Cell-Division-Aware Category-Enhanced Contrastive Learning Model for Embryonic Cleavage Stage Classification

Yukun Li[1], Yang Zhao[1]*, Xu Wang[1], Longlong Fu[2], Jiahui Wu[3], Xuan Yang[1], and Jihong Pei[4]

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. 2023150252@email.szu.edu.cn, {zhaoyang1990, wangxu, yangxuan}@szu.edu.cn

[2] NHC Key Laboratory of Frontiers and Technologies in Reproductive Health, China. fullpumc@163.com

[3] College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, China. wujiahui@szu.edu.cn

[4] College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. jhpei@szu.edu.cn

Abstract. Accurate identification of cleavage stages is crucial for evaluating embryo development quality. However, due to occlusion and morphological confusion caused by continuous changes in the embryonic core region, it is challenging to autonomously learn discriminative features for different cleavage stages directly from image sequences alone. This challenge is particularly pronounced in classifying adjacent stages. This paper proposes a Multi-Image Frame Pair-Driven Cell-Division-Aware Category-Enhanced Contrastive Learning framework for Embryonic Cleavage Stage Classification (3CL-ECSC). The cell-division-aware module (CDA) takes concatenated image frame pairs as input, and outputs a binary indication of whether cell division occurred between them. Meanwhile, a category-enhanced contrastive learning classification (2CLC) module is constructed based on supervised contrastive learning, taking image sequences as input. The cell division predictions from the CDA serve as an explicit prior to perform consistency correction and confidence recalibration on the classification predictions. This process enhances discriminability at stage transitions and improves robustness to sparse stages. During training, we designed a two-stage cell-division-aware learning mechanism. This mechanism learns global cell division cues from multi-order adjacent stage frame pairs while utilizing 1st-order adjacent stage frame pairs to extract local cell division information. This refines stage boundaries and yields more stable, consistent predictions of the stage sequence. The proposed method showed improvements in Accuracy (+3.60%) and Transition Accuracy (+2.99%) for stage classification of human embryos, and gains of +2.46% in Accuracy and +17.86% in Transition Accuracy for mouse embryos, over state-of-the-art methods. Our implementation is publicly available at https://github.com/Yukun-Lee/3CL-ECSC.

* Corresponding author.

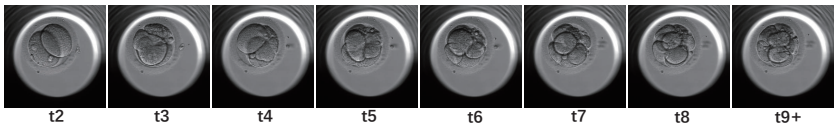

Fig. 1: Examples of typical developmental stages of cleavage-stage embryos.

Keywords: Embryonic cleavage stage classification · Cell-division-aware module · Category-enhanced contrastive learning.

1 Introduction

The rising infertility rates and shifts in reproductive age structures have driven sustained growth in clinical demand for assisted reproductive technologies (ART) and in vitro fertilization (IVF)[1]. During early development, embryonic cleavage refers to successive cell divisions of the fertilized egg, progressing from a few blastomeres to a multicellular state and exhibiting characteristic stages (Fig. 1). Morphological and kinetic characteristics during the cleavage stage, such as time of first cleavage, cleavage intervals, cleavage synchrony, and abnormal cleavage patterns, provide crucial quantitative insights for embryo quality assessment[2,3].

However, during the cleavage stage, changes in the embryo between adjacent stages are often localized. As the number of blastomeres increases, issues such as blastomere overlap and morphological disarray frequently arise. This leads to confusion in the distinctive features of neighboring stages, posing a significant challenge for accurate stage classification.

Structural methods have been explored to support cleavage-stage determination. Zhang et al. used a SAM-based two-branch pipeline for blastomere segmentation and cell counting [4], while He et al. modeled embryonic structures using 3D morphological maps [5]. Since these methods focus on segmentation, cell counting, or 3D structural analysis rather than the temporal classification of cleavage stages within embryo image sequences, we discuss them as related work rather than direct baselines.

In addition to structural analysis, existing research has also explored temporal modeling and sequence-level constraints for cleavage-stage identification. ESOD incorporates temporal learning via CNN+LSTM and improves the detection of critical cleavage stages via global monotonic post-processing[6]. However, it lacks targeted mechanisms to address sparse stages and extreme category imbalance. CNNs-CRF decouples single-frame stage recognition from adjacent-frame cleavage cues into a two-stream architecture, using a linear-chain CRF to fuse monotonicity constraints and improve global sequence-level consistency[7]. However, its cleavage cues primarily rely on locally adjacent frames, making it susceptible to occlusion noise and prone to cumulative errors. Additionally, the method's direct summation of cell division probability and category probability as model output lacks a clear practical interpretation. SCL-Embryo employs

supervised contrastive learning[8], thereby improving representational discriminability and partially mitigating class imbalance. However, it has not yet incorporated stage-monotonicity constraints and sequence-level prediction corrections via division priors.

In summary, existing approaches embed temporal information only superficially in embryo cleavage-stage identification. The constructed models cannot perceive differences between sequential image frames and lack efficient methods to leverage these differences to aid stage classification.

Inspired by the human visual system's decision-making support through perception of image changes, this paper proposes a Multi-Image Frame Pair-Driven **C**ell Division-Aware **C**ategory-Enhanced **C**ontrastive **L**earning for **E**mbryonic **C**leavage **S**tage **C**lassification (3CL-ECSC). The main contributions include:

(1) The 3CL-ECSC framework for embryo cleavage stage classification is proposed. It incorporates a cell-division-aware module CDA that detects cell division events throughout image frames, thereby assisting the classification model in identifying the developmental stage of embryos.

(2) A category consistency correction and confidence recalibration mechanism for image sequences guided by cell-division prior is designed. This mechanism integrates CDA's temporal division prediction as an explicit prior into 2CLC's output correction. Fusing multi-order adjacent-frame cell division information reduces unreasonable stage jumps, enhances sequence category consistency, and improves discrimination capability at stage transitions.

(3) A two-stage cell division perception reinforcement learning mechanism is designed. This mechanism combines learning global cell division cues from multi-frame pairs with local cell division information from single-frame pairs, thereby refining stage boundaries and achieving more stable and consistent stage sequence predictions.

2 Methodology

2.1 Overview of 3CL-ECSC

The overall framework of 3CL-ECSC is shown in Fig. 2, primarily comprising three modules: the Embryonic Core Region Extraction (CRE) module, the Cell Division Aware (CDA) module, and the Category-Enhanced Contrastive Learning Classification (2CLC) module.

Given the original embryo time-lapse image sequence $\mathbf{V}$, a variable-stride sampling strategy is employed to construct training subsequences. A smaller stride s_1 is used at stage edges, while a larger stride s_2 is applied within the same stage, yielding $\mathbf{v} \in \mathbb{R}^{T\times C\times H\times W}$. This improves coverage of splitting events and stage transition segments without significantly increasing computing overhead.

The subsequence $\mathbf{v}$ is then passed into the CRE module to obtain the core region sequence $\mathbf{v}_1$. The core region sequence $\mathbf{v}_1$ undergoes data augmentation to generate $\mathbf{v}_2$, and $\mathbf{v}_1$ and $\mathbf{v}_2$ are concatenated along the temporal dimension to form the dual-view sequence $\mathbf{v}' \in \mathbb{R}^{2T\times C\times H\times W}$.

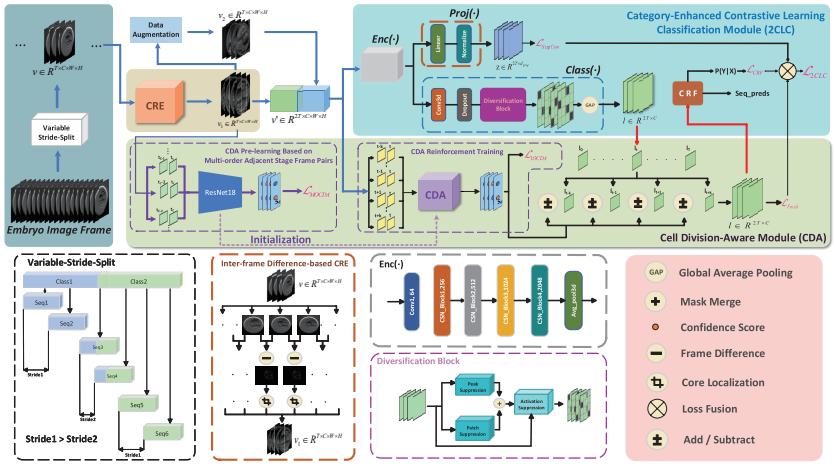

Fig. 2: Overall framework of the proposed 3CL-ECSC.

For the CDA module, training follows a two-stage reinforcement mechanism. In the first stage, multi-order adjacent-stage image frame pairs are constructed from $\mathbf{v}_1$ to learn division cues through stages and obtain initial parameters. In the second stage, 1st-order adjacent-stage image frame pairs are constructed from the dual-view sequence $\mathbf{v}'$ for reinforcement training, focusing on local division changes near stage boundaries.

For the 2CLC module, the model takes $\mathbf{v}'$ as input, learns representations of embryonic developmental stages from temporal images, and outputs frame-by-frame stage prediction logits. It improves robustness for sparsely sampled stages and easily confused adjacent stages by enhancing intra-class compactness and inter-class separability through supervised contrastive learning.

Finally, the cell-division prior from CDA is used for consistency correction and confidence recalibration of 2CLC stage predictions. The corrected logits are then fed into a linear-chain CRF for global sequence decoding with transfer relationships and monotonicity constraints.

2.2 Inter-frame Difference-based Core Region Extraction (CRE)

To reduce artifacts and noise from the culture environment, we developed an embryo core region extraction (CRE) workflow that focuses on localizing the overall embryo region in each image, as shown in Fig. 3.

The difference between consecutive frames $\mathbf{v}^{\{t\}}$ and $\mathbf{v}^{\{t-1\}}$ is computed to generate a difference image $D_t = |\mathbf{v}^{\{t\}} - \mathbf{v}^{\{t-1\}}|$. Canny edge detection is then applied to identify edge candidates E_t followed by morphological opening and closing operations to produce a coherent foreground mask M_t. The largest connected contour is extracted from the mask, and its centroid is used as the center of a circular region of interest (ROI). The ROI radius is defined as the farthest distance from the centroid to any point on the contour. Pixels outside this circle are set to

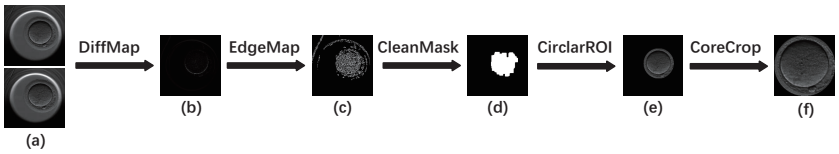

Fig. 3: Inter-frame difference-based core region extraction workflow.

zero to suppress background content. A bounding square window is determined based on the circle's center and radius, and the original image is cropped accordingly to obtain the final embryonic core region image $\mathbf{v}_{core}^{\{t\}}$. By stacking along the time dimension, a sequence $\mathbf{v}_1 = \text{Stack}(\mathbf{v}_{core}^{\{1\}}, \mathbf{v}_{core}^{\{2\}}, \ldots, \mathbf{v}_{core}^{\{T\}}) \in \mathbb{R}^{T\times C\times H\times W}$ free from background interference is obtained.

2.3 Cell-Division-Aware (CDA) Module

The learning process of this module comprises two stages: CDA pre-training on multi-order adjacent stage frame pairs (MOCDA) and CDA reinforcement training on 1st-order adjacent stage frame pairs (1OCDA). Here, order denotes the stage-wise distance between the two frames in a frame pair. The output of the CDA module is a 2-dimensional vector, where the two values represent the probability that the two frames in the concatenated input belong to the same stage and the probability that they belong to different stages.

MOCDA: This paper explicitly constructs multiple cross-stage frame pair combinations. Let $v_{1,i}^{\{p\}}$ denote core region sequence frame from stage indexed by i. The concatenation input $\mathbf{X}_{(t,j)}$ of MOCDA is defined as $\mathbf{X}_{(t,j)} = \{\Phi(\mathbf{v}_{1,j}^{\{t\}}, \mathbf{v}_{1,i}^{\{q\}}) | i = j-2, j-1, j, j+1, j+2, \mathbf{v}_{1,i}^{\{q\}} \in \mathbf{v}_1\}$, where Φ denotes the concatenation operation. The MOCDA with ResNet-18 as its backbone is employed for supervised training on the concatenation input $\mathbf{X}_{(t,j)}$. The output of MOCDA for the ith concatenation input in $\mathbf{X}_{(t,j)}$ is $\mathbf{d}_{(i,t,j)}^{m} = ResNet(\mathbf{X}_{(t,j)}; \theta_m) = [d_{(i,t,j)}^{(0)}, d_{(i,t,j)}^{(1)}]$, which indicates the probability that two frames included in the input belong to the same stage or different stages. The objective function $\mathcal{L}_{\text{MOCDA}}$ for MOCDA is defined based on category-weighted focal loss [11].

1OCDA: This stage is initialized with the parameters obtained from MOCDA. 1OCDA further enhances the model's ability to capture fine-grained variations near stage boundaries. In this stage, we sample two random augmentation operations and concatenate them along the temporal dimension to obtain a two-view sample $\mathbf{v}' \in \mathbb{R}^{2T\times C\times H\times W}$. The concatenation input $\tilde{\mathbf{X}}_{(t,j)}$ of 1OCDA is defined as $\tilde{\mathbf{X}}_{(t,j)} = \{\Phi(\mathbf{v}_{1,j}^{\prime\{t\}}, \mathbf{v}_{1,i}^{\prime\{q\}}) | i = j-1, j, j+1, \mathbf{v}_{1,i}^{\prime\{q\}} \in v'\}$. The paired frames are temporally adjacent and therefore belong either to the same stage or to neighboring stages. Similarly, we employ a category-weighted focal loss for optimization.

2.4 Category-Enhanced Contrastive Learning Classification Module

The 2CLC module takes a dual-view sequence $\mathbf{v}'$ as input to jointly improve frame-wise classification and adjacent-stage discrimination. As shown in Fig. 2, it consists of an encoder, a CDA-enhanced classification head, and a supervised contrastive projection head. The encoder $\mathrm{Enc}_\theta(\cdot)$ extracts spatio-temporal feature maps as $\mathbf{r} = \mathrm{Enc}_\theta(\mathbf{v}') \in \mathbb{R}^{B\times 2048\times 2T\times H'\times W'}$.

CDA-enhanced Classification Head: This branch performs frame-wise stage prediction by integrating the CDA division prior with CRF temporal constraints. The classification head maps the encoded features into the class-logit space through a $1\times1\times1$ 3D convolution, Dropout, Diversification Block [9], and Global Average Pooling (GAP), producing per-frame logits $\mathbf{l} \in \mathbb{R}^{B\times 2T\times K}$.

To incorporate cell-division cues into stage prediction, the CDA output from 1OCDA is used as a frame-pair prior. For frame t and its neighboring frame $t \pm i$ ($i \in [1, k]$), the CDA prediction is denoted as $\mathbf{d}^f_{(t,t\pm i)} = [d^{f,(0)}_{(t,t\pm i)}, d^{f,(1)}_{(t,t\pm i)}]$, where $d^{f,(0)}$ and $d^{f,(1)}$ indicate the probabilities of non-division and division, respectively. We define $p_{(t,t\pm i)} = 0$ if $d^{f,(0)}_{(t,t\pm i)} \geq d^{f,(1)}_{(t,t\pm i)}$, and $p_{(t,t\pm i)} = 1$ otherwise. The neighboring logits are then adaptively fused to obtain the corrected logits:

$$\mathbf{l}'_t = \mathbf{l}_t + \frac{1}{2k}\sum_{\substack{i=-k\\ i\neq 0}}^{k} (-1)^{p_{(t,t+i)}}\, d^{f,(p_{(t,t+i)})}_{(t,t+i)}\, \mathbf{l}_{t+i}. \tag{1}$$

After performing the above updates for all time steps, the corrected sequence logits are obtained, $\mathbf{l}' = \{\mathbf{l}'_t\}_{t=1}^{2T}$. **By perceiving the class relationships between frame pairs, the CDA module assists in refining the class output of the current frame.** For neighboring frames likely belonging to the same stage, their logits are positively fused with the current frame, whereas logits from frames likely crossing stage boundaries are suppressed. The correction direction and strength are adaptively determined by the CDA output.

Supervised Contrastive Projection Head: To improve the discriminability and generalization of stage representations, 2CLC introduces a supervised contrastive projection head to the shared encoder features $\mathbf{r}$. The projection head maps $\mathbf{r}$ into a 128-dimensional embedding space followed by L_2 normalization: $\mathbf{z} = \mathrm{Proj}_\phi(\mathbf{r}) \in \mathbb{R}^{B\times 2T\times 128}$, where $\mathrm{Proj}_\phi(\cdot)$ is supervised by the supervised contrastive loss [10].

Loss Function of 2CLC: The 2CLC module is optimized by combining supervised contrastive loss, frame-wise multi-class focal loss, and CRF negative log-likelihood loss. The CDA branch is independently supervised by a binary focal loss. Specifically, $\mathcal{L}_{\mathrm{Focal}}$ is used to alleviate class imbalance and emphasize hard samples [11], while the linear-chain CRF loss $\mathcal{L}_{\mathrm{crf}} = -\log P(\mathbf{y} \mid \mathbf{x})$ encourages sequence-level consistency. The overall 2CLC loss is defined as:

$$\mathcal{L}_{\mathrm{2CLC}} = w\,\mathcal{L}_{\mathrm{SupCon}} + (1-w)\,[u\,\mathcal{L}_{\mathrm{Focal}} + (1-u)\,\mathcal{L}_{\mathrm{crf}}]\,. \tag{2}$$

Table 1: Embryonic cleavage stage classification results on MED.

	Stage	ESOD [6]	CNNs-CRF [7]	R2+1D [15]	SCL-Embryo [8]	3CL-ECSC
pcF1	t2	98.68 ± 0.27	99.46 ± 0.47	<u>99.50</u> ± 0.30	99.19 ± 1.01	**99.91** ± 0.16
	t3	13.74 ± 30.73	80.60 ± 8.26	77.96 ± 5.18	<u>89.37</u> ± 5.75	**97.19** ± 1.39
	t4	92.09 ± 1.55	96.45 ± 2.81	96.59 ± 1.71	<u>97.03</u> ± 2.32	**98.66** ± 1.96
	t5	0.00 ± 0.00	50.50 ± 20.22	<u>55.19</u> ± 7.11	49.48 ± 18.77	**76.62** ± 21.82
	t6	0.00 ± 0.00	48.89 ± 5.61	43.35 ± 13.21	<u>56.31</u> ± 9.97	**77.84** ± 7.29
	t7	18.51 ± 17.03	61.93 ± 8.38	62.34 ± 11.01	<u>68.39</u> ± 12.54	**77.63** ± 2.93
	t8	70.84 ± 2.08	81.57 ± 8.43	<u>85.52</u> ± 5.37	76.29 ± 10.21	**85.82** ± 6.40
Accuracy	–	91.10 ± 0.77	94.15 ± 1.83	94.41 ± 1.66	<u>94.71</u> ± 1.37	**97.20** ± 1.58
F1	–	88.06 ± 1.56	94.31 ± 1.45	94.25 ± 1.82	<u>94.63</u> ± 1.62	**97.24** ± 1.21
TA	–	23.17 ± 3.52	<u>65.46</u> ± 3.24	57.74 ± 8.80	63.22 ± 9.04	**81.08** ± 3.05

3 Experiments

3.1 Datasets and Preprocessing

To evaluate the performance of the proposed method on embryo cleavage stage classification tasks, we conducted experiments on two time-lapse imaging datasets: the NYU mouse embryo time-lapse dataset (MED) [12] and the Human EmbryoScope time-lapse dataset (HED) [13]. The mouse dataset comprises 100 manually annotated mouse embryo sequences, providing 29134 image frames and cell trajectory annotations; single-frame cropped resolution is approximately 480×480. The human dataset consists of grayscale time-lapse microscopy images, sampled at 20-minute intervals across 7 focal planes. A single embryo is typically recorded over 3-5 days, with approximately 200-350 time points per focal plane. A total of 172551 image frames with a resolution of 500×500 are included.

We employed embryo-level partitioning, dividing the dataset into training, validation, and test sets at an 8:1:1 ratio. Each embryo sequence appeared in only one subset to prevent leakage.

3.2 Implementation Details and Evaluation Metrics

The encoder $Enc(\cdot)$ adopts CSN-50 [14] initialized with MMAction2 pretrained weights. The Diversification Block uses the default hyperparameters [9]. The input is an embryo-level batch of sequences $\mathbf{v} \in \mathbb{R}^{B\times T\times C\times H\times W}$ with $T = 10$, $H = W = 224$, and $B = 16$. The data augmentation operations used in the experiment included brightness and contrast adjustments, horizontal flipping, random rotation, and color jitter. Both branches are optimized with AdamW: 2CLC uses lr $= 1\times10^{-4}$ and weight decay $= 1\times10^{-4}$, and CDA uses lr $= 1\times10^{-5}$ and weight decay $= 1\times10^{-4}$. We employ ReduceLROnPlateau (factor 0.5) with patience 5 (min lr $= 1\times10^{-6}$) for 2CLC and patience 1 (min lr $= 1\times10^{-7}$) for CDA. Training runs for at most 100 epochs with early stopping (patience 10). Loss weights are set to $w = 0.5$ and $u = 0.7$. All experiments were conducted using two NVIDIA RTX 4090 GPUs with 24 GB memory for parallel training. The values reported in the tables are mean ± standard deviation of 5 repeated experiments, where bold indicates the best result and underlining denotes the second-best result.

Table 2: Embryonic cleavage stage classification results on HED.

	Stage	ESOD [6]	CNNs-CRF [7]	R2+1D [15]	SCL-Embryo [8]	3CL-ECSC
pcF1	t2	82.42 ± 0.84	85.38 ± 4.58	88.57 ± 1.49	89.49 ± 3.50	**90.80** ± 1.62
	t3	0.00 ± 0.00	22.82 ± 8.15	26.32 ± 3.58	40.78 ± 7.15	**45.54** ± 4.18
	t4	65.07 ± 2.09	64.18 ± 8.00	73.69 ± 3.41	76.14 ± 2.19	**76.79** ± 2.58
	t5	0.89 ± 1.98	26.26 ± 7.08	14.20 ± 2.67	**34.33** ± 5.26	26.57 ± 1.97
	t6	5.07 ± 7.27	17.01 ± 8.99	17.43 ± 4.47	29.91 ± 3.31	**31.88** ± 5.19
	t7	1.81 ± 1.91	13.41 ± 4.24	17.17 ± 7.12	19.47 ± 2.82	**20.81** ± 2.52
	t8	58.29 ± 2.39	49.19 ± 11.76	64.25 ± 1.85	64.56 ± 3.79	**67.52** ± 4.19
	t9	78.65 ± 2.28	77.73 ± 3.49	79.76 ± 2.34	**82.36** ± 2.42	82.01 ± 2.01
Accuracy	–	64.89 ± 1.90	61.83 ± 1.08	68.86 ± 1.22	69.43 ± 2.39	**71.79** ± 1.79
F1	–	59.33 ± 2.58	60.66 ± 2.08	65.97 ± 1.47	69.16 ± 2.57	**70.26** ± 1.87
TA	–	22.00 ± 1.35	30.17 ± 6.99	36.55 ± 1.71	32.62 ± 1.73	**41.33** ± 4.89

This study employs the following metrics to evaluate classification performance at the cleavage stage: **overall accuracy**, **per-class F1-score** (pcF1), **weighted F1-score** (F1), and **transition accuracy** (TA). To evaluate the identification performance of adjacent stages, we compute transition accuracy $TA = \frac{Correct}{Total}$, where $Correct = \sum_{b=1}^{B}\sum_{t=2}^{T} c_t^{(b)}$, $Total = \sum_{b=1}^{B}\sum_{t=2}^{T} m_t^{(b)}$, where $m_t^{(b)}$ represents the total number of stage transitions, and $c_t^{(b)}$ represents the number of correct predictions of stage transitions. TA focuses on prediction quality at the boundaries of stages, thereby more directly reflecting the model's ability to distinguish between adjacent stages in areas prone to confusion.

3.3 Results and Analysis

Comparison with State-of-the-Art: We compare the proposed 3CL-ECSC with multiple representative methods, including ESOD [6], CNNs-CRF[7], R2+1D[15], and SCL-Embryo[8]. All experiments were conducted on mouse embryo datasets and human embryo datasets. As shown in Tables 1 and 2, 3CL-ECSC achieves the best overall Accuracy, F1-score, and Transition Accuracy (TA) on both MED and HED. The improvement in TA is particularly notable, indicating that division-prior-guided correction improves stage-boundary recognition and sequence consistency. On MED, the gains are mainly observed in sparsely sampled and adjacent confusing stages, while on the more challenging HED, 3CL-ECSC remains robust under stronger inter-embryo variability, imaging noise, and class imbalance.

Ablation Experiments: To validate each component, we conduct ablation experiments on 2CLC, CRE, and CDA. As shown in Table 3, the full model achieves the best Accuracy, F1-score, and Transition Accuracy, indicating that the three modules are complementary. CRE improves overall stability by reducing background interference, while CDA contributes more to Transition Accuracy, demonstrating its effectiveness in stage-boundary discrimination.

4 Conclusion

To address cell overlap and category imbalance in embryonic cleavage-stage classification, this paper proposes 3CL-ECSC, a cell-division-aware category-

Table 3: Ablation study on the human embryo dataset.

2CLC	CRE	CDA	Accuracy	F1	TA
✓	×	×	69.43 ± 2.39	69.16 ± 2.57	32.62 ± 1.73
✓	✓	×	70.11 ± 1.36	69.09 ± 1.69	32.53 ± 2.02
✓	×	✓	70.75 ± 2.01	68.51 ± 2.24	38.32 ± 3.59
✓	✓	✓	**71.79** ± 1.79	**70.26** ± 1.87	**41.33** ± 4.89

enhanced contrastive learning framework. By integrating CDA and 2CLC, the proposed method improves stage-transition discrimination, sparse-stage robustness, and temporal consistency. Experiments on human and mouse embryo datasets show that 3CL-ECSC outperforms existing state-of-the-art methods in both classification and transition accuracy. Nevertheless, the current model assumes monotonic stage progression and does not explicitly handle abnormal patterns such as reverse cleavage, which will be explored in future work.

Acknowledgements. This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (2024A1515010977, 2025A1515060013), Scientific Foundation for Youth Scholars of Shenzhen University, Guangdong Provincial Key Laboratory (2023B1212060076, 2017B030314073).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adamson, G.D., Zegers-Hochschild, F., Dyer, S.: Global fertility care with assisted reproductive technology. Fertility and Sterility **120**(3 Pt 1), 473–482 (2023). https://doi.org/10.1016/j.fertnstert.2023.01.013
2. Lundin, K., Park, H.: Time-lapse technology for embryo culture and selection. Upsala Journal of Medical Sciences **125**(2), 77–84 (2020). https://doi.org/10.1080/03009734.2020.1728444
3. Sayed, S., Reigstad, M.M., Petersen, B.M., Schwennicke, A., Wegner Hausken, J., Storeng, R.: Time-lapse imaging derived morphokinetic variables reveal association with implantation and live birth following in vitro fertilization: A retrospective study using data from transferred human embryos. PLOS ONE **15**(11), e0242377 (2020). https://doi.org/10.1371/journal.pone.0242377
4. Zhang, C., Shi, X., Yin, X., Sun, J., Zhao, J., Zhang, Y.: Cleavage-stage embryo segmentation using SAM-based dual branch pipeline: development and evaluation with the CleavageEmbryo dataset. Bioinformatics **41**(4), btae617 (2025). https://doi.org/10.1093/bioinformatics/btae617
5. He, C., Karpavičiūtė, N., Hariharan, R., Jacques, C., Chambost, J., Malmsten, J., Zaninovic, N., Wouters, K., Fréour, T., Hickman, C., Vasconcelos, F.: Embryo graphs: Predicting human embryo viability from 3D morphology. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) MICCAI 2024. LNCS, vol. 15004, pp. 80–90. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72083-3_8
6. Lockhart, L., Saeedi, P., Au, J., Havelock, J.: Automating embryo development stage detection in time-lapse imaging with synergic loss and temporal learning. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert,

C. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 540–549. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_52
7. Lukyanenko, S., Jang, W.-D., Wei, D., Struyven, R., Kim, Y., Leahy, B., Yang, H., Rush, A., Ben-Yosef, D., Needleman, D., Pfister, H.: Developmental stage classification of embryos using two-stream neural network with linear-chain conditional random field. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 363–372. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_35
8. Hachani, H.Y., Bouthemy, P., Fromont, E., Ruffini, S., Laffont, L., de Paula Reis, A.: Supervised contrastive learning for cell stage classification of animal embryos. arXiv preprint arXiv:2502.07360 (2025)
9. Sun, G., Cholakkal, H., Khan, S., Khan, F.S., Shao, L.: Fine-grained recognition: Accounting for subtle differences between similar classes. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 12047–12054 (2020)
10. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 18661–18673 (2020)
11. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.324
12. Cicconet, M., Gutwein, M., Gunsalus, K.C., Geiger, D.: Label free cell-tracking and division detection based on 2D time-lapse images for lineage analysis of early embryo development. Computers in Biology and Medicine **51**, 24–34 (2014). https://doi.org/10.1016/j.compbiomed.2014.04.011
13. Gomez, T., Feyeux, M., Boulant, J., Normand, N., David, L., Paul-Gilloteaux, P., Fréour, T., Mouchère, H.: A time-lapse embryo dataset for morphokinetic parameter prediction. Data in Brief **42**, 108258 (2022). https://doi.org/10.1016/j.dib.2022.108258
14. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5552–5561 (2019). https://doi.org/10.1109/ICCV.2019.00565
15. Barhoun, A., Balafar, M.A., Golzari Oskouei, A., Sadeghi, L.: Human embryo stage classification using an enhanced R(2+1)D model and dynamic programming with optimized datasets. Biomedical Signal Processing and Control **107**, 107841 (2025). https://doi.org/10.1016/j.bspc.2025.107841