electronics

MDPI

Article

Improving Generation and Evaluation of Long Image Sequences for Embryo Development Prediction

Pedro Celard [1,2,3], Adrián Seara Vieira [1,2,3], José Manuel Sorribes-Fdez [1,2,3], Eva Lorenzo Iglesias [1,2,3] and Lourdes Borrajo [1,2,3,*]

[1] Department of Computer Science, ESEI-Escuela Superior de Ingeniería Informática, Universidade de Vigo, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain; pedro.celard.perez@uvigo.gal (P.C.); adrseara@uvigo.gal (A.S.V.); sorribes@uvigo.gal (J.M.S.-F.); eva@uvigo.gal (E.L.I.)
[2] CINBIO, Nanomaterials and Biomedical Research Centre, Universidade de Vigo, Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain
[3] SING, Next Generation Computer Systems Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, 36312 Vigo, Spain
* Correspondence: lborrajo@uvigo.es

Abstract: Generating synthetic time series data, such as videos, presents a formidable challenge as complexity increases when it is necessary to maintain a specific distribution of shown stages. One such case is embryonic development, where prediction and categorization are crucial for anticipating future outcomes. To address this challenge, we propose a Siamese architecture based on diffusion models to generate predictive long-duration embryonic development videos and an evaluation method to select the most realistic video in a non-supervised manner. We validated this model using standard metrics, such as Fréchet inception distance (FID), Fréchet video distance (FVD), structural similarity (SSIM), peak signal-to-noise ratio (PSNR), and mean squared error (MSE). The proposed model generates videos of up to 197 frames with a size of 128 × 128, considering real input images. Regarding the quality of the videos, all results showed improvements over the default model (FID = 129.18, FVD = 802.46, SSIM = 0.39, PSNR = 28.63, and MSE = 97.46). On the coherence of the stages, a global stage mean squared error of 9.00 was achieved versus the results of 13.31 and 59.3 for the default methods. The proposed technique produces more accurate videos and successfully removes cases that display sudden movements or changes.

Keywords: video generation; embryonic development; long-duration videos; diffusion models; medical imaging; generative models

Citation: Celard, P.; Seara Vieira, A.; Sorribes-Fdez, J.M.; Iglesias, E.L.; Borrajo, L. Improving Generation and Evaluation of Long Image Sequences for Embryo Development Prediction. *Electronics* **2024**, *13*, 476. https://doi.org/10.3390/electronics13030476

Academic Editors: Luca Mesin, Antonio Lanata, Irena Galić and Marija Habijan

Received: 26 December 2023
Revised: 20 January 2024
Accepted: 22 January 2024
Published: 23 January 2024

1. Introduction

The use of artificial intelligence (AI) in healthcare has led to profound advancements that have significantly enhanced medical practices [1]. In particular, the combination of AI and medical imaging analysis has been a significant breakthrough [2]. This integration has transformed the approach of medical professionals to extracting pertinent information from intricate visual data [3] and automated actions that were formerly susceptible to subjectivity and variability [4,5].

Deep learning (DL) techniques have shown great potential in achieving accurate diagnoses, selecting personalized treatments, and improving prognosis [6,7]. DL models have the ability to capture non-linear relationships in medical data and perform tasks such as classifying images [8], localizing and detecting objects [9–11], segmenting data [12–14], and generating synthetic data [15,16]. One of the primary objectives is to personalize these tasks for each individual patient, which can lead to better outcomes and an improved quality of life. Deep learning models for image processing are superior in identifying and classifying objects. However, many studies have failed to account for the importance of

time-dependent variables, resulting in the disregard of changes experienced by the object of study over time. These temporal variables carry significant weight and comprehending how to utilize these correlations is extremely relevant [17].

In order to progress in this domain, it is vital to gather standardized and interoperable time-dependent data, validate and generalize robust models, and examine ethical aspects concerning privacy and bias. The scarcity of this kind of data frequently prohibits the implementation and training of DL models, prompting researchers to focus on data augmentation [18–20] and the advantages of using synthetic data for these tasks [21]. Due to privacy concerns, data collection is commonly anonymized, resulting in loss of traceability. This poses a considerable challenge when conducting studies on evolution over time through images, unless highly specific studies can maintain control over data acquisition and processing over prolonged periods. Many works are now moving toward the use of image sequences, inspired by the success of using DL models in medicine for the treatment and analysis of static images. These sequences can either be located in space [22,23] or within a temporal context as videos [24–26]. Among the most common tasks when working with image sequences is the prediction of frames and filling in gaps between two existing images [27].

This work focuses on the medical field of fertilized embryos for implantation in in vitro fertilization (IVF) procedures. Monitoring development involves acquiring images every few minutes. The process entails controlled ovarian hyperstimulation, egg collection, fertilization, and the cultivation of embryos, which occurs over 2–6 days in environments with regulated conditions. This results in the intrauterine transfer or freezing of embryos that embryologists have identified as having strong implantation potential. The purpose of this monitoring is to determine their viability and the optimal time for extraction and implantation [28,29]. Currently, efforts are underway to identify and recognize cases where embryos can or cannot be successfully implanted after insemination. This task is primarily carried out through the scrutiny and judgment of human experts, whose opinions may not always align [30]. Consequently, the scientific community aims to create classification systems [31] and examine developmental patterns that enable more precise predictions regarding the future development of embryos and their clinical outcomes [32]. The lack of data has impeded the successful use of machine learning techniques in such research; however, there have been significant advances in this field, particularly in forecasting [33] and categorization tasks [34].

Despite the success of generative models applied to biomedical images [35] and in vitro fertilization procedures [36], few studies have exploited their potential to generate images of new synthetic cases of embryos. Dirvanauskas et al. [37] used a model based on a generative adversarial network (GAN) to generate new images of individual embryos. The authors used the stages of the embryos to condition the model, thus generating images of embryos with one, two, or four cells. The results showed that their model can generate realistic individual images while controlling the number and size of cells. In another previous work, Celard et al. [38] explored the generation of time-evolving embryo videos using GAN and diffusion models. They concluded that both models are capable of generating short-duration, unconditioned sequences. Diffusion produces videos of better morphological quality for embryonic development. The GAN-based model generates sequences of stages that are more realistic, but the resulting images of embryos may exhibit deformities or other unrealistic characteristics. On the other hand, the diffusion-based model produces good results for both the sequence of stages and the quality of the individual images.

This work focuses on exploring the capability of a deep learning model for the predictive generation of time-developing videos. Specifically, it focuses on the generation of long-duration videos of embryonic development, but this architecture could be applicable to any situation in which temporal development is important for prediction, detection, or tracking tasks where stages are an essential aspect of the videos. Despite significant advances in synthetic data generation, there is a lack of evaluation measures that are capable of adapting to specific cases to assess the quality of videos [39], so evaluation often relies

on human experts. Therefore, in this study, we propose a method to generate better fake embryo videos, considering the distributions of the represented stages and the relationships between contiguous frames for the autonomous selection of the best video.

The proposed architecture uses two Siamese models to generate long-duration videos. The structure of both models is indistinguishable, but the training process and dataset preprocessing procedure are different. One model specializes in learning short video distributions, enabling the observation of the entire development of the embryos. The other model specializes in filling in new frames for the overview sequence. This Siamese architecture approach has demonstrated its efficacy in previous studies, such as undersampled magnetic resonance imaging reconstruction [40] and automatic eye disease classification [41]. Our main contributions are as follows:

1. We propose a Siamese architecture for the video generation of embryo development based on a video diffusion model, which prevents the loss of quality in synthetic frames when generating long-duration videos;
2. A novel fake video evaluation process is proposed, fully utilizing the prior knowledge of original data and its stage distribution, thus improving resemblance to real data development;
3. The automation of sequence selection is achieved based on the realism of the development of its content, image quality, and movement between neighboring frames, thereby reducing the need for human expert judgment.

The remainder of this paper is organized as follows: in Section 2, we provide the necessary background on the employed techniques of the proposed method and outline relevant related work; Section 3 describes the performed experiments and evaluation metrics; in Section 4, we present the results and discuss them from quantitative and qualitative points of view to analyze the numeric metrics and generated image sequences; finally, we conclude this work in Section 5.

2. Materials and Methods

Models based on generative adversarial networks (GANs) [42–44] and variational autoencoders (VAEs) [45,46] have achieved excellent results in synthetic image generation. Nonetheless, there have recently been significant advancements in static image generation [47,48] and video generation [27,49–51] using diffusion models, even exceeding the performance of GAN and VAE models [27,52].

2.1. Diffusion Models

In the field of image generation, diffusion models are based on the transformation of images into unstructured noise and then reversing this operation to obtain the original images [38]. This means that the input x (real image) is transformed using a Gaussian noise addition process, guided by the hyperparameter λ_t. The result of this transformation, represented as z_t, is both the output of the Gaussian process and the input to a convolutional network that attempts to learn how to remove the noise from z in order to obtain an estimate of the original image [49,53,54]. This reversal involves a convolutional neural network that uses a U-Net architecture. The conventional 2D U-Net [55] has been extensively applied to tasks involving image-to-image generation. It entails a sequence of downsampling convolutional blocks, followed by a proportionate number of upsampling blocks. The outcome of each downsampling block is merged with the corresponding upsampling block via a skip connection. Figure 1 shows the simplified architecture of a diffusion Model for image generation.

However, creating images independently is inadequate for video production since 2D operations are not able to connect the content of one frame to adjacent frames. As a result, 3D convolutional operations are integrated into conventional U-Net models for video generation.

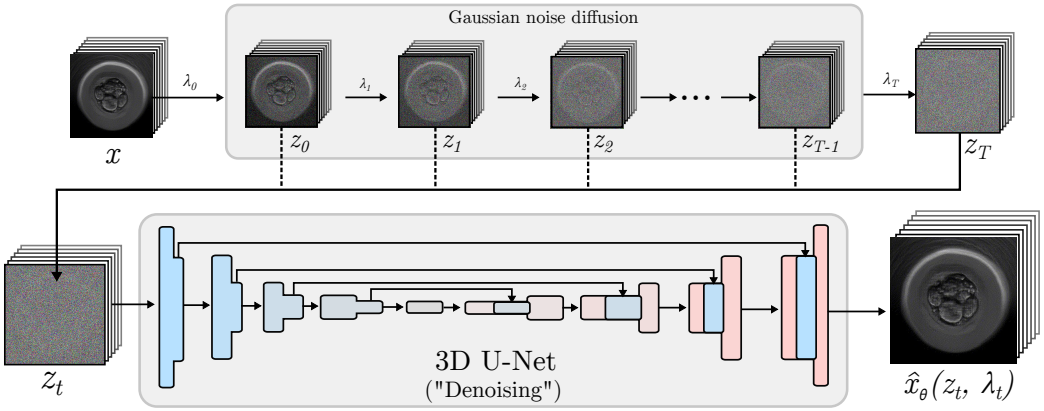

Figure 1. Simplified diffusion model architecture.

The aim of this study was to produce sequences of embryonic growth from real images that corresponded to various developmental phases. Although GAN and VAE models have yielded positive results, diffusion models were selected for this study. Additionally, as explained by Höppe et al. [27] in their proposal of the random-mask video diffusion (RaMViD) model, diffusion models are capable of predicting future frames and generating intermediate frames by taking into account real images at arbitrary positions. This eliminates the need for conditional frames to always be the first frame or always be in the same position.

2.2. Long Video Generation

As claimed by Höppe et al. [27], the RaMViD model can generate long videos iteratively using the last frame as a conditioning factor for the following section. In their work, they conducted tests on the UCF-101 and Kinetics-600 datasets, with videos of 16 and 20 frames; however, as the number of frames increases, the video quality degrades significantly. This problem is exacerbated when videos exceed 30 frames. Figure 2 shows a selection of equally distributed frames extracted from long videos generated by the base model. The figure displays a set of generated fake frames that create long-duration videos produced by the default model. The position of each frame (F) within the videos is indicated at the bottom. All images have the same resolution (128×128) and are a direct output of the default model. It can be observed that the initial frames F_0 exhibit good quality and realism, while by F_{26}, noise starts to appear, making visualization challenging. The loss of quality from F_{52} to F_{182} is significant, to the point that some images may not contain any discernible objects and instead only show blotches and noise.

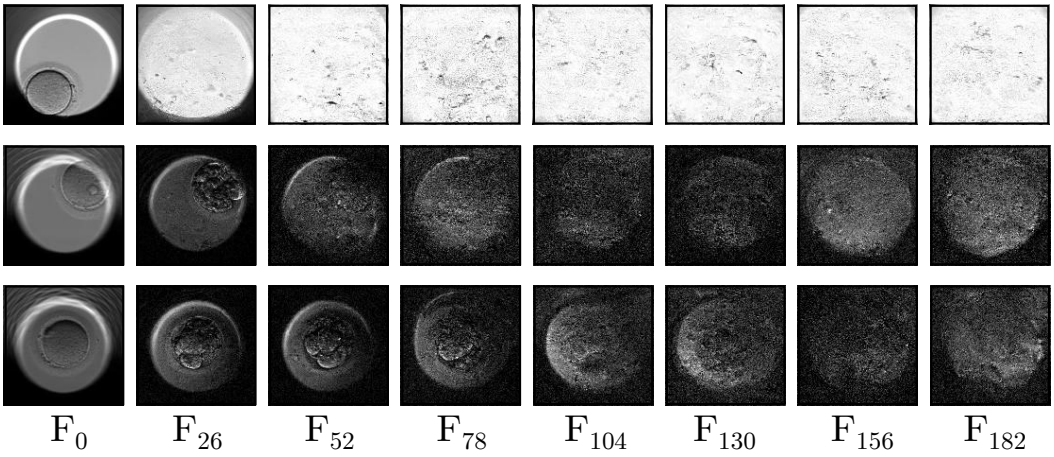

Figure 2. Extracted frames from long videos generated using the default model.

Given this limitation, the proposed model employs a Siamese architecture to generate an overall video as a guide for interpolating new frames, thereby avoiding this loss of quality. The proposed architecture allows for the generation of longer videos, providing a more comprehensive visualization of embryonic development.

As can be seen in Figure 3, the generation of fake videos is carried out using two Siamese models that are specially trained to fulfill different objectives. The first model, called the overview diffusion model (ODM), aims to predict frames from one or more real images. This model specializes in recognizing full embryonic development, capturing the progress of its stages and generating an overview fake video (OF). The ODM is trained using short videos obtained from real datasets, consisting of 15 equally spaced frames. The second model, known as the fill diffusion model (FDM), learns to generate images to interpolate new frames between OF frames, thus generating a filled fake video (FF). Unlike the ODM, the FDM is trained on 200-frame videos and randomly selects two images to serve as conditioning factors.

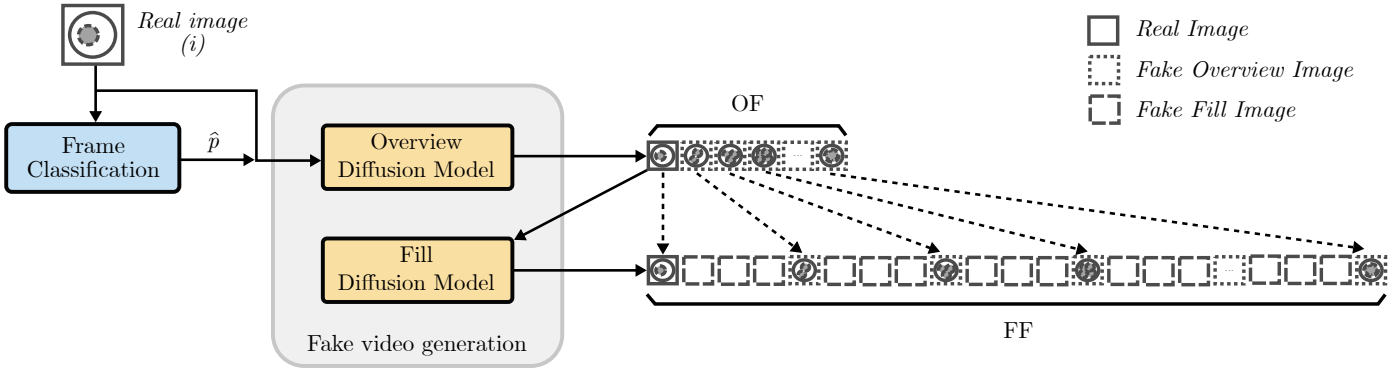

Figure 3. Siamese architecture for long video generation.

In the inference step, the ODM generates a video containing all stages that embryos go through, while the FDM is responsible for filling the gaps between each pair of frames to increase the length of the video. Thanks to this step, the final output goes from 15 frames to 197 frames.

2.3. Initial Frame Placement

In controlled environments, images used as conditioning factors can always be pictures of the initial stage, but this situation may not occur in the real world, either due to a lack of expert personnel or because the data are not annotated. For this reason, the proposed architecture uses an image classifier model to obtain the class represented in each input image. Then, this input image i is assigned a position p based on the probability of it being located in a frame, according to the distribution of the appearance of stages in certain frames in the real dataset. This means that the image is automatically assigned to the position where it would most likely appear in a real video. To calculate this position, the frequency of occurrence of each stage in the frames is first determined and then divided by the total number of occurrences. Equation (1) shows how this frequency $F(p,s)$ is obtained for each position p considering each stage s.

$$F(p,s) = \frac{Count_{p,s}}{TotalCount_p} \tag{1}$$

Finally, the value of s is replaced by the stage returned by the classifier model considering the conditional input image i, represented as $C(i)$. Equation (2) formalizes the initial

frame placement in an expected position $E(P, C(i))$ by considering all possible positions P and the classification output $C(i)$.

$$\hat{p} = E(P, C(i)) = \sum_{p}^{P} (p * F(p, C(i))) \tag{2}$$

2.4. Evaluation and Ranking of Fake Videos

Currently, there are no metrics that can evaluate the overall quality of embryo development sequences created by generative models. This is because when human experts analyze sequences, they take into account image quality, frame differences, object coherence within images, and the consistency of developmental stages. However, employing human experts is a slow and expensive process [39]. In this work, we propose a method for evaluating and ranking predictive sequences generated by the diffusion-based generative model. This method aims to select the best sequence for embryonic biological development without the intervention of experts. This evaluation takes into account the quality of each frame, the coherence of the stages with respect to the real dataset, and the presence of abrupt changes and movements between frames.

2.4.1. Image Quality

The first step is to separately classify each of the frames to determine the stage that they are in. This phase serves two purposes: first, it provides data for analyzing the distribution and coherence of these stages and second, it assesses image quality. The better the quality of the generated images, the better the classification of their stages and, consequently, the higher the quality of the videos. However, if images are blurred or lack quality, the classifier is not able to make a good assessment of the stages contained in the video, thus resulting in poor stage development coherence.

2.4.2. Stage Development Coherence

After classifying all frames, the resulting distribution of classes is compared to the distribution of real videos through the use of the ANCOVA statistical measure (analysis of covariance), which is based on ANOVA (analysis of variance) [56]. As described by Rutherford [57], ANCOVA assesses experimental manipulations on dependent variables objectively and accurately. A single factor single covariate ANCOVA can be formally described as Equation (3), where j represents an observation of class c and $\hat{\mu}$ is the interception point over the y axis of the regression coefficient $\beta_w(Z_{cj} - Z_G)$ of the covariate variable Z influenced by the general covariate mean Z_G [58].

$$Y_{cj} = \hat{\mu} + \hat{\alpha}_j + \beta_w(Z_{cj} - Z_G) + \varepsilon_{cj} \tag{3}$$

Lastly, $\varepsilon_{cj} \sim N(0, \sigma^2)$ stands for the associated unobserved error for the jth member in the c class, taking a value between 0 and the population N variance obtained by Equation (4).

$$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}{N} \tag{4}$$

Therefore, ANCOVA may be formally estimated for a class as the adjusted mean of all observations as follows:

$$\overline{Y}_{ac} = \overline{Y}_c - \hat{\mu} - \beta_w(\overline{Z}_c - Z_G) - \varepsilon_c \tag{5}$$

To obtain the ANCOVA similarity score, a comparison is made between the real video set used to train the generative model and the generated fake videos. The value used is the uncorrected p-value, hereinafter referred to as AP. The p-value relates to various statistical tests within ANCOVA, including tests for the main effects of factors (such as group differences), their interactions, and the effects of covariation. Its purpose is to

determine whether there are significant differences across variables, controlling for the effects of covariates.

The primary objective of video generation models is to produce videos that faithfully represent reality, taking into account both the quality of each image and the coherence of the entire sequences. To conduct this evaluation, we suggest assessing how closely videos resemble real-life situations and the realism of the development of included stages. To better analyze these cases, we propose using Spearman's correlation coefficient, hereafter referred to as Spearman, to check the coherence of stages.

According to Xiao et al. [59], Spearman (S) is a special case of Pearson's correlation coefficient, which uses monotonic relationships between two variables instead of linear relationships. It analyzes how well data fit a monotonic function, that is, a function that is always increasing or decreasing [60,61]. Spearman can be formally written as Equation (6), where N is the total number of samples and $d_i = R(x_i) - R(y_i)$ is the difference between each pair of observations ordered by its position [59].

$$S = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)} \tag{6}$$

Spearman takes into account the slope of the regression while providing a measure of the dispersion of the points examined. As a result, it outputs a measure value between -1 and 1, where 0 means that there is no slope and the observations are highly dispersed, while values close to -1 or 1 indicate very little dispersion and a good fit to a monotonic function [62]. Finally, negative values are obtained from functions with overall negative slopes, while positive values are produced from those with positive slopes. Spearman can lead to situations where videos with high dispersions and slopes obtain similar scores as videos with low dispersions and slopes.

Since this is a frame-by-frame analysis of videos, it is important to note that large dispersions mean that there are large differences between the stages of close frames. This results in the generation of videos showing unrealistic development. For this purpose, we propose using the standard error (SE) of the regression slopes (Equation (7)), so that videos whose points fit well on their regression lines obtain better scores than those whose points are too far apart. Consequently, the calculation of the DAS score can be expressed as $DAS = S + SE$.

$$SE = \sqrt{\frac{1}{N-2} * \frac{\sum (y_i - \hat{y}_i)^2}{\sum (x_i - \overline{x})^2}} \tag{7}$$

Finally, each video is ranked, taking into account both its developmental realism and its similarity to the training dataset. Its formal representation is shown in Equation (8).

$$fss = DAS + (1 - AP). \tag{8}$$

As explained at the beginning of the section, these statistical measures are applied to the results of classifying individual frames to determine the stage that they are in. This entire process involves estimation error. Although not explicitly included in the calculation, the linear regression used, along with the use of the Spearman and ANCOVA measures, indirectly considers this error. That is, if a frame is classified into a particular stage that is not correct, being in a stage close to the correct stage results in a lower decrease in score than being in a stage far from the correct stage. Instead of using discrete values, the further the stage of the frame from its ideal position, the lower the score obtained. This introduces a certain tolerance for error in the final score.

2.4.3. Frame Pair Comparison

Another aspect to consider, in addition to image quality and stage development coherence, is the transition between frames in the videos. Large differences between neighboring frames can happen due to the generation of erroneous frames where the

content is entirely different or the creation of correct content with overly abrupt movements. To identify such cases, we conduct a pairwise analysis of the produced frames, which considers pixel variations at the same image coordinates and motion detection.

Equation (9) shows how the MSE is calculated for each pair of frames f_p and f_{p+1}. In this calculation, H is the number of rows in the images (height) and W is the total number of columns (width).

$$\mathrm{MSE}(f_p, f_{p+1}) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (f_p(h,w) - f_{p+1}(h,w))^2 \tag{9}$$

Changes that occur in the embryo during its development over time significantly impact the mean squared error (MSE) since it undergoes changes and differs from contiguous images. To prevent false positives, the MSE calculation divides each image into sectors of equal height and width. This approach ensures that even if sectors of the developing embryo change, the remainder remain stable, resulting in a more accurate assessment.

For motion detection, the optical flow algorithm described by Farneback [63] is used. This algorithm is based on approximating the neighborhood of each pixel using a quadratic polynomial and then analyzing how this polynomial transforms to estimate displacement. Once the flow for each pixel is obtained, it can be stated that each pixel value in the image f_p is equivalent to the pixel value in the image f_{p+1} if this flow is applied to them (Equation (10)).

$$f_p(h,w) \sim f_{p+1}(h + \mathrm{flow}(h,w)[1], w + \mathrm{flow}(h,w)[0]) \tag{10}$$

It is worth mentioning that there can be variations between successive images and sudden movements in original datasets. To optimize the usage of the MSE and flow algorithm, the corresponding thresholds for each are calculated based on real datasets. Finally, Figure 4 shows the workflow of the proposed model for the evaluation and ranking of generated videos.

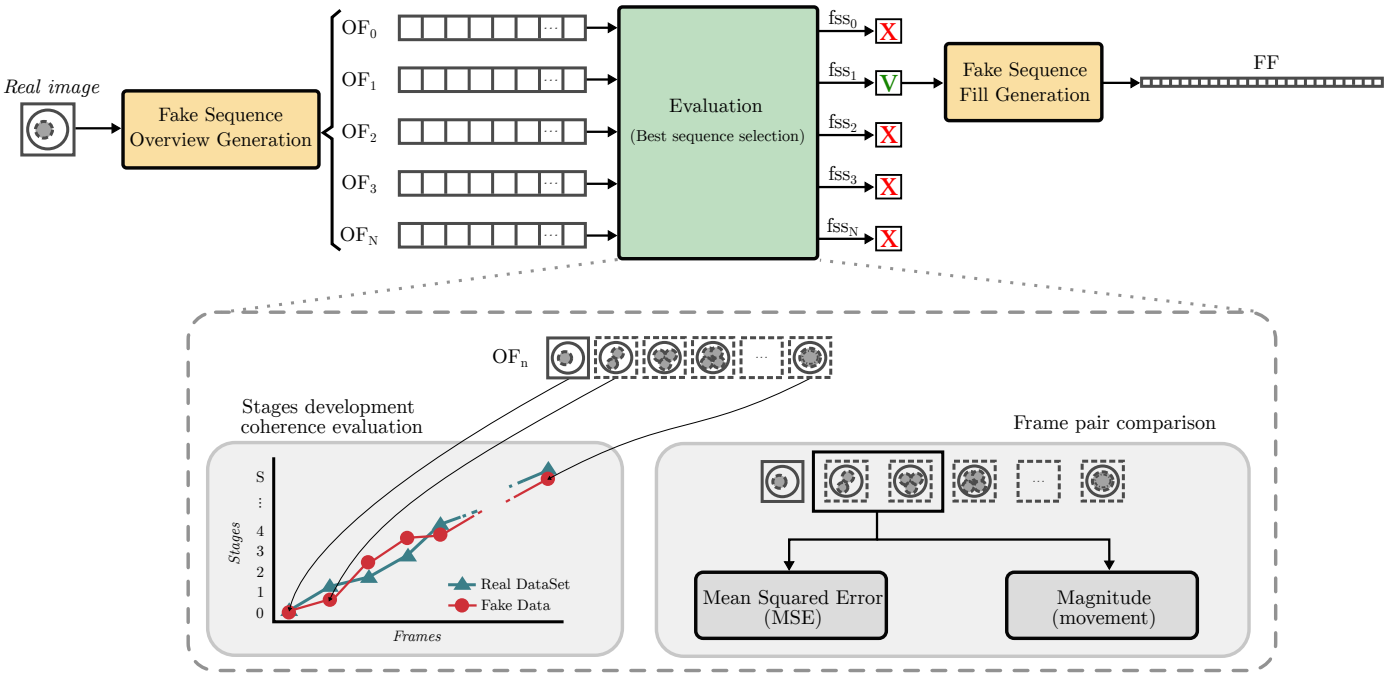

Figure 4. Automatic best sequence selection according to evaluation metrics.

3. Experiments

3.1. Data

Embryo development is a public time-lapse dataset created by Gomez et al. [28] containing cultures of embryos that have undergone IVF. It consists of 704 videos in different focal planes, with a total of 2.4 million images with a resolution of 500 × 500. The collection includes expert annotations for each stage of the embryo: polar body appearance (tPB2); pronuclei appearance and disappearance (tPNa and tPNf); blastomere division from 2-cell stage to 9 (or more)-cell stage (t2, t3, t4, t5, t6, t7, t8 and t9+); compaction (tM); blastocyst formation (tSB; tB), expansion and hatching (tEB and tHB). In addition, this work includes an extra class to detect the absence of an embryo in the image, which is referred to as empty. Due to technical limitations, only one focal plane was used for training.

To conduct the experiments, the total number of videos was split into three subsets for training, validation, and testing, utilizing a split of 80%, 10%, and 10%, respectively. This resulted in a total of 566 videos for training, 69 videos for validation, and 69 videos for testing.

3.2. Hyperparameters and Model Training

The first step in the process proposed in this work is to classify the initial input images to position them in the frames that they have the highest probability of appearing according to the real dataset. For this purpose, a ResNet18 image classifier model [64] was chosen. This is an 18-layer convolutional neural network that uses skip connections to learn residual functions with respect to layer inputs. Table 1 shows the results obtained using this classifier model on the test data.

Table 1. Classification results of ResNet18.

Stage	Precision	Recall	F1-Score
tPB2	0.920	0.933	0.926
tPNa	0.974	0.973	0.974
tPNf	0.864	0.851	0.857
t2	0.950	0.965	0.957
t3	0.946	0.712	0.812
t4	0.924	0.963	0.943
t5	0.884	0.866	0.875
t6	0.886	0.860	0.873
t7	0.949	0.867	0.906
t8	0.943	0.978	0.960
t9+	0.979	0.970	0.974
tM	0.924	0.930	0.911
tSB	0.942	0.924	0.933
tB	0.910	0.913	0.911
tEB	0.945	0.955	0.950
tHB	0.990	0.843	0.904
empty	0.973	0.984	0.978

As previously outlined, this study proposes the use of two Siamese models to produce long-duration videos while retaining control over the depicted stages. The first model, known as the overview diffusion model (ODM), generates a summary of the stages from start to finish utilizing 15 frames. The second model, referred to as the filling diffusion model (FDM), introduces new frames between each stage of the overview. The number of frames generated by both models was set to 15 due to hardware limitations. Nevertheless, using a higher number of frames could start to exhibit signs of quality loss that are inherent to the model. Since the same frame generated by the ODM is used as the final frame of one infilling frame and the initial frame of the next segment, the total number of frames in a sequence is 197. Both models were trained with an image size of 128 × 128 and one channel

as the images from the original dataset are monochrome. The training process incorporated two residual blocks, 128 base model channels, and a learning rate of 2×10^{-5}.

Following the process for the base model proposed by Höppe et al. [27], the ODM was trained by selecting random frames as initial conditional images from a dataset consisting of sequences of 15 frames from beginning to end. On the other hand, the FDM was trained to specialize in interpolation between two frames by using the first and last frames of the sequence to be generated. To accomplish this task, two images that were 15 frames apart were randomly selected from the original dataset, videos in which can contain up to 500 frames.

The ODM model was trained for 200,000 epochs, whereas the FDM model was trained for 300,000 epochs. This was due to the fact that the FDM model needs more iterations to acquire knowledge on how to generate videos displaying greater variability between the different stages of development than the production of the overview videos.

3.3. Evaluation Metrics

The following is an outline of the evaluation methods used in our study to assess the caliber of the generated videos in terms of their image quality and the coherence of the fake videos. The used metrics were Fréchet inception distance (FID) [65], Fréchet video distance (FVD) [39], structural similarity (SSIM) [66], peak signal-to-noise ratio (PSNR) [67], and mean squared error (MSE) [68].

The FID and FVD are based on the standard inception score (IS). The IS uses a pre-trained inception network model on the ImageNet dataset to classify the images to be evaluated and assess their quality based on how accurately these images can be classified. However, this evaluation method lacks utility in situations where the ImageNet collection is incapable of recognizing subject images studied by novel models because they are outside of its training cases, making correct classification impossible. For this reason, Heusel et al. [65] suggest utilizing an inception model to extract the distribution of fake images and contrast them with the distribution of real images. The authors rely on the idea that the more similar these two distributions are, the better the fake images created by the generative model.

This study centered on video generation, where simply evaluating the separate images that make up a video using the FID is inadequate. The learning process of video generative models is a much harder task than synthesizing static images because models must capture the temporal dynamics of scenes in addition to the visual presentations of objects. In this same context, Unterthiner et al. [39] proposed the FVD, which operates similarly to the FID but uses 3D convolution operations to obtain the distribution of videos as indivisible objects. Therefore, the FVD evaluates video quality as a sequence instead of assessing individual frames. Both the FID and FVD are measures of the distances between distributions. Therefore, the closer these values are to zero, the higher the quality of the evaluated object. The FID and FVD metrics have inherent biases. Both metrics are robust to changes in brightness, contrast, and saturation, but are weaker when it comes to changes in object edges and texture. Two of the key distinguishing features when generating the temporal evolution of static objects over time, as in the case of embryos, are texture and shape. Mismatches between human perception and the results of these metrics may occur because of these distinguishing features [69].

In order to evaluate the quality of the prediction videos, the real videos were compared to the generated videos one frame at a time using the mean squared error (MSE) [68], peak signal-to-noise ratio (PSNR) [67], and structural similarity (SSIM) [66].

As Setiadi explains in their work [70], the MSE measures the differences between the values of each pixel in both images at the same coordinates, providing an overall mean assessment of their similarity. The PSNR is based on the MSE but, instead of using a sum, it employs a logarithmic operation. Finally, the SSIM compares the luminance, contrast, and structure of both images. Unlike the MSE, the higher the results of the PSNR and SSIM measurements, the more similar the two images. The MSE, PSNR, and SSIM applied to

individual frames are useful for comparing frames separately, but none of them analyze the temporal evolution of the stages shown in the videos. This is one of the major limitations of these metrics. In addition, all of these measures are highly sensitive to noise, the presence of which can lead to inaccurate evaluations.

In addition, the MSE is used to evaluate the prediction of video stages. Each frame is classified and compared to the corresponding real video stage. Low MSE values suggest closer alignment between the predicted and actual stages, while high MSE values indicate discrepancies between the real and fake videos.

4. Results and Discussion

This section reports both the image and video quality, as well as stage development coherence. The image and video quality was evaluated against well known algorithms, including Fréchet inception distance (FID), Fréchet video distance (FVD), structural similarity (SSIM), peak signal-to-noise ratio (PSNR), and pixel-level mean squared error (MSE). The stage development coherence was assessed using the MSE values of the stages assigned by a classifier model. Furthermore, a qualitative analysis of the generated videos was performed.

4.1. Image and Video Quality

The evaluation of image and video quality typically involves two main types of measures. The first group comprises FID and FVD, which assess the distances between real and fake distributions. In contrast, the MSE, SSIM, and PSNR compare image pairs while disregarding the temporal element.

Table 2 displays the results of experiments comparing real videos to those created using the default base model (RaMViD) and the proposed method. Figure 2 shows that videos of significant length that were generated utilizing the default model exhibited notably poor quality after 30 frames. Consequently, trials were conducted on overview videos. To analyze the results, it is crucial to note that better performance is indicated by lower values of FID, FVD, and MSE. Conversely, higher values for SSIM and PSNR indicate enhancements in the results.

Table 2. Results for image and video quality. ↓ represents that lower values indicate better results; ↑ represents that higher values indicate better results.

Metric	Default Method	Proposed Method
FID (↓)	135.87	129.18
FVD (↓)	817.89	802.46
MSE (↓)	99.64	97.46
SSIM (↑)	0.34	0.39
PSNR (↑)	28.24	28.63

Thanks to the inclusion of the proposed model, any instances of low-quality or unrealistic sequences in the generated images were eliminated. This enhanced the resemblance of both the distribution of individual images (FID) and videos (FVD) to the actual dataset, thereby positively impacting the overall quality of the generated cases.

Improvement was also seen in terms of MSE, SSIM, and PSNR when comparing the real and fake frames. The proposed sequence evaluation method enhanced the accuracy of the represented stage and embryo position, resulting in minimal discrepancies between the predicted and actual images.

4.2. Stage Development Quality

The stages were analyzed using the MSE method. The video frames were classified and compared to the real videos in the test set. To focus on stage analysis rather than image quality (although related), experiments were conducted on both the long videos (197 frames) created by the default model "Default(F)" and the overview videos "Default(O)"

(15 frames). In the case of the proposed method "Proposed", the frames in the overview videos aligned with the frames of the long videos. This occurred due to the way they were constructed, using infilling between frames. The results are presented in Table 3.

Table 3. Results for stage MSE experiments. Default(O) refers to videos of 15 frames generated as overview videos. Default(F) refers to videos of 197 frames with equally distributed frames extracted for testing.

Position	Default(F)	Default(O)	Proposed Method
1	11.54	9.26	**4.22**
2	153.42	12.87	**4.43**
3	152.71	11.90	**7.10**
4	125.43	12.52	**6.36**
5	106.51	14.20	**12.97**
6	94.01	16.81	**8.67**
7	49.91	13.77	**7.36**
8	41.30	12.83	**7.51**
9	35.29	11.17	**7.84**
10	22.54	**9.74**	11.19
11	18.16	16.06	**12.67**
12	16.86	18.70	**15.23**
13	43.56	16.58	**11.74**
14	45.74	16.32	**10.59**
Global Mean	59.30	13.31	**9.00**

The Default(F) model could not provide accurate representations of the stages in the video. It was only able to approach the other models in the first position, closer to the input conditioning image. Positions 10, 11, and 12 also showed similarity to the other models. This situation occurred because these positions usually correspond to stages t9+, tM, tSB, and tB, in which the embryo transitions from having a easily quantifiable number of cells to having a more uniform body. The classifier assigned one of these stages to the images when it was incapable of assigning one of the other more easily recognizable categories. As a result, when the correct stage positions were predicted, they were considered accurate. This situation was particularly noticeable in the case of the Default(O) model as it only outperformed the proposed model in position 10.

In the remaining positions, the proposed model achieved a better fit for the classes and obtained a much lower error. The proposed model was particularly effective in the initial positions, which were closer to the conditioning real images for future predictions. In both models, the position with the highest error was position 12. This position, following the reasoning from the previous paragraph, belonged to stages where the model had more difficulty generating differentiating features. This was due to the embryonic morphology in these stages, where the embryo transitions from being formed of differentiable cells to containing a less defined morula. However, clarity was achieved in the subsequent stages of tEB, tHB, and empty, which offer clear differentiating characteristics and facilitate the generation of related images.

Based on the global MSE, the proposed model demonstrated significant improvement over both default models. These metrics reinforced the proposed model theory. Despite this, as outlined in Section 3.3, there was still substantial dependence on expert judgment in assessing the synthetic videos. Therefore, as well as the statistical measures presented in this section, a qualitative evaluation of the generated videos was also carried out.

4.3. Qualitative Analysis

In this section, we conduct an analysis of the fake videos generated using the two Siamese models to fill in 197 frames. One model generates overview videos, while the other fills in new frames. However, when using the default model, the generated videos experienced significant image quality loss when the number of frames exceeded 20∼30.

This scenario is observable in Figure 2. Therefore, an initial analysis of the overview videos produced was conducted. Figure 5 displays a subset of the sample videos, consisting of eight frames uniformly extracted from each overview video. Images A, B, and C served as the real conditional images on which the 15-frame predictions were made.

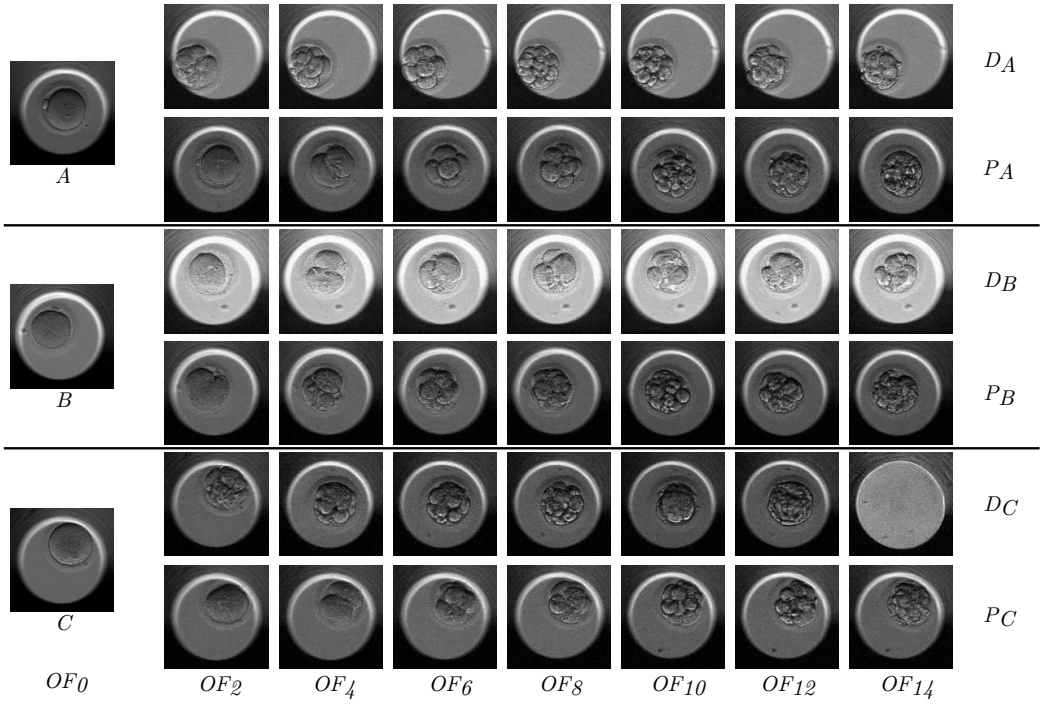

Figure 5. Extracted frames from overview sequences. These sequences are 15 frames long. Default sequences are named D, while videos generated using the proposed method are defined by P.

In Case A, the default model displayed a vastly dissimilar progression from the initial image concerning both color and the position of the embryo. Nevertheless, the recommended model output a video that matched the initial image more accurately. Moreover, the default video D_A displayed highly developed stages in initial positions with minimal development throughout the video, whereas the video generated by the proposed model P_A showcased more differentiated stages and a more realistic progression.

In Case B, the default model D_B maintained a correct position, while the video underwent an excessively sudden color shift that was not present in case P_B. Similarly to the prior case, P_B exhibited a more realistic progression of stages, though the frame OF_2 does not portray a completely accurate morphological entity. This issue arose due to the many transitional frames between stages in the real dataset, in which the morphology of the embryo rapidly undergoes significant changes. During these instances, the model could not produce fully coherent images, but the stage remained identifiable.

Finally, in Case C, both models attained correct positioning and tonality. The movement in case D_C could be encountered in the real world, so it was not considered erroneous. Nevertheless, D_C displayed an early-stage development that was too advanced in the video. Thanks to the proposed work presented in this paper, P_C achieved a more uniform distribution of stages with no stage skipping and was closer to the real data.

Figure 6 shows the three filled sequences corresponding to the cases in Figure 5. Each video comprised 197 images generated from an initial conditional image. Seven frames were randomly selected from each case for an analysis of quality, morphological coherence, and the consistency of the stages depicted in the synthetic videos. The filled sequences complemented the overview sequences with the best scores, as explained in Section 2.2. The model introduced new images between each pair of generated frames, enhancing the level of detail of the temporal evolution of the embryo. Its development over time

remained unaltered, but the filled sequences provided a smoother sequence with fewer abrupt changes, resembling the real videos more closely.

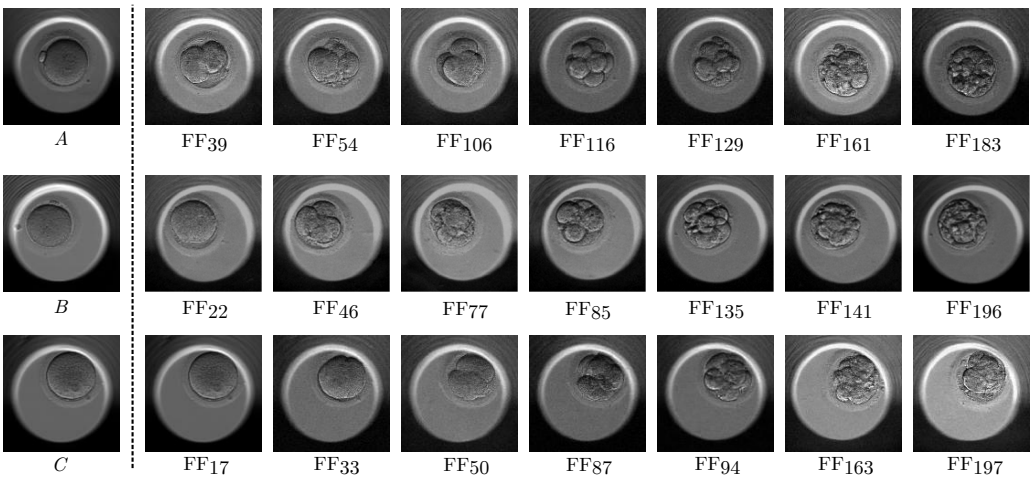

Figure 6. Extracted frames from filled sequences. These sequences are 197 frames long, with 8 random frames extracted as examples.

From a general perspective, it can be observed that the video quality remained consistent from beginning to end without undergoing continuous degradation from frame 20∼30. The image quality exhibited certain tonal changes between some neighboring frames, which are more noticeable in this figure due to the distances between the selected frames. However, the transition was smoother throughout the entire video. The frames displayed high-quality images without any artifacts or noise. The morphology of the embryo was consistent throughout all stages and the location of the embryo did not suffer from any sudden changes. It is important to note that there were transitional frames between each stage that displayed typical morphological changes found in real cases, which may be confusing in still images. With regard to morphology, the overview sequences represented the complete development of the embryo in only 15 frames, so morphological differences between adjacent frames could be substantial when there was a transition between stages. The diffusion model, which aimed to maintain coherence between frames, could introduce some inconsistencies in this situation. However, the second model generated a total of 197 frames, resulting in much smaller differences between adjacent frames. By representing the transition between stages with a larger number of frames, the model could better capture the morphology of the embryo and how it changes during development. Thus, the proposed model could generate videos with smoother transitions.

5. Conclusions

This paper explored the creation of long-duration videos featuring the development of embryos from in vitro fertilization through the generation of predictive videos and frame infilling techniques. A Siamese architecture based on diffusion models and a novel approach to video assessment were introduced to select the frames that match real input images most accurately. The proposal in this study considered both the quality of the frames and the coherence in the progression of the stages featured in the videos. To accomplish this, statistical methods and the stage development distributions in the authentic training set were used.

The results indicated a lack of degradation in long-duration videos, which is a limitation of the base model. In addition, generating overview videos reduced the occurrence of stages appearing in inconsistent sequences, allowing for more realistic depictions of development. As for image quality, the proposed method successfully removed cases that displayed sudden movements or changes and better fit the input images. Improvements in standard evaluation measures (such as FID, FVD, and stage MSE) and pixel-level image

quality metrics (MSE, SSIM, PSNR) supported these outcomes. Additionally, the results obtained from the evaluation measures were reaffirmed by a qualitative analysis of a subset of the generated videos.

Although the proposed model showed good results in generating predictive image sequences for the temporal evolution of embryos, it had several limitations that must be considered. Both generative models extract hidden patterns from data, which requires a large dataset to offer robust predictions. Furthermore, the high number of variables and conditions affecting the temporal development of embryos mean that it may be necessary to incorporate additional multimodal information into the model. Additionally, the model requires significant computational power and extensive training times, particularly when working with high-resolution images. This highlights the importance of investigating optimization strategies and creating methods to decrease the computational demands of the model when working with 3D data.

Future research should seek to improve the consistency of luminance and contrast. Moreover, the generation process may benefit from a preliminary localization task dedicated to ensuring that the images capture solely the positioned embryo occupying the entire frame. This would eliminate issues related to unwanted movements and assist models in focusing on learning to generate the morphological structures of embryos. Localizing the study area would eliminate the generation of non-embryonic areas, which are irrelevant for this research. It would also be interesting to explore the introduction of additional information from developmental predictions using other time series-focused machine learning techniques. This would promote more accurate forecasts, relieving the generative models of this responsibility. Additionally, models correlate with the tolerance limit resulting from the use of probability estimation, so the explicit inclusion of error estimation and its percentile changes in the models could enhance their results. Although the proposed method aims to generate videos depicting embryonic development, statistical measures enable the extrapolation of the proposed model to other areas where the study of temporal development is of interest.

However, the proposed model holds significant potential for evolution prediction in the biomedical field, promoting advancements in predictive analytics and diagnostic tools. The model has the ability to generate long realistic video predictions of embryonic development. Generated synthetic data can help to achieve more precise predictions, enhance the development of more resilient machine learning models, and provide valuable resources for training and enhancing the medical expertise of professional experts. The use of such highly realistic synthetic data, which are not subject to data protection regulations like real data, could aid policymakers in supporting research initiatives that utilize deep learning-based generation models in the biomedical sector. This collaborative approach could facilitate partnerships among academic institutions, healthcare providers, and technology companies. Such partnerships may accelerate progress in understanding complex biological phenomena, ultimately leading to improved patient care and enhanced medical education.

Author Contributions: Conceptualization, P.C. and A.S.V.; methodology, P.C.; software, P.C. and J.M.S.-F.; validation, A.S.V.; formal analysis, J.M.S.-F.; investigation, P.C.; resources, E.L.I.; data curation, A.S.V.; writing—original draft preparation, P.C.; writing—review and editing, E.L.I. and L.B.; visualization, J.M.S.-F.; supervision, L.B.; project administration, E.L.I. and L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Conselleria de Cultura, Educación e Universidade (Xunta de Galicia) through funding from the ED431C 2022/03-GRC Competitive Reference Group and by the Ministerio de Ciencia e Innovación through the State Programmes for Knowledge Generation and Scientific and Technological Strengthening of the R&D&i System (PID2020-113673RB-I00). Pedro Celard was supported by a predoctoral fellowship from the Xunta de Galicia (ED481A 2021/286).

Data Availability Statement: Publicly available datasets were analyzed in this study. These datasets can be found in Gomez et al. [28], available at https://doi.org/10.5281/zenodo.6390798, accessed on 21 November 2023. The data and source codes presented in this study are openly available from https://github.com/pedrocelard/EmbryoPredict, accessed on 18 January 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FID	Fréchet Inception Distance
FVD	Fréchet Video Distance
MSE	Mean Squared Error
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity
GAN	Generative Adversarial Network
DM	Diffusion Model
VAE	Variational Autoencoder
ODM	Overview Diffusion Model
FDM	Fill Diffusion Model

References

1. Galić, I.; Habijan, M.; Leventić, H.; Romić, K. Machine learning empowering personalized medicine: A comprehensive review of medical image analysis methods. *Electronics* **2023**, *12*, 4411. [CrossRef]
2. Celard, P.; Iglesias, E.L.; Sorribes-Fdez, J.M.; Romero, R.; Vieira, A.S.; Borrajo, L. A survey on deep learning applied to medical images: From simple artificial neural networks to generative models. *Neural Comput. Appl.* **2023**, *35*, 2291–2323. [CrossRef] [PubMed]
3. Bhalla, D.; Rangarajan, K.; Chandra, T.; Banerjee, S.; Arora, C. Reproducibility and explainability of deep learning in mammography: A systematic review of literature. *Indian J. Radiol. Imaging* **2023**. [CrossRef]
4. Sultan, S.M.; Mollika, M.T.; Fahim, S.A.; Alam, T.; Mohammed, A.F.Y.; Islam, T. Automated cell counting system using improved implicit activation based U-Net (IA-U-Net). In Proceedings of the 2023 IEEE 5th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), Tainan, Taiwan, 2–4 June 2023; pp. 1–4. [CrossRef]
5. Gomes, R.F.T.; Schuch, L.F.; Martins, M.D.; Honório, E.F.; de Figueiredo, R.M.; Schmith, J.; Machado, G.N.; Carrard, V.C. Use of Deep Neural Networks in the detection and automated classification of lesions using clinical images in ophthalmology, dermatology, and oral medicine—A systematic review. *J. Digit. Imaging* **2023**, *36*, 1060–1070. [CrossRef] [PubMed]
6. Apostolidis, K.; Kokkotis, C.; Karakasis, E.; Karampina, E.; Moustakidis, S.; Menychtas, D.; Giarmatzis, G.; Tsiptsios, D.; Vadikolias, K.; Aggelousis, N. Innovative visualization approach for biomechanical time series in stroke diagnosis using explainable machine learning methods: A proof-of-concept study. *Information* **2023**, *14*, 559. [CrossRef]
7. Daidone, M.; Ferrantelli, S.; Tuttolomondo, A. Machine learning applications in stroke medicine: Advancements, challenges, and future prospectives. *Neural Regen. Res.* **2024**, *19*, 769–773. [CrossRef] [PubMed]
8. Sharma, A.K.; Nandal, A.; Dhaka, A.; Dixit, R. Medical image classification techniques and analysis using deep learning networks: A review. In *Health Informatics: A Computational Perspective in Healthcare*; Springer: Singapore, 2021; pp. 233–258. [CrossRef]
9. Dadoun, H.; Rousseau, A.L.; de Kerviler, E.; Correas, J.M.; Tissier, A.M.; Joujou, F.; Bodard, S.; Khezzane, K.; de Margerie-Mellon, C.; Delingette, H.; et al. Deep Learning for the detection, localization, and characterization of focal liver lesions on abdominal US images. *Radiol. Artif. Intell.* **2022**, *4*, e210110. [CrossRef]
10. Alzaid, A.; Wignall, A.; Dogramadzi, S.; Pandit, H.; Xie, S.Q. Automatic detection and classification of peri-prosthetic femur fracture. *Int. J. Comput. Assist. Radiol. Surg.* **2022**, *17*, 649–660. [CrossRef]
11. Zhu, C.; Liang, J.; Zhou, F. Transfer learning-based YOLOv3 model for road dense object detection. *Information* **2023**, *14*, 560. [CrossRef]
12. Nowakowska, S.; Borkowski, K.; Ruppert, C.M.; Landsmann, A.; Marcon, M.; Berger, N.; Boss, A.; Ciritsis, A.; Rossi, C. Generalizable attention U-Net for segmentation of fibroglandular tissue and background parenchymal enhancement in breast DCE-MRI. *Insights Imaging* **2023**, *14*, 185. [CrossRef]
13. Xia, Y.; Yun, H.; Liu, Y.; Luan, J.; Li, M. MGCBFormer: The multiscale grid-prior and class-inter boundary-aware transformer for polyp segmentation. *Comput. Biol. Med.* **2023**, *167*, 107600. [CrossRef] [PubMed]
14. Wang, J.; Peng, Y.; Jing, S.; Han, L.; Li, T.; Luo, J. A deep-learning approach for segmentation of liver tumors in magnetic resonance imaging using UNet++. *BMC Cancer* **2023**, *23*, 1060. [CrossRef] [PubMed]
15. Yu, M.; Guo, M.; Zhang, S.; Zhan, Y.; Zhao, M.; Lukasiewicz, T.; Xu, Z. RIRGAN: An end-to-end lightweight multi-task learning method for brain MRI super-resolution and denoising. *Comput. Biol. Med.* **2023**, *167*, 107632. [CrossRef]

16. Li, J.; Cairns, B.J.; Li, J.; Zhu, T. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *Npj Digit. Med.* **2023**, *6*, 98. [CrossRef] [PubMed]
17. Sheikhalishahi, S.; Bhattacharyya, A.; Celi, L.A.; Osmani, V. An interpretable deep learning model for time-series electronic health records: Case study of delirium prediction in critical care. *Artif. Intell. Med.* **2023**, *144*, 102659. [CrossRef]
18. Chlap, P.; Min, H.; Vandenberg, N.; Dowling, J.; Holloway, L.; Haworth, A. A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imaging Radiat. Oncol.* **2021**, *65*, 545–563. [CrossRef]
19. Chen, Y.; Yang, X.H.; Wei, Z.; Heidari, A.A.; Zheng, N.; Li, Z.; Chen, H.; Hu, H.; Zhou, Q.; Guan, Q. Generative Adversarial Networks in medical image augmentation: A review. *Comput. Biol. Med.* **2022**, *144*, 105382. [CrossRef]
20. Hashemifar, S.; Marefat, A.; Hassannataj Joloudari, J.; Hassanpour, H. Enhancing face recognition with latent space data augmentation and facial posture reconstruction. *Expert Syst. Appl.* **2024**, *238*, 122266. [CrossRef]
21. Joseph, A.J.; Dwivedi, P.; Joseph, J.; Francis, S.; Pournami, P.N.; Jayaraj, P.B.; Shamsu, A.V.; Sankaran, P. Prior-guided generative adversarial network for mammogram synthesis. *Biomed. Signal Process. Control* **2024**, *87*, 105456. [CrossRef]
22. Sun, P.; Mo, Z.; Hu, F.; Song, X.; Mo, T.; Yu, B.; Zhang, Y.; Chen, Z. 2.5D MFFAU-Net: A convolutional neural network for kidney segmentation. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 92. [CrossRef]
23. Islam, N.U.; Zhou, Z.; Gehlot, S.; Gotway, M.B.; Liang, J. Seeking an optimal approach for Computer-aided Diagnosis of Pulmonary Embolism. *Med. Image Anal.* **2024**, *91*, 102988. [CrossRef] [PubMed]
24. Wong, N.; Ingledew, P.A. The quality of YouTube videos on radiotherapy and prostatectomy for prostate cancer. *Can. Urol. Assoc. J.* **2023**, *18*, 2. [CrossRef]
25. Lee, S.G.; Kim, G.Y.; Hwang, Y.N.; Kwon, J.Y.; Kim, S.M. Adaptive undersampling and short clip-based two-stream CNN-LSTM model for surgical phase recognition on cholecystectomy videos. *Biomed. Signal Process. Control* **2024**, *88*, 105637. [CrossRef]
26. Abdelmotaal, H.; Hazarbassanov, R.M.; Salouti, R.; Nowroozzadeh, M.H.; Taneri, S.; Al-Timemy, A.H.; Lavric, A.; Yousefi, S. Keratoconus detection-based on dynamic corneal deformation videos using Deep Learning. *Ophthalmol. Sci.* **2024**, *4*, 100380. [CrossRef] [PubMed]
27. Höppe, T.; Mehrjou, A.; Bauer, S.; Nielsen, D.; Dittadi, A. Diffusion models for video prediction and infilling. *arXiv* **2022**, arXiv:2206.07696. [CrossRef]
28. 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 Brief* **2022**, *42*, 108258. [CrossRef] [PubMed]
29. Sarandi, S.; Boumerdassi, Y.; O'Neill, L.; Puy, V.; Sifer, C. Intérêt de l'iDAScore (intelligent Data Analysis Score) dans la pratique quotidienne d'un laboratoire de FIV pour la sélection embryonnaire : Résultats d'une étude préliminaire. *Gynécologie Obs. Fertil. Sénologie* **2023**, *51*, 372–377. [CrossRef] [PubMed]
30. Chiappetta, V.; Innocenti, F.; Coticchio, G.; Ahlström, A.; Albricci, L.; Badajoz, V.; Hebles, M.; Gallardo, M.; Benini, F.; Canosa, S.; et al. Discard or not discard, that is the question: An international survey across 117 embryologists on the clinical management of borderline quality blastocysts. *Hum. Reprod.* **2023**, *38*, 1901–1909. [CrossRef]
31. Pons, M.C.; Carrasco, B.; Rives, N.; Delgado, A.; Martínez-Moro, A.; Martínez-Granados, L.; Rodriguez, I.; Cairó, O.; Cuevas-Saiz, I. Predicting the likelihood of live birth: An objective and user-friendly blastocyst grading system. *Reprod. BioMedicine Online* **2023**, *47*, 103243. [CrossRef]
32. Coticchio, G.; Ezoe, K.; Lagalla, C.; Zacà, C.; Borini, A.; Kato, K. The destinies of human embryos reaching blastocyst stage between Day 4 and Day 7 diverge as early as fertilization. *Hum. Reprod.* **2023**, *38*, 1690–1699. [CrossRef]
33. Huang, B.; Zheng, S.; Ma, B.; Yang, Y.; Zhang, S.; Jin, L. Using deep learning to predict the outcome of live birth from more than 10,000 embryo data. *BMC Pregnancy Childbirth* **2022**, *22*, 36. [CrossRef] [PubMed]
34. Theilgaard Lassen, J.; Fly Kragh, M.; Rimestad, J.; Nygård Johansen, M.; Berntsen, J. Development and validation of deep learning based embryo selection across multiple days of transfer. *Sci. Rep.* **2023**, *13*, 4235. [CrossRef] [PubMed]
35. Lan, L.; You, L.; Zhang, Z.; Fan, Z.; Zhao, W.; Zeng, N.; Chen, Y.; Zhou, X. Generative adversarial networks and its applications in biomedical informatics. *Front. Public Health* **2020**, *8*, 164. [CrossRef]
36. Sujata, P.N.; Madiwalar, S.M.; Aparanji, V.M. Machine learning techniques to improve the success rate in in-vitro fertilization (IVF) procedure. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *925*, 012039. [CrossRef]
37. Dirvanauskas, D.; Maskeliūnas, R.; Raudonis, V.; Damaševičius, R.; Scherer, R. HEMIGEN: Human embryo image generator based on generative adversarial networks. *Sensors* **2019**, *19*, 3578. [CrossRef] [PubMed]
38. Celard, P.; Seara Vieira, A.; Sorribes-Fdez, J.M.; Romero, R.; Lorenzo Iglesias, E.; Borrajo Diz, L. Study on synthetic video generation of embryo development. In Proceedings of the Hybrid Artificial Intelligent Systems, Salamanca, Spain, 5–7 September 2023; Springer: Cham, Switzerland, 2023; pp. 623–634.
39. Unterthiner, T.; van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; Gelly, S. Towards accurate generative models of video: A new metric & challenges. *arXiv* **2018**, arXiv:1812.01717. [CrossRef]
40. Yan, Y.; Yang, T.; Zhao, X.; Jiao, C.; Yang, A.; Miao, J. DC-SiamNet: Deep contrastive siamese network for self-supervised MRI reconstruction. *Comput. Biol. Med.* **2023**, *167*, 107619. [CrossRef]
41. Tan, J.; Dong, Y.; Li, J. Automated fundus ultrasound image classification based on siamese convolutional neural networks with multi-attention. *BMC Med. Imaging* **2023**, *23*, 89. [CrossRef]

42. Tulyakov, S.; Liu, M.Y.; Yang, X.; Kautz, J. MoCoGAN: Decomposing motion and content for video generation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1526–1535. [CrossRef]
43. Clark, A.; Donahue, J.; Simonyan, K. Efficient video generation on complex datasets. *arXiv* **2019**, arXiv:1907.06571.
44. Saito, M.; Saito, S.; Koyama, M.; Kobayashi, S. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal GAN. *Int. J. Comput. Vis.* **2020**, *128*, 2586–2606. [CrossRef]
45. Babaeizadeh, M.; Saffar, M.T.; Nair, S.; Levine, S.; Finn, C.; Erhan, D. FitVid: Overfitting in pixel-level video prediction. *arXiv* **2021**, arXiv:2106.13195. [CrossRef]
46. Saxena, V.; Ba, J.; Hafner, D. Clockwork variational autoencoders. In Proceedings of the Neural Information Processing Systems, Virtual, 7–10 December 2021.
47. Mittal, S.; Lajoie, G.; Bauer, S.; Mehrjou, A. From points to functions: Infinite-dimensional representations in diffusion models. In Proceedings of the ICLR Workshop on Deep Generative Models for Highly Structured Data, Virtual, 25–29 April 2022.
48. Dockhorn, T.; Vahdat, A.; Kreis, K. Score-based generative modeling with critically-damped Langevin diffusion. *arXiv* **2022**, arXiv:2112.07068. [CrossRef]
49. Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; Fleet, D.J. Video diffusion models. *arXiv* **2022**, arXiv:2204.03458. [CrossRef]
50. Voleti, V.; Jolicoeur-Martineau, A.; Pal, C. MCVD-masked conditional video diffusion for prediction, generation, and interpolation. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 23371–23385.
51. Yang, R.; Srivastava, P.; Mandt, S. Diffusion probabilistic modeling for video generation. *Entropy* **2023**, *25*, 1469. [CrossRef]
52. Dhariwal, P.; Nichol, A. Diffusion models beat GANs on image synthesis. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 7–10 December 2021; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 8780–8794.
53. Luo, C. Understanding diffusion models: A unified perspective. *arXiv* **2022**, arXiv:2208.11970. [CrossRef]
54. Hagemann, P.; Mildenberger, S.; Ruthotto, L.; Steidl, G.; Yang, N.T. Multilevel diffusion: Infinite dimensional score-Based diffusion models for image generation. *arXiv* **2023**, arXiv:2303.04772. [CrossRef]
55. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention 2015, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
56. Henson, R. Analysis of Variance (ANOVA). In *Brain Mapping*; Toga, A.W., Ed.; Academic Press: Waltham, MA, USA, 2015; pp. 477–481. [CrossRef]
57. Rutherford, A. *ANOVA and ANCOVA: A GLM Approach*; John Wiley & Sons: Staffordshire, UK, 2011.
58. Miller, G.A.; Chapman, J.P. Misunderstanding analysis of covariance. *J. Abnorm. Psychol.* **2001**, *110*, 40–48. [CrossRef]
59. Xiao, C.; Ye, J.; Esteves, R.M.; Rong, C. Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurr. Comput. Pract. Exp.* **2016**, *28*, 3866–3878. [CrossRef]
60. Bakhtiar, A.; Suliantoro, H.; Ningsi, R.H.; Pitipaldi, K. Relationship of quality management system standards to industrial property rights in Indonesia using Spearman Correlation Analysis Method. *IOP Conf. Ser. Earth Environ. Sci.* **2021**, *623*, 012092. [CrossRef]
61. Jia, K.; Yang, Z.; Zheng, L.; Zhu, Z.; Bi, T. Spearman correlation-based pilot protection for transmission line connected to PMSGs and DFIGs. *IEEE Trans. Ind. Inform.* **2021**, *17*, 4532–4544. [CrossRef]
62. Sharma, A.; Suryawanshi, A. A novel method for detecting spam email using KNN classification with Spearman correlation as distance measure. *Int. J. Comput. Appl.* **2016**, *136*, 28–35. [CrossRef]
63. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In Proceedings of the Image Analysis, Halmstad, Sweden, 29 June–2 July 2003; Bigun, J., Gustavsson, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; pp. 363–370.
64. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
65. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; pp. 6629–6640.
66. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
67. Huynh-Thu, Q.; Ghanbari, M. The accuracy of PSNR in predicting video quality for different video scenes and frame rates. *Telecommun. Syst.* **2012**, *49*, 35–48. [CrossRef]
68. Wang, Z.; Bovik, A.C. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.* **2009**, *26*, 98–117. [CrossRef]

69. Borji, A. Pros and cons of GAN evaluation measures: New developments. *Comput. Vis. Image Underst.* **2022**, *215*, 103329. [CrossRef]
70. Setiadi, D.R.I.M. PSNR vs SSIM: Imperceptibility quality assessment for image steganography. *Multimed. Tools Appl.* **2021**, *80*, 8423–8444. [CrossRef]