Depth-Temporal Attention with Dual Modality Data for Walking Intention Prediction in Close-Proximity Front-Following

Chongyu Zhao[1], Guo Lingyu[2], Rongwei Wen, Yanrui Wang, Chuan Wu[†]

Abstract— **The role of robot following is crucial for effective human-robot collaboration. Traditional methods often rely on maintaining a significant distance between the robot and the human, which limits interaction and responsiveness. In contrast, close-proximity front-following facilitates immediate engagement, enhancing user experience and improving human-robot interaction. Nonetheless, it presents challenges in accurately interpreting human walking intentions due to a restricted observational field. In our paper, we introduce an innovative Depth-Temporal Attention Network that takes lower-limb depth images and robot motor signals as input, to accurately predict human walking intentions. This network leverages a depth attention module to capture essential spatial features and integrates a temporal attention mechanism to analyze movement dynamics. To enhance generalization, we use a domain adversarial module that focuses on shared features across diverse walking data, ensuring consistent performance across users. Experimental results demonstrate that our approach achieves an impressive average intention prediction accuracy of 91.09%, significantly surpassing baseline models by 12.59% to 23.66%. Additionally, an ablation study reveals that the depth-attention module substantially improves the model's understanding of depth features, resulting in an 11.44% increase in accuracy. With this high prediction accuracy, smooth front-following is achieved at close-proximity.**

I. INTRODUCTION

Autonomous robot following is a crucial aspect of human-robot interaction in various applications [1], e.g., when robots assist human managers in logistics warehouses, and airport robots carrying luggage for passengers. To achieve human-following, a direct way is to employ distance sensors to detect the user's location and adjust the robot's trajectory accordingly [2][3][4]. Some robots utilize deep neural networks and cameras to recognize the user's body, enabling effective following [5][6][7][8].

However, in certain human-robot interaction scenarios, such as when smart robotic walkers provide walking support, it is essential for the robot to remain in close-proximity to the user, specifically positioned in front [9][10]. The approaches employed by distant-following robots face significant limitations in this context due to their restricted detection range. A common solution to this challenge requires human operators to control the robot using joysticks or force sensor arrays [11][12][13]. However, this not only necessitates additional human involvement but also compels users to keep their hands in a fixed position for extended periods, which can be quite inconvenient. For instance, if a user needs to answer a phone, hold an umbrella, or has an injured hand, it would be far more convenient if robots could intelligently respond while walking to significantly enhance overall efficiency.

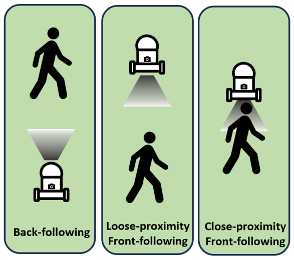

Back-following
Loose-proximity Front-following
Close-proximity Front-following

Fig. 1: Different robot following scenarios. **Back-following**: the robot follows the human at the back in a distance of several meters; **Loose-proximity Front-following**: the robot follows the human at the front in a distance of several meters; **Close-proximity Front-following**: the robot front-follows the human in a distance less than one meter.

Observing, analyzing, and accurately predicting user walking gaits for the automatic navigation of a front-following robot pose significant challenges due to the diverse range of walking patterns. Analyzing foot and leg imaging [14] can lead to difficulties with body overlap and wide-camera distortion, resulting in errors in estimating foot gestures. Convolutional Neural Network (CNN)-based methods provide promising solutions, with various frameworks proposed for predicting human movement intentions [15][16].

Recently, attention-based models have outperformed traditional CNNs in image processing [17][18]. While conventional images capture 2D spatial features, depth images provide valuable 3D spatial knowledge. However, few studies have explored the use of depth images with attention-based models for accurately predicting walking intentions. The challenge lies in enabling these models to effectively interpret depth information for optimal performance. Additionally, the effects of imaging changes caused by robot movement must be considered to enhance the model's understanding of human-robot interaction.

This study presents a front-following system designed for close-proximity scenarios, incorporating intelligent prediction of the user's walking intention. The system utilizes dual modalities of depth image data from a Time-of-Flight (ToF) camera and robot motor signals. We propose a **Depth-Temporal Attention Network** to learn the attention on the depth images and the temporal connections within an observation sequence. This enables a deep understanding of human walking intention and the corresponding changes in robot movement for seamless front-following in a hands-free control scenario, thereby enhancing the overall user

[†] Corresponding Author

This work was supported in part by Hong Kong RGC grants C7004-22G, TRS T43-513/23-N and C5032-23G.

experience. Our contributions are summarized as follows:

▷ The proposed model achieves an average accuracy of 91.09% across eight prediction categories, with accuracy in five categories exceeding 95%. The highest accuracy recorded is 99.38%, while the lowest exceeds 74%. In contrast, baseline models achieve average accuracies of only 78.50% and 67.43%, highlighting significant improvements with our method.

▷ We introduce an innovative depth attention module that enhances the model's understanding of the dynamic distance between the user and the robot during walking, resulting in an 11.44% increase in average accuracy.

▷ We fuse dual modalities of data, including depth images from a ToF camera and motor signals, to predict human walking intentions. This fusion enables the robot to significantly outperform approaches relying solely on imaging data, achieving a 25.14% increase in accuracy.

▷ Based on intention prediction, we design a control system that utilizes model predictions and human location data. Field evaluation indicates that the robot consistently follows the user from the front and aligns well in close-proximity conditions.

II. Related Work

A. Human-Following Robot

Human tracking and following are crucial in robot control, enabling robots to accurately locate and track humans. Conventional methods often use cameras for visual tracking [19][20]. By utilizing computer vision techniques and deep learning algorithms, cameras enable precise human detection and tracking with image and depth data [5][21][22]. While these methods enhance accuracy, they often require maintaining a significant distance from the target user to ensure a clear view. Additionally, camera-based models typically need a wide observation field, and images can be adversely affected by the robot's movement since the camera is mounted on the robot. Thus, the model needs to consider both image data and the robot's movement to improve tracking performance.

B. Spatial-Temporal-based Human Activity Recognition

Human walking is a sequential movement characterized by changing foot orientation and trajectory, allowing intention prediction through current and previous gestures via spatial-temporal analysis. A fundamental approach combines spatial networks, such as Convolutional Neural Networks (CNNs), with recurrent networks like Long Short-Term Memory (LSTM) networks [23]. A straightforward method connects CNNs and LSTMs by applying CNNs to each frame and feeding the resulting embeddings into the LSTM [24][25]. Also, by integrating different levels of CNN embeddings, the LSTM can gain varying insights from the CNN outputs [26]. Alternatively, skeleton data can serve as spatial features when combined with LSTM networks [27][28].

However, LSTMs tend to convey past information through hidden units, which limits their ability to weigh the importance of each data frame in the sequence. Additionally, recurrent computation can reduce the processing speed. Transformer-based models address this limitation through the attention mechanism [29], enabling the model to focus on crucial data within the input sequence while facilitating parallel computation. One method to implement spatial-temporal attention is to use embeddings from images or human skeleton features as input tokens for the transformer encoder [30]. Or, by transposing the input matrix of the transformer encoder, the attention layer can effectively learn spatial dimension features [31].

Despite these advancements, research on using depth image sequences for spatial-temporal attention is limited. The restricted observational range also complicates human skeleton graph calculation in close front-following scenarios. To tackle this, we propose a depth masking procedure to generate a sequence of depth sub-images, enabling the attention layer to learn features in 3D depth space rather than just 2D. Additionally, the motor signals are analyzed alongside the depth images in a temporal attention module to effectively leverage depth-temporal attention for walking intention prediction.

III. Dual-Modality Data

A. Lower-Limb Depth Image

To predict human walking intentions based on walking patterns, a vision modality is essential for observing human body posture and movement. While 2-dimensional spatial information is useful, incorporating depth data significantly enhances the robot's ability to gauge the distance between itself and the human user. For this purpose, we utilize a Time-of-Flight (ToF) camera (Tau Lidar Camera, *2020 Onion Corporation*) that emits safe, invisible lasers and measures the travel time of light to generate depth images. By positioning the camera horizontally at a low angle, the captured images primarily focus on the lower limbs and foot features, minimizing interference from other body parts and clothing. This setup directs the model's attention to the feet, ensures user privacy, and facilitates non-intrusive human-robot interaction.

B. Robot Motor Signals

Since the camera is mounted on the robot, the projection of the human user in the camera view is influenced not only by the user's movements but also by the robot's movements due to their close-proximity. The motor signal modality compensates for this effect by providing critical movement information to the model. Specifically, we collect data on the motor's moving distance and speed.

By integrating these two data modalities, we propose our **Depth-Temporal Attention Network**, which enables the model to understand human walking gestures while accounting for the robot's previous movements. This dual awareness enhances the prediction of human walking intentions relative to the robot's trajectory.

IV. Depth-Temporal Attention Network

The proposed **Depth-Temporal Attention Network** (DTA), illustrated in Fig. 2, consists of two main modules:

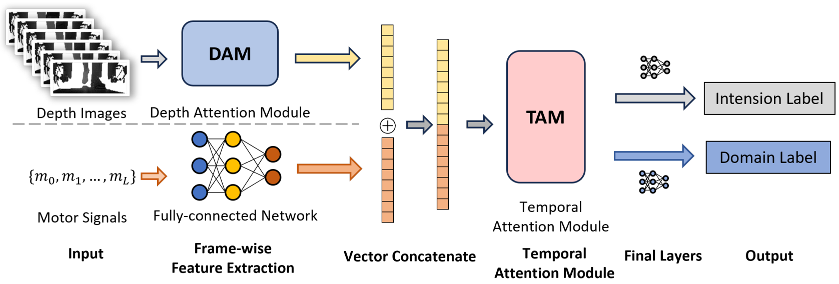


Fig. 2: The **Depth-Temporal Attention Network** predicts walking intentions using lower-limb depth images and robot motor signals. It comprises two key modules: the **Depth Attention Module** (DAM) and the **Temporal Attention Module** (TAM). The DAM masks depth images into sub-images for analysis via a multi-head attention block to generate depth embeddings. The TAM integrates these embeddings with motor signal embeddings to produce a final embedding. This embedding yields the predicted intention label for robot control and a domain label for distinguishing data domains.

the **Depth Attention Module** (DAM) and the **Temporal Attention Module** (TAM). At each time step, a sequence of depth images x_i and motor signals m_i are fed into the model, where $i = 0, 1, \ldots, L$ and L represents the sequence length. Each depth image x_i is processed by the DAM to generate depth embeddings, while the motor signals are embedded through several fully connected layers. These two modality embeddings are then concatenated and fed into the TAM to produce the final embedding. Based on this final embedding, we predict the human walking intention across seven categories: "*Static*", "*Moving Forward*", "*Moving Backward*", "*Turning Left while Moving Forward*", "*Turning Right while Moving Forward*", "*Turning Left on the Spot*", and "*Turning Right on the Spot*". This diverse range of categories enables smooth and accurate movement control of the robot. Additionally, an extra output category, "*No One*", is included to account for instances when no human is present. Ultimately, the model outputs one of these eight category predictions as the final result.

A. Depth Attention Module (DAM)

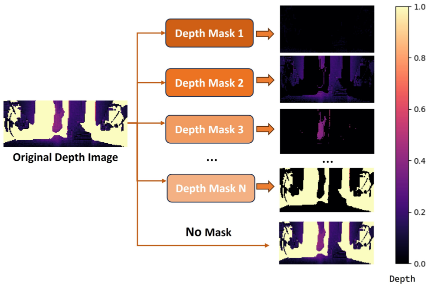


Fig. 3: Depth masking procedure. Each depth image is divided into sub-images based on preset depth thresholds, categorizing distance zones from close to far. The original depth image is included to provide global context.

The input depth image captures spatial features relevant to human walking intentions. Unlike RGB or thermal images, depth images provide essential 3D information, conveying the distance between the human and the robot in our horizontally oriented camera setup. This capability is crucial for accurately assessing proximity and enabling effective human-robot interaction.

To embed depth information, we use a Convolutional Neural Network (CNN) to process the depth image. However, CNNs primarily capture 2D relationships, so we separate each depth image into a sequence of depth sub-images, preserving the original depth order. We apply a depth-masking procedure with predefined thresholds based on the distance from the human to the robot, categorized as follows: `close` $[0m, 0.1m]$, `normal` $[0.1m, 0.5m]$, `far` $[0.5m, 0.8m]$, `out-of-zone` $[0.8m, 0.9m]$, and `background` $[0.9m, +\infty]$, as shown in Fig. 3. The original depth image is also appended to enhance the depth-attention model's understanding of the entire range, resulting in a masked image sequence for the i-th depth image x_i: $\{x_i^j | j = 1, \ldots, N+1\}$, where N is the number of masked zones which is set as 5 in our design.

The masked sequence is fed into a transformer encoder with a multi-head self-attention layer and a feedforward layer, shown in Fig. 4c. As the user's position changes during walking, the self-attention mechanism assesses the importance of different depth zones, capturing long-range dependencies and accommodating varying user distances for better generalization. Additionally, a CNN is applied to each image in the masked sequence before the depth-attention block for improved performance [32]. Inspired by ViT[17], we add an embedding token x_i^{emb} at the sequence's start, enabling the self-attention mechanism to independently gather relevant information from the masked sequence. A learnable positional encoding layer is also included. The output embedding corresponding to x_i^{emb} is selected as the final embedding vector y_i^x for the i-th depth image (see Fig. 4a for details).

Each depth image is processed by the Depth Attention Module (DAM) to generate depth embedding feature vectors y_i^x. For the motor signal vectors m_i, fully connected layers embed them into motor embedding feature vectors y_i^m, ensuring compatibility with y_i^x for effective processing by our Temporal Attention Module (TAM).

B. Temporal Attention Embedding (TAM)

The straightforward approach to combine these two modality embeddings would be to sum them; however, this method does not fully leverage the attention mechanism to weigh different components in the embeddings for the final prediction. Instead, we concatenate the two embeddings and add the concatenated features with learnable positional encodings. Additionally, a classification token is added at the beginning of the sequence.

This concatenated feature vector is then input into a multi-head attention block, which includes both a multi-head attention layer and a feed-forward layer to generate the final embeddings. We select the embedding vector z corresponding to the classification token as the final embedding for the entire sequence of depth images and motor signals. The structure of this process is illustrated in Fig. 4b.

Finally, the calculated attention associated with the classification token is processed by a classifier to output the final human walking intention prediction label. A cross-entropy loss $\mathcal{L}_{cls}$ is applied during training for walking intention classification.

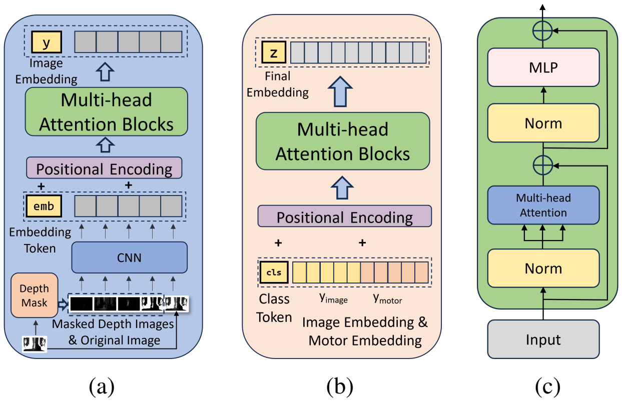


Fig. 4: The structure of (a) DAM and (b) TAM. Both are equipped with additional tokens for final output and apply the positional encoding and (c) multi-head attention block[17].

C. Domain Invariant Training

To enhance the generalization ability of the model, we have incorporated a domain invariant module that generates domain classifications. In our context, different domains can include variations such as different users, clothing, footwear, and walking speeds. A cross-entropy loss $\mathcal{L}_{\text{domain}}$, scaled by a negative value λ, serves as a discriminator, encouraging the model to avoid distinguishing between different domains. This approach prevents the model from achieving high prediction accuracy solely by recognizing domain identities, instead guiding it to learn the common knowledge shared across various walking data domains. Empirically, we set λ to -0.1, as we found that smaller values did not result in significant changes. Finally, the learning loss is the sum of the walking intention classification loss and the domain invariant loss:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\textbf{cls}} + \lambda\mathcal{L}_{\text{domain}} \tag{1}$$

D. Front-following Control

After the network predicts human walking intention labels, a control agent generates different control signals based on these predictions. We utilize a 2D LiDAR sensor to determine the human's position using the K-means algorithm. If the user walks forward, the agent generates a forward speed; if the user turns while moving forward, the agent sets a forward speed with a turning angle. For users turning in place, the robot is controlled to perform a corresponding spot turn. If the user moves backward, the robot will also move backward. To enhance safety, we establish an operation zone. If the user exits this zone, the robot will pause its following behavior. Additionally, if the robot is too close to the human, the control agent will halt the front-following process, ensuring that the human remains centered within the operation zone.

V. Experiments

A. Experimental Settings

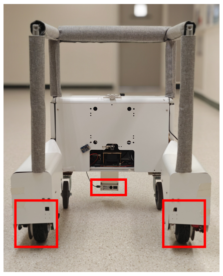
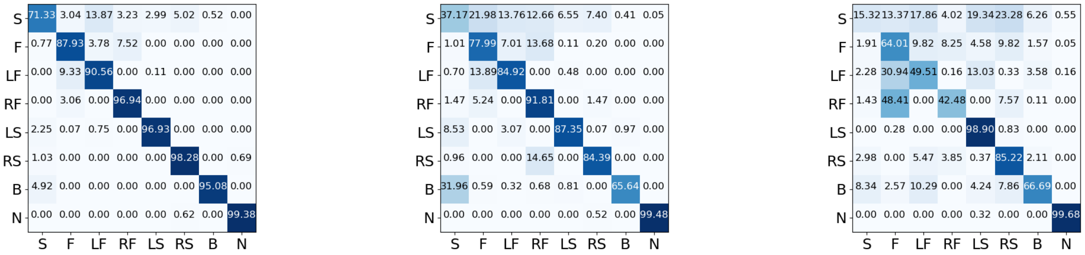


Fig. 5: Our experimental platform is an intelligent autonomous walker. A ToF camera is applied to capture the lower-limb human body. While the two motor wheels record the robot's movement signals.

Platform. We conducted our experiments using an autonomous walker, depicted in Fig. 5. The walker features a Time-of-Flight (ToF) camera mounted along the center line, approximately 5 cm above the ground, directed horizontally backward to capture lower-limb gestures. It is powered by two motors driving the rear wheels. Additionally, a 2D LiDAR, positioned 26 cm above the ground along the same center line, scans leg locations for control purposes.

Dataset. We conducted gait data collection from 25 volunteer participants with diverse body weights and heights and recorded motor signals and walking trajectories for labeling. Each participant performed a 2-minute free walking trial while manually pushing the walker. The sequence of intentions was decided randomly by each participant. Real-time monitoring of gait category distributions enabled dynamic feedback to participants for a balanced label collection. Users walked at three speed levels: fast (>3 m/s), medium (2-3 m/s), and slow (<2 m/s), resulting in at least 75 data domains. Data from 20 users were used for training, while the remaining 5 users' data were reserved for testing. Labels for input sequences are based on motor movement 5 frames later. We also collected data without users in the operation zone, including diverse backgrounds like walls and doors, and intentionally allowed people to walk around the robot to assess the model's ability to distinguish between users and passersby people.

Model Parameters. The input sequence size is set to 15 frames, spanning approximately 0.5 seconds. The embedding

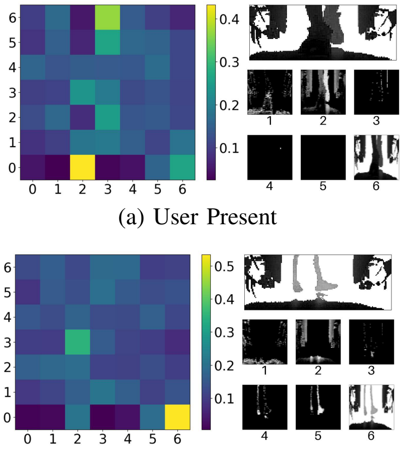


(a) DTA (b) FFLNet[24] (c) Cross Attention

Fig. 6: The confusion matrices of (a) **Depth-Temporal Attention Network**, the proposed model; (b) **FFLNet**, a CNN and LSTM based model; (c) **Cross Attention**, applying motor signal embedding as query and depth image embedding as key and value in the TAM.

CNN comprises two 2D convolutional layers followed by two fully connected layers, using ReLU as the activation function after each layer. The motor signal vector embedding network consists of two fully connected layers, each with 32 elements, producing output vectors of the same dimension.

The attention blocks in DAM and TAM share the same multi-head attention encoder structure shown in Fig. 4c [17]. Each block includes one multi-head attention layer and a feedforward layer with residual connections. We set the number of attention heads to 4 to capture different focusing lengths of the sequence. The hidden dimension of the feedforward layer is configured to 16, and each token's output dimension is set to 32. We use a single attention block, as additional blocks do not enhance accuracy.



(a) User Present

(b) No User

Fig. 7: Depth-attention maps and corresponding depth images for two situations: (a) **User Present**; (b) **No User**. We list the masked sub-images and the original image on the right-hand side: 0 for the embedding token, 1 for `close` distance, 2 for `normal` distance, 3 for `far` distance, 4 for `out-of-zone` distance, 5 for `background` distance, and 6 for the original depth image.

B. Prediction Performance

Fig. 6 presents the confusion matrices for our network performance on the test dataset. The 8 prediction categories are *Static* (S), *Moving Forward* (F), *Turning Left/Right while Moving Forward* (LF/RF), *Turning Left/Right on the Spot* (LS/RS), *Moving Backward* (B), *No User* (N). A number in row i and column j in each matrix represents the percentage of data predicted as category j with true label i in all the data belonging to category i.

The results show that the DTA achieves an average accuracy of 91.09%, with most categories exceeding 95%. This indicates effective recognition of various walking intentions. The lowest accuracy is found in the *Moving Forward* and *Static* categories. The *Static* category suffers from misclassifications due to overlapping data with other movements, making differentiation difficult. Most misclassified samples are predicted as *Turning Right while Moving Forward*, suggesting that monitoring the distance between the human and robot could help resolve this issue. Additionally, the *Moving Forward* and *Turning Left/Right while Moving Forward* categories are visually similar, leading to some misclassifications. However, the model demonstrates high accuracy in recognizing turning intentions, allowing it to correct misclassifications of moving forward.

Fig. 7 shows the depth-attention maps and corresponding depth images in different scenarios. The bottom row (Row 0) in the attention map indicates the level of attention the class token pays to different depth images. Note that in the scenario of a user in operation, depth-attention primarily focuses on the `normal` distance $[0.1m, 0.5m]$, which is the normal operating zone during robot following. Conversely, in the scenario of no user around (or standing out of the operation zone), depth attention focuses more on the `background` distance $[0.9m, +\infty]$ and the original depth image. The shift in attention focus showcases the significant contribution of depth masking in enhancing the model's comprehension of human walking patterns.

Two baseline scenarios are tested. One baseline is the FFLNet from [24] with a CNN embedding and an LSTM for temporal prediction. The other baseline is a cross-attention version of the DTA which uses motor signal embeddings as the query matrix and the depth image embeddings as the key and value matrices in TAM.

The results show that FFLNet only reaches an average accuracy of 78.50%. That is because the CNN in the FFLNet does not fully understand the distance knowledge of the depth images and the LSTM is not powerful enough to learn the temporal features of a human's walking gait. By observing the confusion matrices, we find that the FFLNet performs worse than the DTA in most of the categories. The cross-attention version baseline only achieves a 67.43% accuracy. It cannot fully utilize the attention connection between the depth embeddings and the motor embeddings. Even worse, the motor embeddings do not necessarily have a frame-wise matching with the corresponding depth embeddings, leading to worse performance.

C. Ablation Study

TABLE I: Accuracy in different ablation settings. DTA represents the full setting, while others represent removing one respective module from DTA.

Models	DTA	DTA-D	DTA-ViT	DTA-M	DTA-I	DTA-L
Avg(%)	91.09%	79.65%	85.48%	65.95%	64.57%	90.08%

We have conducted several ablation studies, including the removal of the DAM (DTA-D), the motor signals modality (DTA-M), the depth image modality (DTA-I), and the domain invariant module (DTA-L), respectively. Also, we test replacing the DAM with a ViT [17] model (DTA-ViT). The average accuracy under these settings is given in Table I.

Disabling the DAM results in an 11.44% accuracy drop, indicating that the depth masking procedure and attention-based depth analysis are crucial for fully understanding depth features. While using the ViT [17] yields better performance than using CNNs with an average accuracy of 85.48%, it still falls short of the performance achieved by the DAM design. This is mainly because the ViT model analyzes depth images as standard visual images, lacking a detailed understanding of the depth-modality information.

Without the motor signals modality (DTA-Mot), prediction accuracy falls to 65.95%, highlighting the importance of the robot's movement data in understanding human walking intentions. Additionally, relying solely on motor signals without depth images (DTA-Dep) results in an accuracy of 64.57%, demonstrating that motor signals alone are inadequate for achieving high performance.

Finally, removing the Domain Invariant Module (DTA-Dom) does not harm the model performance a lot, showing that the DTA itself is generalized enough for human walking intention prediction. We maintain the module to further boost model robustness towards new data distributions.

D. Masking Settings Experiment

TABLE II: Average accuracy of different mask settings.

Models	Default	Even	Default*
Average(%)	91.09%	89.64%	87.15%

Another critical parameter setting is the image masking configuration. We evaluate model performance in two scenarios: **Default** and **Even**. The **Default** scenario divides zones by distances (close, appropriate, far) and out-of-operation (allowing the DAM to focus on relevant areas while filtering out background noise). This approach preserves the integrity of the human body, which typically remains in the appropriate zone. Conversely, the **Even** scenario divides zones uniformly, disregarding body integrity, which can harm the accuracy of CNN embeddings. We also assess the impact of excluding the original image from the masked sequence, labeled as **Default***. Results are given in Table II.

The results show that even masking slightly degrades performance due to dense zone separation. However, the attention mechanism helps maintain accuracy by connecting different body parts across masked zones. While the even masking method increases the sequence length, resulting in greater computational overhead, losing the global context from the original depth image significantly decreases accuracy. Thus, the default settings are optimal for maintaining both high accuracy and efficiency in the masking procedure.

E. Field Experiments

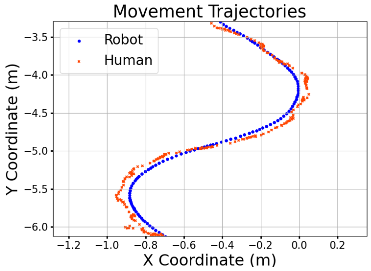


Fig. 8: The trajectory of the robot and the human user when the robot follows the human from the front.

To evaluate front-following in close-proximity, we recorded the robot's trajectory using motor signals and the human's trajectory with a 2D LiDAR mounted on the robot. As shown in Fig. 8, the robot successfully maintains an average distance of under 50 cm from the human and aligns well with user movement, facilitating close and intelligent interaction with the user.

VI. Limitation & Future Research

The DTA network effectively predicts human walking intentions in close-proximity interactions but has limitations. It identifies only eight discrete categories instead of providing continuous spatial prediction, which is better for free movement. This requires generating continuous motor control, prompting future work through reinforcement learning. Also, the DTA's masking thresholds are limited to front-following scenarios, needing expert design for other human-robot interactions. We propose making these thresholds learnable to enhance generalization while maintaining accuracy. Furthermore, incorporating human recognition and clustering of walking patterns could further improve performance among diverse users.

VII. Conclusion

This paper introduces a Depth-Temporal Attention Network that combines depth images and motor signals for accurate robot front-following. Using a masking procedure and depth attention mechanism, the model captures human walking depth features. The motor signals enhance robot movement understanding, boosting prediction accuracy in human-robot interactions. The Temporal Attention Module processes temporal features, prioritizing key information for predictions. Consequently, the model achieves an average prediction accuracy of 91.09% across eight walking intention categories, enabling safe and intelligent front-following.

References

[1] M. Gupta, S. Kumar, L. Behera, and V. K. Subramanian, "A novel vision-based tracking algorithm for a human-following mobile robot," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 7, pp. 1415–1427, 2016.

[2] D. Jin, Z. Fang, and J. Zeng, "A robust autonomous following method for mobile robots in dynamic environments," *IEEE Access*, vol. 8, pp. 150 311–150 325, 2020.

[3] Q. Xiao, F. Sun, R. Ge, K. Chen, and B. Wang, "Human tracking and following of mobile robot with a laser scanner," in *2017 2nd International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, 2017, pp. 675–680.

[4] F. Hoshino and K. Morioka, "Human following robot based on control of particle distribution with integrated range sensors," in *2011 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2011, pp. 212–217.

[5] R. Algabri and M.-T. Choi, "Deep-learning-based indoor human following of mobile robot using color feature," *Sensors*, vol. 20, no. 9, p. 2699, 2020.

[6] B.-J. Lee, J. Choi, C. Baek, and B.-T. Zhang, "Robust human following by deep bayesian trajectory prediction for home service robots," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 7189–7195.

[7] Y. Y. Aye, K. Thiha, M. M. M. Pyu, and K. Watanabe, "A deep neural network based human following robot with fuzzy control," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2019, pp. 720–725.

[8] L. Pang, Y. Zhang, S. Coleman, and H. Cao, "Efficient hybrid-supervised deep reinforcement learning for person following robot," *Journal of Intelligent & Robotic Systems*, vol. 97, pp. 299–312, 2020.

[9] X. Zhao, Z. Zhu, M. Liu, C. Zhao, Y. Zhao, J. Pan, Z. Wang, and C. Wu, "A smart robotic walker with intelligent close-proximity interaction capabilities for elderly mobility safety," *Frontiers in neurorobotics*, vol. 14, p. 575889, 2020.

[10] C. Gonçalves, J. M. Lopes, S. Moccia, D. Berardini, L. Migliorelli, and C. P. Santos, "Deep learning-based approaches for human motion decoding in smart walkers for rehabilitation," *Expert Systems with Applications*, vol. 228, p. 120288, 2023.

[11] M. Martins, C. Santos, A. Frizera, and R. Ceres, "Real time control of the asbgo walker through a physical human–robot interface," *Measurement*, vol. 48, pp. 77–86, 2014.

[12] M. Kumar, A. Vasage, G. Kulkarni, O. Padhye, S. Kerkar, M. Gupta, and K. Singh, "Calibration and optimization of fsr based smart walking assistance device," *Engineering Research Express*, vol. 5, no. 2, p. 025016, 2023.

[13] D.-M. Ding, Y.-G. Wang, W. Zhang, and Q. Chen, "Fall detection system on smart walker based on multisensor data fusion and sprt method," *IEEE Access*, vol. 10, pp. 80 932–80 948, 2022.

[14] L. Yue, L. Zongxing, D. Hui, J. Chao, L. Ziqiang, and L. Zhoujie, "How to achieve human–machine interaction by foot gesture recognition: a review," *IEEE Sensors Journal*, vol. 23, no. 15, pp. 16 515–16 528, 2023.

[15] C. Zhong, L. Hu, Z. Zhang, Y. Ye, and S. Xia, "Spatio-temporal gating-adjacency gcn for human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 6447–6456.

[16] K. M. Abughalieh and S. G. Alawneh, "Predicting pedestrian intention to cross the road," *IEEE Access*, vol. 8, pp. 72 558–72 569, 2020.

[17] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 571–38 584, 2022.

[19] H. Liu, J. Luo, P. Wu, S. Xie, and H. Li, "People detection and tracking using rgb-d cameras for mobile robots," *International Journal of Advanced Robotic Systems*, vol. 13, no. 5, p. 1729881416657746, 2016.

[20] M. Munaro and E. Menegatti, "Fast rgb-d people tracking for service robots," *Autonomous Robots*, vol. 37, pp. 227–242, 2014.

[21] B. X. Chen, R. Sahdev, and J. K. Tsotsos, "Integrating stereo vision with a cnn tracker for a person-following robot," in *Computer Vision Systems: 11th International Conference, ICVS 2017, Shenzhen, China, July 10-13, 2017, Revised Selected Papers 11*. Springer, 2017, pp. 300–313.

[22] T.-H. Tsai and C.-H. Yao, "A robust tracking algorithm for a human-following mobile robot," *IET Image Processing*, vol. 15, no. 3, pp. 786–796, 2021.

[23] S. Hochreiter, "Long short-term memory," *Neural Computation MIT-Press*, 1997.

[24] Z. Chongyu, G. Wenzhi, W. Rongwei, Z. Wang, and C. Wu, "Deep learning-driven front-following within close proximity: a hands-free control model on a smart walker," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 812–818.

[25] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks," *IEEE access*, vol. 6, pp. 17 913–17 922, 2018.

[26] Z. Zhang, Z. Lv, C. Gan, and Q. Zhu, "Human action recognition using convolutional lstm and fully-connected lstm with different attentions," *Neurocomputing*, vol. 410, pp. 304–316, 2020.

[27] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[28] A. Sharma and R. Singh, "Convst-lstm-net: convolutional spatiotemporal lstm networks for skeleton-based human action recognition," *International Journal of Multimedia Information Retrieval*, vol. 12, no. 2, p. 34, 2023.

[29] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[30] D. Ahn, S. Kim, H. Hong, and B. C. Ko, "Star-transformer: a spatio-temporal cross attention transformer for human action recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 3330–3339.

[31] Y. Sun, A. W. Dougherty, Z. Zhang, Y. K. Choi, and C. Wu, "Mixsynthformer: A transformer encoder-like structure with mixed synthetic self-attention for efficient human pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 884–14 893.

[32] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Advances in neural information processing systems*, vol. 34, pp. 30 392–30 400, 2021.