2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-7358-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISBI53787.2023.10230571

MODELING WOUND HEALING USING VECTOR QUANTIZED VARIATIONAL AUTOENCODERS AND TRANSFORMERS

Lenka Backová[1,2,], Guillermo Bengoetxea[2], Svana Rogalla[2], Daniel Franco-Barranco[1,4], Jérôme Solon[2,3], Ignacio Arganda-Carreras[1,2,3,4]*

[1] Dept. of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU)
[2] Instituto Biofisika (CSIC, UPV/EHU) [3] Ikerbasque, Basque Foundation for Science
[4] Donostia International Physics Center (DIPC)

ABSTRACT

Wound healing is a fundamental mechanism for living animals. Understanding the process is crucial for numerous medical applications ranging from scarless healing to faster tissue regeneration and safer post-surgery recovery. In this work, we collect a dataset of time-lapse sequences of *Drosophila* embryos recovering from a laser-incised wound. We model the wound healing process as a video prediction task for which we utilize a two-stage approach with a vector quantized variational autoencoder and an autoregressive transformer. We show our trained model is able to generate realistic videos conditioned on the initial frames of the healing. We evaluate the model predictions using distortion measures and perceptual quality metrics based on segmented wound masks. Our results show that the predictions keep pixel-level error low while behaving in a realistic manner, thus suggesting the neural network is able to model the wound-closing process.

***Index Terms*—** Wound healing, video prediction, deep learning

1. INTRODUCTION

Wound healing is essential for the survival of living animals to repair damaged tissues or cellular structures. Uncovering its underlying biological and physical principles is key for many biomedical applications, such as tissue engineering, regenerative medicine, and healing acceleration [1]. Consequently, the process has been intensively investigated, and several biophysical models have attempted to reproduce experimental observations and uncover the underlying mechanisms [2, 3]. However, due to its complexity, a model able to quantitatively predict the course of healing *in vivo* is yet to be developed.

In this work, we aim to model wound healing in *Drosophila melanogaster* embryos to quantitatively reproduce the kinetics of the process. Embryonic wound healing is a simple scarless model of healing driven by an accumulation of actomyosin, which forms a contractile ring at the leading edge of the wound [4]. The actomyosin ring generates pulling forces deforming the epithelial tissue and promoting wound closure. We aim to utilize different proteins labeled with GFP to examine their behavior during the healing process, starting with myosin. This opens up the possibility of comparing model predictions for different proteins and studying their relative importance in the healing process. Unlike biophysical models, we seek to achieve this without presuming any laws of Physics directly and only by observing the spatiotemporal information of individual protein signals during the process.

In particular, our goal is to extract complex, non-linear kinetics of the healing process by utilizing artificial neural networks (ANNs). Although ANNs have become a standard tool for many bioimage analysis tasks [5], they have rarely been exploited to predict kinetics and multi-cellular dynamic behavior quantitatively. Specifically, we use a set of time-lapse images acquired by a spinning disc microscope as input for a video prediction ANN. Video prediction [6] (also referred to as next-frame prediction) models learn to generate frames of videos by observing information from the initial frames. They have been applied to diverse data where the anticipation of the future is important, e.g., weather forecasting [6], autonomous driving [7] and robotics [8]. Models capturing information from a sub-sequence of a video can be used for early action prediction [9] or cell fate prediction in microscopy [10].

Modeling sequences has been done by transformer models in natural language processing tasks [11], which have recently been successfully applied in other domains [12, 13], including video generation [14]. A two-stage approach has been used for image [15, 13] and video [14] generation. It consists of an autoencoder that learns latent representations of the data and a generative transformer to sample new frames by learning the distribution. An advantage of such architectures is that they can be conditioned on the initial frames of the video. Compared to the approaches used on non-microscopy videos [6, 8, 7, 14], we face several challenges. Namely, small datasets, as the cost of acquiring microscopy videos is expensive and time-exhausting, and longer sequences, since those are needed to predict the entire wound closure.

Corresponding author: lenka.backova@ehu.eus

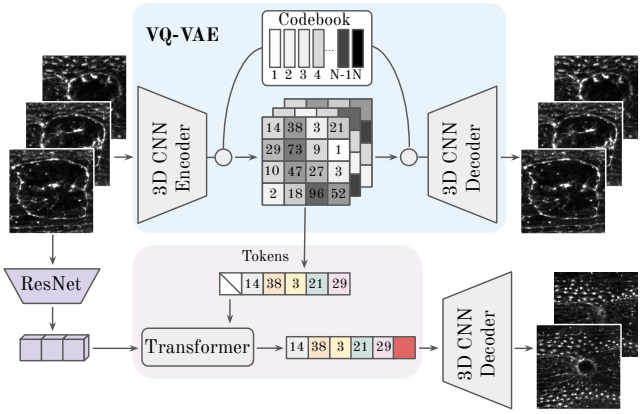


Fig. 1: Our two-stage approach for wound healing video prediction. First, a VQ-VAE is trained to learn discrete latent representations of wound healing time-lapse sequences. Next, a transformer is trained to predict future latents conditioned on a set of frames processed by a 3D ResNet. Finally, the VQ-VAE decoder trained in the first stage converts those latents into new frames.

Contributions. In this work, we acquire a dataset of time-lapse sequences of embryos regenerating their tissue after a laser-induced wound[1], and we model the process using a neural network for video prediction. Finally, we analyze the behavior of the predictions in time and show our approach is able to generate long videos correctly, modeling the process while being conditioned on a handful of frames displaying the initial state of the wound. We quantitatively evaluate our method using both distortion and perceptual quality-like metrics based on the distance of the predicted and observed segmentation masks, as well as examining the area reduction of the wound in time. Our results show the test predictions behave in a similar and realistic manner to experimentally observed wound healing processes. To the best of our knowledge, this is the first time a conditioned video generation neural network has been successfully applied to model complex multi-cellular dynamic behavior and its kinetics, specifically wound healing captured by microscopy time-lapse sequences.

2. METHODS

Following the seminal idea of VideoGPT [14], we phrase the wound healing modeling as a video prediction task using a two-stage architecture that combines a vector quantized variational autoencoder (VQ-VAE) [16] and an autoregressive transformer adapted for videos. The first network learns to (1) downsample and embed the image sequences in a discrete latent space and (2) reconstruct them. The second network learns to generate new frames in the discretized latent space. The generation is conditioned on the past frames. After the frames are predicted in the discretized space, the VQ-VAE reconstructs the predictions to image sequences. A full diagram of our framework is shown in Fig. 1 and described next.

Autoencoder. The VQ-VAE is trained to encode the high-dimensional image sequences into a small set of discrete tokens (integers). It consists of an encoder, a codebook, and a decoder. The encoder is composed of 3D strided convolutions followed by a single attention residual block. This way, the input image sequence is converted into a set of downsampled vectors, which are replaced by the nearest-neighbor vectors in an embedding codebook. Next, the decoder, consisting of an attention residual block and 3D strided transposed convolutions, upsamples the vector representation and reconstructs the input image sequence. The indices of the nearest-neighbor vectors in the codebook are used as the low-dimensional discrete representation of the input image sequence.

Transformer. Given a few starting frames of an image sequence, the transformer is trained to predict the rest of the sequence. Due to the high dimensionality of the image sequences, the transformer works in the low-dimensional discrete representation space computed by the VQ-VAE. Specifically, given an image sequence encoded as N discrete tokens, the transformer is trained to predict the k-th token ($k \leq N$) given the previous $k-1$ tokens. To enrich the context information, the few starting frames are also encoded by a 3D ResNet and added to the transformer's input. For additional details on the architecture and training, we refer to [14].

3. EXPERIMENTAL RESULTS

3.1. Wound healing dataset acquisition and processing

We collect a dataset of time-lapse sequences of *Drosophila* embryos healing after a laser-induced wound. Altogether, 61 sequences are acquired (train, validation, test sets of sizes 44, 12, and 5, respectively.) We image embryos expressing a GFP tagged myosin II (sqhGFP) [17] at stage 17, collected after aging for 16 hours at 18°C The embryos are mounted on a coverslip covered with heptane glue for conventional confocal microscopy and mineral oil to prevent drying. Laser ablation is performed with a pulsed laser. The time-lapse sequences are acquired *in vivo* using an Olympus IXplore SpinSR10, illuminated by laser with 488 nm wavelength and using a 60×1.42 NA oil immersion objective.

The individual frames are taken once every minute. Each sequence captures the entire closure of the wound with an average length of 120 frames. Each frame has 1152×1152 pixels and captures the whole embryo with $4.6\,\mu$m/pixel image resolution. A z-stack of the embryo with a step size of $1\,\mu$m is imaged at each time step. The time-resolved z-stacks are max-projected along the z-axis, and registered by SIFT algorithm [18] in Fiji. The wounds are segmented by a custom stable U-Net-like architecture [19] and manually refined in napari [20]. As each frame includes the entire embryo, it is cropped to a region of 256×256 pixels containing the wound and downsampled to 128×128 by bicubic interpolation.

[1]Dataset available at lenkaback.github.io/wound-healing-modeling

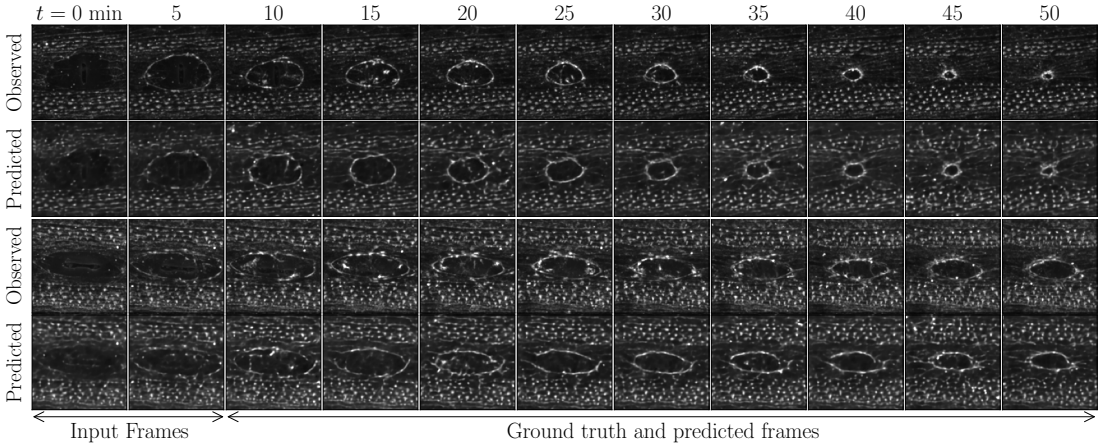


Fig. 2: Predictions on two test sequences. The initial 8 frames are given as context (represented by the first two frames on the left) and the rest is iteratively predicted until wound closure (for visualization purposes, every fifth frame is shown.)

3.2. Training and evaluation regime

First, we train the VQ-VAE to downsample sequences of 20 frames of 128×128 pixels into $5 \times 32 \times 32$ tokens. Next, the transformer is trained on the encoded 20-frame sequences, where the first eight frames are used as the context, and the transformer predicts the 12 following frames. Both networks are trained as in [14], with the following exceptions. To compensate for the small dataset size, we use data augmentation (image transposition, horizontal, vertical flips, and rotation) and random sampling of 20-frame sub-sequences from the original sequences. We select the best VQ-VAE and transformer models based on the validation loss for quantitative and qualitative evaluation. Specifically, we evaluate our proposed method on the test sequences. For each sequence, we use the first eight frames of the sequence as the context and iteratively predict the whole sequence till the wound closure. Given 8 context frames the model predicts 12 frames. Invariably, the last eight predicted frames are used as the context to predict the subsequent 12 frames. This process is repeated until the entire length of the test sequence is reached.

3.3. Evaluation Metrics

Similar to the evaluation of super-resolution and image restoration algorithms, the performance of video generation methods is measured by calculating the similarity between predicted and observed frames. Usually, image similarity is estimated by two types of metrics: distortion error metrics to calculate pixel differences and perceptual quality metrics to estimate human-perceivable differences. Analogously, we evaluate our results using two types of metrics. To compute the metrics, the predicted wounds are segmented. First, we calculate the distances between the predicted and observed wound segmentation masks as a shape distortion measure. Second, we estimate the perceptual quality by comparing the wound area reduction over time in the predicted test sequences with respect to the observed training sequences. This helps us see whether the generated sequences are consistent with the area reduction of the observed data, thus, behaving realistically.

Distance metrics. We use the mean and Hausdorff distances as a proxy for a shape distortion measure between the predicted and observed masks. Specifically, we calculate both the mean distance d_M and the Hausdorff distance d_H between the borders of the segmented masks. If X, Y are point sets of the boundary pixels, we define both metrics as:

$$d_M(X,Y) = \frac{1}{2}\Big(\frac{1}{|X|}\sum_{x\in X}\inf_{y\in Y} d_{xy} + \frac{1}{|Y|}\sum_{y\in Y}\inf_{x\in X} d_{xy}\Big),$$

$$d_H(X,Y) = \max\Big\{\sup_{x\in X}\inf_{y\in Y} d_{xy}, \sup_{y\in Y}\inf_{x\in X} d_{xy}\Big\},$$

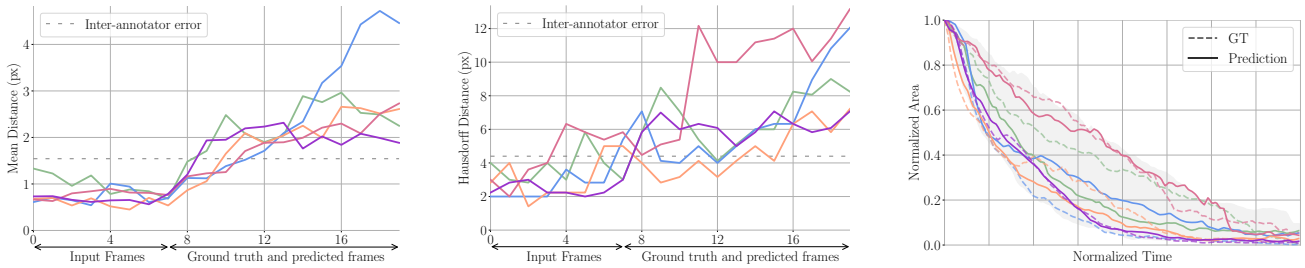
where d_{xy} is the Euclidean distance between points x and y. The mean and the Hausdorff distance show the average and the furthest distance, respectively, of two corresponding pixels on the border, thus measuring how well the shape and position of the predicted and observed wounds match.

Area reduction over time. To estimate the perceptual quality, we examine the area reduction of the wound as it heals. The wound area is calculated as the number of non-zero pixels of the segmented filled mask. The area in each time frame is normalized by the maximum (i.e., initial) wound area.

3.4. Results on unseen data

The results of our proposed method on the test sequences are evaluated both qualitatively (by visual inspection) and quantitatively (by assessing the distortion and perception errors).

Qualitative results. Qualitative results of the predictions on two test sequences are shown in Fig. 2. We can see our method learns to replicate the presence and accumulation of the myosin ring, as well as the reduction of wound area in time. Moreover, the quality of the reconstruction by VQ-VAE is very high, as the conditional frames are visually nearly



(a) Mean Distance (b) Hausdorff Distance (c) Area Reduction

Fig. 3: Quantitative results on test sequences. From left to right, we show the mean (a) and Hausdorff (b) distances between the segmented wound masks of the predicted and observed frames for each test sequence, and (c) the 10-percentile band of area reduction over time of the train and validation dataset of the observed sequences, overlaid with the predicted and observed area reduction (each color represents a different test sequence.)

identical. The predictions are qualitatively better at the beginning as the signal starts to deteriorate slightly over time. Nevertheless, overall the predictions are visually realistic.

Shape distortion of the predicted wounds. The shape of individual wounds varies and evolves during healing. Ensuring our approach can correctly model individual wound healing, the predicted and observed wound borders must behave similarly. To evaluate this, we employ distance metrics, which show shape distortion of the prediction when compared to the observed border of the wound. We evaluate the distance metrics (Section 3.3) for the initial 20 frames (8 context frames and 12 predicted unseen frames) of each test sequence. The distance metrics are reported in Figures 3a and 3b, with an inter-annotator error reported for each metric. We calculate this error between the segmentation borders of 27 sequences segmented by two different human experts as a lower bound for our method. The overall averaged values are ~ 1.5 pixels for the mean and ~ 4.4 pixels for the Hausdorff distance.

In the initial 8 frames, we see a systematic deviation from the observed masks, lower or similar to the inter-annotator error. This could be caused by a segmentation error. As expected, the metrics deviate more notably after the context frames. The 12 predicted frames have average metric values of 2.2 pixels for the mean and 6.8 pixels for the Hausdorff distance. However, the error in prediction is not uniform, as the mean distance increases slightly and peaks between 2-3 pixels; while, one sequence shows a higher deviation above 4 pixels. These results show that the average difference in the wound shape is relatively small compared to the inter-annotator error. The Hausdorff distance increases more drastically, peaking between 6-13 pixels. More sequences show higher deviation.The difference between the mean and the Hausdorff distance suggests the segmented predictions do not differ significantly from the observed on average but include several outlier pixels. It was important to include the 3D ResNet to enrich the input of the transformer, as we saw approximately 50% improvement in the average Hausdorff distance of the 12 predicted frames.

Modeling wound area reduction. In all sequences from our dataset, the area of the wounds reduces over time until closure. We evaluate whether the generated wounds behave realistically and, in particular, whether the area reduction over time is consistent with the observed data. Specifically, we compare the area reduction in the predictions with that of the training and validation videos. To this aim, we process the area of each wound in the whole dataset for each frame as described in Section 3.3. We repeat the calculation for each predicted and observed test sequence. To compare wounds that close at different time points, we normalize the time between the initial and the last frame. We plot a 10-percentile band of the area reduction of the whole dataset over time; and overlay the area reduction curves of the ground truth and predicted test sequences over the dataset band in Fig. 3c. We observe that the area reduction rate in our predictions matches the trend in the whole dataset, as well as most of the individual predictions match their counterpart observations.

How the wound closes does not follow the same trend for every wound, and the area reduction over time curves show variability. This suggests the network not only predicts the closure based on the average kinetics but also predicts wound-specific closure matching the tendency of the observed closure. Altogether suggesting the model generates wounds with viable closure tendencies, generating realistic wounds.

4. CONCLUSION

In this work, we introduce a novel method for modeling wound healing time-lapse sequences with a two-stage architecture combining VQ-VAE and an autoregressive transformer. We evaluated our method on a dataset of wound healing time-lapse sequences of *Drosophila* embryos and show our approach can effectively model wound healing conditioned on a few initial frames. Both qualitative and quantitative results demonstrate the generated sequences present low pixel-level error and display realistic healing patterns when predicted until full closure. This work opens up possibilities of using neural networks to model kinetics and complex dynamic behavior of other processes.

Acknowledgments. This work is supported in part by grants PID2019-109117GB-100 and PID2021-126701OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", by the foundation Biofisika Bizkaia and by grant GIU19/027 funded by the UPV/EHU. Author SR is supported by Human Frontier Science Program Organization (fellowship LT0007/2022-L) and author GB is supported by a fellowship from Ministerio de Ciencia e Innovación (fellowship PRE2020-094463).

Compliance with Ethical Standards. This work is a study for which no ethical approval was required.

5. REFERENCES

[1] Yaiza Belacortu and Nuria Paricio, "Drosophila as a model of wound healing and tissue regeneration in vertebrates," *Developmental Dynamics*, 2011. 1

[2] Adrian R. Noppe, Anthony P. Roberts, Alpha S. Yap, Guillermo A. Gomez, and Zoltan Neufeld, "Modelling wound closure in an epithelial cell sheet using the cellular Potts model," *Integrative Biology*, 2015. 1

[3] Kenta Odagiri, Hiroshi Fujisaki, Hiroya Takada, and Rei Ogawa, "Numerical simulation using cellular Potts model for wound closure with ATP release and the mechanobioligical effects," 2022, arXiv:2209.01354. 1

[4] William Wood, Antonio Jacinto, Richard Grose, Sarah Woolner, Jonathan Gale, Clive Wilson, and Paul Martin, "Wound healing recapitulates morphogenesis in Drosophila embryos," *Nature Cell Biology*, 2002. 1

[5] Erik Meijering, "A bird's-eye view of deep learning in bioimage analysis," *Computational and structural biotechnology journal*, 2020. 1

[6] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *NeurIPS*, 2015. 1

[7] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip Yu, and Mingsheng Long, "PredRNN: A recurrent neural network for spatiotemporal predictive learning," *TPAMI*, 2022. 1

[8] Chelsea Finn, Ian Goodfellow, and Sergey Levine, "Unsupervised learning for physical interaction through video prediction," *NeurIPS*, 2016. 1

[9] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei, "Eidetic 3D LSTM: A model for video prediction and beyond," *ICLR*, 2018. 1

[10] Christopher J. Soelistyo, Giulia Vallardi, Guillaume Charras, and Alan R. Lowe, "Learning biophysical determinants of cell fate with deep neural networks," *Nature Machine Intelligence*, 2022. 1

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *NeurIPS*, 2017. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021. 1

[13] Patrick Esser, Robin Rombach, and Björn Ommer, "Taming transformers for high-resolution image synthesis," *CVPR*, 2021. 1

[14] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas, "VideoGPT: Video Generation using VQ-VAE and Transformers," 2021, arXiv:2104.10157. 1, 2, 3

[15] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever, "Zero-shot text-to-image generation," *ICML*, 2021. 1

[16] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu, "Neural Discrete Representation Learning," *NeurIPS*, 2017. 2

[17] Anne Royou, William Sullivan, and Roger Karess, "Cortical recruitment of nonmuscle myosin II in early syncytial Drosophila embryos : its role in nuclear axial expansion and its regulation by Cdc2 activity ," *Journal of Cell Biology*, 2002. 2

[18] David Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004. 2

[19] Daniel Franco-Barranco, Arrate Muñoz-Barrutia, and Ignacio Arganda-Carreras, "Stable deep neural network architectures for mitochondria segmentation on electron microscopy volumes," *Neuroinformatics*, 2021. 2

[20] Nicholas Sofroniew, Talley Lambert, Kira Evans, Juan Nunez-Iglesias, Grzegorz Bokota, Philip Winston, Gonzalo Peña-Castellanos, Kevin Yamauchi, Matthias Bussonnier, Draga Doncila Pop, and et al., "napari: a multi-dimensional image viewer for python," 2022. 2