technologies

Article

Self-Supervised Human Activity Representation for Embodied Cognition Assessment

Mohammad Zaki Zadeh *, Ashwin Ramesh Babu, Ashish Jaiswal and Fillia Makedon

Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA; ashwin.rameshbabu@mavs.uta.edu (A.R.B.); ashish.jaiswal@mavs.uta.edu (A.J.); makedon@uta.edu (F.M.)
* Correspondence: mohammad.zakizadehgharie@mavs.uta.edu

Abstract: Physical activities, according to the embodied cognition theory, are an important manifestation of cognitive functions. As a result, in this paper, the Activate Test of Embodied Cognition (ATEC) system is proposed to assess various cognitive measures. It consists of physical exercises with different variations and difficulty levels designed to provide assessment of executive and motor functions. This work focuses on obtaining human activity representation from recorded videos of ATEC tasks in order to automatically assess embodied cognition performance. A self-supervised approach is employed in this work that can exploit a small set of annotated data to obtain an effective human activity representation. The performance of different self-supervised approaches along with a supervised method are investigated for automated cognitive assessment of children performing ATEC tasks. The results show that the supervised learning approach performance decreases as the training set becomes smaller, whereas the self-supervised methods maintain their performance by taking advantage of unlabeled data.

Keywords: embodied cognition; cognitive assessment; computer vision; self-supervised learning

Citation: Zaki Zadeh, M.; Ramesh Babu, A.; Jaiswal, A.; Makedon, F. Self-Supervised Human Activity Representation for Embodied Cognition Assessment. *Technologies* **2022**, *10*, 33. https://doi.org/10.3390/technologies10010033

Academic Editor: Fabrizio Stasolla

Received: 1 December 2021
Accepted: 14 February 2022
Published: 17 February 2022

1. Introduction

Executive functions are higher-order mental processes that allows us to successfully plan, multitask, focus, remember instructions, and coordinate, forming the foundation of cognitive development. They mostly rely on brain functions such as working memory, mental flexibility, motor skill, etc. One of the most important skills that humans learn during their childhood is motor skill. This generally involves movements of the large muscles in the arms, legs, and torso. Children affected with neurological conditions such as attention deficit hyperactivity disorder (ADHD) exhibit motor abnormalities [1,2], especially when it comes to balance. When such impairments are not treated in a timely manner, they can have an impact on a person's daily activities. Embodied cognition is the theory that many characteristics of cognition are shaped by the entire body of the organism. High-level mental constructs such as concepts, as well as performance in various cognitive tasks such as reasoning and judgment, are among the cognition aspects. The motor system, the perceptual system, physical interactions with the environment, and the world assumptions built into the organism's structure are all examples of physical features.

The NIH toolbox is a standardized set of tests for cognitive assessment that empowers automated assessment through sensors and mobile applications. The overall goal of this work is to build a low-cost automated assessment system that uses computer vision to analyse participants performing the task and score them on the basis of standard cognitive measures such as gait, balance, bilateral coordination, and response inhibition [3]. Developing an automated system to assess such disorders will aid in closing the gap between detecting clinical phenomena and taking action by utilizing data to trigger, tailor, and deliver personalized digital treatment or prevention interventions. This system should be designed with user equity and privacy in mind.

Due to the immense effort required to manually annotate millions of data samples, the supervised method of learning features from annotated data has nearly reached its limit. This is because most modern supervised computer vision systems attempt to learn some type of image representation by searching large datasets for a pattern between data points and their annotations. Self-supervised methods are at the forefront of efforts to adapt deep learning methods to learn feature representations without costly annotations. In other words, the data itself provides the supervision in self-supervised learning [4]. Contrastive learning (CL) is a self-supervised learning approach for grouping similar samples together and separating dissimilar samples. Its goal is to train a model to distinguish between positive and negative samples. As a result, the model learns input representations that it can use in downstream tasks such as activity recognition or object detection [4–6]. Along with state-of-the-art contrastive methods, this article employs a new family of self-supervised methods that do not require a large number of negative samples and are thus easier to train [6–8].

In this work (Figure 1), the focus of the automated assessment system is on three core tasks: bi-manual ball pass, ball drop to the beat, and tandem gait forward. These tasks are part of a larger system called ATEC—Activated Test of Embodied Cognition [3,9,10], which is described in detail in Section 3. In order to automatically evaluate a subject's performance, first the VIBE [11] human pose estimation system was used to extract 3D-body key-points. Then, a deep learning-based model was trained to classify subject actions. Furthermore, in order to improve the accuracy of the system, the model was pre-trained on the NTU-RGB+D 120 dataset [12,13] and then fine-tuned on our ATEC dataset [3]. Three different state-of-the-art self-supervised learning methods, including those from MoCo [6], SimSiam [14], and VICReg [8], were employed to pre-train the model in a self-supervised manner, and their performances were compared to a supervised learning approach. The results show that a pre-trained model can outperform a supervised learning approach when a small amount of annotated data is available for training.

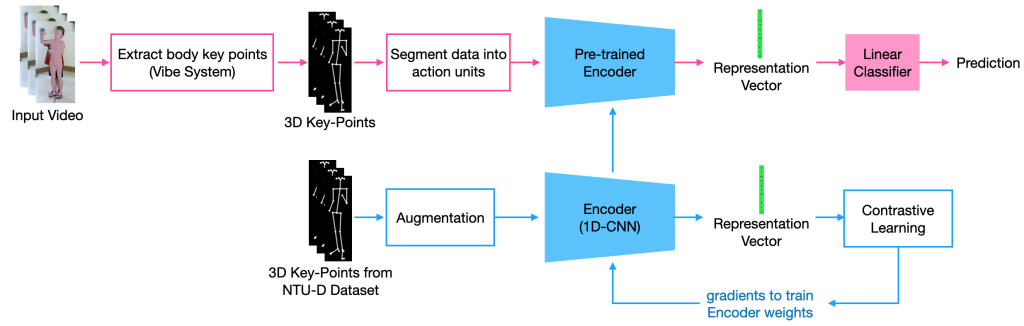

Figure 1. Proposed method architecture: **top** (pink)—supervised classification; **bottom** (blue)—self-supervised pre-training.

The first major contribution of this work is the creation of a dataset containing manually-annotated recorded videos of children for three core ATEC tasks. The videos were then classified using a supervised learning approach into action classes that can be used for cognitive assessments. Furthermore, self-supervised learning methods were used to improve the classifier's efficacy while reducing its reliance on annotated data. It should be mentioned that all of the self-supervised methods used in this work were originally designed to extract features from still images, so we adapted them to extract representations from a sequence of human body key-points. The rest of the paper is structured as follows: Section 2 discusses the related work, Section 3 explains the ATEC system along with presenting information about the recorded dataset, Section 4 describes the proposed methods employed in this work and discusses the results, and finally, the conclusion and future works are mentioned in Section 5.

2. Related Works

Many studies have been undertaken to better understand the relationship between cognitive deficiencies and various psychiatric neuro-developmental diseases, and a number of diagnostic and intervention approaches have been recommended. Traditionally, a diagnosis begins with obtaining comprehensive background information from children, parents, and school teachers, followed by trained psychologists administering standardized tests and a feedback session on performance to explain the findings and make recommendations for possible treatments or interventions. One of the most popular paper-based cognitive assessment tests is the Swanson, Nolan, and Pelham Teacher and Parent Rating Scale (SNAP). It is a 90-question self-report inventory designed to measure attention deficit hyperactivity disorder (ADHD) and oppositional defiant disorder (ODD) symptoms in children and young adults [15]. Each question essentially counts the number of times a particular symptom or behavior occurs. The survey is intended for use by children and young adults. The findings shed light on inattention, hyperactivity, impulsivity, and other factors.

Because physical activities are an important manifestation of cognitive functions [16,17], physical activities can be employed to both assess cognitive skills and to train such skills [18]. Physical activities should be incorporated into cognitive training since research shows that physical fitness and exercise in children leads to quantifiable increases in cognitive skills and academic achievement [19]. Understanding the relationship between physical manifestations of cognitive skills and other sorts of manifestations, such as responses to computer-based problem-solving activities, is still an open problem [20]. There has also been a plethora of research in recent years tackling the problem of analyzing body gait for the prediction and diagnosis of multiple disorders. In [21], machine-learning methods were used for gait assessment through the estimation of spatio-temporal parameters. They used data from inertial measurement units placed at the shank and waist. In [22], the authors explored gait analysis, which include stride length, lateral balance, and effort exerted for a particular class of activity. Although gait has clear links to motor activities, they investigated an interesting link to visual processing, since the visual system is strongly correlated with balance. In [23], the effects of human fatigue due to repetitive and physically challenging jobs that cause work-related musculoskeletal disorder (WMSD) were investigated. This study was designed to monitor fatigue through the development of a methodology that objectively classifies an individual's level of fatigue in the workplace by utilizing the motion sensors embedded in smartphones.

Human activity recognition (HAR) methods use image or body key-point data to model the spatial and temporal features of each class-action [24,25]. Action recognition often involves classifying high-level events with more variation between classes. Researcher have observed [26,27] that many image-based action recognition datasets feature low variability among actions. Thus, body key-points are the most salient high-level features for datasets with high intra-class variance. Recent body pose estimation methods have shown excellent results on benchmarks featuring multiple persons with varying viewpoints and lighting [11,28]. Hand pose estimation is another active area of research in computer vision, with many applications in human–computer interaction and sign language recognition. To reduce the burden of labeled-data acquisition, the authors in [29] proposed a data-driven method for hand part segmentation on depth maps that does not require any additional effort to obtain segmentation labels. The proposed method learns to estimate the hand shape and pose given a depth map by using labels already provided by public datasets in terms of major 3D hand joint locations. In [30], the authors proposed a two-stage pipeline for accurately localizing fingertip position in depth images, even in varying lighting and severe self occlusion. Given a real depth image, a cycle-consistent generative adversarial network (Cycle-GAN) is used to apply unpaired image-to-image translation and generate a depth image with colored predictions on the fingertips, wrist, and palm. Early attempts to build an automated scoring system for ATEC employed supervised learning, which required careful annotation of recorded data by experts. The authors of [9] tried to evaluate children's performance in the ball-drop-to-the-beat and sailor-step tasks by measuring

the distance between relevant body key-points such as wrists and heels. In [10], different deep learning methods were utilized to classify different action categories in the finger opposition task. In another work, [31] a combination of the following three modalities was proposed for HAR: optical flow, objects in the scene, and human poses (skeletal information). Furthermore, an attention-based approach was proposed to combine those three modalities. In order to decrease the reliance of automated systems on manually-annotated datasets, different self supervised learning approaches were developed. In [32], generative adversarial networks (GAN) were augmented with a self-supervised task in order to improve their ability to encode video representations. Furthermore, a contrastive learning framework was employed in [33] to pre-train a classifier in a self-supervised manner for assessment of a subject's performance in tandem gait task. This work is continuation of aforementioned self-supervised learning approaches for tackling more ATEC tasks.

Learning from small datasets is highly difficult and, as a result, remains largely unsolved. Due to the difficulties of generalizing to novel instances, only a few studies have attempted to address the problem of training deep architectures with a small number of examples [34]. The authors of [35] employed a linear program to partition fitted neural networks into ensembles of low-bias sub-networks via empirical decomposition. More recently, Ref. [36] proposed the use of neural tangent kernel (NTK) designs in low-data problems and found that they outperformed all other classifiers. Another proposed approach is to expand the dataset size and, as a result, increase the performance of training generative models such as generative adversarial networks [37,38]. However, in the small sample domain, training a generative model may be computationally demanding or create severe obstacles. Ref. [39] proposed using cosine loss to prevent overfitting, arguing that the cosine loss's $L2$ normalization is a strong, hyper-parameter-free regularizer when sample availability is limited. Self-supervised learning is used in this study to learn feature representations from massive unlabeled datasets found on the internet and to then transfer the acquired features to a target domain that is closely related via transfer learning.

3. ATEC Dataset

The ATEC system has been proposed in order to examine multiple cognitive parameters of children such as working memory, reaction inhibition, and coordination using physical exercises [3,9,10,31]. This system was created to allow both professionals and non-experts to handle it with ease. The ATEC system features a recording and administration interface that was created to keep the assessments running smoothly. Because sensor-based data collection is more expensive and impractical with children, this system simply records video data.

The participants' front and side views were recorded using two Microsoft Kinect V2 cameras. RGB, depth, audio, and skeleton data were all recorded. Figure 2 represents the data collection setup. The recording modules were linked to an Android-based administrative interface that manages the assessment flow and allows the administrator to choose between tasks. To guarantee that the subjects understand the rules of the activity, each task comprises an instructional film as well as practice videos [3,9]. The goal is to develop intelligent software that will allow instructors to view automated system prediction and performance visualizations alongside recorded video of participants, as shown in Figure 3.

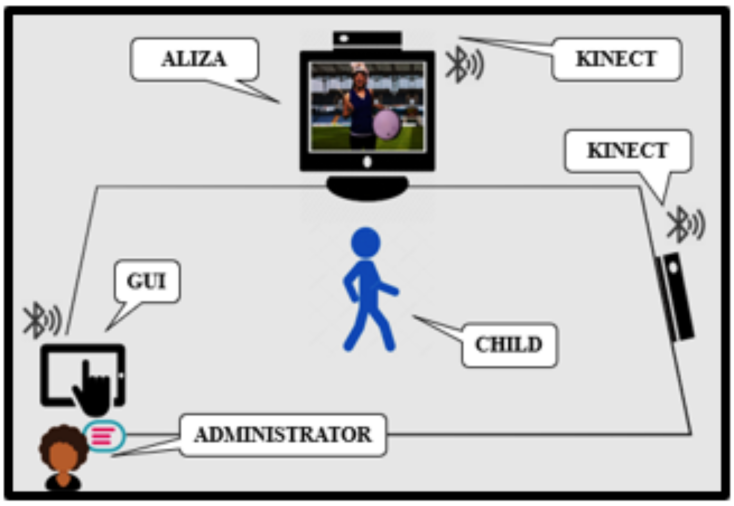

Figure 2. hlData collection setup.

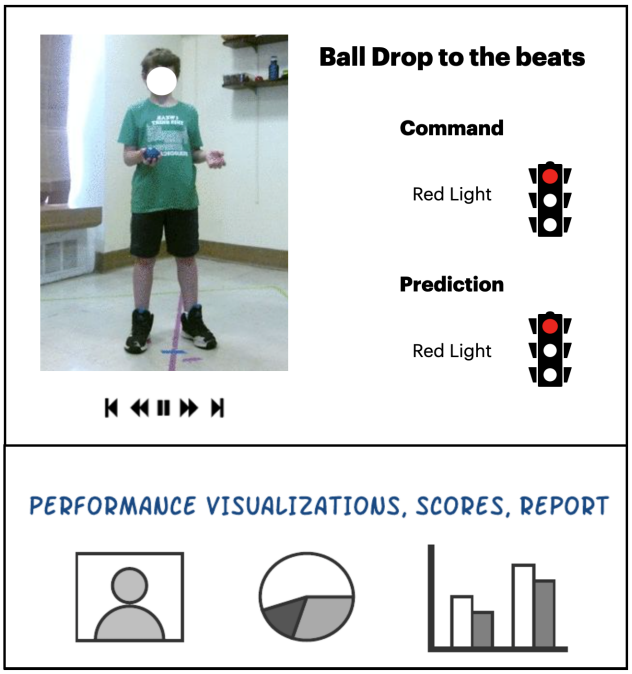

Figure 3. ATEC system GUI.

The instructional video provides a brief demonstration of how the current task should be executed. The recording modules activate once the evaluation is started, and Aliza, the on-screen instructor, assists the students through each task. Annotation software was also created to allow computer scientists and cognitive experts to visualize and annotate the data. An expert examined each recording of the assessment based on a set of task-specific criteria. The automated scoring system was then evaluated using this expert annotation as a baseline [3].

Children aged 5 to 11 (mean (sd) = 8.04 (1.36)) were recruited from the community (N = 55) through local public schools and through fliers displayed on bulletin boards. In accordance with procedures approved by the University's Institutional Review Boards (IRB), their parents provided written informed consent, and the children offered verbal assent. Although all of the children were in regular classes, nine (16.4 percent) of them received additional services through a 504-plan approved by the school. The population was ethnically diverse (56.4 percent Caucasian) and 58.2% male [3].

Before the testing procedure, the parents were required to complete pre-screening paperwork that collected information about the history of the children and family. This pre-screening was followed by paper0based assessment tests such as the Child Behavior Checklist (CBCL) [40], the Social Responsiveness Scale; the Swanson, Nolan and Pelham questionnaire [15]; etc. Then participants were requested to take standard computer tests from the NIH toolbox such as the Flanker and Working Memory Test (WMT) [41] to gauge various cognitive measures such as attention, response inhibition, etc. Finally, the children performed all the tasks from the ATEC program in two trials one week apart.

The ATEC was created to concretely quantify embodied cognition, a concept with broad acceptance, but limited consensus on its precise meaning. The ATEC results show that it has contemporaneous validity with traditional neuro-psychological and parent-reported executive functioning (EF) measures. The ATEC has demonstrated discriminant validity between children at risk for EF-related impairments and children who are not, and it has a strong association with CBCL parent-rated assessment of real-world functioning [3].

ATEC total score alone explains a large amount of variance in CBCL parent-rated functioning according to multiple regression analysis, and no other neuro-psychological measure contributed significantly to the model. None of the other measures entered the model when the ATEC total score was not included. This conclusion could imply that a measure of cognition in action is more closely related to functioning than traditional assessments that do not entail movement [3]. The researchers discovered high test–retest reliability with acceptable practice effects, as well as an expected moderate connection with age, implying that embodied cognition is linked to normal development. Bell et al. [3] provided a more detailed examination of the ATEC system.

The ATEC system consists of 17 physical exercises with different variations and difficulty levels, designed to provide assessment of executive and motor functions including sustained attention, self-regulation, working memory, response inhibition, rhythm, and coordination, as well as motor speed and balance. The measurements are converted to ATEC scores, which describe the level of development (early, middle, full development) [3]. Table 1 represents the list of all ATEC tasks that have been devised for a variety of cognitive measures. Descriptions of the ATEC tasks that have been incorporated in this work are provided below.

Table 1. ATEC tasks to assess various cognitive measures.

Category	Test
Gross Motor, Gait and Balance	Natural Walk, Gait on Toes, Tandem Gait, Stand Arms Outstretched, Stand on One Foot
Synchronous Movements	March Slow, March Fast
Bilateral Coordination and Response Inhibition	Bi-Manual Ball Pass with Green, Red, and Yellow Light
Visual Response Inhibition	Sailor Step Slow, Sailor Step Fast
Cross Body Game	Cross your Body (Ears, Shoulders, Hips, Knees)
Finger-Nose Coordination	Hand Eye Coordination
Rapid Sequential Movements	Foot Tap, Foot-Heel, Toe Tap, Hand Pat, Finger Tap, Oppose Finger Succession

3.1. Ball Drop to the Beat

Bilateral coordination is defined as the ability to coordinate both sides of the body at the same time in a controlled and orderly manner. Bilateral coordination suggests that both sides of the brain are working together successfully. Another important component of cognitive functions is attention. The ability to focus on a certain aspect of information while ignoring other perceptible information is characterized as attention. Similarly, response

inhibition (inhibitory control) is an executive function that allows an individual to regulate their natural or habitual dominant behavioral responses to stimuli by inhibiting their impulses. It enables the individual to adopt a more appropriate conduct that is compatible with their objectives [3].

Ball drop to the beat is a core ATEC task devised to evaluate both audio and visual cue processing while performing upper-body movements. In this task, the participant is required to pass a ball from one hand to the other while following verbal and visual instructions. According to the rules, the participants are required to pass the ball for the green light (pass), keep the ball still in their hand for the red light (No Pass), and move the ball up and down with the same hand for the yellow light (raise). The light colors are presented both audibly and visually to gauge both audio and visual accuracy and response inhibition. The task is assessed at 60 beats per minute (slow trial) and 100 beats per minute (fast trial) for a total of 16 counts for each trial [9]. Examples of ball-drop are presented in Figure 4.

Figure 4. Samples from the Ball drop to the beat task. Each row represents a specific action carried out by the children. (**a**) Ball Pass (**b**) Hand raise (**c**) No pass [9].

Apart from accuracy and response inhibition, ATEC exercises also measure rhythm. During the test, the ATEC on-screen host, Aliza, announces the stimuli in a rhythmic manner by saying green, red or yellow light in two beats: the first beat for the color and the second for the word light. Thus, the subjects are required to perform the actions in two beats. For instance, for the green light (Pass) and yellow light (Raise) commands, the ball is raised on the first beat and either passed or lowered on the second beat. Figure 5 illustrates both the audio and visual stimuli. Each segment (activity) in this diagram is divided by red lines, and each segment contains two beats separated by green lines [9].

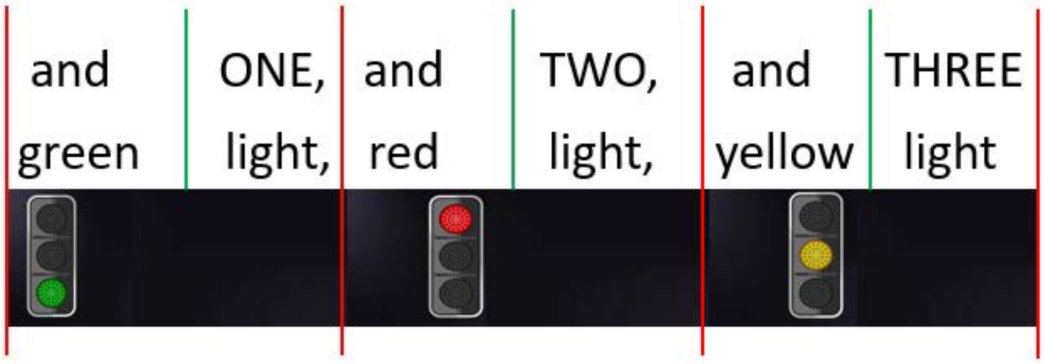

Figure 5. Audio-visual stimuli during the ball-drop-to-the-beat task [9].

3.2. Tandem Gait Forward

In this task, the participants are asked to walk in a straight line where, for every step, the heal of the foot moving forward is expected to touch the toes of the foot behind it. The subject's score is calculated as the total number of correct steps performed out of the total number of 8 expected steps [3,33]. An example of a valid and an invalid step are presented in Figure 6. In these figures, children's body are covered by their estimated SMPL body mesh in order to see VIBE system [11] body pose estimation in action and also to protect their privacy.

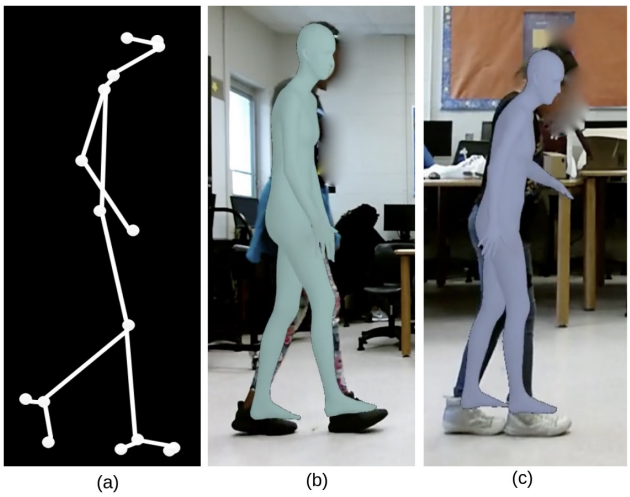

Figure 6. Example of tandem gait task: (**a**) skeleton key points, (**b**) an invalid step, (**c**) a valid step. In a valid step, the heal of the foot moving forward is expected to touch the toes of the leg behind.

3.3. Stand on One Foot

Another task for measuring gross motor, gait and balance is stand on one foot. In this task, the participants are expected to stand on one foot for 10 s. Participants are scored based on their capability to sustain the position for a given period of time. Scores are determined based on the number of seconds the participant can stand without stopping. In the first round, subjects stand on their left foot and in the second round, they stand on their right foot [3]. An example of a participant standing on her right foot is depicted in Figure 7.

Figure 7. Example of stand on one foot.

4. Methodology and Results

4.1. Self-Supervised Learning

Recent advances in deep learning [42] and the challenge of gathering huge amounts of labeled data have encouraged new research in unsupervised or self-supervised learning. In particular, computer vision tasks could greatly benefit from successful models that learn abstract low-dimensional features of images and videos without any supervision, because unlabeled images and video sequences can be gathered automatically without human intervention [43,44]. As a result, significant research effort has been focused on methods that can adapt to new conditions without expensive human supervision. The main focus of this chapter is applying self-supervised visual representation learning for human activity recognition in ATEC-system recorded videos. Self-supervised learning techniques comprising both generative [37] and contrastive approaches [4] have produced state-of-the-art low-dimensional representations on most computer vision benchmarks [6,45–47].

One of the most popular self-supervised approaches is contrastive learning (CL) [4,6]. CL tries to group similar (positive) samples closer and diverse (negative) samples further from each other. Representations are obtained by feeding input data into an encoder network. Contrastive learning focuses on comparing the representations with a variant of the noise contrastive estimation function [48] called InfoNCE [49], which is defined as follows:

$$L = -log\frac{exp(sim(q,k_+)/\tau)}{exp(sim(q,k_+)/\tau) + \sum_{i=0}^{K} exp(sim(q,k_i)/\tau)} \quad (1)$$

where q is the original sample (query), k_+ represents a positive sample, and k_i represents a negative sample. τ is a hyper-parameter used in most of the recent methods and is called the temperature coefficient. The *sim* function can be any similarity function, but generally a cosine similarity is used. The cosine similarity of two vectors is defined as the cosine of the angle between them. The initial idea behind noise contrastive estimation was to perform a non-linear logistic regression that discriminates between observed data and some artificially-generated noise [4].

Since the number of negative samples affects the performance of CL methods [4], different strategies are used for selecting a large number of negative samples. In the first contrastive learning methods, a large batch size is used and all the samples in the batch except for the query and one positive sample are considered as negative. Because large batch sizes inversely affect optimization during training, one possible solution is to maintain a separate dictionary known as a memory bank containing representations of negative samples. However, since maintaining a memory bank during training is complicated, the memory bank can be replaced by a momentum encoder. The momentum encoder (MoCo) (Figure 8 left) generates a dictionary as a queue of encoded samples, with the current mini-batch enqueued and the oldest mini-batch dequeued [6]. The momentum encoder shares the same parameters as the query encoder (θ_q) and its parameters (θ_k) are updated based on the parameters of the query encoder ($\theta_k = m\theta_k + (1-m)\theta_q, m \in [0,1)$: momentum coefficient).

Most self-supervised methods involve specific forms of Siamese networks [50]. Weight-sharing neural networks with two or more inputs are known as Siamese networks. All outputs collapsing to a constant are undesirable trivial solutions to Siamese networks. There have been various general solutions for preventing the collapse of Siamese networks. The SimSiam method [14] proposed by Chen et al. prevents collapsing solutions by directly maximizing the similarity of an image's two views, using neither negative pairs [47], nor a momentum encoder [6]. The authors argue that stop-gradient operation is critical in preventing collapsing solutions. The SimSiam method architecture is depicted in Figure 8 (middle).

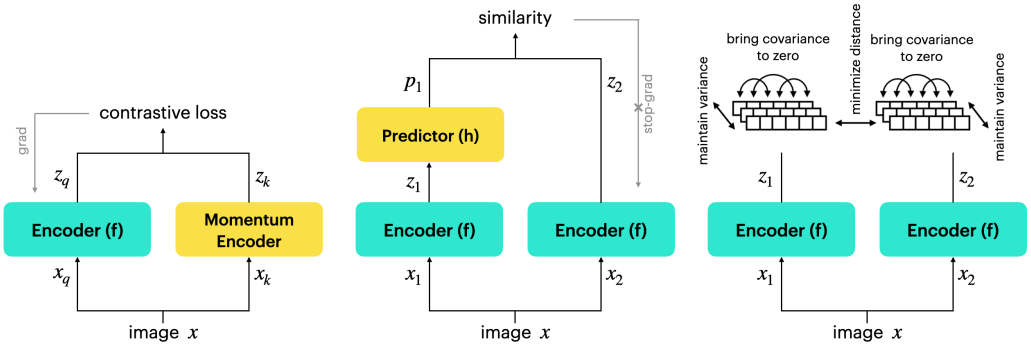

Figure 8. Different self-supervised learning architectures (x_1 and x_2 stand for different augmentations of image x): left—MoCo [6]; middle—SimSiam [14]; right—VICReg [8].

SimSiam methods take as input two randomly augmented views x_1 and x_2 from an image x. The two views are processed by an encoder network f. The encoder f shares weights between the two views. A prediction MLP network h matches the output of one view to the output of another view. The negative cosine similarity of two output vectors $p_1 = h(f(x_1))$ and $z_2 = f(x_2)$ is defined as follows:

$$D(p_1, z_2) = -\frac{p_1}{||p_1||_2} . \frac{z_2}{||z_2||_2} \tag{2}$$

where $||.||_2$ is l2-norm. Finally, the final loss function is defined below:

$$L = \frac{1}{2} D(p1, stopgrad(z2)) + \frac{1}{2} D(p2, stopgrad(z1)) \tag{3}$$

In the first term, the encoder on x_2 receives no gradient from z_2, but in the second term, it receives gradients from p_2 (and vice versa for x_1) [14]. The loss is calculated for each sample, and the total loss is averaged across all samples.

VICReg (Variance-Invariance-Covariance Regularization) proposed by Bardes et al. [8] is another self-supervised method tackling the collapsing solution problem. The VICReg architecture illustrated in Figure 8 (right) is symmetric and is based on three simple principles: variance, invariance, and covariance. The variance principle is a simple but efficient strategy for preventing collapse by constraining the variance of the representations along each dimension independently. Without requiring any negative pairs, the invariance principle learns invariance to various views of an image employing a standard mean-squared Euclidean distance. Finally, the covariance principle uses the Barlow twins' covariance criterion [51], which decorrelates the different dimensions of learned representations with the goal of spreading information across dimensions and avoiding dimension collapse [8].

In the VICReg method, given an image x, two augmented views x_1 and x_2 are encoded using the encoder network f into representations $z_1 = f(x_1)$ and $z_2 = f(x_2)$. The overall loss function is a weighted average of the invariance, variance, and covariance terms [8]:

$$L(z_1, z_2) = S(z_1, z_2) + \lambda(V(z_1) + V(z_2)) + \gamma(C(z_1) + C(z_2)) \tag{4}$$

where λ and γ are hyper-parameters that regulate how important each phrase in the loss is. The overall objective function is computed as the sum of the loss function taken on all samples in the dataset. The variance, invariance, and covariance terms that make up the loss function are described here. First the variance term is defined as follows:

$$V(z) = \frac{1}{d} \sum_{i=1}^{d} max(0, 1 - \sqrt{Var(z) + \epsilon}) \tag{5}$$

where d is the dimension of feature vector z, ϵ is a small scalar for avoiding numerical instabilities, and $Var(x)$ is the unbiased variance estimator.

Inspired by the Barlow twins [51], the covariance regularization term C is defined as the sum of the squared off-diagonal coefficients of covariance matrix of z ($Cov(z)$), with a factor $1/d$ that scales the criterion as a function of the dimension:

$$C(z) = \frac{1}{d}\sum_{i \neq j} Cov(z)_{i,j}^2 \tag{6}$$

This term makes the off-diagonal coefficients close to 0 in order to decorrelate the different dimensions of the projections and prevent them from encoding the same information. Finally, the invariance criteria S between Z_1 and Z_2 is determined as the mean-squared Euclidean distance between each pair of vectors:

$$S(z_1, z_2) = \frac{1}{n}\sum_{i} ||z_{1i} - z_{2i}||_2^2 \tag{7}$$

4.2. Proposed System

The architecture of the proposed computer vision system is depicted in Figure 1. First, the subject's 3D body key-points were extracted using the VIBE system [11]. VIBE (Video Inference for Body Pose and Shape Estimation) is a video pose and shape estimation method that predicts the parameters of SMPL body models for each frame of an input video. From these key-points, 17 of them including head, hands, hip, feet, and toes were selected. Finally, the extracted data were divided into equal segments (with overlap), with each segment corresponding to an action. The numbers of segments for each task were as follows: ball drop to the beat—16; tandem gait forward—8; and stand on one foot—10. Each segment ($X \in \mathbb{R}^{32\times51}$) included 32 samples with 51 features. The features were x,y,z coordinates for each of the 17 key-points rasterized into one vector. Then, the input was fed into an encoder network to obtain the compact representation $z \in \mathbb{R}^{256}$. Finally, a linear classifier was used to classify input segments into action classes according to each task (Figure 1).

In order to pre-train the classifier model, the publicly available NTU-RGB+D 120 [12,13] was used. This dataset contains 120 action classes and 114,480 video samples. In this work, only 3D skeletal data were employed. Similar to the gait dataset, 17 equivalent key-points (head, hands, hip, feet, and toes) were selected. In this work, a four-layer 1D convolutional neural network (CNN) with one penultimate transformer layer [52] was used as the encoder network. For all self-supervised methods, a projector network consisting of three fully connected layers was used. The projector network mapped the representations to a higher dimension of 1024. The SimSiam method also incorporates a two-layer fully connected predictor network that acts as a bottleneck by decreasing dimension of feature vectors to 256 and increasing it back to 1024. All networks used in this work employed batch normalization, except for the last layer.

4.3. Results and Discussion

All of the models used in this work were trained using the Pytorch framework [53] for 200 epochs. Stochastic gradient descent (SGD) was employed as an optimizer with a batch size of 512, learning rate of 0.1, and weight decay of 1×10^{-4}. The learning rate followed a cosine decay schedule [54] with 10 warm-up epochs. Furthermore, for the contrastive learning method MoCo, the temperature hyper-parameter τ and momentum coefficient μ were chosen as 0.1 and 0.999, respectively. For the VICReg method parameters, λ and γ were chosen as 1.0 and 0.04, respectively.

For evaluating the performance of the proposed methods in the case of a small amount of annotated data, three scenarios were defined. In the first scenario, 50% of the data were used for training and the other 50% for testing. In the second scenario, 25% of the data were used for training and the remaining 75% for testing. Finally, for the third scenario,

10% of data was used for training and the remaining 90% was used for testing. The average classification accuracy was calculated by cross-validation. The results for the baseline supervised method are shown in the first row of Table 2.

Table 2. Different methods' top-1 classification accuracy for different training/test splits: 50%—using 50% of the dataset for training and the remaining 50% for testing; 25%—using 25% of the dataset for training and the remaining 75% for testing; 10%—using 10% of the dataset for training and the remaining 90% for testing.

Approach	Ball Drop to the Beat			Tandem Gait			Stand on One Foot		
	50%	**25%**	**10%**	**50%**	**25%**	**10%**	**50%**	**25%**	**10%**
Supervised	**77.61**	61.98	54.79	74.81	60.21	51.72	88.98	79.36	76.93
Multimodal	75.29	52.77	48.13	69.52	55.46	51.11	87.82	75.55	71.96
MoCo	74.70	71.87	70.63	75.52	73.70	72.36	89.63	88.07	85.59
SimSiam	75.44	74.08	71.91	75.81	74.42	73.46	89.76	88.05	87.24
VICReg	77.11	**74.65**	**72.59**	**75.96**	**74.45**	**73.51**	**90.54**	**89.86**	**89.59**

It is clear from the results that the baseline method classification accuracy decreases as training set becomes smaller, whereas the self-supervised methods maintain their performance and even outperform the supervised methods. We also compared the proposed methods to a supervised multi-modal approach that had previously been used successfully on the ball-drop-to-the-beat task [31]. Results show that among self-supervised methods, VICReg attains the highest overall accuracy. When the size of the training set was reduced from 50% of the total dataset to 10% in the ball-drop-to-the-beat task, the accuracy of the supervised approach dropped by about 23%, whereas the accuracy of the VICReg method dropped by about 5%. In the tandem gait task, when the size of the training set decreased from 50% to 10% of the total dataset, the accuracy of the supervised approach declined by around 23%, while the VICReg technique only dropped by about 3%. In the stand-on-one-foot task, the supervised approach's accuracy was reduced by roughly 12% when the training set was decreased from 50% to 10% of the whole dataset, whereas the VICReg technique's performance was lowered by just around 1%. One reason for the multimodal approach's poor performance is that it uses a more complicated model that includes optical flow and object location in addition to the human body skeleton, making it prone to overfitting in cases with small training datasets.

5. Conclusions and Future Works

In this work, the technical description of ATEC, an integrated computer-vision system for assessing embodied cognition in children with executive function disorder, is presented. The ATEC system includes both recording and administrative interfaces, which were designed to streamline the assessments without any interruptions. This system only records video data, since sensor-based data collection can be more expensive and impractical with child participants. The ATEC system consists of variety of physical exercises with different variations and difficulty levels designed to provide assessment of executive and motor functions. The main focus of this work was applying self-supervised visual representation learning for human activity recognition in ATEC system recorded videos. Finding an effective human activity representation will help us to improve the accuracy of the automated computer-vision system with much less annotated training data.

In order to improve the performance of the proposed system, we tried to pre-train the encoder network on large public dataset (NTU) by using self-supervised learning. Different self-supervised methods were investigated to obtain the best representations. The results supported our claim that pre-trained models can outperform supervised learning approaches when small amounts of annotated data are available for training. When the size of the training set was reduced from 50% of the total dataset to 10% in the ATEC task, the accuracy of the supervised approach dropped by about 20%, whereas the accuracy

of the self-supervised methods dropped by less than 5%. Improving the efficacy of the proposed approach in order to deploy it in real-world applications, as well as applying it to the remaining ATEC tasks, such as sailor step [3,9], finger-oppose [3,10], etc., will be the focus of future works. The ultimate goal of this work is to create a comprehensive digital phenotyping framework capable of collecting multimodal data from a variety of sensors such as cameras, wearables, and so on, in order to monitor human behavior. Digital phenotyping will close the loop between detecting clinical phenomena and taking action by using data to trigger and deliver personalized digital treatments or prevention interventions [55].

Author Contributions: Conceptualization, M.Z.Z., A.R.B. and A.J.; writing—original draft preparation, M.Z.Z.; writing—review and editing, M.Z.Z.; visualization, M.Z.Z.; supervision, F.M.; project administration, F.M.; funding acquisition, F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by National Science Foundation grants IIS 1565328 and IIP 1719031.

Institutional Review Board Statement: Participants were recruited from the community at public schools and through posted flyers. Their parents provided written informed consent, and the children provided verbal assent in accordance with procedures approved by the University IRB.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data are not publicly available due to legal and privacy reasons.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Leitner, Y.; Barak, R.; Giladi, N.; Peretz, C.; Eshel, R.; Gruendlinger, L.; Hausdorff, J.M. Gait in attention deficit hyperactivity disorder. *J. Neurol.* **2007**, *254*, 1330–1338. [CrossRef] [PubMed]
2. Buderath, P.; Gärtner, K.; Frings, M.; Christiansen, H.; Schoch, B.; Konczak, J.; Gizewski, E.R.; Hebebrand, J.; Timmann, D. Postural and gait performance in children with attention deficit/hyperactivity disorder. *Gait Posture* **2009**, *29*, 249–254. [CrossRef] [PubMed]
3. Bell, M.D.; Weinstein, A.J.; Pittman, B.; Gorman, R.M.; Abujelala, M. The Activate Test of Embodied Cognition (ATEC): Reliability, concurrent validity and discriminant validity in a community sample of children using cognitively demanding physical tasks related to executive functioning. *Child Neuropsychol.* **2021**, 27, 973–983. [CrossRef] [PubMed]
4. Jaiswal, A.; Ramesh Babu, A.; Zadeh, M.; Banerjee, D.; Makedon, F. A Survey on Contrastive Self-Supervised Learning. *Technologies* **2020**, *9*, 2. [CrossRef]
5. Liu, X.; Zhang, F.; Hou, Z.; Wang, Z.; Mian, L.; Zhang, J.; Tang, J. Self-Supervised Learning: Generative or Contrastive. *IEEE Trans. Knowl. Data Eng.* **2020**. [CrossRef]
6. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Silver Spring, MD, USA, 14–19 June 2020.
7. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.H.; Buchatskaya, E.; Doersch, C.; Pires, B.A.; Guo, Z.D.; Azar, M.G.; et al. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
8. Bardes, A.; Ponce, J.; LeCun, Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *arXiv* **2021**, arXiv:2105.04906.
9. Dillhoff, A.; Tsiakas, K.; Babu, A.R.; Zakizadehghariehali, M.; Buchanan, B.; Bell, M.; Athitsos, V.; Makedon, F. An automated assessment system for embodied cognition in children: From motion data to executive functioning. In Proceedings of the 6th International Workshop on Sensor-Based Activity Recognition and Interaction, Rostock, Germany, 16–17 September 2019; pp. 1–6.
10. Babu, A.R.; Zakizadeh, M.; Brady, J.R.; Calderon, D.; Makedon, F. An Intelligent Action Recognition System to assess Cognitive Behavior for Executive Function Disorder. In Proceedings of the 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), Vancouver, BC, Canada, 22–26 August 2019; pp. 164–169.
11. Kocabas, M.; Athanasiou, N.; Black, M.J. VIBE: Video Inference for Human Body Pose and Shape Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Silver Spring, MD, USA, 14–19 June 2020.
12. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
13. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2684–2701. [CrossRef]
14. Chen, X.; He, K. Exploring Simple Siamese Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Silver Spring, MD, USA, 14–19 June 2020.

15. Atkins, M.S.; Pelham, W.E.; Licht, M.H. A comparison of objective classroom measures and teacher ratings of attention deficit disorder. *J. Abnorm. Child Psychol.* **1985**, *13*, 155–167. [CrossRef]
16. Donnelly, J.E.; Lambourne, K. Classroom-based physical activity, cognition, and academic achievement. *Prev. Med.* **2011**, *52*, 36–42. [CrossRef]
17. Malina, R.M.; Cumming, S.P.; Silva, M.J.C. Physical Activity and Inactivity Among Children and Adolescents: Assessment, Trends, and Correlates. In *Biological Measures of Human Experience across the Lifespan*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 67–101.
18. Dusen, D.P.V.; Kelder, S.H.; Ranjit, N.; Perry, C.L. Associations of physical fitness and academic performance among schoolchildren. *J. Sch. Health* **2011**, *81*, 733–740. [CrossRef] [PubMed]
19. Davis, C.; Cooper, S. Fitness, fatness, cognition, behavior, and academic achievement among overweight children: Do cross-sectional associations correspond to exercise trial outcomes? *Prev. Med.* **2011**, *52* (Suppl. 1), S65–S69. [CrossRef] [PubMed]
20. Hopkins, M.E.; Davis, F.C.; VanTieghem, M.R.; Whalen, P.J.; Bucci, D.J. Differential effects of acute and regular physical exercise on cognition and affect. *Neuroscience* **2012**, *215*, 59–68. [CrossRef] [PubMed]
21. Mannini, A.; Trojaniello, D.; Cereatti, A.; Sabatini, A. A Machine Learning Framework for Gait Classification Using Inertial Sensors: Application to Elderly, Post-Stroke and Huntington's Disease Patients. *Sensors* **2016**, *16*, 134. [CrossRef] [PubMed]
22. Khan, A.; Madden, J.; Snyder, K. Framework Utilizing Machine Learning to Facilitate Gait Analysis as an Indicator of Vascular Dementia. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*. [CrossRef]
23. Karvekar, S. Smartphone-Based Human Fatigue Detection in an Industrial Environment Using Gait Analysis. *Ergonomics* **2019**, *64*, 1–28.
24. Li, C.; Zhang, X.; Liao, L.; Jin, L.; Yang, W. Skeleton-Based Gesture Recognition Using Several Fully Connected Layers with Path Signature Features and Temporal Transformer Module. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
25. Ali, A.; Taylor, G.W. Real-Time End-to-End Action Detection with Two-Stream Networks. In Proceedings of the 2018 15th Conference on Computer and Robot Vision (CRV), Toronto, ON, Canada, 8–10 May 2018; pp. 31–38.
26. Zhang, J.; Li, W.; Ogunbona, P.O.; Wang, P.; Tang, C. RGB-D-Based Action Recognition Datasets: A Survey. *Pattern Recognit.* **2016**, *60*, 86–105. [CrossRef]
27. Piergiovanni, A.; Ryoo, M.S. Fine-grained Activity Recognition in Baseball Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Silver Spring, MD, USA, 14–19 June 2020.
28. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Silver Spring, MD, USA, 14–19 June 2020.
29. Rezaei, M.; Farahanipad, F.; Dillhoff, A.; Elmasri, R.; Athitsos, V. Weakly-Supervised Hand Part Segmentation from Depth Images. In Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), Corfu, Greece, 29 June–2 July 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 218–225. [CrossRef]
30. Farahanipad, F.; Rezaei, M.; Dillhoff, A.; Kamangar, F.; Athitsos, V. A Pipeline for Hand 2-D Keypoint Localization Using Unpaired Image to Image Translation. In Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), Corfu, Greece, 29 June–2 July 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 226–233. [CrossRef]
31. Ramesh Babu, A.; Zadeh, M.; Jaiswal, A.; Lueckenhoff, A.; Kyrarini, M.; Makedon, F. A Multi-Modal System to Assess Cognition in Children from Their Physical Movements; Association for Computing Machinery: New York, NY, USA, 2020. [CrossRef]
32. Zaki Zadeh, M.; Ramesh Babu, A.; Jaiswal, A.; Kyrarini, M.; Makedon, F. Self-Supervised Human Activity Recognition by Augmenting Generative Adversarial Networks. In Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), Corfu, Greece, 29 June–2 July 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 171–176. [CrossRef]
33. Zaki Zadeh, M.; Ramesh Babu, A.; Jaiswal, A.; Kyrarini, M.; Bell, M.; Makedon, F. Automated System to Measure Tandem Gait to Assess Executive Functions in Children. In Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), Corfu, Greece, 29 June–2 July 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 167–170. [CrossRef]
34. Brigato, L.; Iocchi, L. A Close Look at Deep Learning with Small Data. In Proceedings of the 25th International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2021.
35. Olson, M.; Wyner, A.; Berk, R. Modern Neural Networks Generalize on Small Data Sets. In *Advances in Neural Information Processing Systems*; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; 2018; Volume 31. Available online: https://www.semanticscholar.org/paper/Modern-Neural-Networks-Generalize-on-Small-Data-Olson-Wyner/a25bb56506fd1772e17d5b57a75ec838dafb6757 (accessed on 10 February 2022).
36. Arora, S.; Du, S.S.; Li, Z.; Salakhutdinov, R.; Wang, R.; Yu, D. Harnessing the Power of Infinitely Wide Deep Nets on Small-data Tasks. *arXiv* **2019**, arXiv:1910.01663.
37. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. In Proceedings of the International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018.

38. Liu, L.; Muelly, M.; Deng, J.; Pfister, T.; Li, L.J. Generative Modeling for Small-Data Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019.
39. Barz, B.; Denzler, J. Deep Learning on Small Datasets without Pre-Training Using Cosine Loss. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020.
40. Achenbach, T.; Ruffle, T.M. The Child Behavior Checklist and related forms for assessing behavioral/emotional problems and competencies. *Pediatr. Rev.* **2000**, *21*, 265–271. [CrossRef]
41. Zelazo, P.; Anderson, J.; Richler, J.; Wallner-Allen, K.; Beaumont, J.; Weintraub, S. NIH toolbox cognition battery (CB): Measuring executive function and attention. *Monogr. Soc. Res. Child Dev.* **2013**, *78*, 16–33. [CrossRef]
42. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
43. Kong, Y.; Fu, Y. Human Action Recognition and Prediction: A Survey. *arXiv* **2018**, arXiv:1806.11230.
44. Castelló, J.S. A Comprehensive Survey on Deep Future Frame Video Prediction. Master's Thesis, Universitat de Barcelona, Barcelona, Spain, 2018.
45. Chen, T.; Zhai, X.; Ritter, M.; Lucic, M.; Houlsby, N. Self-Supervised GANs via Auxiliary Rotation Loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Silver Spring, MD, USA, 14–19 June 2020.
46. Trinh, T.H.; Luong, M.T.; Le, Q.V. Selfie: Self-supervised Pretraining for Image Embedding. *arXiv* **2019**, arXiv:1906.02940.
47. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the International Conference on Machine Learning, PMLR, Vienna, Austria, 13–18 July 2020.
48. Gutmann, M.; Hyvärinen, A. Noise-Contrastive Estimation: A New Estimation Principle for Unnormalized Statistical Models. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010.
49. van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2018**, arXiv:1807.03748.
50. Bromley, J.; Bentz, J.; Bottou, L.; Guyon, I.; Lecun, Y.; Moore, C.; Sackinger, E.; Shah, R. Signature Verification using a "Siamese" Time Delay Neural Network. *Int. J. Pattern Recognit. Artif. Intell.* **1993**, *7*, 25. [CrossRef]
51. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In Proceedings of the 38th International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021. Available online: http://proceedings.mlr.press/v139/zbontar21a/zbontar21a.pdf (accessed on 10 February 2022).
52. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
53. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8024–8035.
54. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
55. Huckvale, K.; Venkatesh, S.; Christensen, H. Toward clinical digital phenotyping: A timely opportunity to consider purpose, quality, and safety. *Npj Digit. Med.* **2019**, *2*, 1–11. [CrossRef]