<Supplementary Material> Spatio-Temporal Proximity-Aware Dual-Path Model for Panoramic Activity Recognition

Sumin Lee, Yooseung Wang, Sangmin Woo, and Changick Kim

Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
{suminlee94, yswang, smwoo95, changick}@kaist.ac.kr

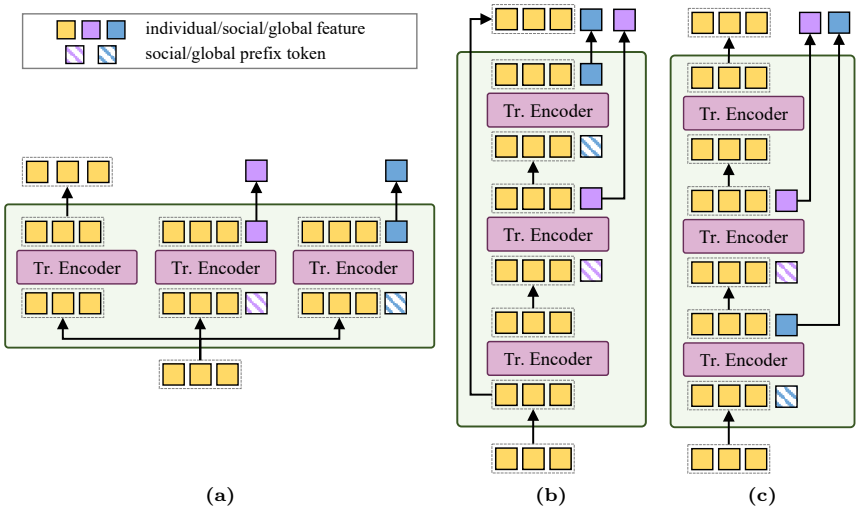


Fig. 1: A detailed overview of (a) parallel, (b) hierarchical, and (c) reverse hierarchical architectures.

A Ablation Experiments

In this section, we conducted additional experiments to demonstrate the effectiveness of the proposed method. We evaluated the experiments for Individual Action Recognition (IAR), Social Group Activity Recognition (SGAR), GloBal Activity Recognition (GBAR), and Social Group Detection (SGDet).

A.1 Dual-Path Activity Transformer

In Sec. 4.3 in the manuscript, we design three types of transformer structures to compare with the proposed Dual-Path Activity Transformer (DPATr). Figure 1 illustrates the detailed mechanism of the ablated architectures: parallel,

Table 1: Ablation experiments on the spatio-temporal individual self-attention. A^t, A^h, and A^w indicate attentions along temporal, height, and width dimensions, respectively. The best scores are marked in **bold** and the second best ones are underlined.

A^t	A^h	A^w	Individual Action			Social Activity			Global Activity			Overall
			$\mathcal{P}_i$	$\mathcal{R}_i$	$\mathcal{F}_i$	$\mathcal{P}_p$	$\mathcal{R}_p$	$\mathcal{F}_p$	$\mathcal{P}_g$	$\mathcal{R}_g$	$\mathcal{F}_g$	$\mathcal{F}_a$
			51.9	38.9	42.7	29.8	23.9	25.6	54.0	39.2	44.0	37.4
	✓	✓	52.4	46.7	47.0	30.6	30.4	29.2	56.2	44.8	48.3	41.5
✓			56.4	44.1	47.5	29.6	28.5	27.8	54.2	42.1	45.6	40.3
✓	✓		57.8	49.5	51.0	31.3	30.2	29.6	59.2	42.8	48.2	42.9
✓		✓	57.5	**50.4**	51.4	32.3	32.4	31.1	58.0	42.5	47.2	43.2
✓	✓	✓	**59.4**	49.7	**51.8**	**36.5**	**34.8**	**34.2**	**63.4**	**48.8**	**53.5**	**46.5**

hierarchical, and reverse hierarchical. Each of these models consists of three transformer encoder blocks [5] dedicated to enhancing features of individual actions, social group activities, and global activities, respectively. In the parallel architecture, these blocks operate independently to capture features related to specific granular activities from the self-attended individual features $\bar{F}^{idv}$. The hierarchical structure sequentially extracts activity information from smaller to larger spatial granularity. In contrast, the reverse hierarchical structure operates conversely, extracting activity information from larger to smaller spatial granularity. As illustrated in Fig. 3.a in the manuscript, each DPATr layer comprises an individual-to-global path and an individual-to-social path. In the individual-to-social path, richer social-level representations are extracted by leveraging the global-local context explored in the individual-to-global path. By mutually reinforcing contextual understanding of multi-spatial activities through multiple layers, SPDP-Net achieves the most superior performances across all metrics (see Sec. 4.3 in the manuscript).

A.2 Spatio-Temporal Individual Attention.

We ablate three attentions across temporal, height, and width axes in the proximity-based relation encoding. The results are shown in Table 1. We observe that SPDP-Net with either the temporal attention A^t or the spatial attention (*i.e.*, A^h and A^w) improves the performances by exploiting informative action features of each individual. Incorporating either A^h or A^w with A^t improves the performance of PAR, particularly IAR and GBAR. Specifically, compared to solely using A^t, a combination of A^t and A^h improves the performances of IAR and GBAR by 3.5% and 2.6% in F1 score, respectively. Similarly, using A^t and A^w achieves 3.3% improvement of SGAR and 1.6% improvement of GBAR, in terms of F1 score. By applying both spatial and temporal attentions, our SPDP-Net achieves the best overall performance, resulting 46.5%.

Moreover, we compare the performances regards to the order of the temporal, height, and width attention in Table 2. We determine the optimal axis order as $T \rightarrow H \rightarrow W$.

Table 2: Ablation study on axis orders in the individual attention

Order	$\mathcal{F}_i$	$\mathcal{F}_p$	$\mathcal{F}_g$	Order	$\mathcal{F}_i$	$\mathcal{F}_p$	$\mathcal{F}_g$
H → W→ T	50.1	30.5	51.6	W → H→ T	49.1	30.8	51.6
H → T→ W	50.6	30.3	51.0	W → T→ H	50.4	30.6	52.3
T → W→ H	52.7	33.8	53.4	T → H→ W	**51.8**	**34.2**	**53.5**

Table 3: Ablation experiments on clustering algorithm, spectral clustering and K-means clustering. The best scores are marked in **bold**.

clustering	IoU@0.5	IoU@AUC	Mat.IoU
Spectral [6]	49.1	34.8	27.7
K-means	**56.4**	**42.5**	**34.4**

A.3 Clustering Algorithms

Compared with previous works [1, 3] employing a graph-based Spectral clustering [6], we utilize a parametric-based clustering scheme with the predicted number of the social groups. Table 3 shows the results of SPDP-Net with Spectral clustering and K-means clustering, which is a parameter-based method. SPDP-Net using K-means clustering outperforms using Spectral clustering in SGDet. In particular, K-means clustering demonstrates performance improvements across various metrics: achieving 56.4% in IoU@0.5, 42.5% in IoU@AUC, and 34.4% in Mat.IoU. These results signify enhancements of 7.3%, 7.7%, and 6.7%, respectively. Spectral clustering encounters challenges in determining the optimal cluster number due to its sensitivity to the kernel function. Additionally, it may face scalability and stability issues. For these reasons, K-means clustering, which utilizes the predicted number of clusters, exhibits greater robustness than spectral clustering in social group activity detection in a crowded scene.

A.4 Loss functions

We ablate the auxiliary loss L_{aux} and the relation loss L_R functions and summarize the results in Table 4. While L_{aux} encourages the individual self-attention to learn individual action information, L_R drives the visual similarity matrix R_s to capture social relationships among individuals. While solely using L_R results in slight improvements in multi-granular activity recognition compared with the baseline, utilizing $\mathcal{L}_{aux}$ achieves performance improvements by 3.9% in $\mathcal{F}_i$, 2.9%p in $\mathcal{F}_p$, and 3.2% in $\mathcal{F}_i$. With cooperatively synergistic effects of $\mathcal{L}_{aux}$ and $\mathcal{L}_R$, SPDP-Net achieves the best performance in both multi-granular activity recognition and social group detection.

A.5 Balancing Hyper-parameters

In Table 5, we summarize the results of experiments with varying balancing hyper-parameters of the individual action loss $\mathcal{L}_{idv}$, the relation loss $\mathcal{L}_R$, and the auxiliary loss $\mathcal{L}_{aux}$ functions. When λ_{idv} and λ_{aux} are increased, we observe

Table 4: Ablation experiments on the auxiliary loss $\mathcal{L}_{aux}$ and the relation loss $\mathcal{L}_R$ functions. The best scores are marked in **bold**.

L_{aux}	L_R	Individual Action			Social Activity			Global Activity			IoU@0.5
		$\mathcal{P}_i$	$\mathcal{R}_i$	$\mathcal{F}_i$	$\mathcal{P}_p$	$\mathcal{R}_p$	$\mathcal{F}_p$	$\mathcal{P}_g$	$\mathcal{R}_g$	$\mathcal{F}_g$	
		50.7	45.5	45.5	29.4	29.7	28.3	58.6	43.3	48.4	51.8
	✓	53.2	48.7	47.8	30.4	29.9	28.9	59.9	43.9	49.2	51.4
✓		56.7	47.3	49.4	33.9	31.0	31.2	60.4	47.7	51.6	53.3
✓	✓	**59.4**	**49.7**	**51.8**	**36.5**	**34.8**	**34.2**	**63.4**	**48.8**	**53.5**	**56.4**

Table 5: Ablation experiments on the balancing parameters between loss functions.λ_{idv}, λ_R, and λ_{aux} represents weights of the individual action loss, the relation loss, and the auxiliary loss functions, respectively

$\lambda_{idv} : \lambda_R : \lambda_{aux}$	Activity Recognition				Social Group Detection		
	$\mathcal{F}_i$	$\mathcal{F}_p$	$\mathcal{F}_g$	$\mathcal{F}_a$	IoU@0.5	IoU@AUC	Mat.IoU
1 : 1 : 3	53.1	30.4	55.2	46.2	50.0	35.8	28.1
1 : 3 : 1	50.0	32.1	52.6	44.9	56.1	42.0	34.6
3 : 1 : 1	52.5	29.8	52.9	45.1	51.7	39.3	29.8
1 : 1 : 1	51.8	34.2	53.5	46.5	56.4	42.5	34.3

that a slight improvement in IAR and GBAR, but a significant performance decrease in SGAR. Conversely, when λ_R is increased, the overall performance $\mathcal{F}_a$ is decreased by 1.6%. When the proportions of λ_{idv}, λ_R, and λ_{aux} are equal, SPDP-Net achieves the best performance in social group detection performance and overall multi-granular activity recognition ($\mathcal{F}_a$).

A.6 Comparison on Social-CAD Dataset

We additionally evaluate SPDP-Net on Social-CAD dataset [2] and summarize the results in Table 6. SPDP-Net achieves 67.7%, 96.3%, and 97.7% accuracy in individual, social-group, and global activity recognition, which are comparable to or higher than existing methods. Note that videos in Social-CAD have more narrower views than JRDB-PAR [3].

B Additional Visualization

B.1 Relation Matrix

Figure 2 shows more visual comparisons between the ground truth and predicted social relation matrix R with the proximity relation matrix R_p and the similarity matrix R_s. Those matrices have 1 for individuals belonging to the same social group and 0 for otherwise. We note that R_p closely corresponds to the ground-truth social relation compared to R_s (see Fig. 2a, 2b, 2c, and 2d). In contrast, in panoramic scenes with relatively larger bounding boxes and fewer people, we observe that R_s is effective (see Fig. 2e and 2f).

Table 6: Accuracy comparison on Social CAD dataset.

Method	Individual	Social Group	Global
Tamura [4]	66.6	96.3	-
Composer [7]	-	-	96.2
Ours	67.7	96.3	97.7

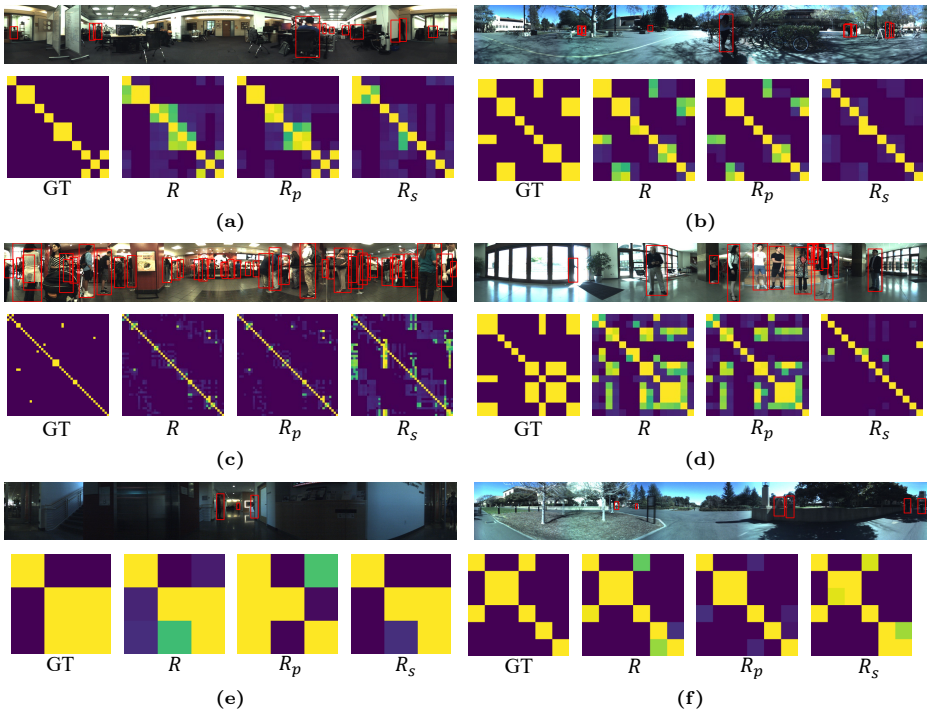


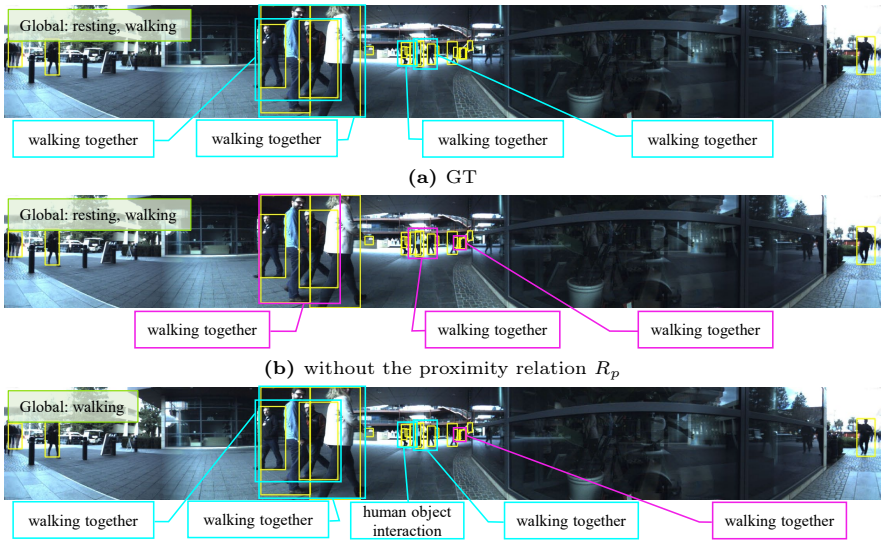
Fig. 2: Visualization of the ground-truth (GT) and predicted relation matrix R, the proximity relation matrix R_p, and the similarity matrix R_s. Best viewed zoomed in on screen.

B.2 Prediction Results

Figure 3, 4, 5, and 6 shows the visual comparisons between the ground-truth and SPDP-Net with and without the social proximity relation R_p on JRDB-PAR dataset [3]. We note that the absence of utilizing R_p leads to inaccurate or missed detections of social groups.

C Limitation and Future Work.

There are still unresolved problems. In Table 6 in the manuscript, we observed that the performance of social group activity detection is enhanced when using the ground-truth number of social groups. To address this, it is necessary to



(a) GT

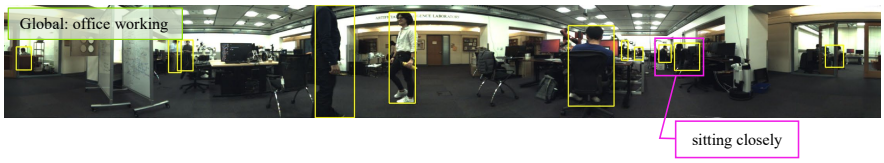
(b) without the proximity relation R_p

(c) Ours

Fig. 3: Visual comparisons of the social group activity detection and global activity recognition between of (a) ground-truth, (b) SPDP-Net without the social proximity relation R_p, and (c) SPDP-Net. Misclassified social group detections are indicated in magenta, while ground-truth and correctly predicted bounding boxes are in cyan.

develop strategies to adjust to varying group densities and complexities. Moreover, real-world datasets, such as JRDB-PAR [3], exhibit significant biases in their class distributions. Overcoming these biases is crucial for improving the robustness and generalization of the proposed method. We leave this intriguing challenge to future work.

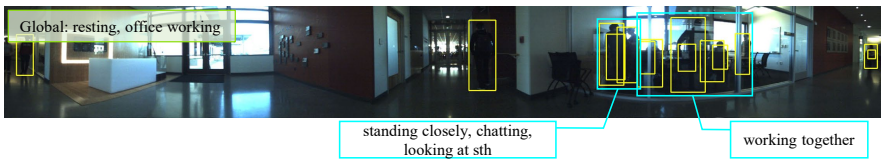
(a) GT



(b) without the proximity relation R_p

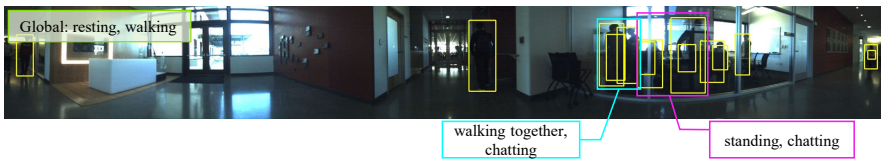
(c) Ours

Fig. 4: Visual comparisons of the social group activity detection and global activity recognition between of (a) ground-truth, (b) SPDP-Net without the social proximity relation R_p, and (c) SPDP-Net. Misclassified social group detections are indicated in magenta, while ground-truth and correctly predicted bounding boxes are in cyan.



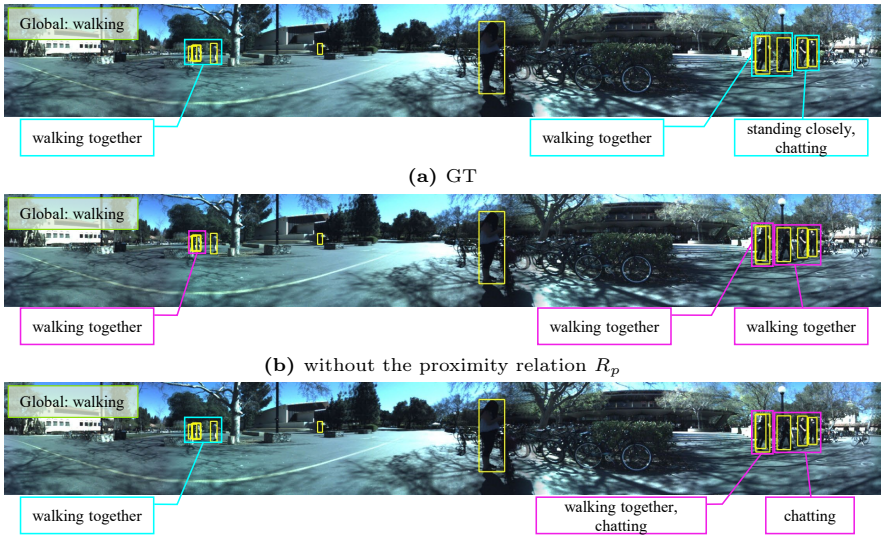
(a) GT

(b) without the proximity relation R_p



(c) Ours

Fig. 5: Visual comparisons of the social group activity detection and global activity recognition between of (a) ground-truth, (b) SPDP-Net without the social proximity relation R_p, and (c) SPDP-Net. Misclassified social group detections are indicated in magenta, while ground-truth and correctly predicted bounding boxes are in cyan.



(a) GT

(b) without the proximity relation R_p

(c) Ours

Fig. 6: Visual comparisons of the social group activity detection and global activity recognition between of (a) ground-truth, (b) SPDP-Net without the social proximity relation R_p, and (c) SPDP-Net. Misclassified social group detections are indicated in magenta, while ground-truth and correctly predicted bounding boxes are in cyan.

References

1. Cao, M., Yan, R., Shu, X., Zhang, J., Wang, J., Xie, G.S.: Mup: Multi-granularity unified perception for panoramic activity recognition. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7666–7675 (2023)
2. Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C.G.: Actor-transformers for group activity recognition. In: CVPR. pp. 839–848 (2020)
3. Han, R., Yan, H., Li, J., Wang, S., Feng, W., Wang, S.: Panoramic human activity recognition. In: ECCV. pp. 244–261. Springer (2022)
4. Tamura, M.: Design and analysis of efficient attention in transformers for social group activity recognition. IJCV pp. 1–20 (2024)
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
6. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D.: Two-person interaction detection using body-pose features and multiple instance learning. In: CVPRW. pp. 28–35 (2012)
7. Zhou, H., Kadav, A., Shamsian, A., Geng, S., Lai, F., Zhao, L., Liu, T., Kapadia, M., Graf, H.P.: Composer: compositional reasoning of group activity in videos with keypoint-only modality. In: ECCV. Springer (2022)