A. Supplementary Material

A.1. Related work

In this section, we briefly review the literature most relevant to our study.

A.1.1 Deep learning-based ASD detection

Recent studies on computer-aided diagnosis of ASD based on deep learning utilizing several cues such as grasping gestures, repetitive behaviors, eye movement, abnormal gait, and facial features have attracted significant research interests [22–24, 32, 33, 36, 43]. For example, Zunino *et al.* [43] introduced a novel dataset of simple grasping gestures labeled according to ASD/TD subject and demonstrated that a simple LSTM model with an attention module can effectively classify whether a grasping act is performed by ASD or TD subjects. Jiang *et al.* [24] claimed that differences in eye movement patterns between ASD and TD can be discriminative features and proposed a deep learning model trained by using their eye-tracking data in free image viewing. In another work, Syeda *et al.* [36] utilized the face scanning cues based on the observation that children with ASD show less attention to facial features (i.e. eyes, nose, mouth) during face scanning. Jaiswal *et al.* [22] proposed a method for diagnosing ASD in test subjects through automatic video analysis using facial cues such as head motion, head pose, and facial expression. In addition, Rihawi *et al.* [33] focused on autistic behaviors such as rocking, flapping, spinning, etc., and provided a publicly available 3D skeleton dataset of repetitive behaviors (3D-AD). Likewise, Jazouli *et al.* [23] investigates repetitive behaviors (i.e. hand flapping, hand on the face, hands behind back, fingers flapping, and body rocking) and proposed a DNN network utilizing 3D skeleton features.

A.1.2 Self-supervised learning

To overcome the limitation of conventional supervised methods requiring laborious annotation effort, recent self-supervised learning (SSL) methods that learn feature representations from data itself have attracted significant research interest. Early studies on SSL try to learn feature embeddings by introducing handcrafted pretext tasks [1, 15, 27, 30, 31]. In particular, Jigsaw [30] introduced the context-free network (CFN) and train the network in an unsupervised manner by training the CFN to solve Jigsaw puzzles. Counting [31] defines counting as a pretext task and relates it to the counted visual primitives. Likewise, RotNet [27] proposed a simple framework for SSL that trains a CNN to recognize the image rotation that is applied to input images.

Further to the success of the pretext task-based SSL methods, recent state-of-the-art approaches are exploiting contrastive learning that attracts the semantically positive samples while repulsing the negative pairs [9, 20]. MoCo [20] introduced a momentum encoder for robust representations learning of negative pairs drawn from a memory bank. SimCLR [9] showed the effect of different design choices and presented a simple framework for contrastive learning without a memory bank. Rather than relying on a large number of negative samples, recent non-contrastive methods such as BYOL [19] proposed two CNN referred to as online and target networks, where the target network is updated with a slow-moving average of the online network. In a similar approach, SimSiam [10] proved that a simple design of a weight-sharing Siamese network can prevent collapse without negative pairs or a momentum encoder. In a different study, Barlow Twins [40] showed that by using a loss function that brings a cross-correlation matrix between the feature embeddings of two augmented views close to the identity matrix, distortion-invariant features can be learned without using any asymmetry mechanism. In addition to advances within convolutional networks, recent studies for effectively applying self-supervised learning to Transformer architectures such as MoBY [11], MoCo v3 [39] and DINO [8] are also widely being explored.

A.1.3 Domain generalization

Our work is motivated by recent progress in domain generalization. Different from domain adaptation (DA) where access to datasets from a target domain is required in a training stage, the goal of domain generalization (DG) is to learn models that are robust to domain shift and generalize well to arbitrary unseen target domain. In DG, there are two major approaches. The first involves the methods that generate out-of-distribution (OOD) samples with data or feature augmentation techniques [37, 38, 41, 42]. For example, Zhou *et al.* [42] proposed a MixStyle, where instance-level feature statistics are mixed across multiple source domains to broaden the domain distribution of the source domain. Wang *et al.* [38] designed a style complement module for generating diverse OOD images from a single source domain by utilizing an iterative min-max mutual information optimization strategy. Volpi *et al.* [37] and Zhou *et al.* [41] adopt adversarial learning schemes to generate diverse sets of source domain images.

The second approach involves methods that aim to learn domain-invariant features [7, 12, 26, 28] which is the scope of this paper. Choi *et al.* [12] introduced an instance selective whitening loss to disentangle the domain-specific style and domain-invariant content and only remove the style information. Li *et al.* [28] proposed an adversarial autoencoder-based framework to learn a domain-invariant feature representation by minimizing Maximum Mean Discrepancy (MMD). Among these approaches, our work is most relevant to SSL-based DG [7, 26]. JiGen [7] ex-

Figure 5. Examples of result images on the *ASD-Pointing* dataset. The green and red colors represent test cases where pointing is performed and not performed, respectively. The videos were captured with four Azure Kinect cameras in three living lab spaces.

ploits pretext task-based SSL scheme (i.e. Jigsaw [30]) as an auxiliary regularization to learn domain invariant features. Likewise, SelfReg [26] proposed a new regularization method for domain generalization utilizing a self-supervised contrastive regularization loss. Motivated by these works, in our paper, we propose the SSL-based DG framework to recognize the pointing gestures of children without directly using target domain data, where access to large datasets with fine-grained annotations is limited.

A.2. Experiments

A.2.1 Additional examples of result images

Fig. 5 shows additional examples of result images on the *ASD-Pointing* dataset during the SIIC-based testing from four different camera views.

A.2.2 Varying the number of frames in the ensemble block

In this subsection, we analyzed the effect of varying the number of frames, T, in the ensemble block. As shown in Fig. 6, due to the trade-off relation, recall of the Proposed$_{\text{SimSiam}}$ decreases while precision increases when using a larger number of frames within a sliding window. For the accuracy and F1-score, the usage of temporal ensemble strategies provides an overall performance improvement. The performance improvement was most dramatic when T is changed from 0 to 1, where the accuracy and F1-score of the Proposed$_{\text{SimSiam}}$ were improved about 9%p, respectively. Through these experiments, we observed that more robust prediction results can be obtained by adopting a simple ensemble method that aggregates frame-level predictions into video-level predictions. Considering both accuracy and efficiency, we set the number of frames, $T = 2$, in the rest of our paper.

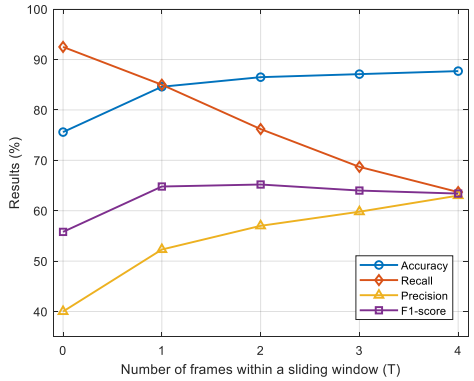

Figure 6. The effect of varying the number of frames within a sliding window, T, in the ensemble block.