Embryo Staging with Weakly-Supervised Region Selection and Dynamically-Decoded Predictions

Tingfung Lau   TINGFUNL@CS.CMU.EDU
Carnegie Mellon University

Nathan Ng   NNG@FB.COM
Facebook AI Research

Julian Gingold   GINGOLJ@CCF.ORG
Cleveland Clinic Foundation

Nina Desai   DESAIN@CCF.ORG
Cleveland Clinic Foundation

Julian McAuley   JMCAULEY@ENG.UCSD.EDU
University of California San Diego

Zachary C. Lipton   ZLIPTON@CMU.EDU
Carnegie Mellon University

Abstract

To optimize clinical outcomes, fertility clinics must strategically select which embryos to transfer. Common selection heuristics are formulas expressed in terms of the durations required to reach various developmental milestones, quantities historically annotated manually by experienced embryologists based on time-lapse *EmbryoScope* videos. We propose a new method for automatic embryo staging that exploits several sources of structure in this time-lapse data. First, noting that in each image the embryo occupies a small subregion, we jointly train a region proposal network with the downstream classifier to isolate the embryo. Notably, because we lack ground-truth bounding boxes, our we weakly supervise the region proposal network optimizing its parameters via reinforcement learning to improve the downstream classifier's loss. Moreover, noting that embryos reaching the blastocyst stage progress monotonically through earlier stages, we develop a dynamic-programming-based decoder that post-processes our predictions to select the most likely monotonic sequence of developmental stages. Our methods outperform vanilla residual networks and rival the best numbers in contemporary papers, as measured by both per-frame accuracy and transition prediction error, despite operating on smaller data than many.

1. Introduction

Following its introduction in 1978, in vitro fertilization (IVF), in which an egg and sperm are combined outside the body, has rapidly emerged as one of the most successful assisted reproductive technologies, contributing to roughly 1.7% of all births in the United States

(Center for Disease Control, 2016). A single cycle of IVF may lead to the growth of multiple ovarian follicles, each of which may contain an oocyte (egg cell). These oocytes are aspirated with a fine needle using ultrasound guidance while the patient is under anesthesia and subsequently fertilized with sperm. Only a fraction of the oocytes fertilize, and a smaller fraction continue to grow and develop normally as embryos before being considered ready for transfer into the uterus (typically after 5-6 days, though some labs use embryos grown only 3 days). Although this process typically generates multiple embryos, most infertility clinics strongly encourage (and some require) transfer of only one embryo at a time because of the greater maternal and fetal risks associated with multi-fetal gestation.

Unfortunately, even under the best circumstances and with genetic testing, the implantation rate following embryo transfer is around 70% (Simon et al., 2018) and may be significantly lower without genetic testing, meaning that patients may be forced to undergo multiple transfers of embryos generated from IVF cycle(s) in order to achieve a single normal pregnancy. This leads to the clinical challenge of identifying and prioritizing those embryo(s) most likely to lead to a normal pregnancy with the fewest total transfers, in the least time and at the lowest cost. These priorities are often in direct conflict, leading to the wide variability in clinical decisions made between doctors, clinics and countries.

To prioritize embryo selection, embryologists typically incorporate scores based on a morphological evaluation of each embryo (Cetinkaya and Kahraman, 2016). Historically, embryos were removed from the incubator for assessment under a microscrope by a trained embryologist one to two times daily. The development of incubators with built-in time-lapse monitoring has enabled non-invasive embryo assessment with comparatively fine-grained detail, inspiring significant interest in applying embryo "morphokinetics" to score and prioritize embryos (Chamayou et al., 2013). Informally, the morphokinetics comprise the timing and morphologic appearance of embryos as they grow and pass through a series of sequential developmental stages, with earlier stages corresponding to cell divisions and subsequent stages corresponding to larger structural milestones, e.g. formation of the blastocyst.

Modern incubators use a high-powered microscope to capture images of a developing embryo approximately every 15 minutes. Currently, embryologists must perform the morphokinetic analysis manually, viewing a sequence of photographs and annotating the time stamps at which each embryo achieves various developmental milestones. These scores are combined according to heuristic formulas to rank the embryos by their putative viability for transfer into a prepared endometrium.

1.1. Clinical Relevance

In this paper, we investigate machine learning techniques for automatically performing morphokinetic analysis on embryo using time-lapse video. We developed a machine learning algorithm that labels embryo developmental stages and detects the transition times. The transition times when the embryo reach different stages can be used in order to score the embryos based on standard heuristic formulas. Our system could potentially help to reduce the human effort required presently to annotate the transition times manually.

Our experiments focus on a dataset consisting of 1309 time-lapse videos extracted from EmbryoScope^{TM}(Vitrolife, Sweden) incubators at the Cleveland Clinic Fertility Center. Each frame in the raw videos has 500×500 resolution. We downsize these to the standard 224×224 ImageNet dimensions for compatibility with pretrained nets and corresponding hyperparameters.

Morphokinetic analysis is the first step towards accessing implantation potential. We hope that the models we proposed for classifying developmental stages could be further extended to solve the problem of embryo selection to maximize transfer viability.

1.2. Technical Significance

In order to build an accurate model for morphokinetic prediction, we propose several methods to (i) learn region proposal models from weak supervision to discard background so that the classifier can focus on the region corresponding to the embryo; (ii) incorporate the temporal context of the video into the model architecture; and (iii) post-process our predictions at the sequence level, using dynamic programming to determine the most likely monotonically-increasing sequence of morphokinetic stages.

Compared to a baseline deep residual network (ResNet) (He et al., 2016), we find significant benefits from each of our three proposed techniques. Our region proposal network selects a 112×112 region and is optimized by reinforcement learning, following the policy gradient algorithm, using the cross entropy loss of the downstream classifier as a reward signal. This technique improves frame-level accuracy from 83.65% to 86.34%. Adding a Long Short-Term Memory (LSTM) recurrent neural network to post-process the predictions, we achieve an additional gain, boosting frame-level accuracy to 88.28%. Finally, we evaluate two variants of our dynamic programming technique for decoding monotonic predictions, one based on finding the *most likely* monotonically increasing sequence, and another that minimizes the expected distance between the predicted and actual states (e.g. when the ground truth is 'stage 4', we prefer to predict stage 3 over stage 2). When applied to the raw classifier, the dynamic programming post-processing methods confer improvements to frame-level accuracy of 85.54% and 86.34%, respectively. Notably, our techniques yield complementary benefits: altogether, they combine to achieve a frame-level accuracy of 90.23%, and reduce the transition-level error (number of frames off) of the raw classifier by 30% to 4.37.

2. Dataset

The EmbryoScope time-lapse system is an embryo incubator capable of holding up to 12 wells simultaneously, each containing one embryo. Built in to the device are both a high-powered microscope and a camera used jointly to photograph each embryo on a 15-minute cycle. Each frame of the resulting time-lapse video consists of a 500×500 resolution grayscale image with a well number in the lower-left corner and the time superimposed in the lower-right corner, as seen in Figure 1. The system also captures each image in multiple focal planes, although the central focal plane alone is used in this study.
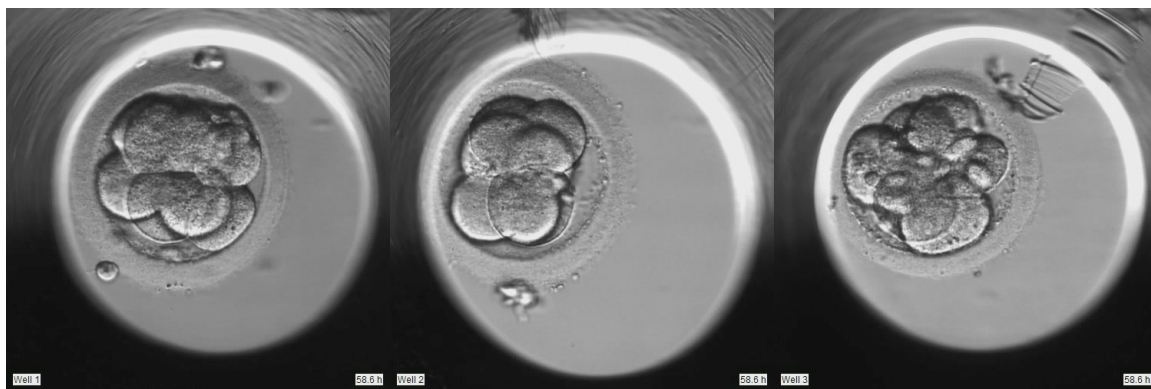
Figure 1: A sample frame containing 3 wells in an EmbryoScope video from our dataset.

Our dataset consists of 1309 EmbryoScope time lapse videos extracted from incubators at Cleveland Clinic Fertility Center. These videos span 113 different patients, each with 11 to 12 wells and corresponding videos. Videos begin roughly 18 hours after fertilization, and end roughly 140 hours after fertilization. Annotations in the videos correspond to 15 distinct morpho-kinetic stages, with the embryologist marking the time at which each embryo was first observed in each developmental stage. Among cells that mature successfully (the rest are discarded), the stages are monotonically increasing, meaning that among non-discarded embryos the ground truth labels never regress from a more advanced stage to a less mature stage. We transform our stage transition annotations into per-frame stage labels by applying the most recently assigned stage. We also assign a special **tStart** stage label to frames before the first stage. The first observed stage corresponds to the moment when two pronuclei are visible (**tPNF**), and the next several stages correspond to cell divisions. After the embryo reaches 9+ cells, the subsequent stages correspond to higher-level features, namely the formation of the morula (**tM**), start of blastulation (**tSB**), blastocyst (**tB**), expanded blastocyst (**tEB**) and hatched blastocyst (**tHB**). We present the distribution of embryo stages in Figure 2.

Because most embryo selection heuristics depend only on the time to reach the cell division milestones, in this study, we focus on the first six stages of development for each embryo, cutting off each video at hour 60. Moreover, these stages admit a cleaner problem because for a significant portion of our videos, expert (ground truth) annotations are missing for the latter 9 stages. The 6 stages that we address include the initial stage (`tStart`), the appearance and breakdown of the male and female pronucleus (`tPNf`), and the appearance of 2 through 4+ cells (`t2, t3, t4, t4+`). Any stage after `t4` within first 60 hour of videos is labeled as stage `t4+`. Among these frames, the class distribution is 11.48%, 6.11%, 25.70%, 4.36%, 25.85%, 26.50%. Because we want to be sure that our models generalize not only across frames, or even across embryos but also across patients (mothers), we stratify the dataset by patient, creating training/validation/test splits by randomly selecting 93/10/10 patients and their respective wells. This yields 510/60/56 embryos in the respective splits, corresponding to 117553/14606/13305 frames.
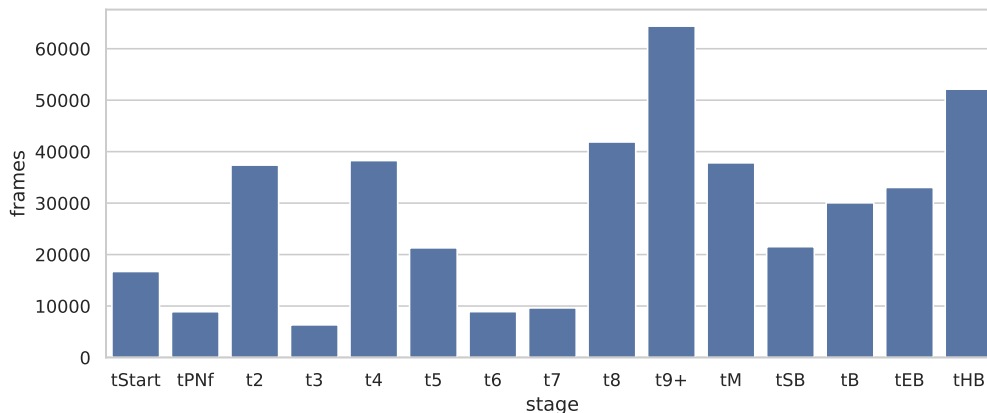
Figure 2: Distributions of morphokinetic stages across all labeled video frames.

3. Methods

We cast predicting embryo morphokinetics as a multiclass classification problem, where the input is a time-lapse EmbryoScope$^{\text{TM}}$video, and the output is a sequence of labels indicating the predicted stage of the embryo at each frame of the video.

Our simplest method consists of applying standard image recognition tools to predict the stage label y_t for each frame given the image x_t. Image classification is now a mature technology, and for all known related tasks, the current best-performing methods are deep convolutional neural networks (CNNs). All of our approaches are based upon convolutional neural networks. Specifically, we choose the ResNet-50 architecture as our base model due to (He et al., 2016). By default, this model takes as input a 224×224 resolution image which we downsize from the original 500×500 image. The output consists of a 6-dimensional softmax layer corresponding to the 6 class labels, and we optimize the network in the standard fashion to minimize cross entropy loss. We initialize the network with pre-trained weights learned on the benchmark ImageNet image recognition challenge (Russakovsky et al., 2015), a practice widely known to confer significant transfer learning benefits (Yosinski et al., 2014). We suspect that given a more relevant source task with a comparably-large dataset (ideally, concerning gray-scale images from cellular microscopy), we might get even greater benefits, although we leave this investigation for future work.

3.1. Weakly-Supervised Embryo Detection

Motivating our first contribution for improving performance of the embryo classifier, we observe that the embryo's cell(s) lie in a small region of the image, and that the rest of the image, containing the rest of the well and surrounding background consists only of imaging artifacts that have no relevance to stage prediction. We postulate that by first detecting where the embryo is, and then subsequently basing classifications on the cropped region containing only the embryo, we could filter out the background noise, improving predictive

performance. Moreover, since the subsequent classification is based on a smaller region, we could either (i) save computation, or (ii) refer back to the original image to extract a higher-resolution zoom on the cropped region, providing greater detail to the classifier.

The most standard way to cast the bounding box detection task is to train a model with labeled data corresponding to the height and width of the box as well as an x and y coordinates to locate the box. For typical detection tasks, current deep learning-based object detection systems require large annotated datasets with bounding box labels. However, we do not have any such labels available for our task.

To learn embryo-encapsulating bounding boxes without explicitly annotated boxes, we propose a new approach that relies only on image-level class labels, optimizing the region proposal model via weak supervision using reinforcement learning. To begin, noting that the embryo size does not vary much, we fix the box dimensions to 112×112 (a $.5 \times .5$ crop), focusing only on identifying the box center. Since we only have the image-level label for the image classification task, the training objective of the detector G is to help a downstream classifier F to better classify the image. Our two-step detect-then-classify algorithm is described below:

1. Given an input image x, the detector predicts a probabilistic distribution $G_\theta(\, \cdot \, ; x)$ over a 14×14 rectangular grid of candidate box centers.

2. Sample a region $R \sim G_\theta(x)$ and get the cropped subregion x_R.

3. Feed x_R to the classifier to predict probabilities for each class $F_\phi(x_R)$.

Let y be the label and ℓ be the usual cross entropy loss function. The expected loss of the two-step classification algorithm is

$$\mathcal{L}_{cls}(\theta, \phi) = \mathbb{E}_{R \sim G_\theta(\, \cdot \, ; x)}[\ell(F_\phi(\text{crop}(x, R)), y)] \tag{1}$$

Note that both the detector G and classifier F share the objective of minimizing the expected classification loss $\mathcal{L}_{cls}$. The intuition behind this objective is that if the image crop has a larger intersection with the cell, it is easier for the classifier to classify the image. On the other hand, if a large part of the cropped image is background, the classifier should not perform much better than random guessing. Note that our detector outputs a probability distribution over grid-cells. At test time we make predictions by centering the bounding box at the expected x and y coordinates.

The loss function involves computing the expectation with respect to all possible regions. We use the Monte Carlo method to estimate the loss by drawing K sample regions $R_1, R_2, \cdots, R_K \sim G_\theta(\, \cdot \, ; x)$. The optimization problem becomes

$$\arg\min_{\theta, \phi} \sum_{i=1}^{K} \ell(F_\phi(\text{crop}(x, R_i)), y). \tag{2}$$

The gradient for the classifier's parameter ϕ is

$$\nabla_\phi \mathcal{L}_{cls}(\theta, \phi) = \frac{1}{K} \sum_{i=1}^{K} \nabla_\phi \ell(F_\phi(\text{crop}(x, R_i)), y). \tag{3}$$
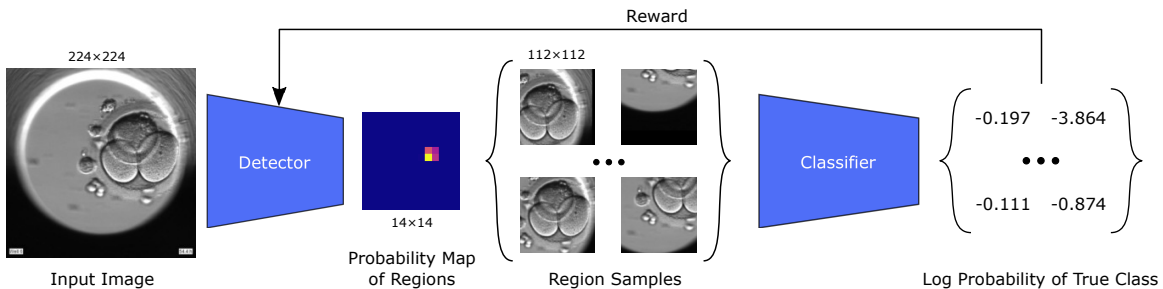
Figure 3: The pipeline of learning a detector in a weakly-supervised way using reinforcement learning.

The gradient for the detector's parameter θ is estimated using the policy gradient, a common reinforcement learning algorithm. Moreover, we incorporate a standard technique for variance reduction, use average rewards as a baseline b. This gives us the gradient for θ,

$$b = \frac{1}{K} \sum_{i=1}^{K} \ell(F_\phi(\text{crop}(x, R_i)), y) \tag{4}$$

$$\nabla_\theta \mathcal{L}_{cls}(\theta, \phi) = \frac{1}{K} \sum_{i=1}^{K} [\ell(F_\phi(\text{crop}(x, R_i)), y) - b] \nabla_\theta \log G_\theta(R_i; x). \tag{5}$$

In preliminary experiments, we found that solely relying on the objective (2) converges quickly to an unsatisfactory local optimum where the distributions of regions $G_\theta(\,\cdot\,; x)$ are always peaked on one specific region proposal. To overcome this issue, we encourage exploration in the reinforcement learning objective by adding the negative entropy of the region distribution $G_\theta(\,\cdot\,; x)$, a technique made by Mnih et al. (2016). The augmented overall loss function is

$$\mathcal{L}(\theta, \phi) = \mathcal{L}_{cls}(\theta, \phi) - \lambda H(G_\theta(\,\cdot\,; x)) \tag{6}$$

where λ is the weight to balance the term.

Network structure The detector predicts the region distribution using a sliding window method based on a Region Proposal Network (RPN) (Ren et al., 2015). The region proposal is computed from a 14×14 intermediate feature map at `conv4_2` in Resnet-50. Based on our exploration of the data, we found that the embryo is typically contained in a rectangular region that is roughly one quarter the size of the image. To simplify the distribution, we fix the width and height of the rectangle region to be 50% of the size of the image using this prior knowledge. We assume that the center of the region proposal lies on 14×14 grid, so that we only need to predict the probability of the region lying at each position in that grid. The probability is computed by applying a 3×3 convolutional filter W to the feature map followed by a softmax operation

$$\gamma = \text{softmax}(W * X_{\texttt{conv4\_2}}), \tag{7}$$

where γ_{ij} indicates the probability of selecting the box center at the i-th row and j-th column of the 14×14 grids.

Our base classifier is a Resnet-50 convolutional network that takes a 112×112 cropped image as its input. We remove the layers in `conv5` to speed up computation.

Test time   While the detector outputs a distribution of regions, at test time we want to use only "the best" region. Some early experiments revealed the heuristic of choosing the expected center coordinates of the predicted distribution. The average box center $(\bar{c}_x, \bar{c}_y)$ is computed by

$$\bar{c}_x = \mathbb{E}_{c_x \sim G_\theta(c_x, c_y; x)}[c_x] \tag{8}$$

$$\bar{c}_y = \mathbb{E}_{c_y \sim G_\theta(c_x, c_y; x)}[c_y]. \tag{9}$$

3.2. LSTM

Our first idea to incorporate context across adjacent frames is to employ recurrent neural networks with Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) units. The LSTM takes as input a sequence of inputs, updates its internal state at each time, and predicts a sequence of outputs. The inputs to the LSTM consist of 2048-dimensional feature vectors extracted from the hidden layers of a vanilla CNN. We then feed the feature vector to a bi-directional LSTM layer with 100 units for each direction. We apply a linear mapping of the LSTM output at each time step to 6 classes to get a sequence of predictions $y^{(1)}, y^{(2)}, \cdots, y^{(T)}$. We set T to 9 optimizing the model to predict accurate on the middle 5 frames. We do not use predictions made on the first 2 or last 2 frames because they lack sufficient context.

3.3. Structured Decoding with Dynamic Programming (DP)

For embryos that successfully reach the blastocyst stage, ground truth stages in our selected data set are monotonically non-decreasing, reflecting the condition that any viable embryo must continue to grow and develop rather than arrest and die. The predictions of frame level CNNs or LSTMs with short sequences cannot learn this constraint since the model does not have enough context. Therefore we impose this inductive bias through a dynamic programming decoder that enforces monotonicity of predictions. For each video, our model predicts the probability of the embryo stages $\hat{p}^{(t)} \in \mathbb{R}^6$ at every frame $t = 1, \cdots, T$, where T is the total number of frames in the video. We want to find a decoded label sequence $\hat{\mathbf{y}}$ such that $\hat{y}^{(t+1)} \geq \hat{y}^{(t)}$ and $\hat{y}^{(t)}$ most match the frame prediction $\hat{p}^{(t)}$ for each frame. We define a potential function $\phi(y^{(t)}, \hat{p}^{(t)})$ to measure how much the decoded label $\hat{y}^{(t)}$ deviates from $\hat{p}^{(t)}$ and turn the decoding to the following optimization problem:

$$\arg\min_{\hat{\mathbf{y}}} \sum_{t=1}^{T} \phi(\hat{y}^{(t)}, \hat{p}^{(t)}) \quad \text{s.t.} \quad \hat{y}^{(t+1)} \geq \hat{y}^{(t)} \ \forall \ t \in \{1, \cdots, T-1\}. \tag{10}$$

We investigate two potential functions, the negative log likelihood (NLL) and the earth mover's distance (EMD), defined by $\phi_{NLL}(\hat{y}, \hat{p}) = -\log(\hat{p}_{\hat{y}})$ and $\phi_{EMD}(\hat{y}, \hat{p}) = \sum_{s=1}^{S} \hat{p}_s |\hat{y} - s|$, respectively, where $S = 6$ is the number of development stages. This optimization problem

can be solved in polynomial time using Dynamic Programming (DP) with a forward pass and a backward pass.

4. Experiments

4.1. Embryo Detection

We train the region proposal network with SGD with momentum 0.9 and learning rate 0.01, with batch size set to 16. The image is first downsampled to 224×224 before feeding into the detector. For each image in the batch, we sample 10 regions, extracted as 112×112 images cropped the from 224×224 input image, and feed the cropped images into the classifier. We train two detectors with and without entropy regularization ($\lambda = 0.01, 0$ respectively) to measure the effect of using the augmented loss function.

We also compare another approach to learn the detector using differentiable bi-linear sampling. The idea is that the detector only predicts a single region that is fed to the classifier. We use differentiable bi-linear sampling when cropping the image at that region so that the gradient with respect to the classification loss can be back-propagated to the detector. We change the last layer of the detector to be a fully connected layer to predict the coordinates c_x, c_y of the center of the box. We were unable to make this alternative approach converge using SGD, so we eventually settled on the Adam optimizer with default parameters.

To evaluate the performance of the learned detector, we manually label a tiny data set with 120 images randomly sampled from the validation set, corresponding to 20 images from each embryo stage, and use these ground truth labels to get a quantitative evaluation of the detector. We report the Jaccard index, which is calculated by the intersection over union between the ground truth box and predicted box, as well as the euclidean distance between the ground truth box center and predicted box center, measured in pixels in the 500×500 raw image. We also include the classification accuracy of a two-step detector-classifier on the selected 120 images as this is our actual training objective.

Detection results are shown in Table 1. Our RL training with entropy loss achieves a Jaccard index of 0.6957 and a center distance only 11.58 pixels from the manually labeled 500×500 images. The detector trained without the entropy term underperforms the detector with entropy, reflecting network convergence to some local optimum based on the current best performing region at an early stage. The differentiable sampling approach performs poorly for detection; this shows that using a stochastic region proposal in our RL training is crucial for successfully training a detector.

We visualize the detection results of a random sample of images in Figure 4. We see that the predicted boxes contain the region of the ground truth box in almost all images and are only fractionally larger than the ground truth boxes.

Table 1: Quantitative results of various detector training methods.

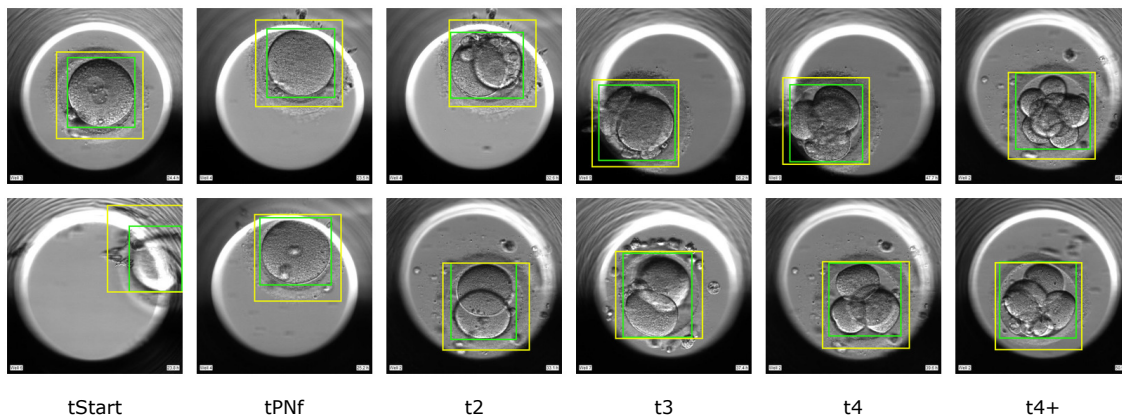| Training Method	Jacc. Index	Distance	Accuracy
RL w/ Entropy Loss	**0.6957**	**11.58**	**82.50%**
RL w/o Entropy Loss	0.6876	18.83	75.83%
Differentiable Sampling	0.0779	228.5	71.67%

tStart   tPNf   t2   t3   t4   t4+

Figure 4: Visualization of detection results by model trained using RL with entropy loss. The green boxes are ground truth bounding boxes and the yellow boxes are network predictions.

4.2. Embryo Staging

The baseline model is ResNet-50 applied to raw image resized to 224×224. Our method (*DetCls*, 'detect then classify') first uses the detector learned in the previous section to identify the region of the embryo on 224×224 resized input. We experiment with two image cropping methods. The first method crops the 112×112 region on the resized input, while the other crops the raw 500×500 image and resizes it to 224×224. The cropped image with size 112×112 or 224×224 will then be fed into the same ResNet-50 as the baseline. We also try to add LSTM to our DetCls method with 224×224 crops.

After successfully training the detector in an end-to-end manner, we subsequently use the same detector to compare all downstream models, The detector is set to test mode to predict only one region. We initialize each classifier as a ResNet-50 with pretrained ImageNet weights, updating all weights using the Adam optimizer with a learning rate of 0.001 and default parameters. The validation data set is used for early stopping and all metrics are evaluated on test data.

Data augmentation has often proven critical for achieving competitive results in deep learning. For both the baseline and our proposed method, we apply random rotation augmentation. The rotation angle is randomly chosen from 0 to $360°$.

Table 2:   Quantitative results for various architectures and output decoding schemes.

Model	Raw. Acc.	DP: label likelihood s.t. monotonicity			DP: earthmover's distance s.t. monotonicity		
		Accuracy	MAE	RMSE	Accuracy	MAE	RMSE
Naïve Baseline	66.87%	-	-	-	-	16.42	24.34
ResNet Baseline	83.65%	85.54%	6.582	17.26	86.34%	6.288	16.72
DetCls112	86.34%	89.63%	4.594	13.38	89.13%	4.842	13.98
DetCls224	86.18%	89.34%	4.842	14.10	89.79%	4.452	**12.97**
DetCls224+LSTM	**88.28%**	89.92%	4.688	14.47	**90.23%**	**4.370**	13.67

We report the per-frame accuracy of our raw predictions as well as the per-frame accuracy of the DP predictions (for both objectives). We also report the mean absolute error (MAE) and root mean squared error (RMSE) (measured in frames) of the predicted stage transition times after post-processing. To better justify these results, we include the result of a naive baseline that simply labels each frame using the mode stage among all frames captured at the same time in the training set, and predicts the transition time for each stage using the median transition time among all embryo videos in the training set. Table 2 summarizes the results of four models.

Effect of detection.   Two of our single frame DetCls models significantly outperformed the baseline before and after post-processing in all metrics. Of note, the gain in accuracy due to detection *after post-processing* is typically as great or greater than the gains seen in raw accuracy (without post-processing). The performance of DetCls112 and DetCls224 is comparable. The model with high resolution cropping performs only slightly better after post-processing, suggesting that the performance gains with respect to the baseline are mainly due to removing irrelevant background in the raw input and not due to enabling higher-resolution inputs.

Using temporal information.   DP post-processing yields an accuracy improvement of 2.69% to 3.61% to all three single-frame models and allows us to generate a monotonic prediction sequence to predict the stage transition time. DP using Earthmover's distance achieves slightly better performances on three models (Baseline, DetCls224, DetCls224+LSTM) than DP using likelihood. Adding LSTM to DetCls224 further improves the raw accuracy and two metrics (accuracy, MAE) after post-processing. The improvement is less significant after post-processing. This suggests that the DP decoders already encode most of the temporal relationships between frames.
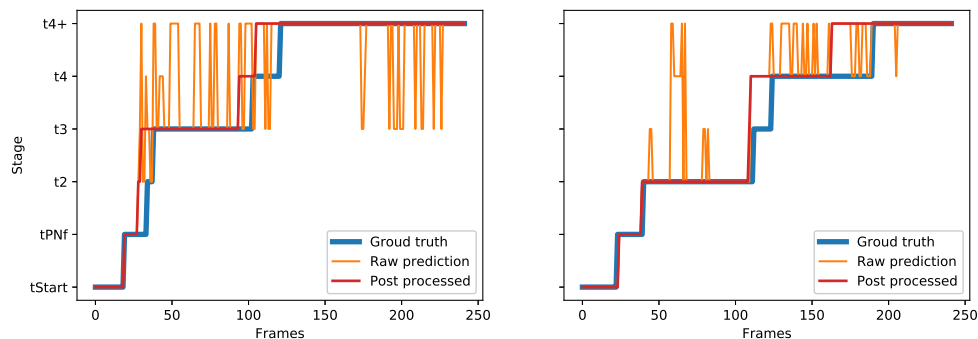
Figure 5: The results of DetCls224 model on two embryo videos, before and after DP post-processing using the earth mover's distance. Post-processing removes the fluctuations in raw predictions to obtain a smooth (monotonic) result.

5. Related Work

5.1. Computer Vision Methodology Papers

Over the past several years, a variety of papers have made rapid progress on both single and multiple object detection using convolutinoal neural networks. Two of the most popular approaches are Faster R-CNN (Ren et al., 2015) and YOLO (Redmon et al., 2016), from which we draw loose inspiration in designing our region proposal network for predicting the bounding box. Traditional object recognition methods are trained on large datasets where the true bounding boxes are annotated, data which is not freely available in many domains, including ours. Several previous works seek to address this problem, learning weakly-supervised object detection that relies only on the image-level class label (Bilen et al., 2015; Bilen and Vedaldi, 2016). Unlike our method, these approaches are not end-to-end trainable with a classifier.

Ba et al. (2015), Sermanet et al. (2014) and Yeung et al. (2016) use reinforcement learning to learn an attention mechanism for selecting most relevant image regions or video frames for downstream visual recognition tasks. Jaderberg et al. (2015) propose an alternative method that learns a geometric transform on the input image, which is fed into a classifier using differentiable bi-linear sampling. Our work shares a similar idea of localizing the object before classifying it.

The idea of extending models to include temporal information has been explored extensively in recent years. Simonyan and Zisserman (2014) used a two-stream architecture applied to a single frame as well as multi-frame optical flow in order to combine spatial and temporal information. Tripathi et al. (2016) studied techniques for using RNNs to improve frame-level object detection by incorporating context from adjacent frames. They also introduce several additional losses, e.g. to encourage smoothness in the predictions across adjacent frames.

5.2. Embryology Applications Papers

The problem of predicting embryo annotations from time lapse videos has been addressed in the literature by Khan et al. (2016). They use an 8-layer convolutional network to count the number of cells in an embryo image (up to 5 cells), a related but different setting from ours. To incorporate temporal information, they use conditional random fields and similarly use dynamic programming to enforce monotonicity constraints. Milewski et al. (2016) collected annotations of time-lapse morphokinetic data and used principal component analysis and logistic regression analysis to predict pregnancy with an AUC of 0.70. Pribenszky et al. (2017) performed a meta-analysis of RCTs comparing use of a morphokinetic algorithm versus single time-point embyro evaluations and found an improved ongoing clinical pregnancy rate with use of the technology. Zaninovic et al. (2018) demonstrated that there is significant variability in some morphokinetic intervals between IVF clinics, suggesting that the parameters used to select embryos may require tuning for each particular clinic. The human-selected morphorkinetic annotations are in near perfect agreement across repeated exams (Adolfsson and Andershed, 2018)

Multiple applications of CNNs to embryo assessment were presented at the 2018 American Society of Reproductive Medicine Annual Meeting. Zaninovic et al. (2018) analyzed 18,000 images of blastocyts using CNNs trained on raw time-lapse images and was able to classify the quality of the embryos into three morphologic quality grades with 75% accuracy. Iwata et al. (2018) performed a similar analysis to predict good-quality embryos with 80% accuracy. Malmsten et al. (2018) built a CNN using raw time-lapse from images of 11,898 human embryos to classify up to the 8-cell stage with 82% reported accuracy. They also reported that the cell-division transition times predicted within 5 frames of when the embryologist annotates the transition for 91% of transitions.

To our knowledge, our work is the first to use deep learning to predict embryo morphokinetics (the above works were published after our work was first made public), the first to improve performance by localizing the embryo through a weakly-supervised reinforcement learning method, and the first to demonstrate the benefit of incorporating contextual frames via LSTMs.

Beyond embryology, CNN-based classification techniques have emerged as popular tools in the clinical literature, with successes such as image-based classification of skin lesions (Esteva et al., 2017) including keratinocyte carcinomas versus benign seborrheic keratoses and malignant melanomas versus benign nevi, and the detection of diabetic retinopathy from retinal fundus imaging (Gulshan et al., 2016).

6. Conclusion

This paper introduced a suite of techniques for recognizing stages of embryo development, achieving 90.23% frame-level accuracy. We also achieve a mean average error for predicting stage transition times of 4.370. We believe that several directions realizable in future work could bring this technology to the level of clinical utility. To begin, our results are achieved using only 510 embryos for training. Given that performance of deep learning methods

tends not to saturate quickly with dataset size, we plan to access a considerably larger dataset for future studies, to test the limits of our current methodology. Additionally, our models are initialized by using pretrained weights from an ImageNet classifier originally on full color photographs. We suspect that transfer from a comparably large dataset of more relevant images (gray-scale microscopy of cells) might yield additional gains. Identifying such a dataset for transfer remains a challenge. Moreover, even if we can access such a dataset of unlabeled images, deciding upon a (possibly unsupervised) objective for the source task could pose an interesting research problem. Additionally, we plan to extend our experiments to predict not only the 6 stages useful for current embryo selection heuristics but to predict all 15 stages of development. And finally, we hope to use the models learned from the morphokinetic predictions as themselves a source task, fine-tuning the models to the more pressing downstream problem of assessing implantation potential directly. We note that assessing the viability of an embryo represents an interesting off-policy learning problem. Outcomes are only observed for those embryos that were transferred. Success on this task may require not only representation learning, but also estimating counterfactual quantities.

Acknowledgments

References

E Adolfsson and AN Andershed. Morphology vs morphokinetics: a retrospective comparison of inter-observer and intra-observer agreement between embryologists on blastocysts with known implantation outcome. *JBRA Assist Reprod*, pages 228–237, 2018.

Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. In *Proceedings of the International Conference on Learning Representations*, 2015.

Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with convex clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1081–1089, 2015.

Center for Disease Control. Assisted reproductive technology (art). https://www.cdc.gov/art/artdata/index.html, 2016. Accessed: 2019-03-27.

Caroline Pirkevi Cetinkaya and Semra Kahraman. Morphokinetics of embryos - where are we now? *Journal of Reproductive Biotechnology and Fertility*, pages 1–8, 2016.

Sandrine Chamayou, Pasquale Patrizio, Giorgia Storaci, Venera Tomaselli, Carmelita Alecci, Carmen Ragolia, Claudia Crescenzo, and Antonino Guglielmino. The use of morphokinetic parameters to select all embryos with full capacity to implant. *Assisted Reproductive Genetics*, pages 703–710, 2013.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997.

K Iwata, M Sato, I Matsumoto, T Shimura, K Yumoto, A Negami, and Y Mio. Deep learning based on images of human embryos obtained from high-resolusion time-lapse cinematography for predicting good-quality embryos. *Fertility and Sterility*, 110(4):e213, 2018.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

Aisha Khan, Stephen Gould, and Mathieu Salzmann. Deep convolutional neural networks for human embryonic cell counting. In *ECCV*, 2016.

J Malmsten, N Zaninovic, Q Zhan, M Toschi, Z Rosenwaks, and J Shan. Automatic prediction of embryo cell stages using artificial intelligence convolutional neural network. *Fertility and Sterility*, 110(4):e360, 2018.

R Milewski, AJ Milewska, A Kuczyska, B Stankiewicz, and Kuczyski W. Do morphokinetic data sets inform pregnancy potential? *Assisted Reproductive Genetics*, pages 357–365, 2016.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

C Pribenszky, AM Nilselid, and M Montag. Time-lapse culture with morphokinetic embryo selection improves pregnancy and live birth chances and reduces early pregnancy loss: a meta-analysis. *Reprod Biomed Online*, pages 511–520, 2017.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for fine-grained categorization. *arXiv preprint arXiv:1412.7054*, 2014.

Alexander L Simon, Michelle Kiehl, Erin Fischer, J Glenn Proctor, Mark R Bush, Carolyn Givens, Matthew Rabinowitz, and Zachary P Demko. Pregnancy outcomes from more than 1,800 in vitro fertilization cycles with the use of 24-chromosome single-nucleotide polymorphism–based preimplantation genetic testing for aneuploidy. *Fertility and sterility*, 110(1):113–121, 2018.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

Subarna Tripathi, Zachary C Lipton, Serge Belongie, and Truong Nguyen. Context matters: Refining object detection in video with recurrent neural networks. *arXiv preprint arXiv:1607.04648*, 2016.

Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2678–2687, 2016.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

N Zaninovic, P Khosravi, I Hajirasouliha, JE Malmsten, E Kazemi, Q Zhan, M Toschi, O Elemento, and Z Rosenwaks. Assessing human blastocyst quality using artificial intelligence (ai) convolutional neural network (cnn). *Fertility and Sterility*, 110(4):e89, 2018.