Real-Time 3D Reconstruction in Robot-Assisted Therapy for Autistic Children

Bernardo Silva Miguel Fernandes de Carvalho

***Abstract*—This thesis focuses on enhancing Robot-Assisted Therapy (RAT) for children with Autism Spectrum Disorder (ASD), a condition whose prevalence has been increasing, now affecting 1 in every 36 children. An important aspect of RAT is observing the child's behavior through non-intrusive sensors, among which the Kinect system has been prevalently used. However, this system often encounters difficulties in handling occlusions, particularly during real-time therapeutic interactions. To overcome this limitation, our research focused on 3D real-time skeleton reconstruction models, using a single camera. First, we adapted the existing Coherent 3D Reconstruction of Multiple Humans (CRMH) model to reconstruct children's bodies, creating the CRMH-p. Subsequent comparative analyses against other state-of-the-art models confirmed CRMH-p's superior balance between accuracy and run-time while handling occlusions. Our next challenge was integrating CRMH-p into the RAT framework, ensuring harmonious operation with the NAO robot and the existing gesture recognition system. After a comprehensive analysis, we encapsulated the CRMH-p model within a Docker container, utilizing the RabbitMQ communication system for efficient data exchange. This integration maintained the integrity of real-time processes critical for RAT. Finally, we tested the complete system in environments that mirrored live therapy sessions. CRMH-p, integrated within the RAT setup, achieved high-performance metrics, maintaining a robust run-time of 4.98 FPS in these simulated conditions. These results are an important step towards offering consistent, non-disruptive feedback for the children's continuous engagement and progress in therapy sessions. In conclusion, the developed system is primed for incorporation in upcoming clinical acquisitions, ensuring its impact in the therapeutic realm.**

***Index Terms*—Autism Spectrum Disorder, Robot-Assisted Therapy, 3D skeleton reconstruction, Occlusions, Real-time.**

I. Introduction

AUTISM Spectrum Disorder (ASD) is a neurodevelopmental condition that varies widely in its impact on social interaction and communication skills. With the prevalence of ASD rising, management strategies have evolved, focusing on individualized therapeutic interventions. A notable advancement is Robotic-Assisted Therapies (RAT), where Socially-Assistive Robots (SAR) are employed to enhance social engagement through gestural coaching, addressing a significant challenge faced by individuals with ASD. This research delves into integrating sophisticated skeleton recognition models to enable real-time engagement between the child and the robot.

This collaborative research between Instituto Superior Técnico, Lisbon, Portugal, and Politecnico di Milano, Milan, Italy, aims to enhance therapeutic interventions using technological integrations. At the heart of this project is a therapeutic protocol where a robot mirrors a child's movements to improve gestural communication. Given the heightened sensitivity of children with ASD, especially regarding touch, the incorporation of non-intrusive sensors in therapy becomes vital [1]. Consequently, we use a single Kinect camera for capturing motions. However, the Kinect system [2] faces challenges in handling occlusions, which are common during therapy sessions, as shown in Figure 1. A past study [3] revealed that Kinect lost over two-thirds of data due to occlusions, resulting in suboptimal engagement between the children and the robot. This thesis targets the development of a real-time 3D reconstruction and pose estimation system capable of operating in cluttered environments and effectively handling occlusions, aiming to improve the visual and vocal feedback in RAT.

Fig. 1. Example of challenging position during therapy captured by the Kinect camera. The represented skeleton is a hybrid of the upper body of the therapist and the lower body of the child (extracted from ([4]).

This research starts with a comprehensive analysis of state-of-the-art models addressing occlusions, and identifies potential candidates for integration into our single-camera RAT environment. Special attention was directed towards understanding the subsystems within therapy sessions, with a spotlight on the gesture recognition system. The research transitioned to the integration of the selected model, ensuring compatibility with the existing ecosystem and augmenting the precision of skeleton reconstructions. The ultimate goal is to forge a more responsive, effective, and seamless RAT experience for children with ASD, that is ready for testing in live clinical acquisitions.

[1]B. Carvalho is with Instituto Superior Técnico, Lisboa, Portugal.

II. State of the Art

The inherent complexity of ASD, along with the diverse ways it affects individuals, renders a singular best therapy approach elusive [5]. While various therapeutic strategies exist, each is tailored to specific needs, but all share a common goal: to enhance and stimulate interaction.

Research indicates that individuals with ASD often exhibit enhanced responsiveness and positive behaviors when interacting with mechanical devices, computers, and especially robots [6]. In settings where robots are integrated into the therapy, there is a notable uptick in the child's engagement, attention span, and the emergence of specific social behaviors such as cooperative enthusiasm and spontaneous imitation [7]. Some children even exhibit behaviors like attempting eye contact or showing signs of empathy, positioning robots in a unique category—somewhere between inanimate toys and sentient beings. This heightened engagement can be linked to the novel sensory stimulations robots introduce [8].

Current sensor technologies, while advanced, still face challenges in accurately capturing and interpreting the complex range of human emotions and behaviors [9]. The challenge complicates even further with individuals with ASD who may exhibit non-standard behavioral patterns [10]. The limitations in precision and the scope of these sensors mean that certain subtle gestures or uncommon expressions might go undetected or be misinterpreted, impacting the therapy's effectiveness.

Due to the challenge posed by the necessity for non-intrusive sensors for ASD children, our project opted for the utilization of a Kinect camera system for this task. The initial approach aimed at reconstructing the 3D scene leveraging the Kinect system, however, it faltered in adequately handling occlusions, as showcased in Figure 1, which led to data loss and suboptimal performance. This bottleneck has driven our exploration towards three distinct models that proficiently tackle occlusions from a single RGB image, thereby promising a more robust solution for real-time action observation and understanding.

Jiang et al. [11] introduced the Coherent Reconstruction of Multiple Humans (CRMH) model. It uses an end-to-end framework for 3D pose and shape estimation of all the individuals in an image, able to deal with interactions between people, thanks to the integration of challenging positions in training. This model takes an RGB image as input and starts by detecting all people through a Faster-RCNN, providing bounding boxes around them. Following this, each individual is reconstructed through their skeletons, utilizing an additional input of focal length and the SMPL model parameters. For a deeper understanding of SMPL consult [12]. Ultimately, the process culminates in a 3D scene, delivering an accurate mesh representation of the individuals [11]. This final step is achieved due to the incorporation of the actual translation of each individual and the use of a full perspective camera at each bounding box's center with a corresponding focal length f.

By defining these virtual camera models, the following translation vector was derived with (1). This translation vector represents the relationship from each bounding box's local camera to the overarching scene's primary camera. Notably, this configuration ensures that a point, when projected through a camera positioned at a bounding box's center, aligns with its projection from a centrally-placed image camera.

$$t_i = \begin{bmatrix} \frac{d_i\left(x_i\alpha_i + c_{i,x} - \frac{w}{2}\right)}{f} & \frac{d_i\left(y_i\alpha_i + c_{i,y} - \frac{h}{2}\right)}{f} & d_i \end{bmatrix} \quad (1)$$

Here, α_i is the larger size of the bounding box (width w or height h), while x and y denote the bounding box's image coordinates. The centers of each bounding box are represented by $c_{i,x}$ and $c_{i,y}$ and d_i indicates the scene depth for each person, calculated with

$$d_i = \frac{2f}{s_i\alpha_i} \quad , \quad (2)$$

where s_i is a model-estimated scale factor.

On the other hand, Sun et al. [13] proposed the Regression Of Multiple People (ROMP) model which employs a one-stage approach, analyzing entire images instead of individual bounding boxes using a body-center-guided pixel-level representation. It produces three maps per image: a body center heatmap, a camera map, and an SMPL map for mesh parameters. For each 2D body center, the 3D mesh parameters are extracted for the SMPL model to create 3D body meshes. Notably, this method was only used on adults. To deal with this problem, [14] adapted it for children by adding the SMIL [15] model to SMPL, using an age-related parameter. [14] trained this on a new dataset including children and introduced a Bird's eye view map for depth reasoning, indicating probable body centers in depth.

The choice of the best approach for 3D reconstruction depends considerably on the application in a real-world therapeutic context. In our specific case, we collaborated with health professionals to design a protocol where the robot NAO presents eight different gestures: big, little, me, hello/goodbye, giving, pointing, yes, and no. The sessions have a triadic format, involving the robot, the therapist, and the child. Each session has a structured flow, maintaining uniformity in the gestures performed. However, it is important to note that while this protocol provides a guideline, therapists have full liberty to adapt it according to the unique needs and progress of each child.

To provide performance feedback to the child, a gesture recognition model developed by [16] is used. The system was designed to process Kinect skeletons as its input, subsequently analyzing their poses to deduce a final gesture prediction. [16] focuses solely on the upper body joints and is trained to recognize 19 distinct gestures.

In summary, the Kinect system's limitations have previously affected the effectiveness of NAO robot's integration into therapy, mainly due to challenges in accurate skeleton identification. Each of the three alternative models presented has its own limitations. The CRMH model [11] addresses occlusions but is constrained by its reliance on an adult-only dataset, potentially affecting its accuracy for younger users. ROMP [13], with its faster processing, uniquely addresses occlusions but shares the same dataset constraint. BEV [14], by incorporating age factors, presents a solution to the skeletal

reconstruction of children, a gap observed in both CRMH and ROMP. However, its processing time might be compromised due to additional layers for age consideration.

There is a gap in the current literature regarding a direct comparison of the three models concerning their 3D reconstruction accuracy and processing speed, which is crucial for our project's objectives. To address this, the next step was implementing and evaluating these models to provide a comparative analysis providing ground in understanding the performance of each model, aiding in the selection of the best fit for our project to enhance the therapy experience of ASD children.

Following we present the process to ensure compatibility of the chosen model with the gesture recognition system. Finally, we discuss its implementation of our RAT framework testing the final performance of the all system in resembling live therapy scenarios.

III. Methodology

A. CRMH personalized

Using the CRMH model on child session images revealed a depth-based translation between the child and adult. Given that the translation vector in (1) relies on parameter f, we adjusted it to better suit children. We found f varied with height, prompting us to create a height-based regression model for estimating it. This approach, using data from the CADin study videos captured by Kinect, ensured fast testing times, essential for real-time applications.

To create the regression model, we used the RANSAC method to associate the focal lengths and heights of six therapists who participated in our study. This decision was made due to RANSAC's robustness in handling outliers within the dataset, which are common occurrences in real-world measurement data. To calculate f for each therapist, we selected a variable window of frames in which they were at known positions. The positions chosen were 2.2 m in depth for four of the therapists, 2.5 m, and 3.1 m for the other two. Then, as a loss function to decide the consensus set, we used the weighted sum of squares. The weights considered the uncertainty associated with each of our known positions and caused by the perspective projection. Thus using the parameter f as the focal length, the uncertainty in the depth direction increases quadratically in (3), where h is the height of the person and Z is the respective depth.

$$\begin{aligned} y &= \frac{hf}{Z} \\ \frac{\delta y}{\delta Z} &= -\frac{hf}{Z^2} \\ \delta Z &= -\frac{Z^2}{hf}\delta y \end{aligned} \tag{3}$$

Therefore, our loss function is given by (4), in which $w_i = \frac{1}{\sigma_i^2}$, $\sigma_i \propto Z_i^2$, f_i represents the focal length measured for each therapist and $\hat{f}_i$ the focal length estimated by the regression model.

$$L = \frac{\sum_i w_i (f_i - \hat{f}_i)^2}{\sum_i w_i} \tag{4}$$

To choose the best model from those generated by the RANSAC method, we used the coefficient of determination R^2 as a metric. An R^2 value closer to 1 suggests a stronger fit, thus better explaining the variance in the data. Using this, we pinpointed the model that best correlated the focal lengths with the heights of the therapists.

B. Comparison with ROMP and BEV

Due to the absence of a direct 3D reconstruction comparison between CRMH-p, ROMP, and BEV models, we designed a rigorous evaluation. To achieve this, we enlisted three participants of distinct heights: 1.41m (child), 1.62m, and 1.75m, and juxtaposed their captured data against the OptiTrack system [17], a gold standard in this domain due to its accuracy achieved through 12 strategically placed cameras in a well-lit room.

In our controlled setting, each participant began 1.3m away from the Kinect, adopting an open-armed stance. After holding this pose for 5 seconds, they progressively retreated backward, stopping at the following positions in sequence: 1.3m, 1.9m, 2.2m, 2.5m, and 3.0m. The captured data, especially focusing on the coordinates of the upper body joints, was analyzed in context to the Gesture Recognition model, as illustrated in Figure 2.

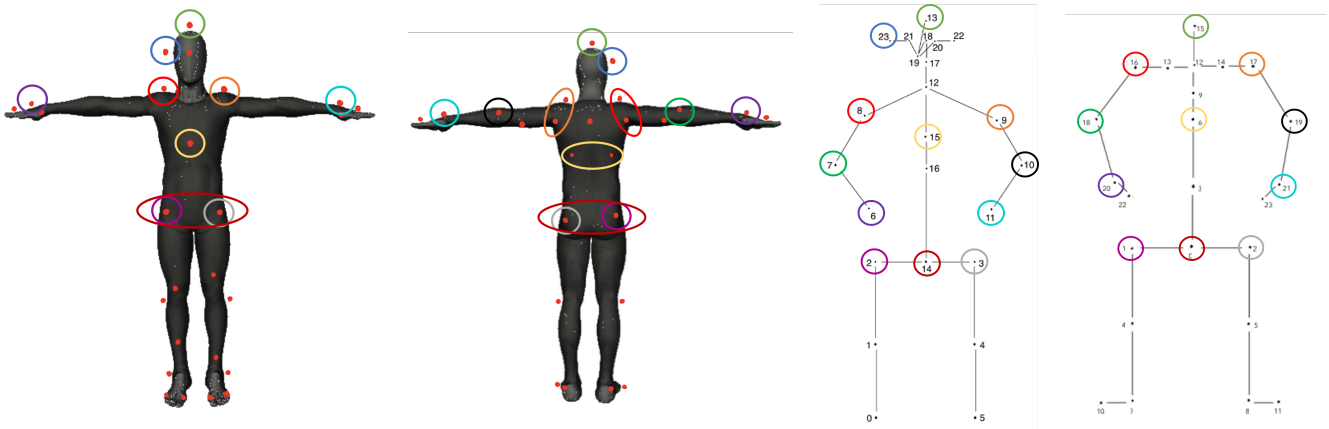
The metric used for comparison was the average 3D Root Mean Squared Error (RMSE) obtained by calculating the RMSE for each frame and then computing the average across all frames. This metric allowed us to assess the differences between the estimated positions obtained from both systems, providing an overall measure of accuracy for the constructed model within a constrained environment.

In addition, we performed a depth analysis focusing on the z-axis (depth direction). We measured the behavior of the child's hip joint depth throughout the experience to be able to compare the performance of the models when reconstructing children. This joint was chosen as a representative point of interest for our analysis, since it had a clear movement in z, allowing a better comparison of the models. Furthermore, the operational feasibility of each model in real-time scenarios is evaluated using their frames per second (FPS) rates.

C. Gesture Recognition

From our earlier experiment, the CRMH-p model emerged as the optimal choice due to its balance of accuracy and operational efficiency. Thus, we focused on adapting it to the pre-existing gesture recognition system, initially tailored for Kinect.

During integration with the gesture system, we grappled with the discrepancy in skeletal representations between the two models. To address this, we introduced an 'Encoder'—a computational mechanism that transforms CRMH-p skeleton data into a format aligning with Kinect. This encoder was built using a feed-forward neural network. Its input of 72 nodes, representing 3D coordinates of 24 joints, is processed to produce an output of 75 nodes, mirroring the Kinect skeletal structure.



(a) Optitrack markers' scheme (front view). (b) Optitrack markers' scheme (back view). (c) CRMH skeleton. (d) ROMP and BEV skeleton.

Fig. 2. Skeletons' scheme of the systems. The correspondences between joints used to evaluate are marked by circles of the same color. Since the systems' skeletons are not a perfect match, for some joints (for example 14 in (a)) we had to use the mean of two associated markers.

For its training, we orchestrated a controlled experiment where five participants mimicked therapy sessions. Their gestures, in both standing and seated stances, were recorded and processed through CRMH-p, using Kinect predictions as a reference. To evaluate the encoder, we utilized real therapy data and a controlled set featuring a neurotypical child. Due to the frequent occlusions in therapy data, the controlled dataset was crucial to ensure accurate skeletal predictions from both Kinect and CRMH-p.

To further substantiate our model's real-world utility, we organized a structured experiment around six key gestures: big, little, me, happy, giving, and pointing. Thirteen participants were involved over multiple sessions, mimicking potential therapeutic scenarios with occlusions. The first session had five participants—a child and four adults. In subsequent sessions, with two participants in the camera view, only one was actively assessed. The other was free-moving, introducing challenges to skeleton reconstruction.

During the experiment, participants sat on the ground, reflecting on the setup of the established protocol. A single Kinect camera was used to record approximately 20-second-long video segments, with each participant enacting a specified gesture.

The experiment was divided into four distinct settings, each representing a unique challenge for the gesture recognition model:

A. **One Subject**:

1. **Solo Setting** A straightforward scenario where a lone participant performed the chosen gestures.

B. **Two subjects**:

1. **Without Occlusion**: Two participants sat side by side, emulating the standard therapy setup. Only one of the two was evaluated, although both were visible in the frame.
2. **Visual Clutter**: Designed to replicate instances where participants might inadvertently overlap in the therapist's field of view, the evaluated individual sat slightly in front of another.
3. **Partial Occlusion**: A reversal of the previous setting, the participant under evaluation was seated slightly behind another, partially obscured from direct view.

Given the continuous output of gesture predictions by the gesture recognition model, the evaluation metric used for this experiment is uniquely tailored. Specifically, for any given video, a successful recognition is registered if the correct gesture is predicted at least once throughout the video's duration. Conversely, if the correct prediction does not occur, that session is considered unsuccessful. Using this criterion, the metric of success is defined as the average percentage of successful gesture recognitions over all the videos for each gesture type. The results are then compared between data acquired from Kinect and CRMH-p. Beyond accuracy, understanding the data availability (skeletons detected) is also relevant. Hence, the mean percentage of detected skeletons is also computed grouped by gesture types.

Through this evaluation process, we expect to highlight CRMH-p's advanced capabilities in maintaining accurate skeleton reconstructions, regardless of environmental complexities and potential occlusions, thus, improving the gesture recognition accuracy. This direct comparison with the Kinect system solidifies the argument for CRMH-p's suitability in enhancing non-verbal communication during therapy sessions.

IV. Real-Time Skeleton Reconstruction

After ensuring the compatibility of the CRMH-p model with the gesture recognition model in offline, we studied its integration in a real-time system to be used in our RAT framework. Central to our approach to meeting this real-time requirement is the deployment of the CRMH-p model within a Docker environment [18]. This technology allows us to encapsulate the model's functionalities, ensuring consistent, isolated, and efficient execution, irrespective of the underlying hardware.

However, encapsulating the CRMH-p model within Docker introduces a significant challenge: the necessity for robust communication between the local environment and the Docker

container. To overcome this challenge we opted to implement RabbitMQ system, a renowned open-source message broker that has gained traction in numerous industries due to its performance and reliability [19].

In order to ensure the best processing time, we investigated sequential and parallel paradigms. Therefore, we devised the following systems.

- **System 1 - Sequential Processing:** In this setup, each frame captured by the Kinect is sent to the Docker container through RabbitMQ, processed in a first-come-first-serve manner, and the results are sent back similarly. While this system ensures order preservation and consistency, it faces challenges in high-throughput scenarios due to the inherent latency in processing frames one at a time.
- **System 2 - Parallel Processing:** In contrast, this system processes multiple frames concurrently. It uses RabbitMQ for its messaging system, handling a higher volume of data exchanges more efficiently. This system, while geared for high efficiency, introduces complexity in managing thread synchronization and maintaining data consistency due to the simultaneous processing of multiple frames. To navigate these challenges, we designated a dedicated thread for dispatching the processed skeleton results back to the local environment. Although this approach requires additional computational resources, it strategically circumvents concurrency issues, ensuring controlled data transmission and system stability.

To test our skeleton reconstruction methodology's robustness, we used three distinct therapy sessions from past acquisitions. We evaluated the frame rate each system achieved across these videos, to measure both consistency and robustness under varying session conditions.

Additionally, we analyzed the factors contributing to the processing time between the Kinect camera's frame capture and receiving the detected skeletons from Docker. We determined the average time each system required per frame, dividing it into communication time and CRMH-p processing time, thereby evaluating RabbitMQ communication performance.

In selecting the best system, we weighed frame rates, time delay, and also computational costs, especially given the demands of integrating with resource-intensive systems like the NAO robot and gesture recognition. Ensuring runtime efficiency, stability, and computational balance is vital for effective therapy. After choosing and implementing the system, we analyzed its performance, confirming its seamless integration with the NAO and gesture recognition while upholding the standards and conditions for real-time therapeutic application.

V. Experimental Results and Discussion

A. CRMH personalized

Using the RANSAC method to avoid outliers, and the heights of six therapists, we chose the model with the highest coefficient of determination presented in Table I in green. From this table, we can observe that when using the weighted sum of squares, the model is more robust and achieves a higher coefficient of determination. Thus, extending the CRMH capability of accurately regressing children. Figure 3 shows the final output difference between the original CRMH (a) and our proposed model, CRMH-p (b). The correction of the position of the child is noticeable.

TABLE I
RESULTS FROM THE RANSAC MODEL USING TWO DIFFERENT LOSS FUNCTIONS (SUM OF SQUARES OR WEIGHTED SUM OF SQUARES). EACH ROW CORRESPONDS TO A LINEAR MODEL ($f = Slope \times height + Intercept$) GENERATED BY THE RANSAC MODEL WITH A COEFFICIENT OF DETERMINATION R^2. THE SELECTED MODEL IS UNDERLINED IN GREEN.

	R^2	Slope	Intercept
Sum of Squares	0.7660	250.51	-4.51
	0.7548	271.56	-38.58
	0.8383	207.05	65.66
	0.7789	242.16	8.93
	0.9780	696.14	-750.66
Weighted sum of squares	0.9943	158.20	145.72
	0.9959	164.47	135.23
	0.9959	164.47	135.23
	0.9959	164.47	135.23
	0.9959	164.47	135.23

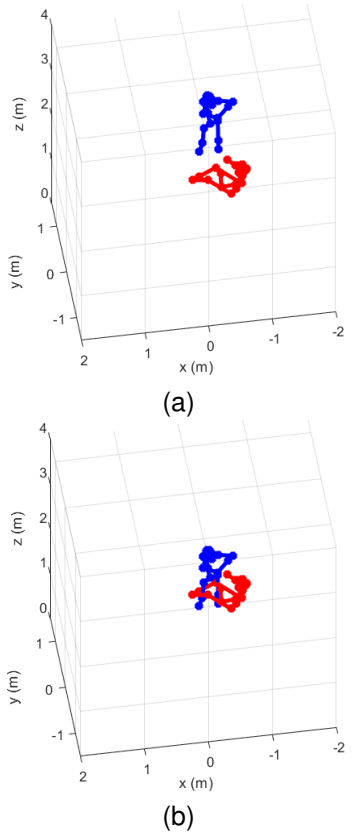


Fig. 3. Correcting pose of the child skeleton through personalized focal length. The child's skeleton (blue) with CRMH had its skeleton further away. With CRMH-p we can observe he is side-by-side with the therapist (red).

B. Comparison between Skeleton Reconstruction Models

Building on our development of the CRMH-p model, we delve into a comparative analysis with the ROMP and BEV models. We primarily focus on the average 3D RMSE and FPS rates to determine each model's accuracy and real-time

applicability. The Average 3D RMSE evaluates the precision of skeletal joint location estimations against the Optitrack system's ground truth. Table II presents the results for each model.

TABLE II
COMPARATIVE ANALYSIS OF RMSE VALUES (m) FOR DIFFERENT METHODS ACROSS HEIGHTS AND DEPTH RANGES.

Height (m)	**Depth Range** (m)	**CRMH-p**	**ROMP**	**BEV**
1.41	$z < 1.9$	0.28	0.25	0.24
	$1.9 < z < 2.5$	0.20	0.33	0.35
	$z > 2.5$	0.27	0.37	0.42
1.62	$z < 1.9$	0.25	0.25	0.25
	$1.9 < z < 2.5$	0.23	0.28	0.27
	$z > 2.5$	0.25	0.37	0.38
1.75	$z < 1.9$	0.28	0.24	0.25
	$1.9 < z < 2.5$	0.14	0.20	0.19
	$z > 2.5$	0.17	0.29	0.28

A breakdown of these results is instrumental in gleaning significant insights into the robustness and adaptability of each model under diverse real-world conditions.

The first observation is the CRMH-p model's steadfast accuracy across all participant heights. Unlike other models, its performance remains relatively unaltered regardless of the individual's height, maintaining consistent readings, especially in deeper zones ($1.9 < z < 2.5$ and $z > 2.5$). This trait underscores its reliability and broad applicability, essential for therapeutic settings involving diverse participants.

On the other hand, BEV, designed to enhance skeletal tracking in children using an age factor, was anticipated to perform better with younger subjects. However, its accuracy decreases with depth, despite having a slight advantage at closer ranges. Both ROMP and BEV are more accurate at closer depths ($z < 1.9$) across all heights, but their precision drops at greater depths, especially beyond $z > 2.5$. This suggests their restricted use in broad movement situations or when subjects are far from the sensor. Figure 4 depicts the hip joint's depth change during a child's session, highlighting this depth sensitivity rather than skeletal detection issues.

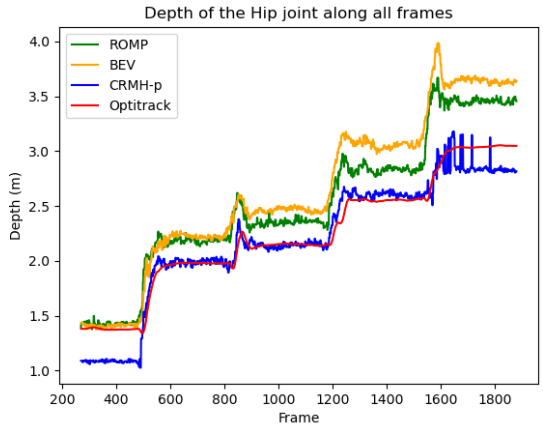


Fig. 4. Depth values for the child's hip joint using different systems: CRMH-p (blue), BEV (orange), ROMP (green) and Optitrack (red).

Considering the therapeutic context, precise motion tracking across comprehensive depths is essential. While ROMP and BEV can be adequate for restricted or near-range motions, their accuracy diminishes significantly for activities requiring more extensive movement. Conversely, CRMH-p demonstrates superior, reliable performance, making it a more fitting choice for therapeutic applications. Its exceptional performance in the standard therapeutic range ($1.9 < z < 2.5$) underlines its practical efficacy for real-world therapy sessions.

While the precision of skeletal tracking is undeniably crucial, it does not solely define the efficacy of a model, especially in real-time applications where processing speed is equally vital. Thus, to construct a more holistic evaluation, we extend our analysis to investigate the frame rates at which each model operates, as encapsulated in Table III.

TABLE III
COMPARATIVE ANALYSIS OF FRAMES PER SECOND (FPS) VALUES FOR DIFFERENT METHODS ACROSS VARIOUS HEIGHTS.

Height (m)	**CRMH-p**	**ROMP**	**BEV**
1.41	8.02	8.11	5.25
1.62	7.96	8.03	5.24
1.75	7.92	8.01	5.12
Mean	7.97	8.05	5.20
Std Dev	0.050	0.053	0.057

In the frame rate analysis, each model underscores different priorities in their design and functionality, highlighting a critical aspect of real-time applications: the balance between run-time and accuracy.

From Table III, ROMP achieves an average of 8.05 FPS, clearly prioritizing speed in its design, making it suitable for applications demanding instant feedback. Not far behind is the CRMH-p, registering an average of 7.97 FPS. Despite its intricate design centered on accuracy, the CRMH-p still manages to offer commendable speed, reflecting its well-rounded efficiency tailored for real-time contexts.

In contrast, the BEV model logs an average of just 5.20 FPS, signaling its preference for detailed analytical accuracy, often sacrificing speed. While this meticulousness can benefit certain applications, its slower performance raises considerations for real-time scenarios, where timely feedback is essential.

CRMH-p, with its methodological refinement that includes personalized focal length adjustments, shows a marked improvement in accuracy, placing it on par with leading models. Even with this enhanced accuracy, it retains a commendable operational speed, making it highly suitable for real-time applications. This balance solidifies CRMH-p's position as a pragmatic tool tailored for therapeutic contexts.

In the subsequent sections, we discuss the practical application of CRMH-p within our therapeutic framework, highlighting its significance in real-world scenarios and exploring the essential variables for its successful implementation.

C. Gesture Recognition

Table IV enumerates the RMSE values, observed at the culmination of both training and validation phases of the encoder.

From our results, we deduced that while the validation loss may not be irrelevant, it remains within acceptable bounds,

TABLE IV
FINAL LOSS VALUES (RMSE) AFTER THE TRAINING AND VALIDATION PHASES, INDICATING THE OPTIMAL EPOCH CHOSEN BASED ON VALIDATION LOSS.

	RMSE (m)	**Best Epoch**
Training Loss	0.078	32
Validation Loss	0.209	

confirming that the encoder model can reliably transform CRMH-p skeletal data to be compatible with Kinect.

Moreover, the controlled experiment aimed to assess the recognition capabilities of the Kinect and CRMH-p models in standardized settings. By monitoring the models under controlled conditions, we endeavored to ascertain the inherent strengths and weaknesses of each, while also understanding their performance vis-a-vis data availability.

Table V presents the mean percentage and standard deviation of successful gesture recognitions, capturing the performance intricacies of both Kinect and CRMH-p models across diverse settings.

TABLE V
COMPARISON OF AVERAGE GESTURE RECOGNITION RATES ACROSS DIFFERENT SETTINGS BETWEEN KINECT AND THE PROPOSED MODEL, CRMH-P.

Settings	**Kinect (%)**		**CRMH-p (%)**	
	Mean	Std Dev	Mean	Std Dev
Solo Setting	73	24	57	29
Without Occlusion	75	18	63	18
Visual Clutter	63	16	65	28
Partial Occlusion	40	17	61	23

In clear, unobstructed environments, especially with solo participants or multiple participants without occlusion, the Kinect model surpasses CRMH-p. Specifically, Kinect boasts an accuracy rate of 73% compared to CRMH-p's 57% in isolated gesture scenarios. This disparity suggests that Kinect flourishes in simpler settings without interference, while CRMH-p's performance might be hampered due to the complexities of its encoder model. However, the narrative shifts in visually challenging or partially occluded settings. Here, CRMH-p starts to shine, often matching or even outperforming Kinect. Such observations hint at CRMH-p's strengths in processing imperfect or obscured visual data.

A gesture-specific breakdown reveals varying degrees of recognition success. For example, both models proficiently recognize 'Pointing' in most conditions. Conversely, 'Giving' is more challenging, particularly for CRMH-p in less complex environments. This variance underscores the inherent intricacies of certain gestures.

Importantly, CRMH-p's consistent performance across multiple scenarios highlights its potential value for therapy sessions that might involve a variety of conditions, including obstructions or visual distractions.

In conclusion, while both models, Kinect and CRMH-p, showcase their unique strengths and capabilities, their overall performance is still below the required for ensuring smooth and reliable gesture recognition during therapy sessions. This prompts an exploration into potential causes for these deviations.

Given the expansive capability of the Gesture Recognition model, which can recognize up to 19 gestures, juxtaposed against the limited 8 gestures used in our therapy protocol, a pertinent question arises: Are the unused gestures impacting the accuracy of the model's predictions? To investigate this, we analyzed the data from sessions with just one participant present in Table VI. This approach helped us avoid errors from extra participants or outside disturbances in the scene.

With this analysis along with the understanding of each gesture dynamics, we present a set of justified groupings:

- **Big - Happy**: *Big* is frequently mistaken for *Happy* at 85%. This deviation is mainly caused since both gestures are executed with the two upper limbs that move simultaneously.
- **Little - Peekaboo**: The model tends to predict *Little* as *Peekaboo* 46% of the time. The similarity between the gestures is clear since both have the two hands close to each other.
- **Me - Hungry**: *Me* is misinterpreted as *Hungry* 25% of the time. The gestures have the participant put one hand in front of the upper body. The near 0% rate of *Hungry* being mistaken for other therapy gestures supports this pairing.
- **Hello**: 71% of the predictions for *Hello* lean towards *Pointing*. Despite both gestures being executed by lifting one arm, *Hello* has its limb moving while in *Pointing* it is steady. This result suggests the window of frames used may not have the number of frames necessary to have the arm moving and therefore misinterpret the gesture. However, as *Pointing* is a primary therapy gesture, the two cannot be combined.
- **Giving - Where - Waiting**: *Giving* is often misidentified as *Where* (57%). Since the gesture *Where* is composed by part of the gesture *Giving*, a small deviation in the skeleton can induce this error. The model also predicts *Giving* as *Waiting* in 25% of instances. This misinterpretation is comprehensible because they differ only by the rotation of the hand.
- **Pointing**: *Pointing* achieves a 74% recognition on its own, emphasizing its uniqueness among the gestures.

To provide a clear perspective on the effect of these groupings, Table VII respectively illustrates the prediction rates after the suggested modifications. It is important to mention that due to the omission of the rest of the other gestures that were not part of this analysis, the sum of the percentages in each row may not total 100%.

Given the comparative data presented in Tables VI and VII, we further discuss the impacts and insights from the suggested gesture groupings:

The gesture grouping strategy noticeably elevates recognition capabilities. Notably, the gesture *Big* has seen a transformative improvement, shifting from a mere 2% recognition rate to a commendable 87%, a change largely attributed to its frequent misidentification as *Happy*. Similarly, *Little*, which was previously confused with *Peekaboo* almost half the time, now stands at a 73% recognition rate. However, it's crucial to understand that grouping doesn't serve as a universal solution. This is underscored by gestures like *Me*. Even after being

TABLE VI
PREDICTION PERCENTAGE DISTRIBUTION BY THE GESTURE RECOGNITION MODEL USING SOLO SETTING DATA FOR ALL TRAINED GESTURES. IN EACH VIDEO, THE GESTURE RECOGNITION MODEL MAKES MULTIPLE PREDICTIONS ACROSS THE WHOLE SESSION. THE CORRECT PREDICTION IS IN GRAY AND THE COMMON MISIDENTIFICATION IN GREEN

Gestures (%)	Big	Little	Me	Hello	Giving	Pointing	Where	Angry	Listening	Coming	Peekaboo	Waiting	Hungry	Tall	Happy	Kissing	Short	Yes	No
Big	2	3	1	0	0	1	6	0	0	0	2	1	0	0	85	0	0	0	0
Little	0	27	1	1	0	0	16	0	0	0	46	0	0	2	0	0	0	0	7
Me	2	22	30	1	0	0	8	0	0	3	0	3	25	0	3	0	0	3	0
Hello	0	0	0	7	0	71	6	0	0	0	0	0	0	1	15	0	0	0	0
Giving	0	2	2	1	3	8	57	0	0	1	0	15	2	0	3	0	4	2	0
Pointing	0	3	0	11	0	74	4	0	0	0	0	0	0	0	3	0	2	3	0

TABLE VII
MODEL'S GESTURE RECOGNITION RATES AFTER GROUPING SIMILAR GESTURES BASED ON FREQUENT MISINTERPRETATIONS. THE MERGED GESTURES ARE INDICATED IN THE HEADERS. EACH ROW PROVIDES THE PERCENTAGE OF TIMES A GESTURE WAS RECOGNIZED OR MISINTERPRETED AS ANOTHER, WITH THE TRUE GESTURE OR GESTURE GROUP HIGHLIGHTED IN GRAY. THE SUM OF PERCENTAGES IN EACH ROW MAY NOT TOTAL 100% DUE TO THE OMISSION OF OTHER GESTURES FROM THE COMPLETE SET OF 19 POSSIBLE GESTURES.

Gestures (%)	**Big** - Happy	**Little** - Peekaboo	**Me** - Hungry	**Hello**	**Giving** - Where - Waiting	**Pointing**
Big	87	4	1	0	6	1
Little	3	73	1	1	16	0
Me	5	22	55	1	11	0
Hello	15	0	0	7	6	71
Giving	3	2	4	1	75	8
Pointing	3	3	0	11	4	74

paired with *Hungry*, its recognition improvement is moderate, moving from 30% to just 55%, indicating inherent limitations in the skeleton reconstruction model.

After analyzing the prediction distribution in the Solo setting, we evaluated the impact of a group strategy on both Kinect and CRMH-p models across all settings. Using this strategy may reveal distinct advantages and challenges of each model. Table VIII shows the gesture recognition success rates for each condition when applying the group strategy.

TABLE VIII
COMPARISON OF GESTURE RECOGNITION RATES' MEAN AND STANDARD DEVIATION ACROSS DIFFERENT SETTINGS BETWEEN KINECT AND THE PROPOSED MODEL, CRMH-P.

Settings	**Kinect (%)**		**CRMH-p (%)**	
	Mean	Std Dev	Mean	Std Dev
Solo Setting	93	16	80	22
Without Occlusion	86	16	77	18
Visual Clutter	80	15	84	15
Partial Occlusion	61	17	77	18

When comparing the results from Table V with the post-grouping figures in Table VIII, several key insights arise. There is a notable uptick in mean accuracies across all scenarios. This indicates that our group strategy, derived solely from solo setting data, has a robust nature that generalizes well across various conditions.

Moreover, the relative standing between the Kinect and CRMH-p models remains consistent both before and after applying the group strategy. Kinect performs particularly well in simpler environments such as the Solo Setting and Without Occlusion. However, its performance wanes in settings where there is 2D interference, especially during partial occlusion. On the other hand, the CRMH-p model showcases remarkable consistency, maintaining its performance across an array of scenarios

Furthermore, data availability can be seen as an indicator of the reliability or accuracy of the models in detecting and interpreting specific gestures. An efficient model should consistently recognize and track skeletons across various scenarios. Table IX presents the results obtained.

TABLE IX
COMPARISON OF AVERAGE DETECTED SKELETONS PERCENTAGE ACROSS DIFFERENT SETTINGS BETWEEN KINECT AND THE PROPOSED MODEL, CRMH-P.

Settings	**Kinect (%)**		**CRMH-p (%)**	
	Mean	Std Dev	Mean	Std Dev
Solo Setting	63	2.3	77	6.4
Without Occlusion	54	1.5	81	1.2
Visual Clutter	53	2.5	79	3.4
Partial Occlusion	50	4.2	81	1.0

In our evaluation of data availability across different settings, the CRMH-p model consistently outperforms the Kinect by an average of roughly 20 percentage points. This disparity aligns with our initial expectations, given the differences between the two models' capabilities.

While one might expect the Kinect's performance to decline in complex environments, especially those with occlusions, the data shows the Kinect maintaining relatively consistent frame availability across scenarios. However, it is essential to note that this evaluation focuses on the number of detected skeletons without considering their accuracy. Thus, the Kinect might maintain a certain detection rate, but the reliability of these detections remains questionable. The distinction between quantity and quality of detections is pivotal when considering the models' real-world applicability.

From the data we have collected, even after undergoing the skeletal transformation process, our proposed model demonstrates unwavering consistency and reliability across varied scenarios typical in therapeutic settings. The introduction of the group strategy further bolsters its capabilities, consistently achieving an approximate 80% detection rate for both gestures and skeletons. This resolves the data loss issues previously encountered with the Kinect-based acquisitions.

D. Real-Time Skeleton Reconstruction

Our focus progresses to identifying the optimal architecture to implement the CRMH-p model to process in real-time sessions integrated into our RAT framework. The two systems were evaluated to determine the optimal processing paradigm (sequential or parallel) using RabbitMQ communication between the local setup and the Docker container. This assessment excluded NAO and the gesture recognition model to focus solely on performance comparison. Tables X and XI showcase a detailed comparison of run-time and time-delay achieved by each system across three therapy sessions.

TABLE X
COMPARATIVE FRAMES PER SECOND (FPS) ANALYSIS BETWEEN SEQUENTIAL AND PARALLEL PROCESSING USING RABBITMQ COMMUNICATION SYSTEM

System	Mean (FPS)	Std Dev (FPS)
Sequential	7.59	0.07
Parallel	8.65	0.38

TABLE XI
COMPARISON OF COMMUNICATION, PROCESSING, AND TOTAL DELAY TIMES ACROSS SEQUENTIAL AND PARALLEL SYSTEMS USING RABBITMQ COMMUNICATION SYSTEM.

Split	Sequential		Parallel	
	Mean (s)	Std Dev (s)	Mean (s)	Std Dev (s)
Communication	0.027	0.018	0.063	0.112
Processing	0.104	0.015	0.163	0.057
Total Delay	0.130	0.012	0.226	0.143

Parallel processing, while achieving higher FPS, does not manifest a substantial advantage as one might anticipate from a system handling frames concurrently. This modest FPS enhancement can be ascribed to the computational overhead and communication requirements inherent to parallel methods. Furthermore, the parallel approach displays greater variability in performance, as evidenced by its higher standard deviation.

In evaluating Table XI, although the parallel method takes longer for individual frame processing, it compensates by analyzing multiple frames simultaneously. This added complexity is reflected in the elongated communication delays. Efficiently coordinating multiple frames and ensuring proper thread synchronization contribute to this extended communication time.

Considering therapeutic applications where consistency and reliability are vital, the results from sequential processing—with nearly equivalent FPS, reduced time delay, and lesser variation—make it the more favorable choice for such contexts.

To validate the performance of the chosen architecture integrated into the full system in a real-world setting, we recorded five sessions following the therapy protocol. Table XII presents the frame rates registered across these sessions, while Table XIII provides a comprehensive breakdown of the system's time delay when in full operation. Together, these results offer concrete evidence of the efficiency and reliability of the sequential processing with RabbitMQ and the overall practicality of the final system in our therapeutic setting.

When examining the data from Table XII and Table X, the system's performance evidently shifts based on its operational context. Without the robot, peak frame rates are achieved, indicating optimal computational efficiency. Yet, the integration of multiple subsystems leads to an average frame rate of 4.98 FPS. Simultaneously, time metrics from Table XIII and Table XI reveal that standalone operations result in swifter data exchanges and calculations. However, the presence of additional subsystems introduces minor increases in communication and processing times, despite which, the overall delay in the integrated system remains commendably low at 0.198s, with a low standard deviation underscoring consistent performance.

TABLE XII
FRAMES PER SECOND (FPS) ANALYSIS OF THE FULL SYSTEM USING SEQUENTIAL PROCESSING AND RABBITMQ COMMUNICATION TECHNIQUE.

	Mean (FPS)	Std Dev (FPS)
Full System	4.98	0.12

TABLE XIII
COMMUNICATION, PROCESSING, AND TOTAL DELAY TIMES FOR THE FULL SYSTEM USING SEQUENTIAL PROCESSING AND RABBITMQ COMMUNICATION.

Split	Mean (s)	Std Dev (s)
Communication	0.063	0.007
Processing	0.135	0.009
Total Delay	0.198	0.013

This holistic view emphasizes the system's resilience and adaptability, managing complexities while maintaining a dependable frame rate and minimal time delays, vital for seamless therapeutic sessions. In conclusion, we successfully implemented a real-time 3D pose reconstruction model, adept at handling occlusions, into our RAT framework.

VI. CONCLUSIONS AND FUTURE WORK

Central to this research was the creation of a system capable of precise 3D skeleton reconstruction and dealing with occlusions. This objective was realized by synergizing CRMH-p with a comprehension of the delicacies inherent in ASD therapy. The decision to employ non-intrusive devices, grounded in the need for a natural and free setting for child interactions, led to the adoption of a strategically positioned camera as the essential medium for engagement between the robot and the child. This necessity, while fundamental, introduced complexities, particularly in distant skeleton estimation and occluded scenarios.

A comparative analysis between different skeleton reconstruction models identified the CRMH adaptation, personalized for each participant's height, CRMH-p model, as instrumental in our therapeutic infrastructure. Harmonizing precision with instantaneous feedback, CRMH-p outperformed its peers. However, this study was performed without including scenarios with occlusions in the tests which might influence these findings. Despite this, CRMH-p's efficacy was palpable, integrating seamlessly into the comprehensive system and providing the necessary conditions for an optimal therapeutic approach.

Subsequent efforts centered on the integration of the CRMH-p with the gesture recognition model and comparing it with the Kinect system. The Kinect system had its limitations, especially with occlusions. CRMH-p, on the other hand, showcased versatility and reliability, even when interfacing with the Kinect-based system. The innovative group strategy further enhanced gesture recognition accuracy.

In real-time skeleton reconstruction, we used the CRMH-p model in a Docker environment for consistent and efficient operation, free from hardware limitations. This decision necessitated a communication system between the local operation and the Docker setup. Integrating a sequential processing approach with the RabbitMQ communication method was ideal for our therapeutic context, ensuring smooth live session interactions. This system remained robust even when integrated into the RAT framework, reaching a frame rate of 4.98 FPS.

In summary, this thesis entailed several significant contributions that advanced the current understanding and technical capabilities of the project:

- Co-designing the therapy protocol alongside therapists from the CADIn association, enhancing the interactive engagement tailored for therapy sessions.
- Adapting the CRMH model to extend its performance to individuals of all heights, thereby improving the accuracy of skeleton recognition across diverse participant groups.
- Developing an encoder model to convert the CRMH skeleton structure to the Kinect's format, enabling compatibility with the Gesture Recognition model.
- Improving the gesture recognition accuracy through a group strategy where we associated unused gestures with similar ones that are part of the therapy protocol, therefore, minimizing the dispersity of results.
- Implementing the RabbitMQ communication system with sequential processing which facilitated the real-time operation of the CRMH-p model, notably improving the system's ability to handle occlusions in real-time scenarios.

As we reflect on the advancements achieved through this research, there are certain interesting avenues to follow. A critical frontier for subsequent studies lies in the evaluation of state-of-the-art models in regular occlusion settings usually found in therapy sessions. Authenticating their robustness in real-world therapy requires both recreating these scenarios and implementing the models in real-time alongside other therapy subsystems.

A promising research avenue is creating a gesture recognition model tailored to CRMH-p skeleton structures. The present need to convert CRMH-p skeletons for Kinect compatibility can introduce error propagation. Direct training with CRMH-p may simplify the process and improve accuracy.

Additionally, the current model prioritizes Kinect's RGB data, neglecting depth information due to limited datasets. Creating synthetic datasets with depth information could refine CRMH-p and enhance 3D reconstruction. However, this should not compromise run-time performance.

In conclusion, with upcoming clinical acquisitions, the real-world impact of the contributions made in this thesis is about to be realized. The methodologies and systems developed will now be applied in actual therapeutic settings, providing a tangible measure of their efficacy. It is an exciting point where the theoretical and practical aspects of this research will meet real-world clinical needs, potentially making a meaningful difference in the lives of children with ASD.

REFERENCES

[1] Y. Kim, S. Lee, and S. Cho, "The effects of robot-assisted therapy on children with autism spectrum disorder: A systematic review and meta-analysis," *Research in Developmental Disabilities*, vol. 81, pp. 25–34, 2018.

[2] B. Teke, M. Lanz, J.-K. Kamarainen, and A. Hietanen, "Real-time and robust collaborative robot motion control with microsoft kinect ® v2," 07 2018, pp. 1–6.

[3] B. Silva, "Gaze analysis in robotic therapy for autistic children," Master's thesis, Instituto Superior Técnico, Lisbon, 2021.

[4] C. Silva, "2d and 3d reconstruction of skeletons in robot assisted therapy for autistic children," Master's thesis, Instituto Superior Técnico, Lisbon, 2022.

[5] National Institute of Mental Health, "Autism spectrum disorder," https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd#part_145441, March 2022.

[6] J. J. Diehl, L. M. Schmitt, M. Villano, and C. R. Crowell, "The clinical use of robots for individuals with autism spectrum disorders: A critical review," *Research in Autism Spectrum Disorders*, vol. 6, no. 1, pp. 249–262, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1750946711000894

[7] A. Kouroupa, K. Laws, K. Irvine, S. Mengoni, A. Baird, and S. Sharma, "The use of social robots with children and young people on the autism spectrum: A systematic review and meta-analysis," *PLoS ONE*, vol. 17, no. 6, p. e0269800, 2022.

[8] A. Alabdulkareem, N. Alhakbani, and A. Al-Nafjan, "A systematic review of research on robot-assisted therapy for children with autism," *Sensors*, vol. 22, no. 3, p. 944, 2022.

[9] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, 2020.

[10] M. K. Yeung, "A systematic review and meta-analysis of facial emotion recognition in autism spectrum disorder: The specificity of deficits and the role of task characteristics," *Neuroscience & Biobehavioral Reviews*, vol. 133, p. 104518, 2022.

[11] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, and K. Daniilidis, "Coherent reconstruction of multiple humans from a single image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[12] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, oct 2015. [Online]. Available: https://doi.org/10.1145/2816795.2818013

[13] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei, "Monocular, one-stage, regression of multiple 3d people," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 179–11 188.

[14] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, "Putting people in their place: Monocular regression of 3d people in depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 243–13 252.

[15] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger *et al.*, "Learning an infant body model from rgb-d data for accurate full body motion analysis," in *MICCAI*, 2018, pp. 792–800.

[16] A. Giubergia and A. Ivani, "Guess My Gesture. A gesture recognition algorithm in a robot therapy for ASD children." Ph.D. dissertation, Politecnico di Milano, 2020.

[17] A. M. Aurand, J. S. Dufour, and W. S. Marras, "Accuracy map of an optical motion capture system with 42 or 21 cameras in a large measurement volume," *Journal of Biomechanics*, vol. 58, pp. 237–240, 2017.

[18] M. Moravcik and M. Kontsek, "Overview of docker container orchestration tools," in *2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA)*. IEEE, 2020, pp. 475–480.

[19] A. Ćatović, N. Buzaija, and S. Lemes, "Microservice development using rabbitmq message broker," *Science, Engineering and Technology*, vol. 2, no. 1, pp. 30–37, 2022.