Automated Sperm Assessment Framework and Neural Network Specialized for Sperm Video Recognition

Takuro Fujii[1] Hayato Nakagawa[1] Teppei Takeshima[2] Yasushi Yumura[2] Tomoki Hamagami[1]

[1]Yokohama National University [2]Yokohama City University Medical Center

{tkr.fujii.ynu, haya.nakagawa.ynu, hamagami.ynu}@gmail.com {yumura, teppei_t}@yokohama-cu.ac.jp

Abstract

Infertility is a global health problem, and an increasing number of couples are seeking medical assistance to achieve reproduction, at least half of which are caused by men. The success rate of assisted reproductive technologies depends on sperm assessment, in which experts determine whether sperm can be used for reproduction based on morphology and motility of sperm. Previous sperm assessment studies with deep learning have used datasets comprising images that include only sperm heads, which cannot consider motility and other morphologies of sperm. Furthermore, the labels of the dataset are one-hot, which provides insufficient support for experts, because assessment results are inconsistent between experts, and they have no absolute answer. Therefore, we constructed the video dataset for sperm assessment whose videos include sperm head as well as neck and tail, and its labels were annotated with soft-label. Furthermore, we proposed the sperm assessment framework and the neural network, RoSTFine, for sperm video recognition. Experimental results showed that RoSTFine could improve the sperm assessment performances compared to existing video recognition models and focus strongly on important sperm parts (i.e., head and neck). Our code is publicly available at `https://github.com/FTKR12/RoSTFine`.

1. Introduction

Infertility is a critical problem around the world. This afflicts one in six couples, at least half of whom are caused by men [19, 16]. Assisted reproductive technologies (ARTs), such as in-vitro-fertilization (IVF) and intracytoplasmic sperm injection (ICSI), are used depending on the cause and severity of infertility. However, ARTs are currently successful in only approximately 33% of cases, and this main reason is suboptimal sperm selection [27]. In the sperm selection process, at least three fertility factors are typically examined; sperm concentration, motility and morphology [20]. In sperm selection, motility and sperm concentration are assessed using computer-aided semen analysis (CASA) systems, which are sensitive to sample preparation and equipment setup [37, 2]. Morphology is assessed manually by experts, which are inconsistent among individuals and clinics owing to subjective criteria, in addition to being time-consuming and labor-intensive [13, 8, 25, 22]. Therefore, an End2End sperm assessment framework, considering all three factors, is in high demand and promising for improving reproductive success.

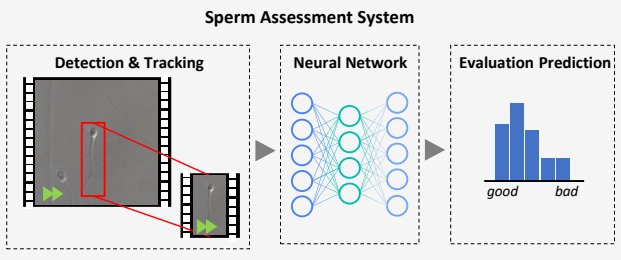

Figure 1. Proposed sperm assessment framework whose inputs are sperm videos. It detects and tracks a target sperm in a video taken from a microscope to make an input of a neural network, and then the neural network predicts the *grade distribution* of the sperm.

Deep learning has shown promise for standardizing and automating sperm assessments [40, 26, 24]. Several studies have addressed sperm assessment using deep learning [32, 30, 35, 18, 39]. However, there are two problems. First, at the viewpoint of an End2End sperm assessment framework, despite the fact that the head morphology as well as other morphologies and motility are important factors in sperm selection [28], these studies assessed sperm by classifying only head morphology, ignoring neck and tail morphology and sperm motility. Second, at the viewpoint of support for experts, the annotation of the data used in these studies was a one-hot label for classification tasks. As the labels of a dataset for sperm assessment, soft-labels are better because there is no absolute answer and experts have inconsistent assessments. Soft-labels enable to aid experts in flexible decision-making because soft-labels are informative.

For video application tasks, it is straightforward ap-

proach to apply existing video recognition models directly. Their models, however, are designed for general videos whose domain is different from sperm videos. Therefore, we should design a model specialized for sperm video recognition, but there are two challenges. First, it is difficult for a model to capture only a target sperm because videos are background dominant, and dust, air bubbles and non-target sperm can interfere with sperm recognition. Second, the model must capture the diverse characteristics of sperm, such as motility, morphologies and dependencies of the head, neck and tail. It is particularly difficult to capture the sperm tail, which is often assimilated into the background.

To solve these problems, in this study, we constructed a video dataset annotated with soft-labels, and proposed an End2End framework for sperm assessment and a neural network for sperm video recognition. When constructing the dataset, each of the 40 experts annotated one of the five grades for each sample, thus, the labels are soft-labels, 5-grade histograms, which we refer to as *grade distribution*. The details of the dataset are presented in §3. The proposed framework, illustrated in Figure 1, detects and tracks a target sperm in a video captured from a microscope to provide an input to a neural network, and then predicts the *grade distribution* of the sperm. The proposed neural network, *Role-Separated Transformer for Fine-Grained and Diverse Sperm Feature Extraction* (RoSTFine), can focus only on a target sperm and extract fine-grained and diverse sperm features. The details of RoSTFine are presented in §4. The experimental results (§5) show that RoSTFine achieves a higher performance than existing video recognition models, such as TimeSformer and SlowFast [12]. Further Analysis showed that RoSTFine can attend strongly to the sperm head and neck which are important for sperm assessment, and can generate effective features.

Our main contributions are summarized as follows: (1) To address reproduction that is an important issue but little-studied in computer vision fields, we constructed a video dataset annotated with soft-labels for sperm assessment; (2) We developed an automated framework for sperm assessment; (3) We developed a sperm-specific model, RoSTFine, to capture important sperm characteristics; (4) Experimental results showed that RoSTFine improved assessment performances on three evaluation metrics; (5) RoSTFine can focus on important sperm parts, such as the head and neck.

2. Related Work

2.1. Datasets and Methods for Sperm Assessment

There are three publicly available datasets for sperm assessment. The Sperm Morphology Image Dataset (SMIDS) [17] comprises 3000 images of sperm head, and is annotated in three classes of normal, abnormal, and non-sperm. The human sperm head morphology (HuSHeM) dataset [33] comprises 216 images of stained sperm head, and is annotated in four classes of normal, tapered, pyriform, and amorphous. The Laboratory for Scientific Image Analysis Gold-standard for Morphological Sperm Analysis (SCIAN) dataset [7] comprises 1132 images of sperm head, and is annotated in five classes of small and the same 4 classes as the HuSHeM dataset. The classes of the HuSHeM and SCIAN dataset are subsets of the categories of sperm head morphology which World Health Organization (WHO) provided in the semen analysis manual [28].

In some studies, these datasets have been used to train deep learning models. Riordon *et al*. [32] fine-tuned VGG16 pretrained on ImageNet to classify sperm head morphology. Spencer *et al*. [35] used a stacked ensemble comprising VGG16, VGG19, ResNet-34, and DenseNet-161. Yüzkat *et al*. [39] designed and fusioned six CNN-based models. Ilhan *et al*. [18] proposed a computational framework that includes multistage cascade-connected preprocessing techniques, region-based descriptor features, and nonlinear kernel SVM-based learning.

However, these datasets and studies are incomplete for sperm assessment because they focus only on sperm head morphology, although other morphologies and motility are also important. Additionally, although there are no absolute answer and assessment results are sometimes inconsistent, the labels of the datasets are one-hot labels. In this study, we constructed a sperm video dataset annotated with soft-labels, and developed a sperm recognition model to capture sperm morphology and motility.

2.2. Video Recognition Models

Video recognition is one of the most popular computer vision fields, and many neural networks have been developed. Video recognition models are classified into two categories: CNN- and Transformer-based models. In CNN-based models, Two-Stream I3D [5] has an RGB stream and an optical flow stream, and its backbone model is a 3D convolutional network. SlowFast [12] involves a slow pathway to capture shapes and slow motion and a fast pathway to capture fast motion and movement. In Transformer-based models, ViViT [3] combines spatial and temporal attention in various ways. TimeSformer [4] proposes various methods for calculating temporal attention. Althogh CNN-based models can achieve high performance even with small datasets, owing to their strong inductive bias, they have a narrow receptive field and are poor at capturing long dependencies [38]. Transformer-based models require large dataset for high performance owing to their weak inductive bias, however, they have a wide receptive field and can capture long dependencies [36]. Furthermore, the model structures are flexible and easy to operate, making them useful not only in computer vision but also in various fields, such

as natural language processing [10], speech recognition [6], and multi-modal processing [31].

These models achieve high performances in some benchmark datasets, such as Kinetics [5], Diving-48 [21], and Something-Something-V2 [14], but these models and datasets have been developed for general video recognition. However, sperm videos are different from general videos because they are captured using a microscope. Therefore, sperm-specific models must be developed. In [32, 35], the effectiveness of fine-tuning models pretrained on general images in sperm image recognition was shown. Inspired by this and the high operability of local features, we developed Transformer-based model to utilize a pretrained model.

3. Dataset and Task Definition

We constructed a sperm video dataset for sperm assessment. When constructing the dataset, each of the 40 experts annotated one of the five grades for each sample. The grades are as follows: A (best); B (good); C (neither); D (bad); and E (worst). To replicate the actual variability, the experts graded them based on their knowledge and senses. Therefore, the label of the dataset is a soft-label, which is a 5-grade histogram. We refer to this soft-label as the *grade distribution*. The dataset includes 615 videos captured using a microscope. The videos are 175-frame clips at 15 frames per second with 1392 × 976-pixel crops.

We applied sperm detection and tracking to all videos to create inputs for the neural network because they were taken from a microscope, the target sperm was considerably small, and debris, such as air bubbles and other sperm, were reflected. This preprocessing is shown on the left side of Figure 1. We tagged the target sperm in the first frame of all videos, and tracked it by template matching to detect and track sperm. The tracked videos are 16-frame clips with 150 × 150-pixel crops. The dataset statistics is shown in Table 1, and a sample of the tracking videos and labels are shown in Figure 2. More samples, including original videos, tracking videos, and labels, are shown in Figure A.2.

We define the sperm assessment task in our dataset as the 5-point regression task, because information of the most selected grade as well as that of the other grades are important for decision support. Given an input video $\mathcal{V} \in \mathbb{R}^{H\times W\times 3\times T}$, a neural network predicts the *grade distribution* $\hat{Y} \in \mathbb{R}^5$.

train / test	Grade A	B	C	D	E	Total
492 / 123	45	194	356	9	11	615

Table 1. Statistics of the dataset. Each value of Grade is the number of the samples for which the most experts selected the grade.

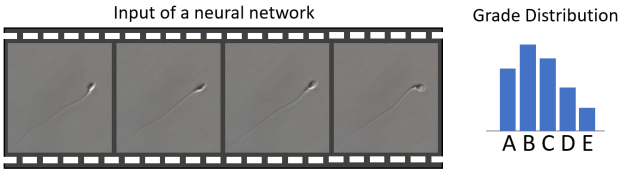

Figure 2. Sample tracking video and label of the dataset.

4. Method

4.1. Method Overview

Transformer models capture global dependencies through self-attention, and have achieved high performances in various vision tasks. However, self-attention treats each local patch uniformly to calculate the attention score, and then computes a weighted sum of all local patches. A global feature is dominated by all local patches, thus, simultaneously considering all local patches may reduce the influence of some important local patches. Particularly, in a sperm recognition task that require the capture of fine-grained shapes and motions, this method may cause serious discriminative deficiencies. Therefore, we propose *Patch Selection Module* (PSM) to select only important and informative patches.

Another challenge in sperm recognition is the extraction of diverse sperm features, such as the morphologies, motions and dependencies of various sperm parts. We expect to extract diverse features using global and local features effectively. We propose *Role-Separated Branch* (RSB) to effectively use local patches obtained by PSM and extract the spatial and temporal features separately.

We propose *Role-Separated Transformer for Fine-Grained and Diverse Sperm Feature Extraction* (RoSTFine), putting PSM and RSB on the head of TimeSformer. The details of RoSTFine are described below and its architecture is illustrated in Figure 3.

4.2. Video Encoder

TimeSformer [4] is used as the sperm video encoder. An input video $\boldsymbol{X} \in \mathbb{R}^{H\times W\times 3\times T}$ comprising T RGB frames of size $H \times W$ is decomposed into N non-overlapping patches for each frame, each of size $P \times P$. Each patch $\boldsymbol{x}_{(t,p)} \in \mathbb{R}^{3P^2}$ is linearly mapped into an embedding $\boldsymbol{z}_{(t,p)} \in \mathbb{R}^d$ using a learnable matrix $\boldsymbol{E} \in \mathbb{R}^{d\times 3P^2}$.

$$\boldsymbol{z}_{(t,p)} = \boldsymbol{E}\boldsymbol{x}_{(t,p)} + \boldsymbol{e}_{(t,p)} \tag{1}$$

where $\boldsymbol{e}_{(t,p)}$ denotes the learnable positional embeddings added to encode the spatiotemporal position of each patch. The input of TimeSformer is represented by Eq. 2.

$$\boldsymbol{Z} = \{\boldsymbol{z}_{cls}, \boldsymbol{z}_{(1,1)}, ..., \boldsymbol{z}_{(T,N)}\} \in \mathbb{R}^{(1+NT)\times d} \tag{2}$$

where $\boldsymbol{z}_{cls}$ is a learnable `[CLS]` embedding. The input $\boldsymbol{Z}$ is fed into TimeSformer blocks, which comprises a spatial attention, a temporal attention, and a multi-layer-perceptron.

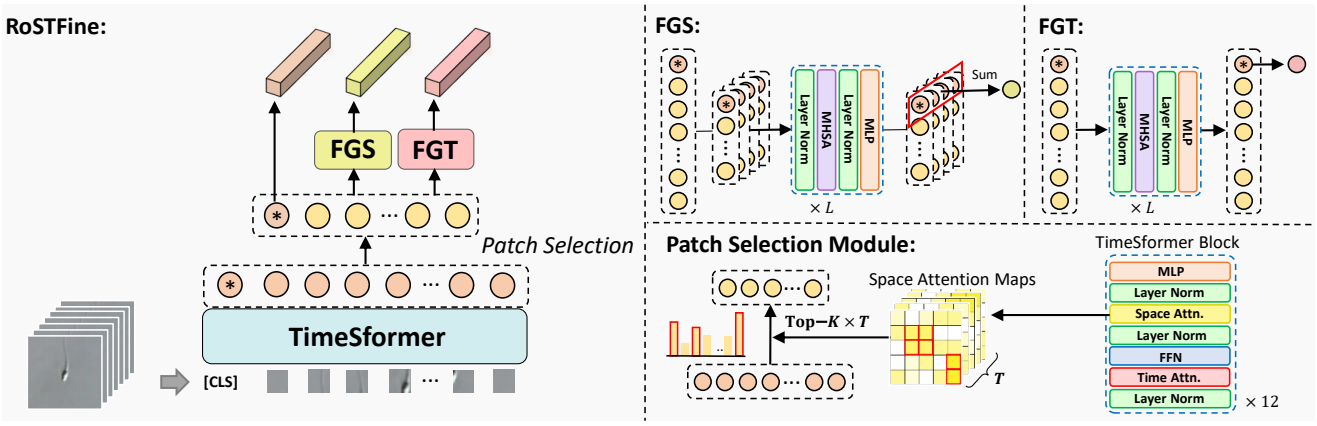

Figure 3. Architecture of RoSTFine. The video encoder of RoSTFine is TimeSformer. *Patch Selection Module* (PSM) selects important and informative local patches based on attention maps obtained from TimeSformer. *Fine-Grained Spatial Feature Extraction Branch* (FGS) applies multi-head self attention to each frame. *Fine-Grained Temporal Feature Extraction Branch* (FGT) applies multi-head self attention across all frames.

The output of TimeSformer is represented by Eq. 3.

$$\boldsymbol{V} = \{\boldsymbol{v}_{cls}, \boldsymbol{v}_{(1,1)}, ..., \boldsymbol{v}_{(T,N)}\} \in \mathbb{R}^{(1+NT)\times d} \tag{3}$$

4.3. Patch Selection Module

Transformer models solve various tasks using `[CLS]` embedding aggregated by all local patches. Using only the `[CLS]` embedding may cause serious discriminative deficiencies, especially in a sperm recognition task, whose data are background dominant and which requires some fine-grained parts. Certain critical local patches have the potential to capture fine-grained sperm features.

We propose *Patch Selection Module* (PSM) to select important local patches based on attention scores. Specifically, the spatial attention of each TimeSformer block generates an attention map $\boldsymbol{A}^l \in \mathbb{R}^{T\times(1+N)\times(1+N)}$, which represents the correlation with all patches. Here, l is the number of layers and ranges from 1 to L. To select patches, PSM uses the attention scores $\boldsymbol{A}^*$ calculated as the sum of the attention scores of `[CLS]` of the last two layers.

$$\boldsymbol{A}^* = \boldsymbol{A}^{L-1}[:,0,1:] + \boldsymbol{A}^{L}[:,0,1:] = \{\boldsymbol{A}^*_1, ..., \boldsymbol{A}^*_T\} \in \mathbb{R}^{T\times N\times N} \tag{4}$$

The top K patches are selected in N patches in each frame corresponding to the top K highest scores in $\boldsymbol{A}^*_i$. The selected patches are represented by Eq. 5.

$$\boldsymbol{F} = \{\boldsymbol{f}_{(1,1)}, \boldsymbol{f}_{(1,2)}, ..., \boldsymbol{f}_{(T,K)}\} \in \mathbb{R}^{TK\times d} \tag{5}$$

where $\boldsymbol{f}_{(t,k)}$ denotes the embeding of the k-patch in t-frame. The process of PSM is shown in the bottom right of Figure 3.

4.4. Role-Separated Branch

In a sperm recognition task, it is important to capture diverse sperm features, such as shapes, motions and their dependencies. However, `[CLS]` embedding may miss these features, because it is calculated using all local patches in the same manner and includes a large background. Therefore, we propose *Role-Separated Branch* (RSB) to effectively use the local patches and extract fine-grained spatial and temporal features separately and explicitly, using the local patches selected in PSM. RSB comprises *Fine-Grained Spatial Feature Extraction Branch* (FGS) and *Fine-Grained Temporal Feature Extraction Branch* (FGT).

Fine-Grained Spatial Feature Extraction Branch (FGS) obtains spatial sperm features, such as shape, texture, and dependencies within a frame. The inputs $\boldsymbol{G}^{s,(0)}$ of FGS are obtained by dividing $\boldsymbol{F}$ into frame units, and then attaching `[CLS]` embedding to each unit. The i-frame unit is represented by Eq. 6.

$$\boldsymbol{G}^{s,(0)}_i = \{\boldsymbol{v}_{cls}, \boldsymbol{g}^s_{(i,1)}, ..., \boldsymbol{g}^s_{(i,K)}\} \in \mathbb{R}^{(1+K)\times d} \tag{6}$$

Each $\boldsymbol{G}^{s,(0)}_i$ is fed into L Attention blocks, and then we obtain the output $\boldsymbol{G}^{s,(L)}_i$ (Eq. 7).

$$\boldsymbol{G}^{s,(l)}_i = \mathrm{MLP}(\mathrm{MHSA}(\mathrm{LN}(\boldsymbol{G}^{s,(l-1)}_i))) \tag{7}$$

where LN, MHSA and MLP denote Layer Normalization, Multi-Head Self-Attention, and Multi-Layer-Perceptron, respectively, and l represents the number of attention blocks. Finally, the fine-grained spatial feature $\boldsymbol{v}^s$ is obtained from the mean of all `[CLS]` embeddings $\boldsymbol{G}^{s,(L)}_i[0]$ $(i = 1, ..., F)$ of the last layer (Eq. 8).

$$\boldsymbol{v}^s = \frac{1}{KT}\sum_{i=1}^{F} \boldsymbol{G}^{s,(L)}_i[0] \tag{8}$$

The FGS process is shown in the top right of Figure 3.

Fine-Grained Temporal Feature Extraction Branch (FGT) obtains temporal features (*e.g.*, motions, movements,

and dependencies between frames). The input $\boldsymbol{G}^{t,(0)}$ of FGT is obtained by attaching `[CLS]` embedding $\boldsymbol{v}_{cls}$ to $\boldsymbol{F}$.

$$\boldsymbol{G}^{t,(0)} = \{\boldsymbol{v}_{cls}, \boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_{K\times T}\} \in \mathbb{R}^{(1+KT)\times d} \quad (9)$$

$\boldsymbol{G}^{t,(0)}$ is fed into L Attention blocks, and then we obtain the output $\boldsymbol{G}^{t,(L)}$ (Eq. 10). Finally, we obtain the `[CLS]` embedding as the fine-grained temporal feature $\boldsymbol{v}^t$ (Eq. 11).

$$\boldsymbol{G}^{t,(l)} = \text{MLP}(\text{LN}(\text{MHSA}(\text{LN}(\boldsymbol{G}^{t,(l-1)})))) \quad (10)$$

$$\boldsymbol{v}^t = \boldsymbol{G}^{t,(L)}[0] \quad (11)$$

where l denotes the number of the attention blocks. The FGT process is shown in the top right of Figure 3.

4.5. Training and Inference

Training. We regard the estimation of the *grade distribution* of the sperm as a 5-point regression task because the *grade distribution* is a 5-point histogram. We use Mean Squared Error (MSE) as the training objective, represented by Eq. 12.

$$\text{MSE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i)^2 \quad (12)$$

where n denotes the number of data points, and $n = 5$ in this case. The training objective $\mathcal{L}_{mse}$ is calculated as the mean of all MSE of $\hat{\boldsymbol{y}}^g$, $\hat{\boldsymbol{y}}^s$, and $\hat{\boldsymbol{y}}^t$ (Eq. 13).

$$\mathcal{L}_{mse} = \left(\text{MSE}(\hat{\boldsymbol{y}^g}, \boldsymbol{y}) + \text{MSE}(\hat{\boldsymbol{y}^s}, \boldsymbol{y}) + \text{MSE}(\hat{\boldsymbol{y}^t}, \boldsymbol{y})\right)/3 \quad (13)$$

where $\hat{\boldsymbol{y}}^g$, $\hat{\boldsymbol{y}}^s$, and $\hat{\boldsymbol{y}}^t$ are the predicted *grade distributions* obtained by linear projection of $\boldsymbol{v}^g$, $\boldsymbol{v}^s$, and $\boldsymbol{v}^t$, respectively.

However, the *grade distribution* estimation task can be regarded as a distribution distance minimizing problem. Therefore, we also use JS-divergence as the training objective. JS-divergence is represented by Eq. 15, which is a variant of KL-divergence, represented by Eq. 14, to smooth out the divergence and maintain symmetry.

$$\text{KL}(P, Q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) \quad (14)$$

$$\text{JS}(P, Q) = \frac{1}{2}\text{KL}\left(P, \frac{P+Q}{2}\right) + \frac{1}{2}\text{KL}\left(Q, \frac{P+Q}{2}\right) \quad (15)$$

The training objective $\mathcal{L}_{js}$ is calculated by the mean of all JS-divergence of $\hat{\boldsymbol{y}}^g$, $\hat{\boldsymbol{y}}^s$, and $\hat{\boldsymbol{y}}^t$ (Eq. 16).

$$\mathcal{L}_{js} = \left(\text{JS}(\hat{\boldsymbol{y}^g}, \boldsymbol{y}) + \text{JS}(\hat{\boldsymbol{y}^s}, \boldsymbol{y}) + \text{JS}(\hat{\boldsymbol{y}^t}, \boldsymbol{y})\right)/3 \quad (16)$$

Optional. In addition, we use the diversity loss as an optional training objective. $\boldsymbol{v}^g$, $\boldsymbol{v}^s$, and $\boldsymbol{v}^t$ may be similar vectors because their vectors are generated from the same `[CLS]` embedding. To diversify them, we apply the orthogonal constraint (Eq. 17) to any pairs of $\boldsymbol{v}^g$, $\boldsymbol{v}^s$ and $\boldsymbol{v}^t$, represented by Eq. 18. This constraint enables the cosine similarity between two vectors to be brought close to zero, and effectively facilitates them becoming different vectors.

$$\text{div}(\boldsymbol{v}^i, \boldsymbol{v}^j) = \left|\frac{\boldsymbol{v}^i \cdot \boldsymbol{v}^j}{||\boldsymbol{v}^i||_2||\boldsymbol{v}^j||_2}\right| \quad (17)$$

$$\mathcal{L}_{div} = \left(\text{div}(\boldsymbol{v}^g, \boldsymbol{v}^s) + \text{div}(\boldsymbol{v}^g, \boldsymbol{v}^t) + \text{div}(\boldsymbol{v}^s, \boldsymbol{v}^t)\right)/3 \quad (18)$$

We expect each vector to reclaim diverse features of sperm by the diversity loss $\mathcal{L}_{div}$. When applying the diversity loss, the training objective is represented as Eq. 19

$$\mathcal{L} = \mathcal{L}_{mse/js} + \alpha\mathcal{L}_{div} \quad (19)$$

where α is the weight of the diversity loss.

Inference. The final predicted *grade distribution* $\hat{y}$ is obtained from the mean of all predicted *grade distributions* $\hat{\boldsymbol{y}}^g$, $\hat{\boldsymbol{y}}^s$, and $\hat{\boldsymbol{y}}^t$ (Eq. 20).

$$\hat{\boldsymbol{y}} = (\hat{\boldsymbol{y}}^g + \hat{\boldsymbol{y}}^s + \hat{\boldsymbol{y}}^t)/3 \quad (20)$$

The architecture of RoSTFine is expected to have an effect similar to that of ensemble learning because $\hat{\boldsymbol{y}}^g$, $\hat{\boldsymbol{y}}^s$, and $\hat{\boldsymbol{y}}^t$ are optimized, respectively. $\hat{\boldsymbol{y}}^g$ is optimized based on the entire sperm, $\hat{\boldsymbol{y}}^s$ is optimized based on the sperm shape, and $\hat{\boldsymbol{y}}^t$ is optimized based on the sperm motion and dependencies across frames. We will show that this training and inference strategy is the best in §5.5.

5. Experiment

5.1. Experimental Setup

Dataset and Task. We use the dataset described in §3, and the task is 5-point histogram value regression.

Metrics. We use Mean Squared Error (MSE) and JS-divergence between predicted *grade distributions* and ground truth distributions as metrics. Additionally, we evaluate models on classification task by assigning a class (*e.g.*, the most selected grade) into a sperm. We use Balanced Accuracy as a metric due to the long-tail distributed dataset (*cf.* §3). Specifically, we measure Balanced Accuracies on the classification task to predict the n-th most selected grade. The worst or the second most selected grades as well as the most selected one are important for decision support.

Implementation Details. We conduct experiments on NVIDIA A5000 24GB Single GPU using the Pytorch library [29]. We download the pretrained weight of TimeSformer from `https://github.com/facebookresearch/TimeSformer`. In the training process, we optimize the models with Stochastic Gradient Descent (SGD) optimizer with learning rate of 1e-3, momentum of 0.9, and weight decay of 5e-4. Each model is trained with a batch size of 8 and lasted for 200 or 300 epochs. We train and evaluate the models on five-fold cross-validation. The default settings for RoSTFine are $K = 60, L = 6$, and $H = 8$, where K is the number of

Method	MSE$^{(10^{-2})}$	JS div.$^{(10^{-2})}$	Balanced Accuracy (%) 1st	2nd	3rd	4th	5th	Avg.
VGG16†	1.517 $\pm$ 0.09	5.628 $\pm$ 0.27	26.49 $\pm$ 3.40	22.72 $\pm$ 4.44	22.31 $\pm$ 5.11	22.25 $\pm$ 4.12	33.44 $\pm$ 6.59	25.44
R3D	1.365 $\pm$ 0.10	4.978 $\pm$ 0.38	28.72 $\pm$ 4.16	31.03 $\pm$ 2.50	29.70 $\pm$ 4.30	28.46 $\pm$ 6.35	34.68 $\pm$ 3.29	30.52
R(2+1)D	1.702 $\pm$ 0.09	7.043 $\pm$ 0.28	29.83 $\pm$ 3.42	21.83 $\pm$ 1.49	20.26 $\pm$ 0.51	22.64 $\pm$ 2.78	26.96 $\pm$ 5.50	24.30
X3D	1.808 $\pm$ 0.07	7.186 $\pm$ 0.23	27.65 $\pm$ 4.33	20.46 $\pm$ 0.99	22.33 $\pm$ 1.67	22.74 $\pm$ 2.45	32.12 $\pm$ 7.03	25.06
I3D	1.376 $\pm$ 0.13	5.206 $\pm$ 0.40	30.28 $\pm$ 4.77	31.39 $\pm$ 1.74	29.16 $\pm$ 1.67	27.44 $\pm$ 4.07	36.65 $\pm$ 10.20	30.98
SlowFast	1.346 $\pm$ 0.13	5.059 $\pm$ 0.54	31.58 $\pm$ 5.36	30.00 $\pm$ 2.41	29.15 $\pm$ 4.46	24.85 $\pm$ 5.06	34.65 $\pm$ 8.03	30.05
ViViT	1.406 $\pm$ 0.09	4.987 $\pm$ 0.39	27.37 $\pm$ 0.73	28.34 $\pm$ 3.28	28.26 $\pm$ 4.65	**31.46** $\pm$ 4.55	39.50 $\pm$ 5.70	30.99
TimeSformer	1.186 $\pm$ 0.10	4.283 $\pm$ 0.17	31.43 $\pm$ 1.32	35.62 $\pm$ 3.72	32.49 $\pm$ 4.05	28.47 $\pm$ 5.19	41.68 $\pm$ 5.46	33.94
RoSTFine$_{\alpha=0}$	**1.121** $\pm$ 0.11	**4.145** $\pm$ 0.26	**33.48** $\pm$ 6.97	35.56 $\pm$ 2.00	**33.40** $\pm$ 3.15	30.17 $\pm$ 5.59	**43.59** $\pm$ 6.39	**35.24**
RoSTFine$_{\alpha=1}$	**1.104** $\pm$ 0.12	**4.109** $\pm$ 0.26	**35.61** $\pm$ 4.19	**36.71** $\pm$ 3.55	**35.97** $\pm$ 5.51	30.29 $\pm$ 6.10	**43.29** $\pm$ 8.58	**36.37**

Table 2. Comparing RoSTFine to the baselines. The value is average $\pm$ standard deviation of 5 folds. RoSTFine$_{\alpha=0,1}$ achieves the best performance in MSE, JS-divergence (JS div.) and most Balanced Accuracies (BA). † denotes the lower bound, which is trained on the first frame image. RoSTFine$_{\alpha=0,1}$ have statistical significance ($p < .05$, *Kruskal-Wallis test*) among the other models except TimeSformer on MSE and JS-divergence. Overall, RoSTFine$_{\alpha=1}$ obtains 0.082×10^{-2}, 0.174×10^{-2}, 2.43% higher on MSE, JS-div. and average of BA.

patches picked out in PSM, L is the number of FGS and FGT branch attention blocks, and H is the number of heads of the multi-head attention.

5.2. Comparison with Baselines

We compare the performances of our RoSTFine$_{\alpha=0,1}$ with those of the baselines to confirm its superiority for sperm videos. Where α denotes a weight of the diversity loss (*cf.* Eq.19). The baselines are R3D [15], R(2+1)D [15], X3D [11], I3D [5], SlowFast [12], ViViT [3], and TimeSformer [4]. The models[1] used in this experiment are pretrained on Kinetics-400 [5]. Furthermore, we consider VGG16 [34], pretrained on ImageNet [9], as the lower bound, whose inputs are the first frame image.

First, in Table 2, RoSTFine$_{\alpha=0}$ outperforms the baselines on MSE, JS-divergence (JS) and most Balanced Accuracy (BA), and has statistical significances ($p < .05$, *Kruskal-Wallis test*) among the other models except TimeSformer on MSE and JS. Although we cannot obtain statistical significance between RoSTFine$_{\alpha=0}$ and TimeSformer, RoSTFine$_{\alpha=0}$ outperform TimeSformer in four out of five folds in MSE and JS (*cf.*§A.3). Therefore, our RoSTFine model are better than TimeSformer for the sperm videos. Second, RoSTFine$_{\alpha=1}$ outperforms the baselines and RoSTFine$_{\alpha=0}$, and obtains the best results on MSE, JS and most BA, which indicates that the diversity loss is effective. Then, although R(2+1)D and X3D are trained on videos, and VGG16 is trained on only one frame image, the performances of R(2+1)D and X3D are worse than those of VGG16 in Table 2. One possible reason for this is that sperm motion factor is negative for sperm recognition when the model cannot capture sperm motion accurately. This suggests that caution should be exercised when designing sperm recognition models.

[1]We downloaded the pretrained weights of R3D, R(2+1)D, X3D, I3D and SlowFast from `https://github.com/facebookresearch/pytorchvideo` and that of ViViT from `https://github.com/mx-mark/VideoTransformer-pytorch`.

5.3. Visualizing Attention Map

In this section, we analyze the space attention visualizations. Figure 4 presents the space attention maps of 1st, 3rd, 5th, 7th, 9th, 11th, 13th, and 15th frames obtained by TimeSformer and RoSTFine($K = 5$). To visualize attention maps, we use the Attention Rollout scheme [1].

In Figure 4, we observe that TimeSformer focuses on a wide area around the sperm and strongly on the sperm head, whereas RoSTFine focuses hardly on background and only on the sperm, such as the head and neck. Moreover, we observe that RoSTFine captures the sperm head and neck as well as the sperm tail. These results suggest two good points of RoSTFine. First, PSM can reduce redundancy. Second, RoSTFine can focus on critical parts of the sperm, because the sperm head contains deoxyribonucleic acid, which carries the genetic instructions necessary for reproduction, and the neck of the sperm contains mitochondria, which supply the energy necessary for movement to the tail [28]. Furthermore, The tip of the sperm tail, captured across most frames, and the middle of the sperm tail, captured when the tail moves vigorously, are necessary for motility.

5.4. Effectiveness of Diversity Loss

To confirm the effectiveness of the diversity loss (§4.5), we compare the task performance of RoSTFine in the range of $\alpha = \{0, 0.01, 0.05, 0.1, 0.5, 1.0, 1.5\}$.

We observe that RoSTFine$_{\alpha=1.0}$ is 0.017 point higher than RoSTFine$_{\alpha=0}$ in Table 5 (left). Furthermore, we observe that the higher the α value, the higher the performance in the range of $0 < \alpha < 1$. The performance is higher than that of the baseline ($\alpha = 0$) even when the α value is the

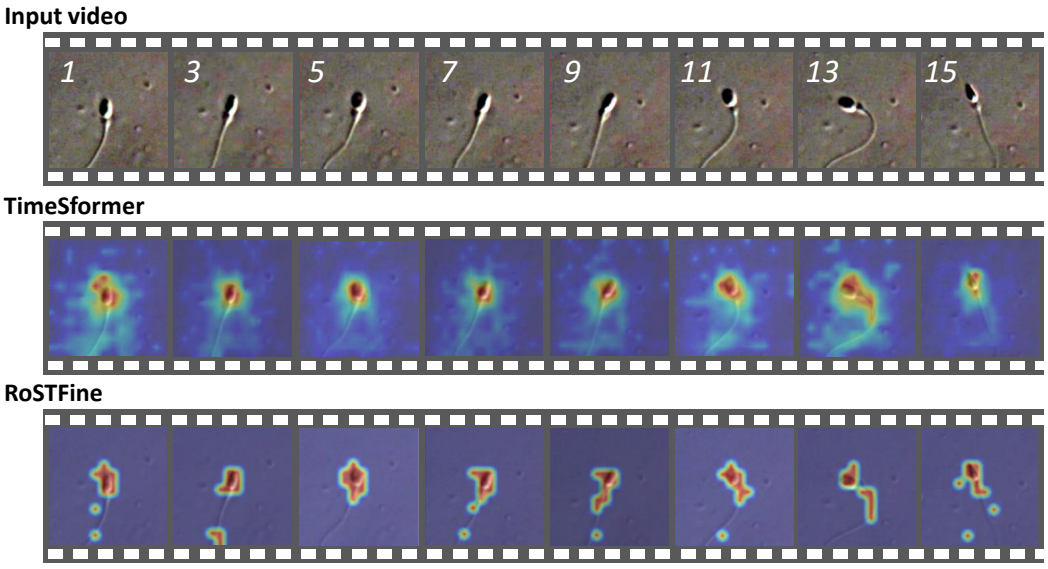

Figure 4. Attention map visualizations of odd number frames of TimeSformer and RoSTFine($K = 5$). While TimeSformer attend to a wide area around the sperm, RoSTFine attend strongly to only sperm. RoSTFine can capture particularly the sperm head and neck, and can capture the tip of the tail across frames and the middle of the tail when the tail moves vigorously.

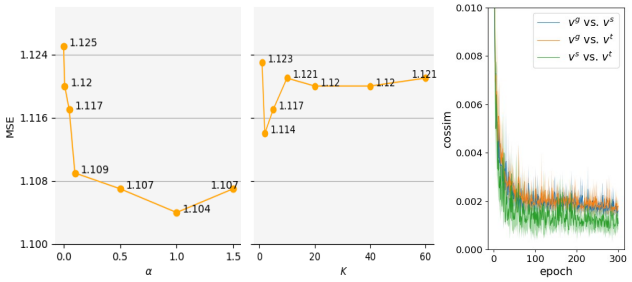

Figure 5. The transition of the MSE loss with the degree α of $\mathcal{L}_{div}$ (left), that with the number K of selected patches (center), and cosine similarity of any features combinations in each epoch (right).

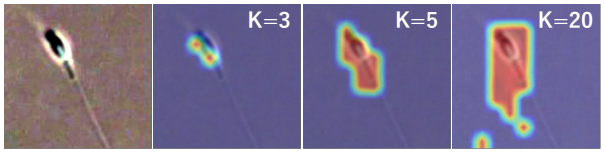

Figure 6. Attention map visualizations of a frame using space attentions of RoSTFine in $K = 3, 5, 20$. We observe that more patches contain background in $K = 20$ than that in $K = 3, 5$.

smallest ($\alpha = 0.01$). These results indicate that the diversity loss is effective. Then, Figure 5 (right) shows the transition of the cosine similarities of any features combinations. We confirm that all cosine similarities are brought close to zero. We consider that the diversity loss makes $\boldsymbol{v}^g$, $\boldsymbol{v}^s$, and $\boldsymbol{v}^t$ different from each other to obtain diverse sperm features, which facilitate sperm representation learning.

5.5. Additional Ablations

In this section, we investigate RoSTFine in detail by conducting additional experiments to answer the four Research Questions.

RQ1: How many tokens should be selected in PSM?
We are interested in the best number of patches to be selected. We hypothesize that if the value of K is considerably large, RoSTFine cannot extract the fine-grained features, and if it is significantly small, RoSTFine may miss some important factors, both of which decreases performances. To verify this hypothesis, we compare the task performances in the range of $K = \{1, 2, 5, 10, 20, 40, 60\}$ and attention maps of RoSTFine in $K = 3, 5, 20$.

Figure 5 and 6 show the task performances and attention maps, respectively. In Figure 5 (center), we observe that the performances remain almost the same in range from $K = 10$ to 60, the best is in $K = 3$, and the worst is in $K = 1$. In Figure 6, we observe that RoSTFine captures more background in $K = 20$ than that in $K = 3, 5$. These results indicate that the best number of K is 3. When $K = 3, 5$, selected patches have low redundancy, and RoSTFine focuses almost only on sperm and extracts fine-grained features. When $10 \leq K \leq 60$, selected patches are redundant, contain background, and RoSTFine cannot extract the fine-grained features. When $K = 1$, selected patches are so few that they contains insufficient information.

RQ2: Are the features $\boldsymbol{v}^s$ and $\boldsymbol{v}^t$ really effective?
We confirm the effectiveness of the features $\boldsymbol{v}^s$ and $\boldsymbol{v}^t$ generated by FGS and FGT, respectively, by comparing their

performances in any combinations of $\boldsymbol{v}^g$, $\boldsymbol{v}^s$, and $\boldsymbol{v}^t$. We experiment for both training and inference using the same combinations. When using only $\boldsymbol{v}^g$, the model is the same as TimeSformer.

In Table 3, we observe that the performances of only $\boldsymbol{v}^s$ and only $\boldsymbol{v}^t$ is higher than that of only $\boldsymbol{v}^g$, which suggests that $\boldsymbol{v}^s$ and $\boldsymbol{v}^t$ generated by FGS and FGT are effective. Moreover, we observe that combinations in any other features perform better, and the performance of the combination of all features is the best. This result demonstrates the effectiveness of the feature combination.

$\boldsymbol{v}^g$	$\boldsymbol{v}^s$	$\boldsymbol{v}^t$	MSE$^{(10^{-2})}$
✓†			1.186 $\pm$ 0.10
	✓		1.138 $\pm$ 0.10
		✓	1.145 $\pm$ 0.07
✓	✓		1.135 $\pm$ 0.11
✓		✓	1.139 $\pm$ 0.11
	✓	✓	1.134 $\pm$ 0.10
✓	✓	✓	**1.121** $\pm$ 0.11

Table 3. Performances of any combinations of $\boldsymbol{v}^g$, $\boldsymbol{v}^s$ and $\boldsymbol{v}^t$. † denotes the same model as TimeSformer. The performance when using all features is the best.

RQ3: How should $\boldsymbol{v}^g$, $\boldsymbol{v}^s$ and $\boldsymbol{v}^t$ be aggregated?
There are some way to aggregate $\boldsymbol{v}^g$, $\boldsymbol{v}^s$ and $\boldsymbol{v}^t$. We test the following aggregation strategies: (1) concatenating the features (Concat); (2) Summing the features (Sum); (3) Calculating losses separately in training, and the mean of the outputs $\{\boldsymbol{y}^g, \boldsymbol{y}^s, \boldsymbol{y}^t\}$ in inference (Ours).

The results in Table 4 show that our strategy achieves the best performance, which can be attributed to the method of calculating losses separately for multiple features generated on different architectures, thereby enabling the use of more diverse features, such as ensemble learning.

Aggregation type	MSE$^{(10^{-2})}$	JS divergence$^{(10^{-2})}$
Concat	1.228 $\pm$0.01	4.238 $\pm$0.15
Sum	1.197 $\pm$ 0.10	4.153 $\pm$0.24
Ours	**1.121** $\pm$ 0.11	**4.145** $\pm$0.26

Table 4. Performances in three aggregation strategies. (1) concatenating the features (Concat); (2) Summing the features (Sum); (3) Calculating losses separately in training and the mean of the outputs $\boldsymbol{y}^g$, $\boldsymbol{y}^s$ and $\boldsymbol{y}^t$ in inference (Ours). Ours achieves the best performance.

RQ4: What impact do differences in pretraining data have on sperm recognition?
Pretraining data are an important factor in designing a model because they critically affect the nature of the model. We investigate the best pretraining data for a sperm recognition task using TimeSformer pretrained on Kinetics [5], Something-Something-V2 (SSv2) [14], and HowTo100M (HT100M) [23], publicly available at `https://github.com/facebookresearch/TimeSformer`. While spatial cues are more important than temporal information in achieving high performance in Kinetics and HT100M, spatial and temporal information are important in SSv2.

In Figure 7, both TimeSformer and RoSTFine achieve the lowest loss in SSv2, which suggests that both spatial and temporal sperm characteristics are important. Therefore, a model pretrained on a dataset that requires capturing temporal information is suitable for sperm recognition.

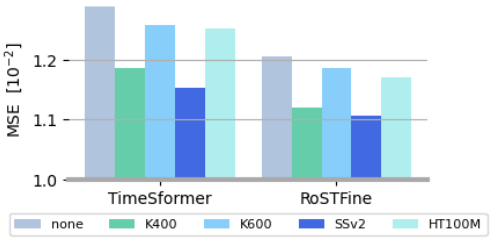

Figure 7. MSE loss of each pretraining dataset. In both of TimeSformer and RoSTFine, SSv2 achieves the best performance.

6. Conclusion

To assist clinicians to assess sperm and select optimal sperm, in this study, we constructed a sperm video dataset annotated with soft-labels and proposed an automated framework and a neural network, RoSTFine, for sperm assessment. In designing the network, to extract fine-grained and diverse sperm features, Patch Selection Module (PSM) and Role-Separated Branch (RSB) are placed on the head of TimeSformer. PSM filters patches to obtain features that focus on fine-grained sperm characteristics. RSB can obtain spatial and temporal fine-grained sperm features. Our experimental sesults showed the superiority of RoSTFine and the effectiveness of PSM and RSB. We addressed reproduction, an important medical issue in human life but little-studied in computer vision fields, and our study has the potential to make a contribution to human well-being.

Limitations. (1) We used comprehensive models, in particular publicly available pretrained models, (§5.2) as far as we know, but might have overlooked or updated the other state-of-the-art models. (2) We cannot confirm the model's robustness to sample preparation and microscope settings, because all samples of our dataset were taken in the same ways and settings. We will develop our work by collecting more samples in various ways and settings.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4190–4197, 2020.

[2] Rupert P. Amann and Dagmar Waberski. Computer-assisted sperm analysis (casa): Capabilities and potential developments. *Theriogenology*, 2014.

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 813–824, 2021.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] Feng-Ju Chang, Martin Radfar, Athanasios Mouchtaris, Brian King, and Siegfried Kunzmann. End-to-end multichannel transformer for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888, 2021.

[7] Violeta Chang, Alejandra Garcia, Nancy Hitschfeld, and Steffen Härtel. Gold-standard for computer-assisted morphological sperm analysis. *Computers in Biology and Medicine*, pages 143–150, 2017.

[8] Russell O. Davis, David E. Bain, Rebecca J. Siemers, David M. Thal, Jane B. Andrew, and Curtis G. Gravance. Accuracy and precision of the ceiiform-human* automated sperm morphometry instrumentt†‡. *Fertility and Sterility*, pages 763–769, 1992.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019.

[11] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[13] N Gatimel, J Moreau, J Parinaud, and RD Léandri. Sperm morphology: assessment, pathophysiology, clinical relevance, and state of the art in 2017. *Andrology*, pages 845–862, 2017.

[14] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5842–5850, 2017.

[15] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017.

[16] C.A. Holden, R.I. McLachlan, R. Cumming, G. Wittert, D.J. Handelsman, D.M. de Kretser, and M. Pitts. Sexual activity, fertility and contraceptive use in middle-aged and older men: Men in Australia, Telephone Survey (MATeS). *Human Reproduction*, pages 3429–3434, 2005.

[17] Hamza Ilhan, İbrahim Sığırcı, Gorkem Serbes, and Nizamettin Aydin. A fully automated hybrid human sperm detection and classification system based on mobile-net and the performance comparison with conventional methods. *Medical & Biological Engineering & Computing*, 2020.

[18] Hamza Osman Ilhan, Gorkem Serbes, and Nizamettin Aydin. Automated sperm morphology analysis approach using a directional masking technique. *Computers in Biology and Medicine*, 2020.

[19] Aldo Isidori, Maurizio Latini, and Francesco Romanelli. Treatment of male infertility. *Contraception*, pages 314–318, 2005.

[20] Neeraj Kumar, Kuljeet Kaur, Anish Jindal, and Joel J.P.C. Rodrigues. Providing healthcare services on-the-fly using multi-player cooperation game theory in internet of vehicles (iov) environment. *Digit. Commun. Networks*, pages 191–203, 2015.

[21] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[22] R Menkveld, FS Stander, TJ Kotze, TF Kruger, and JA van Zyl. The evaluation of morphological characteristics of human spermatozoa according to stricter criteria. *Human reproduction (Oxford, England)*, pages 586–592, 1990.

[23] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[24] Simcha K Mirsky, Itay Barnea, Mattan Levi, Hayit Greenspan, and Natan T Shaked. Automated analysis of individual sperm cells using stain-free interferometric phase microscopy and machine learning. *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, pages 893–900, 2017.

[25] J Neuwinger, HM Behre, and E Nieschlag. Computerized semen analysis with sperm tail detection. *Human reproduction (Oxford, England)*, pages 719–723, 1990.

[26] Alberto Nogales, Álvaro J García-Tejedor, Diana Monge, Juan Serrano Vara, and Cristina Antón. A survey of deep learning models in medical therapeutic areas. *Artificial intelligence in medicine*, 2021.

[27] Reza Nosrati, Percival J Graham, Biao Zhang, Jason Riordon, Alexander Lagunov, Thomas G Hannam, Carlos Escobedo, Keith Jarvi, and David Sinton. Microfluidics for sperm analysis and selection. *Nature reviews. Urology*, pages 707–730, 2017.

[28] World Health Organization. *WHO laboratory manual for the examination and processing of human semen*. World Health Organization, 2021.

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[30] Xingyu Pei, Shuntao Huang, Jiangjing Cui, Wei Qiu, Danbing Liu, and Anbo Meng. An improved convolutional neural network used in abnormality identification of indicating lighting in cable tunnels. In *International Conference on Power System Technology (POWERCON)*, pages 4550–4554, 2018.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[32] Jason Riordon, Christopher McCallum, and David Sinton. Deep learning for the classification of human sperm. *Computers in Biology and Medicine*, 2019.

[33] Fariba Shaker, S. Amirhassan Monadjemi, Javad Alirezaie, and Ahmad Reza Naghsh-Nilchi. A dictionary learning approach for human sperm heads classification. *Computers in Biology and Medicine*, pages 181–190, 2017.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Clinical Orthopaedics and Related Research (CoRR)*, 2014.

[35] Lindsay Spencer, Jared E. M. Fernando, Farzan Akbaridoust, Klaus Ackermann, and Reza Nosrati. Ensembled deep learning for the classification of human sperm head morphology. *Advanced Intelligent Systems*, 2022.

[36] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision? *Clinical Orthopaedics and Related Research (CoRR)*, 2021.

[37] Yihe Wang, Jason Riordon, Tian Kong, Yi Xu, Brian Nguyen, Junjie Zhong, Jae Bem You, Alexander Lagunov, Thomas G Hannam, Keith Jarvi, and David Sinton. Prediction of dna integrity from morphological parameters using a single-sperm dna fragmentation index assay. *Advanced science (Weinheim, Baden-Wurttemberg, Germany)*, 2019.

[38] Daniel Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, pages 8619 – 8624, 2014.

[39] Mecit Yüzkat, Hamza Osman Ilhan, and Nizamettin Aydin. Multi-model cnn fusion for sperm morphology analysis. *Computers in biology and medicine*, page 104790, 2021.

[40] Jae Bem You, Christopher McCallum, Yihe Wang, Jason Riordon, Reza Nosrati, and David Sinton. Machine learning for sperm selection. *Nature reviews. Urology*, pages 387–403, 2021.