# Evaluating Head Pose Estimation for Assessing Visual Attention in Children with Special Needs During Robot-Assisted Therapy

Rusnani Yahya[1], Rozita Jailani[2,*], Nur Khalidah Zakaria[3], Fazah Akhtar Hanapiah[4]
Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Malaysia [1, 2, 3]
Faculty of Medicine, Universiti Teknologi MARA, Sungai Buloh, Malaysia [4]
Center for Medical Electronic Technology, Politeknik Sultan Salahuddin Abdul Aziz Shah, Shah Alam, Malaysia [1]

***Abstract*—This study investigates the application of head pose estimation (HPE) to assess visual attention in children with special needs (CwSN) during robot-assisted therapy sessions, focusing on its effectiveness and the attention patterns exhibited by these children. CwSN often faces unique challenges, such as sensory processing difficulties or delayed cognitive processing. Age and therapy duration also influenced attention levels, with younger children generally exhibiting shorter attention spans than older participants. Additionally, familiarity with technology, such as prior screen time at home, positively impacted engagement during robot-assisted therapy. An experimental study was conducted with 30 children aged 2 to 7 years, including those with autism spectrum disorder (ASD), speech delay (SD), and attention-deficit/hyperactivity disorder (ADHD). Using an integrated camera, head movements were tracked to analyse forward-facing head direction as an indicator of attention. The system achieved an overall accuracy of 82% and an average attention percentage of 65%, highlighting that visual attention varies significantly based on the type of disability, age, and therapy duration. The integration of the robot enhanced visual engagement across all groups, fostering improved interaction and attention. These findings emphasise the importance of tailoring robot-assisted therapy (RAT) to the specific needs and attention patterns of children with different disabilities, ages, and therapy histories, underscoring the potential of assistive robotics to optimise therapeutic outcomes in special education settings. This research highlights the potential of personalised RAT to improve social, cognitive, and motor skills. It offers evidence-based strategies for integrating assistive robotics into special education and therapeutic settings for CwSN.**

***Keywords*—Head pose estimation; visual attention; robot-assisted therapy; children with special needs**

## I. Introduction

Visual attention assessment is critical for understanding and enhancing therapeutic outcomes for CwSN. These children often exhibit atypical visual attention patterns, which impact their social interactions, communication skills, and learning processes [1] [2]. Accurate assessment of visual attention provides valuable insights into cognitive and behavioural states, enabling therapists to tailor interventions that promote engagement and skill development. Nevertheless, traditional observation-based assessments predominantly remain subjective and demonstrate variability across different sessions, thereby limiting the accuracy and reliability of therapeutic evaluations.

To address these limitations, there is an increasing need for an AI-driven methodology capable of objectively quantifying attention through head pose estimation (HPE). The understanding and supervision of visual attention using such computational techniques are essential for improving engagement in therapeutic and educational settings, in alignment with the global movement towards intelligent and inclusive educational technologies. The HPE, specifically analysing head direction patterns, is significant in assessing attention in CwSN [1]. Monitoring head movements helps therapists understand where a child is focusing their attention during therapy sessions [2], enabling real-time feedback and adaptive strategies that foster improved social and cognitive outcomes [3].

However, quantifying and accurately assessing head direction in CwSN presents significant challenges. These children may struggle to maintain a consistent gaze or display rapid and unpredictable head movements, adding complexity to understanding head direction patterns unique to this group[4]. Gaining more profound insights into these patterns is crucial for tailoring interventions to support their developmental needs [5].

Despite significant advancements in head pose estimation, few studies have utilised these methodologies to evaluate visual attention among children during real-world therapeutic sessions. Current models primarily focus on adult datasets or controlled laboratory environments, thus limiting their applicability in educational and clinical contexts. This research aims to address this limitation by developing and validating a real-time head pose estimation framework designed to assess visual attention within the scope of robot-assisted therapy.

This study evaluates head direction patterns in CwSN during RAT sessions. Using an integrated camera system to track head movements toward a service robot, we aim to identify distinct attention patterns associated with CwSN. The novelty of this work lies in the integration of a hybrid HPE algorithm into real-world therapy sessions, as well as the multi-factor analysis of attention patterns across disability type, age, and therapy duration. These contributions advance the field of RAT by providing real-time, evidence-based insights that can

* Corresponding author.

guide the design of more personalised engagement strategies and improve therapeutic outcomes for CwSN.

The remainder of this study is organised as follows: Section II reviews related work on head pose estimation and visual attention assessment. Section III describes the methodology, including the proposed framework, data collection, and analysis procedures. Sections IV and V presents the results and discussion, while Section VI concludes the study with implications, limitations, and directions for future research.

## II. Literature Review

### A. Head Pose Estimation Techniques

Head pose estimation (HPE) is pivotal in interpreting visual attention and social interactions, especially among CwSN. Various methods have been developed to estimate head pose, broadly categorised into feature-based, appearance-based, and hybrid approaches. Feature-based methods rely on detecting facial landmarks such as the eyes, nose, and mouth to infer head orientation [6]. The algorithms and techniques associated with these methods can be categorised into classical approaches, machine learning-based models, deep learning algorithms [16], and hybrid techniques. Classical algorithms, such as Active Shape Models (ASM) and Active Appearance Models (AAM), focus on statistical shape modelling and texture analysis [7][8].

At the same time, machine learning-based approaches utilise classifiers like Support Vector Machines (SVM) and Random Forests for head pose prediction [9][10][11]. Deep learning algorithms, including Convolutional Neural Networks (CNNs) and heatmap regression methods, leverage advanced architectures to enhance accuracy and robustness in landmark detection and pose estimation [12]. Appearance-based methods employ machine learning techniques to analyse pixel intensity patterns across the face without explicitly detecting facial features. CNNs have been widely adopted in this category due to their robustness in handling variations in lighting and facial expressions. These methods have advanced with the development of deep learning, enabling more accurate and efficient head pose estimation.

Significant advancements in HPE techniques for assessing visual attention in CwSN have been achieved. Deep learning approaches, including CNNs and transformer-based models, have gained prominence due to their ability to handle the complex and variable head movement characteristic of CwSN populations. These models have been optimised for real-time performance, enabling seamless integration into interactive therapy sessions without causing delays or disruptions [13][14].

Hybrid techniques, such as the Perspective-n-Point (PnP) algorithm, integrate 2D facial landmarks with 3D head models for precise orientation estimation [6][10]. These advancements contribute significantly to the field of HPE, with applications spanning robotics, healthcare, and human-computer interaction. While effective, these methods can face challenges with occlusions and variations in facial expressions, which are common in real-world settings.

Researchers have developed non-intrusive methods using robot cameras and AI-based frameworks such as dlib and MediaPipe to capture head movements without requiring children to wear any devices. Techniques that leverage facial landmark detection and 3D modelling have improved the accuracy of HPE, even in uncontrolled environments with varying lighting conditions and occlusions. These advancements are essential for creating responsive therapeutic tools that can adapt to the unique behaviours of CwSN.

In the context of CwSN, real-time and adaptive HPE techniques are crucial. CwSN may exhibit rapid or atypical movements, making robust and flexible estimation methods essential. Recent developments have focused on enhancing these methods to increase accessibility and practicality for use in therapeutic settings, ensuring minimal intrusion while maintaining maximum accuracy.

### B. Visual Attention in Children with Special Needs

Children with special needs (CwSN) often face challenges in maintaining attention, particularly in social contexts, where they may focus less on facial features and eye regions than typically developing peers, instead directing their attention toward objects or patterns. This atypical visual attention can hinder the development of social and communication skills, making it a critical area of interest for researchers. Studies employing eye-tracking technology have revealed that these children spend significantly [15] less time engaging with social stimuli, such as people, and more time attending to non-social elements within their environment. These findings suggest that interventions to redirect their visual attention toward socially relevant stimuli could enhance social skill acquisition [17]. Additionally, difficulties with joint attention and the ability to share a focus on an object or activity with another person are common and significantly impact language development. Targeted strategies to improve joint attention could, therefore, play a vital role in supporting the developmental needs of these children [18].

Recent advancements in research and technology have paved the way for innovative interventions tailored to the unique requirements of CwSN. Visual dysfunction, prevalent in this population, necessitates diagnostic approaches that extend beyond traditional visual acuity assessments. [19]. For example, integrating Augmented Reality (AR) in learning environments has demonstrated the potential to improve visual attention by providing multi-sensory experiences that actively engage children [20]. Additionally, tools such as 3D-printed toys have been shown to improve attention spans and fine motor skills in children with autism [21]. Environmental factors also play a critical role, with studies indicating that exposure to greenery, such as classroom windows overlooking natural landscapes, can positively impact sustained attention levels. Such findings highlight the importance of creating enriched environments and using adaptive tools to foster visual and cognitive engagement in CwSN [22] [23].

Educational strategies and structured interventions further enhance visual attention and related skills. Visual media, such as flashcards, have been shown to improve memory and recognition in autistic children [24]. Art provides an effective medium for visual expression among CwSN [25]. These strategies align with cognitive load theory, which emphasises tailoring instructional methods to manage cognitive demands

and optimise learning outcomes. By understanding how CwSN allocate their visual attention, therapists and educators can design targeted interventions that enhance engagement with social stimuli and promote meaningful improvements in communication, social interaction, and overall learning.

### C. Robot-Assisted Therapy

Robot-assisted therapy (RAT) has emerged as a promising tool for supporting CwSN, particularly enhancing their social, emotional, and cognitive skills. Recent advancements have focused on designing robots to address specific challenges faced by children with autism. For instance, robot functionalities tailored for Applied Behaviour Analysis (ABA) therapy demonstrate the ability to stimulate cognitive skills while adapting to individual limitations [26]. A dual-cycle therapy model emphasising the integration of therapists as teleoperators, which enhances the effectiveness of robot-mediated interactions with neurodivergent children [27].

Social assistive robots have also been shown to facilitate emotional expression and secure social interactions in therapy sessions. These robots provide a judgment-free environment, significantly improving engagement and therapeutic outcomes in children with autism spectrum disorders (ASD) [28]. Furthermore, the use of social humanoid robots as mediators in interventions emphasises their potential to promote communication skills and interaction between children, teachers, and therapists [29]. The importance of advancing robot design lies in better accommodating the needs of CwSN.

In addition to fostering social and emotional growth, robot-assisted therapy is being explored in broader contexts, such as mental health interventions and physical skill development. The potential of robots to engage adolescents in therapeutic settings, with implications for similar applications in CwSN [30]. Integrating interactive elements in robot-mediated therapy has proven effective in enhancing engagement, indicating that these tools hold significant promise for complementing traditional therapy methods across diverse therapeutic contexts.

In summary, previous research has demonstrated promising methodologies for head pose estimation and attention tracking; however, most are confined to controlled environments or adult populations. Furthermore, the integration of these techniques into robot-assisted therapy remains insufficiently explored. To address these limitations, this study presents a hybrid framework combining Dlib and MediaPipe for robust head pose estimation and employs it to assess visual attention among children with special needs.

## III. Methodology

The study involved an in-situ experiment at the Kizzu Kids Rehabilitation and Enrichment Centre, focusing on CwSN. Service robots were integrated into the participants' standard therapy routines in these sessions. Data was collected during these sessions, capturing video recordings, images, and tabular data in CSV files. These multimodal data sources were utilised to investigate visual attention patterns, particularly analysing how the children directed their gaze and engaged with their surroundings during therapy.

The collected data were used to evaluate a real-time system for measuring visual attention. Section III A describes the experimental procedure, outlining the data acquisition process and the specific formats for the captured data. In Section III B, the study introduces system architecture, detailing its components and incorporating a widely adopted HPE algorithm. This algorithm enables the assessment of gaze direction and engagement levels, thereby facilitating an analysis of the children's visual attention during RAT sessions.

### A. Data Gathering from the In-situ Experiment

This section describes an experiment in which therapy sessions were recorded to create a database for machine learning research. These sessions involved the use of service robots to support CwSN, including those diagnosed with ASD, ADHD, and SD. The robot facilitated app-based activities that complemented therapy sessions by providing structured tasks across multiple domains, including communication (AutiSpark), social-emotional (LogicLike), cognitive (Khan Academy Kids), and motor domains (ABC Kids). These apps functioned primarily as mediators, ensuring the children remain engaged with the robot interface. At the same time, the core focus of the study was the integration of the hybrid head pose estimation (HPE) system. During these sessions, the robot's integrated camera and HPE workflow (dlib, MediaPipe, and PnP) simultaneously recorded and analysed head movements to quantify visual attention in real-time.

Thirty CwSN underwent the RAT to train skills across multiple domains, highlighting the potential of integrating technology and robotics into therapeutic intervention.

*1) Participants:* Thirty children diagnosed with ASD, ADHD, and SD were selected who are currently receiving treatment at the Kizzu Kids Rehabilitation and Enrichment Centre (Malaysia), a specialised institution for the rehabilitation of CwSN.

Ethical approval was obtained from the ethics committees of Universiti Teknologi MARA and Kizzu Kids Rehabilitation and Enrichment Centre. All the parents signed consent forms before their children were included in the study. Children were free to leave the experiment at any time and were always supported by a professional educator, other than the researcher.

*2) The robot therapist: Temi 3 service robot:* The robot that led the RAT was the Temi 3 service robot (see Fig. 1), a telepresence robot and a trustworthy autonomous personal assistant focused on high-quality video. By combining artificial intelligence and autonomous navigation, it can recognise and follow people on demand, memorise predefined locations, and navigate effortlessly in various settings, including hotels, restaurants, shops, businesses, educational institutions, healthcare facilities, and more.

This study utilised the default settings and standard equipment of the Temi 3 robot unless otherwise specified. The Temi 3 has multiple advanced cameras designed to enhance its functionalities. It features a 13-megapixel high-resolution camera with autofocus, capable of recording 1080p video at 30 frames per second (FPS), a 120-degree field of view (FOV), a

5-element lens, and a hybrid infrared (IR) filter. Additionally, it includes a wide-angle 13-megapixel camera for remote navigation, offering a 95-degree FOV and 1080p video recording at 30 FPS. Complementing these is a Time-of-Flight (TOF) depth camera, which operates at 30 FPS with a 90-degree FOV and an effective range of up to 5 meters, enabling depth perception essential for navigation and object detection. Among its software features, Temi 3 includes face detection and tracking capabilities, which are employed in *in-situ* experiments to guide the robot toward the child during interactions.


Fig. 1. The Temi 3 Robot. The camera used for recording the child's head pose is the one on top.

*3) Protocol for the In-situ experiments:* To evaluate the children's attention, this study focused on participants who were already enrolled in therapy sessions of varying durations and had been diagnosed by specialists before commencing the program. In this research, pre-existing app-based learning modules were downloaded and integrated into a robot interface to facilitate interaction between the children and the robot during therapy sessions. These modules were carefully selected to align with the therapeutic needs of the children, incorporating tasks adapted to their developmental levels. These tasks are designed to observe how children visually engage with the robot and respond to visual stimuli presented through the robot's interface.

The robot served as a mediator during the therapy sessions, integrating pre-existing apps and a camera system to facilitate interaction and observation. It was included in the children's daily activities and identified through a specific "visual schedule". A visual schedule communicates the sequence of upcoming activities or events using objects, photographs, icons, words, or a combination of these supports. During each session, the children engaged in tasks presented through the robot, with the apps delivering visual stimuli and the camera capturing data to observe and analyse the children's visual attention.

To facilitate interaction, the robot-led sessions were conducted in the same room where the children typically received their therapy, ensuring a familiar and comfortable environment. The robot was positioned in front of the child and initially placed at a distance of at least 0.5 meters. Children were allowed to adjust their seating position to enhance comfort. Each session focused on a single activity to evaluate visual attention, with tasks presented in a randomised sequence to minimise repetitive or predictable patterns. The robot introduced each activity using clear and straightforward verbal instructions and visual prompts to engage the children in the tasks effectively (see Fig. 2).

Fig. 2. An example of a child–robot interaction during the therapeutic session. A therapist was always present nearby to support the child.

A therapist familiar with the children's daily treatment routines was present during the sessions to provide security and support. The therapist offered positive reinforcement through verbal cues, such as "good" or "right", and, in some instances, physical reinforcement, such as a gentle touch or pat. These reinforcement strategies were individualised to meet each child's needs and address their unique behaviours.

Before the main sessions, an introductory session was conducted to familiarise the children with the robot and minimise potential novelty effects. During this preliminary session, the robot was introduced in a non-therapeutic context for approximately ten minutes, allowing the children to acclimate to its presence in a relaxed and informal manner.

*4) Video recording and annotation*

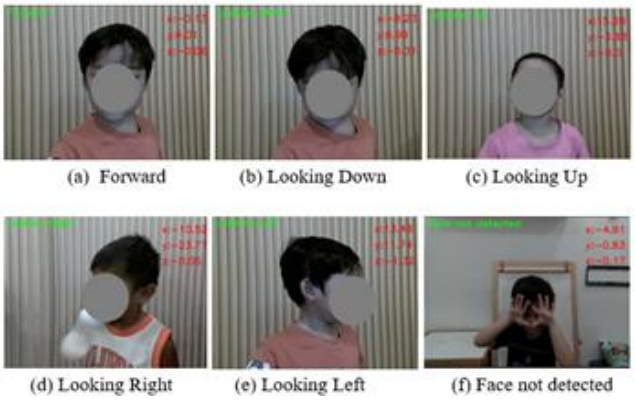


Fig. 3. The frame is extracted from four videos recorded by the robot's camera (the child interacts with the robot, while the system records and annotates the head position in real-time).

Fig. 3 above showcases a series of frames extracted from videos recorded by the Temi 3 robot's camera, capturing various head poses of a child during the interaction. Each frame represents a different head position, annotated in real-time by the system. The poses include: a) Forward, where the child looks directly at the camera; b) Looking Down, where the child tilts their head downward; c) Looking Up, with the child tilting their head upward; d) Looking Right, with the head turned to the right; e) Looking Left, where the child turns their head to the left; and f) Face Not Detected, where the face is obscured or not visible to the system. These annotated frames demonstrate the system's capability to record and categorise head positions during live interactions, essential for analysing visual attention and engagement.

### B. Head Direction Detection Algorithm

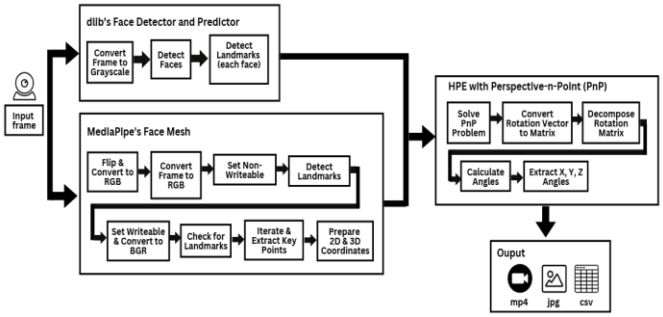


Fig. 4. Comprehensive workflow of the head pose estimation.

Fig. 4 illustrates a comprehensive workflow for HPE, integrating dlib's face detector and MediaPipe Face Mesh as complementary modules for robust facial landmark detection and analysis. The hybrid approach, which combines Dlib's 68-point landmark predictor with MediaPipe Face Mesh, was selected to leverage Dlib's geometric precision alongside the real-time tracking stability provided by MediaPipe. This integration enhances accuracy across diverse lighting conditions and head movements, which are commonly encountered during therapeutic sessions involving children.

To estimate the three-dimensional head orientation, the Perspective-n-Point (PnP) algorithm was employed to assess the three-dimensional head pose due to its efficiency and appropriateness for monocular camera input, thereby ensuring compatibility with standard RGB webcams integrated into the robotic platform.

Input video frames from a camera are preprocessed through two distinct pathways. The first pathway utilises dlib, where frames are converted to grayscale, facial regions are detected, and 68-point facial landmarks are extracted. This pathway ensures accurate detection and tracking of facial features under various lighting and environmental conditions. In the second pathway, MediaPipe's Face Mesh processes the input frames by converting them to RGB format and, if necessary, flipping them for consistency. The frames are then set to a non-writable state to optimise computational performance before detecting facial landmarks. These landmarks undergo iterative refinement to extract key points, which are further processed to generate 2D and 3D coordinate mappings. This dual-pathway approach enhances the system's robustness, enabling it to accommodate diverse input formats while maintaining precision across various scenarios.

The extracted key points from both pathways are input into the HPE module, which employs the Perspective-n-Point (PnP) algorithm for pose computation. This algorithm calculates head orientation by solving for rotation and translation vectors, which are then decomposed to derive X, Y, and Z rotation angles. These angles represent the real-time spatial orientation of the head, enabling precise monitoring and tracking of head movements. The integration of the dlib and MediaPipe modules ensures that the system achieves a balance between computational efficiency and detection accuracy.

Finally, the outputs from the workflow are presented in multiple formats to support diverse analytical needs. Annotated videos are generated in MP4 format, individual frame images are saved as JPG files, and a CSV file is produced to log timestamps, rotation angles, and head directions data. These multimodal outputs provide a comprehensive dataset for analysing head pose dynamics, thereby supporting applications in visual attention assessment and behavioural research. The modular design of this workflow ensures adaptability and effectiveness for real-time analysis in therapeutic environments.

### C. Performance Measures

The forward-facing head pose measure was calculated to evaluate engagement and focus of the participants during therapy sessions, providing a quantitative metric for assessing visual attention. This was achieved by determining the total number of instances where the forward head pose was detected in real-time using head pose estimation algorithms. The forward-facing pose, which indicates that the child is looking directly at the task or stimuli, was analysed explicitly as a key performance measure of attention.

To calculate the frequency of forward-facing head poses as a measure of attention, the following formula was used:

$$\text{Frequency of Forward Head Pose (\%)} = \frac{T}{(T+F)} \text{ x } 100 \quad (1)$$

In this formula, T represents the total number of instances where the system correctly detected the forward head pose and verified its accuracy based on video frame analysis, indicating the participant's direct focus on the task or stimulus. F represents the total number of instances where the system detected a non-forward head pose (left, right, up, or down). Still, upon reviewing the video frames, these instances were verified to be correct forward head poses. This formula provides a normalised percentage of forward-facing head poses, enabling a consistent evaluation of visual attention during the therapy session. This formula provided a normalised percentage of forward-facing poses, enabling a consistent and comparable assessment across participants and sessions.

The forward-facing head pose was considered a reliable indicator of attention, reflecting how participants maintained their focus on the task or robotic intervention during therapy. Higher frequencies of forward-facing poses were interpreted as higher levels of engagement and attention, highlighting the ability of the participant to stay visually connected to the activity.

This analysis offered valuable insights into engagement patterns, emphasising forward-facing head poses as a direct measure of visual attention. Using this metric, the study demonstrated the importance of monitoring forward-facing head poses as a performance measure, particularly for children with special needs. This approach provides an objective and real-time method for evaluating engagement, offering actionable data to optimise therapy strategies and improve outcomes.

## IV. Results

This section presents the experimental outcomes obtained from the head pose estimation framework, including detection accuracy, frame rate performance, and attention classification metrics across participant groups.

### A. Accuracy of Head Direction Detections

The head direction detection algorithm accurately identified and categorised the five predefined head orientations: forward, left, right, up, and down. Validation against the automated annotated dataset showed an overall accuracy of 82% across all participants and sessions. Fig. 5 shows the accuracy percentage for each participant.

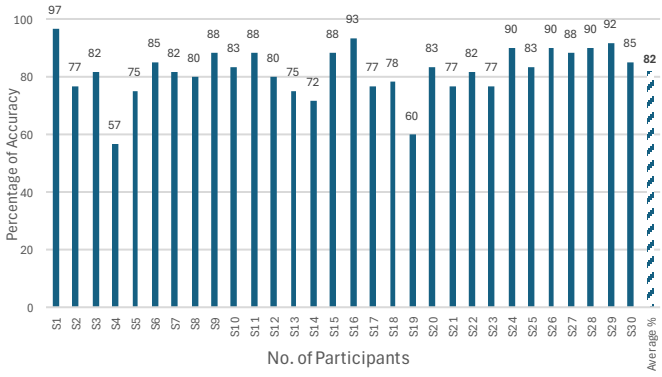


Fig. 5. The overall accuracy percentage for each participant.

Among the 30 participants, a clear contrast in accuracy rates was observed, with S4 and S19 recording the lowest rates and S1 and S16 achieving the highest. S4 achieved 57% accuracy, while S19 recorded 60%. Both participants are relatively inexperienced in therapy, with S4 having attended only a single session and S19 completing four sessions. This limited exposure may have contributed to their lower performance, as they are likely still developing familiarity and engagement with the tasks. Additionally, their young ages, 2 years for S4 and 4 years for S19, may reflect less developed cognitive and attentional abilities than older participants, further influencing their accuracy levels.

In contrast, S1 and S16 demonstrated the highest accuracy rates, with S1 attaining 97% and S16 achieving 93%. Their superior performance can be attributed to their extensive therapeutic exposure and structured intervention programs. S1, a 4-year-old female, has undergone eight one-to-one therapy sessions for personalised guidance and consistent engagement. Similarly, S16, a 6-year-old male, has participated in an intensive Early Intervention Program (EIP), attending sessions five days per week for 24 months. This rigorous and targeted approach likely enhanced their cognitive and attentional abilities, as well as their familiarity with therapeutic tasks, enabling them to outperform their peers.

### B. Percentages of Attention Among Children with Special Needs

The attention percentages of CwSN were analysed to understand their engagement levels during interactions. Focusing on the proportion of time they maintained a forward head pose, a key indicator of attention, the analysis aimed to provide insights into their focus patterns. This measure is particularly significant in assessing how children with special needs respond to stimuli in structured settings, highlighting their ability to stay engaged and attentive over time.

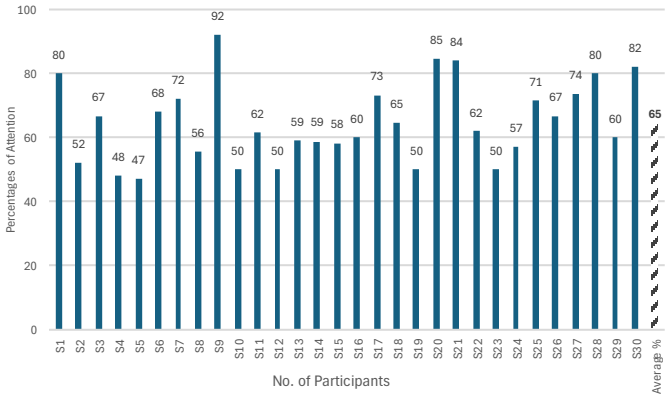


Fig. 6. The overall visual attention percentage for each participant.

Fig. 6 illustrates the attention percentage for each participant during the study, showcasing a wide range of engagement levels. The average attention percentage is 65%, with notable variations at both extremes. These differences highlight individual factors that influence the participant's ability to remain attentive during therapy sessions, including therapy duration, the severity of autism, age, and home environment.

The participant with the highest percentage of attention is S9, achieving an impressive 92%. S9 is a 5-year-old child with mild autism who has undergone 10 months of therapy and regularly experiences screen time at home. These factors may contribute to their higher attention level, as mild autism and longer therapy duration will likely enhance their ability to focus and engage with the task. Additionally, familiarity with screen-based interactions at home could make the experimental setup more engaging and comfortable for them.

In contrast, the participant with the lowest attention percentage is S5, recording just 47%. S5 is a 2-year-old child with ASD who has only undergone 9 months of therapy and does not have screen time at home. These factors might explain their lower attention level, as young ages and limited therapy duration can pose more significant challenges in maintaining focus. Furthermore, the lack of exposure to screens may make the experimental setup unfamiliar or less engaging for S5. These findings emphasise the importance of tailoring interventions to individual needs and considering factors such as therapy duration, different disabilities, and home environment when analysing engagement levels.

The differences in attention percentages highlight the importance of accounting for individual variability when interpreting these results. Participants with higher attention percentages, such as S9, may reflect the effectiveness of the experimental design in fostering engagement, particularly for children who benefit from longer therapy and familiar environments. Conversely, participants with lower attention levels, like S4, point to areas for improvement, such as minimising distractions or adapting activities better to suit the needs of children with more severe challenges. Understanding these patterns provides valuable insights for designing more inclusive and effective interventions, ensuring all participants can maximise their engagement during similar studies or therapeutic settings.

### C. Type of Disability and Attention Pattern Analysis

The analysis of head direction among thirty children with special needs revealed distinct patterns among the three diagnostic groups: Autism Spectrum Disorder (ASD), Sensory Disorder (SD), and Attention Deficit Hyperactivity Disorder (ADHD).

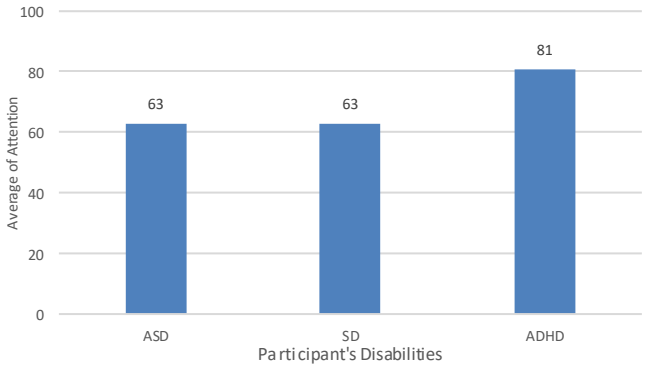


Fig. 7. The average of visual attention for different types of disability.

Fig. 7 illustrates the average percentages of visual attention across three categories of disabilities. The results indicate that children with ADHD (S20, S21, and S27) exhibit the highest average attention percentage at 81%, followed by those with SD at 63% and ASD at 63%. This suggests that the type of disability significantly influences attention levels, with children with ADHD demonstrating comparatively higher focus during activities.

From observations and discussions with a therapist during the therapy sessions, it was noted that children with ADHD tend to exhibit hyperfocus on tasks that are highly stimulating or engaging. These behaviours align with their preference for dynamic and interactive activities that effectively capture their attention. Additionally, shorter, highly interactive sessions are efficient in maintaining focus. These structured and engaging therapy strategies likely contribute to the higher average attention percentages observed in children with ADHD compared to other disability groups. Children with ASD and SD might face different challenges that affect attention, such as sensory processing difficulties or delayed cognitive processing, which could contribute to slightly lower average attention levels compared to children with ADHD. These findings may highlight a strength within the ADHD population with an ability to hyperfocus under the right conditions, which could be leveraged to improve therapeutic and educational outcomes.

### D. Age-Based Variations in Attention Percentages Analysis

Visual attention plays a critical role in understanding the engagement and focus levels of CwSN during therapy sessions. By analysing attention percentages across different age groups, developmental patterns influencing their ability to concentrate on tasks can be identified. Below is a graph (see Fig. 8) illustrating the average attention percentages for children aged 2 to 7, which highlights the relationship between age and visual attention capabilities.

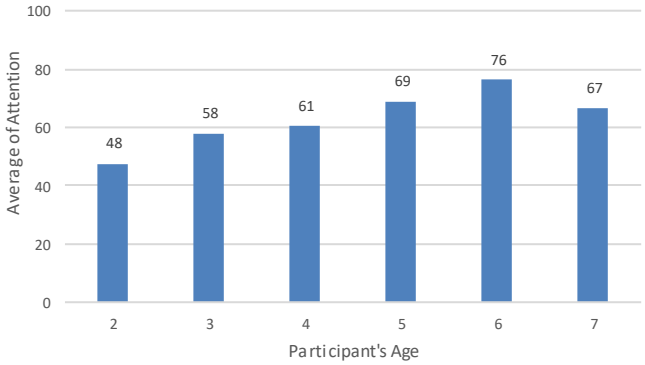


Fig. 8. The average visual attention for different ages.

The results reveal a noticeable trend where attention percentages increase with age, peaking at 76% for 6-year-olds before slightly declining to 67% for 7-year-olds. Younger children, particularly 2-year-olds (S4 and S5), exhibit the lowest attention levels at 48%, indicating developmental differences in focus and engagement during therapy sessions. This pattern suggests that as children grow older, their ability to maintain attention improves, likely due to increased cognitive development and maturity. However, the slight decrease in the oldest age group (7 years) may reflect varying individual differences or a plateau in attention development. These findings demonstrate that attention levels tend to align with developmental maturity as children grow older, highlighting the natural progression of focus and engagement with increasing age.

### E. Impact of Therapy Duration on Attention

Therapy duration plays a significant role in influencing the attention levels of CwSN. Analysing attention percentages across different therapy durations provides valuable insights into how the length of therapy impacts focus and engagement over time. It is important to note that the children in this study were interacting with a robot for the first time, and their unfamiliarity with the robotic environment may have influenced their attention levels. The novelty of the robot could have initially captured their curiosity and engagement.

Fig. 9 illustrates the average percentages of visual attention for CwSN across varying therapy durations, ranging from 6 months to 36 months. The data reveal that children in the 6-month therapy group exhibited the highest attention levels at 67%. However, as therapy duration increased to 12, 24, and 36 months, attention percentages slightly decreased and stabilised

at around 63-64%. This pattern suggests that shorter therapy durations may initially result in heightened attention, potentially due to the novelty effect and intensive engagement during the early stages of therapy. In contrast, stabilising attention levels over longer therapy durations may indicate that children adapt to the therapy routines or experience a plateau in attentional improvement. These findings underscore the need to revise and adjust therapeutic strategies over time to sustain engagement and optimise long-term outcomes.

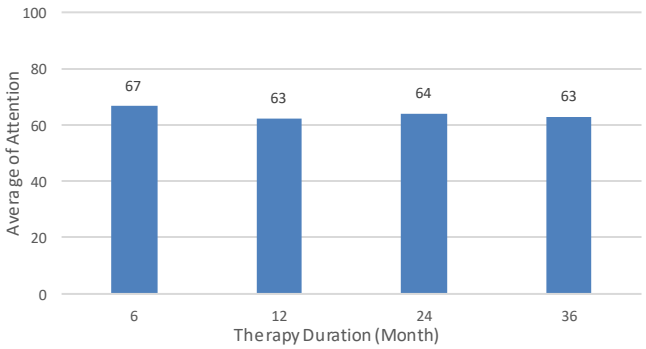


Fig. 9. The percentages of visual attention for different therapy durations.

## V. Discussion

The results indicate that the proposed framework effectively identifies head orientation with high accuracy and stability. This section discusses the implications of these findings, compares them with previous studies, and highlights potential applications as well as existing limitations.

The outcomes of this study demonstrate the feasibility of integrating a head pose estimation algorithm within the context of robot-assisted therapy (RAT) to evaluate visual attention in children with special needs (CwSN). The system achieved an overall head orientation detection accuracy of 82%, alongside an average attention rate of 65%, indicating consistent performance across diverse participant groups and experimental scenarios. These results align with previous studies on real-time head pose estimation, which reported accuracy rates ranging from 78% to 85% under controlled conditions [31][32][33]. This consistency supports the robustness of the hybrid Dlib–MediaPipe–PnP methodology, demonstrating its resilience in the dynamic and unpredictable environment of therapy sessions.

Notably, the findings also reveal considerable variability in the results, suggesting that factors such as age, therapy experience, and type of disability have a substantial impact on engagement levels. This variation highlights the importance of considering developmental and contextual variables when interpreting performance metrics, as children with greater therapy exposure and familiarity with structured activities tend to achieve higher accuracy and attention scores.

Beyond technical validation, the study emphasises the broader implications of integrating robotics into therapeutic practice. RAT not only offers an objective and real-time method for attention monitoring but also facilitates opportunities for tailoring interventions to individual needs. The observed differences across various diagnostic groups and therapy durations indicate that a "one-size-fits-all" strategy may be inadequate, and that customised approaches are imperative to maintain long-term engagement.

These findings contribute to the growing body of evidence supporting artificial intelligence (AI) and robotics as effective tools for enhancing therapeutic precision, inclusivity, and personalisation. By integrating computer vision with behavioural analytics, the proposed framework demonstrates the potential of AI-driven systems to objectively quantify visual attention, thereby providing therapists with actionable data to support informed clinical decision-making. This integration represents a significant step toward data-informed, child-centred therapy environments that foster measurable developmental outcomes and elevate the overall quality of interventions for children with special needs.

Although the proposed framework demonstrates robust performance, several limitations should be acknowledged. The sample size was relatively small, and the controlled experimental setting may not fully reflect the variability and complexity of real-world therapeutic environments. Furthermore, external factors such as inconsistent lighting conditions and occasional occlusion intermittently influenced detection accuracy, suggesting the need for further refinement and validation in more diverse and naturalistic contexts.

Notwithstanding these limitations, the findings hold significant implications for therapeutic practice and special education. By enabling objective monitoring of attention, the framework allows therapists to tailor interventions and track engagement patterns over time, thereby enhancing therapeutic precision and potentially improving developmental outcomes for children with special needs.

In summary, this discussion highlights that the proposed hybrid head pose estimation framework demonstrates strong accuracy, real-time performance, and adaptability across diverse therapeutic contexts. The variations observed among participants underscore the importance of personalised and data-driven therapy strategies, wherein AI-based tools serve to complement rather than replace human expertise. By transforming head orientation data into meaningful attention metrics, this study bridges computational analysis with behavioural interpretation, enabling a more precise understanding of engagement in children with special needs. These insights establish a solid foundation for advancing intelligent robot-assisted systems and inform the development of future research directions in data-driven therapeutic interventions.

## VI. Conclusion

The findings of this study highlight the significance of head direction patterns in assessing visual attention among children with special needs (CwSN), with variations in engagement, such as the high focus of S9 versus the challenges faced by S5, underscoring the influence of disability severity, therapy duration, and age. These results have important implications for robot-assisted therapy (RAT), as head direction data can inform the design of therapeutic activities better aligned with individual needs. Nonetheless, the study is limited by its relatively small sample size, variability in participant profiles,

and occasional inaccuracies in head pose classification. These limitations could be mitigated through larger, multicenter studies, the inclusion of multimodal behavioural measures (e.g., eye tracking, body posture), and further optimisation of the hybrid HPE algorithm to improve robustness in real-world conditions. Overall, the study emphasises the potential of head direction analysis in improving visual attention assessment and optimising RAT, offering a promising direction for enhancing therapeutic outcomes for CwSN.

This study presents a validated head pose estimation framework designed to quantify visual attention in children with special needs during robot-assisted therapy. Theoretically, it advances the understanding of how AI-based visual cues can represent attentional behaviour in non-verbal or minimally responsive participants. Practically, the system provides a non-intrusive solution that enhances engagement assessment within therapeutic and inclusive learning settings. Future research will involve larger participant groups, integration with eye-gaze and emotion detection technologies, and deployment in real therapeutic environments to validate scalability and adaptability.

Overall, the study establishes a solid foundation for the integration of AI-driven attention assessment in therapy, bridging computational modelling with behavioural science. By promoting more personalised, data-informed, and child-centred approaches, this research contributes to the evolving landscape of intelligent robot-assisted therapy, paving the way for future innovations that enhance developmental outcomes for CwSN.

## ACKNOWLEDGMENT

This work was funded by the Strategic Research Partnership (SRP) grant (100-RMC 5/3/SRP INT (047/2023)) and (100-RMC 5/3/SRP INT (048/2023)). The authors would like to thank the Institute of Research Management Centre (RMC) and the Faculty of Engineering, Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia, for the financial support, instrumentation, and experimental facilities provided. This study was undertaken during the study leave of the first author under the 2022 Federal Training (HLP) scholarship scheme awarded by the MOHE Malaysia.

## REFERENCES

[1] M. Chita-Tegmark and M. Scheutz, "Assistive Robots for the Social Management of Health: A Framework for Robot Design and Human–Robot Interaction Research," Int. J. Soc. Robot., vol. 13, no. 2, pp. 197–217, 2021, doi: 10.1007/s12369-020-00634-z.

[2] A. Kouroupa, K. R. Laws, K. Irvine, S. E. 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 June, pp. 1–25, 2022, doi: 10.1371/journal.pone.0269800.

[3] T. Singh, "Attention Span Prediction Using Head-Pose Estimation with Deep Neural Networks," IEEE Access, vol. 9, pp. 142632–142643, 2021, doi: 10.1109/ACCESS.2021.3120098.

[4] Y. F. Kaul, K. Rosander, C. von Hofsten, K. Strand Brodd, G. Holmström, and L. Hellström-Westas, "Visual tracking at 4 months in preterm infants predicts 6.5-year cognition and attention," Pediatr. Res., vol. 92, no. 4, pp. 1082–1089, 2022, doi: 10.1038/s41390-021-01895-8.

[5] F. Yuan, E. Klavon, Z. Liu, R. P. Lopez, and X. Zhao, "A Systematic Review of Robotic Rehabilitation for Cognitive Training," Front. Robot. AI, vol. 8, no. May, pp. 1–24, 2021, doi: 10.3389/frobt.2021.605715.

[6] D. A. Kaliukhovich et al., "Social attention to activities in children and adults with autism spectrum disorder: effects of context and age," Mol. Autism, vol. 11, no. 1, pp. 1–14, 2020, doi: 10.1186/s13229-020-00388-5.

[7] Z. Shi, T. R. Groechel, S. Jain, K. Chima, O. Rudovic, and M. J. Matarić, "Toward Personalized Affect-Aware Socially Assistive Robot Tutors for Long-Term Interventions with Children with Autism," ACM Trans. Human-Robot Interact., vol. 11, no. 4, 2022, doi: 10.1145/3526111.

[8] Q. Wu and L. Zhang, "A Real-Time Multi-Task Learning System for Joint Detection of Face, Facial Landmark and Head Pose," pp. 1–12, 2023, [Online]. Available: http://arxiv.org/abs/2309.11773

[9] N. Van Nam and N. T. N. Quyen, "FLASH: Facial Landmark Detection Using Active Shape Model and Heatmap Regression BT - Industrial Networks and Intelligent Systems," N.-S. Vo and H.-A. Tran, Eds., Cham: Springer Nature Switzerland, 2023, pp. 170–182.

[10] G. Tzimiropoulos, J. Alabort-I-Medina, S. Zafeiriou, and M. Pantic, "Generic active appearance models revisited," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7726 LNCS, no. PART 3, pp. 650–663, 2013, doi: 10.1007/978-3-642-37431-9_50.

[11] H. Su, W. Qi, J. Chen, C. Yang, J. Sandoval, and M. A. Laribi, "Recent advancements in multimodal human–robot interaction," Front. Neurorobot., vol. 17, 2023, doi: 10.3389/fnbot.2023.1084000.

[12] N. Efthymiou et al., "ChildBot: Multi-robot perception and interaction with children," Rob. Auton. Syst., vol. 150, 2022, doi: 10.1016/j.robot.2021.103975.

[13] S. Malek and S. Rossi, "Head pose estimation using facial-landmarks classification for children rehabilitation games," Pattern Recognit. Lett., vol. 152, pp. 406–412, 2021, doi: 10.1016/j.patrec.2021.11.002.

[14] A. Singh, K. Raj, T. Kumar, S. Verma, and A. M. Roy, "Deep Learning-Based Cost-Effective and Responsive Robot for Autism Treatment," Drones, vol. 7, no. 2, pp. 1–18, 2023, doi: 10.3390/drones7020081.

[15] N. Wedasingha, P. Samarasinghe, L. Senevirathna, M. Papandrea, A. Puiatti, and D. Rankin, "Automated anomalous child repetitive head movement identification through transformer networks," Phys. Eng. Sci. Med., vol. 46, no. 4, pp. 1427–1445, 2023, doi: 10.1007/s13246-023-01309-5.

[16] F. M. Talaat, "Real-time facial emotion recognition system among children with autism based on deep learning and IoT," Neural Comput. Appl., vol. 35, no. 17, pp. 12717–12728, 2023, doi: 10.1007/s00521-023-08372-9.

[17] S. Congiu, G. Doneddu, and R. Fadda, "Attention toward Social and Non-Social Stimuli in Preschool Children with Autism Spectrum Disorder: A Paired Preference Eye-Tracking Study," Int. J. Environ. Res. Public Health, vol. 21, no. 4, 2024, doi: 10.3390/ijerph21040421.

[18] C. F. Pinto, H. Mohan, R. Shenoy, V. Guddattu, and S. Tiwari, "The Effect of Parent-Mediated Joint Attention Intervention on Joint Attention and Language Skills in Children with Autism Spectrum Disorder - A Systematic Review," Child Fam. Behav. Ther., vol. 46, no. 3, pp. 272–297, 2024, doi: 10.1080/07317107.2024.2338741.

[19] M. Musleh, A. Green, A. Mankowska, C. Viner, and R. Pilling, "Spectrum of Visual Dysfunction Detected by a Novel Testing Protocol Within a Special School Eye Care Service," Br. Ir. Orthopt. J., vol. 20, no. 1, pp. 219–225, 2024, doi: 10.22599/bioj.391.

[20] G. Nie et al., "An Immersive Computer-Mediated Caregiver-Child Interaction System for Young Children with Autism Spectrum Disorder," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 29, pp. 884–893, 2021, doi: 10.1109/TNSRE.2021.3077480.

[21] N. M. S. -, S. N. A. M. B. -, L. Z. Q. -, R. T. Y. S. -, and N. S. A. R. -, "Fine Motor Training using Conventional Tools, 3D-printed Toys and Digital Platform," Int. J. Multidiscip. Res., vol. 5, no. 6, pp. 1–8, 2023, doi: 10.36948/ijfmr.2023.v05i06.9934.

[22] K. D. C. M. Nushara and A. Hettiarachchi, "Impact of Greenery in the Window View on Visual Sustained Attention with Special Reference to Classrooms of Children with Down Syndrome," J. Real Estate Stud., vol. 20, no. 2, 2023, doi: 10.31357/jres.v20i2.6749.

[23] S. Bektaş and Z. G. Ercan, "A Study of Visual Motor Skills of Children With Special Needs," Eur. J. Educ. Stud., vol. 10, no. 8, pp. 341–359, 2023, doi: 10.46827/ejes.v10i8.4930.

[24] Y. Y. Subaihah, S. Alfinuha, M. Hasanah, and I. Indrawati, "Improved Ability to Recognize Letters Through Flashcard Media With Maximization Memory for Children With Autistic Special Needs," J. Univ. Muhammadiyah Gresik Eng. Soc. Sci. Heal. Int. Conf., vol. 2, no. 1, p. 95, 2023, doi: 10.30587/umgeshic.v2i1.5128.

[25] A. S. Dewi and E. Sugiarto, "Visual Expression of Children with Special Needs in SLB Muhammadyaah Surya Gemilang in the Image of the Picture," Gondang J. Seni dan Budaya, vol. 7, no. 1, p. 127, 2023, doi: 10.24114/gondang.v7i1.38420.

[26] T. S. F. Valenca, E. A. N. Carvalho, E. O. Freire, L. Molina, and J. G. N. De Carvalho Filho, "Specification of Robots for the Treatment of Autistic Children," Proc. - 2023 Lat. Am. Robot. Symp. 2023 Brazilian Symp. Robot. 2023 Work. Robot. Educ. LARS/SBR/WRE 2023, pp. 672–677, 2023, doi: 10.1109/LARS/SBR/WRE59448.2023.10333018.

[27] S. Elbeleidy, T. Mott, D. Liu, E. Do, E. Reddy, and T. Williams, "Beyond the Session: Centering Teleoperators in Socially Assistive Robot-Child Interactions Reveals the Bigger Picture," Proc. ACM Human-Computer Interact., vol. 7, no. CSCW2, 2023, doi: 10.1145/3610175.

[28] Polyxeni Ntaountaki, Georgia Lorentzou, Andriana Lykothanasi, Panagiota Anagnostopoulou, Vasiliki Alexandropoulou, and Agathi Stathopoulou, "Robotics for Autistic Children," Int. J. Sci. Res. Arch., vol. 9, no. 2, pp. 548–559, 2023, doi: 10.30574/ijsra.2023.9.2.0556.

[29] A. Puglisi et al., "Social Humanoid Robots for Children with Autism Spectrum Disorders: A Review of Modalities, Indications, and Pitfalls," Children, vol. 9, no. 7, 2022, doi: 10.3390/children9070953.

[30] P. Alves-Oliveira, T. Budhiraja, S. So, R. Karim, E. Björling, and M. Cakmak, "Robot-mediated interventions for youth mental health," Des. Heal., vol. 6, no. 2, pp. 138–162, May 2022, doi: 10.1080/24735132.2022.2101825.

[31] H. Lee, B. Oh, and S. C. Kim, "Recognition of Forward Head Posture Through 3D Human Pose Estimation With a Graph Convolutional Network: Development and Feasibility Study," JMIR Form. Res., vol. 8, pp. 1–12, 2024, doi: 10.2196/55476.

[32] S. Zhou, W. Zhang, Y. Liu, X. Chen, and H. Liu, "Real-Time Driver Attention Detection in Complex Driving Environments via Binocular Depth Compensation and Multi-Source Temporal Bidirectional Long Short-Term Memory Network," Sensors, vol. 25, no. 17, pp. 1–19, 2025, doi: 10.3390/s25175548.

[33] M. C. Ye and J. J. Ding, "Real-Time Head Orientation and Eye-Tracking Algorithm Using Adaptive Feature Extraction and Refinement Mechanisms †," *Eng. Proc.*, vol. 92, no. 1, 2025, doi: 10.3390/engproc2025092043.