MICCAI

Time-Lapse Video-Based Embryo Grading via Complementary Spatial-Temporal Pattern Mining

Yong Sun[1], Yipeng Wang[2], Junyu Shi[1], Zhiyuan Zhang[1], Yanmei Xiao[3], Lei Zhu[1], Manxi Jiang[3], and Qiang Nie[1]

The Hong Kong University of Science and Technology (Guangzhou)
Shanghai Jiao Tong University
Center for Reproductive Medicine, Guangdong Second Provincial General Hospital
qiangnie@hkust-gz.edu.cn
manxijiang@yahoo.com

Abstract. Artificial intelligence has recently shown promise in automated embryo selection for In-Vitro Fertilization (IVF). However, current approaches either address partial embryo evaluation lacking holistic quality assessment or target clinical outcomes inevitably confounded by extra-embryonic factors, both limiting clinical utility. To bridge this gap, we propose a new task called **Video-Based Embryo Grading** - the first paradigm that directly utilizes full-length time-lapse monitoring (TLM) videos to predict embryologists' overall quality assessments. To support this task, we curate a real-world clinical dataset comprising over 2,500 TLM videos, each annotated with a grading label indicating the overall quality of embryos. Grounded in clinical decision-making principles, we propose a Complementary Spatial-Temporal Pattern Mining (**CoSTeM**) framework that conceptually replicates embryologists' evaluation process. The CoSTeM comprises two branches: (1) a morphological branch using a Mixture of Cross-Attentive Experts layer and a Temporal Selection Block to select discriminative local structural features, and (2) a morphokinetic branch employing a Temporal Transformer to model global developmental trajectories, synergistically integrating static and dynamic determinants for grading embryos. Extensive experimental results demonstrate the superiority of our design. This work provides a valuable methodological framework for AI-assisted embryo selection. The source code is available at https://github.com/RIL-Lab/CoSTeM.

Keywords: Time-lapse Video · In-Vitro Fertilization · Deep Learning.

1 Introduction

Infertility has become a significant global issue [13], with in-vitro fertilization (IVF) as the primary treatment. The key to IVF success is selecting high-potential embryos, traditionally done through embryologist manual assessments on microscopy images of in-vitro fertilized embryos. However, the evaluation process remains inherently time-consuming and subject to a certain level of inter-observer variability, prompting the development of AI-assisted automated

embryo selection systems [23, 20, 6, 9, 11]. Existing approaches [14, 26, 20, 23] primarily focused on the automated analysis of isolated microscopic images. However, in-vitro fertilized embryos require multiple days of dynamic developmental progression before transfer. Consequently, image-centric methodologies overlook holistic developmental information and fundamentally disregard morphokinetic features, such as cell division timing, which are essential for comprehensive evaluations of embryo quality.

The advent of Time-Lapse Monitoring (TLM) technology has enabled the continuous recording of embryonic development process, providing essential morphological and morphokinetic cues for embryo selection [17]. A few studies [22, 16, 5, 18, 15] have explored video-based methods for automating specific aspects of embryo selection. Some [22, 16] focused on blastocyst quality assessment using day 5–6 video clips, while others aimed to predict blastocyst formation from early-stage TLM videos [5, 24]. Additionally, research [18, 19] utilized TLM videos for developmental stage classification by modeling temporal correspondences. Although these methods have demonstrated great potential, they nevertheless fail to provide a prediction of the overall embryonic quality necessary for clinical decision-making regarding transplantation. More recently, Kim *et al.* sought to bypass expert evaluation by directly predicting clinical outcomes for automated embryo selection, integrating TLM videos with electronic health records (**EHR**) based on ViViT [1]. However, clinical outcomes are often confounded by a variety of external factors (*e.g.*, uterine environment, maternal age), which increases modeling complexity and leads to limited prediction accuracy. Currently, a common clinical practice for selecting embryos among clinicians is to first evaluate the embryonic development process by holistically analyzing TLM videos, synthesizing morphological and morphokinetic features according to [2]. Embryos with high developmental potential are then prioritized for transfer. Considering this, a viable alternative for automated embryo selection is directly predicting embryologists' holistic evaluation results based on TLM videos, which is clinically translational and remains unexplored.

To close this gap, we propose a new task named **Video-Based Embryo Grading** - a data-driven paradigm that automatically categorizes in-vitro fertilized embryos into three quality grades (poor/fair/good) by holistically analyzing TLM videos. This paradigm takes TLM videos as input and systematically models embryologists' comprehensive evaluation criteria, generating quality assessments aligned with clinical standards. To support this task, we curate a dataset of 2,596 TLM videos from real-world clinical IVF practices. Each video is paired with a grading label indicating the overall quality of the corresponding embryo. A notable characteristic of this dataset is the variation in video recording lengths, which stems from the natural inconsistencies in embryonic development rates and various clinical operational circumstances. This presents AI systems with two challenges: adaptive temporal modeling to detect critical morphokinetic events without manual annotations, and robust spatiotemporal reasoning to identify fragmented morphological features across various developmental stages for comprehensive quality assessment.
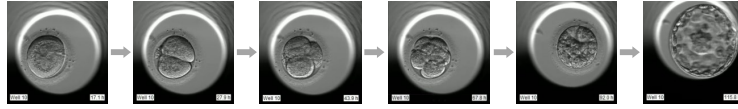
Fig. 1. Representative frames extracted at key developmental stages of a embryo from "good quality" category

As a powerful baseline, we introduce a clinically motivated framework termed CoSTeM that jointly models complementary morphological and morphokinetic patterns for grading embryos. Built upon CLIP pretrained visual encoder [21], we first employ a Bypass Adaption Network to modulate and integrate multiple intermediate features for each frame. These features are decomposed and processed through two specialized streams: 1) A **morphological branch** that progressively identifies and consolidates discriminative morphological patterns through Mixture of Cross-Attentive Experts Layer and Temporal Selection Block, and 2) a **morphokinetic branch** that explores embryo developmental trajectories via Temporal Transformer. Extensive comparisons with state-of-the-art embryo selection systems and general video action recognition models demonstrate the superiority of our approach. In summary, our contributions are three-fold:

1. We introduce a new task termed video-based embryo grading that predicts embryologists' holistic embryo evaluations directly from TLM videos, accompanied by a curated benchmark dataset to catalyze automated embryo selection research in reproductive medicine.
2. We propose a clinically inspired dual-branch framework CoSTeM that synergistically models complementary morphological and morphokinetic patterns for automatically grading in-vitro fertilized embryos.
3. We conduct comprehensive experiments demonstrating the superiority and the rationality of our designs, establishing a foundational benchmark for future research in the field of AI-assisted embryo selection.

2   Dataset

Our dataset contains 2,596 TLM videos collected from the Center for Reproductive Medicine at Guangdong Second Provincial General Hospital. Each video lasts over 110 hours after fertilization with natural variations in duration reflecting real-world clinical conditions.

In IVF cycles, TLM videos of each embryo are subjected to daily evaluations by embryologists. Specifically, key parameters such as pronucleus number, cell fragmentation degree, and inner cell mass (ICM) quality are scored using the standardized Istanbul Consensus criteria. Based on these daily assessments, embryologists integrate the morphological and morphokinetic features in accordance with clinical embryo transfer guidelines to classify the embryos into three quality categories: Good, Fair, and Poor. Embryos classified as Good possess the
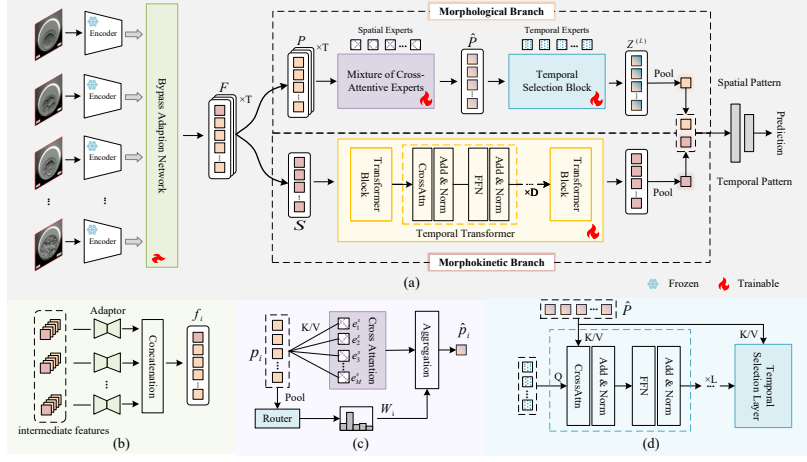
Fig. 2. Architecture of the proposed CoSTeM framework. (a) Schematic overview of our framework. (b) Bypass Adaption Network for feature modulation and integration. (c) Mixture of Cross-Attentive Experts (MCAE) layer for adaptive spatial feature selection. (d) Temporal Selection Block (TSB) for local frame selection.

best developmental potential as displayed in Fig. 1. For algorithm development, the dataset is split into training and test sets (7: 3 ratio) using patient-stratified sampling to maintain a consistent class distribution and ensure that all embryos from the same patient are assigned to only one set.

3   Methods

3.1   Overall Framework

The fundamental concept of our CoSTeM framework is to emulate the evaluation principle of embryologists for prediction by concurrently mining both static structural characteristics and dynamic developmental patterns.

Given a TLM sequence $V = \{x_t\}_{t=1}^{T}$ with T frames, each frame x_i is partitioned into N spatial tokens and prepended with a $[cls]$ token. These tokens are sent to a pretrained image encoder to extract hierarchical features. To reduce domain gap and construct holistic representations for frames from various developmental stages, a Bypass Adaptation Network, as shown in Fig. 2 (b), modulates intermediate features from layers $\mathcal{L} = \{3, 6, 9, 12\}$ through MLP adaptors [12], resulting in frame feature sequence $\mathbf{F} = \{f_t\}_{t=1}^{T}$, $f_t \in \mathbb{R}^{(N+1) \times C}$, where $C = 768$ is the feature dimension. Subsequently, $\mathbf{F}$ is decomposed into two complementary streams: a patch token sequence $\mathbf{P} = \{p_t\}_{t=1}^{T}$, $p_t \in \mathbb{R}^{N \times C}$ encoding local spatial details for the **Morphological Branch**, and a class token sequence $\mathbf{S} \in \mathbb{R}^{T \times C}$ capturing global semantics for the **Morphokinetic Branch**.

The morphological branch aims to selectively capture critical morphological cues (*i.e.*, spatial pattern) from $\mathbf{P}$. Specifically, a Mixture of Cross-Attentive Experts (MCAE) layer dynamically selects discriminative spatial features per frame using M spatial experts, followed by a Temporal Selection Block to identify critical local frames using K temporal experts.

The morphokinetic branch aims to capture global developmental trajectories (*i.e.*, temporal pattern) using $\mathbf{S}$. In particular, a Temporal Transformer consisting of $D = 2$ cascaded Transformer Block is employed as in ViViT [1] to perform global temporal modeling through self-attention.

Ultimately, the spatial pattern and temporal pattern are concatenated together along the channel dimension and sent to an MLP classifier for prediction.

3.2 Mixture of Cross-Attentive Experts

During embryonic development, various stages exhibit distinct discriminative features. To adaptively focus on stage-specific morphological features, we propose the Mixture of Cross-Attentive Experts (MCAE) layer, inspired by the Mixture of Experts [7]. As in Fig. 2 (c), a set of spatial experts $\mathbf{E^s} = \{e_1^s, ..., e_M^s\} \in \mathbb{R}^{M \times C}$ is defined, each conceptually representing a feature selector of a particular embryonic developmental stage. For the i-th frame's patch tokens $p_i \in \mathbb{R}^{N \times C}$, each expert e_j^s acts as a query to attend to p_i for feature selection:

$$\mathbf{H}_i^{[j]} = \mathrm{CrossAttn}\left(e_j^s, p_i\right) \in \mathbb{R}^C, \quad \forall j \in \{1, \ldots, M\} \tag{1}$$

Concurrently, a MLP router calculates belonging weights $\mathbf{W}_i \in \mathbb{R}^M$ based on the frame's global averaged context. The $\mathbf{W}_i^{[j]}$ approximates the probability of i-th frame belongs to j-th developmental stage. Afterwards, the final condensed feature for i-th frame is a weighted combination:

$$\hat{p}_i = \sum_{j=1}^{M} \mathbf{W}_i^{[j]} \cdot \mathbf{H}_i^{[j]} \in \mathbb{R}^C \tag{2}$$

Note that these learnable spatial experts and the cross-attention layer are shared across frames, and feature selection is performed in parallel for each frame. The output of MCAE is spatially selected feature sequence $\hat{\mathbf{P}} = \{\hat{p}_t\}_{t=1}^T \in \mathbb{R}^{T \times C}$.

3.3 Temporal Selection Block

According to clinical guidelines, different frames within the TLM video have varying degrees of decisive impact on the final embryo quality, and the lengthy frame sequence may also contain potential noise. Therefore, after conducting stage-specific feature selection in the spatial dimension, it is necessary to further select critical local frames in the temporal dimension.

Inspired by DETR [4], we further define K learnable temporal experts $\mathbf{E^t} = \{e_1^t, ..., e_K^t\} \in \mathbb{R}^{K \times C}$ to select key local frames. As depicted in Fig. 2 (d), we

consider $\mathbf{Z}^{(0)} = \mathbf{E^t}$ as queries and $\hat{\mathbf{P}}$ as key and values, passing them through $L = 2$ Temporal Selection Layer contains cross-attention and feed-forward network,

$$\mathbf{Z}^{(l)} = \text{FFN}\left(\text{CrossAttn}(\mathbf{Z}^{(l-1)}, \hat{\mathbf{P}})\right) \quad \text{for } l = 1, 2, ..., L \tag{3}$$

where $\mathbf{Z}^{(L)} \in \mathbb{R}^{K \times C}$ is the final temporally selected features. Through cross-attention, learnable temporal experts can adjust their representations and adaptively select frames that contribute most to the final prediction.

3.4   Loss Functions

The training objective combines a standard cross-entropy loss $\mathcal{L}_{cls}$ and a diversity loss $\mathcal{L}_{div}$ that minimizes pairwise cosine similarities between temporal experts $\mathbf{E^t}$ and encourages them to select distinct local frames for richer morphological cues. The total loss function is:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{div}, \;\; \mathcal{L}_{\text{div}} = \sum_{i=1}^{K} \sum_{j=1}^{K} \left| \frac{(e_i^t)^\top e_j^t}{\|e_i^t\| \|e_j^t\|} \right| \tag{4}$$

Here, $|\cdot|$ denotes absolute value operation, $\|\cdot\|$ denotes L_2 norm, and λ is the balancing coefficient of the two objectives.

4   Experiments

4.1   Experiment Setup

Implementation Details CLIP-pretrained ViT-B/16 [21] is chosen as the feature extractor. The M and K are both set to 8. The λ is set to 0.1. The training comprises 50 epochs with a batch size of 8, using the AdamW optimizer and a Cosine learning rate scheduler. The learning rate increases from 1.0e-6 to 2.5e-5 after a 5-epoch warmup. Label smoothing of 0.1 is applied to the cross-entropy loss to mitigate overconfidence. For data processing, we uniformly sample 64 frames at 2-hour intervals between 16 and 142 hours post-fertilization from TLM videos, with zero-padding applied if frames are unavailable at specific time points. Original 500×500 frames are resized to 256×256. During training, we apply random 224×224 cropping and data augmentations including horizontal flip, rotation, and color jitter. At inference, frames are center-cropped to 224×224. We normalized each frame using the statistical information of the collected dataset.

Evaluation Metrics We trained all the models on the partitioned training set and reported their best performances on the test set. There were no overlapping samples between the training and test sets. Accuracy, Precision, Recall, and F1-score were selected as evaluation metrics to make comprehensive comparisons. The macro-averaged F1-score was prioritized as it balances two critical clinical risks: preventing non-viable transfers (precision-driven) and preserving high-potential embryos (recall-sensitive), while accounting for class imbalance.

Table 1. The classification performance comparisons between selected methods and our proposed CoSTeM framework.

Model	Accuracy	Precision	Recall	F1-score
Methods for AI-assisted Embryo Selection				
Xception-LSTM [16]	0.8005	0.6726	0.6982	0.6831
KeyFrameFusion [3]	0.8082	0.6880	0.7122	0.6982
EmbryoViViT [15]	0.8350	0.7197	0.7229	0.7213
Methods for Video Action Recognition				
ResNet50-3D [10]	0.8350	0.7444	0.6910	0.7039
SlowFast [8]	0.8120	0.6959	0.7105	0.6994
AIM [25]	0.8606	0.7524	0.7448	0.7477
CoSTeM **(ours)**	**0.8606**	**0.7746**	**0.7595**	**0.7618**

Table 2. Ablation study results for our key designs.

Model	Accuracy	Precision	Recall	F1-score
w.o. Morphological Branch	0.8082	0.6921	0.6942	0.6862
w.o. Morphokinetic Branch	0.8350	0.7226	0.7475	0.7327
w.o. Bypass Adaption Network	0.8414	0.7283	0.7318	0.7297
w.o. $\mathcal{L}_{div}$	0.8542	0.7730	0.7439	0.7487
CoSTeM	**0.8606**	**0.7746**	**0.7595**	**0.7618**

4.2 Comparisons with Advanced Methods

We benchmarked our CoSTeM against state-of-the-art AI-assisted embryo selection methods [16, 3, 15] and general video action recognition systems [10, 8, 25] for comparison, implementing all approaches using their official codebases or paper specifications under identical training protocols. As shown in Table 1, our proposed CoSTeM framework achieves the best performance. Compared with previous AI-assisted embryo selection methods that lack tailored designs, our framework demonstrates marked superiority. The introduced embryo grading task necessitates the use of extended frame sequences unlike action recognition which typically analyzes sparse frame sequences, posing challenges to the model's adaptive feature extraction capabilities. Among models for video action recognition, the AIM [25] performs best, achieving an accuracy on par with ours. However, through selective and complementary pattern mining in the proposed CoSTeM, we surpass AIM in other critical metrics: precision (+2.95%), recall (+1.97%), and F1 score (+1.89%). This underscores the superior efficacy of our clinically inspired framework.

4.3 Ablation Study

Effectiveness of Core Designs We performed an ablation study by iteratively removing key components. As shown in Table 2, removing any single branch significantly degrades performance. This highlights their complementary roles and
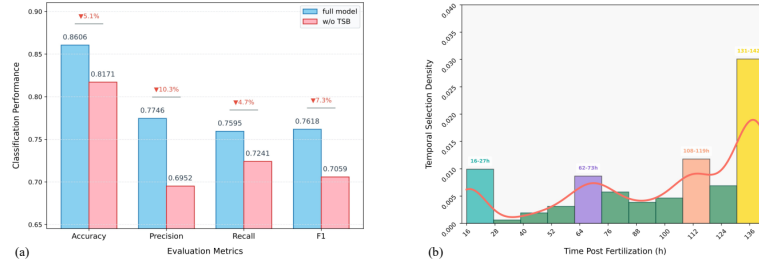
Fig. 3. Ablation results of TSB. (a) is the performance comparisons of Temporal Selection Block and average pooling. (b) is the temporal density distribution of selected frames by temporal experts in TSB.

Table 3. Comparison results of different frame feature selection methods.

Model	Accuracy	Precision	Recall	F1-score
Average Pooling	0.8440	0.7288	0.7411	0.7338
Max Pooling	0.8376	0.7205	0.7427	0.7278
Shared Expert	0.8453	0.7391	0.7389	0.7385
MCAE	**0.8606**	**0.7746**	**0.7595**	**0.7618**

validates the framework's alignment with clinical decision logic. Moreover, the results show that morphological features contribute more than morphokinetic features in our framework. Disabling the Bypass Adaption Network disrupts holistic frame representation, causing performance drops. Removing the $\mathcal{L}_{div}$ obviously reduces performance, showing that temporal expert constraints encourage diverse and complementary features.

Effectiveness of MCAE We evaluated the MCAE layer for selecting stage-specific morphological features using various operations. Table 3 shows that MCAE outperforms others, proving its effectiveness in spatial feature selection. Average pooling treats each patch equally, failing to extract stage-specific features for each frame. Max pooling, on the other hand, is susceptible to noise and can result in information loss. The shared expert approach, which uses only one shared expert across frames, also struggles to capture stage-specific features.

Effectiveness of TSB Fig. 3 (a) compares the performance when the Temporal Selection Block (TSB) is replaced with average pooling. Notably, there is a significant decline in all evaluation metrics, highlighting the crucial role of the TSB in accurately assessing embryo quality. Fig. 3 (b) shows the density distribution of frames selected by the temporal experts in the first Temporal Selection Layer. Remarkably, these experts mainly focus on key embryonic developmental stages: pronucleus formation stage, cleavage stage, and blastocyst stage, which

aligns perfectly with standard embryologist protocols, confirming that our TSB can adaptively capture and select biologically diagnostic features.

5   Conclusion

In conclusion, this paper introduces a novel task named Video-Based Embryo Grading, aiming at bridging the divide between current AI-assisted embryo selection paradigms and clinical practices, thereby enabling end-to-end, trustworthy automated embryo selection. To underpin this task, we further curate a dataset that aligns with real-world clinical scenarios. Additionally, we propose CoSTeM, a robust baseline model inspired by clinical decision-making principles. Experimental results highlight the substantial superiority of our method and underscore its potential to propel advancements in AI-assisted embryo selection.

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

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6836–6846 (2021)
2. Balaban, B., Brison, D., Calderon, G., Catt, J., Conaghan, J., Cowan, L., Ebner, T., Gardner, D., Hardarson, T., Lundin, K., et al.: Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. Reproductive Biomedicine Online **22**(6), 632–646 (2011)
3. Borna, M.R., Sepehri, M.M., Maleki, B.: An artificial intelligence algorithm to select most viable embryos considering current process in ivf labs. Frontiers in Artificial Intelligence **7**, 1375474 (2024)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
5. Chen, T., Cheng, Y., Wang, J., Yang, Z., Zheng, W., Chen, D.Z., Wu, J.: Automating blastocyst formation and quality prediction in time-lapse imaging with adaptive key frame selection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 445–455. Springer (2022)
6. Eswaran, M., Balasubramanie, P., Salman, J.M.: Deep learning algorithms for time-lapse image sequence-based automated blastocyst quality detection. In: 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC). pp. 1246–1252. IEEE (2023)
7. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)

8. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6202–6211 (2019)

9. Garg, K., Dev, A., Bansal, P., Mittal, H.: An efficient deep learning model for embryo classification. In: 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence). pp. 358–363. IEEE (2024)

10. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018)

11. He, C., Karpavičiūtė, N., Hariharan, R., Jacques, C., Chambost, J., Malmsten, J., Zaninovic, N., Wouters, K., Fréour, T., Hickman, C., et al.: Embryo graphs: Predicting human embryo viability from 3d morphology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 80–90. Springer (2024)

12. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)

13. Inhorn, M.C., Patrizio, P.: Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21st century. Human Reproduction Update **21**(4), 411–426 (2015)

14. Khosravi, P., Kazemi, E., Zhan, Q., Malmsten, J.E., Toschi, M., Zisimopoulos, P., Sigaras, A., Lavery, S., Cooper, L.A., Hickman, C., et al.: Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. NPJ Digital Medicine **2**(1), 21 (2019)

15. Kim, J., Shi, Z., Jeong, D., Knittel, J., Yang, H.Y., Song, Y., Li, W., Li, Y., Ben-Yosef, D., Needleman, D., et al.: Multimodal learning for embryo viability prediction in clinical ivf. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 542–552. Springer (2024)

16. Kragh, M.F., Rimestad, J., Berntsen, J., Karstoft, H.: Automatic grading of human blastocysts from time-lapse imaging. Computers in Biology and Medicine **115**, 103494 (2019)

17. Liao, Q., Zhang, Q., Feng, X., Huang, H., Xu, H., Tian, B., Liu, J., Yu, Q., Guo, N., Liu, Q., et al.: Development of deep learning algorithms for predicting blastocyst formation and quality by time-lapse monitoring. Communications Biology **4**(1), 415 (2021)

18. Lockhart, L., Saeedi, P., Au, J., Havelock, J.: Automating embryo development stage detection in time-lapse imaging with synergic loss and temporal learning. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 540–549. Springer (2021)

19. Nguyen, T.P., Pham, T.T., Nguyen, T., Le, H., Nguyen, D., Lam, H., Nguyen, P., Fowler, J., Tran, M.T., Le, N.: Embryosformer: Deformable transformer and collaborative encoding-decoding for embryos stage development classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1981–1990 (2023)

20. Ou, Z., Zhang, Q., Li, Y., Meng, X., Wang, Y., Ouyang, Y., Li, Z., Zeng, K.: Classification of human embryos by using deep learning. In: International Conference on Cloud Computing, Performance Computing, and Deep Learning (CCPCDL 2023). vol. 12712, pp. 342–347. SPIE (2023)

21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)

22. Rajendran, S., Brendel, M., Barnes, J., Zhan, Q., Malmsten, J.E., Zisimopoulos, P., Sigaras, A., Ofori-Atta, K., Meseguer, M., Miller, K.A., et al.: Automatic ploidy prediction and quality assessment of human blastocysts using time-lapse imaging. Nature Communications **15**(1), 7756 (2024)

23. Truong, V.V., Le, M.H., Phu, P.L., Phan, H., Le, T., Nguyen, H., Pham, V.D., Le, X.H.: A deep learning approach to embryo quality assessment. In: The International Conference on Intelligent Systems & Networks. pp. 452–461. Springer (2024)

24. Xie, X., Yan, P., Cheng, F.Y., Gao, F., Mai, Q., Li, G.: Early prediction of blastocyst development via time-lapse video analysis. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2022)

25. Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., Li, M.: Aim: Adapting image models for efficient video action recognition. In: The Eleventh International Conference on Learning Representations (2023)

26. Zeman, A., Maerten, A.S., Mengels, A., Sharon, L.F., Spiessens, C., de Beeck, H.O.: Deep learning for human embryo classification at the cleavage stage (day 3). In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I. pp. 278–292. Springer (2021)