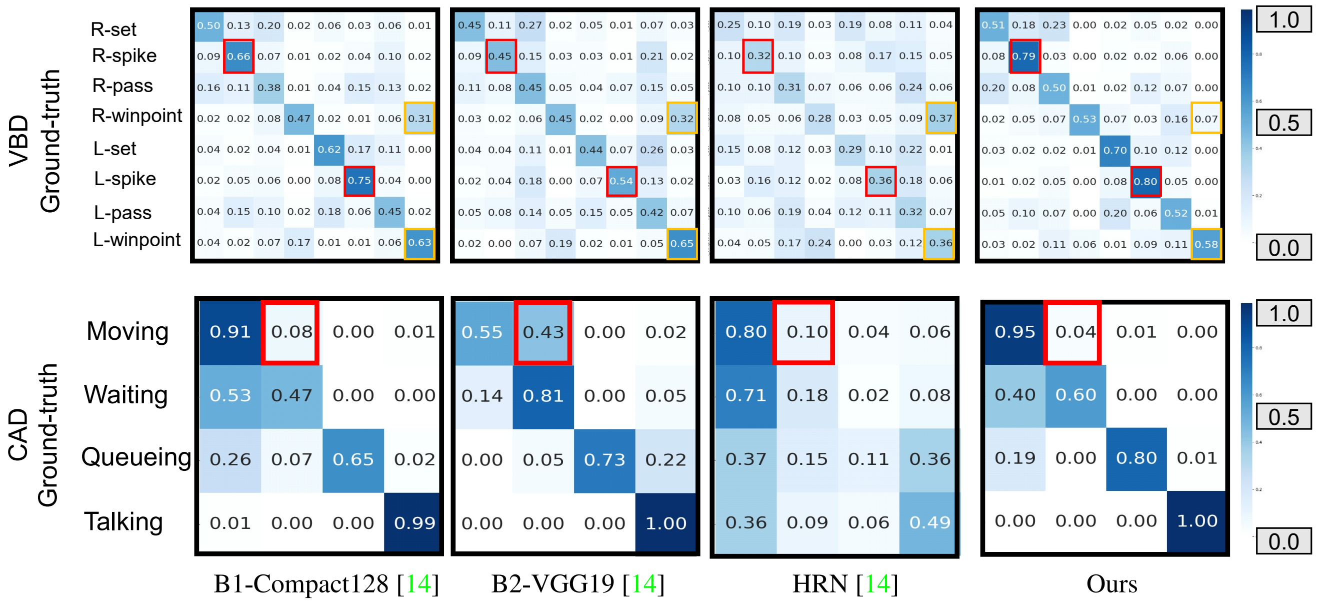
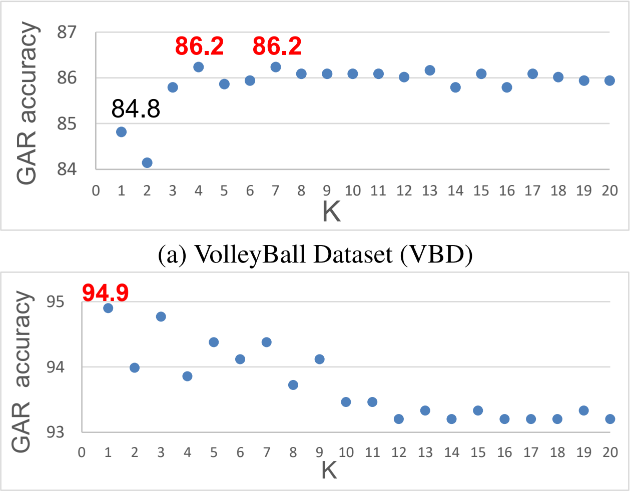


Figure 9. Confusion matrices of GAR by nearest neighbor retrieval on the VolleyBall Dataset (VBD) in GAFL-PAF. Each row and column show the ground-truth of the GA label and the recognized GA, respectively. While the results obtained by "B2-VGG19" and "Ours" are only shown in Fig. 4 of the main paper, this figure shows the results obtained by all methods used in a comparison.

sion matrices in GAFL-PAC on the volleyball and collective activity datasets are also shown in Fig. 8. As shown in Fig. 8 (top), our method is better in all group activity classes on the volleyball dataset except for L-set in Dual-AI. In the L-set, however, the accuracy difference between "Dual-AI" and "Ours" is small (i.e., 0.01 %). Furthermore, "Ours" achieves high-performance gain in L-pass, R-pass, L-winpoint, and R-winpoint compared with Dual-AI. The average performance gain in these four group activity classes is 0.46 %, so the performance drop in L-set (i.e., 0.01 %) can be regarded as relatively small. This is because our method learns visually subtle but important differences between these group activities, as demonstrated in the main paper (e.g., Fig. 6).



(a) VolleyBall Dataset (VBD)

(b) Collective Activity Dataset (CAD)

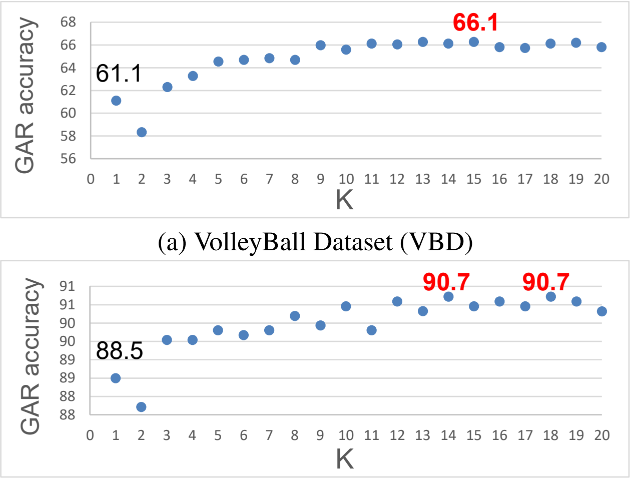
Figure 10. GAR accuracy curve by the KNN classification in GAFL-PAC on the VolleyBall Dataset (VBD) and Collective Activity Dataset (CAD). K changes from 1 to 20 in our experiment.

Figure 8 (bottom) shows that our method is the best in all group activity classes on the collective activity dataset. In particular, while the previous methods get many false negatives in Waiting, our method correctly recognizes Waiting. We can interpret this reason that some Waiting scenes are visually similar to specific Moving scenes in which the movement of people is relatively small. Such visually subtle differences are well discriminated in our method, as mentioned in the main paper.

Confusion matrices in GAFL-PAF. In addition to the results obtained by “B2-VGG19” shown in Fig. 4 of the main paper, we further show the results obtained by “B1-Compact128” and “HRN” in Fig. 9 of this supplementary material. Regarding the additional results, we also see the same superiority of our method (i.e., Spike activity is well recognized in VBD, and false negatives in Moving are quite low on CAD), as noted in the main paper.

KNN for Group Activity Recognition in GAFL-PAC. In addition to the results of Group Activity Recognition (GAR) obtained by the 1-nearest neighbor classification shown in Figs. 8 and 9, the results obtained with other neighbor numbers (e.g., 2) on the volleyball and collective activity datasets in GAFL-PAC are also shown in Fig. 10. In our experiments, K is changed from 1 to 20 for the K-nearest neighbor classification.

Fig. 10 (a) shows that $K \geq 3$ achieves better GAR accuracy than $K = 1$ on the volleyball dataset. Specifically, the best results obtained by $K = 4, 7$ are 1.4% better than that of $K = 1$. The results show that using KNN for GAR is simple but effective for accuracy. Results on the collective activity dataset shown in Fig. 10 (b) show that the result obtained by $K = 1$ (i.e., 94.9) is the best. The results indicate that the 1-nearest neighbor classification shown in Fig. 8 (bottom) is accurate enough in GAFL-PAC on the collective activity dataset.



(a) VolleyBall Dataset (VBD)

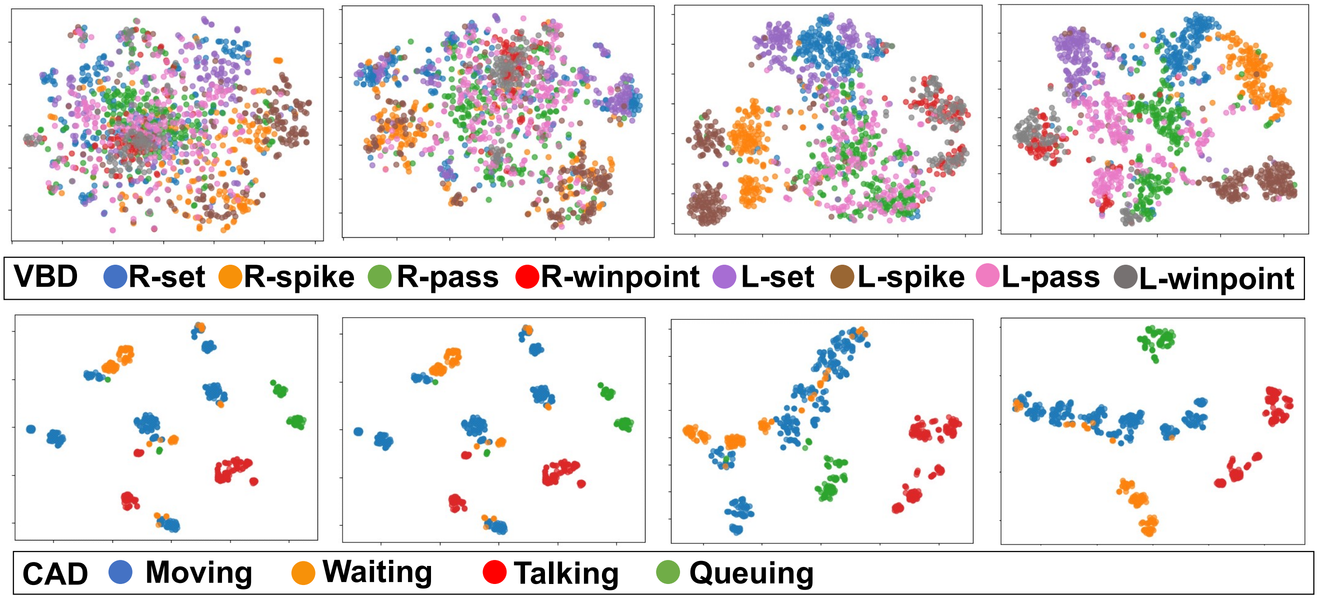
(b) Collective Activity Dataset (CAD)

Figure 11. GAR accuracy curve by the KNN classification in GAFL-PAF on the VolleyBall Dataset (VBD) and Collective Activity Dataset (CAD). K changes from 1 to 20 in our experiment.

KNN for Group Activity Recognition in GAFL-PAF. As with the results above, the results obtained on the volleyball and collective activity datasets in GAFL-PAF are also shown in Fig. 11. Figure 11 shows that $K \geq 3$ achieves better GAR accuracy than $K = 1$ on the volleyball and collective activity datasets. In particular, the best result obtained by $K = 15$ on the volleyball dataset is 5.0% better than $K = 1$. The results indicate that the KNN is more effective when the performance obtained by $K = 1$ is not highly accurate (i.e., 61.1 % on VBD in GAFL-PAF) compared with the accurate results (i.e., 84.8 % on VBD in GAFL-PAC). This difference may come from the fact that a more abstract supervision signal (i.e., person appearance features) is used to learn in GAFL-PAF, so the GAF includes redundant information for representing manually annotated group activity classes. For such GAF, ensembling ($K \geq 2$) is more effective due to its robustness.

2.1.3 Visualization of Learned Group Activity Feature

GAF visualization in GAFL-PAC. As shown in Fig. 5 of the main paper, the distribution of learned GAFs in GAFL-PAC is visualized in Fig. 12. Figure 12 shows that our method can learn the GAFs better than the other methods on the VolleyBall Dataset (VBD) and Collective Activity



HiGCIN [37] DIN [39] Dual-AI [11] Ours

Figure 12. Visualization of the learned GAF by t-SNE on the VolleyBall Dataset (VBD) and Collective Activity Dataset (CAD) in GAFL-PAC. The color of each sample shows the annotated group activity class corresponding to each test sample. Results obtained by "Ours-grp" are regarded as "Ours" in VBD and CAD.

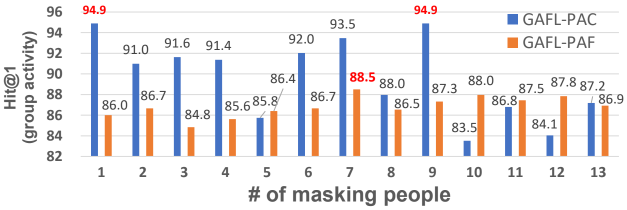


Figure 13. Performance changes depending on the number of masking people on the Collective Activity Dataset (CAD).

Dataset (CAD) in GAFL-PAC, as with the results in GAFL-PAF shown in the main paper. The results on VBD shown in Fig. 12 (upper) reveal that our GAF is useful for discriminating L-pass and R-pass, which are also mentioned in the above confusion matrices (Fig. 8 in this supplementary material). Figure 12 (bottom) shows the results on CAD and indicates that our method learns GAF well in terms of representing visually similar Waiting and Moving activities. The results are also validated in the above confusion metrics (Fig. 8 in this supplementary material).

2.2. Detailed Analysis

Comparison of the number of masked persons. As shown in Fig. 7 of the main paper, the performance changes depending on N_{mask} on the Collective Activity Dataset (CAD) in GAFL-PAC and GAFL-PAF are shown in Fig. 13. The results in GAFL-PAC (denoted by blue bars) show that our MPM is not important in the group activity retrieval performance. This is because the GAF trained with person action labels is enough to represent group activities defined by the maximum number of person actions in CAD, even without our MPM. In contrast, the results in GFL-PAF (denoted by orange bars) validate that our MPM improves the group activity retrieval performance. Specifically, the results obtained by $\boldsymbol{N}_{mask} = 7$ is 2.5% better than those without our MPM. These results reveal that our MPM is also effective for learning GAF for such general scenes included in CAD.

Table 7. Effectivenss of our location-guidance in our GAF learning on the Collective Activity Dataset (CAD). Results obtained by "Ours-grp" and "Ours-ind" are shown as "Ours" in GAFL-PAC and GAFL-PAF, respectively.

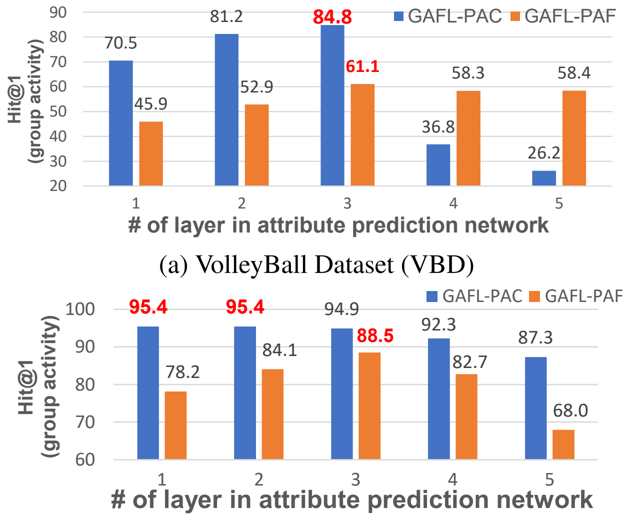
	Retrieval type	Action set (IoU)	Action set (AF-IDF)	Group activity
	Method	Hit@1	Hit@1	Hit@1
GAFL-PAC	Ours w/o $\boldsymbol{F}_{loc}$	80.5	95.0	92.7
	Ours	81.8	96.1	94.9
GAFL-PAF	Ours w/o $\boldsymbol{F}_{loc}$	42.1	56.5	57.1
	Ours	67.6	83.7	88.5

Effect of location-guidance in our GAF learning. While the results of the ablation study for $\boldsymbol{F}^{p}_{loc}$ on VBD in GAFL-PAC and GAFL-PAF are shown in Table 3 of the main paper, we further show the results on CAD in Table 7. On both GAFL-PAC and GAFL-PAF, "Ours" is better than "Ours w/o $\boldsymbol{F}^{p}_{loc}$" in all metrics. In particular, the performance gain in the GAFL-PAF is larger than the one in GAFL-PAC. We can interpret the reason as follows. In GAFL-PAC, "Ours w/o $\boldsymbol{F}_{loc}$" learns person action distribution of a scene, as mentioned in Sec. 4.5 of the main paper, is enough to represent group activities observed in CAD. This is because the location-related group activities (e.g., "Waiting" and "Queuing") can be represented by the distribution of person actions, which already includes location information in their class definition. Furthermore, the group activities are defined by the maximum number of person actions in a scene on CAD. Therefore, the group activities can be understood from person action distribution without $\boldsymbol{F}_{loc}$. In GAFL-PAF, however, person appearance feature distribution learned in "Ours w/o $\boldsymbol{F}_{loc}$" is not sufficient to represent group activities because people's appearances are similar to each other on CAD (e.g., the appearance of "Waiting" and "Queuing" actions are similar to each other). To discriminate these similar appearances, our location guidance is effective. For example, when people stand in line, their actions are likely to be regarded as "Queuing" even if their appearance features are similar to "Waiting".

In contrast to CAD, on VBD, our location guidance is crucial even in GAFL-PAC, as shown in Table 3 of the main paper. The results can be attributed to group activities where spatial relationships are meaningful (e.g., spiker and blocker are close in spike activity) on VBD.

Optimal layer number for our attribute prediction network. Figure 14 shows that the retrieval performance changes depending on the number of layers in our attribute prediction network on the VolleyBall Dataset (VBD) and Collective Activity Dataset (CAD). In general, such prediction performance is improved by increasing the number of layers. However, we find that the performance decreases when the layer number is larger than four in Fig. 14, while the performance increases until the layer number is three. These results may come from the complexity of our person attribute prediction using the GAF compared with general person attribute prediction in which features of the target person are directly used. The complexity may cause the model to overfit, so employing a shallow fully-connected network (e.g., 3-layer in this dataset) is adequate as our attribute prediction network.

The results on CAD in Fig. 14 (b) show that the performance is saturated even in the small number of layers (i.e., 1 and 2) in GAFL-PAC. This is because the attribute prediction using GAF in GAFL-PAC on CAD is easier than the one on the VBD. Therefore, such a shallow full connection network is enough for the attribute prediction network on CAD in GAFL-PAC.



(a) VolleyBall Dataset (VBD)

(b) Collective Activity Dataset (CAD)

Figure 14. Performance changes depending on the number of masking people on the VolleyBall Dataset (VBD) and Collective Activity Dataset (CAD).

Fine granularity of our GAF. As with the Fig. 6 of the main paper, we further show additional examples that validate the fine granularity of our GAF in Figs. 15 and 16.

In Fig. 15, we can see that the brown data points (i.e., L-spike) are divided into two sub-categories. The two sub-categories differ in where the spiker hits the ball on the court. Figure 16 shows that the gray data points (i.e., L-winpoint) are split into two sub-categories due to whether the receiver touched the ball or not. We further find the upper and bottom sub-categories are close to green (i.e., R-pass) and red (i.e., R-winpoint) data points, respectively. The reason for this closeness can be interpreted as follows. In the samples of this upper sub-category, the receiver touched a ball, so these samples are regarded as being similar to R-pass in which someone always touches a ball. In the samples of this bottom sub-category, the receiver focuses on but never touches the ball. Therefore, the team on the right side can get the score if the ball goes out of the court. This relationship with the scoring possibility in the bottom sub-categories makes the closeness with R-winpoint.

Furthermore, the effectiveness of our location-guidance for the fine granularity of our GAF is validated in Fig. 17. This visualization shows that similar but subtly different group activities (e.g., the location of spiker is different in R-spike and L-spike while both represent spike activity) are

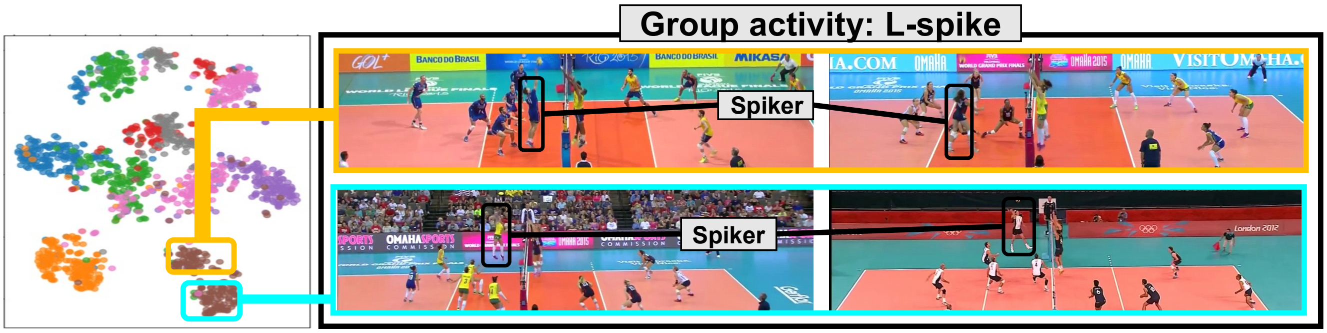


Figure 15. Visualization of the learned GAF by t-SNE in GAFL-PAC. The brown data points (i.e., "L-spike") are divided into two sub-categories based on the context (i.e., where the spiker is located in the left side court).

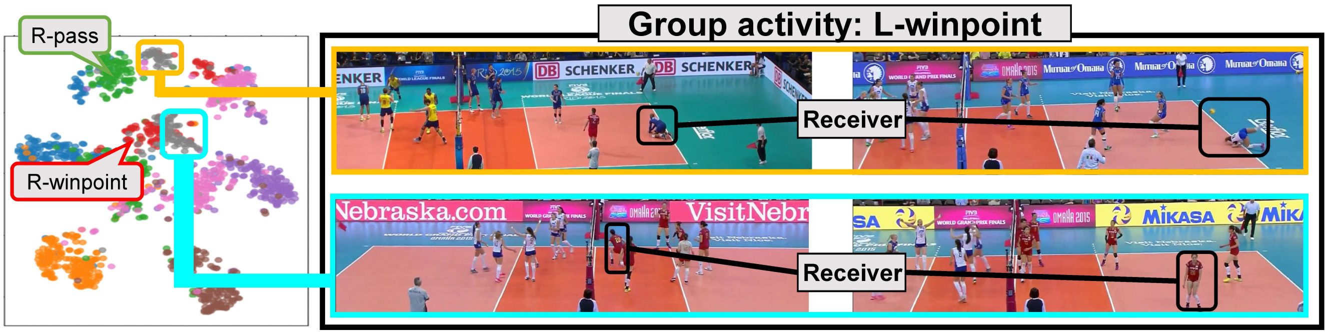
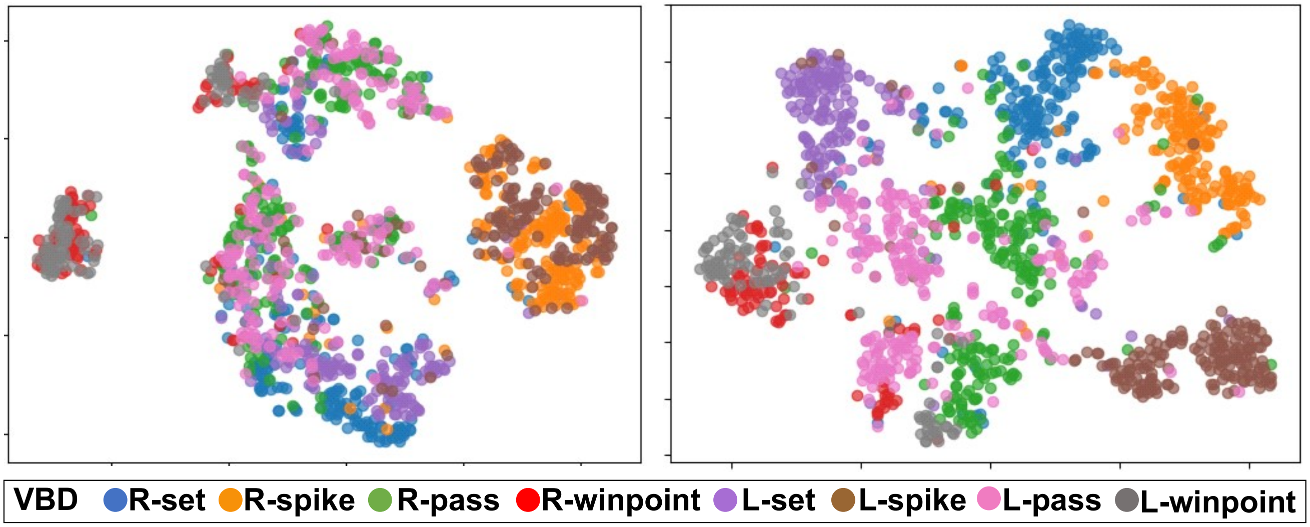


Figure 16. Visualization of the learned GAF by t-SNE in GAFL-PAC. The gray data points (i.e., "L-winpoint") are divided into two sub-categories based on the context (i.e., whether the receiver touched the ball or not).

not separated in "Ours w/o $\boldsymbol{F}_{loc}$." From the results, we can interpret that our location-guidance is essential for learning where the group activity is happening, as also confirmed in Fig. 15.



Ours w/o $\boldsymbol{F}_{loc}$

Ours

Figure 17. Effectiveness of our location-guidance in our GAF on the VolleyBall Dataset (VBD) in GAFL-PAC. Results obtained by "Ours-grp" are regarded as "Ours."