2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Worker Tracking Using Skeletal Graphs for Agricultural Support Robot in Narrow Furrows

Kosuke Murakami[1], Akihisa Ohya[2], and Ayanori Yorozu[3]

Abstract— **This study proposes an innovative approach for tracking agricultural workers in narrow furrows, aimed at enhancing the performance of agricultural support robots. The method integrates RGB-D camera-based skeleton extraction with a Space-Time-Separable Graph Convolutional Network (STS-GCN) for lower limb motion prediction, and introduces a novel fusion algorithm that combines these predictions with real-time observations. Experimental results demonstrate the proposed method's ability to maintain accurate worker tracking even in occluded scenarios by complementing observations with predictions. This research provides insights into the effectiveness of combining skeletal graph-based motion prediction with real-time observations for robust worker tracking in challenging agricultural environments.**

I. Introduction

The aging and declining population of agricultural workers in Japan has become a critical issue with far-reaching implications for food security and rural economies [1]. To address this challenge, the development of robots capable of recognizing and following human workers in agricultural fields has gained significant attention. These robots aim to provide direct support and augment human capabilities, potentially alleviating labor shortages and increasing agricultural productivity.

However, the implementation of such robots in narrow crop furrows presents unique challenges. As illustrated in Fig. 1, these robots must navigate between crop rows while accurately determining the worker's position to avoid damaging crops. This task is further complicated by frequent occlusions caused by dense foliage, as shown in Fig. 2, where traditional camera-based recognition and tracking often fail.

To address these challenges, this study proposes a novel approach for robust worker tracking in narrow crop furrows using skeletal graphs and deep learning techniques. Our method integrates RGB-D camera-based skeleton extraction with a Space-Time-Separable Graph Convolutional Network (STS-GCN) for lower limb motion prediction. We further introduce an innovative fusion algorithm that combines these predictions with real-time observations to maintain robust tracking even under severe occlusion.

The primary objectives of this research are to: 1) Develop a 3D skeletal estimation method that leverages temporal information for accurate identification of the tracking point, and 2) Achieve robust recognition in scenarios with partial occlusions, where only a portion of the worker's body is visible. The tracking accuracy of the 3D joint positions of the lower limbs using the proposed method is evaluated with observational data simulating scenarios where the worker is obscured from the camera in narrow furrows densely covered with crops.

Fig. 1: Scene of agricultural robot assisting with harvest transport by following worker

Fig. 2: Challenging scenario where human detection is difficult due to leaf occlusion

II. Related Works

A. Human Following Robots

Human following robots have been widely studied in various environments. In indoor settings, Mandischer et al. developed a radar-based leg tracking for service robots, achieving stable following in home environments [2]. Similarly, Kim et al. proposed a vision-based tracking for mobile robots in indoor spaces [3].

Compared to these approaches, agricultural environments as shown in Fig. 1 present unique challenges such as frequent occlusions by crops, requiring more robust tracking solutions. Arai et al. proposed an agricultural robot designed

[1]Kosuke Murakami is with the Institute of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan murakami-k@roboken.cs.tsukuba.ac.jp

[2]Akihisa Ohya is with the Institute of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan ohya@cs.tsukuba.ac.jp

[3]Ayanori Yorozu is with the Institute of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan yorozu@cs.tsukuba.ac.jp

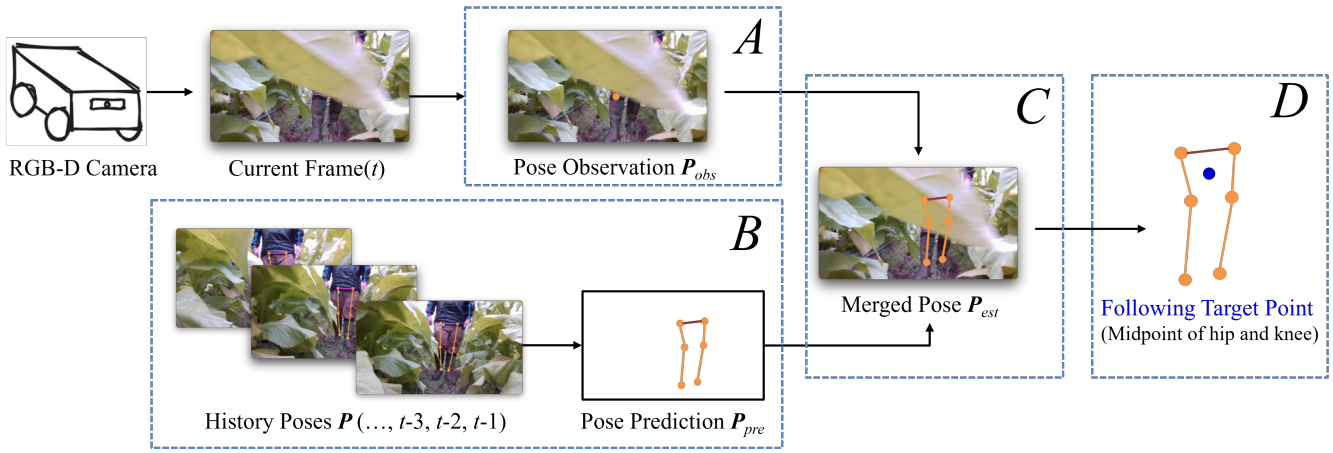


Fig. 3: Framework of the proposed method

Fig. 4: Example of correspondence between RGB image and lower limb skeleton.

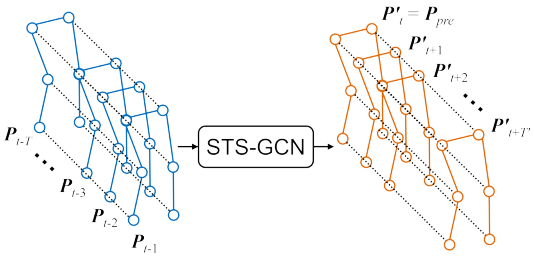


Fig. 5: Overview of the pose prediction pipeline. Detected lower limb joint positions are processed as skeletal sequences through STS-GCN to predict future poses.

to follow workers in narrow furrows during harvesting operations [4]. They addressed the challenge of detecting the worker's body center position using a deep learning model (PoseNet) [5] to detect worker joints from RGB images, even when only part of the body was visible due to the limited field of view of the RGB-D camera sensor. This approach demonstrated the feasibility of worker following in narrow furrows without the robot straying onto the crop beds. However, this method encounters difficulties when the worker's lower limbs are significantly occluded by crops, resulting in the inability to obtain observations of the tracking target, making it challenging for the robot to follow.

B. Human Pose Estimation Using Skeletal Graphs

Several researches have been proposed that models the human skeleton as a graph structure and applies Graph Neural Networks (GNN) [6] to predict movements. Yan et al. proposed a method that divides human skeletal data into spatial and temporal graphs, extracting features through Graph Convolution to consider both the relationships between joints and temporal changes [7]. Theodoros et al. built upon Yan et al.'s work, introducing a new graph convolutional network (STS-GCN) for human pose prediction [9]. Their method improved accuracy by processing temporal and spatial graph data in separate dimensions.

Yang et al. proposed an approach to learn pose dynamics (how posture changes over time) to enable estimation even in occluded states [8]. Their method performs human skeleton recognition on each frame, stores the results as history, and uses a GNN model to predict poses for each tracklet in the history memory based on the pose history and the current frame. However, in robot following scenarios, where often only the lower body is captured, a unique challenge arises. While [9] predicted relative motion in a coordinate system with the waist as the origin, motion prediction in the robot's coordinate system is needed, considering that the robot follows the human. Moreover, relying solely on predictions cannot adapt to changes in human movements.

Therefore, this study proposes applying a modified STS-GCN to predict lower limb motions in the robot coordinate system. In addition, a framework that integrates the prediction results with observation data to enable robust tracking even in occluded scenarios is proposed. This approach is crucial for practical robot following applications in various environments where only partial human body information is available.

III. Proposed Method

Based on the approaches of Yang and Theodoros, we propose a method for continuous worker tracking. Fig. 3 shows the framework of the proposed method. Our method consists of four main processes (A-D shown in Fig. 3).

A. 3D Pose Observation of Lower Limb

YOLOv8 model is used for initial 2D joint estimation, focusing on six key points: both hips, both knees, and both ankles. Fig. 4 shows an example of the detected lower limb skeleton overlaid on the corresponding RGB image. The 2D coordinates are transformed to robot's coordinate system then combined with depth information from an RGB-D camera to obtain six 3D joint positions $\boldsymbol{P}_{obs} \in \mathbb{R}^{6\times3}$.

B. GNN-based Motion Prediction from History Poses

Fig. 5 shows the overview of our pose prediction pipeline. The lower limb joints are detected from input images and processed as a sequence of skeletons through STS-GCN to predict future poses.

STS-GCN architecture is applied to focus specifically on lower limb motion prediction. Our modified model takes as input a sequence of lower limb joint positions in the robot's coordinate system, represented as $\boldsymbol{P} = \{\boldsymbol{P}_{t-T+1}, \ldots, \boldsymbol{P}_{t-1}\}$, where $\boldsymbol{P}_{t-1} \in \mathbb{R}^{6\times3}$ represents the 3D coordinates of the six lower limb joints at time $t-1$. The model outputs a sequence of predicted joint positions $\boldsymbol{P'} = \{\boldsymbol{P'}_t, \ldots, \boldsymbol{P'}_{t+T'-1}\}$, where $\boldsymbol{P'}_t := \boldsymbol{P}_{pre} \in \mathbb{R}^{6\times3}$ represents the predicted 3D coordinates of the lower limb joints at time *t*, also in the robot coordinate system. Here, *T* and *T'* denote the number of input and output frames, respectively.

C. 3D Pose Estimation by Prediction-Observation Merging

We propose a fusion method that integrates observed and predicted poses based on their confidence levels. The estimated (merged) pose $\boldsymbol{P}_{est} \in \mathbb{R}^{6\times3}$ at current time t. The estimated position of each joint $\boldsymbol{P}_{est,j}$ is calculated using the following equation:

$$\boldsymbol{P}_{est,j} = \omega_j \boldsymbol{P}_{obs,j} + (1-\omega_j)\boldsymbol{P}_{pre,j} \tag{1}$$

where $\boldsymbol{P}_{obs,j}$ and $\boldsymbol{P}_{pre,j}$ represent the observed and the predicted position of the j-th joint at time t, respectively. ω_j is determined by a sigmoid function:

$$\omega_j = \frac{1}{1+e^{-20(c_j-0.65)}} \tag{2}$$

where c_j is the confidence score of the observed j-th joint position. This weighting scheme smoothly transitions between prediction and observation based on the observation confidence. When observations are unavailable due to occlusions, the system naturally relies on predictions. The merged skeletal graph serves as input for the next prediction step, enabling continuous tracking through occluded sequences.

D. Computation of Following Target Point

In their research on robot following in narrow furrows, Arai et al. [4] defined the robot's following target point as the midpoint of the hip and knee coordinates to achieve accurate following while preventing the robot from mounting the furrows and damaging crops. We adopt this approach in our research. Based on this tracking point, we issue depth and speed instructions to enable the robot to follow the worker effectively.

IV. Experiments and Results

A. Experimental Overview

This study evaluates our proposed method for robust worker tracking in agricultural settings, which combines a modified STS-GCN model for lower limb motion prediction with a novel approach for merging predictions and observations. We conducted two main experiments:

1. Evaluation of the lower limb motion prediction model: We assessed the accuracy of our modified STS-GCN in predicting worker movements, particularly focusing on its performance in normal walking scenarios and its limitations in handling sudden direction changes.

2. Performance analysis of the integrated tracking: We evaluated our proposed method, which merges prediction results with real-time observations, in challenging scenarios including direction changes and occlusions. This experiment aimed to demonstrate the robustness of our approach in maintaining accurate tracking even when direct observations are partially or fully obstructed.

Fig. 6 shows the experimental environment. Assuming an RGB-D camera is mounted at the front of the robot, the camera center is defined as the robot's coordinate system shown in Fig. 6. The measurement accuracy of the tracking target point, based on the estimation of the lower limbs in the coordinate system is validated. For our agricultural robot following application, errors within 0.1 m for both horizontal (X-coordinate) and forward (Y-coordinate) measurements are considered acceptable. This tolerance was set based on the ability to avoid mounting the furrows while maintaining a safe following distance from the worker in the environment shown in Fig. 1. In this experimental setup, we focused exclusively on the walking motion of agricultural workers in furrows, without considering other work-related movements such as harvesting or pruning. While the basic motion is walking, our experiments include scenarios where the worker changes walking direction, which is a common occurrence in actual agricultural settings.

Through these experiments, we aim to validate the effectiveness of our integrated approach in addressing the unique challenges of worker tracking in narrow furrows, particularly in scenarios with frequent occlusions and varying worker movements.

B. Verification of Lower Limb Prediction

1) Experimental Setup: We created a dataset of approximately 1,400 frames of walking worker data using the Yolov8 model-x in Fig. 6. Our modified STS-GCN model was trained on this dataset to construct a prediction model. The training process focused on learning the dynamics of lower limb motions in the robot coordinate system, allowing for accurate predictions even when only partial body information is available. We employed a sequence-to-sequence learning approach, where the model learns to predict future frames based on a sequence of input frames, all within the robot's frame of reference.

The computations were performed on a desktop PC with Intel Core i7-13700KF CPU and NVIDIA GeForce RTX

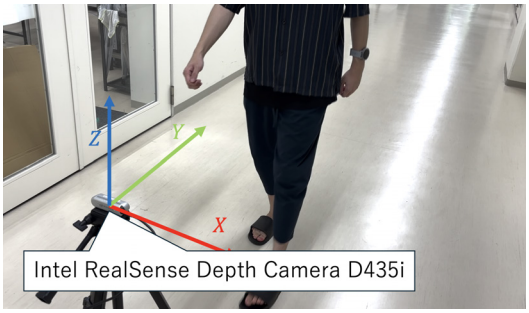


Fig. 6: The experimental situation

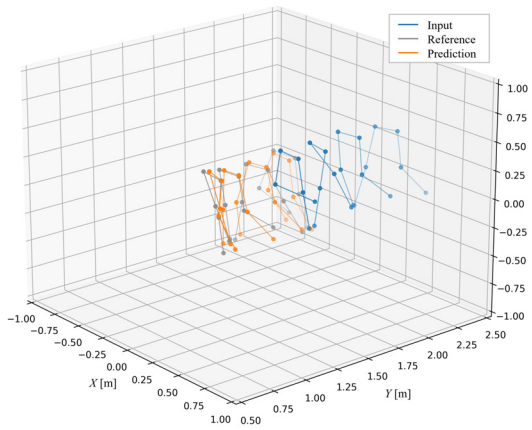


Fig. 7: Prediction example: Model input, prediction results, and reference

4090 GPU. The average processing time per frame was 1.2 ms for observation, 29 - 33 ms for prediction, and 0.1 ms for merging, allowing the calculation of the target point at a frame rate sufficient for robot following.

2) Results and Analysis: To evaluate the prediction performance, we conducted predictions on test data and analyzed the results. Here, the estimated skeleton positions when the joints are not occluded are used as the reference and compared with the predicted values. Fig. 7 shows an example of the prediction results. In the figure, orange represents input frames, gray represents reference, and purple represents the model's predictions.

The midpoint between both knees and both hips was analyzed as the tracking point. As shown in Fig. 8, the average error in the X-coordinate prediction for this tracking point was found to be 0.0125 m, with a maximum error of 0.0224 m. Furthermore, Fig. 9 demonstrates that the average error in the Y-coordinate for the same tracking point was measured at 0.0167 m, with a maximum error of 0.0330 m. These results indicate that temporal consistency was maintained through our prediction method.

However, this prediction model is unable to accommodate sudden changes in direction. The prediction errors that occur when the worker changes direction are illustrated in Fig. 10. As shown, it is evident that the model fails to adapt to direction changes, resulting in significant distance errors.

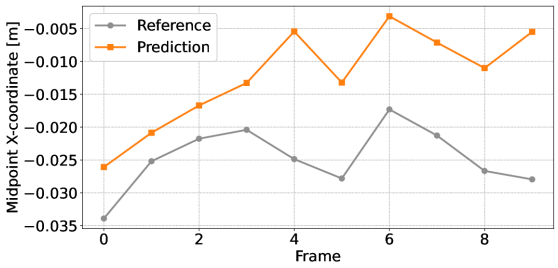


Fig. 8: Prediction result in X-coordinate of the midpoint of both knees and hips

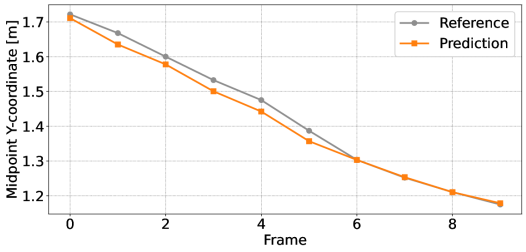


Fig. 9: Prediction result in Y-coordinate of the midpoint of both knees and hips

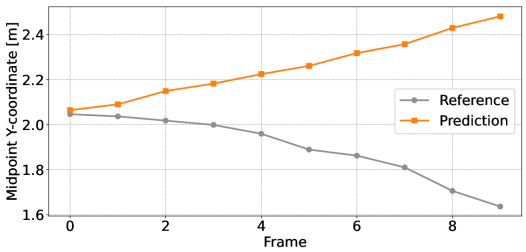
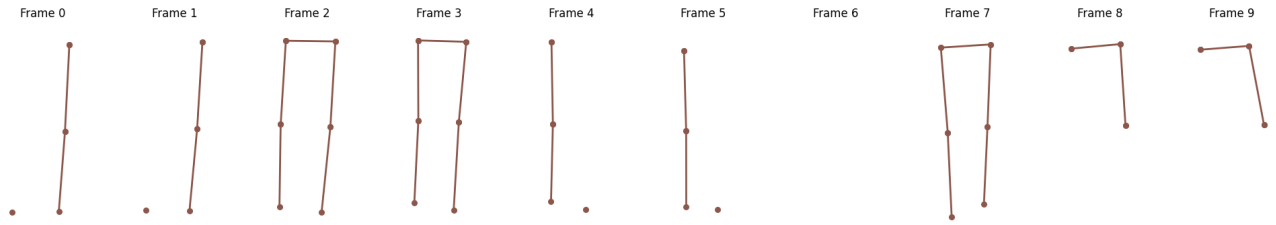


Fig. 10: Prediction result in Y-coordinate of the midpoint of both knees and hips when the worker changes direction.

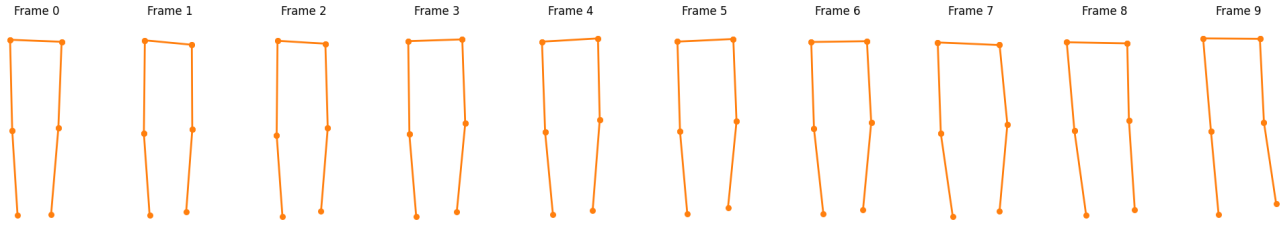
C. Verification of Proposed Result with Occlusion

1) Experimental Setup for Direction Changes and Occlusions: To address the limitations of the prediction model, particularly its inability to handle sudden direction changes as observed in Section IV-B.2, and to evaluate our method's performance in occlusion scenarios, we devised an integrated approach. This approach combines recognition results with prediction results to calculate the tracking point, specifically targeting scenarios with direction changes and occlusions. We conducted simulations to estimate joint positions for workers making sudden directional changes and when obscured by leaves and crops, mirroring the challenging conditions identified in our initial prediction analysis. Here, the estimated skeleton positions when the joints are not occluded are used as the reference and compared with the merged results. To simulate scenarios where the worker's lower limbs are partially obscured by crops, we conducted evaluations using data with partially missing observations.

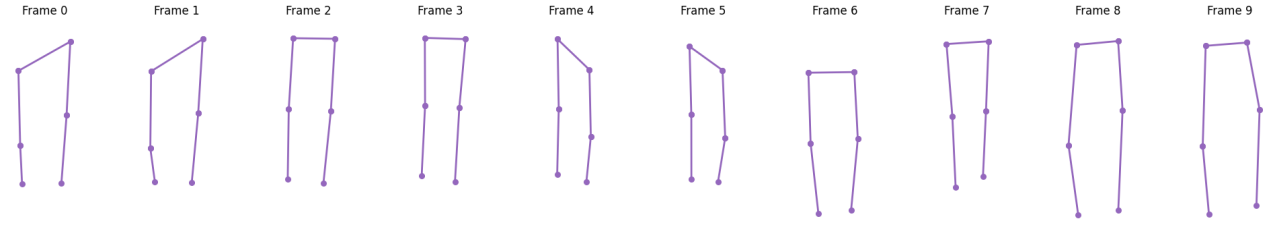
2) Results of Merged Prediction and Observation: Fig.11a through Fig.11d depict the skeletal graphs from model input to prediction and integration. Out of 10 total



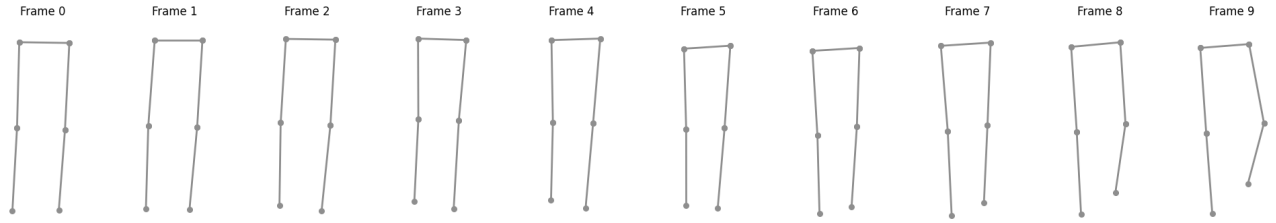
(a) Observation results: Frames 0,1,4,5,6,8,9 were simulated partial missing observations to account for occlusion



(b) Prediction results



(c) Merged results



(d) Reference

Fig. 11: Comparison of prediction, occlusion handling, merged results, and reference for lower limb.

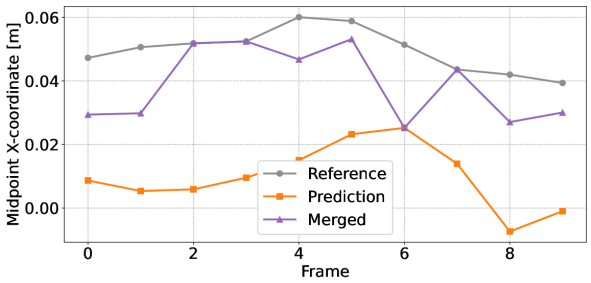


Fig. 12: Comparison of merged results, reference, and prediction results in X-coordinate

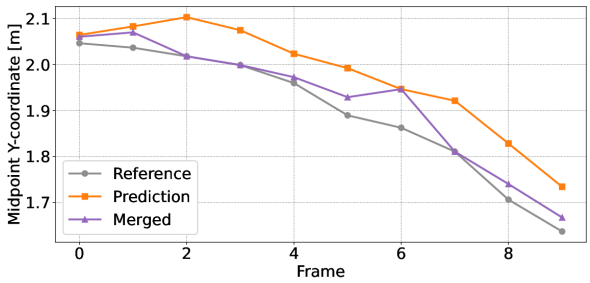


Fig. 13: Comparison of merged results, reference, and prediction results in Y-coordinate

frames, 7 frames exhibit joint loss or recognition failure. To simulate realistic occlusion scenarios commonly encountered in agricultural settings, we systematically introduced occlusions in different frames. In frames 0 and 1, the right side of the body was occluded, resulting in the loss of detection for the right hip and right knee. As the worker moved, these occlusions were temporarily resolved. Subsequently, in frames 4 and 5, the left side became occluded, causing the left hip and left knee to be undetectable. Frame 6 presented a complete occlusion of the lower body, followed by clear visibility in frame 7. Finally, in frames 8 and 9, the lower portion of the body was occluded, simulating a scenario

where the right knee, right ankle, and left ankle became undetectable due to dense foliage near the ground. As shown in Fig. 11c, even in cases where observation values could not be obtained, it can be observed that a complete lower limb is obtained through the merging process using the prediction results.

3) Comparison of Merged and Prediction Results: For this example, Fig. 12 and Fig.13 illustrate the comparison of merged results, reference, and prediction results for the tracking point (the midpoint between both knees and both hips) in X-coordinate and Y-coordinate, respectively. Both figures demonstrate that the merged results show smaller errors compared to the prediction results alone, indicating an improvement in accuracy through our merge approach. In terms of specific values, from Fig. 12, it can be seen that the maximum error in the X-coordinate is 0.0248 m for the merged results. Additionally, as shown in Fig. 13, the maximum error in the Y-coordinate for the same tracking point is 0.0840 m for the merged results. These results confirm that the integrated values of predictions and observations fall within an acceptable error range, even in challenging scenarios with direction changes and occlusions.

V. Discussion

The proposed method, utilizing a Space-Time-Separable Graph Convolutional Network (STS-GCN) for lower limb motion prediction and merging, has demonstrated promising results for human-following tasks in agricultural support robots.

The performance of the prediction model warrants careful consideration. For the tracking point (midpoint between both knees and hips), an average error of 0.0125 m and a maximum error of 0.0224 m were observed in X-coordinate prediction. Similarly, for distance prediction, an average error of 0.0167 m and a maximum error of 0.0330 m were recorded. These results fall well within the established acceptable error range (0.1 m), indicating practical following performance. Notably, temporal consistency was maintained, which is expected to contribute to stable following behavior.

However, limitations in the model's adaptability to sudden directional changes were identified. This issue can be attributed to the limited diversity of motion patterns in the training dataset. Future improvements could be achieved by training the model on a dataset encompassing a wider variety of motion patterns.

The primary strength of the proposed method lies in its novel approach to integrating prediction results with observational data. As demonstrated by the simulation results, complete lower limb skeletons were successfully reconstructed through the merging process, despite joint loss or recognition failure in 7 out of 10 frames. This suggests effective following capabilities even in scenarios where the worker is obscured by leaves or crops.

When the integrated results were compared with reference data, maximum errors of 0.0248 m in X-coordinate and 0.0840 m in Y-coordinate were observed for the tracking point. These values fall within the predetermined acceptable error range, thus validating the effectiveness of the proposed method. The particularly low error in X-coordinate prediction is indicative of the robot's ability to follow without deviating from furrows, enhancing its applicability in real agricultural settings.

It should be acknowledged that the experiments were conducted in a relatively controlled environment, which represents a limitation of this study. Actual agricultural fields present more complex factors, such as uneven terrain and varying weather conditions. Future research should involve experiments that account for these factors to further verify the robustness of the proposed method.

Additionally, evaluation from the perspective of computational cost and real-time performance remains a future challenge. Given that agricultural robots must operate with limited computational resources, model optimization and inference acceleration are crucial areas for further investigation.

VI. Conclusion

This study proposed a novel method for human-following tasks in agricultural support robots, utilizing a Space-Time-Separable Graph Convolutional Network (STS-GCN) for lower limb motion prediction and an integration technique for combining predictions with observational data. The method demonstrated high accuracy in predicting motion and effective tracking performance, even in scenarios with occlusions caused by leaves or crops. The results highlight the potential for improving human-following capabilities in narrow furrows, making the method applicable to real agricultural settings.

Future research should focus on enhancing responsiveness to sudden movements, validating the method in real agricultural environments, optimizing the model for real-time operation, and addressing more complex factors such as weather and terrain variations.

References

[1] Ministry of Agriculture, Forestry and Fisheries, "Basic Agricultural Data Set" (in Japanese)

[2] N. Mandischer *et al.*, Radar tracker for human legs based on geometric and intensity features, 2021.

[3] H. Kim *et al.*, Detection and tracking of human legs for a mobile service robot, 2010.

[4] L. Arai, R. Yorozu, A. Ohya, T. Tsubouchi, "Worker Recognition for Agricultural Support Robot Following in Narrow Furrows - Recognition of Human Center of Gravity Using RGB-D Camera and PoseNet -," Proceedings of the 2021 JSME Conference on Robotics and Mechatronics, June 2021. (in Japanese)

[5] G. Papandreou, T. Zhu, L. Chen, S. Gidaris, J. Tompson, K. Murphy, "PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model," 2018.

[6] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, "The Graph Neural Network Model," IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 61-80, Jan. 2009.

[7] S. Yan, Y. Xiong, D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," 2018.

[8] Y. Yang, Z. Ren, H. Li, C. Zhou, X. Wang, G. Hua, "Learning Dynamics via Graph Neural Networks for Human Pose Estimation and Tracking," 2021.

[9] T. Sofianos, A. Sampieri, L. Franco, F. Galasso, "Space-Time-Separable Graph Convolutional Network for Pose Forecasting," 2021.