Strong Eye Closure Detection in Children with Profound Intellectual and Multiple Disabilities Using Robust Temporal Difference Features

Kaito Kosaki*, Teppei Nakano*, Mari Wakabayashi†, Tomomi Sato† and Tetsuji Ogawa*
* Waseda University, Tokyo, Japan
E-mail: kosaki@pcl.cs.waseda.ac.jp
† Yokohama City University, Yokohama, Japan

***Abstract*—In this study, we propose a model for detecting strong eye closure from sequences of periocular images. This motion is characterized by a temporal pattern in which the eyelids gradually close over time. While Transformers are well suited for modeling such temporal patterns, they primarily capture global dependencies across an entire sequence and may overlook subtle local changes. In addition, slight misalignments caused by camera shake, subject movement, or environmental variations across frames can make these fine-grained patterns even more difficult to detect. To address these issues, we aim to extract and exploit fine-grained local temporal variations in the periocular region while maintaining robustness to environmental changes and misalignment. Specifically, the proposed approach combines detailed, semantically meaningful feature extraction using GC^{2}SA-Net, fine-grained inter-frame difference extraction via a token-exchange-based difference evaluation (TEDE) module, and Transformer-based sequence modeling. This design complements the Transformer's strength in modeling long-range temporal context by enhancing its sensitivity to subtle local changes and improving robustness to environmental variation and frame misalignment. Experiments on care-scene video data of a child with profound intellectual and multiple disabilities, who expresses discomfort or pain through strong eye closure, demonstrate that incorporating TEDE-based difference information significantly improves detection accuracy for strong eye closure.**

I. Introduction

Facial expressions are critical indicators of a person's internal state and serve as an essential channel for nonverbal communication in diverse contexts. Among these, *strong eye closure* is recognized as an important facial cue that can convey emotional states or physical discomfort, giving it broad practical significance. For example, strong eye closure is known to occur as an involuntary reaction to acute pain or distress [1] and has been reported as a facial expression associated with severe pain during events such as myocardial infarction or acute aortic dissection [2], suggesting its potential as an objective indicator for detecting intense pain or sudden physiological abnormalities.

In driver monitoring systems, real-time detection of such abnormal reactions could help prevent accidents by enabling timely alerts or emergency interventions when signs of physical distress are observed. For children with profound intellectual and multiple disabilities (PIMD), who face substantial barriers to verbal communication, strong eye closure can, in some cases, serve as an important nonverbal cue to express discomfort or pain [3]–[5]. Accurately detecting this subtle signal is therefore critical for ensuring appropriate care.

This study focuses on developing a framework for detecting strong eye closure, with the primary aim of supporting communication for children with PIMD. In this context, these children rely heavily on nonverbal cues, such as facial expressions and gestures, to convey their emotional and physical states. However, the ways in which such nonverbal signals are expressed can vary greatly from child to child, making it challenging for caregivers other than the child's parents to interpret them accurately [6]. Automatic detection and interpretation of these cues could help third-party caregivers provide more appropriate care while also reducing the burden on primary caregivers [7]. Moreover, strong eye closure itself is a subtle periocular movement, and compared to typically developing individuals, children with severe disabilities often display facial expressions that are even more subtle and shorter in duration [8]. These factors make such expressions particularly difficult for conventional facial expression recognition models to detect reliably. As a result, accurately detecting strong eye closure in this population remains a significant technical and societal challenge.

To tackle this challenge with machine learning, we identify three essential requirements:

- **Requirement 1: Temporal dynamics modeling** — Strong eye closure is not a static state but a dynamic process in which the eyes gradually close. Detecting it requires capturing both the magnitude and the temporal pattern of these visual changes across frames.
- **Requirement 2: High spatial resolution for feature extraction** — Unlike ordinary eye closure, strong eye closure involves subtle local changes, such as eyelid tension and glabellar contraction, which appear as fine-grained pixel variations. High-resolution and high-sensitivity visual feature extraction is therefore required.
- **Requirement 3: Robustness to input variations** — In real-world scenarios, head movements, changes in lighting, or face detection inaccuracies can introduce noise or misalignment in the input images. The model must remain robust under such conditions.

979-8-3315-7206-8/25/$31.00

The Transformer architecture [9]–[13] is well-suited for modeling long-term (global) temporal context and provides robustness against environmental variations and noise [14], [15], thus addressing Requirements 1 and 3. However, Transformers do not explicitly retain fine-grained local inter-frame differences, information about *which regions change and by how much*. This structural limitation suggests that, to reliably detect our target behavior, strong eye closure, an additional mechanism capable of capturing subtle, localized motion is required.

To address this issue, we focus on modeling fine-grained local temporal variations in the periocular region by explicitly combining detailed periocular feature extraction with inter-frame difference estimation. However, temporal difference features are inherently sensitive to minor misalignments caused by camera shake, head movement, or detection errors, which can introduce noise and obscure meaningful changes. To mitigate this, we employ feature representations that embed detailed states of periocular components, rather than relying solely on the raw spatial configuration of local structures. For this purpose, we utilize GC^{2}SA-Net [16], which is designed to capture subtle periocular cues and encode semantically and functionally meaningful information. Subtle temporal variations are then extracted using the token-exchange-based difference evaluation (TEDE) module [17], which enhances difference extraction by emphasizing inconsistencies between changing regions and their surrounding context. TEDE has demonstrated strong performance in remote sensing change detection tasks that require the detection of subtle differences despite significant variations in viewpoint, position, or illumination. By effectively integrating these two components, our approach is well-suited for robust detection of fine periocular movements, such as strong eye closure, thereby satisfying Requirements 1, 2, and 3.

In this study, we validate the proposed approach using real-world care-scene video data of a child with PIMD, demonstrating the effectiveness of integrating difference information through the TEDE module. We expect that this sequence modeling approach, designed to meet the three identified requirements, will not only advance precise recognition of subtle facial signals for applications such as emotion or intent estimation in children with PIMD but also offer broader implications for difference-enhanced sequence modeling.

The remainder of this paper is organized as follows. Section II introduces the proposed strong eye closure detection model. Section III presents experimental comparisons using real video data to validate the model's effectiveness. Finally, Section IV concludes the paper.

II. Modeling for Detecting Strong Eye Closure

While typical eye opening and closing can be detected using temporal information from sequences of eye-region images, reliably identifying the specific sign of "strong eye closure" requires capturing subtle, fine-grained changes in the periocular region.

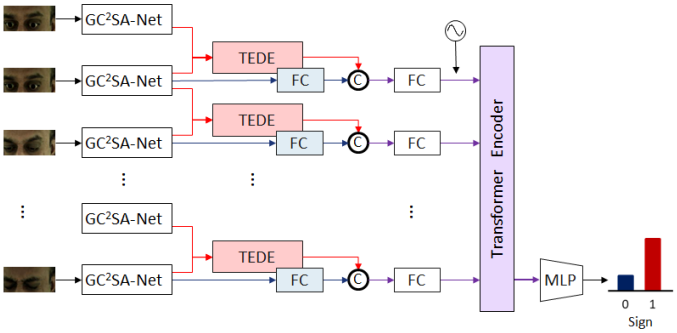

Fig. 1. Architecture of proposed strong eye closure sign detector. Periocular images are first processed by GC^{2}SA-Net-based feature extractor to obtain periocular feature embeddings. Inter-frame difference features are then extracted using TEDE module and concatenated with corresponding static feature embeddings. Resulting vectors are fed into Transformer encoder with positional encoding to capture temporal patterns associated with target sign. "FC" denotes fully-connected layer, and Ⓒ indicates concatenation operation.

To address this challenge, we propose an integrated approach that combines the following three technologies:

1) GC^{2}SA-Net, which is designed to capture and embed the states of periocular components, such as eye openness, gaze direction, periocular wrinkles, and eyebrow shape, rather than their spatial configurations, thereby enabling the extraction of semantically meaningful and discriminative features;
2) TEDE, which accurately extracts inter-frame differences in these features to capture fine-grained temporal variations in periocular states; and
3) a Transformer module, which models the temporal dynamics of the target behavior in an end-to-end fashion.

This section provides an overview of the proposed system architecture (Section II-A) and details how the framework incorporates GC^{2}SA-Net (Section II-B) and TEDE (Section II-C) as its core components.

A. System Overview

Figure 1 illustrates the architecture of the proposed model for detecting strong eye closure. In this framework, the eye and eyebrow regions are first detected in each frame and processed by GC^{2}SA-Net to extract periocular feature embeddings. Next, TEDE is applied to the feature representations from the current and previous frames to explicitly capture inter-frame differences. The resulting TEDE-based difference embeddings are then concatenated with the corresponding static per-frame embeddings to form a sequence of feature vectors, which is fed into a Transformer encoder. Finally, the Transformer's output at the last time step is passed through a fully connected (FC) network to determine whether the sequence contains the sign of strong eye closure.

B. Periocular Feature Extraction

To detect subtle facial changes associated with eye states, we employ GC^{2}SA-Net [16] as the periocular feature extractor. Although GC^{2}SA-Net was originally developed to focus on

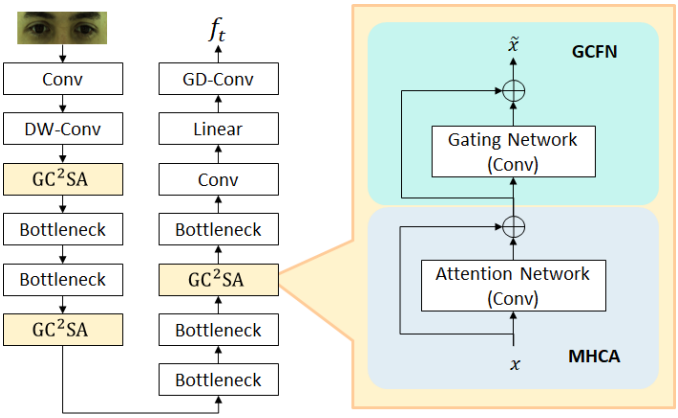

Fig. 2. Schematic of GC^{2}SA-Net architecture, which functions as periocular feature extractor, and detailed view of GC^{2}SA block within it. Attention mechanism in MHCA captures global dependencies, while gating structure in GCFN preserves overall features and enables extraction of finer details.

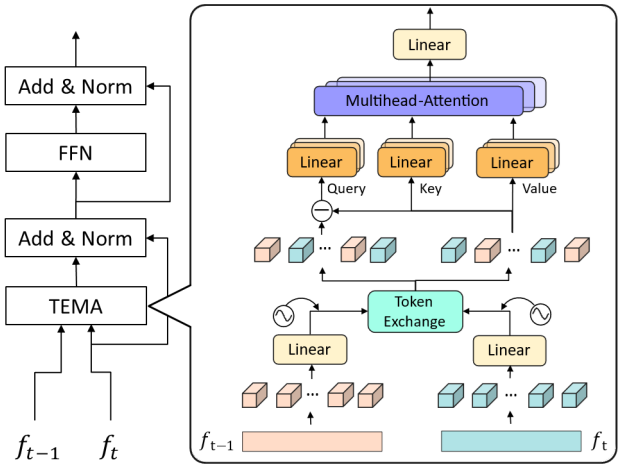

Fig. 3. Schematic of TEDE model used to obtain inter-frame difference features, along with detailed view of TEMA module within TEDE. In TEMA, features are partitioned into patches to generate tokens, which are exchanged, and difference information is extracted using attention mechanism.

iris-based information for discriminative representation learning in personal identification tasks, it is also well suited for capturing fine-grained local features across the entire periocular region.

The architecture of the GC^{2}SA-Net module used in our framework is illustrated in Fig. 2. The network extends MobileFaceNet [18] by incorporating the GC^{2}SA module, which comprises a multi-head channel-wise attention (MHCA) mechanism and a gated convolutional feedforward network (GCFN).

MHCA integrates channel-wise self-attention with depth-wise convolution to simultaneously capture local spatial dependencies and global inter-channel relationships. This design enables the model to generate semantically meaningful feature representations, specifically, the states of periocular components rather than their mere spatial arrangement, while preserving fine-grained periocular details, resulting in robust and discriminative embeddings. GCFN further refines these representations through dynamic gating with GELU activation, which suppresses irrelevant information and highlights salient patterns. The outputs of MHCA and GCFN are then fused via element-wise addition, enhancing the expressive capacity of the extracted features while preserving their essential characteristics. Consequently, the resulting periocular features are expected to remain robust to environmental variations while effectively capturing subtle changes. Given the demands of our task, which requires high sensitivity to fine-scale periocular dynamics and resilience to input variations, GC^{2}SA-Net is particularly well suited for this application.

GC^{2}SA-Net was pre-trained on a large-scale dataset containing 166737 periocular and facial images from 1054 individuals, demonstrating strong generalization capabilities. However, while the original model was designed for personal identification using periocular information, it is not explicitly optimized for detecting visual cues specific to strong eye closure, such as the degree of eyelid closure, eyebrow positioning, and fine-grained texture variations like wrinkles or skin deformation from muscle contraction. To address this, we fine-tune the pre-trained model to better capture these task-specific features.

C. Extraction of Temporal Difference Features

To effectively capture subtle temporal variations in the eye region associated with eye movements, we adopt TEDE, a token-exchange-based differential feature extraction technique originally proposed in DiFormer [17]. DiFormer was developed for change detection in remote sensing imagery, where it enhances the discriminability of changes by emphasizing both the altered regions and the inconsistencies with their surroundings across images captured at different time points.

The original TEDE mechanism works by dividing spatial feature maps extracted from input images into patches, converting these into token sequences, and partially exchanging corresponding tokens between two sequences derived from different frames. This token exchange process trains the attention mechanism to highlight meaningful differences caused by token mismatches, while maintaining consistency in unchanged areas, thereby suppressing irrelevant variations.

In this study, we adapt this TEDE mechanism to operate on sequences of periocular feature vectors. The overall architecture of the TEDE module employed in our framework is illustrated in Fig. 3. Specifically, each 512-dimensional periocular feature vector extracted by GC^{2}SA-Net for each frame is divided into 32 tokens of 16 dimensions. These tokens are subsequently subjected to linear transformation and positional encoding. To emphasize temporal differences, corresponding tokens between consecutive frames are then partially exchanged.

Unlike conventional approaches that define tokens based on the spatial configuration of eye-region components (e.g., eyes and eyebrows), our method constructs token sequences from semantically meaningful features derived from GC^{2}SA-Net. This design enables the model to robustly extract inter-frame

differential features, even in the presence of positional misalignments across frames.

Following the original TEDE strategy, every even-indexed token is replaced with the corresponding token from the adjacent frame. This operation strengthens the representation of changing regions and guides the multi-head attention mechanism to focus on these variations. Attention is then computed using the exchanged tokens as queries, while the original tokens serve as keys and values, allowing the model to effectively capture fine-grained temporal dynamics.

By explicitly emphasizing temporally dynamic semantic features, TEDE is particularly well suited for detecting subtle facial movements, such as strong eye closure.

III. Sign Detection Experiments

To evaluate the effectiveness of the proposed strong eye closure sign detection model shown in Fig. 1, we conducted comparative experiments using video recordings of caregiving scenes involving a child with PIMD. The primary objective of this experiment is to investigate the impact of incorporating differential features, with a particular focus on the effectiveness of the TEDE module compared to a simpler frame-level differencing approach. For this purpose, we compared the following three methods:

1) **No Difference Features**: A baseline model that inputs only the periocular features extracted from each frame into the Transformer encoder, without any temporal difference information.
2) **Delta Difference Model**: A model that inputs the periocular features concatenated with simple difference features into the Transformer encoder. In this model, the difference features are computed as the direct frame-to-frame difference between periocular features.
3) **TEDE Difference Model (Proposed)**: A model that inputs the periocular features concatenated with difference features extracted using the TEDE module into the Transformer encoder. The TEDE module extracts semantically and locally meaningful difference information, which is expected to improve robustness to noise or misalignment.

It is important to note that the detection system was trained exclusively on periocular motion data extracted or synthesized from facial video datasets of typically developing individuals; no data from the target child was used for training. The remainder of this section describes the synthesis of periocular motion data (Section III-A), the experimental setup (Section III-B), and the experimental results (Section III-C).

A. Periocular Motion Data Synthesis for Data Augmentation

1) Dataset: To train the model for detecting strong eye closure signs, we used the Multiface dataset [19], which contains facial images of twelve typically developing individuals captured from multiple viewpoints. This dataset includes various sequence categories representing periocular expressions such as "closed eyes" and "strong eye closure." However, the range of periocular motion patterns captured in Multiface is somewhat limited.

2) Frame-Level Annotation of Expression Classes: To expand the diversity of periocular motion patterns, particularly the target pattern of strong eye closure, we synthesized periocular videos using images and video segments from the Multiface dataset, guided by probabilistic class transition rules. To generate these pseudo videos, we manually annotated the source data (i.e., Multiface) with the following 14 frame-level periocular expression labels:

- **Seven States**: 1) eyes open with raised eyebrows, 2) eyes open with lowered eyebrows, 3) eyes open, 4) squinting, 5) eyes closed, 6) moderate strong eye closure, 7) intense strong eye closure
- **Six Transitions**: 8) eyes open with raised eyebrows $\leftrightarrow$ eyes open, 9) eyes open with lowered eyebrows $\leftrightarrow$ eyes open, 10) eyes open $\leftrightarrow$ squinting, 11) squinting $\leftrightarrow$ eyes closed, 12) eyes closed $\leftrightarrow$ moderate strong eye closure, 13) moderate strong eye closure $\leftrightarrow$ intense strong eye closure
- **One Action**: 14) eyes open $\rightarrow$ blink $\rightarrow$ eyes open

This additional labeling step was necessary because the original Multiface dataset provides periocular expression labels only at the sequence level; detailed frame-level annotations are not available (for example, frames showing open eyes may appear within a sequence labeled "eyes closed").

3) Class Transition Probability: In this process, the next class C_j following a given class C_i is sequentially selected from among 14 classes, including states, transitions, and an action, based on the class transition probability $P(C_j|C_i)$. The transition probabilities were manually defined and adjusted to reflect the intended frequency of occurrence. Transitions that were physically implausible or discontinuous were explicitly prohibited by assigning them a probability of zero. To prevent the model from remaining in a single class for an excessively long duration, the self-transition probability $P(C_i|C_i)$ was dynamically decreased over time using a Gaussian decay function, thereby encouraging smooth transitions to other classes.

4) Periocular Video Synthesis: After generating the label sequence, pseudo-video data with temporal consistency was created by mapping dataset frames to each label. Specifically, for the state classes, multiple frames corresponding to the assigned label were randomly selected and arranged consecutively. As a result, each frame sequence contained slight variations in facial expressions or viewpoints. For the transitions and the action class, a continuous sequence of frames matching the relevant label was randomly selected and arranged in temporal order, either forward or backward. In total, we generated between 590 and 1071 pseudo frames per individual, thereby augmenting the training data and addressing the limited diversity of class transitions in the original dataset.

B. Experimental Setups

1) Input to Model: The input to the sign detection model consisted of sequences of 20 consecutive frames, with the

window size fixed at 20 and a shift size of five used during training. To improve the model's generalization performance, we applied on-the-fly data augmentation, including random rotations and brightness adjustments, during training. The Adam optimizer was used for optimization, with a batch size of eight.

2) Training: For training, we used the Multiface dataset along with the synthetic video data generated from it. Sequences representing states and transitions associated with "moderate strong eye closure" and "intense strong eye closure" (e.g., "eyes elosed ↔ moderate strong eye closure," " moderate strong eye closure ↔ intense strong eye closure") were treated as positive examples, while all other states and transitions were treated as negative examples. The resulting training set comprised 17408 positive sequences and 34920 negative sequences, all derived from multi-directional recordings of twelve typically developing individuals.

3) Evaluation: For evaluation, we used video recordings of the target child with PIMD who was not included in the training data. Data collection was conducted with approval from the child's parents and the university's ethics committee[1]. Each frame of the collected video was annotated to indicate whether the eyes were strongly closed. In this evaluation, instances of strong eye closure were treated as positive examples and all other instances as negative. The evaluation dataset consisted of 140 positive sequences and 14844 negative sequences, with the shift size during evaluation set to one. For all frames, periocular regions were extracted based on facial landmarks detected using Face-alignment [20], horizontally aligned, and then fed into the sign detection model.

4) Developed System: The experiment employed a hierarchical detection framework combining lightweight preprocessing with precise classification. First, for each frame, we extracted 52 BlendShape values [21] using the Face Landmarker in MediaPipe [22]. Among these, two scores related to eye openness (eyeBlinkLeft and eyeBlinkRight) were used to estimate whether the eyes were open. If, for a given sequence, the average of both eyeBlinkLeft and eyeBlinkRight scores across all frames was below 0.5, the sequence was classified as "eyes open" and designated as a negative example, skipping the detailed detection step. Sequences that did not meet this condition were passed to the strong eye closure sign classification model shown in Fig. 1 for fine-grained classification.

5) Evaluation Criteria: Final detection performance was evaluated by comparing the model predictions with the annotated ground truth labels, taking into account both the sequences filtered out as negative by the BlendShape-based pre-processing and those classified (as positive or negative) by the proposed method. Evaluation metrics included precision, recall, F1 score and balanced accuracy, along with an error reduction rate to quantify the effectiveness of false positive suppression.

[1]Approved by Yokohama City University's Ethical Review Committee for Life Science and Medical Research Involving Human Subjects (Approval No. F22080013-Med).

TABLE I
ESTIMATION RESULTS FOR STRONG EYE CLOSURE SIGN ON VIDEO DATA OF TARGET CHILD WITH PIMD USING THREE SIGN DETECTION MODELS. P/R/F INDICATE PRECISION, RECALL, AND F1. BA INDICATES BALANCED ACCURACY. ERR INDICATES ERROR REDUCTION RATE.

Models	P↑	R↑	F↑	BA↑	ERR↑
No Difference Features	0.62	0.20	0.30	0.60	-
Delta Difference Model	0.86	0.31	0.46	0.66	20.2
TEDE Difference Model	0.66	0.46	0.54	0.73	31.8

C. Experimental Results

Table I summarizes the estimation results for the strong eye closure sign using the dataset of the target child with PIMD.

First, the baseline model (**No Difference Features**), which does not explicitly utilize differential features, showed a limited response to the transient and subtle nature of the "strong eye closure" sign, resulting in a significantly low recall. This result indicates that relying solely on static features makes it challenging to detect signs involving fine-grained temporal variations, even when using a Transformer to model the temporal dynamics.

The **Delta Difference Model**, which incorporates simple motion-based information, reduced false detections and thus improved precision. However, its recall for detecting sign occurrences remained low, with many instances still missed. This suggests that simple frame-to-frame differences do not provide sufficient representational capacity to fully capture the emergence of subtle signs.

By contrast, the proposed **TEDE Difference Model** achieved a substantial improvement in recall and a corresponding increase in F1 score, striking the best overall balance between precision and recall while significantly reducing missed detections. It further attained the highest balanced accuracy, indicating that class-wise recall improved on average rather than being limited to the minority class. This performance can be attributed to the model's ability to emphasize local and structural changes in eye and eyebrow movements, allowing it to capture not only whether motion is present but also how it occurs. Moreover, the **TEDE Difference Model** yielded the highest error reduction rate, greatly lowering the total number of false positives and false negatives compared to the baseline. This demonstrates that the proposed method maintains stable and reliable performance in practical scenarios, without compromising either precision or recall.

In summary, the proposed approach based on structural difference extraction effectively models temporal and visual variations, making it well suited for detecting transient and non-stationary behaviors such as the "strong eye closure" sign.

IV. CONCLUSION

In this study, we proposed a model for detecting the subtle motion of strong eye closure from sequences of periocular images and demonstrated the effectiveness of incorporating temporal difference information using care-scene video data of

a child with PIMD. To address the high degree of individual variation in facial expressions, we augmented the facial video dataset to enhance diversity. We also designed a periocular feature extractor that preserves alignment of the periocular region while effectively capturing fine-grained temporal differences, and integrated this with a sequence modeling framework. Experimental results confirmed that explicitly extracting subtle visual changes with the TEDE module and incorporating them into Transformer-based temporal pattern recognition significantly improves detection performance.

V. Acknowledgment

This study was supported by JSPS Scientific Research Grant JP20K10860. Also, we would like to express our sincere gratitude to the child and their parents for their invaluable cooperation.

References

[1] M. Kunz, D. Meixner, and S. Lautenbacher, "Facial muscle movements encoding pain – a systematic review," *PAIN*, vol. 160, p. 1, Oct. 2018.

[2] K. Kimura *et al.*, "JCS 2018 guideline on diagnosis and treatment of acute coronary syndrome," *Circulation Journal*, vol. 83, no. 5, pp. 1085–1196, 2019.

[3] V. R. D. M. Herbuela *et al.*, "Children with PIMD/SMID expressive behaviors: Development and testing of ChildSIDE app, the first step for independent communication and mobility," 2020. arXiv: 2009.00260.

[4] V. R. D. M. Herbuela *et al.*, "Integrating behavior of children with profound intellectual, multiple, or severe motor disabilities with location and environment data sensors for independent communication and mobility: App development and pilot testing," *JMIR Rehabil Assist Technol*, vol. 8, no. 2, e28020, Jun. 2021.

[5] M. Roemer, E. Verheul, and F. Velthausz, "Identifying perception behaviours in people with profound intellectual and multiple disabilities," *Journal of Applied Research in Intellectual Disabilities*, vol. 31, no. 5, pp. 820–832, 2018.

[6] T. Sato, "Creation of care through communication by nurses, welfare workers, and persons (children) with profound intellectual multiple disabilities at a day care center: Emancipation from the japanese "shame culture" ," *Advances in Nursing Science*, vol. 45, no. 2, E69–E93, 2022.

[7] K. Mochida *et al.*, "Exploring robust and explainable design for facial expression-based emotional state estimation in children with profound intellectual multiple disabilities," in *32nd European Signal Processing Conference*, 2024, pp. 481–485.

[8] E. Yokozeki *et al.*, "Differences in facial muscle movements affected by respiratory status in children with severe motor and intellectual disabilities during sputum suction," *Annual Bulletin of the Research Institute of Interdisciplinary Research, Shikoku University*, vol. 4, pp. 93–100, 2023.

[9] N. Mubashira and A. James, "Transformer network for video to text translation," in *2020 International Conference on Power, Instrumentation, Control and Computing (PICC)*, 2020, pp. 1–6.

[10] R. Karthik, S. Adithya, P. Shalmiya, and V. Subramaniyaswamy, "Video anomaly detection using factorized self-attention transformer," in *2024 International Conference on Computational Intelligence and Network Systems*, 2024, pp. 1–6.

[11] A. Arnab *et al.*, "ViViT: A video vision Transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Oct. 2021, pp. 6836–6846.

[12] A. Lakkapragada *et al.*, "The classification of abnormal hand movement to aid in autism detection: Machine learning study," *JMIR Biomed Eng*, vol. 7, no. 1, e33771, Jun. 2022.

[13] X. Ma *et al.*, "Latte: Latent diffusion transformer for video generation," *Transactions on Machine Learning Research*, 2025.

[14] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[15] Y.-L. Hsieh *et al.*, "On the robustness of self-attentive models," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Association for Computational Linguistics, Jul. 2019, pp. 1520–1529.

[16] T. Ng, J. Chai, C. Low, and A. Beng Jin Teoh, "Self-attentive contrastive learning for conditioned periocular and face biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3251–3264, 2024.

[17] H. Lin, R. Hang, S. Wang, and Q. Liu, "Diformer: A difference transformer network for remote sensing change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[18] C. Sheng, L. Yang, G. Xiang, and H. Zhen, *Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices*, 2018.

[19] C. Wuu *et al.*, "Multiface: A dataset for neural face rendering," in *arXiv*, 2022.

[20] B. Adrian and T. Georgios, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.

[21] J. P. Lewis *et al.*, "Practice and Theory of Blendshape Facial Models," in *Eurographics 2014 - State of the Art Reports*, S. Lefebvre and M. Spagnuolo, Eds., The Eurographics Association, 2014.

[22] *Face landmark detection guide — google ai edge — google for developers*, (Accessed on 06/28/2025). [Online]. Available: https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker.