Timed Data Incrementation: A Data Regularization Method for IVF Implantation Outcome Prediction from Length Variant Time-lapse Image Sequences

Mehryar Abbasi*†, Parvaneh Saeedi*‡, Jason Au§¶, and Jon Havelock§‖
*School Of Engineering Science,Simon Fraser University, Burnaby, BC, Canada
§Pacific Centre for Reproductive Medicine, Burnaby, BC, Canada
Email:†mabbasib@sfu.ca, ‡psaeedi@sfu.ca,¶jau@pacificfertility.ca,‖jhavelock@pacificfertility.ca

***Abstract*—Identifying embryos with the highest implantation potential is one of the most critical tasks in In Vitro Fertilization (IVF) treatment. We propose a classifier for predicting an embryo's implantation outcome by analyzing time-lapse images during the blastocyst stage. We report a novel data regularization method called Timed Data Incrementation (TDI) to address the length variation in time-lapse image sequences. Sequences with variable length could add significant bias in any training-based method and ultimately lead to an over-fitting or a non-converging system. Our proposed system outperforms the reported state-of-the-art by 2.18% in accuracy (73.08%). The current state-of-the-art is only trained and tested on a single blastocyst image from many embryos. However, to the best of our knowledge, we are the first to utilize time-lapse sequences for embryo implantation outcome prediction. Finally, We show that TDI could benefit other AI-based systems requiring analyzing videos with different capture frequencies or lengths in various applications.**

***Index Terms*—IVF, Machine Learning, Time-lapse Image Sequence, Classification, Video Analysis**

I. Introduction

Infertility is a reproductive disease that affects over 180 million people around the world [1]. In-Vitro Fertilization (IVF) is one of the most common fertility treatments done worldwide [2]. The success rate for IVF remains at 30% despite the improvement in techniques and technologies. IVF involves fertilization of multiple ova retrieved from a female patient in-vitro. Fertilized ova (zygotes) stay in an incubator with a controlled environment for five days. The development process is then recorded and monitored through time-lapse imaging systems. On the 5th day, one or more embryos are selected to be transferred to the patient's uterus. Transferring multiple embryos was a strategy to increase the IVF success rate. It, however, increased the risk of multiple pregnancies and complications at birth [3]. Today the primary strategy for optimizing IVF's outcome is transferring one embryo with the highest implantation potential. The most common selection process is grading an embryo at its blastocyst stage (day-5 embryo) [4]. Such grading is performed by embryologists based on the characteristics of a blastocyst's main components. The grading process is subjective and requires an expert embryologist. Additionally, it could be inconsistent among experts/clinics with different grading systems [5].

Here, we propose a system for predicting implantation outcomes using time-lapse sequences of an embryo during blastulation. Utilizing time-lapse video sequences in Artificial Intelligence-based applications is the next natural trend in deep learning-based applications, especially in the medical image analysis field. The only reported work for analyzing an embryo's time-lapse video sequences is in [6], where time-lapse video sequences of human embryos were used for predicting fetal heart pregnancy. We introduce Timed Data Incrementation (TDI) to reduce a potential bias that arises from image sequences with variant lengths. Feeding videos of high redundancy to a model without proper regulation of the length will lead to a bias in the model toward lengthier videos. The effect of such a process is similar to repeating some of the training images of one class when training an image classifier model. Here, the apparent imbalance is not between the classes and is instead between the samples. Such bias can be even more destructive when limited training data is available. We utilized a dual-path structure to process blastocysts at the early and advanced development phases. Our model consists of two separated processing paths with a shared section. Therefore, we suggest a new data loading method, Random Repetitive Training (RRT), that prepares dual data batches to avoid repeating frames. Our main contributions include:

1) Timed Data Incrementation (TDI), An epoch-wise data injection strategy of the training phase for a structural increase in training data, according to the training stage and the length of each data.
2) A data load and batch preparation technique, called Random Repetitive Training (RRT), for applications with simultaneous multipath model training.

The remainder of this paper is structured as follows. The following related works section reviews the researches that focused on using machine learning-based strategy for automatic grading or outcome prediction of embryos. Section III examines the available data and its format, then the proposed classifier, TDA, and the RRT algorithms are presented. Finally, Section IV showcases our results on our embryo time-lapse image dataset and the effect of the addition of TDA and RRT

978-1-6654-3288-7/21$31.00 ©2021 IEEE

on the performance of our classifier. These results are then compared with the reported state-of-the-art. This section also includes experimental results highlighting the effect of the TDI algorithm by testing it on another time-lapse image dataset.

II. Related works

A. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have seen significant growth in applications of medical image analysis such as detection or classification of breast cancer cells [7], stem cells [8], and skin lesions [9]. Deep CNN architectures, Resnet [10] and Inception [11], have offered better accuracy in image classification tasks. They have been extended to applications in videos and time-lapse image datasets. Some examples include CNN+RNN models [12] and 3D CNN models [13]. CNN+RNN models spatial and temporal information separately, where 3D CNN models combine spatial and temporal processing.

B. Machine-learning based human Embryo analysis

Previous works focus on assessing human embryos' quality using computer-based algorithms. They are divided into two categories:(i) grading the quality (ii) predicting the implantation or live-birth outcome.

1) Embryo grading: Embryo grading is a scheme for embryologists to select the best embryos for transfer. Commonly, day-5 embryos (blastocysts) are graded visually using their morphological attributes [14]. These attributes are:

(i) Expansion level of embryo's cavity (2 to 6)
(ii) Inner Cell Mass's (ICM) cells' distinctiveness and compactness (A to C)
(iii) Trophectoderm's (TE) cell regularities and formation (A to C) [4].

Many attempted to automatically grade embryos at different growth stages via processing single images [15]–[17]. [16] graded an embryo by classifying ICM components separately with an accuracy of 75.36%. [17] extended grade classification mechanism to time-lapse video sequences of day 1 to day 5. The structure in [17] utilized a CNN-based model as the feature vector extractor with individual frames as its input. An RNN structure performed the final grading of TE and ICM with an accuracy of 74.15%, which is lower than their single image counterpart method [16]. [15] introduced a deep learning-based model for single embryo images to classify blastocysts as good or poor with a precision of 95.7%.

2) predicting implantation or live-birth outcome: Since embryo grading is subjective, the actual outcome seems to be a better indicator of embryos' quality. Here we define implantation as achieving a positive pregnancy test and live-birth as the live birth of a baby. The dataset used in this group of works includes images of embryos transferred with known outcomes. [18] introduced a conventional machine learning classifier for live-birth prediction. It required manual segmentation of boundaries of blastocysts along with handcrafted feature extraction. The extracted features are then fed into an ensemble of classifiers for further analysis (This work is refered to as HFC). [19] proposed a multi-path deep learning-based model that used implantation annotations. It reported an accuracy of 70.9% for implantation outcome. In [20], a dataset with live-birth annotation was used to train a multi-layer CNN model combined with the Conventional Embryo Evaluation(CEE) algorithm as the live-birth predictor with an average accuracy of 68.9%.

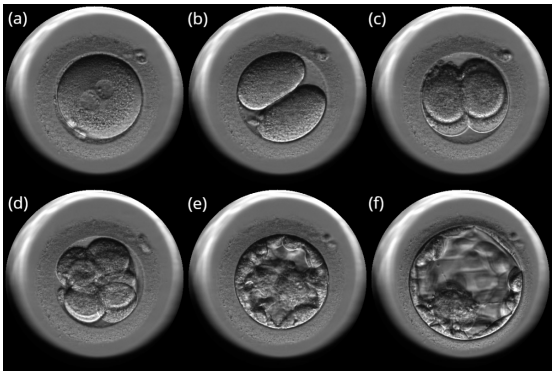

Fig. 1. Embryo development stages(a) pronuclear, (b) two-cells, (c) Three or four-cell, (d) Over five cells, (e) Morula,(f) Blastocyst.

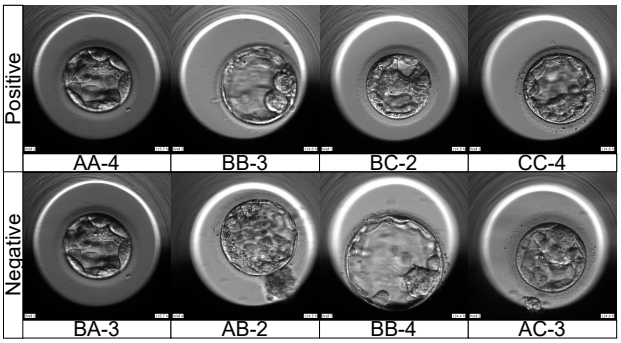

Fig. 2. Examples of Embryo grades of the last frame of a sequence. The top row samples have positive implantation outcome. The bottom row samples have negative implantation outcome.

All the above methods are based on a single blastocyst image of embryos right before transfer. Our goal here is to present a system that utilizes a time-lapse image sequence of an embryo during its blastulation and predicts the implantation outcome.

III. Methodology

A. Data

The dataset used in our work comprises 127 time-lapse video sequences of embryos developed in an EmbryoScopeTM time-lapse incubator (Vitrolife) for five days.

The image capture frequency was at 15 minutes. Each frame is an 8-bit image of 500×500 pixels. Expert embryologists annotated each sequence into six sections/intervals according to its development stage, including pronuclear, two-cells, three or four-cell, over five-cell, morula, and blastocyst (Fig. 1). The embryologist also grades the embryo based on the last

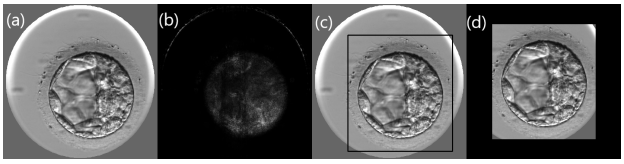

Fig. 3. (a) Raw input, (b) optical flow sum, (c) detected borders, (d) Final image.

frame of each sequence. Sample of graded embryos and their implantation outcome is shown in Fig. 2.

Here, we only utilized and crop frames from the blastocyst section. The cropped videos of blastocyst frames are referred to as sequences in the remainder of this paper. A blastocyst's morphological attributes are different at the beginning of its formation from later when transferred to the body (Fig. 4). Visual features at the beginning are more similar to those of day-4 (morula stage). Therefore, each blastocyst video section is divided into two at its mid-frame, referred to early and advanced sub-sequences. This is the main reason for our dual structure. Each sequence's length varies from 2 to 154 frames due to embryos' different progress rates. Out of 127 sequences, 59 (47%) resulted in positive, and 68 (53%) negative implantation.

B. Data Preparation- Hard Attention Process

The first step for the data preparation is to identify the image region that includes embryonic cells. We used the thresholded optical flow [21] of every two consecutive frames. The threshold value was set to $0.25 \times median(optical\ flow\ sum)$ across the image. We also utilized the contour detection in the OpenCV library [22] to identify cells' contours. The coordinates of the contours were then used for hard attention cropping [23] of the image that removed areas outside the region of interest. Next, an image patch was created centered at the region of interest. The above process is illustrated visually through Fig. 3-a to -d.

The following standard data augmentations are applied on each patch: crop at a scale ratio of 0.85 to 1 (randomly selected ratio), spatial shift on both dimensions (randomly selected between 0-5% of the image size), rotate by an angle (randomly chosen between 0 to 45 degrees) and flip in one of the four directions (up, down, left, right) randomly. The images were then resized to 224×224 pixels before their intensities were normalized.

C. Model architecture

The proposed CNN architecture consists of two separate processing pathways, each for one category of the early or the advanced blastocyst sequences (Fig. 4). Our tests show that such a separation improves the model's performance, as the visual characteristics of the two are considerably different. The feature vectors of both paths were concatenated before being fed into the Fully Connected layer.

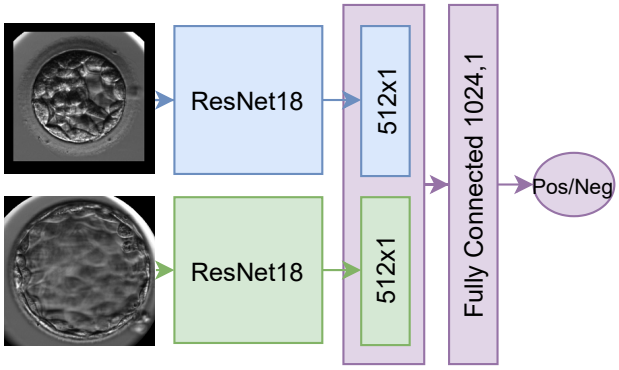

Fig. 4. The dual path model: bottom- advanced stage , top- early stage.

D. Timed Data Incrementation

TDI is a novel strategy that controls input sequences in the training data based on the training's progress. Without TDI in the training phase, the model might be exposed to more frames from videos with longer lengths. Such imbalance exposure causes a destructive bias in the model. TDI specifies and ranks sequences for training according to their lengths. The main goal is to limit the training data to videos with shorter lengths and gradually add the other videos as the training progresses. The training data in the early epochs comprises samples from sequences that are shorter than a given length. The length limit is modified once a certain number of training epochs are achieved. The process continues until all training sequences are integrated into the training operation. In short, TDI filters out and limits the starting training sequences, and upon reaching specified injection epochs, the filtering criterion expands to include more sequences. This process continues until all of the data is included in the training. The filtering criterion for TDI is the sequences' length. The frequency of the injection rates is a hyperparameter that can be modified based on the problem.

E. Random Repetitive Training (RRT)

Random Repetitive Training (RRT) is an algorithm that creates fresh combinations for our model's last layer on the same epoch that could have repeated frames for the separate section of each path. If the early and the advanced sections of the sequence each have L frames, a frame from one section can be randomly paired with another frame from the other section, creating a total of L image pairs in an epoch. This strategy ensures that none of the images are repeated in that epoch. We called this Non-Repetitive Training (NRT). In RRT, however, L pairs are created randomly on the spot without the knowledge of the other pairs. When the model is trained, elements of the pair might have already passed the model on the same epoch, but the combination would be new for the last layer. Until this position, we have been trying to avoid repeating samples to lessen the bias buildup. However, at this stage, the error of a new combination in the last layer delays the bias buildup through the separate model paths.

TABLE I
COMPARISON OF THE IMPLANTATION PREDICTION ACCURACY IN PER SEQUENCE MANNER.

Model	TDI	RRT	PRC[1]	RCL[2]	JI[3]	ACC[4]
Single path	✗	✗	65.9	61.67	44.65	62.82
		✓	67.57	63.17	46.25	64.36
	✓	✗	66.07	64.43	47.48	65.87
		✓	71.52	66.7	50.18	67.07
Dual path	✗	✗	69.79	64.52	46.25	63.46
		✓	71.81	65.87	49.33	66.67
	✓	✗	73.65	69.79	53.74	71.15
		✓	**74.11**	72.62	**57.22**	**73.08**
Single Stream [19]			63.6	63.6	46.7	62.8
Raw+Mask CNN [19]			71.1	**72.7**	56	70.9
HFC[**][18]			61.5	60.5	44	62
CEE + CNN (all age ranges)[**][20]			66.2	69.5	50.2	68.9

[**] Live-birth [1] Precision [2] Recall [3] Jaccard-Index [4] Accuracy

F. Network Training

During the training, the input sequences are divided into image pairs either through NRT or RRT, and their corresponding outcome labels are passed to the model.

G. Network Testing

Steps to predict the outcome of an input test sequence are as follows: First, the sequence is cut in half. Second, all possible combinations of image pairs are created. In each pair, one image is selected from each half. For a sequence with $2L$ length, the number of created sets would be L^2. Finally, the predicted probabilities of all pairs are averaged to create a single prediction for the input sequence.

IV. EXPERIMENTAL RESULTS

Our model was trained and tested on our dataset described in Section III-A. We used the Cedar cluster of Compute Canada [24] that is equipped with NVIDIA V100 Volta (32G HBM2 memory). The models were trained with mini-batches of size 256 over 200 epochs. We set the starting learning rate at $1e{-}4$, which was then reduced by a factor of 0.1 on epochs 30 and 80 (the respective values were $1e{-}5$ and $1e{-}6$).

TDI algorithm's filtering length upper limit was set to 15 initially (only those sequences with lengths shorter than the 15 were let through). The upper limit was then increased to 30, 45, and 200 at epochs 5, 10, and 15, respectively. The number of sequences between these video lengths accounted for roughly 25% of the entire dataset. To assess TDI's performance, we also included experiments with non-TDI configurations. To demonstrated RRT's impact, we included experiments with NRT configuration (described in Section III-F) too. Additionally, we presented test configurations with a single path model for an entire blastocyst sequence.

TABLE II
EFFECT OF TDI ALGORITHM ON I3D TRAINED ON SBU RGB.

Algorithm	PRC	RCL	JI	ACC
TDI Off	95.2	94.0	90.5	94.5
TDI On	**95.9**	**94.8**	**91.6**	**95.3**

The prediction outcome for the Dual-path model was generated using the method described in Section III-G. In the single-path model, all the frames of each sequence were given to a Resnet18 [10]. Then the single frame predictions were averaged together to create each sequence's predicted outcome. Table I details results for 5-fold cross-validation. We divided our dataset into five sets. We used one section as the test and the other four as the training datasets in each run. This selection is shuffled five times, each time different from the times before. The outcomes of all runs are averaged and reported. The last two rows of Table I display two reported systems that predict the live-birth outcome. Please note that it is not appropriate to compare these two systems directly with our proposed method in this paper as the outcomes are different (positive pregnancy vs. live-birth), not every positive pregnancy leads to live birth. However, it might be useful to the readers to know about those works in the context of our research.

Table I leads to the following conclusions. The Dual-path model with RRT and TDI delivers the best performance over the other configurations. Moreover, TDI improvement is independent of the model's configuration.

A. Application of TDI on time-lapse sequences in deep-learning

To demonstrate the effectiveness of the TDI method, we present its application on a different system that utilizes RGB videos of the SBU Kinect Interaction dataset [25]. We used SBU Kinect 'RGB video data to achieve an action recognition task. We fine-tuned a pre-trained I3D model [26] with 5-fold cross-validation on that dataset's predetermined folds. Results for these tests are shown in Table II. Our tests were carried on under identical conditions (100 epochs, starting learning rate of 0.0001) once with and once without TDI.

These results demonstrate that TDI is beneficial to other types of applications and models where the input includes video sequences with variable length.

V. CONCLUSION

This paper proposed a fully automated approach for predicting human embryo implantation outcomes from time-lapse image sequences in the IVF process. We presented a dual-path CNN that processed blastocyst sequences in two different classified stages. The dual-path model showed improvement over the single-path model. This improvement shows that the evolution of an embryo's morphological attributes during the blastocyst stage is substantial and therefore, it justifies increasing the computational cost. We also presented a novel algorithm (TDI), which regulated the training data based on

input sequences' lengths, reducing the potential bias associated with them due to the frame redundancy and the slow-varying scenes. Experimental results showed that our system delivers an implantation outcome prediction accuracy of 73.07% for our time-lapse image sequences. TDI increases the mean prediction accuracy by 4.96% from non-TDI (64.32%). Our experimental results also showed that the TDI method could be used in other applications for processing image sequences with variable length. Moreover, its application is independent of the model's configuration.

REFERENCES

[1] M. C. Inhorn and P. Patrizio, "Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21st century," *Human reproduction update*, vol. 21, no. 4, pp. 411–426, 2015.

[2] E. Santos Filho, J. Noble, and D. Wells, "A review on automatic analysis of human embryo microscope images," *The open Biomed. Eng. journal*, vol. 4, p. 170, 2010.

[3] P. Saeedi, D. Yee, J. Au, and J. Havelock, "Automatic identification of human blastocyst components via texture," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 12, pp. 2968–2978, 2017.

[4] C. Manna, L. Nanni, A. Lumini, and S. Pappalardo, "Artificial intelligence techniques for embryo and oocyte classification," *Reproductive biomedicine online*, vol. 26, no. 1, pp. 42–49, 2013.

[5] T. Puga-Torres, X. Blum-Rojas, and M. Blum-Narváez, "Blastocyst classification systems used in latin america: is a consensus possible?" *JBRA Assisted Reproduction*, vol. 21, no. 3, p. 222, 2017.

[6] D. Tran, S. Cooke, P. Illingworth, and D. Gardner, "Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer," *Human Reproduction*, vol. 34, no. 6, pp. 1011–1018, 2019.

[7] N. Bayramoglu, J. Kannala, and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," in *2016 23rd Int. Conf. on pattern recognition*. IEEE, 2016, pp. 2440–2445.

[8] A. Witmer and B. Bhanu, "Multi-label classification of stem cell microscopy images using deep learning," in *2018 24th Int. Conf. Pattern Recognit.* IEEE, 2018, pp. 1408–1413.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. Comput. vision and pattern Recognit.*, 2016, pp. 770–778.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. of the IEEE Conf. Comput. vision and pattern Recognit.*, 2015, pp. 1–9.

[12] G. An, W. Zhou, Y. Wu, Z. Zheng, and Y. Liu, "Squeeze-and-excitation on spatial and temporal deep feature space for action recognition," in *2018 14th IEEE Int. Conf. on Signal Process.* IEEE, 2018, pp. 648–653.

[13] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool, "Temporal 3d convnets: New architecture and transfer learning for video classification," *arXiv preprint arXiv:1711.08200*, 2017.

[14] D. K. Gardner, M. Lane, J. Stevens, T. Schlenker, and W. B. Schoolcraft, "Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer," *Fertility and sterility*, vol. 73, no. 6, pp. 1155–1158, 2000.

[15] P. Khosravi, E. Kazemi, Q. Zhan, J. E. Malmsten, M. Toschi, P. Zisimopoulos, A. Sigaras, S. Lavery, L. A. Cooper, C. Hickman *et al.*, "Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization," *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–9, 2019.

[16] T.-J. Chen, W.-L. Zheng, C.-H. Liu, I. Huang, H.-H. Lai, and M. Liu, "Using deep learning with large dataset of microscope images to develop an automated embryo grading system," *Fertility & Reproduction*, vol. 1, no. 01, pp. 51–56, 2019.

[17] M. F. Kragh, J. Rimestad, J. Berntsen, and H. Karstoft, "Automatic grading of human blastocysts from time-lapse imaging," *Computers in biology and medicine*, vol. 115, p. 103494, 2019.

[18] A. Chavez-Badiola, A. F.-S. Farias, G. Mendizabal-Ruiz, R. Garcia-Sanchez, A. J. Drakeley, and J. P. Garcia-Sandoval, "predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning," *Scientific Reports*, vol. 10, no. 1, pp. 1–6, 2020.

[19] R. M. Rad, P. Saeedi, J. Au, and J. Havelock, "Predicting human embryos' implantation outcome from a single blastocyst image," in *2019 41st Annu. Int. Conf. of the IEEE Eng. in Med. and Biol. Society*. IEEE, 2019, pp. 920–924.

[20] Y. Miyagi, T. Habara, R. Hirata, and N. Hayashi, "Feasibility of predicting live birth by combining conventional embryo evaluation with artificial intelligence applied to a blastocyst image in patients classified by age," *Reproductive Medicine and Biology*, vol. 18, no. 4, pp. 344–356, 2019.

[21] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scand Conf. Image Anal.* Springer, 2003, pp. 363–370.

[22] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[23] A. Kosiorek, A. Bewley, and I. Posner, "Hierarchical attentive recurrent tracking," in *Adv in neural Inf. Process. Sys.*, 2017, pp. 3053–3061.

[24] O. A. https://www.computecanada.ca/.

[25] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *2012 IEEE Conf. on Comput. Vis. and Pattern Recognit WorkShop.* IEEE, 2012, pp. 28–35.

[26] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2017, pp. 6299–6308.