Automated Hand Prehension Assessment From Egocentric Video After Spinal Cord Injury

Nicholas Zhao and José Zariffa, *Senior Member, IEEE*

***Abstract*—Hand function assessments in a clinical setting are critical for upper limb rehabilitation after spinal cord injury (SCI) but may not accurately reflect performance in an individual's home environment. When paired with computer vision models, egocentric videos from wearable cameras provide an opportunity for remote hand function assessment during real activities of daily living (ADLs). This study demonstrates the use of computer vision models to predict clinical hand function assessment scores from egocentric video. SlowFast, MViT, and MaskFeat models were trained and validated on a custom SCI dataset, which contained a variety of ADLs carried out in a simulated home environment. The dataset was annotated with clinical hand function assessment scores using an adapted scale applicable to a wide range of object interactions. An accuracy of 0.551±0.139, mean absolute error (MAE) of 0.517±0.184, and F1 score of 0.547±0.151 was achieved on the 5-class classification task. An accuracy of 0.724±0.135, MAE of 0.290±0.140, and F1 score of 0.733±0.144 was achieved on a consolidated 3-class classification task. This novel approach, for the first time, demonstrates the prediction of hand function assessment scores from egocentric video after SCI.**

***Index Terms*—Prehension, spinal cord injury (SCI), egocentric video, automated assessment, deep learning.**

I. Introduction

SPINAL cord injury (SCI) is a devastating event that drastically impacts affected individuals, their families, and the healthcare system. Over 50% of all cases of SCI result in upper limb impairment, which greatly affects the ability to live independently [1]. In fact, the restoration of upper limb function was reported to be the top priority recovery target for individuals with SCI [2].

Within the rehabilitative process, assessment of function plays a key role in both clinical and research applications, for purposes such as monitoring progress or evaluating the efficacy of novel interventions. Current hand function assessments, such as the Graded Redefined Assessment of Strength, Sensibility, and Prehension (GRASSP) [3] are typically performed in-person at the clinic. This poses several limitations, as these assessments are performed in a highly standardized environment and manner, and therefore do not accurately reflect a patient's performance in their home environment [4]. Currently, evaluating hand performance at home relies primarily on self-report, which is susceptible to biases [5], [6], [7]. In-person assessments are also inaccessible for many patients, as transportation is a critical barrier that hinders individuals with SCI from obtaining essential needs [8].

To address the limitations described above, wearable sensors are a promising solution that can potentially perform hand function assessment within the home. In particular, wearable cameras have become widely accessible to the public and are capable of capturing a wealth of spatiotemporal data in the form of recorded first-person perspective (egocentric) video. Egocentric video contains information about the wearer's functional movements as well as valuable contextual information about the movements, such as objects that are being interacted with and environmental cues [9]. This is a critical advantage over other wearable sensors, such as inertial measurement units (IMUs), which are common within the field of rehabilitation research [10], [11], [12], [13]. Furthermore, egocentric video can easily capture detailed information about the precise movements of the hand whereas IMUs must employ more complex solutions, such as instrumented gloves, which may interfere with tactile feedback [14], [15], [16].

Although egocentric video allows for extensive recording of an individual's activities of daily living (ADLs), manually browsing through hours of raw video footage is prohibitive, especially for clinicians whose time is often at a premium. Thus, there is a need to extract key biometric data from the egocentric video footage to provide a high-level summary of the video. Deep learning models are an emerging method to automate the extraction of biometric data from egocentric video. Yet, previous works have primarily focused on extracting metrics based on detection, such as detecting hand-object

Manuscript received 24 October 2023; revised 22 March 2024, 23 June 2024, and 30 July 2024; accepted 31 July 2024. Date of publication 5 August 2024; date of current version 12 August 2024. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant RGPIN-2020-06246 and in part by the Praxis Spinal Cord Institute. *(Corresponding author: José Zariffa.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Board of the Institution (Research Ethics Board, University Health Network) under Application No. 15-8830.

Nicholas Zhao is with the Institute of Biomedical Engineering, University of Toronto, Toronto, ON M5S 3G9, Canada, and also with the KITE Research Institute, Toronto Rehabilitation Institute – University Health Network, Toronto, ON M5G 2A2, Canada (e-mail: nick.zhao@mail.utoronto.ca).

José Zariffa is with the KITE Research Institute, Toronto Rehabilitation Institute – University Health Network, Toronto, ON M5G 2A2, Canada, and also with the Institute of Biomedical Engineering, the Rehabilitation Sciences Institute, and the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G9, Canada (e-mail: jose.zariffa@utoronto.ca).

Digital Object Identifier 10.1109/TNSRE.2024.3438436

interactions [9], [17], [18], usage of compensatory grasping postures [19], and specific grasp types [20] in SCI populations. There has yet to be any work that has attempted to directly assess the quality of hand function from egocentric video using deep learning models. Thus, we are interested in answering the following question: *can we use deep learning models to extract information about the quality of hand function from egocentric video*? In this study, we framed the problem of assessing the quality of hand function as the estimation of hand function scores, derived from scales used in commonly employed clinical outcome measures. We therefore developed computer vision models to predict clinical hand function assessment scores from egocentric video. A successful implementation of such a model will contribute to the development of a fully automated hand function assessment method for individuals with SCI. This will allow hand function assessment to be more accessible and relevant to a patient's daily life, for both research and clinical applications.

II. Methods

A. Dataset

To develop the deep learning model, a specialized dataset was first constructed to train, validate, and test the model. The ANS-SCI dataset [9], which was previously collected by our group was used for this study. It contains over 1200 minutes of egocentric video footage of 17 participants with cervical SCI performing approximately 38 common interactive ADL tasks, which have been identified by the American Occupational Therapy Association (AOTA) as important [21]. The participants' inclusion criteria encompassed individuals with cervical SCI whose AIS grades ranged from A-D. The participants' had an average age of 50±12 years and included 15 males and 2 females, with AIS grades from A-D. The footage was recorded at the HomeLab home simulation environment at the KITE Research Institute. The tasks were recorded in several home environments, including a kitchen, living room, bedroom, and bathroom. The study participants provided written consent prior to participation in the study, which was approved by the Research Ethics Board of the institution (Research Ethics Board, University Health Network: 15–8830).

Figure 1 depicts four example frames of participants performing various tasks in the different environments in the HomeLab. This dataset was chosen for this study because of the standardization of the ADLs found in the dataset. As will be seen in the next section, an adapted version of the GRASSP Prehension Performance subtest was used to annotate the dataset with clinically relevant hand performance assessment scores. These annotations require every possible expected grasping posture to be identified for each ADL found in the dataset, thus the standardization of ADLs across all participants is required for feasible annotation. Other datasets, particularly datasets that are recorded at the participant's home, contain many different ADLs that differ between each participant. Identifying the expected grasping posture for possibly hundreds of unique ADLs is impractical. Thus, the standardization of ADLs was vital for the practicality of this study.

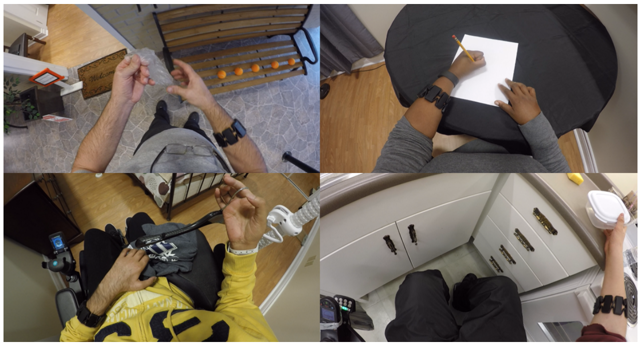

Fig. 1. Example frames from the ANS-SCI dataset. Top-left: Placing tennis balls into a plastic bag in the walkway. Top-right: Writing on paper in the dining room. Bottom-left: Hanging clothes in the bedroom. Bottom-right: Grabbing a plastic container in the kitchen.

TABLE I
Scoring Criteria for the GRASSP Prehension Performance Subtest [24]

Score	Criterion
0	The task cannot be conducted at all.
1	The task cannot be completed (less than 50% of the task).
2	The task is not completed (50% or more of the task).
3	The task is conducted (completed) using tenodesis or an alternative grasp other than the expected grasp.
4	The task is conducted using the expected grasp with difficulty (lack of smooth movement or difficult slow movement).
5	The task is conducted without difficulties using the expected grasping pattern and unaffected hand function.

B. Scoring Methodology

Out of the potential SCI hand performance measures, the GRASSP Prehension Performance subtest [3] was chosen for this study. Firstly, GRASSP was primarily chosen because of its proven reliability and validity, as well as its ongoing adoption as the gold-standard measure for SCI hand performance in many locations. This study required a quantitative measure of hand performance that can be assessed from video footage alone. The measure must also be easily adapted for the various ADLs found in the dataset, which may conflict with assessments that require specific methods of scoring hand performance. For example, the Capabilities of Upper Extremity Test (CUE-T) [22] test is another widely used SCI hand performance assessment but has unique scoring criteria for each of the tasks, making it difficult to translate to generic ADLs. Although the GRASSP Prehension Performance subtest defines specific tasks to be performed, it has universal scoring criteria that are used for every task, shown in Table I. These criteria are based on task completion, usage of the expected grasp, smoothness of movement, and speed of movement. For each task, a 50% completion checkpoint, a 100% completion checkpoint, and a list of expected grasps are defined for the examiner.

For the purposes of this study, an adapted version of the Prehension Performance subtest was created. The scoring

TABLE II

50% Completion Checkpoint, 100% Completion Checkpoint, and Expected Grasp Types for 3 Example Tasks Found in the ANS-SCI Dataset. Expected Grasp Types Follow the GRASP Taxonomy of Human Grasp Types [23]

Task	**50% Completion Checkpoint**	**100% Completion Checkpoint**	**Expected grasp types** [23]
Writing	Pencil lead makes contact with paper	A legible character is written on the paper	Prismatic 2 finger Prismatic 3 finger Prismatic 4 finger Writing tripod
Swipe card	Card enters the card slot but does not travel the entire length of the reader	Card is swiped through the entire length of the card reader	Lateral
Coin and slot	Coin is lifted off the table	Coin is placed into the slot	Tip pinch Palmar pinch

method of the original subtest was preserved and used to score hand performance of the various ADLs found in the dataset. Following the GRASSP scoring criteria, 50%/100% completion checkpoints and expected grasps were defined for each ADL. The expected grasps were taken from the GRASP taxonomy of human grasp types [23], and each ADL found in the ANS-SCI dataset was assigned a list of expected grasps. Following GRASSP, the checkpoints were defined by empirically determining key stages within each task and setting the checkpoints accordingly [24]. For reference, the completion checkpoints and expected grasp types from 3 example tasks in the ANS-SCI dataset are shown in Table II. These 3 tasks were chosen to demonstrate the variety of expected grasp types found in the dataset.

A consolidated scoring scheme was also investigated for this study, where scores 0-2 were combined (task not completed), score 3 was left unchanged (task completed with alternative grasp), and scores 4-5 were combined (task completed with expected grasp), yielding a 3-point scale. Two methods of implementing class consolidation were explored in this study. The first method consolidated the classes on the model level by training a new model with 3 output neurons instead of 5. The second method retroactively consolidated the classes by applying the consolidation scheme to the predictions of a model that was trained with 5 classes.

Using the adapted GRASSP Prehension Performance scoring, the entirety of the ANS-SCI dataset was annotated with GRASSP scores (0-5 scale), along with timestamps for each of the tasks. The timestamps were used to extract only the active ADL footage from the dataset, and each ADL clip was labelled with a GRASSP score.

The class distribution of the annotated dataset can be seen in Figure 2. After extracting the active ADL footage, the dataset consisted of over 40,000 frames. There is an evident class imbalance in this dataset, as GRASSP scores 3 and 4 comprise over 70% of the entire dataset. The consolidated labels show a more balanced distribution. Furthermore, there were no tasks that were scored as 0, thus score 0 was removed, making the classification a 5-class problem, from scores 1-5.

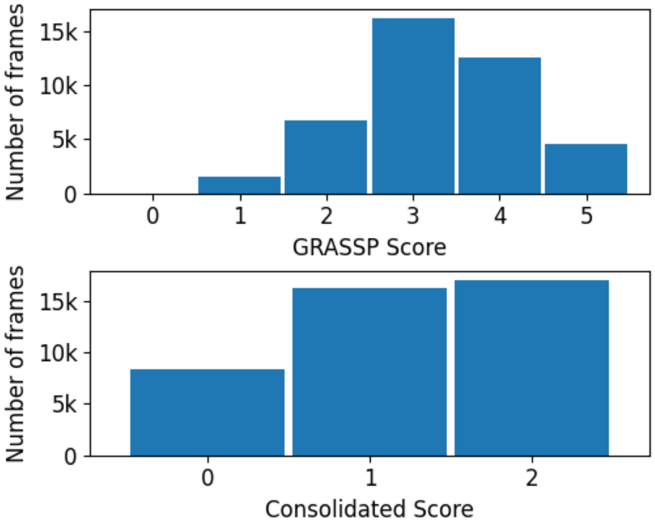

Fig. 2. Class distribution of the dataset with the original GRASSP scoring (top) and the consolidated scoring (bottom).

To validate the clinical relevance of the adapted GRASSP scoring, the mean GRASSP score for each participant was compared to their Upper Extremity Motor Score (UEMS) component of the International Standards for the Neurological Classification of SCI [25]. Concurrent validity was determined via the Spearman's rank correlation coefficient, where the mean GRASSP score achieved a coefficient of 0.81 ($p < 0.001$), demonstrating strong concurrence with UEMS. The inter-annotator reliability of the adapted GRASSP scoring was also evaluated on a randomized subset of the dataset, consisting of 4 randomly selected tasks from each participant. Cohen's Kappa scores of 0.73 and 0.78 were achieved for the original and consolidated scoring, respectively.

C. Deep Learning Model

To predict GRASSP scores from egocentric video footage, state-of-the-art video classification architectures were chosen as candidates for investigation of their performance in this novel task. Video classification architectures were chosen over frame-level image classification, due to the capacity to integrate temporal information. Furthermore, frame-level image classification performance is sensitive to occlusions, which can be mitigated by the variety of hand angles that appear in video data. Video classification architectures that demonstrated high performance in action recognition datasets, such as Kinetics-400 [26], were particularly favored in selection, as action recognition likely requires the extraction of visual features that are also relevant to hand performance assessment. Furthermore, video classification architectures were chosen over frame-level image classification, since the GRASSP scores consider temporal features, such as speed of movement. The following video classification architectures were chosen for this study:

The SlowFast network is a convolutional neural network-based architecture proposed by Feichtenhofer et al. in 2019 and was one of the most promising architectures for this study, as it exceeded state-of-the-art performance in a variety of action recognition datasets, including Kinetics-400 [27]. The usage of a low temporal resolution (slow) pathway and a high temporal resolution (fast) pathway was identified as potentially useful for detecting the nuances of impairment in individuals with SCI. SlowFast consists of 33.6 million parameters, making it the smallest model in the study.

The MViT architecture, proposed by Fan et al. in 2021 is a transformer-based architecture that hierarchically expands the channel capacity while reducing spatial resolution [28]. MViT exceeded state-of-the art performance in many video datasets, including the Kinetics-400 dataset. Li et al. then released the MViTv2 architecture in 2022, which exceeded the state-of-the-art in Kinetics-400 again [29]. This study used the MViTv2-B architecture but is referred to as MViT in this document. The MViTv2-B architecture consists of 50.9 million parameters, making it the largest model in the study.

The MaskFeat model, proposed by Wei et al., uses the MViT architecture, but utilizes a self-supervised pretraining method, which has been proposed to initialize models for robust semantic understanding of visual data [30]. This self-supervised pretraining method prepares typically data-demanding transformer architectures to be readily trained on smaller datasets using transfer learning. The MaskFeat model demonstrated state-of-the-art performance on smaller datasets such as Something-Something v2, using transfer learning. The MaskFeat model in this study used the MViT-S backbone and was pretrained on the Kinetics-400 dataset using the self-supervised method. The MaskFeat architecture consists of 36.2 million parameters.

D. Hyperparameter Optimization

Model hyperparameters were optimized by first identifying and optimizing high-level hyperparameters, which have significant architectural effects on the model, then optimizing low-level hyperparameters, which are model values such as learning rate and weight decay. We explored 3 high-level hyperparameters in this study: transfer learning methods, temporal sampling methods, and ordinal regression methods.

The transfer learning methods explored in this study were training from scratch, fine-tuning, and feature extraction. Models that used feature extraction had all layers but the final layer frozen, while fine-tuned models allowed all model parameters to be updated. All pretrained weights used for transfer learning were obtained from publicly available sources and were trained on common action recognition and egocentric video datasets. SlowFast models were pretrained on Epic Kitchens [31], MViT models were pretrained on Kinetics 400 [26] and Something-Something v2 [32], and MaskFeat models were pretrained on Kinetics 400 [26].

We explored two different temporal sampling methods in this study. The first method we employed was typical temporal window sampling, where a video clip was sampled with a moving window of 32 frames, frame stride of 2, and window stride of 32. Sparse temporal sampling, as described by Wang et al. [33], was also explored in this study. This method divides the entire video clip into K equal segments and samples N consecutive frames from each segment, starting at a random index, resulting in KxN total frames. From experimentation, we found that $K = 8$ and $N = 4$ resulted in optimal performance, and thus were used for the remainder of the study. Due to the randomization of index selection for the starting frame of each segment, sparse temporal sampling could be applied to the same video multiple times and yield outputs with significantly different video content. Thus, videos whose classes were underrepresented in the dataset were also oversampled with sparse temporal sampling to balance the class distribution. The minority classes were oversampled to generate an equal class distribution for each participant. However, since the class balancing was done on a participant-basis, if a particular participant did not demonstrate any instances of a single class, the class would still be absent from the dataset, as there would be no video to oversample.

The classification task in this study is of ordinal nature, thus we explored two ordinal regression methods that attempt to exploit the ordinal relationship between the classes. The methods we explored were Consistent Rank Logits (CORAL) [34] and Conditional Ordinal Regression for Neural Networks (CORN) [35], as well as the absence of any ordinal regression method.

The low-level hyperparameters that were optimized in this study were learning rate, weight decay, label smoothing, batch size, optimizer, and warmup batches. Optimization was performed by beginning at the values that were presented in the original study for each architecture, then sweeping an order of magnitude above and below the original value.

E. Model Evaluation

All models were evaluated using Leave-One-Subject-Out cross validation (LOSO-CV), meaning 17 versions of the same model were trained, with each participant being left out of the training set in turn, to be used as the validation set. Then, the performance of the 17 models was averaged to give the final model performance. Model performance was determined with the following classification metrics: accuracy, mean absolute error (MAE), and weighted F1 score. Particularly, MAE was used to distinguish the magnitude of misclassifications (i.e., predicting a score of 5 as 1 gives more error compared to predicting a score of 5 as 4).

Rather than evaluating model performance by averaging the results of all the individual predictions, this study used an aggregate task score to determine performance. For a video clip containing the footage of a single task being performed, the clip was sampled multiple times depending on its duration (see above for sampling methods). Each sample was passed through the model to make a prediction and the predictions for a single task clip were aggregated via a majority vote. The aggregated prediction was then evaluated with the aforementioned metrics (i.e., MAE). This aggregate prediction method was used instead of the individual sample predictions, as it emulates how the model would likely be used in practice.

In addition to LOSO-CV, leave-one-task-out cross validation (LOTO-CV) and leave-one-background-out cross validation (LOBO-CV) were also implemented in this study, as an investigative tool. LOTO-CV extracts all the videos of a specific task and uses them as the validation set, while training on

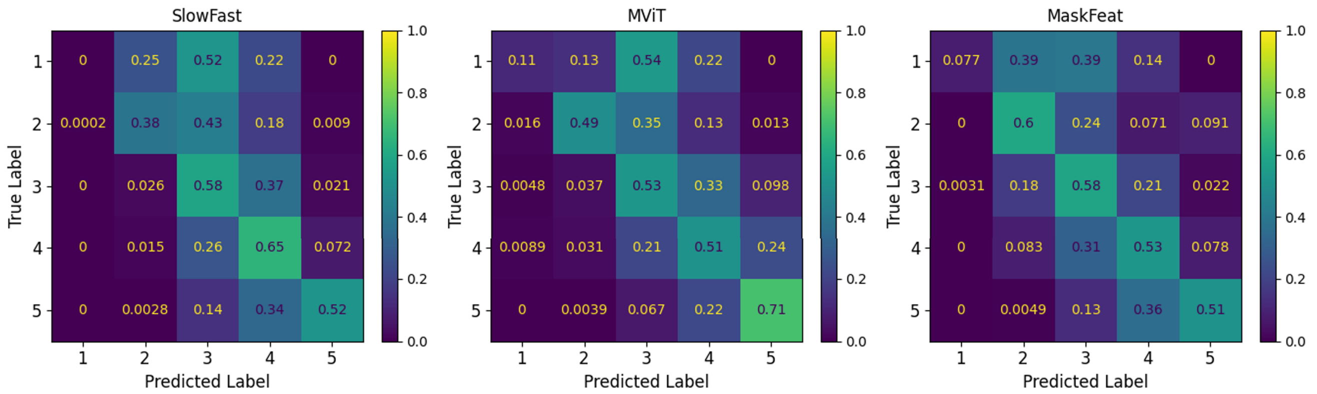

Fig. 3. Confusion matrices of the top performing Slowfast (left), MViT (middle), and MaskFeat (right) models. Cell values are normalized across each row. Each cell value represents the proportion of predicted labels for each true class label.

the rest of the data. This process is repeated for every task in the dataset. Similarly, LOBO-CV extracts all the videos that were filmed at a specific setting in the HomeLab and uses them as the validation set. Due to the added computational load of these methods, LOSO-CV and LOBO-CV were only performed with the SlowFast architecture.

Due to the highly resource intensive nature of LOSO-CV, hyperparameter optimization was performed on a representative subset of 5/17 participants in the dataset. The models were trained on the entire training set, but only 5 models were trained instead of 17. The 5 participants in the representative subset were chosen to have a wide range of average GRASSP scores, UEMS, and AIS grades, ranging from 2.4-4.7, 9-24, and B-D, respectively. The final optimized models were then evaluated on all 17 participants using LOSO-CV.

To compare the performance metrics between models, statistical significance was determined as follows. Normality of the data distribution was first checked using the Shapiro-Wilk test. To determine statistical significance between the means of two groups, a t-test was performed on normally distributed data and the Wilcoxon rank-sum test was performed on non-normal data. For groups of 3 or more, one-way ANOVA followed by the Tukey HSD post-hoc test was used for normal data and the Kruskal-Wallis H-test followed by Dunn's post-hoc test was used for non-normal data. For all tests, a p-value of $\alpha = 0.05$ was used to determine whether to reject the null hypothesis.

III. Results

After performing hyperparameter optimization, the optimized models used fine-tuning and sparse temporal sampling across all architectures. As for ordinal regression, the SlowFast and MViT architectures used CORAL and the MaskFeat architecture used CORN.

The average performance metrics achieved by the optimized models can be found in Table III. All reported results are from models trained with 5 classes and evaluated using LOSO-CV, except where noted otherwise. The differences between any performance metric across all architectures were not statistically different from one another ($p > 0.05$ for all). Notably, the SlowFast model performed statistically similar to the other models, while also being the smallest model at 33.6 million parameters. The SlowFast model also demonstrated an accuracy of 0.551, surpassing the even chance accuracy of a 5-class classification task, 0.2, and the accuracy of only guessing the majority class, 0.382.

TABLE III
Mean Absolute Error, Accuracy, and Weighted F1 Score for the Top Performing Models of Each Architecture

Architecture	Accuracy	MAE	Weighted F1
SlowFast	**0.551±0.139**	**0.517±0.184**	**0.547±0.151**
MViT	0.516±0.122	0.558±0.161	0.534±0.121
MaskFeat	0.520±0.123	0.570±0.209	0.535±0.119

Figure 3 shows the confusion matrices of the top performing models of each architecture, normalized row-wise, across the true labels. The values on the diagonal of the matrix can be considered an accuracy for the respective true class label, as they represent the proportion of the validation set for that label that were correct predictions. All architectures had difficulty in classifying samples with a score of 1, particularly SlowFast, which did not make any correct predictions for that class. The SlowFast model performed the best at predicting scores 3-5, while showing poor performance for scores 1-2. The MViT model demonstrated better performance at predicting scores 1, 2, and 5, but worse performance for scores 3 and 4. Notably, the MViT model showed considerably high accuracy for predicting scores of 5.

In addition to comparing the model output to their respective ground truth labels, we also calculated the ranked correlation of the average GRASSP scores as an alternative evaluation method. An average predicted GRASSP score was calculated for each participant, by taking the mean of the model's aggregate task scores. Spearman's rank correlation coefficient between the predicted GRASSP scores and the mean and median annotated GRASSP scores were calculated. Correlation was also calculated between the predicted GRASSP scores and the participant's UEMS. The correlation coefficients and p-values can be found in Table IV.

Figure 4 shows the comparative performances between consolidating classes on the model-level, retroactively, and the 5-class implementation, across the three architectures. There were no statistical differences between the two class consolidation methods, in all cases ($p > 0.05$). Both consolidation methods demonstrated statistically significant improvements

TABLE IV
SPEARMAN'S RANK CORRELATION COEFFICIENT BETWEEN THE AVERAGE PREDICTED GRASSP SCORES FOR EACH ARCHITECTURE (ROWS) AND ANNOTATED/MEASURED HAND FUNCTION ASSESSMENT SCORES (COLUMNS)

	Mean GRASSP Score	Median GRASSP Score	UEMS

SlowFast	ρ = 0.82
	p = 2.4e-5
		ρ = 0.87
		p = 3.4e-6
			ρ = 0.56
			p = 9.7e-3

MViT	ρ = 0.85
	p = 9.5e-6
		ρ = 0.87
		p = 2.5e-6
			ρ = 0.71
			p = 7.2e-4

MaskFeat	ρ = 0.85
	p = 9.5e-6
		ρ = 0.84
		p = 1.2e-5
			ρ = 0.65
			p = 2.5e-3

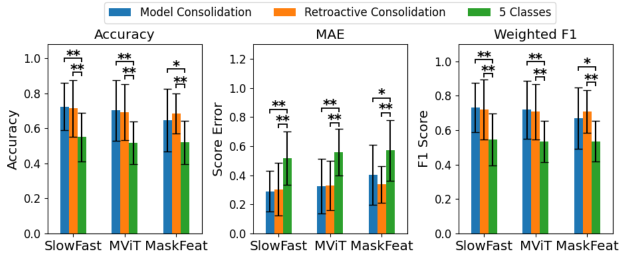

Fig. 4. Accuracy, MAE, and weighted F1 score for each class consolidation method, including the 5-class implementation, and architecture. * $p < 0.05$, ** $p < 0.01$.

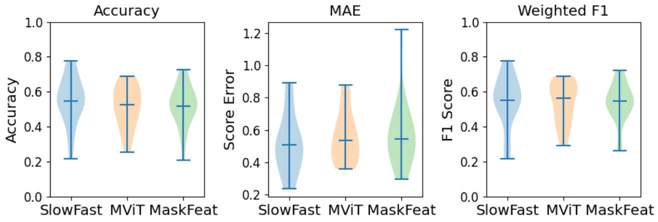

Fig. 5. Distribution of performance metrics across participants, using LOSO-CV.

to model performance, compared to the 5-class case. All improvements demonstrated a p-value of less than 0.01, except between the model-level consolidation and the 5-class case in MaskFeat, which demonstrated a p-value of less than 0.05 in all metrics. The top performing model in all metrics was SlowFast with model-level class consolidation, demonstrating an average accuracy, MAE, and weighted F1 score of 0.724, 0.290, and 0.733, respectively. These results surpass the even chance accuracy of a 3-class classification task, 0.333, and the accuracy of only guessing the majority class, 0.410.

Figure 5 shows the distribution of performance metrics across all participants for each architecture. As seen, model performance greatly varied between participants. SlowFast demonstrated the widest distribution in accuracy and weighted F1 score, ranging from 0.220-0.778 and 0.217-0.776, respectively. MaskFeat saw the highest variation in MAE, ranging from 0.296-1.22.

Figure 6 shows the average SlowFast performance metrics of the two alternative cross validation methods (LOBO-CV and LOTO-CV), as well as the default cross validation method (LOSO-CV). There were no significant differences found between the average performance of LOSO-CV and LOBO-CV in all metrics ($p > 0.05$). LOTO-CV was found to have a statistically significant increase in accuracy and weighted F1 score, as well as a decrease in MAE compared to the other two cross validation methods.

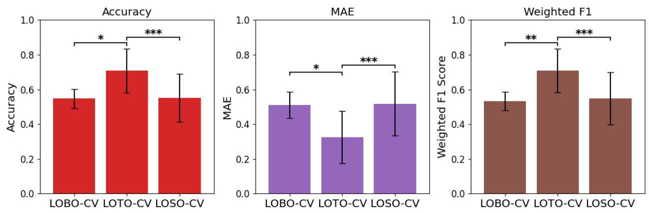

Fig. 6. Average SlowFast accuracy, MAE, and weighted F1 score for each cross-validation method; leave-one-background-out cross-validation (LOBO-CV), leave-one-task-out cross-validation (LOTO-CV), and leave-one-subject-out cross-validation (LOSO-CV).

IV. DISCUSSION

A. Model Performance

The classification performances achieved by the highest-performing models support the hypothesis that automated assessment of hand function quality from egocentric videos is possible via deep learning models. Achieving an average classification accuracy of 0.551 on the 5-class classification task and 0.724 on the 3-class classification task demonstrates that the model learned some meaningful semantic representation of the egocentric video data and was able to recognize key visual features related to hand function quality.

The confusion matrices in Figure 3 reveal considerable information about the differences in behavior between the different architectures. One clear behavior of interest is the tendency for the SlowFast and MViT models to misclassify a score of 1 as a score of 3. Due to the class imbalance of the dataset, it was expected that the minority classes of 1 and 5 would result in a higher misclassification rate, which was certainly the case for the score of 1. Interestingly, the models performed well at classifying samples with a true score of 5 –achieving higher class-specific accuracy than the score of 2, in the case of SlowFast and MViT, despite having a lesser representation in the dataset. This may be the result of transfer learning: the top performing models were pretrained on data from uninjured individuals, which would be more similar to the samples with a score of 5.

Although the models demonstrated imbalanced class-wise performance, the optimized models all used sparse temporal sampling, which balanced the class distribution. However, the class balancing was achieved through oversampling, thus no new information was truly added to the dataset. The underlying class distribution of the dataset is still evident in the class-specific performance of the models, such as the poor performance for scores of 1. These results demonstrate the limitations of sampling-based class balancing methods and suggest that additional egocentric video data should be collected to effectively balance the dataset.

The results from Table IV propose an alternative evaluation method for the models, by calculating the ranked correlation between the predicted and annotated average GRASSP scores of each participant. An advantage of this evaluation method is that it ignores the numerical difference between the scores. This is desirable, as the difference between GRASSP scores does not have a numerical interpretation – there is no meaningful interpretation of whether the difference between GRASSP scores 1 and 2 is greater or less than the difference between scores 3 and 4. The MAE is flawed in this way, as it treats all errors as numerically equivalent. Accuracy and F1 score similarly do not necessarily capture the full behavior of the models, as they only evaluate predictive performance. Yet, predictive performance is not the only desirable property of these models. For instance, concurrence with other validated outcome measures is also a desirable property. The predicted average GRASSP scores demonstrated correlation coefficients with UEMS ranging from 0.560-0.709, which can be considered as moderate to good concurrence [36]. The predicted average GRASSP scores also demonstrated correlation coefficients with the mean and median annotated scores ranging from 0.824-0.846 and 0.841-0.872, respectively, which are considered as good to excellent concurrence [36]. These results suggest that model outputs could potentially be used to develop an outcome measure that is independent from GRASSP. It should be noted, however, that the annotated scores are merely adapted from a validated outcome measure and are not truly validated themselves.

The two class consolidation methods tested in Figure 4 did not demonstrate significant effects on model performance between each other but did improve performance compared to the 5-class classification case. The top performing consolidated model demonstrated promising performance, with an accuracy of 0.724 on the 3-class classification task. The boost in model performance is likely attributed to two main factors: reduction of bias due to a more balanced dataset and simplification of the classification task. The lack of significant differences between the two consolidation methods suggests that the effect of consolidation on the model itself is minimal. The 5-class model may be implicitly grouping the classes in a similar manner to the consolidated scoring, thereby demonstrating similar results between retroactive and model consolidation. The performance of the consolidated model, compared to the original model, suggests that the 5-class classification task is too difficult with respect to the dataset size. Healthcare tools generally demand strong and predictable performance in order to be adopted in the field. Focusing on the 3-class classification task may be warranted for future investigations.

The models exhibited their strongest performance for participants 4 and 10. These participants have similar levels of hand function, as their mean GRASSP scores were 3.30 and 3.79 respectively, and their UEMS were both 20. It is not surprising that the model excelled with these participants, as scores 3 and 4 were overrepresented in the dataset. This suggests that additional data for the remaining scores may result in similarly strong performances for the other participants. Conversely, the models exhibited their worst performance for participants 14 and 16. Their mean GRASSP scores were 2.42 and 4.34, and their UEMS were 16 and 20, respectively. Again, it appears that the model predictably performs worse on validation participants who lie on the tails of the class distribution of the dataset. These results further suggest that future studies require more data of the minority classes in the dataset.

The main purpose of conducting the alternative cross-validation methods LOTO-CV and LOBO-CV was to investigate whether the model was biased towards particular aspects of the video data, such as objects and backgrounds. LOTO-CV involves validating on a task that the model has never seen before, thus demonstrating a strong performance on LOTO-CV strongly suggests that the model is not memorizing objects within the videos. Similarly, LOBO-CV involves validating on a background that the model has never seen before, thus a lack of performance degradation suggests that the model is not memorizing backgrounds to make its prediction.

The average performance metrics for LOTO-CV were significantly higher than the performance when using LOSO-CV. One main reason for this is likely due to the smaller size of the validation set. Furthermore, the validation set's variability was only due to differing participants, rather than differing tasks, making the validation task arguably easier than for LOSO-CV.

The average results of the LOBO-CV SlowFast models were not statistically different from the LOSO-CV model's results, which suggests that the models are identifying visual features that are agnostic to specific backgrounds or participants in the dataset. This is promising for the generalizability of the model, as the learned features may be able to generalize past the ANS-SCI dataset used in this study.

B. Limitations and Future Work

A major limitation for this study was the size of the dataset, as state-of-the-art video classification models are typically trained on datasets that are orders of magnitude larger, such as Ego4D [37], Kinetics400 [26], Something-Something v.2 [32], and AVA v2.2 [38]. A sufficient amount of data is a critical requirement for a machine learning model to generalize effectively – otherwise, models naturally tend to overfit to the training data.

Class consolidation was an attempt to balance the dataset, while simultaneously provide additional samples for each class label. However, the 3-point scale used for consolidation has not been formally validated, despite being derived from a validated measure.

Another potentially limiting factor in this study was the high variability in impaired hand postures. In uninjured individuals, hand postures generally converge to a few commonly expected hand postures for a given task. However, hand impairment is highly heterogenous and can manifest into less standardized grasping strategies, depending on level of injury, severity of injury, and recovery profiles. As a result, impaired hand postures often are unique for each individual, thus introducing high variability in the dataset. Figure 7 shows an example of hand posture heterogeneity during a page flip in the "read_book" task. Although each of the tasks in the figure were given an adapted GRASSP score of 2, the hand postures are unique for each participant. For instance, the four participants all demonstrated different contact points for the flipped page; participant 1 held the page between the index and thumb on the left hand, participant 8 held the page between two closed fists, participant 9 held the page between the middle and ring fingers on the right hand, and participant 14 held

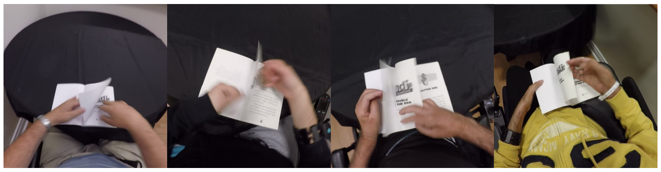

Fig. 7. (Left to right) Participants 1, 8, 9, and 14 flipping a page during the "read_book" task. The depicted tasks were given an adapted GRASSP score of 2.

the page between the index and middle fingers on the right hand. Furthermore, the book is also held steady with varying supporting hand postures.

Due to the high variability within a single class definition, the ability to learn meaningful representations of the video content may be compromised. The combination of small dataset size and high intraclass variability results in a scarcity of consistent visual features for the model to detect and affiliate with a particular class label. This can lead the model towards the memorization of common visual content in the videos, such as objects and backgrounds. Although objects and backgrounds provide valuable contextual information about hand function, if they dominate the model's focus, the model consequently learns a suboptimal representation of the video content. Ideally, the model should primarily focus on the hands, while taking contextual clues such as objects and backgrounds to supplement the decision making. Larger datasets could be used in the future to train models on more homogeneous sub-groups of individuals with similar levels and severities of lesions, and thus more similar postures. This approach may provide additional insights into the impact of grasp variability on model performance.

Previous works using the ANS-SCI dataset experienced similar high variability between impaired participants. Dousty et al. found that the high variability between participants in LOSO-CV results in non-identical distributions between the test and validation set, which violates the critical assumption that the test and validation distributions must be identical in order for a machine learning model to generalize [20]. These findings coincide with our results, as Figure 5 also demonstrates high variability in model performance between validation subjects.

The proposed method for automated hand function detection yielded promising results, but there are clear areas of improvement. The end-to-end learning method used in this study mainly relies on the loss function and propagation of the subsequent gradients to do all the work of learning meaningful representations of the video data. With sufficient data, this approach may be effective, but with the limited size and high variability of the dataset, the model may have tended to overfit. As such, in order to work with limited data, a feature extraction pipeline may be a more effective method, which can guide the model towards learning the relevant aspects of the video. There are many possibilities for the design of such a pipeline. Recently, the Segment Anything Model (SAM) demonstrated remarkable performance in semantic segmentation of images [39]. Semantic segmentation could be used to isolate the hands and active objects in the video scene to direct model learning towards focusing on the hands. Another possibility is to use a pose estimation network to obtain postural information of the hand. Dousty et al. demonstrated that combining postural and contextual information allowed for strong performance in predicting hand grasp types [20]. A similar architecture may also be employed for automated hand function assessment.

However, a major limitation of a pipeline approach is the propagation of error. Each model in the pipeline has an associated level of error, thus the performances of the subsequent models are capped by the models before them. Furthermore, one of the main advantages of egocentric video is the capability of capturing rich contextual information, thus segmentation approaches can omit valuable contextual features, and thus severely mitigating this advantage.

A logical next step is to use a larger dataset representative of hand impairments to train and validate a new model. Previously, Bandini et al. collected over 65 hours of egocentric video footage from individuals with cervical SCI performing their normal daily routines at home [18]. This dataset is a very strong candidate for future work in this field, as it can greatly bolster the data content for training new models. However, this dataset is much more heterogenous than the ANS-SCI dataset, since it recorded the participants' daily routines in their real home environments. As such, there is much more variability in this dataset, such as a wider breadth of tasks, lighting conditions, backgrounds, and objects. This variability makes annotation more costly but will aid in the development of a model that is robust to these variations.

Due to the cost of annotation, self-supervised pretraining has become a popular technique to initialize deep neural networks, as it can make use of unlabeled data for pretraining. The MaskFeat architecture leveraged self-supervised pretraining on the Kinetics-400 dataset in an attempt to learn universally relevant spatiotemporal features, but did not result in significant improvements in this work compared to the other architectures. This may be due to the domain differences between the source and target datasets. Consequently, self-supervised pretraining could potentially be used with the dataset above from Bandini et al. to initialize the networks. This would alleviate the high annotation cost due to the high variability of the dataset, while performing novel self-supervised pretraining on an egocentric SCI dataset. Domain adaptation is a significant barrier that limits transfer learning from publicly available datasets collected with uninjured individuals, thus performing transfer learning between egocentric SCI datasets is an interesting future direction to pursue.

Encoder-decoder architectures aim to construct a latent representation of the data, which is agnostic to specific labels [40]. Leveraging a latent representation of hand usage from egocentric video to develop an independent SCI outcome measure is an exciting and promising future direction.

V. Conclusion

The objective of this study was to determine whether deep learning models are capable of extracting information about hand function quality from egocentric video. The results presented in this work strongly support that this notion is true. We presented 3 modern video classification architectures, each demonstrating acceptable performance on the novel 5-class and 3-class classification tasks, after hyperparameter optimization via a guided grid search. Through further explorations, we demonstrated that the models learned meaningful, semantic representations of the video data, rather than simple memorization of extraneous features. This work demonstrated that computer vision models can make valuable contributions to the future of SCI rehabilitation by enabling assessments that are practical, relevant, and accessible.

Acknowledgment

The authors would like to give thanks to Dr. Sukhvinder Kalsi-Ryan for her consultation regarding the GRASSP scale and to Dr. Damian Manzone for annotating a subset of the dataset to perform the reliability study.

References

[1] Nat. Spinal Cord Injury Stat. Center. (2020). *Facts and Figures at a Glance*. [Online]. Available: https://www.nscisc.uab.edu/Public/Facts

[2] K. D. Anderson, "Targeting recovery: Priorities of the spinal cord-injured population," *J. Neurotrauma*, vol. 21, no. 10, pp. 1371–1383, Oct. 2004, doi: 10.1089/neu.2004.21.1371.

[3] S. Kalsi-Ryan et al., "The graded redefined assessment of strength sensibility and prehension: Reliability and validity," *J. Neurotrauma*, vol. 29, no. 5, pp. 905–914, Mar. 2012, doi: 10.1089/neu.2010.1504.

[4] C. E. Lang et al., "Improvement in the capacity for activity versus improvement in performance of activity in daily life during outpatient rehabilitation," *J. Neurologic Phys. Therapy*, vol. 47, no. 1, pp. 16–25, Jan. 2023, doi: 10.1097/npt.0000000000000413.

[5] M. Itzkovich et al., "SCIM III (Spinal cord independence measure version III): Reliability of assessment by interview and comparison with assessment by observation," *Spinal Cord*, vol. 56, no. 1, pp. 46–51, Jan. 2018, doi: 10.1038/sc.2017.97.

[6] N. M. Bradburn, L. J. Rips, and S. K. Shevell, "Answering autobiographical questions: The impact of memory and inference on surveys," *Science*, vol. 236, no. 4798, pp. 157–161, Apr. 1987, doi: 10.1126/science.3563494.

[7] S. A. Adams, "The effect of social desirability and social approval on self-reports of physical activity," *Amer. J. Epidemiology*, vol. 161, no. 4, pp. 389–398, Feb. 2005, doi: 10.1093/aje/kwi054.

[8] C. Craven et al., *Rehabilitation Environmental Scan Atlas: Capturing Capacity in Canadian SCI Rehabilitation*. Vancouver, BC, Canada: Rick Hansen Institute, 2012.

[9] J. Likitlersuang, E. R. Sumitro, T. Cao, R. J. Visée, S. Kalsi-Ryan, and J. Zariffa, "Egocentric video: A new tool for capturing hand use of individuals with spinal cord injury at home," *J. NeuroEng. Rehabil.*, vol. 16, no. 1, p. 83, Dec. 2019, doi: 10.1186/s12984-019-0557-1.

[10] M. Noorkõiv, H. Rodgers, and C. I. Price, "Accelerometer measurement of upper extremity movement after stroke: A systematic review of clinical studies," *J. NeuroEng. Rehabil.*, vol. 11, no. 1, p. 144, 2014, doi: 10.1186/1743-0003-11-144.

[11] P. S. Lum et al., "Improving accelerometry-based measurement of functional use of the upper extremity after stroke: Machine learning versus counts threshold method," *Neurorehabilitation Neural Repair*, vol. 34, no. 12, pp. 1078–1087, Dec. 2020, doi: 10.1177/1545968320962483.

[12] M. Brogioli et al., "Novel sensor technology to assess independence and limb-use laterality in cervical spinal cord injury," *J. Neurotrauma*, vol. 33, no. 21, pp. 1950–1957, Nov. 2016, doi: 10.1089/neu.2015.4362.

[13] W. L. Popp et al., "Wearable sensors in ambulatory individuals with a spinal cord injury: From energy expenditure estimation to activity recommendations," *Frontiers Neurol.*, vol. 10, p. 1092, Nov. 2019, doi: 10.3389/fneur.2019.01092.

[14] P.-F. Xu, Z.-X. Liu, F. Li, and H.-P. Wang, "A low-cost wearable hand gesture detecting system based on IMU and convolutional neural network," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 6999–7002, doi: 10.1109/EMBC46164.2021.9630686.

[15] J. Connolly, J. Condell, B. O'Flynn, J. T. Sanchez, and P. Gardiner, "IMU sensor-based electronic goniometric glove (iSEG-Glove) for clinical finger movement analysis," *IEEE Sensors J.*, vol. 18, no. 3, pp. 1273–1281, Nov. 2017, doi: 10.1109/jsen.2017.2776262.

[16] S. I. Lee, X. Liu, S. Rajan, N. Ramasarma, E. K. Choe, and P. Bonato, "A novel upper-limb function measure derived from finger-worn sensor data collected in a free-living setting," *PLoS One*, vol. 14, no. 3, Mar. 2019, Art. no. e0212484, doi: 10.1371/journal.pone.0212484.

[17] J. Likitlersuang, R. J. Visée, S. Kalsi-Ryan, and J. Zariffa, "Capturing hand use of individuals with spinal cord injury at home using egocentric video: A feasibility study," *Spinal Cord Ser. Cases*, vol. 7, no. 1, p. 17, Mar. 2021, doi: 10.1038/s41394-021-00382-w.

[18] A. Bandini, M. Dousty, S. L. Hitzig, B. C. Craven, S. Kalsi-Ryan, and J. Zariffa, "Measuring hand use in the home after cervical spinal cord injury using egocentric video," *J. Neurotrauma*, vol. 39, nos. 23–24, pp. 1697–1707, Dec. 2022, doi: 10.1089/neu.2022.0156.

[19] M. Dousty and J. Zariffa, "Tenodesis grasp detection in egocentric video," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1463–1470, May 2021, doi: 10.1109/JBHI.2020.3003643.

[20] M. Dousty, D. J. Fleet, and J. Zariffa, "Hand grasp classification in egocentric video after cervical spinal cord injury," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 2, pp. 1–11, Apr. 2023, doi: 10.1109/JBHI.2023.3269692.

[21] C Boop et al., "Occupational therapy practice framework: Domain and process-fourth edition," *Am. J. Occup. Ther.*, vol. 74, no. 2, Aug. 2020, Art. no. 7412410010, doi: 10.5014/ajot.2020.74S2001.

[22] R. J. Marino et al., "Development of an objective test of upper-limb function in tetraplegia: The capabilities of upper extremity test," *Amer. J. Phys. Med. Rehabil.*, vol. 91, no. 6, pp. 478–486, Jun. 2012, doi: 10.1097/phm.0b013e31824fa6cc.

[23] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The GRASP taxonomy of human grasp types," *IEEE Trans. Hum.-Mach. Syst.*, vol. 46, no. 1, pp. 66–77, Feb. 2016, doi: 10.1109/THMS.2015.2470657.

[24] *Graded and Redefined Assessment of Strength, Sensation and Prehension Version 1 Manual*, Neural Outcomes Consulting, Toronto, ON, Canada, 2019.

[25] R. Rupp et al., "International standards for neurological classification of spinal cord injury," *Topics Spinal Cord Injury Rehabil.*, vol. 27, no. 2, pp. 1–22, Mar. 2021, doi: 10.46292/sci2702-1.

[26] W. Kay et al., "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[27] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6201–6210.

[28] H. Fan et al., "Multiscale vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6804–6815.

[29] Y. Li et al., "MViTv2: Improved multiscale vision transformers for classification and detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4794–4804.

[30] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14648–14658.

[31] D. Dima et al., "Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100," *Int. J. Comput. Vis.*, pp. 1–23, 2022.

[32] R. Goyal et al., "The 'something something' video database for learning and evaluating visual common sense," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5843–5851.

[33] L. Wang et al., "Temporal segment networks: Towards good practices for deep action recognition," 2016, *arXiv:1608.00859*.

[34] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognit. Lett.*, vol. 140, pp. 325–331, Dec. 2020, doi: 10.1016/j.patrec.2020.11.008.

[35] X. Shi, W. Cao, and S. Raschka, "Deep neural networks for rank-consistent ordinal regression based on conditional probabilities," *Pattern Anal. Appl.*, vol. 26, no. 3, pp. 941–955, Aug. 2023, doi: 10.1007/s10044-023-01181-9.

[36] L. G. Portney, M. P. Watkins, *Foundations of Clinical Research: Applications to Practice*, vol. 892. Upper Saddle River, NJ, USA: Prentice-Hall, 2009.

[37] K. Grauman et al., "Ego4D: Around the world in 3,000 hours of egocentric video," 2021, *arXiv:2110.07058*.

[38] C. Gu et al., "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6047–6056.

[39] A. Kirillov et al., "Segment anything," 2023, *arXiv:2304.02643*.

[40] C. Feichtenhofer, H. Fan, Y. Li, and K. He, "Masked autoencoders as spatiotemporal learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 35946–35958.