Human Embryo Classification Using Self-Supervised Learning

1st R. Satrio Hariomurti Wicaksono
Faculty of Science and Technology
Universitas Al Azhar Indonesia
Jakarta, Indonesia
wicaksono.satrio@if.uai.ac.id

2nd Ali Akbar Septiandri
Faculty of Science and Technology
Universitas Al Azhar Indonesia
Jakarta, Indonesia
aliakbar@if.uai.ac.id

3rd Ade Jamal
Faculty of Science and Technology
Universitas Al Azhar Indonesia
Jakarta, Indonesia
adja@uai.ac.id

***Abstract*—The implementation of computer vision generally focuses on resolving a problem of one of two types of image data, whether it was natural image datasets or medical image datasets. Self-supervised learning (SSL), as the most recent approach in computer vision, is typically used to solved tasks related only to natural image datasets such as ImageNet, CIFAR-10, and PASCAL. This study focuses on analyzing the involvement of natural images in the SSL implementation when classifying medical images. We used the ResNet50 architecture comparing the results of human embryo image classification using the conventional transfer learning method and SSL methods. We apply three SSL approaches, namely Rotation, Jigsaw Puzzle, and SimCLRv1. The results of our research, with 1226 embryo images divided into three classes, show that the SSL implementations cannot be said to be efficient enough compared to the conventional model produced by ResNet50 with pretrained weight ImageNet. However, when the models were trained using random weight initialization, the Rotation model with an average accuracy of 78.15% and the Jigsaw Puzzle model with an average accuracy of 80.30% are significantly superior to the conventional model with average accuracy is 71.07%. Moreover, related to our embryos dataset, we discuss the differences between the SSL context-context contrastive learning approach and the context instance contrastive learning approach.**

***Keywords*—self-supervised learning, convolutional neural network, transfer learning, images classification**

I. Introduction

The specific objective of implementing machine learning for image classification tasks is to obtain consistent and optimal results, which is quite difficult to achieve manually by humans. The ubiquitous practice in machine learning mainly focuses on natural images than medical ones. However, in recent years, the implementation of medical images in machine learning tasks has been increasing. The background of this practice is the high demand for faster, more objective, and consistent results. Medical images have different characteristics compare to natural images. Specific color composition, contrast, and other attributes make medical images hard to notice and indistinguishable for humans. It became few of many factors that manual task related to medical images is relatively more arduous to do by humans than by machine [1].

978-1-6654-1726-6/21/$31.00 ©2021 IEEE

The classification of human embryos in the process of in-vitro fertilization is one example of machine learning's implementation on medical images that we will discuss in this research. The human embryo quality assessment process carried out in in-vitro fertilization is divided into three types according to the sequence of the process. The first is the Zygotes Scoring System which is effectuated 16-18 hours after the oocyte insemination process [2], [3]. The second is the Cleaved Embryos Scoring which is carried out on the third day after the oocyte insemination process [4]. The last is the Blastocyst Scoring System which is carried out on the fifth day after the oocyte insemination process [5]. This research is a continuation of previous research done by Septiandri et al. [6] that discusses the implementation of several deep learning architectures on embryo image classification on the third day. The standard used as a reference for the assessment is proposed by Veeck, which divides the quality of embryos into five different classes [7]. The results of the previous study concluded that the best architecture for classifying the quality of human embryos is ResNet50 with ImageNet pretrained weight with an accuracy of $91.79\% \pm 0.48\%$ [6].

Thus, this research attempts to focus on the implementation of the Self-Supervised Learning (SSL) approach into the ResNet50 architecture and comparing the classification results between the model from the conventional method (without SSL) and the model from the SSL methods. We also analyzed the effect of using pretrained models with ImageNet and using random weight initialization on classifying embryo images.

II. Self-Supervised Learning

Self-Supervised Learning (SSL) is an approach derived from unsupervised learning. A procedural sequence used in SSL implementation consists of two distinct phases, namely pretext task training and downstream task training [8]. In the pretext task phase, each image was uniquely augmented depending on the type of SSL used. Data augmentation serve to generate pseudo labels used throughout pretext task training. We fed augmented images into CNN for the training process to learn the representation of augmented images [9]. The model trained in this pretext phase was used as a pretrained model (hereafter referred to as pretrained SSL model) in the downstream task training phase. Downstream task training

is a phase where the model is trained using the original label as in supervised learning [8]. In this study, conventional model training (hereafter referred to as benchmark models) was assimilated into the downstream task phase due to the benchmark models, which we used as a comparison to SSL models do not have special image augmentation and training as explained in the pretext phase.

The principle behind the SSL approach makes it possible to implement SSL into deep learning architectures. Based on the previous research [6], we used ResNet50 as it was able to produce the highest accuracy in the human embryos classification. In addition, ResNet50 has the lowest loss compared to other deep learning architectures such as DenseNet, MobileNet, and Xception. It shows that ResNet50 is the most optimal deep learning architecture for classifying human embryos.

Stating the implementation of SSL, there are three SSL approaches in this research, namely Rotation [10], Jigsaw Puzzle [11] [12], and SimCLRv1 [13]. We created two versions for every model from each approach, such as models built using ImageNet pretrained weights and models built with random weight initialization. The creation of these two versions resulted in eight different models at the end of the study, and two of them are benchmark models.

We grouped one pretext phase and five iterations of the downstream phase as one batch. An overview of the general pipeline for each batch is shown in Fig. 1. We repeated the experiment in three different batches with the same parameters to get the general idea of the robustness of each model.

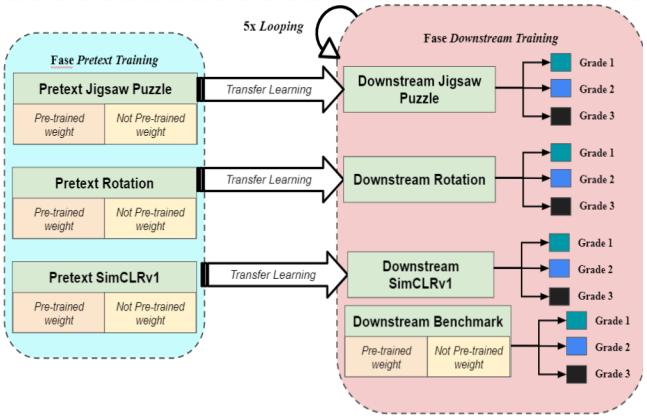

Fig. 1: Initialization of two versions (pretrained weight ImageNet and random weight initialization) for the SSL approach is carried out in the pretext task phase, while the initialization of two versions for the conventional approach (benchmark) is carried out in the downstream task phase. The whole process resulted six models in the pretext task phase, and the models became eight other models in the downstream task phase.

The representation learning in SSL was emphasized in the pretext task phase. We discussed two types of contrastive learning as part of SSL representation learning, namely context-instance contrastive learning and context-context contrastive learning [14]. Context-instance contrastive learning is a model learning that explains the relationship between local features of the sample and the global representation context of a respectable sample [14]. One practice of context-instance contrastive learning is Predict Relative Position (PRP). PRP focuses on studying the relative positions between local components of the sample. In this study, we apply Jigsaw Puzzle and Rotation as a representative of PRP.

On the contrary, context-context contrastive learning is a type of learning that focuses only on the global representation of the sample used [14]. Instance Discrimination is one of the practices of context-context contrastive learning. Global representation of Instance Discrimination is obtained by making several different perspectives from several samples and comparing the results of these perspectives. We used SimCLRv1 as a representative of the Instance Discrimination approach.

A. *Jigsaw Puzzle*

In Jigsaw Puzzle, local features of an image are represented by dividing the image into nine small tiles. Each tile will be reordered based on a predefined subset of permutations. This subset of permutations is then used as a pseudo label in the pretext task. Images and corresponded pseudo labels are fed into CNN and treated as a classification problem. This approach is based on a study conducted by Carlucci et al. [12]. We did not use Context-Free Network (CFN) as described by Noroozi et al. [11] for simplification reasons. The general pipeline of the Jigsaw Puzzle is shown in Fig. 2.

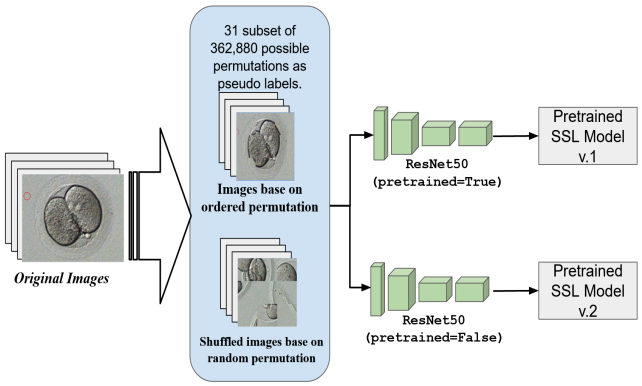

Fig. 2: The general pipeline of Jigsaw Puzzle implementation in the pretext task training phase. Each original image would be stripped from its label. The unlabeled image was augmented base on the ordered permutation and one out of 30 possible permutations. Total images fed to CNN are twice as much as many original images. Then, the images were trained using ResNet50 with pretrained weight ImageNet and ResNet50 with random weight initialization.

B. *Rotation*

Pretext task rotation is an attempt to predict the relative position of an image in the context of its rotation angle. The pseudo label in the pretext phase represents the rotation angle that is implemented in the embryo image. Gidaris et al. [10] stated in their research that 0°, 90°, 180°, and 270° rotation angles were considered good enough to get an optimal feature representation of natural images. We used these four rotation angles as our reference in pretext rotation training in spite of using medical images as our dataset. The general pipeline of Rotation is shown in Fig. 3.

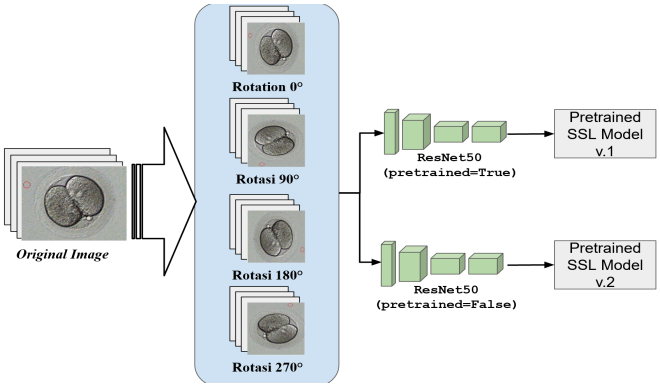

Fig. 3: The general pipeline of rotation implementation in the pretext task training phase. Each original image would be stripped from its label. Unlabeled images were rotated into four different angels. Then, the images were trained using ResNet50 with pretrained weight ImageNet and ResNet50 with random weight initialization.

C. SimCLRv1

SimCLRv1 is a learning method that tries to learn the representation of image objects by maximizing the "agreement" between different perspectives. Different perspectives are obtained from different combinations of augmentation on each sample, such as random crop, color distortion and, Gaussian blur [15]. The general pipeline of SimCLRv1 is shown in Fig. 4.

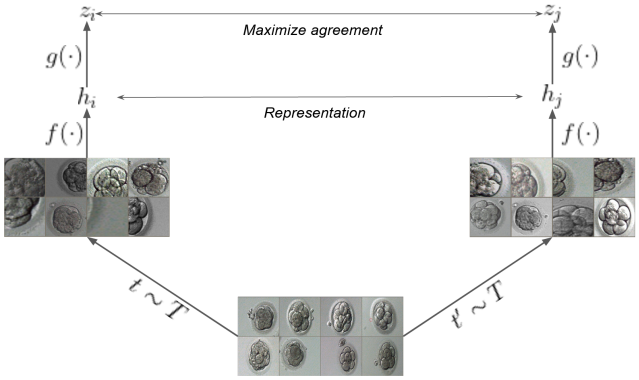

Fig. 4: The image samples were processed through two similar random augmentations process (random crop, colour distortion, and Gaussian blur) to produce two respective versions of original images. The $f(\cdot)$ encoder is used to get the representation from each version of the image and the projection head $g(\cdot)$ is used to see the similarity between representation of each image version in the same batch. High similarity indicates two samples of images versions were originated from the same original image, thus classified as positive pair. Sample images with low similarity were classified as negative pair. By following this rules, there is always 1 positive pair for $2(N-1)$ negative pairs in a batch consisting of N original images. This pipeline was implemented in both ResNet50 with pretrained weight ImageNet and ResNet50 with random weight [13].

III. Dataset

The embryo images used in this study were the same as the previous study. There are 1226 original embryo images [6]. All data were classified according to three of five Veeck standards with a distribution of 459 grade 1, 620 grade 2, and 147 grade 3 embryos. We divided the dataset into train and test sets with 75:25 ratio to model training in the downstream task phases. The training set is further divided into training and validation sets with a 70:30 ratio. This ratio was applied to all approaches of SSL and benchmark in the downstream task phase.

As previously explained about SSL, the image dataset used in the pretext task phase is different for each SSL approach. We took some references from previous research for the datasets creation and pseudo labels used in pretext task training.

A. Jigsaw Puzzle Pretext Dataset

We used a subset of 30 permutations selected based on the Hamming distance algorithm [11]. Each embryo image for a total of 1226 images was divided into nine image tiles with dimensions 63x63 pixels. Image augmentation in the form of randomization of nine tiles was carried out based on one of the random permutations in a pre-determined subset. In addition, the images were also arranged sequentially according to the original image, so that the total data used is 2452 images. We divided the training data and test data in a ratio of 75:25. Throughout the model training process, we implemented a 70:30 ratio for the training set and validation set. An example of one image batch consists of 16 shuffled images shown in Fig. 5.

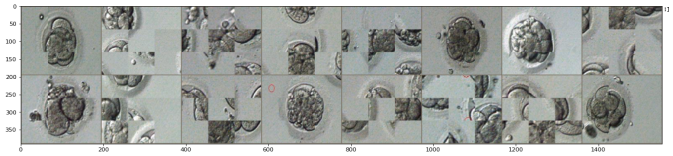

Fig. 5: A batch of Jigsaw Puzzle pretext dataset.

B. Rotation Pretext Dataset

We created four new images from each embryo image which each new image was generated from the rotation of the original image. We used four rotation angles, such as 0° , 90° , 180° , and 270° . Each rotated image had a pseudo label according to the rotation angle. At the end of the augmentation, we had 4904 rotated images. We still used the same ratio as Jigsaw Puzzle. We divided the training data and test data in a ratio of 75:25. Over the model training process, we applied a 70:30 ratio for the training set and validation set. The example of one image batch consists of 16 rotated images is shown in Fig. 6.

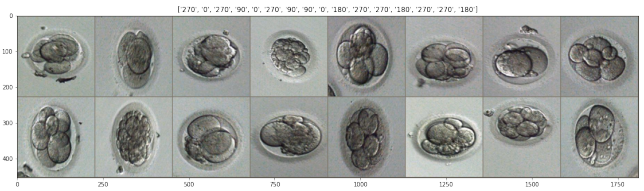

Fig. 6: A batch of Rotation pretext dataset.

C. SimCLRv1 Pretext Dataset

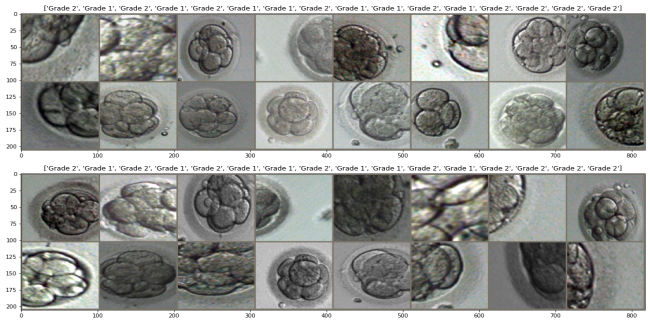

Fig. 7: A batch of SimCLRv1 pretext dataset.

Each embryo image was processed through a set of random augmentations to produce two new images associated with the original image. Therefore, the total images used in the SimCLRv1 pretext task was twice as much as the original data. We arranged the training data and test data in a ratio of 75:25. As the model training process, we utilized a 70:30 ratio for the training set and validation set. An example of one image batch consists of 16 augmented images shown in Fig. 7.

IV. Result and Discussion

The establishment of benchmarks in this study could be interpreted as an effort to replicate the deep learning model from research conducted by Septiandri et al. [6]. The implementation of the fast.ai library in the previous studies makes the PyTorch library an excellent choice to implement in this study. It is because fast.ai is a library built on PyTorch and several other supporting libraries [16]. Even though fast.ai has many advantages in terms of speed and the implementation of other best practices behind every function of fast.ai, we think that SSL customization is still easier to do and observe when using the PyTorch library. After making several parameter adjustments, we managed to train the benchmark model with an accuracy that is close to that of previous studies.

We found that training with 30 epochs and a learning rate of 5×10^{-3} can form a benchmark model that is relatively as good as the previous study [6]. We used these parameters throughout the downstream phase to train another six SSL models and one benchmark model with random weight initialization. According to Table I and Table II, we could assume that using the random weight initialization, the SSL models perform better than benchmark models. However, these advantages are not enough to rival the high accuracy produced by pretrained ImageNet models.

TABLE I: Pretrained ImageNet model comparison

Model	Accuracy		
	***Batch* 1**	***Batch* 2**	***Batch* 3**
Jigsaw	**91.86**%±**0.46**%	90.88%±1.12%	91.40%±0.81%
Rotation	91.01%±1.14%	**90.94**%±**0.86**%	91.40%±1.16%
SimCLRv1	73.55%±4.06%	75.57%±1.98%	68.14%±6.82%
Benchmark	91.14%±1.52%	90.22%±2.57%	**91.66**%±**1.12**%

TABLE II: Random weight initialized model comparison

Model	Accuracy		
	***Batch* 1**	***Batch* 2**	***Batch* 3**
Jigsaw	78.63%±5.40%	**79.67**%±**4.02**%	**82.61**%±**3.02**%
Rotation	**80.33**%±**3.59**%	77.52%±3.48%	76.61%±7.61%
SimCLRv1	75.24%±3.01%	75.83%±2.69%	73.29%±1.84%
Benchmark	70.10%±6.26%	72.05%±6.39%	71.07%±6.85%

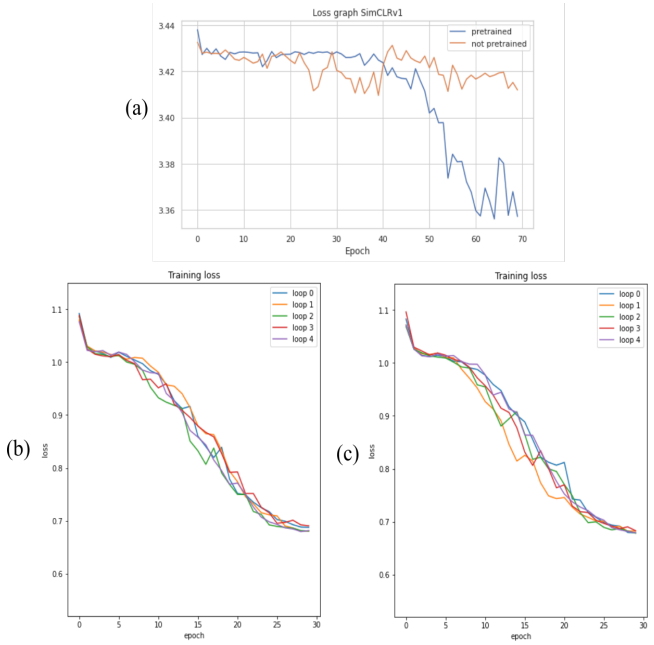

Fig. 8: SimCLRv1 training loss throughout : (a) pretext task with 70 epochs and loss interval from 3.35 to 3.44. (b) downstream using pretext ImageNet with 30 epochs and loss interval from 0.52 to 1.15. (c) downstream using random weight initialization with 30 epochs and loss interval from 0.52 to 1.15.

Furthermore, we are interested in the relatively adjacent difference between the SimCLRv1 accuracy with pretrained ImageNet and SimCLRv1 with random weight initialization. We argue that this is due to the nature of SimCLRv1 as context-context contrastive learning focuses only on global representations of related images. The use of medical images of embryos in the SimCLRv1 pretext made the model rather ignore the usage of natural images and only rely on the embryo images as a global representation. That is presumably one of many possible causes of both SimCLRv1 models behaved as the ImageNet has never been implemented in the training process. That leads to the identically low accuracy produced by both models.

The low accuracy for the SimCLRv1 models also could be due to the potential for under-fitting during pretext and downstream training. Loss showed in Fig. 8 indicates a slow decline throughout training. At the end of the downstream phase, SimCLRv1 still showed a bit of slope, indicating a possibility of local minimum at the end of model training.

Each model has five accuracy values in each batch. It means each model has 15 accuracy values in total. We tested the hypothesis using a Paired Student's T-Test to see whether the SSL model was superior to the benchmark model or the

TABLE III: PAIRED STUDENT'S T-TEST

Model	Mean Diff.	P-Value
Benchmark - Jigsaw	-0.003	0.482
Benchmark - Rotation	-0.001	0.857
Benchmark - SimCLRv1	0.186	0.0**
Benchmark^{RW} - Jigsaw^{RW}	-0.092	0.001**
Benchmark^{RW} - Rotation^{RW}	-0.070	0.005**
Benchmark^{RW} - SimCLRv1^{RW}	-0.037	0.087*
SimCLRv1 - SimCLRv1^{RW}	-0.023	0.168

*$p < 0.1$; **$p < 0.05$; ^{RW} Model initialized with random weight

opposite, based on the difference in the average accuracy of all models [17]. We also tested the hypothesis between the two versions of the SimCLRv1 model to strengthen earlier assumptions regarding context-context contrastive learning. Null hypothesis (H_0) in Paired Student's T-Test stated that the difference between the means of the two samples is not significant [17].

Table III proves the superiority of SSL over the benchmark only seen in models with random weight initialization, except for the SimCLRv1 model. Hypothesis testing also showed that the two versions of the SimCLRv1 model have no significant differences. It could be concluded that the assumptions related to the accuracy which was delivered by SSL and benchmark models were supported by the hypothesis test with SimCLRv1 as an exception.

V. CONCLUSIONS

We have proven in this study that the implementation of SSL with ImageNet pretrained weight does not have a significant impact when compared to the benchmark model. SSL Rotation and Jigsaw Puzzle models show significant benefit only if all models are trained using random weight initialization. If we consider resources used throughout this study, it can be said that the benchmark models are still superior compared to the SSL models on both versions. It is because it is unnecessary for the benchmark to wasted time on pretext training.

We also observed that the implementation of SimCLRv1 as context-context contrastive learning tends to ignore ImageNet pretrained weight when training a model, so it consumed a significantly long time to trained and low accuracy result. Nevertheless, the results of the two SimCLRv1 models tend to be identical. It shows the potential of SimCLRv1 to become an equitably robust model by relying only on more medical images. Moreover, a deeper study related to SimCLRv1 on medical images is needed to support the claim.

Each result in this study is highly dependent on the specific hyperparameter values and construction of pseudo-labels in the pretext task. Thus, an in-depth study is still needed to determine the impact of different configurations on each approach. However, this research was able to represent the general effect and efficiency of SSL implementation on medical images with a relatively small size.

REFERENCES

[1] P. Khosravi, E. Kazemi, Q. Zhan, J. E. Malmsten, M. Toschi, P. Zisimopoulos, A. Sigaras, S. Lavery, L. A. D. Cooper, C. Hickman, M. Meseguer, Z. Rosenwaks, O. Elemento, N. Zaninovic, and I. Hajirasouliha, "Deep Learning Enables Robust Assessment and Selection of Human Blastocysts After In Vitro Fertilization," *npj Digital Medicine*, vol. 2, 2019.

[2] L. Scott, R. Alvero, M. Leondires, and B. Miller, "The Morphology of Human Pronuclear Embryos is Positively Related to Blastocyst Development and Implantation," *Human Reproduction*, vol. 15, no. 11, pp. 2394–2403, 2000.

[3] J. Tesarik and E. Greco, "The Probability of Abnormal Preimplantation Development Can be Predicted by a Single Static Observation on Pronuclear Stage Morphology," *Human Reproduction*, vol. 14, no. 5, pp. 1318–1323, 1999.

[4] T. Baczkowski and W. Kurzawa Rafałand Głabowski, "Methods of Embryo Scoring in In Vitro Fertilization." *Reproductive biology*, vol. 4, no. 1, pp. 5–22, 2004.

[5] 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.

[6] A. A. Septiandri, A. Jamal, P. A. Iffanolida, O. Riayati, and B. Wiweko, "Human Blastocyst Classification after In Vitro Fertilization Using Deep Learning," 2020.

[7] M. I. Hsu, J. Mayer, M. Aronshon, S. Lanzendorf, S. Muasher, P. Kolm, and S. Oehninger, "Embryo Implantation in In Vitro Fertilization and Intracytoplasmic Sperm Injection: Impact of Cleavage Status, Morphology Grade, and Number of Embryos Transferred," *Fertility and Sterility*, vol. 72, no. 4, pp. 679–685, 1999.

[8] L. Jing and Y. Tian, "Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[9] I. Misra and L. van der Maaten, "Self-Supervised Learning of Pretext-Invariant Representations," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 6706–6716.

[10] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised Representation Learning by Predicting Image Rotations," in *ICLR 2018*, 2018.

[11] M. Noroozi and P. Favaro, "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles," *CoRR*, 2016.

[12] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 2224–2233, mar 2019.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," 2020.

[14] X. Liu, F. Zhang, Z. Hou, Z. Wang, L. Mian, J. Zhang, and J. Tang, "Self-supervised Learning: Generative or Contrastive," 2020.

[15] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *CoRR*, vol. abs/2006.10029, 2020.

[16] J. Howard and S. Gugger, "Fastai: A layered api for deep learning," *Information (Switzerland)*, vol. 11, no. 2, feb 2020.

[17] J. Brownlee, *Statistical Methods for Machine Learning Discover how to Transform Data into Knowledge with Python*. Machine Learning Mastery, 2019.