2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Exploring Transformers and Visual Transformers for Force Prediction in Human-Robot Collaborative Transportation Tasks

J. E. Domínguez-Vidal and Alberto Sanfeliu

***Abstract*— In this paper, we analyze the possibilities offered by Deep Learning State-of-the-Art architectures such as Transformers and Visual Transformers in generating a prediction of the human's force in a Human-Robot collaborative object transportation task at a middle distance. We outperform our previous predictor by achieving a success rate of 93.8% in testset and 90.9% in real experiments with 21 volunteers predicting in both cases the force that the human will exert during the next 1 s. A modification in the architecture allows us to obtain a second output from the model with a velocity prediction, which allows us to improve the capabilities of our predictor if it is used to estimate the trajectory that the human-robot pair will follow. An ablation test is also performed to verify the relative contribution to performance of each input.**

***Index Terms*— Physical Human-Robot Interaction, Object Transportation, Human-in-the-Loop, Force Prediction**

I. Introduction

The field of robotics has always made use of the latest advances in fields such as control, psychology and, more recently, artificial intelligence (AI) to provide robots with more and better capabilities enabling them first to work autonomously and then in collaboration with humans in increasingly less controlled environments.

Thus, during the last decade, different methods have been developed to improve this collaboration, seeking to better understand the human's preferences [1], [2] or their intentions. This has allowed to made significant advances in multiple tasks: predicting the next object the human will choose [3], the path to follow [4]–[6], the next area of interest in a collaborative search [7], [8] or even which action will be the next to be performed by the human [9].

In this paper we will focus on a specific task such as human-robot collaborative transport of objects. Specifically, this article is a continuation of our previous work [10], [11]. In [10] we explored this task in a classical way: developing a controller that would combine the robot's input with that of the human in a way that would be comfortable and intuitive for the latter. The experimental data obtained in that work allowed us to take a different approach from the usual one in the literature. Instead of developing a controller that tries to adapt to the human, we took advantage of the fact that this is a task where the information exchange is mainly through forces and we developed a first version of a force predictor based on a simple Deep Learning architecture (see Fig. 1) [11]. This first predictor proved to be effective in predicting the force to be exerted by the human during the next second but also presented some limitations. In this paper, we explore more recent architectures based on Transformers [12] by testing their capabilities in this task.

In the remainder, Section II presents the related work. Section III presents the architectures of the force predictors used in this article. Section IV shows the results obtained regarding the performance of each predictor in dataset and real experiments. Finally, Section V presents the conclusions.

II. Related Work

The way in the literature to solve tasks involving collaborative manipulation or transportation between a human and a robot is by using controllers. Both admittance [13] and impedance [14]–[16] control systems have been used for decades, using both one level of control [17] or several [18] and generating their reference in multiple ways [14], [19]. Most of these controllers are oriented to make the robot adapt to the actual human's actions. However, it is also possible to find works which include a prediction of future human behavior such as the trajectory that the human desires for themselves [20] or for the transported object [21].

The development of different Deep Learning models and methods has improved these predictions. Thus, [22] uses Reinforcement Learning (RL) to adapt the damping coefficient in an admittance controller and in [23], [24] they use Learning from Demonstration (LfD) to learn a model of the task or to predict the speed profile that the human would like to follow to complete the task. While these works use Deep Learning models as input to a controller, it is also possible to find systems that use the output of a neural network directly as in [25], [26] where Radial Basis Function Neural Networks are used to estimate the human's trajectory.

Although [24], [25] use the force exerted by the human as an input to their model, to the best of our knowledge only our previous work [11] seeks to predict the force to be exerted by the human rather than the trajectory the human wishes to follow or where they would like to carry the object. This is because with this force prediction we can both detect rapid changes in the human's intention and obtain an estimate of the human's desired trajectory as demonstrated in [11].

If in our previous work we were inspired by [27]–[29] to combine Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) units to process both visual and sequential information, in this work we will rely on

Work supported under the European project CANOPIES (H2020-ICT-2020-2-101016906) and by JST Moonshot R & D Grant Number: JPMJMS2011-85. The first author acknowledges Spanish FPU grant with ref. FPU19/06582.

The authors are with the Institut de Robòtica i Informàtica Industrial (CSIC-UPC). Llorens i Artigas 4-6, 08028 Barcelona, Spain and with Universitat Politècnica de Catalunya - BarcelonaTech (UPC). Jordi Girona, 31, 08034, Barcelona, Spain. {jdominguez, sanfeliu}@iri.upc.edu. The first one is the corresponding author.

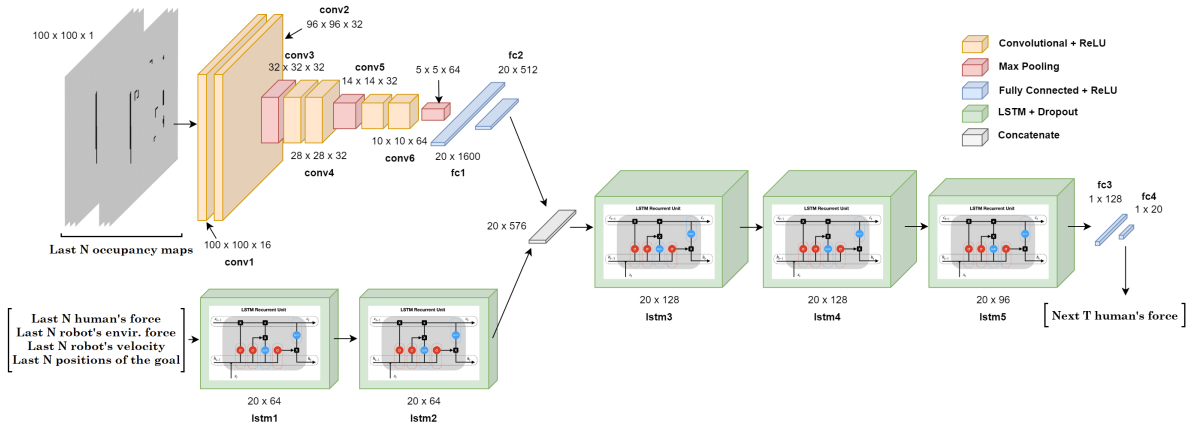

Fig. 1. **Model architecture of original force predictor in [11].** Two streams in parallel. Top one to process occupancy map obtained from LiDAR and bottom one to process other inputs (previous human's force, environment force, robot's velocity and task's goal position). Both streams concatenated to finish the processing obtaining prediction of next human's force.

architectures based on Transformers [12], thought originally to process sequential information. Two of its evolutions, the Vision Transformer (ViT) [30] and the Swin Transformer [31], are the ones that allow to use the concept of attention in images by processing them as sequences of patches. Subsequently, their respective evolutions such as the Video Vision Transformer (ViViT) [32] and the Video Swin Transformer [33] allow to apply the same logic to sequences of images. Thus, in this work we rely on the Transformer and ViViT architectures to enhance our force predictor.

III. Transformer-based Force Predictors for Collaborative Object Transportation

To make the results comparable with [11], we will use as a use case the same collaborative transport of a light object between the human and the robot and the same scenario in which several walls and columns are placed as obstacles.

A. Problem definition

The problem is equivalent to the one presented in [11], since we will use the same five sources of information. The main difference is that this improved version of our force predictor will have two output streams instead of only one. Thus, the predictor will generate an estimate of both the force that will be exerted by the human and the velocity that the human-robot pair will present. As it will be shown Section IV, this velocity prediction will improve the capabilities of our predictor when used as an estimator of the trajectory to be followed by both agents.

Thus, the first source of information used will be an occupancy map obtained from the robot's LiDAR/LaserScan readings. The size of this map is 100x100 pixels representing each one if the equivalent area of 10x10 cm in the real environment is occupied or not. The second source of information used also represents the environment in which the pair is working. Based on the Social Force Model (SFM) [34], the environment is represented as a set of attractive and repulsive forces. Specifically, each of the O obstacles detected in the environment generates a repulsive force $\boldsymbol{f_{C,obs}} \in \mathbb{R}^2$. At the same time, the robot has a global planner that generates a series of waypoints to the pre-established position to which the pair will take the object. Each of these generates an attractive force $\boldsymbol{f_{C,goal}} \in \mathbb{R}^2$ and the weighted sum of these two types of forces gives rise to the force $\boldsymbol{f_{E,C}} \in \mathbb{R}^2$ which is the second input to our model. More details about how to calculate each of these forces in [35], [36].

Third, the force exerted by the human and measured by the robot using a force sensor on its wrist is considered, $\boldsymbol{F_{H,C}} \in \mathbb{R}^2$. The weighted sum of $\boldsymbol{f_{E,C}}$ and $\boldsymbol{F_{H,C}}$, $\boldsymbol{F_{Task,C}}$, is what the robot uses to generate its own linear and angular velocity commands being these velocities the fourth input to our model. Finally, the fifth source of information is the distance to the goal in modulus and angle[1]. These last four inputs are normalized to the range $[-1, 1]$ considering a maximum modulus for each force of 12 N, maximum velocities of 0.65 m/s and 1 rad/s respectively and a maximum distance to the goal of 7 m. Same criteria used in [11].

To predict the next T forces exerted by the human, $Y^{force}_{N+1:N+T} \in \mathbb{R}^{2,T}$, we will use the last N occupancy maps, $X^{map}_{1:N}$, and the last N concatenations of the other four information sources, $X^{f}_{1:N} = [x^f_1, x^f_2, ..., x^f_N]$ with $x^f_i \in \mathbb{R}^8$. These same inputs will be used to obtain the T following velocities shown by the human-robot pair, $Y^{vel}_{N+1:N+T} \in \mathbb{R}^{2,T}$. The system's working frequency is 10 Hz so we will use the information from the last 2 s ($N = 20$) to predict the next second ($T = 10$).

B. Force/Velocity Predictor Models

Fig. 1 shows an scheme of the architecture of our original force predictor. Based on [28], [29], two parallel streams are used processing $X^{map}_{1:N}$ and $X^{f}_{1:N}$ respectively. The workhourses used in that case were CNNs and LSTMs, as well as Fully Connected (FC) layers to reduce dimensionality and obtain the $Y^{force}_{N+1:N+T}$ vector. This original model will be called the *CNN+LSTM* version.

[1]Example of how to calculate $\boldsymbol{F_{Task,C}}$ and the performed experiments: `https://youtu.be/Aub8WPKHJi0`

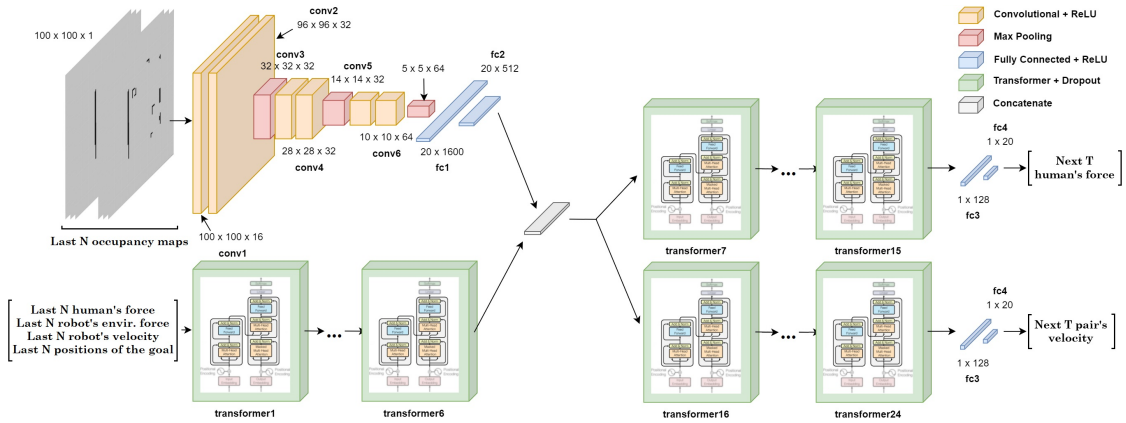

Fig. 2. **Model architecture for CNN+Transformer force predictor.** Two input and two output streams in parallel. CNNs to process occupancy map obtained from LiDAR and Transformers to process other inputs. Both streams concatenated and processed by two parallel Transformers streams to obtain a prediction of next human's force and next human-robot pair's velocity.

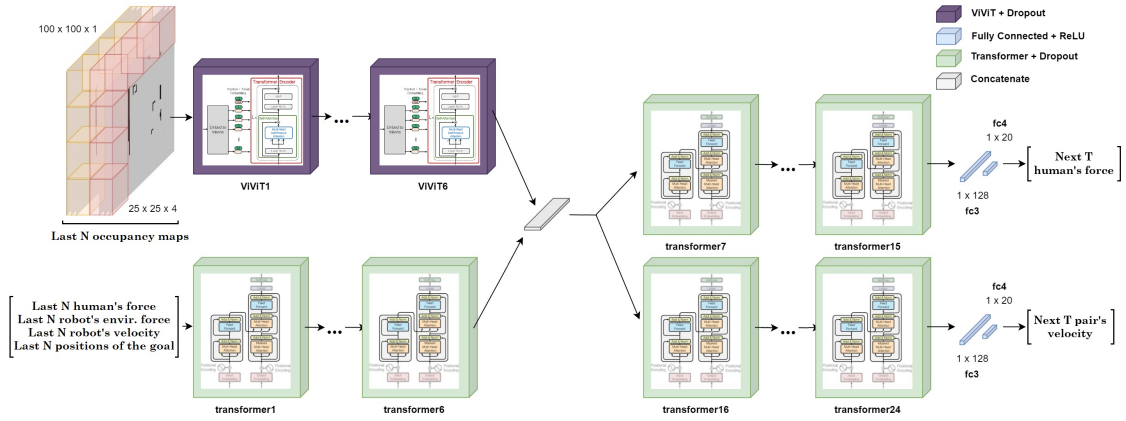

Fig. 3. **Model architecture for ViViT+Transformer force predictor.** Two input and two output streams in parallel. ViViT to process occupancy map obtained from LiDAR and Transformers to process other inputs. Both streams concatenated and processed to obtain a prediction of next human's force and next human-robot pair's velocity.

Fig. 2 shows the first modifications made on this model. All LSTM units are replaced by layers of Transformers. To process the input $X^f_{1:N}$ six layers of Transformers are used each with $h = 8$ self-attention heads, 64 as the dimensionality of the linearly projected queries, keys and values, 512 for the dimensionality of the inner FC layer and 128 for the dimensionality of the sub-layers' outputs. Dropout is also used with a probability $p = 0.3$. The second difference with the original model consists in sending the concatenation of the two input streams not to a single output stream but to two in parallel in order to obtain both $Y^{force}_{N+1:N+T}$ and $Y^{vel}_{N+1:N+T}$ vectors. Both are made up of nine layers of Transformers with the same parameters except $h = 10$ and $p = 0.35$ now. This one will be called the *CNN+T* version. The second output stream could also be added to *CNN+LSTM*, but we keep it as it was in its original article.

Fig. 3 shows the second modificacion made. As with LSTMs, all the CNN layers are replaced by another architecture that has proven its efficiency in processing image sequences such as ViViT. In this case, six ViViT layers are used, each with $h = 8$ self-attention heads, 128 as the projection dimensionality and $p = 0.35$ as the Dropout probability. For its part, the input $X^{map}_{1:N}$ is sequenced into $L = 4$ consecutive patches of size 25x25 pixels. The rest of the structure is equivalent to that used in the CNN+T version. The selection of number of layers, patch size or attention heads was made iteratively until getting the best results. We will call this version *ViViT+T*.

The version using ViViT with LSTMs has also been trained. However, and for the sake of brevity, its scheme is omitted since it is the worst performing version as it will be shown in Section IV. We will call this version *ViViT+LSTM*.

C. Dataset Acquisition and Training

The same dataset used in [11] is extended using the samples obtained in [37] in which our first predictor is compared with other system also performing a human-robot collaborative transport in the same scenario. In this way, a dataset with 14120 sub-sequences is obtained. These sub-sequences are obtained splitting each recorded experiment in blocks of $N+T$ samples with an overlapping of $(N+T)/2$ samples between sub-sequences. The input of each model are the first N samples with which they try to predict the next T human's forces and human-robot pair's velocities. This sub-sequences are divided into the classic splits of training (90%:

TABLE I

EVOLUTION OF MEAN ERROR AND PERCENTAGE OF CORRECT PREDICTIONS IN TESTSET. VARIABLE Y REPRESENTS FORCE (F) OR VELOCITY (Vel).

Measure		Time [ms]							
		Force (Y = F)				Velocity (Y = Vel)			
		100	300	500	1000	100	300	500	1000
Error Y_x [N or m/s]	CNN+LSTM	0.208	0.241	0.248	0.280	–	–	–	–
	CNN+T	**0.188**	**0.200**	**0.204**	**0.239**	**0.0037**	**0.0049**	**0.0062**	**0.0073**
	ViViT+LSTM	0.315	0.349	0.388	0.466	0.0102	0.0110	0.0129	0.0151
	ViViT+T	0.234	0.289	0.360	0.440	0.0061	0.0074	0.0085	0.0104
Error Y_y [N or rad/s]	CNN+LSTM	0.099	0.123	0.124	0.150	–	–	–	–
	CNN+T	**0.085**	**0.094**	**0.096**	**0.121**	**0.0025**	**0.0033**	**0.0042**	**0.0049**
	ViViT+LSTM	0.167	0.182	0.200	0.239	0.0071	0.0083	0.0101	0.0125
	ViViT+T	0.114	0.151	0.193	0.221	0.0042	0.0049	0.0059	0.0072
Error $\lvert\boldsymbol{Y}\rvert < 0.1 \cdot Y_{max}$ & Error $\angle\boldsymbol{Y} < 18°$ [%]	CNN+LSTM	94.5	93.7	93.4	92.4	–	–	–	–
	CNN+T	**95.6**	**94.9**	**94.7**	**93.8**	**98.9**	**98.4**	**97.8**	**96.9**
	ViViT+LSTM	90.8	89.0	86.6	83.0	95.6	95.3	93.9	92.7
	ViViT+T	94.0	91.8	87.7	84.2	97.9	97.0	96.5	95.4

12708 sub-sequences), validation (5%: 706 sub-sequences) and testing (5%: 706 sub-sequences) datasets.

Adam with its default parameters is the optimizer used for training each model. Additionally, learning rate decay is added with a minimum $lr = 3 \times 10^{-5}$ and a decay factor of 0.96. The maximum number of epochs is 80, although using early stopping to avoid overfitting. Models were not observed to exceed epoch 65 (CNN+LSTM), 70 (CNN+T and ViViT+LSTM) or 75 (ViViT+T) respectively. An NVIDIA RTX 2080 Ti graphics card was used, training for 85-130 minutes depending on the model.

IV. RESULTS

First, we test the ability of each trained model to predict the force that the human will exert as well as the velocity profile that the human-robot pair will follow. For this we use the testset split. Having done this, we perform a new round of experiments with 21 volunteers (age: $\mu = 27.45$, $\sigma = 4.02$). These volunteers perform the same collaborative transport task in the same scenario used in [11] a total of three times: once without any predictor in order to obtain more samples with which to continue expanding our dataset, once with the original predictor, and once with the predictor that gives the best results in the testset. With these experiments we can check the real improvement obtained with different humans who have not performed the experiment before. All the experiments reported in this work have been performed after getting the approval of the ethics committee of the Universitat Politècnica de Catalunya (UPC) in accordance with all the regulations and relevant guidelines (ID: 2023.05).

To do this, we use the same ROS (Robot Operating System) node used to encapsulate and format all the inputs and outputs necessary to operate the original predictor using in each case the architecture and weights of the model of interest. Each of the predictors conditions the robot's planner in a different way. The original CNN+LSTM predictor sends its prediction of the human's force to a controller equivalent to the one used to generate the robot's speed command from the $\boldsymbol{F_{Task,C}}$ force, thus generating an estimate of the human's desired speed. In turn, the remaining predictors can directly generate a prediction of the human-robot pair's velocity. Both velocity estimates are integrated to obtain an estimation of the trajectory to follow which is finally used to condition the robot's planner. All the real experiments reported in this section, were performed using an MSI GS66 laptop with an RTX 3060 Mobile (80 W).

A. Force/Velocity Predictor Performance

As in [11], we compute the absolute error in each Cartesian axis between the prediction of the next human's force and its actual value. Likewise, we also compute the absolute error between the prediction of the linear and angular velocity and the real value of this variable for the human-robot pair. Additionally, we also calculate the percentage of samples that present an error lower than a 10% both in modulus and angle. This means an error less than 1.2 N for the force prediction and an error less than 0.065 m/s and 0.1 rad/s for the velocity. The results are shown in Table I.

First, the results provided by the CNN+LSTM model differ slightly from those shown in [11] as this model has been retrained for more epochs with a larger dataset. Among the three models designed in this paper, CNN+T wins in all the measurements performed both in predicting the human's force and the pair's velocity during the next second. It is closely followed by the ViViT+T model and ViViT-LSTM is the one with worst performance. Among them, only CNN+T manages to outperform CNN+LSTM. This implies that the use of ViViT is counterproductive contrary to what might initially appear. This result should not be so surprising since it is a known issue in the literature that ViT (and therefore also ViViT) can improve the performance offered by CNNs in image classification tasks but only if it is trained with very large datasets [30], [38]. This result seems to indicate that our dataset is not large enough. Likewise, the substitution of

TABLE II

ABLATION STUDY WITH CNN+T REMOVING EACH INPUT. VARIABLE Y REPRESENTS FORCE (F) OR VELOCITY (Vel).

Measure		Time [ms]							
		Force ($Y = F$)				Velocity ($Y = Vel$)			
		100	300	500	1000	100	300	500	1000
Error $\lvert\boldsymbol{Y}\rvert < 0.1 \cdot Y_{max}$ & Error $\angle\boldsymbol{Y} < 18°$ [%]	Without ccupancy map	86.6	84.0	81.3	76.4	94.6	92.1	89.0	85.3
	Without env. force	93.5	91.3	89.1	85.6	97.1	96.0	93.9	91.2
	Without human's force	90.5	88.8	86.0	81.3	97.0	95.8	93.7	90.8
	Without robot's velocity	93.4	91.0	88.5	84.6	96.2	95.0	92.8	89.5
	Without goal position	94.5	93.4	92.8	91.5	97.9	97.1	96.3	95.0

TABLE III

MEAN ERROR AND PERCENTAGE OF CORRECT PREDICTIONS IN REAL EXPERIMENTS. VARIABLE Y REPRESENTS FORCE (F) OR VELOCITY (Vel).

Measure		Time [ms]							
		Force ($Y = F$)				Velocity ($Y = Vel$)			
		100	300	500	1000	100	300	500	1000
Error Y_x [N or m/s]	CNN+LSTM	0.281	0.310	0.316	0.350	–	–	–	–
	CNN+T	**0.188**	**0.200**	**0.204**	**0.239**	**0.0063**	**0.0074**	**0.0085**	**0.0112**
Error Y_y [N or rad/s]	CNN+LSTM	0.151	0.163	0.170	0.180	–	–	–	–
	CNN+T	**0.085**	**0.094**	**0.096**	**0.121**	**0.0043**	**0.0050**	**0.0059**	**0.0084**
Error $\lvert\boldsymbol{Y}\rvert < 0.1 \cdot Y_{max}$ & Error $\angle\boldsymbol{Y} < 18°$ [%]	CNN+LSTM	92.3	91.4	90.8	89.3	–	–	–	–
	CNN+T	**93.5**	**92.9**	**92.3**	**90.9**	**97.6**	**96.9**	**96.4**	**95.2**

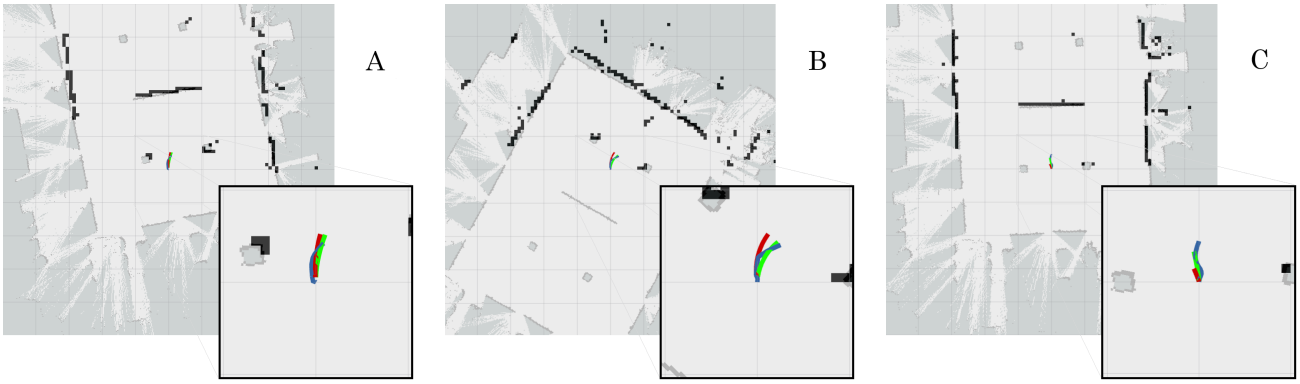

Fig. 4. **Comparison of human-robot pair's real trajectory and trajectory estimation for** $1\ s$ **in different situations.** Actual trajectory in blue. Trajectory estimation from CNN+LSTM in red and trajectory estimation from CNN+T in green. A - Normal situation. B - Human's force against normal operation, in this case, over-avoiding the obstacle in the right. C - Human exerting an extremely low force.

LSTM cells by Transformers does improve the performance due to its richer ability to encode features. Since CNN+T is the model that gives the best results, this will be the only one used in real experiments together with the original CNN+LSTM model to check the real improvement obtained.

On the other hand, it is also observed that it is easier to make predictions for velocity than for force. This is because the velocity is somewhat proportional to the integral of the force. This causes the velocity to vary more slowly. First, because there is a delay between when a substantial change in force occurs and when it makes a difference in the robot's motion. Secondly, because this integration is the equivalent of a low-pass filter that eliminates the small instantaneous variations that may occur in the force exerted by the human. At the same time, this phenomenon underlines the usefulness of our force predictor since a good prediction of the force allows to obtain a good estimation of the velocity, while the opposite case does not necessarily occur.

The same testset can be used to make an ablation study. In this case, eliminating each input, one at a time, and checking the performance drop. Table II shows the percentage of samples considered as correct when each of the inputs are missing using only the CNN+T model since it is the one with the best performance and, therefore, the one we compare with the original CNN-LSTM in the real experiments. The most relevant input is the occupancy map causing a drop of up to 17.4% in the force prediction and up to 11.6% in the velocity prediction. The next most relevant inputs are the human's force when predicting the next force they will exert with a drop of up to 12.5% and the robot's speed when predicting

TABLE IV

COMPARISON OF MEAN ERROR ESTIMATING HUMAN TRAJECTORY WITH DIFFERENT MODELS. * MARKS VALUES OBTAINED BY INTERPOLATION FROM LAPLAZA EL AL. [6].

Model	L2 [m]	
	500 ms	1000 ms
Martinez el al. [4]	0.159*	0.317*
Mao et al. [5]	0.081*	0.161*
Laplaza et al.. [6]	0.072*	0.142*
2nd order polynomial	0.123	0.277
CNN+LSTM [11]	0.093	0.199
CNN+T	**0.061**	**0.138**

the pair's speed with up to 7.4%. The environmental force, $\boldsymbol{f_{E,C}}$, affects both predictions with drops of up to 8.2% and 5.7%, respectively. The least influential variable is the position of the goal with maximum drops of 2.3% and 1.9%.

The first experiment performed by the 21 new volunteers in which no predictor is used allows us to perform the same measures as with the testset. Table III shows a reduction in both predictors of between 2.1% and 3.1% in the percentage of samples considered as correct for the prediction of the human's force and between 1.3% and 1.7% for the velocity prediction, in this case, only with the CNN+T predictor. This reduction already occurred in [11] and it is mainly due to the participation of new volunteers with preferences that may differ slightly from those present in the dataset. In any case, the CNN+T model still performs better than CNN+LSTM.

B. Force/Velocity Predictor used for Movement Estimation

As commented at the beginning of this Section, both predictors can be used to obtain an estimate of the desired trajectory. In the case of CNN+LSTM, by generating an estimate of the human's desired velocity from the prediction of its force (using for this purpose the same controller used by the robot to generate its velocity commands from $\boldsymbol{F_{Task,C}}$) and subsequently integrating this velocity. In the case of CNN+T, directly integrating the velocity prediction generated by the model.

It is worth mentioning that the trajectory estimate obtained from CNN+LSTM refers to the trajectory that the human would like to follow based on the force exerted. Meanwhile, the one obtained from CNN+T refers to the one that the model estimates that the human-robot pair will follow. This difference will be negligible as long as human and robot interpret the scenario in the same way and collaborate with each other in equal proportion. Fig. 4 - A shows an example of this situation. However, if there is any discrepancy between the human and robot contribution, this estimate does show variations between one model and the other. Fig. 4 - B and Fig. 4 - C show two examples of this situation.

Fig. 4 - B shows the case where the human wishes to avoid a particular obstacle at all costs, even if that means bringing the robot closer to another obstacle. The robot will take into account the human's input and move away from the first obstacle but will also make its own contribution by avoiding getting too close to the second obstacle. This discrepancy causes the trajectory estimated with CNN+LSTM to differ from the trajectory finally followed. The same does not occur with the estimate obtained from the predicted velocity with CNN+T, since it takes into account the contribution expected by both agents and is closer to the real trajectory.

Fig. 4 - C shows the case where the human contribution is significantly lower than that of the robot exerting a very low force. The trajectory estimated using the CNN+LSTM output generates a smaller displacement by considering only the human contribution. However, the trajectory generated from the CNN+T velocity prediction does take into account that the robot will compensate for the low human contribution.

The Table IV shows a comparison between the movement estimate generated from our predictors and other models. The values shown in the first three methods are obtained by interpolation of the values reported by Laplaza et al. in [6]. The comparison with these models is not totally fair since these are used to predict the movement of only a human in tasks other than ours, but it allows to have a view of the typical values obtained in the State-of-the-Art. Comparing the result obtained using CNN+T with the one previously generated with CNN+LSTM or with the approximation of a second-order curve to the previous trajectory, a remarkable improvement is observed. It is worth mentioning that this is not only due to the use of more efficient architectures such as Transformers (as shown in Table I) but to the modification in the general architecture to have a dual output system with a prediction of both the human's force and the pair's velocity.

V. CONCLUSIONS

We have explored the usefulness of State-of-the-Art Deep Learning architectures such as Transformers and its image sequence processing oriented version, ViViT, in the specific task of collaborative human-robot transport of objects. With them, we have improved the performance of our force predictor in all the objective variables analyzed, being this the main contribution. Thus, in real experiments, predictions for the next $1\ s$ are achieved with an acceptable error over 90.9% of the time for the force to be exerted by the human and over 95.2% for the speed of the human-robot pair.

Second, we have corroborated that architectures such as ViViT need very large datasets to fully displace well-established architectures such as CNNs. Thirdly, through an ablation study we have verified the low relevance of the goal position on the performance of our predictor. This allows us to consider modifying our predictor so that it is not dependent on this variable, thus allowing the task to be developed without a predetermined goal but one that the human can decide and change at any time.

As for the applicability of our work in other scenarios, our predictor can be useful in any task where physical forces are exchanged. We also believe that the improvement observed when using Transformers will hold in those setups where multiple inputs of diverse nature requiring richer encoding capabilities are considered.

REFERENCES

[1] E. A. Sisbot and R. Alami, "A human-aware manipulation planner," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1045–1057, 2012.

[2] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, "Human-aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1726–1743, 2013.

[3] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 83–90.

[4] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2891–2900.

[5] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 474–489.

[6] J. Laplaza, F. Moreno-Noguer, and A. Sanfeliu, "Context and Intention aware 3D Human Body Motion Prediction using an Attention Deep Learning model in Handover Tasks." IEEE, 2022, pp. 4743–4748.

[7] M. Dalmasso, A. Garrell, J. E. Domínguez-Vidal, P. Jiménez, and A. Sanfeliu, "Human-Robot Collaborative Multi-Agent Path Planning using Monte Carlo Tree Search and Social Reward Sources," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2021.

[8] J. E. Domínguez-Vidal, I. J. Torres-Rodríguez, A. Garrell, and A. Sanfeliu, "User-Friendly Smartphone Interface to Share Knowledge in Human-Robot Collaborative Search Tasks," in *30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2021, pp. 913–918.

[9] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2015.

[10] J. E. Domínguez-Vidal, N. Rodríguez, and A. Sanfeliu, "Perception-Intention-Action Cycle as a Human Acceptable Way for Improving Human-Robot Collaborative Tasks," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, p. 567–571.

[11] J. E. Domínguez-Vidal and A. Sanfeliu, "Improving Human-Robot Interaction Effectiveness in Human-Robot Collaborative Object Transportation using Force Prediction," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7839–7845.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] S. Tarbouriech, B. Navarro, P. Fraisse, A. Crosnier, A. Cherubini, and D. Sallé, "Admittance control for collaborative dual-arm manipulation," in *2019 19th International Conference on Advanced Robotics (ICAR)*. IEEE, 2019, pp. 198–204.

[14] D. J. Agravante, A. Cherubini, A. Bussy, P. Gergondet, and A. Kheddar, "Collaborative human-humanoid carrying using vision and haptic sensing," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 607–612.

[15] Z. Li, J. Liu, Z. Huang, Y. Peng, H. Pu, and L. Ding, "Adaptive impedance control of human–robot cooperation using reinforcement learning," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 10, pp. 8013–8022, 2017.

[16] X. Yu, B. Li, W. He, Y. Feng, L. Cheng, and C. Silvestre, "Adaptive-constrained impedance control for human–robot co-transportation," *IEEE transactions on cybernetics*, vol. 52, no. 12, pp. 13 237–13 249, 2021.

[17] K. Kosuge and N. Kazamura, "Control of a robot handling an object in cooperation with a human," in *Proceedings 6th IEEE International Workshop on Robot and Human Communication. RO-MAN'97 SENDAI*. IEEE, 1997, pp. 142–147.

[18] O. M. Al-Jarrah and Y. F. Zheng, "Arm-manipulator coordination for load sharing using reflexive motion control," in *Proceedings of International Conference on Robotics and Automation*, vol. 3. IEEE, 1997, pp. 2326–2331.

[19] A. Bussy, P. Gergondet, A. Kheddar, F. Keith, and A. Crosnier, "Proactive behavior of a humanoid robot in a haptic transportation task with a human partner," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 962–967.

[20] X. Yu, W. He, Y. Li, C. Xue, J. Li, J. Zou, and C. Yang, "Bayesian estimation of human impedance and motion intention for human–robot collaboration," *IEEE transactions on cybernetics*, vol. 51, no. 4, pp. 1822–1834, 2019.

[21] C. N. Mavridis, K. Alevizos, C. P. Bechlioulis, and K. J. Kyriakopoulos, "Human-robot collaboration based on robust motion intention estimation with prescribed performance," in *2018 European Control Conference (ECC)*. IEEE, 2018, pp. 249–254.

[22] F. Dimeas and N. Aspragathos, "Reinforcement learning of variable admittance control for human-robot co-manipulation," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 1011–1016.

[23] E. Gribovskaya, A. Kheddar, and A. Billard, "Motion learning and adaptive impedance for robot control during physical interaction with humans," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 4326–4332.

[24] A. Al-Yacoub, Y. Zhao, W. Eaton, Y. M. Goh, and N. Lohse, "Improving human robot collaboration through force/torque based learning for object manipulation," *Robotics and Computer-Integrated Manufacturing*, vol. 69, p. 102111, 2021.

[25] S. S. Ge, Y. Li, and H. He, "Neural-network-based human intention estimation for physical human-robot interaction," in *2011 8th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*. IEEE, 2011, pp. 390–395.

[26] Y. Li and S. S. Ge, "Human–robot collaboration based on motion intention estimation," *IEEE/ASME Transactions on Mechatronics*, vol. 19, no. 3, pp. 1007–1014, 2013.

[27] Y. Cheng, L. Sun, C. Liu, and M. Tomizuka, "Towards efficient human-robot collaboration with robust plan recognition and trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2602–2609, 2020.

[28] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[29] Q. Xiong, J. Zhang, P. Wang, D. Liu, and R. X. Gao, "Transferable two-stream convolutional neural network for human action recognition," *Journal of Manufacturing Systems*, vol. 56, pp. 605–614, 2020.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[32] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[33] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[34] D. Helbing and P. Molnar, "Social Force Model for Pedestrian Dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[35] J. E. Domínguez-Vidal, N. Rodríguez, R. Alquézar, and A. Sanfeliu, "Perception-Intention-Action Cycle in Human-Robot Collaborative Tasks," *arXiv preprint arXiv:2206.00304*, 2022.

[36] J. E. Domínguez-Vidal, N. Rodríguez, and A. Sanfeliu, "Perception-Intention-Action Cycle in Human-Robot Collaborative Tasks: the Collaborative Lightweight Object Transportation Use-Case," *International Journal of Social Robotics*, p. to appear, 2024.

[37] J. E. Domínguez-Vidal and A. Sanfeliu, "Inference VS. Explicitness. Do We Really Need the Perfect Predictor? The Human-Robot Collaborative Object Transportation Case," in *32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 1866–1871.

[38] L. Deininger, B. Stimpel, A. Yuce, S. Abbasi-Sureshjani, S. Schönenberger, P. Ocampo, K. Korski, and F. Gaire, "A comparative study between vision transformers and cnns in digital pathology," *arXiv preprint arXiv:2206.00389*, 2022.