Spinal Disease Classification Using Deep Learning on Dual-View Videos

Tsuguhiro Tsukumo[1], Kaixu Chen[1], Tomoyuki Asada[2,3], Kousei Miura[2], Hideki Kadone[2,4], Kotaro Sakashita[2], Takahiro Sunami[2], Masashi Yamazaki[2], Naoto Ienaga[5], Yoshihiro Kuroda[5]

***Abstract*— Adult spinal deformity (ASD) is characterized by spinal deformities that cause symptoms such as balance disorders. Although gait observation is used for the diagnosis of ASD, no quantitative evaluation protocol has been established. In this study, we propose a novel approach for distinguishing ASD from non-ASD using a neural network with gait videos captured from two perspectives, enabling the detection of unique gait fluctuations specific to each patient. By integrating spatiotemporal gait dynamics, the proposed method enhances quantitative diagnostics and provides a more comprehensive gait assessment. The experimental results indicate that, relative to a model employing a singular perspective, there was a modest enhancement in classification accuracy accompanied by a notable improvement in the F1 score. Specifically, the F1 score increased by 10% over the single-view model, achieving an F1 score of 71.86%. These results confirm the efficacy of integrating two-perspective videos for gait analysis in ASD and suggest that further investigation into fusion techniques may be beneficial. Codes and models are available at `https://github.com/TsuguTsukumo/2stream_3D_CNN_Walk_Pytorch.git` .**

I. INTRODUCTION

Adult Spinal Deformity (ASD) is a condition characterized by abnormal kyphosis or scoliosis of the spine. These deformities can cause balance disturbances and gait impairments, and in severe cases, they pose a risk of multiple complications. Effective treatment of ASD requires early detection and timely intervention. Traditional diagnostic methods primarily rely on X-ray imaging, which is effective for patients whose symptoms are pronounced in a static posture. However, some cases of ASD present mild symptoms at rest, but worsen during gait. As a result, diagnosing ASD based solely on static images is challenging. In practice, physicians often observe the patient's gait and integrate self-reported symptoms for a comprehensive diagnosis. However, this diagnostic process depends heavily on the physician's experience, and no standardized quantitative diagnostic protocol has been established to date.

*This work was supported by JSPS KAKENHI (JP24K02969, JP24K22316).

[1]Tsuguhiro Tsukumo and Kaixu Chen are with the Degree Programs in Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan s2420780@u.tsukuba.ac.jp

[2]Tomoyuki Asada, Kousei Miura, Hideki Kadone, Kotaro Sakashita, Takahiro Sunami, and Masashi Yamazaki are with the Department of Orthopaedic Surgery, Institute of Medicine, University of Tsukuba, Tsukuba, Japan

[3]Tomoyuki Asada is with the Hospital for Special Surgery, New York, USA

[4]Hideki Kadone is with the Center for Cybernics Research, University of Tsukuba, Tsukuba, Japan

[5]Naoto Ienaga and Yoshihiro Kuroda are with the Institute of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan

To address this limitation, various quantitative approaches for ASD diagnosis have been proposed. For example, methods that use electromyography (EMG) sensors [1] and motion capture systems [2] have shown promise in analyzing gait data from ASD patients. However, these methods require specialized equipment and controlled environments, which restrict their applicability in routine clinical settings. On the other hand, neural network models applied to lateral-view gait videos captured using easily deployable RGB cameras [3], [4] have been investigated. These methods face difficulties in accurately capturing the lateral sway patterns that are characteristic of ASD patients.

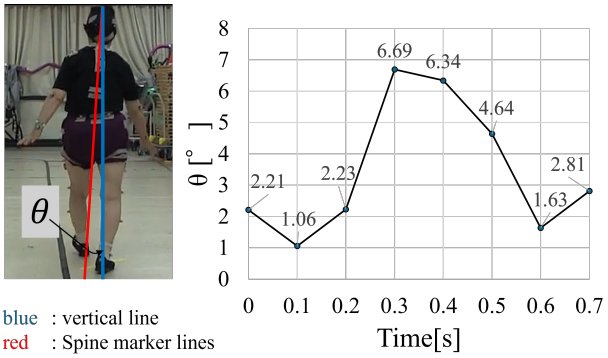

Fig. 1. Body swaying of ASD patients while walking. The time between frames is 0.1 second

As illustrated in Fig. 1, ASD patients frequently exhibit upper body tilting during gait; however, such sway patterns are insufficiently captured in lateral-view videos. In this study, we propose a novel approach for ASD classification using a dual-view neural network model trained on both frontal-view and lateral-view gait videos. Unlike conventional methods that rely on a single viewpoint, our dual-view model integrates information from both perspectives, enabling a more comprehensive analysis of gait patterns. The integration of lateral sway observations from the frontal view with postural alignment assessments from the lateral view holds significant potential to augment the accuracy of ASD classification.

II. RELATED WORKS

A. Medical diagnostic methods

The diagnosis of ASD is typically performed by using X-ray images to identify abnormalities in spinal alignment [5]. Although this method is effective for patients who can be evaluated in a static posture, gait observation is used as

979-8-3315-8618-8/25/$31.00 ©2025 IEEE

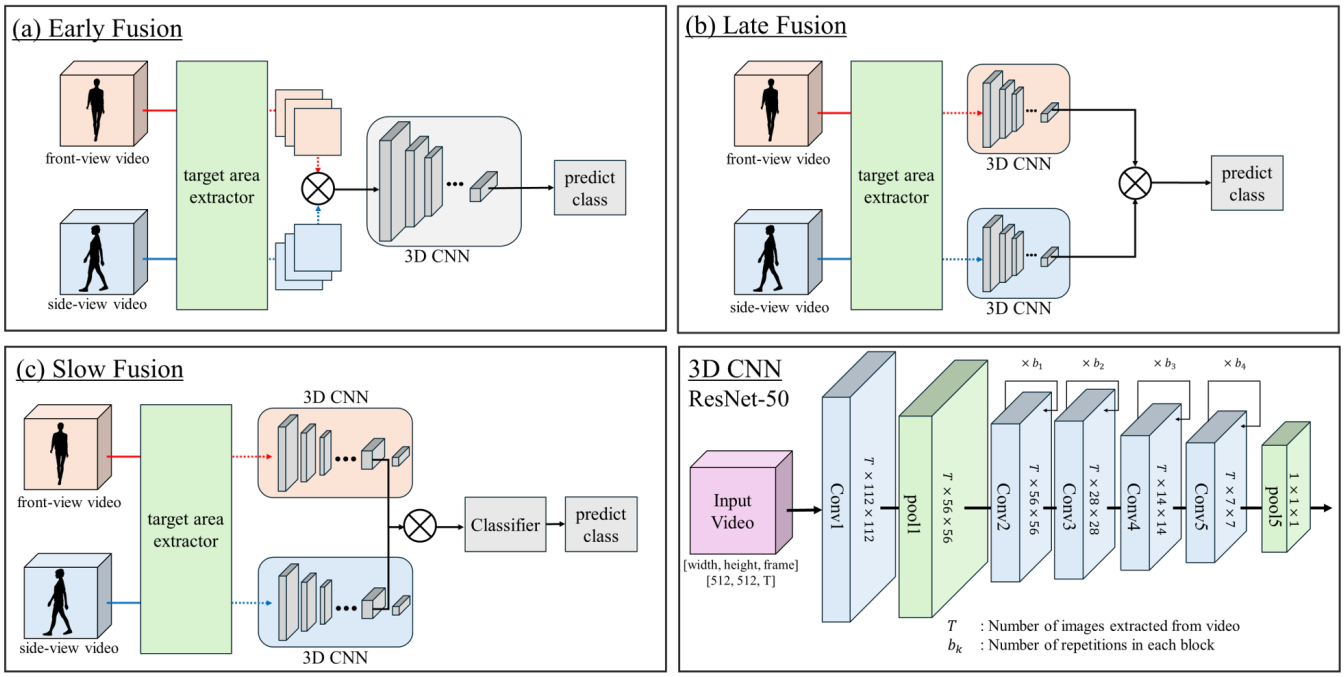

Fig. 2. Architecture of dual-view fusion model. The target area extractor extracts the target person's region from the input video and evenly samples frames. (a) The early fusion model fuses two videos by doubling the number of channels and inputs them into a single 3D CNN. (b) The late fusion model inputs the two videos into separate 3D CNNs and fuses the resulting two scores. (c) The slow fusion model inputs the two videos into separate 3D CNNs and fuses the extracted features. All 3D CNNs are based on ResNet-50 [12].

a diagnostic tool for those requiring dynamic assessment. For the quantitative analysis of gait, 3D motion capture systems have been utilized [3]. Severijins et al. [6] analyzed spinal and lower limb movements in patients who underwent corrective spinal surgery, detecting deviations from normal gait patterns to evaluate the effectiveness of interventions and treatments. However, the requirement for specialized equipment and dedicated software limits the widespread clinical application of 3D motion capture systems.

B. Deep Learning Methods

Neural networks for recognizing human actions from videos have been extensively studied. One representative approach is 3D CNN [7], which leverages convolutional operations to capture both spatial and temporal features, making it applicable to various tasks. Chen et al. [3] developed a 3D CNN model using side-view videos to detect gait disorders in ASD patients, achieving a certain level of classification performance. However, the rationale behind the model's decisions remained unclear, and side-view videos alone were insufficient to fully capture whole-body movements. Chen [4] also explored ASD classification using single-view videos by focusing on gait periodicity. They constructed a model to evaluate symmetry within the gait cycle; however, their results suggested that single-view videos still lacked sufficient information for a comprehensive analysis.

In this study, we constructed a fusion model integrating both side-view and front-view videos as inputs to a 3D CNN. There are multiple approaches for multimodal data fusion in action recognition [8], [9], [10], [11]; however, considering scalability, we evaluate and compare three fusion strategies: early fusion, late fusion, and slow fusion.

III. Proposed Method

A. Models

Fig. 2 illustrates the three fusion models constructed in this study, along with the details of the ResNet-50 [12] used.

1) **Early Fusion** combines the front and side video channels and inputs them into a single 3D CNN. The method integrates information from both videos early on, allowing the model to learn from both inputs simultaneously.
2) **Late Fusion** inputs the front and side videos into separate 3D CNNs, subsequently merging the probability scores output by each network. The features of each video are processed independently, while the prediction results are obtained by integrating the averaged probability scores.
3) **Slow Fusion** uses separate 3D CNNs for front and side videos, merges extracted features before the fully connected layer, and classifies them. This method highlights the videos' spatial features by combining information at the feature level.

B. ResNet-50

In this study, we utilize a 3D CNN-based ResNet-50 [12], depicted in Fig. 2. Videos are split into 1-second segments, with T frames extracted at equal intervals. We set T to 8 for experiments. Each frame is resized to 512×512 pixels

and processed through convolutional and pooling layers for feature extraction. Parameter k indicates the repetition count of each convolutional layer in a block, while b_k denotes block depth. ResNet enhances learning stability and addresses the vanishing gradient issue with residual connections (skip connections), supporting effective training in deep networks. The model has a total of 25,557,032 trainable parameters.

C. Target Area Extractor

To minimize background influence, we use YOLO v11 [13] to crop video segments based on body dimensions, scaling the height to 512 pixels while keeping the aspect ratio. Black fills any resizing gaps.In our setting, YOLO detects a person in approximately 30 milliseconds per frame.

IV. Experiment

In this experiment, we train single-view models as a comparison to the proposed dual-view fusion model. The single-view models are evaluated using two input patterns: frontal-view videos and side-view videos. The architecture of the single-view model is shown in Fig. 3. Similar to the dual-view fusion model, the 3D CNN employs ResNet-50 as its backbone.

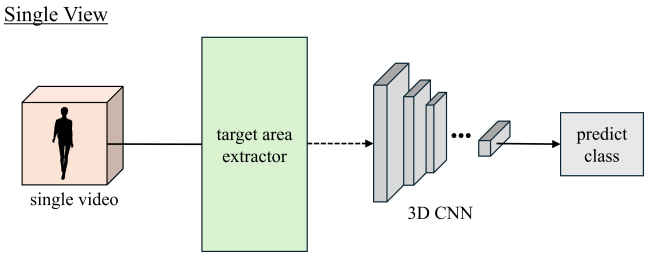

Fig. 3. single-view video model

A. Evaluation metrics

The evaluation metrics include commonly used classification metrics: accuracy, precision, and F1 score. The F1 score is key for assessing a model's performance, as it is the harmonic mean of precision and recall, indicating prediction accuracy and coverage.

B. Dataset

Fig.4 illustrates the walking course of the participants and the placement of the two cameras. The participants include not only those with ASD but also individuals with LCS (Lumbar Canal Stenosis), DHS (Dropped Head Syndrome), and HipOA (Hip Osteoarthritis). This study was approved by the Ethics Committee of the University of Tsukuba Hospital (H30-087). This study was performed in accordance with the contemporary amendments to the Declaration of Helsinki and within an appropriate ethical framework. Prior to participation, written informed consent was obtained from all participants after providing a detailed explanation regarding the purpose of the study, the nature of video recording, and the use of recorded data for research purposes. Specific attention was given to privacy protection due to the use of video data, including information about how facial and bodily features might be recorded. Participants were informed of their right to withdraw consent at any time without any disadvantages. The collected data was stored securely on encrypted servers with restricted access. The details of the dataset are presented in TableI. The duration of the videos varies depending on the severity of the symptoms for each patient. Since this study focuses on ASD, the labels are classified into two categories: ASD and non-ASD. The collected dataset length varies across different conditions, leading to an imbalance. However, the total duration of the non-ASD group was adjusted to 4,358 seconds, which is close to the 4,673 seconds of the ASD group.

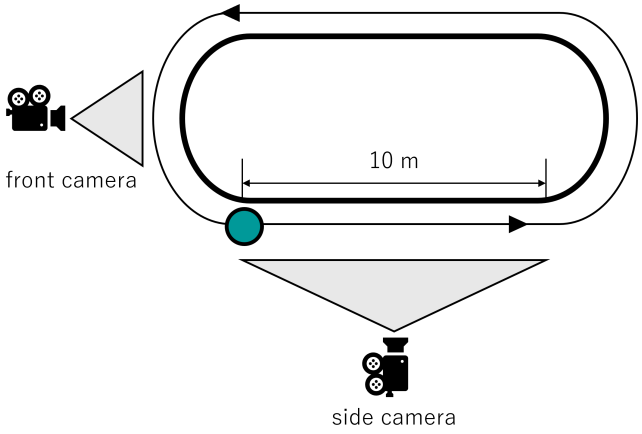

Fig. 4. Pedestrian course and camera position

TABLE I
Details of the dataset

Category	Subcategory	Participants	Duration (sec)
ASD	-	27	4,673
non-ASD	DHS	9	2,959
	LCS	6	1,026
	HipOA	2	373

In classification tasks using machine learning, datasets are typically divided into training, validation, and test sets. However, due to the small size of our dataset, we use only training and validation sets. To ensure a fair evaluation, the data is split participant-wise, maintaining an 8:2 ratio for the number of participants and videos, and five-fold cross-validation is performed. In each fold, the validation set contains participants who do not appear in the training set, preventing data leakage. Although a separate test set is not used, five-fold cross-validation helps to mitigate the risk of overfitting and provides an estimate of the model's generalization performance. The model is trained using the Adam optimizer with a learning rate of 0.0001. Early stopping is applied with a patience of 5 and a maximum of 50 epochs to prevent overfitting.

TABLE II

PERFORMANCE METRICS (MEAN ± STD) FOR DIFFERENT CONDITIONS

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
single (front)	53.39 ± 10.85	62.02 ± 9.81	14.19 ± 9.97	23.09 ± 16.03
single (side)	69.12 ± 11.01	**80.65** ± 10.23	49.20 ± 11.03	61.12 ± 15.89
Early Fusion	63.04 ± 11.48	59.64 ± 11.55	**77.53** ± 17.53	67.42 ± 11.50
Late Fusion	**70.28** ± 11.78	67.41 ± 10.23	76.94 ± 14.80	**71.86** ± 10.67
Slow Fusion	70.12 ± 7.11	73.48 ± 14.62	61.67 ± 16.11	67.06 ± 12.77

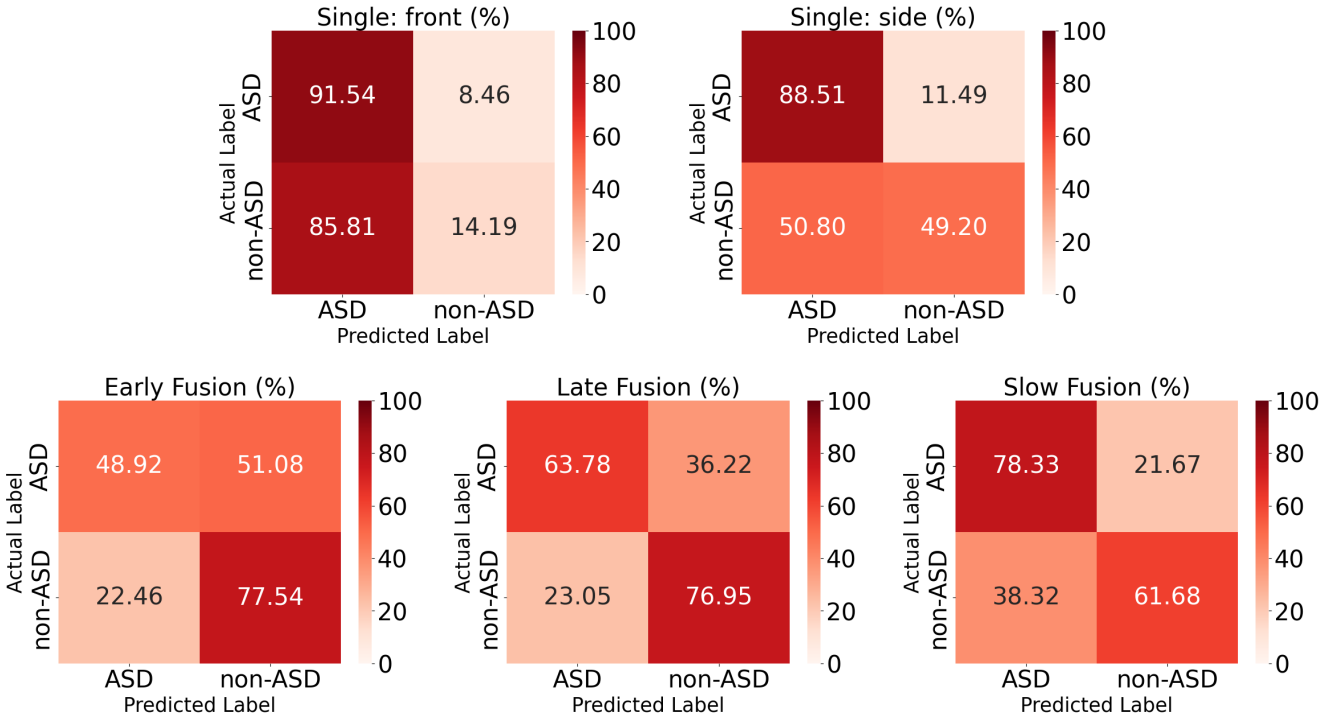

Fig. 5. Confusion matrix for each method

C. Results and Discussions

Table II presents the training results for different models. It should be noted that the standard deviations across all evaluation metrics were relatively large. This variability likely reflects differences in individual gait characteristics as well as the limited dataset size, both of which may have caused performance fluctuations across cross-validation folds.

Among the Single-frame models, the side (lateral) view demonstrated higher accuracy than the front view (69.12% vs. 53.39%). This suggests that the side view contains more discriminative features that aid classification, making it more informative than the front view, where movement characteristics may be less distinct. Early Fusion achieved an accuracy of 63.04%, which was lower than the side-view model alone. This suggests that direct concatenation of front and side view features might introduce noise or misalignment, leading to suboptimal feature representation.

In contrast, Late Fusion achieved the highest accuracy of 70.28%, surpassing all other models. This implies that training separate classifiers for each viewpoint and integrating their outputs at the decision level allows for more effective utilization of complementary information while avoiding the drawbacks of direct feature fusion.

Slow Fusion also demonstrated strong performance with an accuracy of 70.12%, closely matching Late Fusion. However, its F1 score (67.06%) was lower than that of Late Fusion (71.86%), despite achieving a higher Precision (73.48%). This suggests that Slow Fusion may be better at correctly identifying positive cases but might struggle with balanced classification across classes.

Fig. 5 illustrates the confusion matrices for each method. To evaluate the performance of each method, we selected the fold with the highest accuracy for visualization. The two Single-frame models, particularly the side-view model, achieved relatively high accuracy in detecting ASD cases.

However, they exhibited a tendency to misclassify non-ASD cases as ASD, likely due to class imbalances or insufficient feature representation in the front view. On the other hand, the fused models demonstrated a more balanced classification performance. Although the accuracy for ASD detection slightly decreased compared to the single-view model (side), the overall performance improved, particularly in correctly classifying non-ASD cases. Late Fusion, in particular, achieved the best balance, making it a promising approach for future research.

The results confirm that fusion models are effective for ASD classification, as dual-view models outperformed single-view models. Moreover, the side view provided more useful classification cues than the front view. While Early Fusion did not perform well, both Late Fusion and Slow Fusion achieved better performance. The poor results of Early Fusion suggest that the model may have struggled to effectively learn the relationship between the two viewpoints when trained on merged features. In contrast, the success of Late Fusion and Slow Fusion indicates that independently learning from each viewpoint and then optimally combining their outputs is a more effective strategy. This finding emphasizes the importance of carefully designing multi-view integration strategies, ensuring that the relationships between different viewpoints are properly captured rather than simply merging raw features.

Future research should explore refined fusion strategies, such as attention-based mechanisms or adaptive weighting techniques, to further improve classification accuracy. Additionally, investigating the effects of different temporal resolutions or alternative sensor modalities could yield valuable insights. To gain deeper insights, expanding the dataset and exploring more advanced methods for linking the two viewpoints, such as explicit spatial-temporal alignment techniques, may be necessary.

V. Conclusion

In this study, we proposed a neural network-based approach for ASD classification using gait videos captured from two viewpoints. By leveraging synchronized dual-view videos, we evaluated the performance of three fusion models through comparative experiments. Our results showed that both Late Fusion and Slow Fusion outperformed single-view models, with improvements in classification accuracy and F1 score. In contrast, Early Fusion struggled, indicating that the model had difficulty learning the relationship between the two viewpoints when trained on merged features. This underscores the importance of independent learning from each viewpoint before combining their outputs. The findings highlight the potential of dual-view gait analysis for ASD classification. Further research is needed to refine dual-view fusion strategies, expand the dataset, and improve the interpretability of the classification process for clinical applications. In addition, real-time applicability remains an important consideration for clinical deployment. While the current study focused on classification accuracy, future work should also evaluate computational efficiency and latency, particularly in scenarios where rapid feedback is required (e.g., during routine clinical screening). In our current implementation, inference for a single video takes approximately 90 to 120 s, which limits immediate feedback. However, depending on the clinical setting and available computational resources, such latency may still be acceptable for semi-automated assessments or batch processing. To move toward clinical application, it will be necessary to explore practical system integration strategies including workflow design, user interface development, and hardware optimization to ensure smooth adoption in real-world environments. Notably, our experimental results exhibited relatively large standard deviations across evaluation metrics, indicating performance variability potentially caused by inter-subject differences and the limited size of the dataset. To address this, future studies should focus on expanding the dataset and improving its diversity, which may contribute to more stable and generalizable model performance.

References

[1] K. Miura, H. Kadone, M. Koda, K. Nakayama, H. Kumagai, K. Nagashima, K. Mataki, K. Fujii, H. Noguchi, T. Funayama, *et al.*, "Visualization of walking speed variation-induced synchronized dynamic changes in lower limb joint angles and activity of trunk and lower limb muscles with a newly developed gait analysis system," *Journal of Orthopaedic Surgery*, vol. 26, no. 3, p. 2309499018806688, 2018.

[2] P. Severijns, L. Moke, T. Overbergh, E. Beaucage-Gauvreau, T. Ackermans, K. Desloovere, and L. Scheys, "Dynamic sagittal alignment and compensation strategies in adult spinal deformity during walking," *The Spine Journal*, vol. 21, no. 7, pp. 1059–1071, 2021.

[3] K. Chen, T. Asada, N. Ienaga, K. Miura, K. Sakashita, T. Sunami, H. Kadone, M. Yamazaki, and Y. Kuroda, "Two-stage video-based convolutional neural networks for adult spinal deformity classification," *Frontiers in Neuroscience*, vol. 17, p. 1278584, 2023.

[4] K. Chen, J. Xu, T. Asada, K. Miura, K. Sakashita, T. Sunami, H. Kadone, M. Yamazaki, N. Ienaga, and Y. Kuroda, "Phasemix: A periodic motion fusion method for adult spinal deformity classification," *IEEE Access*, 2024.

[5] S. D. Glassman, S. Berven, K. Bridwell, W. Horton, and J. R. Dimar, "Correlation of radiographic parameters and clinical symptoms in adult scoliosis," *Spine*, vol. 30, no. 6, pp. 682–688, 2005.

[6] K. Miura, H. Kadone, M. Koda, T. Abe, T. Funayama, H. Noguchi, K. Mataki, K. Nagashima, H. Kumagai, Y. Shibao, *et al.*, "Thoracic kyphosis and pelvic anteversion in patients with adult spinal deformity increase while walking: analyses of dynamic alignment change using a three-dimensional gait motion analysis system," *European Spine Journal*, vol. 29, pp. 840–848, 2020.

[7] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

[9] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "Mmtm: Multimodal transfer module for cnn fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13289–13299, 2020.

[10] M. Abavisani, H. R. V. Joze, and V. M. Patel, "Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1165–1174, 2019.

[11] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1159–1168, 2018.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[13] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024.