Embryo Grading With Unreliable Labels Due to Chromosome Abnormalities by Regularized PU Learning With Ranking

Masashi Nagaya and Norimichi Ukita, *Member, IEEE*

***Abstract*— We propose a method for human embryo grading with its images. This grading has been achieved by positive-negative classification (i.e., live birth or non-live birth). However, negative (non-live birth) labels collected in clinical practice are unreliable because the visual features of negative images are equal to those of positive (live birth) images if these non-live birth embryos have chromosome abnormalities. For alleviating an adverse effect of these unreliable labels, our method employs Positive-Unlabeled (PU) learning so that live birth and non-live birth are labeled as positive and unlabeled, respectively, where unlabeled samples contain both positive and negative samples. In our method, this PU learning on a deep CNN is improved by a learning-to-rank scheme. While the original learning-to-rank scheme is designed for positive-negative learning, it is extended to PU learning. Furthermore, overfitting in this PU learning is alleviated by regularization with mutual information. Experimental results with 643 time-lapse image sequences demonstrate the effectiveness of our framework in terms of the recognition accuracy and the interpretability. In quantitative comparison, the full version of our proposed method outperforms positive-negative classification in recall and F-measure by a wide margin (0.22 vs. 0.69 in recall and 0.27 vs. 0.42 in F-measure). In qualitative evaluation, visual attentions estimated by our method are interpretable in comparison with morphological assessments in clinical practice.**

***Index Terms*— Deep convolutional networks, positive-unlabeled learning, learning-to-rank, mutual information.**

I. Introduction

IN AN artificial fertilization process, medical doctors select good embryos, each of which has a high probability of live birth, based on their visual features. This process requires expert skill because several embryo images of live birth and non-live birth are similar to each other. Furthermore, we still have insufficient knowledge about visual features for this classification [1]–[4]. The goal of this work is to identify embryos each of which has a high probability of live birth by visual features extracted from each embryo image. While Deep Neural Networks (DNNs) improve such classification, general DNNs require supervised training data. As with other problems, the classification of embryos is achieved with supervised data [5]–[7]. In these papers, (I) the visual features of embryo images are labeled by medical doctors [6] or (II) a birth result (i.e., live birth or non-live birth) is used as a class label for supervised learning [5]. However, both schemes have the following unreliabilities:

1) **Manual annotation [6], [7]:** Medical doctors may give inconsistent labels. This may happen due to time-consuming difficult annotations for the embryo grading.
2) **Supervision by birth results [5]:** In general, embryos do not develop normally if they are visually graded lower. However, even if an embryo has visual features that are observed in those resulting in live birth, this embryo cannot develop normally if it has chromosome abnormalities [8]–[10]. That is, visual features are shared between live birth and non-live birth. This overlap results in difficulty in utilizing the birth result as a label for supervised learning.

To avoid these two unreliabilities, our method employs semi-supervised classification with Positive-Unlabeled (PU) learning [11]–[15]. This approach is further improved by two more contributions, namely PU learning with efficient ranking-based objectives and PU learning with unsupervised regularization. Furthermore, the interpretation of our classification result is visualized for supporting medical doctors. These fourfold contributions are summarized as follows:

- **PU learning for unreliable samples:** Since many medical image problems have a small amount of supervised training data, semi-supervised learning is widely used [16]. PU learning [17] as semi-supervised learning is useful for learning a limited number of labeled samples [11]–[15]. While an original scenario for PU learning is that labeled positive samples and unlabeled samples are provided for training, our proposed method employs PU learning for suppressing an adverse effect of unreliably-labeled samples.

Manuscript received July 6, 2021; revised August 24, 2021; accepted September 6, 2021. Date of publication November 8, 2021; date of current version February 2, 2022. This work was supported in part by the Toyota Riken Scholarship. *(Corresponding author: Norimichi Ukita.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Nagoya City University under Application No. 60-18-0111.

The authors are with the Graduate School of Engineering, Toyota Technological Institute, Nagoya 468-8511, Japan (e-mail: dc.03290122@gmail.com; ukita@toyota-ti.ac.jp).

Digital Object Identifier 10.1109/TMI.2021.3126169

- **PU learning with efficient ranking-based objectives:** Only a few PU learning methods are developed for DNNs (e.g., [13], [14]), because of difficulty in designing proper differentiable loss functions. We improve PU learning on DNNs with learning-to-rank, which is validated with classic machine learning methods [18].
- **PU learning with unsupervised regularization:** Overfitting is a major problem in PU learning [13]. We tackle this problem with maximization of mutual information (MI) [19]. Unlike [19], an end-to-end PU learning jointly with all other losses is proposed.
- **Visual interpretations of time-lapse embryo images:** Interpretability [20], [21] of DNNs is important in clinical practice [22]–[24] as well as other real-world problems. Our method refines such interpretability in time-lapse images by a simple smoothness loss.

II. Related Work

A. DNNs for Blastocyst Scores Given by Medical Doctors

As with many medical image problems, DNNs are useful for the diagnosis of human preimplantation embryo viability. In the current clinical practice, morphological assessments such as Veeck criteria [25] and Gardner criteria [26] are used in general. With the blastocyst score of each sample image provided by medical doctors, DNNs are trained in [5], [6], [27], [28], However, the blastocyst scores given by medical doctors are inconsistent and insufficient for a better embryo selection process.

B. DNNs for Live-Birth Probabilities

While real live birth and non-live birth embryo images are used as training samples in [5], standard DNNs are just utilized in these papers for feasibility studies. It is revealed that these standard DNNs yield very low recall values (e.g., recall = 0.148 in [29]). Such low recall values are unacceptable in clinical practice. One of the major reasons for the performance degradation is unreliable and noisy labels included in live birth and non-live birth results, which are unavoidable due to chromosome abnormalities in embryos. On the other hand, we propose a training scheme that is more appropriate for a realistic case where reliable labels are unavailable. The properties of these methods including our method, which are described in this paragraph, are summarized in Table I.

C. Semi-Supervised Learning for a Small Amount of Training Data

Since many medical problems have a small amount of supervised data, semi-supervised learning is useful [16]. In general semi-supervised learning, a model trained by supervised data (e.g., training data with positive and negative labels) is given. This model is used for classifying unlabeled data, and then is re-trained by the classified unlabeled data.

D. PU Learning as Semi-Supervised Learning

Unlike the aforementioned semi-supervised learning, PU learning [11] assumes that only a part of the positive data is labeled and the remaining positive and all negative data are unlabeled. See a survey paper [17]. PU learning works better than Positive-Negative (PN) supervised learning empirically [13] and in theory in a specific condition [12]. This advantage motivates us to use PU learning (e.g., vascular lesion detection [30], ROI localization [31] and video and volume segmentation [32]). While these methods simply follow a general assumption of PU learning (i.e., only a part of positives are labeled in training data), our proposed method employs PU learning for suppressing an adverse effect of unreliably-labeled data. While a huge number of methods are proposed in order to cope with unreliably-labeled data (a.k.a training data with noisy labels), they focus on loss functions [33], noise modeling/estimation [34], and clean data selection [35]. Furthermore, all of these previous methods are designed for PN learning. On the other hand, our proposed method is designed so that PN learning with noisy labels is regarded as PU learning.

However, due to the complexity of loss functions for PU learning, its performance is still limited for DNNs. For example, only a single type of loss, composite loss [36], is applicable to DNNs in [15]. A non-negative risk estimator, which estimates the risk of misclassification, proposed in [13] allows us to use any losses for DNNs, while its performance is limited due to its heuristic design. Instead of any risk estimator, generalized expectation criteria [37] for posterior regularization are used as an additional constraint in [14].

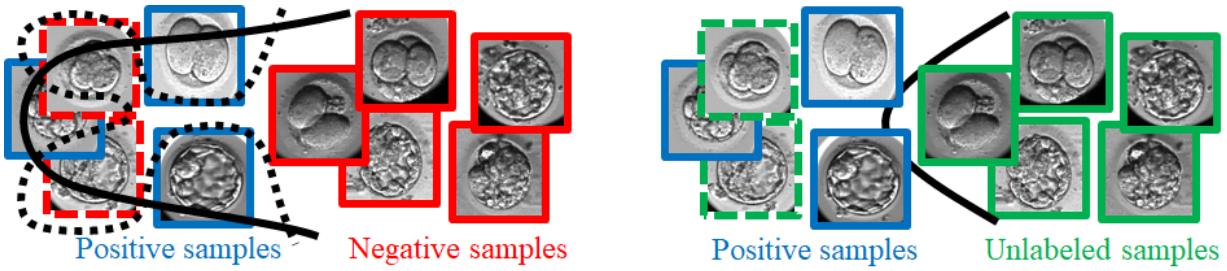
While PU learning is achievable by carefully-designed loss functions and risk estimators as mentioned above, PU learning can be improved also by reducing it to a ranking problem defined by meaningful multivariate performance measures [11] such as the balanced accuracy (i.e., Area Under Curve, AUC), the precision-recall product (i.e., F-measure), and the mean average precision (mAP) [11]. However, the AUC loss cannot be used directly for DNNs because it is indifferentiable. This problem is avoided by approximating the AUC loss with differentiable ones [38]. While these approximate losses are poor in performance and computationally expensive, the AUC loss can be represented by more efficient approximations using upper/lower bounds [39].

E. Regularized PU Learning

One of the major problems in PU learning is overfitting, while unlabeled training data can be utilized for regularization [40]. For example, unbiased PU learning [36] tends to fall into overfitting so that most unlabeled data is classified as negatives, as validated in [13] where a heuristic solution for overfitting is proposed. While overfitting can be suppressed by regularization, powerful regularization techniques such as the contrastive loss [41] and the triplet loss [42] cannot be used in PU learning because these techniques require all training data to be labeled. Among various approaches for regularization such as those with KL divergence, entropy, and self-supervised teacher-student learning, mutual information [19] is employed in our proposed method because mutual information outperforms others in several scenarios, as demonstrated in [19].

TABLE I
SUMMARIZED PROPERTY COMPARISON BETWEEN PREVIOUS MACHINE LEARNING-BASED APPROACHES AND OUR PROPOSED METHOD. FROM A VIEWPOINT OF MACHINE LEARNING, BETTER PROPERTIES ARE COLORED WITH RED

	Score given by doctors [6]	Live birth and non-live birth classification by PN learning [5]	Live birth and non-live birth classification by PU learning (Ours)
Goal	Imitation of medical doctors	Prediction	Prediction
Task difficulty	Easy	Difficult	Difficult
Annotation cost	Large	Small	Small
The number of samples	Large	Small	Small
Adverse effect of chromosome abnormalities	No	Yes	Decreased



(a) **Positive-Negative** (PN) learning (b) **Positive-Unlabeled** (PU) learning

Fig. 1. PU learning for unreliable negative labels in embryo classification. Blue, red, and green rectangles indicate positive, negative, and unlabeled samples, respectively. In (a) PN learning, samples labeled negative due to chromosome abnormalities (i.e., red dashed rectangles) are mixed with positive samples in the feature space. In (b) PU learning, such negative samples (indicated by green dashed rectangles) are regarded as positive.

F. Visual Interpretability of DNNs

Interpretability (a.k.a. visual attention) of DNNs [20], [21] is a major issue especially in clinical practice (e.g., Alzheimer's disease classification [22], 3D imaging data [23], and pathology localization [24]). While these methods visualize attentions in a still image, our method improves the consistency of spatio-temporal visual attentions in time-lapse embryo images.

III. PROPOSED METHOD

Section III describes the proposed method as follows:

- **Section III-A** explains our basic strategy using PU learning for coping with unreliable positive-negative labels observed due to chromosome abnormalities in embryos.
- **Section III-B** introduces two existing methodologies closest to our method.
- **Section III-C** proposes our PU learning using learning-to-rank extended by overfitting suppression.
- **Section III-D** describes how our PU learning exploits unsupervised metric learning for regularization.
- **Section III-E** shows how we improve the interpretability of classification results with temporal consistency.

A. PU Learning for Unreliable Labels

Live birth and non-live birth embryos are labeled as positive and negative, respectively. As mentioned before, an embryo results in non-live birth (i.e., negative) independently of its visual features, if it has chromosome abnormalities [8]–[10]. Since such negative samples may have visual features that are observed also in positive samples, PN learning leads to poor performance on test data both with generalized and overfitted boundaries, which are depicted by black solid and black dashed curves in Figure 1 (a), respectively.

In our method, only positive labels are used, and all non-live birth samples are regarded as unlabeled, as illustrated in Figure 1 (b). These unlabeled samples may include both visually-positive and visually-negative samples, which are indicated by green dashed and green solid rectangles, respectively. Assume that these visually-positive non-live birth images are fewer than visually-negative non-live birth images,[1] the visually-positive non-live birth images can be regarded as noisy outliers. Our proposed method neglects these noisy outliers for avoiding undesirable overfitting, as illustrated in Figure 1 (b). Under this assumption, we expect that our strategy with PU learning improves the classification performance stochastically rather than PN learning, even if the visually-positive non-live birth embryos are misclassified to positive.

What happens with our strategy in clinical practice? For implantation, in general, it is important not to miss any embryos that have a high probability of live birth even though false-positive embryos are detected. This is why our strategy with PU learning is better than PN learning in terms of clinical practice as well as classification performance.

B. Existing PU Learning Methods Improved by Overfitting Suppression and Rank Learning

In [13], DNN (denoted by $N(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^c$ where $\boldsymbol{x}$ is a d-dimensional query and c denotes the number of classes; $c = 2$ for binary classification) is trained by the gradient of a risk estimator, R, instead of directly using a loss function. Given $N(\boldsymbol{x})$:

$$R(N(\boldsymbol{x})) = \pi_+ R_+^+(N(\boldsymbol{x})) + \max\left\{0, R_u^-(N(\boldsymbol{x})) - \pi_+ R_+^-(N(\boldsymbol{x}))\right\},$$

[1]The literature [1] reported that only 6.8 % embryos each of whose morphological quality is excellent are aneuploid.

$$R_+^+(N(\boldsymbol{x})) = \frac{1}{|S^+|}\sum_{\boldsymbol{x}\in S^+}\ell_s\left(N(\boldsymbol{x}),+1\right),$$
$$R_+^-(N(\boldsymbol{x})) = \frac{1}{|S^+|}\sum_{\boldsymbol{x}\in S^+}\ell_s\left(N(\boldsymbol{x}),-1\right),$$
$$R_u^-(N(\boldsymbol{x})) = \frac{1}{|S^u|}\sum_{\boldsymbol{x}\in S^u}\ell_s\left(N(\boldsymbol{x}),-1\right), \quad (1)$$

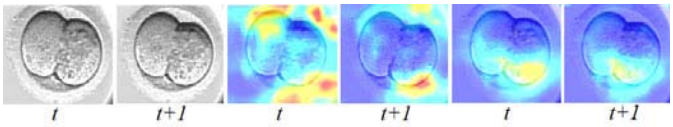
where $R_{c_q}^{c_p}(N(\boldsymbol{x}))$ denotes the following marginal probability of samples labeled as c_q. S^+ and S^u are a set of positive and unlabeled samples, respectively. Given a mini-batch for learning, $R_{c_q}^{c_p}(N(\boldsymbol{x}))$ is the mean of losses, ℓ_s, over each mini-batch in the case that a sample labeled as c_q is predicted to be c_p by $N(\boldsymbol{x})$; $c_p \in \{+,-\}$ and $c_q \in \{+,u\}$ where $+$, $-$, and u denote positive, negative, and unlabeled, respectively. We use the sigmoid function as ℓ_s based on the holistic analysis shown in [13]. π_+ denotes the positive-class-prior probability. $R_u^-(N(\boldsymbol{x})) - \pi_+ R_+^-(N(\boldsymbol{x}))$ in the second term is a transformation of the risk of negative samples, R_-^-. This second term is clipped by zero, in order to avoid overfitting where all unlabeled samples are regarded as negative samples. The risk estimator (1) is named the non-negative PU (nnPU) loss after the zero clipping. Refer to [13] for more details. While the effectiveness of nnPU learning is validated with simple shallow networks, (e.g., 6-layer perception and 13-layer CNN at most), the risk estimator expressed by Eq. (1) is based on a heuristic (i.e., zero clipping) for avoiding overfitting.

For more complex networks, we propose employing ranking-based measures that have the following two properties. (I) The ranking-based measures allow us to directly and accurately improve classification [38], [39] rather than standard classification losses such as soft-max cross-entropy. (II) Ranking-based measures can be employed to improve PU learning [11].

Among several ranking-based measures, our method employs the approximated AUC of a Precision-Recall (AUCPR) curve [39]. This is because AUCPR is robust to an imbalance between positive and negative samples [43]. This property is beneficial for artificial fertilization because the number of negative samples (i.e., non-live birth) is much more than the number of positive samples (i.e., live birth). In addition, this approximated AUCPR [39] has advantages in efficiency for learning and in applicability to any learning architectures.

Given $N(x)$, a PR curve is drawn by varying a threshold of the score for binary classification. In a discrete manner, (I) given a threshold for precision $= \alpha$, its recall is computed, and (II) the sum of the recalls of varying α is regarded as AUCPR. For simplicity, α is varied at k regular intervals between π_+ and 1 so that $\alpha_t = \pi_+ + \frac{(1-\pi_+)t}{k}$ where $t = \{1,\ldots,k\}$. The following saddle-point problem is resolved for optimizing AUCPR:

$$\min_N \max_{\lambda_1,\ldots,\lambda_k} L(N,\lambda), \quad (2)$$
$$L(N,\lambda) = \sum_{t=1}^{k}\Delta_t\Big((1+\lambda_t)\mathscr{L}_+^+(N(\boldsymbol{x}),b_t) + \lambda_t\frac{\alpha_t}{1-\alpha_t}\mathscr{L}_-^-(N(\boldsymbol{x}),b_t) - \lambda_t|S^+|\Big), \quad (3)$$



(a) Input image (b) Framewise attention (c) Temporal-smooth attention

Fig. 2. Examples of visual attentions.

where λ_t and b_t denote a Lagrange multiplier and a score threshold for "precision $\geq \alpha$" at t-th interval, respectively. $\mathscr{L}_+^+(N,b_t)$ and $\mathscr{L}_-^-(N,b_t)$ denote the sum of errors on positive and negative samples, respectively, as follows:

$$\mathscr{L}_+^+(N(\boldsymbol{x}),b_t) = \sum_{\boldsymbol{x}\in S^+}\ell_h(N(\boldsymbol{x}),b_t,+1),$$
$$\mathscr{L}_-^-(N(\boldsymbol{x}),b_t) = \sum_{\boldsymbol{x}\in S^-}\ell_h(N(\boldsymbol{x}),b_t,-1),$$
$$\ell_h(N(\boldsymbol{x}),b_t,y) = \max\left(0,1-y\cdot(N(\boldsymbol{x})-b_t)\right) \quad (4)$$

where ℓ_h is the hinge loss of $N(\boldsymbol{x}) - b_t$ with its label $y \in \{-1,1\}$. We can resolve the above saddle-point problem (3) by the following iterative stochastic gradient descent updates using a subgradient method for approximate saddle-points [44], if $\mathscr{L}_+^+$ and $\mathscr{L}_-^-$ are convex, as proven in [39], [45]:

$$N^{(i+1)} = N^{(i)} - \gamma\nabla L(N^{(t)},\lambda^{(i)}),$$
$$\lambda^{(i+1)} = \lambda^{(i)} - \gamma\nabla L(N^{(t+1)},\lambda^{(i)}),$$

where i and γ denote the number of iterations and a constant stepsize, respectively.

C. PU Learning Improved by Extended Rank Learning

While Eq. (4) is satisfied in PN learning, S^- denotes a set of negative samples but the negative samples are unavailable in PU learning. We need to modify Eq. (4) for unlabeled samples. To this end, $\mathscr{L}_-^-(N(\boldsymbol{x}),b_t)$ is replaced in accordance with [13]:

$$\mathscr{L}_-^-(N(\boldsymbol{x}),b_t) = -\mathscr{L}_+^-(N(\boldsymbol{x}),b_t) + \mathscr{L}_u^-(N(\boldsymbol{x}),b_t),$$
$$\mathscr{L}_u^-(N(\boldsymbol{x}),b_t) = \sum_{\boldsymbol{x}\in S^u}\ell_h(N(\boldsymbol{x}),-1) \quad (5)$$

By substituting Eq. (5) to Eq. (3), the following surrogate function for optimizing AUCPR in PU learning is obtained:

$$\max_{\lambda_1,\ldots,\lambda_k}\sum_{t=1}^{k}\Delta_t\Big((1+\lambda_t)\mathscr{L}_+^+(N(\boldsymbol{x}),b_t) + \lambda_t\frac{\alpha_t}{1-\alpha_t}\left(\mathscr{L}_u^-(N(\boldsymbol{x}),b_t) - \mathscr{L}_+^-(N(\boldsymbol{x}),b_t)\right) - \lambda_t|S^+|\Big), \quad (6)$$

However, the second term of the above AUCPR-PU loss, Eq. (6), may tend to overfit so that most unlabeled samples are classified as negative [13]. This overfitting is caused in a way that $\left(\mathscr{L}_u^-(N(\boldsymbol{x}),b_t) - \mathscr{L}_+^-(N(\boldsymbol{x}),b_t)\right)$ in Eq. (6) keeps

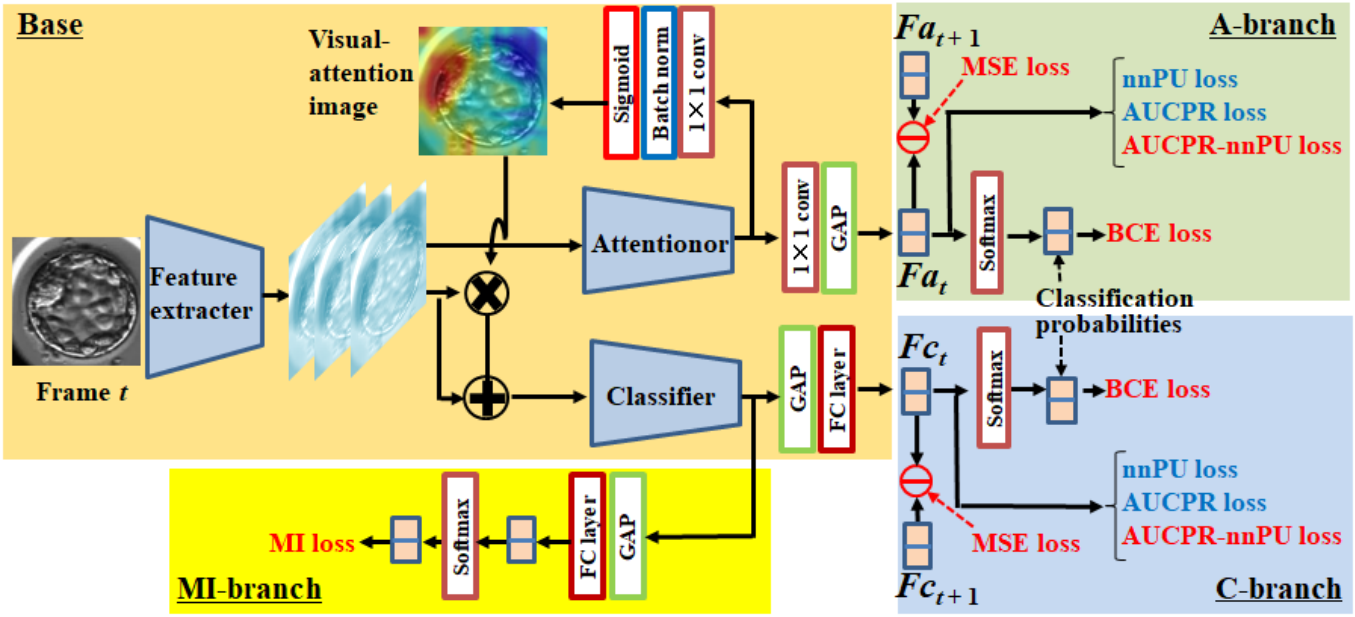


Fig. 3. Network for framewise processing in our method. Parts with a background colored with light orange come from the base network [21].

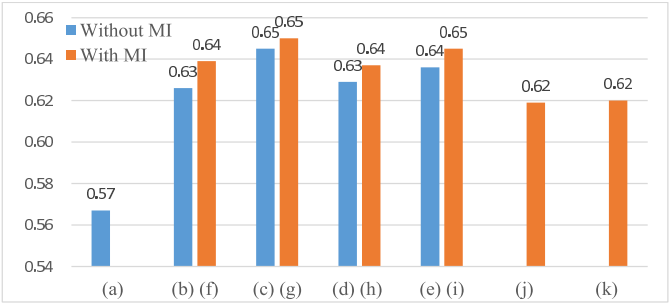


Fig. 4. Performance evaluation with AUCROC.

decreasing. This overfitting can be avoided by clipping this term:

$$\max_{\lambda_1,\ldots,\lambda_k} \sum_{t=1}^{k} \Delta_t \Big((1+\lambda_t)\mathscr{L}_+^+(N(\boldsymbol{x}), b_t) + \max\left\{ \beta, \lambda_t \frac{\alpha_t}{1-\alpha_t} \left(\mathscr{L}_u^-(N(\boldsymbol{x}), b_t) - \mathscr{L}_+^-(N(\boldsymbol{x}), b_t) \right) \right\} - \lambda_t |S^+| \Big), \tag{7}$$

where β is a hyperparameter for clipping. $\beta = 4$ in our experiments. We call the loss expressed by Eq. (7) the AUCPR-nnPU loss.

D. Regularization With Mutual Information Maximization

In our proposed method, PU learning is further improved by metric learning. Metric learning [42], [46] allows us to optimize the feature space so that samples in the same class get close together and those in different classes depart from each other. This property is also effective for our PU problem. However, most metric learning methods are designed for supervised learning so that all training data is labeled. This assumption is not satisfied in our PU problem.

For metric learning in our PU problem, unsupervised metric learning [19] can be employed. This metric learning [19] makes the clusters of samples so that the distributions of intra-cluster and inter-cluster distances are small and large, respectively.

For this clustering, a sample z is slightly changed to a new sample z'. For example, if z is an image, z is rotated, noised, and/or blurred in order to generate z', as with image deformations for the general data augmentation. Since z and z' should be close to each other in the feature space, the following mutual information is maximized for the clustering:

$$I(z, z') = H(z) - H(z|z'), \tag{8}$$

where $H(z)$ and $H(z|z')$ denote the entropy function and the conditional entropy, respectively. While there should be a trade-off between $H(z)$ and $H(z|z')$, the feature space is trained so that similar samples make the cluster and other samples separate from each other. The mutual information of two discrete random variables, X and Y, is expressed as follows:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \cdot \ln \frac{p(x, y)}{p(x)p(y)}, \tag{9}$$

where $p(x, y)$ denotes the joint distribution function of X and Y. $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y, respectively.

Assume that C denotes the number of clusters, and $\Phi(z)$ is a set of C probability scores of z. The i-th score is a probability that z belongs to i-th cluster. Φ is a DNN in our method. With Eq. (9), the loss function for maximizing the mutual information can be expressed as follows:

$$L_{MI} = -I(z, z') = -\sum_{c=1}^{N_C} \sum_{c'=1}^{N_C} \boldsymbol{P}_{c,c'} \cdot \ln \frac{\boldsymbol{P}_{c,c'}}{\boldsymbol{P}_c \cdot \boldsymbol{P}_{c'}}, \tag{10}$$

$$\boldsymbol{P} = \frac{1}{N_B} \sum_{i=1}^{N_B} \Phi(x_i) \cdot \Phi(x_i')^{\mathrm{T}}, \tag{11}$$

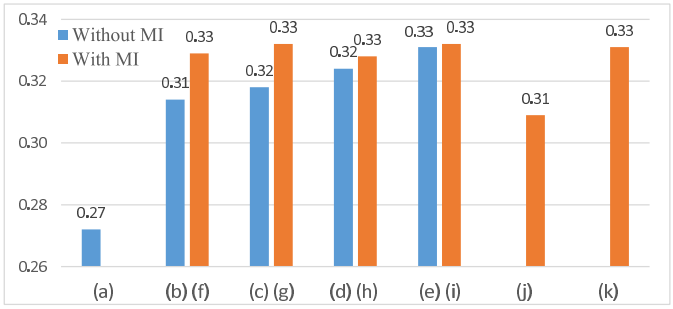


Fig. 5. Performance evaluation with AUCPR.

where $\boldsymbol{P}$, N_C, and N_B denote a $N_C \times N_C$ matrix, the number of the clusters, and the number of samples in a mini batch, respectively. $\boldsymbol{P}_c$ and $\boldsymbol{P}_{c'}$ denote a sum of values in the c-th row and the c'-th column of $\boldsymbol{P}$, respectively.

While the original L_{MI} is proposed to use any kind of image deformations for making z' from z, some of them are inappropriate for embryo classification. If z' is away from the distribution of possible embryo images, metric learning using the loss (10) is corrupted. Indeed, the importance of appropriate image deformations for metric learning is validated in the literature [47]. We empirically validated the appropriate image deformations for embryo image classification; see Table III.

E. Temporal Visual Attentions

While our embryo grading is done in time-lapse images rather than in each frame, an embryo is not changed significantly in most sequential frames (see Figure 2 (a)), because an embryo cleaves at around 12-24 hours intervals before implantation. For this temporally-sparse embryo cleavage, dense temporal attention is not beneficial. Instead of such temporal attentions, our method focuses more on spatial attention.

In time-lapse images, spatial attentions on sequential frames should be almost equal because these images are almost equal visually. However, a framewise attention mechanism may produce inconsistent results between the sequential frames, as shown in Figure 2 (b). While these unreliable results can be possibly resolved by more training samples, the availability of medical data is often limited.

Our method achieves temporally-consistent spatial attentions by smoothness constraint. This temporal smoothness loss is given by the mean squared error between features used for generating attention maps at t-th and $(t+1)$-th frames:

$$L_{MSE} = \frac{1}{N_B} \sum_{i=1}^{N_B} \left(\Phi(x_i^{t+1}) - \Phi(x_i^t) \right)^2, \qquad (12)$$

where x_i^t denotes i-th image captured at t-th frame. The results of this loss are shown in Figure 2 (c).

F. Implementation Details

Figure 3 shows the architectures of our network. It consists of the base network (including feature extractor, classifier, and attentionor) and three branches for loss computation. An input frame at t is fed into the feature extractor.

Fig. 6. Classification performance evaluation (threshold = 0.5).

1) Base: The feature extractor, classifier, and (framewise) attentionor are constructed as with [21] as described below.

ResNet56 [48] pretrained on ImageNet [49] is divided in a 46-th layer into two parts. The front and back parts are used as the feature extractor and classifier, respectively. While ResNet with any layers is applicable in our proposed method, a relatively-deep ResNet56 (e.g., deeper than ResNet34 and ResNet50) is employed for improving the representation ability. Finally, the output of the classifier, which is a 2048-dimensional feature, is fed into the global average pooling (GAP) layer and the fully-connected (FC) layer in order to get a 2D feature, $\boldsymbol{Fc}_t$ in Figure 3, from a t-th frame (i.e., "Frame t" in the figure). The attentionor consists of 10 convolutional layers. The output of the attentionor, which is also a 2048-dimensional feature, is fed into a 1×1 conv layer and the GAP layer in order to get a 2D feature, $\boldsymbol{Fa}_t$ in Figure 3. In addition to $\boldsymbol{Fa}_t$, a visual-attention image is produced from the output of the attentionor through a 1×1 conv layer, the batch normalization, and the sigmoid activation.

2) Ours: $\boldsymbol{Fc}_t$ and $\boldsymbol{Fa}_t$ are used for computing Binary Cross-Entropy (BCE), nnPU, AUCPR, and AUCPR-nnPU losses in the C-branch (colored with the blue background in Figure 3) and the A-branch (colored with the green background in Figure 3), respectively. In all of Eq. (1) of nnPU loss, Eq. (3) of AUCPR loss, and Eq. (7) of AUCPR-nnPU loss, these features are represented as $N(\boldsymbol{x})$. While PN learning uses only BCE, the risk estimator R in Eq. (1), AUCPR loss in Eq. (3), and AUCPR-nnPU loss in Eq. (7) are used for nnPU learning, AUCPR learning, and our proposed AUCPR-nnPU learning, respectively. The classifier and attentionor are augmented by the temporal-smoothness loss. This loss is computed by MSEs (indicated by "MSE loss" in the figure) between $\boldsymbol{Fc}_t$ and $\boldsymbol{Fc}_{t+1}$ and between $\boldsymbol{Fa}_t$ and $\boldsymbol{Fa}_{t+1}$, as expressed by Eq. (12). When updating the network weights with frame t, frame at $t+1$ is feed-forwarded to the network to get $\boldsymbol{Fc}_{t+1}$ and $\boldsymbol{Fa}_{t+1}$, while the frame $t+1$ itself is not used for the update. For MI loss, the output of the classifier is also fed into another branch consisting of the GAP layer and the FC layer. While the structure of this MI-branch (colored with the yellow background in Figure 3) is equal to that of the C-branch, these two branches do not share their weights for optimizing each branch for each loss.

Our code is available at https://github.com/masashi-nagaya/embryo-analysys/tree/master.

IV. Experiments

A. Dataset

All experiments were done with gray-scale time-lapse image sequences. 9%, 69%, and 22% of patients are in their twenties, thirties, and forties, respectively. All sequences were collected in OB-GYNs under either of the two conditions below:

(A) A frame-capture interval is 10 mins. A frame size is 250×250 pixels.
(B) A frame-capture interval is 15 mins. A frame size is 500×600 pixels.

101 and 542 sequences were collected under (A) and (B), respectively. While these different conditions are useful for validating the robustness of each method against the domain gap, it is better to guarantee spatial and temporal alignment between different sequences. In our experiments, spatial alignment is done, so that (I) a rectangle enclosing an embryo is annotated at each frame and (II) this rectangle is cropped out and rescaled to 224×224 pixels, which is the input size of our network shown in Figure 3. By extracting frames every 60 mins (i.e., every six and four frames from sequences in (A) and (B), respectively), six and four sub-sequences are produced from original sequences in (A) and (B), respectively. Since these sub-sequences have the same frame-capture interval (60 mins), they are roughly aligned temporally. In total, $(6 \times 101 = 606)$ and $(4 \times 542 = 2,168)$ sub-sequences were produced from the original sequences in (A) and (B), respectively. In each original sequence, the first sub-sequence begins around 25 hours after fertilization. The number of frames in each original sequence varies between around 100 and 700. While the visually-similar temporal frames of each embryo are useless for representing a temporal history, even subtle variations in these frames are useful for data augmentation. This idea is supported by Figure 2 (b) where almost equal frames produce different attentions probably because a DNN finds subtle differences between these frames (i.e., t and $t+1$) due to overfitting. For suppressing this overfitting, all sub-sequences were used as independent training sequences in our experiments. These sub-sequences are flipped and rotated for further data augmentation. In what follows, these sub-sequences are called *sequences*.

TABLE II
The Number of Image Sequences in Each Group for the Cross Validation

	group-1	group-2	group-3	group-4	group-5	total
Live-birth	28	28	28	28	28	140
Non live-birth	100	100	101	101	101	503

The sequences are randomly divided into five groups for the cross-validation, as shown in Table II. A fifth group was always used as validation data. The validation data was used for optimizing parameters (i.e., learning rate, epochs, the dimension of a feature vector $=$ 2048, optimizer $=$ Adam) before evaluation on test data. The learning rate and the number of epochs differ among the variants of our method (i.e., (a) – (k) in Section IV-B). Each of the remaining four groups was used as test data, while the other three groups were used as training data. All embryo images are automatically annotated based on their birth results. While images labeled as non-live birth in the training data are used as unlabeled samples in our PU learning, those in the validation and test data are used as negative samples for evaluation. The mean of these four trials is shown in Section IV-B.

B. Classification Performance Evaluation

We conducted evaluation experiments with the following 10 variants of our method, (b) – (k). All of these 10 variants are implemented on our proposed network shown in Figure 3. For comparison, a previous PN classifier (a) is also evaluated.

(a) **F-PN-CNN [5]:** A CNN-based PN classifier [5].
(b) **F-PN:** Framewise PN learning trained by BCE.

(c) **F-nnPU:** Framewise nnPU learning trained by the nnPU loss expressed by Eq. (1).
(d) **F-AUCPR:** Framewise PU learning trained by the AUCPR loss expressed by Eq. (3).
(e) **F-AUCPR-nnPU:** Framewise PU learning trained by Eq. (7).
(f) **F-PN-MI:** (b) + the MI loss expressed by Eq. (10).
(g) **F-nnPU-MI:** (c) + the MI loss expressed by Eq. (10).
(h) **F-AUCPR-MI:** (d) + the MI loss expressed by Eq. (10).
(i) **F-AUCPR-nnPU-MI:** (e) + the MI loss expressed by Eq. (10).
(j) **S-Mean:** The classification probabilities of all time-lapse features in each sequence are averaged with weights for classifying this sequence. The weight of each frame is equal to its frame number (i.e., the weight of f-th frame is f). The feature vector (i.e., $\boldsymbol{Fc}_t$) is trained by PU learning in (i).
(k) **S-GRU:** The feature vector (i.e., $\boldsymbol{Fc}_t$) trained by our PU learning (i) is extracted from each frame in the sequence. These feature vectors are sequentially fed into a Gated Recurrent Unit (GRU) [50]. The output of the GRU at the last frame is fed into a fully-connected layer followed by the softmax layer to get the classification probabilities.

Except for (a) and (b), all of these variants used PU learning. While binary classification is done framewise (i.e., in all frames independently) in (a) – (i), all frames in each sequence are used for its classification in (j) and (k).

Figures 4 and 5 show classification results evaluated by AUCPR and AUC of a Receiver Operating Characteristic (AUCROC), which is independent of a binarization threshold for the classification probability, respectively. In these results, we can see the following observations:

- In both metrics, all PU learning methods (c) – (e) and (g) – (i) work better than PN learning (a) and (b).
- In both metrics, each method with MI is superior to the one without MI.
- In general, a sequence of time-lapse frames has rich information for a recognition task. However, in our experiments, classification methods using time-lapse image sequences (i.e., (j) and (k)) are not superior to those with frames (i.e., (b) – (i)). This might be because the variation of sequences is larger compared to the amount of training data.
- In both metrics, (f), (g), (h), and (i) are almost comparable.

While AUCPR and AUCROC are appropriate for evaluating the performance of each method independently of a threshold, medical doctors may require binarized classification results (i.e., live birth or non-live birth) in clinical practice. Figure 6 shows the precision, recall, F-measure, false-positive rate, and negative-predictive rate provided with a binarization threshold = 0.5 (i.e., the mid value of 0 and 1).

In the F-measure (i.e., the harmonic mean of the precision and recall), our proposed methods (e) and (i) get the best performance in scores without and with MI, respectively. It can also be seen that (i) with MI is superior to (e) without MI; 0.423 vs. 0.405. However, we pay more attention to the recall. This is because it is more important to find positive-class

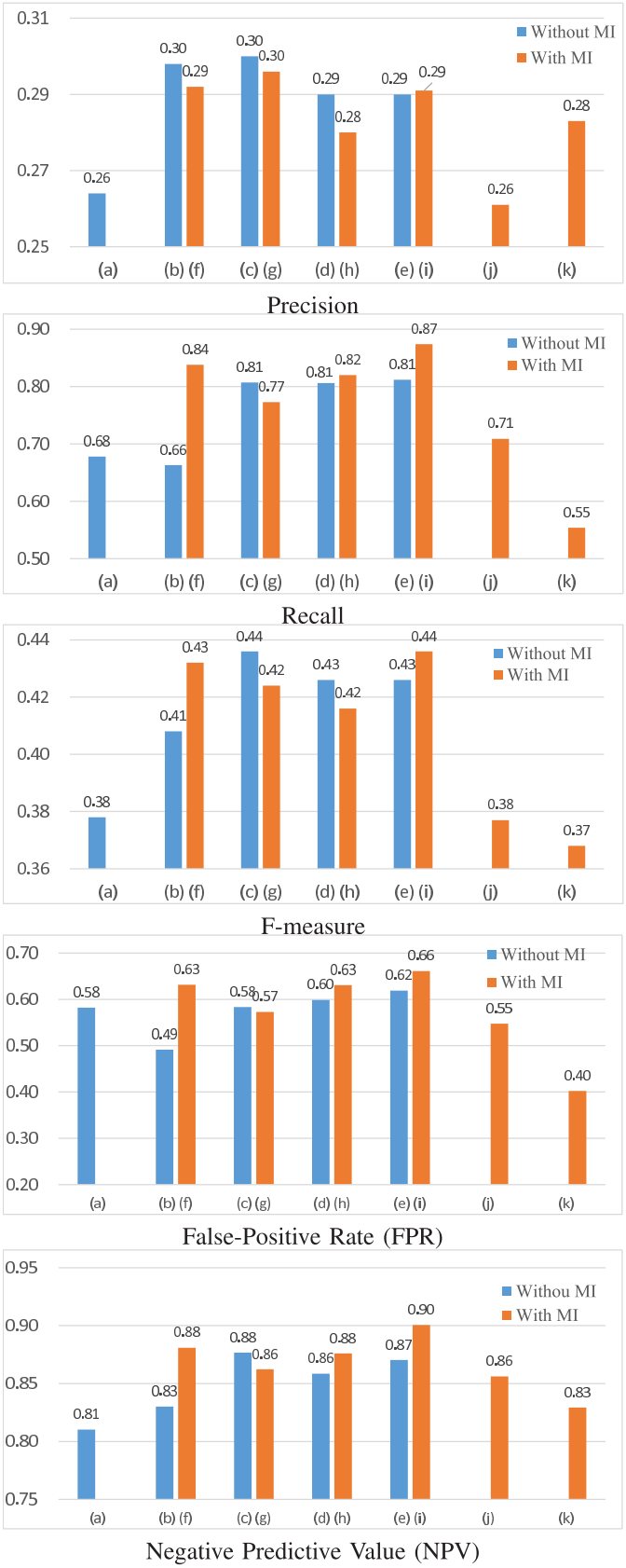
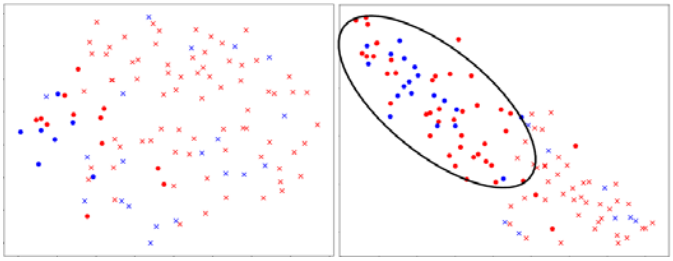


Fig. 7. Classification performance evaluation. The best threshold was determined based on the validation data.



(b) PN learning (i) Our proposed method

Fig. 8. Feature distributions. The features are visualized with their 2D representations obtained by t-SNE. The left and right figures show the results of PN learning (b) and our proposed method (i), respectively. Blue and red marks indicate ground-truth live birth and non-live birth embryos, respectively. Predicted live birth and non-live birth embryos are indicated by $\cdot$ and $\times$, respectively.

embryos, which have a high probability of being live birth, in order not to miss the pregnancy opportunity, as described in Section III-A. In the recall, our proposed method (i) can outperform others with large margins.

While the aforementioned results are obtained with a manually-given threshold ($= 0.5$), we can select a threshold based on the validation data so that the F-measure is the highest with the selected threshold. The best thresholds were selected in each method independently. These thresholds were used for evaluating the classification results, shown in Figure 7. While margins between our proposed method and others is not large in this case, our proposed method (i) with MI is still the best among all variants in terms of the recall and the F-measure.

Since our focus is to detect embryos each of which has a high probability of live birth, the false-positive rate is inevitably increased. However, the increase can be suppressed with a threshold given by the validation data; "0.04 in (c) vs. 0.41 in (i) with a threshold $= 0.5$" and "0.58 in (c) vs. 0.66 in (i) with a threshold given by the validation data." On the other hand, the negative-predictive rate is successfully gained by our method (i) compared with other framewise methods (i.e., (a) – (h)) in both of Figures. 6 and 7.

1) Run-Time Analysis: The run time for inference with the proposed network architecture shown in Figure 3 (i.e., methods (b) – (i)) is 0.0045 seconds per frame, which is sufficiently fast for analyzing many sequential frames.

C. Feature Distribution Analysis

In order to visually see the effect of PU learning compared to PN learning, the distributions of positive and negative test samples are visualized by t-distributed Stochastic Neighbor Embedding (t-SNE) [51]. t-SNE obtained the 2-D feature vectors from the 2048-D feature vectors used in our network. Figure 8 shows the obtained distributions of the features in PN learning (b) and our proposed PU learning (i), which are shown in the left and right figures, respectively. These results are obtained with the binarization threshold determined based on the validation data.

In PN learning, positive and negative test samples (i.e., blue and red marks) are almost uniformly distributed. As a result, only a few test samples are classified to be positive (i.e., indicated by dots in Figure 8). In our proposed method, on the other hand, many positive samples are located where test samples are classified to be positive (i.e., indicated by dots in Figure 8). In our proposed method, while many false-positives (indicated by red dots) are detected, many true-positives (indicated by blue dots) are also detected, as enclosed by a solid black ellipse in the figure. This is a big difference between the results of PN learning and our proposed method. This observation is identical to the results shown in Figure 7. That is, the recall of our proposed method (i) is much better than that of PN learning (b).

D. Effect of Image Deformations on the MI Loss

As described in Sec. III-D, the types of image deformations used in the MI loss, expressed by Eq. (10), affect the performance of the MI loss. Table III shows the performance change depending on the image deformations. We verified two types of the image deformations:

- V1: Only random image rotation and random image flipping were given to training images.
- V2: In addition to V1, random color jittering was also provided. We used the Gaussian blur ($\sigma \in [0.1, 2.0]$) and changed brightness ($\times[0.2, 1.8]$), contrast ($\times[0.2, 1.8]$), saturation ($\times[0.2, 1.8]$), and hue ($\times[-0.2, 0.2]$).

Our motivation in this experiment was to verify whether or not the color jittering schemes in V2 change positive/negative images so that they deviate from the real distribution of the positive/negative images.

In Table III, we can see the following observations:

- In our proposed method (i), the threshold-independent measures (i.e., AUCPR and AUCROC) and the threshold-dependent measures (i.e., recall and F-measure) are better in V2 and V1, respectively. This suggests that image jittering should not be used in clinical practice where medical doctors require the binary classification results obtained by a threshold.
- Compared with the threshold-dependent measures where binarization is done only with better thresholds chosen, difficult binarization where positive and negative samples are mixed is required in other thresholds in the threshold-independent measures. Therefore, our interpretation about the superiority of V2 in the threshold-independent measures is that, for such mixed positive and negative samples, color jittering schemes in V2 might produce embryo images that are useful for correct binarization.
- While the threshold-independent measures are superior in V2 than V1 in our proposed method (i), it is possible that this effect is obtained not by image jittering for the MI loss but by general data augmentation. In order to identify the cause of the performance gain in V2, the performance measures in (e) are also shown in Table III. Since (i) with V2 is better than (e) with V2 in all cases except precision (i.e., 0.311 in (i) vs. 0.315 in (e)), we conclude that the performance gain with V2 is mainly obtained by the MI loss.

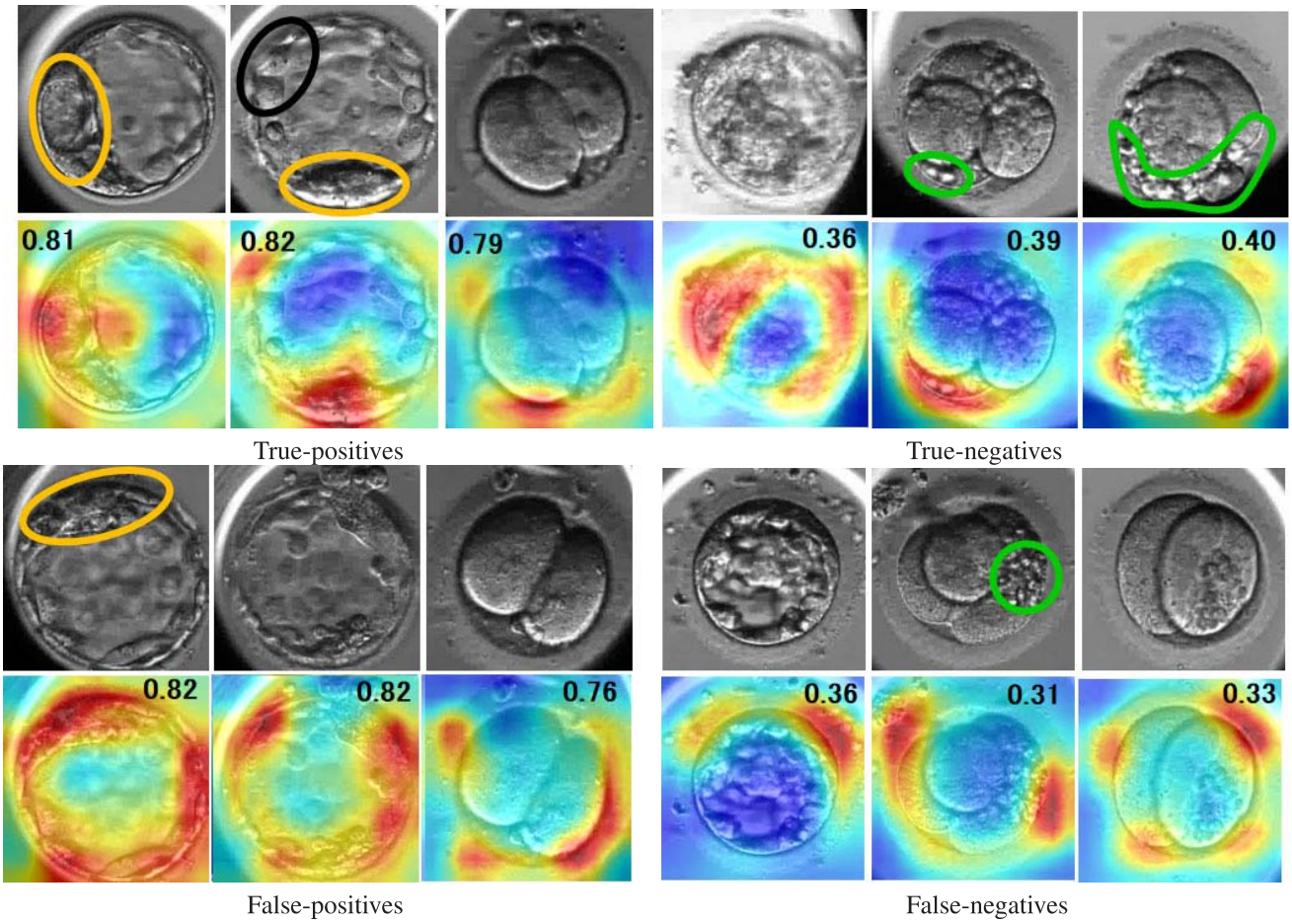


Fig. 9. Visual attentions of true-positives, true-negatives, false-positives, and false-negatives obtained by our proposed method (i). Upper and lower rows in each case show observed images and their corresponding visual attentions, respectively. Regions enclosed by orange, black, and green lines indicate dense inner cell mass, nutrient ectoderm, and fragments, respectively. A value overlaid in each image is a probability score of being positive.

TABLE III
PERFORMANCE CHANGE DEPENDING ON THE IMAGE DEFORMATIONS USED IN THE MI LOSS. THE HIGHEST SCORE IN EACH COLUMN IS COLORED WITH RED

Methods	AUCPR	AUCROC	Precision	Recall	F-measure
(i) AUCPR-nnPU+MI (V1)	0.332	0.645	0.311	0.689	0.423
(i) AUCPR-nnPU+MI (V2)	0.361	0.654	0.311	0.639	0.415
(e) AUCPR-nnPU (V1)	0.328	0.637	0.330	0.398	0.355
(e) AUCPR-nnPU (V2)	0.356	0.648	0.315	0.615	0.414

E. Visualization Results

As described in Sec. III-F, a visual-attention image is produced in the base network [21]. In order to validate the reasonability of the visual attentions, those of true-positives, true-negatives, false-positives, and false-negatives are shown in Figure 9.

1) True-Positives: While the images are significantly different depending on the cleavage stage, dense inner cell mass and nutrient ectoderm are highly activated in the visual attentions of the left two examples. The rightmost example might be classified to be positive because the blast is uniform and no fragments are observed. In that sense, we can interpret this result so that no fragments are observed in highly-activated regions where fragments are likely to appear.

2) True-Negatives: In the right two examples, fragments are highly activated as the evidence of being negative in early cleavage stages. In the leftmost example, since the visual attentions are activated around the boundary of the embryo, the density of the inner cell mass might be considered to be not dense enough.

3) False-Positives: In the left two examples, inner cell mass might be considered to be dense by mistake. Since these misrecognitions might be avoided by medical doctors, more training images allow our proposed method to correctly visualize the evidences. The rightmost example is a negative test sample, while its appearance seems to be positive. This negative test sample might have a chromosome abnormality, or it is too difficult to classify this test sample as negative because it is still in the early cleavage stage.

4) False-Negatives: While all of these test samples are positive, there are fragment-like regions in these images. These regions might confuse the classifier.

V. Concluding Remarks

A. Summary and Discussions

This paper proposed a method for embryo classification by PU learning improved by learning-to-rank. Since our proposed method employs real live birth and non-live birth results as labels given to training data, the training data can be efficiently collected with no manual annotations. While such birth results are unreliable/noisy for classification using visual features, PU learning allows us to alleviate an adverse effect of unreliable/noisy labels in our method unlike a general usage of PU learning. Experimental results demonstrate the positive impacts of our proposed loss function for PU learning (i.e., loss expressed by Eq. (7)), unsupervised metric learning for PU learning (i.e., achieved by loss expressed by Eq. (10)), and appropriate data augmentation for this metric learning.

In binary classification (i.e., live birth or non-live birth) for supporting medical doctors in clinical practice, the recall value is important in order to not miss good embryos. Our proposed method gets a higher recall value than PN classification (0.87 vs. 0.68), while the precision value is also improved a little, as shown in Figure 7.

Visual attentions can be also improved even in an insufficient number of training images by our temporal smoothness loss (Eq. (12)), as shown in Figure 2. Its effectiveness is validated in comparison with well-known morphological assessments [25], [26], as shown in Figure 9. This consistency between the decisions of medical doctors and our proposed method allows the doctors to confirm the reliability of each classification result.

B. Limitations and Future Work

Although our proposed method outperforms previous ones, its performance is expected to be further improved. While our proposed method employs only positive and unlabeled samples, the effectiveness of a combination of positive, *negative*, and unlabeled samples are demonstrated recently [52]. Such negative samples are available in the artificial fertilization process also so that discarded embryos, which are selected as apparently low-quality ones by medical doctors, are regarded as negative samples.

Embryo grading with time-lapse images is done in clinical practice [53]. The time-lapse images are employed for automatic embryo grading using graphical models [54] and DNNs [55], [56], while the effectiveness of the time-lapse image evaluation is not clear yet [3], [4]. For the time-lapse analysis, while a recurrent network using features extracted by our proposed method (i.e., (k) S-GRU in Section IV-B) is also evaluated in our experiments, it is inferior to framewise classification. Our future work includes more exploration of effective time-lapse analyses.

In experiments shown in our paper, every embryo region was manually extracted as a rectangle. This manual annotation should be avoided in clinical practice. In addition, the rectangle contains background pixels as well as embryo pixels. Such background pixels are irrelevant to the embryo classification task. These problems can be avoided by employing pixelwise region segmentation such as [57], [58] instead of the manual annotation.

An essential limitation of our proposed method is that it cannot discriminate live birth embryos from visually-positive non-live birth embryos including chromosome abnormalities. In order to resolve this problem, another approach is required. For example, a larger number of embryo images might reveal subtle visual differences between embryos with and without the chromosome abnormality. For finding such subtle visual differences, genetic analysis such as Preimplantation Genetic Diagnosis probably provides us with reliable supervised data about the chromosome abnormality.

Acknowledgment

The authors appreciate Dr. Yuki Sawada, Dr. Takeshi Sato, and Dr. Mayumi Sugiura at Nagoya City University and Sawada Women's Clinic for providing them precious time-lapse images of embryos, their annotation data, and informative knowledge of implantation. The authors also thank Dr. Makoto Miwa for discussions about PU learning.

References

[1] A. Capalbo *et al.*, "Correlation between standard blastocyst morphology, euploidy and implantation: An observational study in two centers involving 956 screened blastocysts," *Hum. Reproduction*, vol. 29, no. 6, pp. 1173–1181, Jun. 2014.

[2] M. G. Minasi *et al.*, "Correlation between aneuploidy, standard morphology evaluation and morphokinetic development in 1730 biopsied blastocysts: A consecutive case series study," *Hum. Reproduction*, vol. 31, no. 10, pp. 2245–2254, Oct. 2016.

[3] J. Zhang *et al.*, "Morphokinetic parameters from a time-lapse monitoring system cannot accurately predict the ploidy of embryos," *J. Assist. Reproduction Genet.*, vol. 34, no. 9, pp. 1173–1178, Sep. 2017.

[4] A. Reignier, J. Lammers, P. Barriere, and T. Freour, "Can time-lapse parameters predict embryo ploidy? A systematic review," *Reproductive Biomed. Online*, vol. 36, no. 4, pp. 380–387, Apr. 2018.

[5] Y. Miyagi, T. Habara, R. Hirata, and N. Hayashi, "Feasibility of artificial intelligence for predicting live birth without aneuploidy from a blastocyst image," *Reproductive Med. Biol.*, vol. 18, no. 2, pp. 204–211, Apr. 2019.

[6] P. Khosravi *et al.*, "Deep learning enables robust assessment and selection of human blastocysts after *in vitro* fertilization," *NPJ Digit. Med.*, vol. 2, no. 1, Dec. 2019, Art. no. 21.

[7] C. L. Bormann *et al.*, "Consistency and objectivity of automated embryo assessments using deep neural networks," *Fertility Sterility*, vol. 113, no. 4, pp. 781–788, 2020.

[8] M. Sugiura-Ogasawara, Y. Ozaki, K. Katano, N. Suzumori, T. Kitaori, and E. Mizutani, "Abnormal embryonic karyotype is the most frequent cause of recurrent miscarriage," *Hum. Reproduction*, vol. 27, no. 8, pp. 2297–2303, Aug. 2012.

[9] K. Kirkegaard, A. Ahlström, H. J. Ingerslev, and T. Hardarson, "Choosing the best embryo by time lapse versus standard morphology," *Fertility Sterility*, vol. 103, no. 2, pp. 323–332, Feb. 2015.

[10] J. L. Eaton, M. R. Hacker, D. Harris, K. L. Thornton, and A. S. Penzias, "Assessment of day-3 morphology and euploidy for individual chromosomes in embryos that develop to the blastocyst stage," *Fertility Sterility*, vol. 91, no. 6, pp. 2432–2436, Jun. 2009.

[11] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2008, pp. 213–220.

[12] G. Niu, M. C. Du Plessis, T. Sakai, Y. Ma, and M. Sugiyama, "Theoretical comparisons of positive-unlabeled learning against positive-negative learning," in *Proc. NIPS*, 2016, pp. 1199–1207.

[13] R. Kiryo, G. Niu, M. C. Du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," in *Proc. NIPS*, 2017, pp. 1–17.

[14] T. Bepler *et al.*, "Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs," *Nature Methods*, vol. 16, no. 11, pp. 1153–1160, Nov. 2019.

[15] E. Sansone, F. G. B. De Natale, and Z.-H. Zhou, "Efficient training for positive unlabeled learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2584–2598, Nov. 2019.

[16] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280–296, May 2019.

[17] J. Bekker and J. Davis, "Learning from positive and unlabeled data: A survey," *Mach. Learn.*, vol. 109, no. 4, pp. 719–760, Apr. 2020, doi: 10.1007/s10994-020-05877-5.

[18] D. Zhang and W. S. Lee, "Learning classifiers without negative examples: A reduction approach," in *Proc. 3rd Int. Conf. Digit. Inf. Manage.*, Nov. 2008, pp. 638–643.

[19] X. Ji, A. Vedaldi, and J. F. Henriques, "Invariant information clustering for unsupervised image classification and segmentation," in *Proc. ICCV*, 2019, pp. 9865–9874.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[21] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10705–10714.

[22] C. Yang, A. Rangarajan, and S. Ranka, "Visual explanations from deep 3D convolutional neural networks for alzheimer's disease classification," in *Proc. AMIA*, 2018, p. 1571.

[23] G. Zhao, B. Zhou, K. Wang, R. Jiang, and M. Xu, "Respond-CAM: Analyzing deep models for 3D imaging data by visualizations," in *Proc. MICCAI*, 2018, pp. 485–492.

[24] S. Shinde, T. Chougule, J. Saini, and M. Ingalhalikar, "HR-CAM: Precise localization of pathology using multi-level learning in CNNs," in *Proc. MICCAI*, 2019, pp. 298–306.

[25] M.-I. Hsu *et al.*, "Embryo implantation *in vitro* fertilization and intracytoplasmic sperm injection: Impact of cleavage status, morphology grade, and number of embryos transferred," *Fertility Sterility*, vol. 72, no. 4, pp. 679–685, Oct. 1999.

[26] D. K. Gardner and W. B. Schoolcraft, "Culture and transfer of human blastocysts," *Current Opinion Obstetrics Gynaecol.*, vol. 11, no. 3, pp. 307–311, Jun. 1999.

[27] J. C. Rocha *et al.*, "A method based on artificial intelligence to fully automatize the evaluation of bovine blastocyst images," *Sci. Rep.*, vol. 7, no. 1, Dec. 2017, Art. no. 7659.

[28] Q. Zhan, E. T. Sierra, J. Malmsten, Z. Ye, Z. Rosenwaks, and N. Zaninovic, "Blastocyst score, a blastocyst quality ranking tool, is a predictor of blastocyst ploidy and implantation potential," *F&S Rep.*, vol. 1, no. 2, pp. 133–141, Sep. 2020.

[29] Y. Miyagi, T. Habara, R. Hirata, and N. Hayashi, "Feasibility of deep learning for predicting live birth from a blastocyst image in patients classified by age," *Reproductive Med. Biol.*, vol. 18, no. 2, pp. 190–203, Apr. 2019.

[30] M. A. Zuluaga, D. Hush, E. J. F. D. Leyton, M. H. Hoyos, and M. Orkisz, "Learning from only positive and unlabeled data to detect lesions in vascular CT images," in *Proc. MICCAI*, 2011, pp. 9–16.

[31] P. Pati *et al.*, "Deep positive-unlabeled learning for region of interest localization in breast tissue images," *Proc. SPIE Med. Imag.*, vol. 10581, Mar. 2018, Art. no. 1058107.

[32] L. Lejeune, J. Grossrieder, and R. Sznitman, "Iterative multi-path tracking for video and volume segmentation with sparse point supervision," *Med. Image Anal.*, vol. 50, pp. 65–81, Dec. 2018.

[33] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *NeurIPS*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018.

[34] C. G. Northcutt, T. Wu, and I. L. Chuang, "Learning with confident examples: Rank pruning for robust classification with noisy labels," in *UAI*, G. Elidan, K. Kersting, and A. T. Ihler, Eds., 2017.

[35] B. Han *et al.*, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. NeurIPS*, 2018, pp. 1–13.

[36] M. C. Du Plessis, G. Niu, and M. Sugiyama, "Convex formulation for learning from positive and unlabeled data," in *Proc. ICML*, 2015, pp. 1386–1394.

[37] G. S. Mann and A. McCallum, "Generalized expectation criteria for semi-supervised learning with weakly labeled data," *J. Mach. Learn. Res.*, vol. 11, pp. 955–984, Dec. 2010. [Online]. Available: https://dl.acm.org/citation.cfm?id=1756038

[38] C. Cortes and M. Mohri, "AUC optimization vs. error rate minimization," in *Proc. NIPS*, 2003, pp. 313–320.

[39] E. Eban, M. Schain, A. Mackey, A. Gordon, R. Rifkin, and G. Elidan, "Scalable learning of non-decomposable objectives," in *Proc. AISTATS*, 2017, pp. 832–840.

[40] X. Chen *et al.*, "Self-PU: Self boosted and calibrated positive-unlabeled training," in *Proc. ICML*, 2020, pp. 1510–1519.

[41] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. SoC. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 539–546.

[42] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015, pp. 815–823.

[43] K. Boyd, J. Davis, D. Page, and V. S. Costa, "Unachievable region in precision-recall space and its effect on empirical evaluation," in *Proc. ICML*, 2012, p. 349.

[44] K. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Linear and Non-Linear Programming*. Palo Alto, CA, USA: Stanford Univ. Press, 1958.

[45] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *J. Optim. Theory Appl.*, vol. 142, no. 1, pp. 205–228, 2009.

[46] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. CVPR*, 2014, pp. 1386–1393.

[47] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020, pp. 1597–1607.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[49] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.

[50] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1–15.

[51] L. V. D. Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[52] Y. Hsieh, G. Niu, and M. Sugiyama, "Classification from positive, unlabeled and biased negative data," in *Proc. ICML*, 2019, pp. 2820–2829.

[53] D. J. Kaser and C. Racowsky, "Clinical outcomes following selection of human preimplantation embryos with time-lapse monitoring: A systematic review," *Human Reproduction Update*, vol. 20, no. 5, pp. 617–631, Sep. 2014.

[54] F. Moussavi, Y. Wang, P. Lorenzen, J. Oakley, D. Russakoff, and S. Gould, "A unified graphical models framework for automated mitosis detection in human embryos," *IEEE Trans. Med. Imag.*, vol. 33, no. 7, pp. 1551–1562, Jul. 2014, doi: 10.1109/TMI.2014.2317836.

[55] Y. Wang, F. Moussavi, and P. Lorenzen, "Automated embryo stage classification in time-lapse microscopy video of early human embryo development," in *Medical Image Computing and Computer-Assisted Intervention*, K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds., 2013, pp. 460–467.

[56] M. F. Kragh, J. Rimestad, J. Berntsen, and H. Karstoft, "Automatic grading of human blastocysts from time-lapse imaging," *Comput. Biol. Med.*, vol. 115, Dec. 2019, Art. no. 103494, doi: 10.1016/j.compbiomed.2019.103494.

[57] R. Zhao *et al.*, "Rethinking dice loss for medical image segmentation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, H. Wang, A. Cuzzocrea, C. Zaniolo, and X. Wu, Eds., Nov. 2020, pp. 851–860.

[58] T. Eelbode *et al.*, "Optimization for medical image segmentation: Theory and practice when evaluating with dice score or Jaccard index," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3679–3690, Nov. 2020, doi: 10.1109/TMI.2020.3002417.