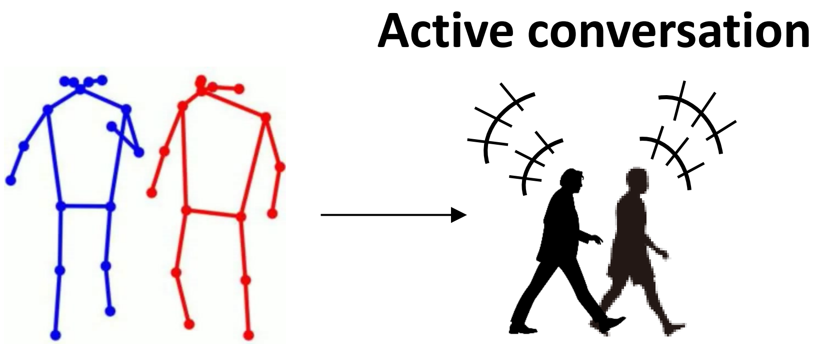
Reducing Computational Cost in Pedestrian Conversation Activity Recognition through Skeleton Spatiotemporal Graphs

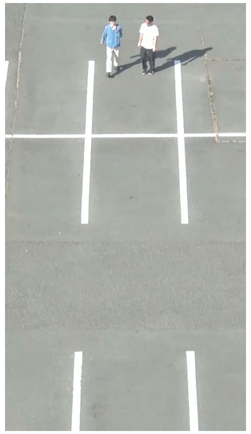
Tsubasa Kondo∗, Michiko Inoue∗, Shunsuke Yoneda†
and Masashi Nishiyama∗

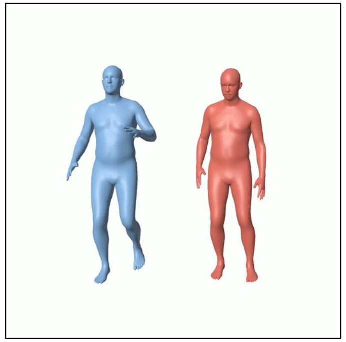
∗Graduate School of Sustainability Science, Tottori University, Japan
†Organization for Information Strategy and Management, Tottori University, Japan



Introduction

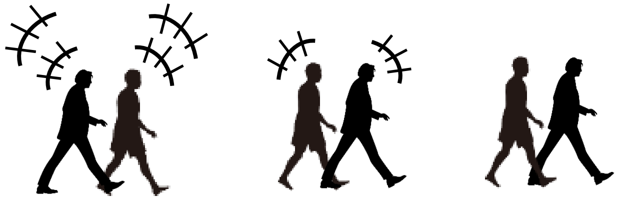
- A growing demand exists for a technique that automatically recognizes conversation activity inside pedestrian groups walking outdoors.
- Only one existing method has addressed conversation activity recognition for walking groups using video sequences. [Ganaha+, ICPR'24]





Interaction video sequence

Conversation activity



Active Inactive None

✓ In addition to its high accuracy, this design enables developers to visually confirm which body parts in the 3D model contribute to conversation activity recognition.

- This method uses a succession of whole-body movements, termed body interaction, as the visual features based on McNeill's finding that gestures are bodily movements that accompany speech and are helpful for the analysis of conversation. [Mcneill, the University of Chicago Press'94]

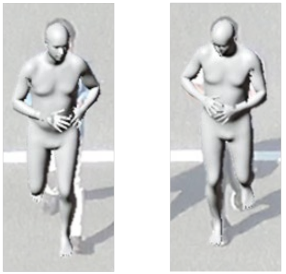
Bottlenecks in the existing method

The existing method was not designed for scenarios with
[Ganaha+, ICPR'24]
a limited computation time and GPU memory usage .

Bottlenecks

The existing method estimates the SMPL model pose-and-shape parameters and converts them into a mesh structure.

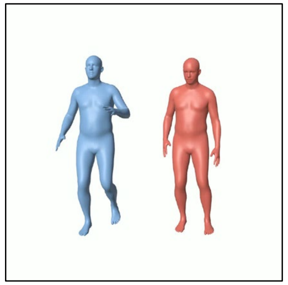
- Although low-dimensional SMPL parameters can be inferred stably and accurately, the computation required is high.



SMPL model

The existing method renders each mesh structure onto image planes to form an interaction video sequence.

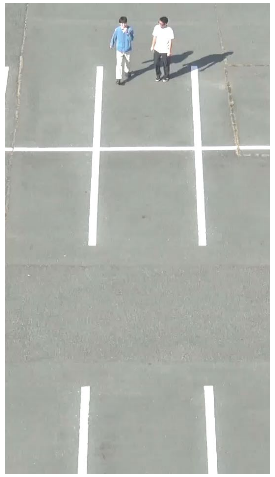
- The rendering cost increases as the video sequence length increases. It demands extra computation and GPU memory.

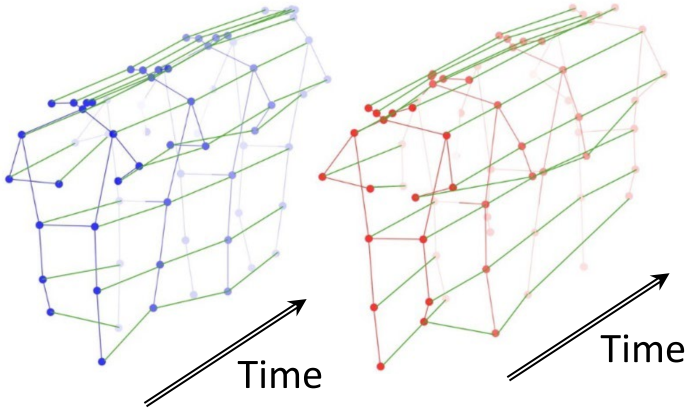


Interaction video sequence

Purpose

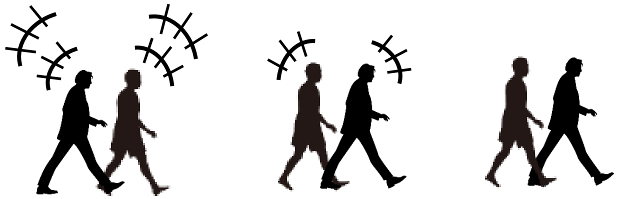
We encode the pedestrian group's body gesture interaction as a skeleton spatiotemporal graph built from body-joint keypoints and evaluate its effectiveness for conversation activity recognition.





✓ Skeleton spatiotemporal graph

Conversation activity



Active Inactive None

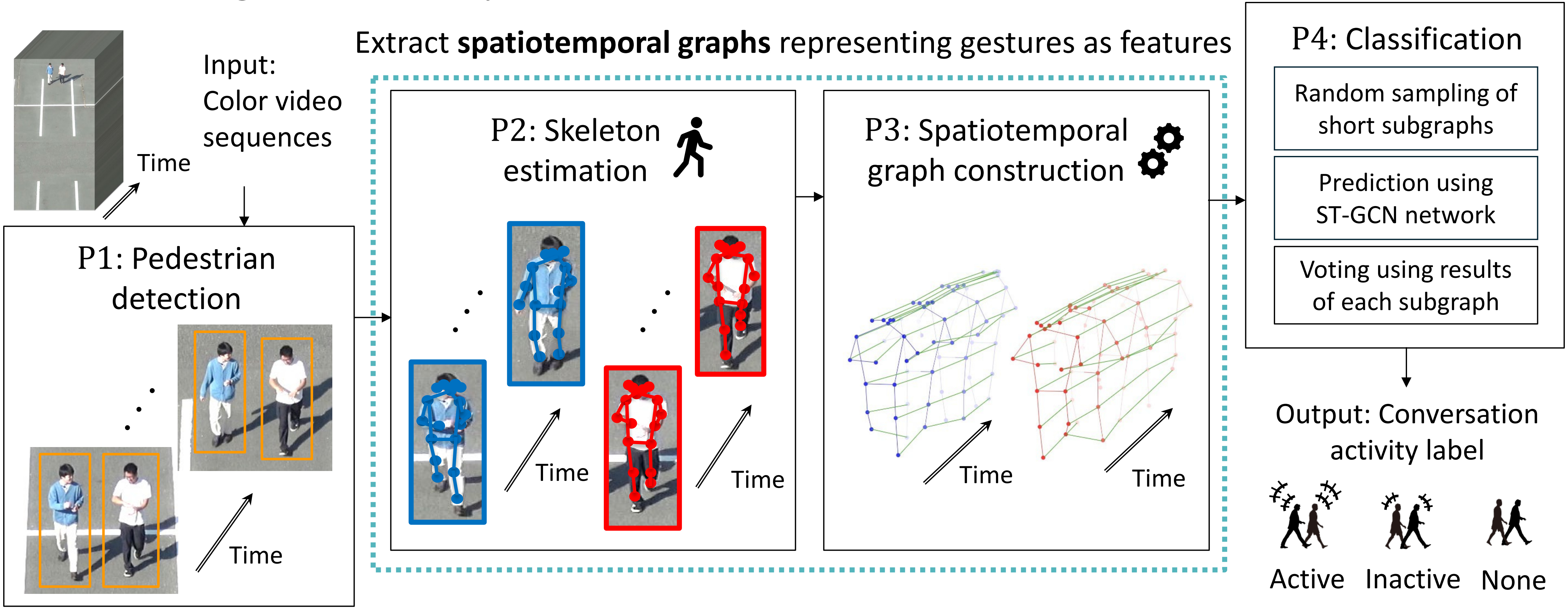
[I1] Body movement parameter inference stage
[I2] Feature extraction stage

Our method lowers computation time and GPU memory usage compared with the existing method.

We confirm that our method achieves recognition accuracy on a par with, or superior to, the existing method.

Overview of the proposed method

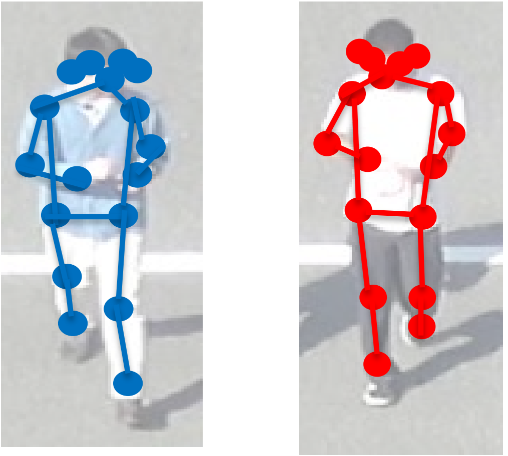
We design a new approach to achieve points I1 and I2 while preserving the recognition accuracy.



Skeleton for body movement parameter inference

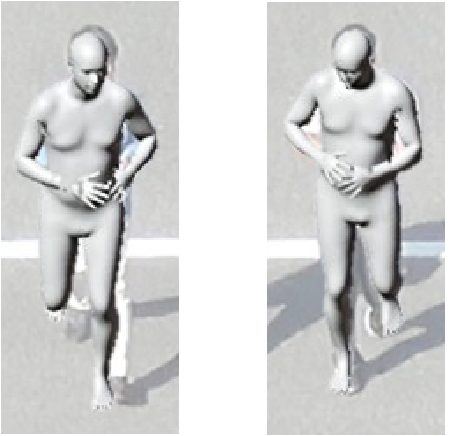
- We represent the skeleton as the image-plane positions of keypoints, such as the center of mass of the head and the body joints, together with connectivity among these keypoints.
  - The body can be represented by the skeleton at each time point without the computationally intensive, high precision estimation of SMPL pose-and-shape parameters.
  - We expect to reduce the computation time required to estimate the bodies of pedestrians in a video sequence.

At each time point t



At each time point t

✓ Skeleton

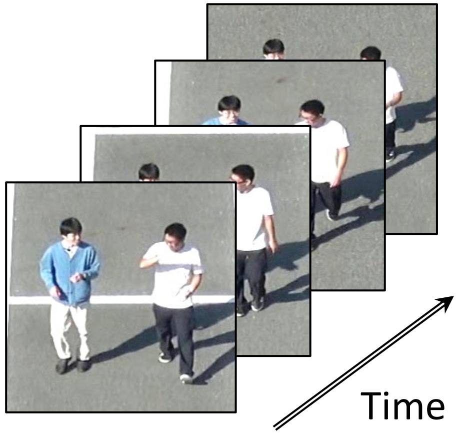


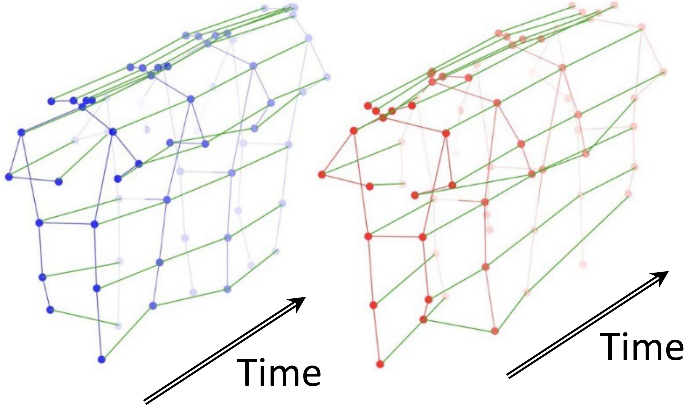
At each time point t

✗ SMPL model

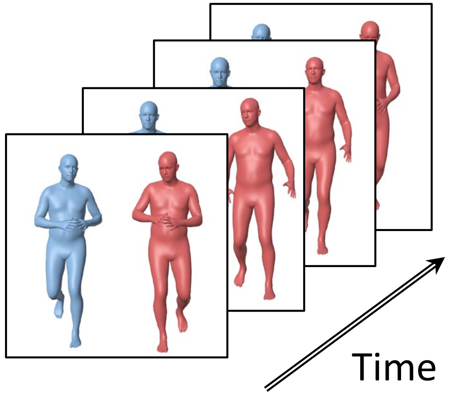
Skeleton spatiotemporal graph for feature extraction

- Our method suppresses the large amount of GPU memory usage by linking skeletons along the temporal axis to form a skeleton spatiotemporal graph.
  - Within the same frame of the video sequence, we connect the keypoints of each skeleton in the spatial domain, and between temporally adjacent frames, we connect the corresponding keypoints in the temporal domain, thereby constructing the graph.
  - The skeleton spatiotemporal graph **expresses body interaction without performing 3D rendering**, and thus this graph reduces both computation time and GPU memory usage.





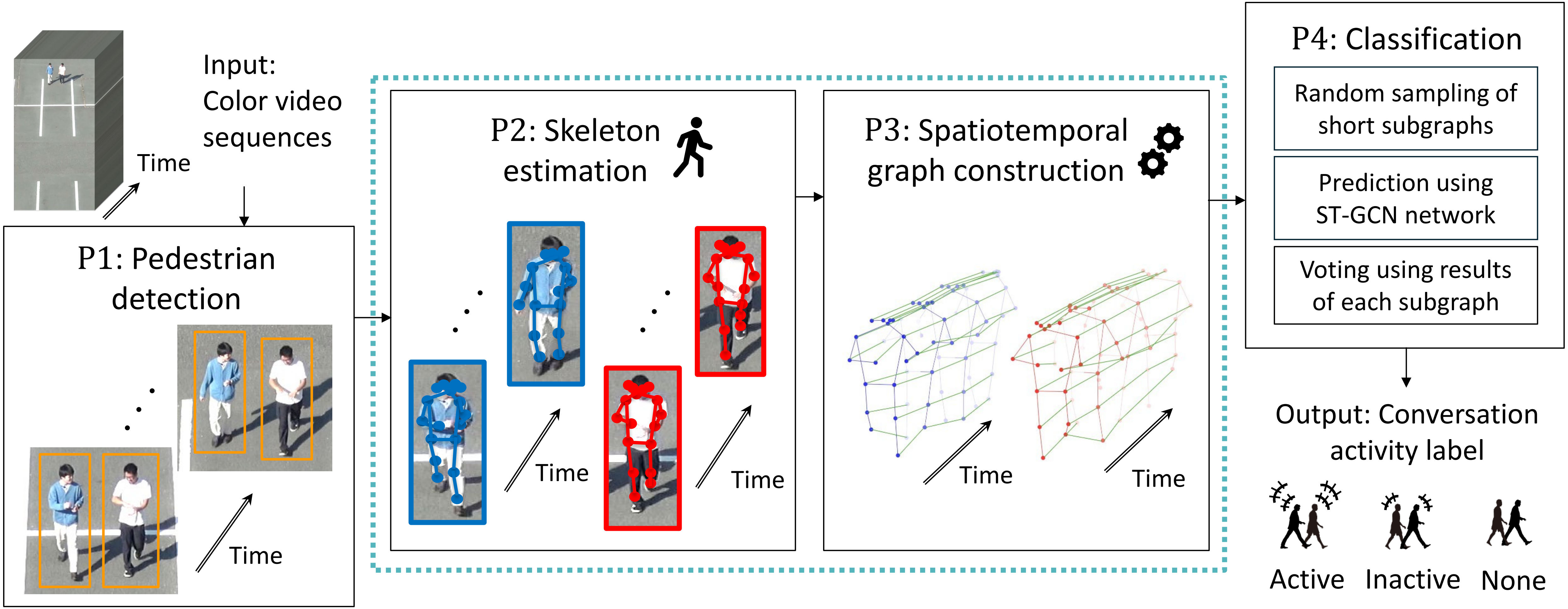
✓ Skeleton spatiotemporal graph



✗ Interaction video sequence

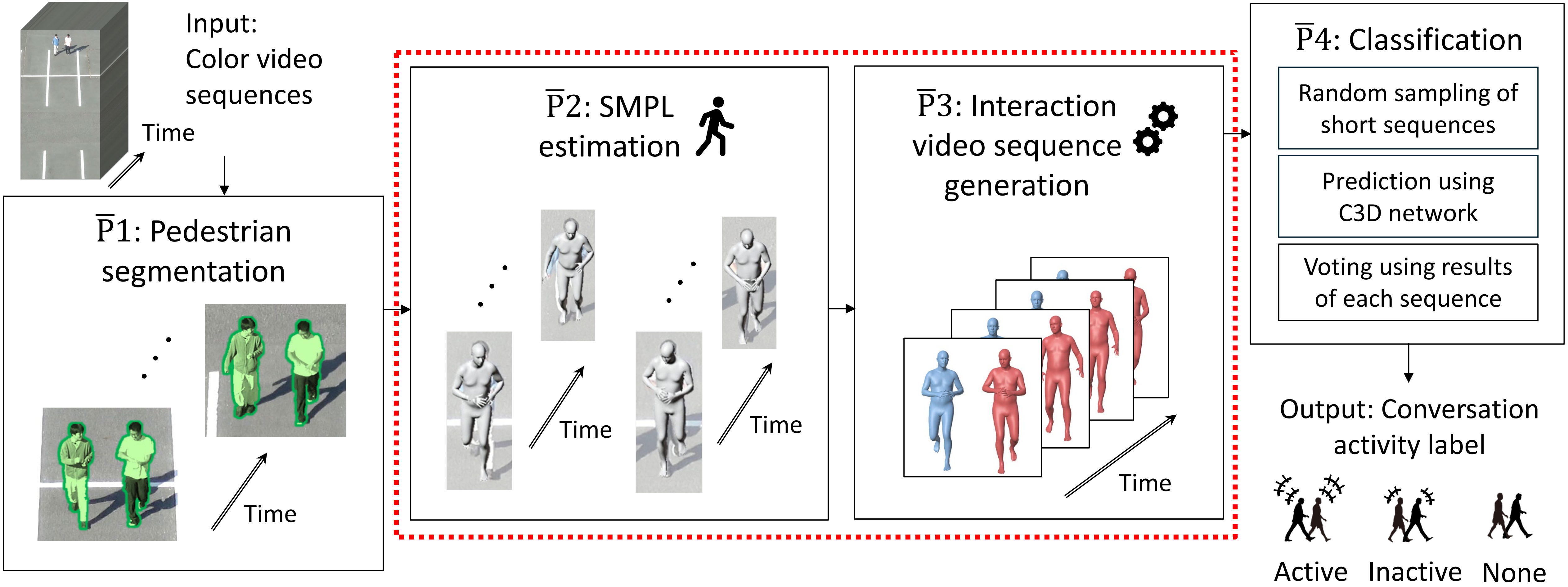
Overview of the proposed method (Reprinted)

Extract **spatiotemporal graphs** representing gestures as features



Overview of the existing method [Ganaha+, ICPR'24]

Extract **interaction video sequences** representing gestures as features



The dataset used in the experiment

- To verify the effectiveness of our method, we used the dataset collected in experiments for the existing method. [Ganaha+, ICPR'24]
- We prepared three conversation activity labels.

Active

The active label indicates that the pedestrian group is engaged in a lively conversation on topics of mutual interest.

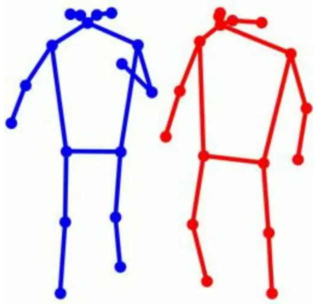
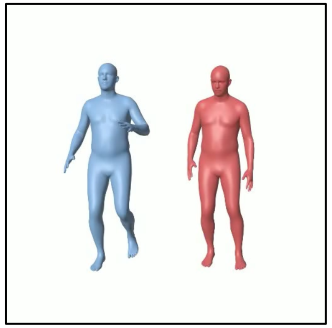
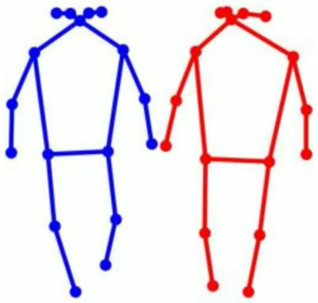
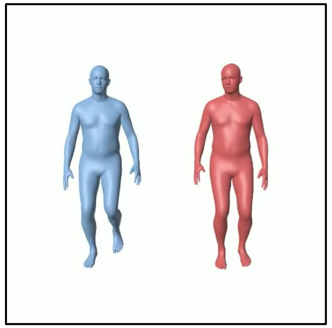
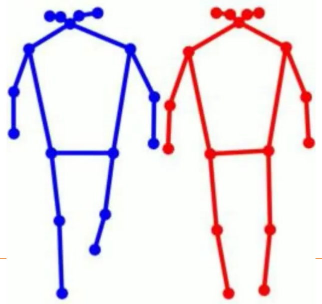
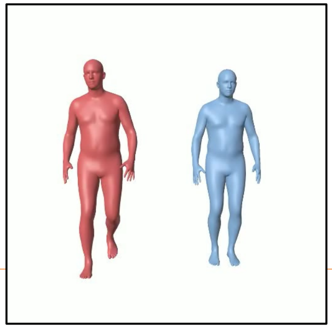
Inactive

The inactive label indicates that the group is not engaged in a lively conversation, for example, because the topic is of little interest.

No conversation

The no conversation label indicates that no conversation is taking place.

Characteristics of gestures in conversation activity

		Color video sequence	Skeleton	Interaction video sequence
	**Active**: Large gestures occur at a high frequency			
	**Inactive**: Small gestures occur occasionally			
	**No conversation**: Gestures almost never occur			

How to collect conversation activity labels

When collecting color video sequences, we only instructed the participants on the topic of the conversation and did not give any explanation or instructions regarding the physical body interaction.

Active

We instructed the participants to **introduce their hobbies** while walking.

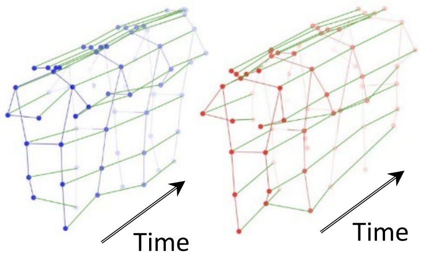
Inactive

We instructed the participants to talk about **topics of little interest** to each other while walking. The topic was chosen by the participants from several candidate topics prepared in advance (e.g.,economic situation and political situation in a country that the participants had never visited and had almost no knowledge of).

No conversation

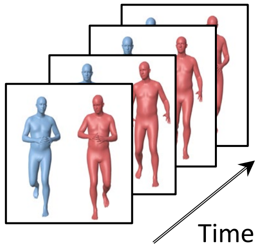
We instructed the participants **not to engage in any conversation** while walking.

Comparison of computation time



Our method		Time [seconds per frame] ↓
P1	Pedestrian detection	0.085
P2	Skeleton estimation	**0.021**
P3	Spatiotemporal graph construction	**0.000**
P4	Classification	*0.001
Total		0.107

*Seconds per short subgraph

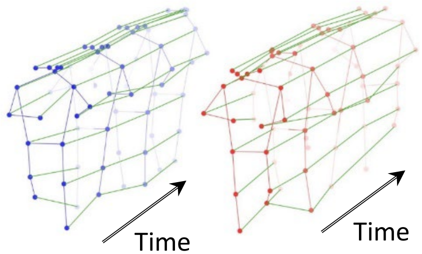


Existing method [Ganaha+, ICPR'24]		Time [seconds per frame] ↓
$\overline{\mathrm{P}}1$	Pedestrian segmentation	0.249
$\overline{\mathrm{P}}2$	SMPL estimation	0.326
$\overline{\mathrm{P}}3$	Interaction video sequence generation	0.542
$\overline{\mathrm{P}}4$	Classification	*0.007
Total		1.124

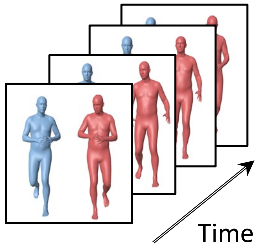
*Seconds per short video sequence

When the computation times were summed, our method was approximately 10.5 times faster than the existing method.

Comparison of GPU memory usage



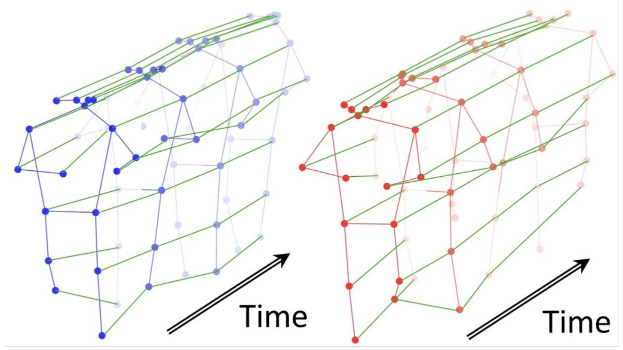
Our method		Memory usage [MiB] ↓
P1	Pedestrian detection	978
P2	Skeleton estimation	**1002**
P3	Spatiotemporal graph construction	**0**
P4	Classification	1194
Total		3174



Existing method [Ganaha+, ICPR'24]		Memory usage [MiB] ↓
$\overline{\mathrm{P}}1$	Pedestrian segmentation	1020
$\overline{\mathrm{P}}2$	SMPL estimation	1220
$\overline{\mathrm{P}}3$	Interaction video sequence generation	1227
$\overline{\mathrm{P}}4$	Classification	1690
Total		5157

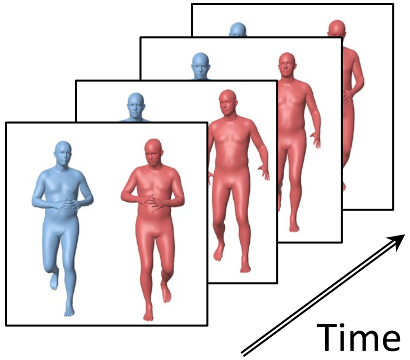
A comparison of total GPU memory usage indicated that our method reduced the requirement to approximately 61% of that of the existing method.

Comparison of conversation activity recognition accuracies



Accuracy [%] ↑

Our method	**80.3±0.8**



Accuracy [%] ↑

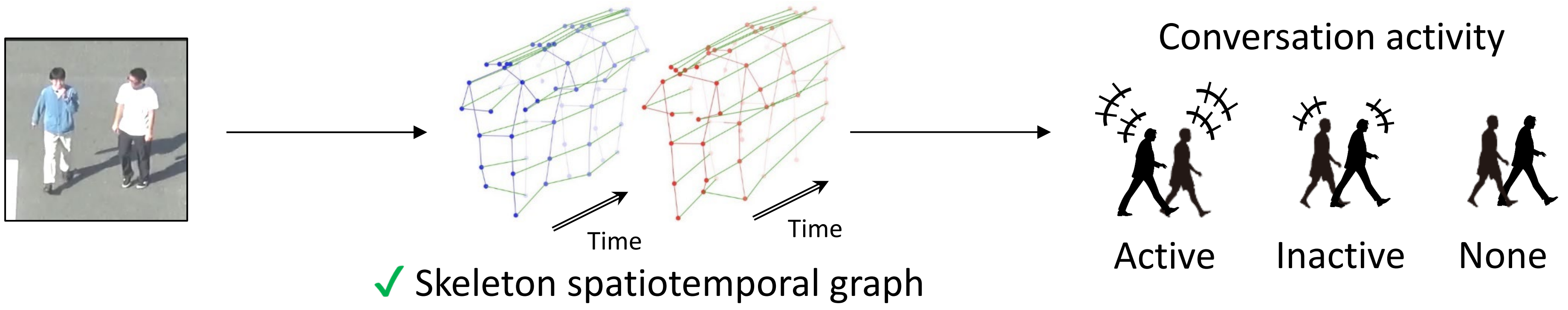
Existing method	76.2±0.7

[Ganaha+, ICPR'24]

The results confirm that our method's recognition accuracy was comparable to or higher than that of the existing method.

Conclusions

- We investigated the effectiveness of a technique that recognizes conversation activity using skeleton spatiotemporal graphs, estimated from color video sequences, as informative and compact features.
- The experimental results confirmed that our method significantly reduced computation time and GPU memory usage while achieving recognition accuracy comparable to or higher than the existing method.



- Future work:
  - We intend to apply conversation activity recognition to various video sequences collected in more practical scenarios.
  - We extend the framework to pedestrian groups with three or more members and scenes in which group membership changes over time.
  - We investigate adaptive spatiotemporal graph construction that handles occlusion, missing keypoints, and large pose variations.