Technical Appendices for ImageFlowNet: Forecasting Multiscale Trajectories of Disease Progression with Irregularly-Sampled Longitudinal Medical Images

Chen Liu[1*] **Ke Xu**[1*] **Liangbo L. Shen**[2] **Guillaume Huguet**[3,4] **Zilong Wang**[3,5]
Alexander Tong[3,4§] **Danilo Bzdok**[3,5§] **Jay Stewart**[2§] **Jay C. Wang**[1,2,6§]
Lucian V. Del Priore[1§] **Smita Krishnaswamy**[1§]

[1]Yale University [2]University of California, San Francisco [3]Mila - Quebec AI Institute
[4]Université de Montréal [5]McGill University [6]Northern California Retina Vitreous Associates

* These authors are joint first authors: {chen.liu.cl2482, k.xu}@yale.edu. § Senior authors.
Please direct correspondence to: smita.krishnaswamy@yale.edu or lucian.delpriore@yale.edu.

Main paper can be found at https://arxiv.org/pdf/2406.14794.

Table of Contents

A Ablation Studies

A.1 Flow Field Formulation

The first experiment we considered was whether formulating the flow field f_θ as $f_\theta(z_t, t)$ or $f_\theta(z_t)$ would affect the performance. As mentioned earlier, we decided on the $f_\theta(z_t)$ formulation analytically, and here we support our decision with empirical evidence (Table S3).

Table S3: Effect of flow field formulation.

	PSNR↑	SSIM↑	MAE↓	MSE↓	DSC↑	HD↓
$f_\theta(z_t, t)$	22.42	0.643	0.123	0.027	0.872	48.38
$f_\theta(z_t)$	**22.63**	**0.646**	**0.119**	**0.024**	**0.874**	**42.68**

A.2 Single-scale vs Multiscale ODEs

The UNet architecture uses hierarchical hidden layers to extract multiscale representations. Starting at the image resolution and ending at the bottleneck layer (bottom of the "U"), the model produces increasingly higher-level and more global representations. In this study, we analyze the advantages of multiscale ODEs. Moreover, there might be multiple hidden layers at the same resolution. On which representations should we perform trajectory inference?

Table S4: Selection of latent representations for ODE inference.

	PSNR↑	SSIM↑	MAE↓	MSE↓	DSC↑	HD↓
bottleneck only	22.33	0.639	0.122	0.026	0.850	48.13
all unique resolutions	22.49	0.643	0.122	0.025	0.859	43.39
all unique layers	**22.63**	**0.646**	**0.119**	**0.024**	**0.874**	**42.68**

To study this, we explored the following settings: (1) infer a single-scale z_t from the bottleneck layer, (2) infer multiscale $\{z_t\}$ at all layers and use distinct f_θ for each resolution, but all hidden layers of the same resolution share the same f_θ, and (3) infer multiscale $\{z_t\}$ at all layers and use distinct f_θ for each hidden layer. The empirical results as shown in Table S4 indicate that modeling all representations separately would lead to the best performance.

Note: To avoid confusion, all of these hidden layers produce outputs that are bridged by skip connections from the contraction path to the expansion path.

A.3 Effects of Regularizations

We experimented with the effect of visual feature regularization under different λ_v (Table S5), contrastive learning regularization under different λ_c (Table S6), and trajectory smoothness regularization under different λ_s. (Table S7).

The results showed that regularization on visual features through a pre-trained vision encoder, contrastive learning regularization, and constraining on trajectory smoothness all yielded some improvements. The final set of weighting coefficients is $\lambda_v = 0.001$, $\lambda_c = 0.01$, and $\lambda_s = 0.1$.

Table S5: Effect of visual feature regularization.

λ_v	PSNR↑	SSIM↑	MAE↓	MSE↓	DSC↑	HD↓
0	22.63	0.646	0.119	**0.024**	**0.874**	**42.68**
0.001	**22.65**	**0.658**	**0.118**	**0.024**	0.872	44.27
0.01	22.64	0.650	0.120	0.025	0.872	45.89
0.1	22.57	0.647	0.120	0.025	0.869	50.69
1	22.54	0.634	0.124	0.027	0.867	48.13
λ_v = 0 & no MSE loss	21.53	0.593	0.133	0.031	0.860	49.78

Table S6: Effect of contrastive regularization.

λ_c	PSNR↑	SSIM↑	MAE↓	MSE↓	DSC↑	HD↓
0	22.63	0.646	0.119	**0.024**	0.874	42.68
0.001	22.63	0.646	0.119	0.025	0.872	46.23
0.01	**22.65**	**0.652**	**0.118**	**0.024**	**0.875**	**42.18**
0.1	22.38	0.651	0.121	0.025	0.871	45.30
1	22.25	0.644	0.121	0.025	0.868	46.85

Table S7: Effect of smoothness regularization.

λ_s	PSNR↑	SSIM↑	MAE↓	MSE↓	DSC↑	HD↓
0	22.63	0.646	0.119	**0.024**	0.874	**42.68**
0.001	22.38	0.649	0.123	0.027	0.870	46.91
0.01	22.65	0.648	0.119	0.024	0.870	45.71
0.1	**22.70**	**0.657**	**0.118**	**0.024**	**0.878**	47.44
1	22.69	0.655	**0.118**	**0.024**	0.875	45.16

B Propositions and Proofs

B.1 Proposition 3.2

Proposition B.1. *Let f_θ be a continuous function that satisfies the Lipschitz continuity and linear growth conditions. Also, let the initial state $y(t_0) = y_0$ satisfy the finite second moment requirement ($\mathbb{E}[|y(t_0)|^2] < \infty$). Suppose $z(t_0)$ is the latent representation learned by ImageFlowNet at the initial state corresponding to t_0. Then, our neural ODEs (Eqn (3a)) are at least as expressive as the original neural ODEs (Eqn (1a)), and their solutions capture the same dynamics.*

We recall the two dynamic systems for original neural ODEs and our ODEs:

Original neural ODEs:

$$\frac{\mathrm{d}y(\tau)}{\mathrm{d}\tau} = f_\theta(y(\tau), \tau), \quad f_\theta : \mathbb{R}^n \times [0, T] \to \mathbb{R}^n$$

Our neural ODEs, with (1) superscript $\cdot^{(b)}$ omitted without loss of generality, (2) z_τ equivalently replaced by $z(\tau)$ for notation consistency, and (3) f_θ replaced by $\tilde{f}_\theta$ for distinction:

$$\frac{\mathrm{d}z(\tau)}{\mathrm{d}\tau} = \tilde{f}_\theta(z(\tau)), \quad \tilde{f}_\theta : \mathbb{R}^m \to \mathbb{R}^m$$

Proof.

Theorem B.2 (Picard-Lindelöf [52])**.** *Let $D \subset \mathbb{R}^n$ be an open set, and let $f : D \times [0, T] \to \mathbb{R}^n$ be a continuous function that satisfies a Lipschitz condition in y uniformly in τ. Then, for any initial condition $y(t_0) = y_0$, there exists a unique solution to the initial value problem:*

$$\frac{\mathrm{d}y(\tau)}{\mathrm{d}\tau} = f(y(\tau), \tau), \quad y(t_0) = y_0.$$

Lipschitz Condition:

$$|f_\theta(y_1, \tau) - f_\theta(y_2, \tau)| \leq L|y_1 - y_2|$$

Linear Growth Condition:

$$|f_\theta(y, \tau)| \leq K(1 + |y|)$$

Given these conditions, both the original neural ODE and the Latent Space Neural ODE have unique strong solutions.

Since both the original ODE and the Latent Space Neural ODE have unique solutions, we could then construct a bijective and sufficiently smooth mapping $h : \mathbb{R}^n \times [0, T] \to \mathbb{R}^m$ such that $z(\tau) = h(y(\tau), \tau)$.

We define a function $h(y, \tau)$ that maps the state $y(\tau)$ and time τ to a new latent state $z(\tau)$ as

$$h(y, \tau) := y(\tau) \oplus \tau,$$

where $\oplus$ denotes the concatenation of the state and time.

Then, as h is bijective, the inverse function h^{-1} maps $z(\tau)$ back to $y(\tau)$ and τ. Given $h(y, \tau) = y \oplus \tau$, the inverse is:

$$h^{-1}(z) = (y(z_{\text{time}}), z_{\text{time}})$$

By the chain rule, the derivative of $z(\tau)$ with respect to τ is:

$$\frac{\mathrm{d}z(\tau)}{\mathrm{d}\tau} = \frac{\partial h}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}\tau} + \frac{\partial h}{\partial \tau}$$

Substituting the ODE for $y(\tau)$, we get:

$$\frac{\mathrm{d}z(\tau)}{\mathrm{d}\tau} = \frac{\partial h}{\partial y} f_\theta(y(\tau), \tau) + \frac{\partial h}{\partial \tau}$$

We can then simply define the function $\tilde{f}_\theta$ in the latent space such that it incorporates the dynamics from the original space:

$$\tilde{f}_\theta(z(\tau)) := \frac{\partial h}{\partial y} f_\theta(y(\tau), \tau) + \frac{\partial h}{\partial \tau}$$

The universal approximation theorem ensures that there exists a neural network parameterized by θ that can approximate any continuous function, including $\tilde{f}_\theta(z(\tau))$.

Existence of Equivalent Function Since the neural network can approximate $\tilde{f}_\theta(z(\tau))$, there exists a function $\tilde{f}_\theta(z(\tau))$ in the latent space that can represent the same system behavior governed by $f_\theta(y(\tau), \tau)$ in the original space.

Proving Equivalence: Given $z(\tau) = h(y(\tau), \tau)$ and the corresponding functions f_θ and $\tilde{f}_\theta$, we have shown that the new ODE formulation:

$$\frac{\mathrm{d}z(\tau)}{\mathrm{d}\tau} = \tilde{f}_\theta(z(\tau))$$

captures the same dynamics as the original ODE:

$$\frac{\mathrm{d}y(\tau)}{\mathrm{d}\tau} = f_\theta(y(\tau), \tau)$$

□

B.2 Proposition 3.3

Proposition B.3. *If we consider an image as a distribution over a 2D grid, ImageFlowNet is equivalently solving a dynamic optimal transport problem, as it meets three essential criteria: (1) matching the density, (2) smoothing the dynamics, and (3) minimizing the transport cost, where the ground distance is the Euclidean distance in the latent joint embedding space.*

Proof. ImageFlowNet can alternatively be viewed in the context of a dynamic optimal transport framework, which aims to determine the optimal plan π to transport mass from an initial distribution μ to a target distribution ν for a fixed state interval $[\tau_i, \tau_j]$. The task meets three requirements of dynamic optimal transport: (1) matching the density, (2) smoothing the dynamics, and (3) minimizing the transport cost. The ground distance in the latent joint embedding space is the Euclidean distance.

Matching the density The image is a 2D grid, and the distribution for the pixel intensities is μ at τ_i and ν at τ_j on this grid. μ and ν are defined on measure space $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^n$ respectively. The set of all joint probability measures on $\mathcal{X} \times \mathcal{Y}$ is denoted as $\Pi(\mu, \nu)$ and $c(x, y)$ is the cost of moving a mass unit from the original distribution μ at state τ_i to the target distribution ν at state τ_j. Then, the distance between the two distributions μ and ν is the p-Wasserstein distance:

$$W(\mu, \nu)_p := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y) \right)^{\frac{1}{p}}, \text{ where } p \geq 1$$

Benamou & Brenier [23] present a dynamic view of optimal transport, which links to differential equations. For the state interval $[\tau_i, \tau_j]$, there is a smooth and status-dependent density $P(z, \tau) \geq 0$ with $\int_{\mathbb{R}^n} P(z, \tau) dz = 1, \forall \tau \in [\tau_i, \tau_j]$, and a velocity fields $f(z, \tau)$ that obeys the continuity equation:

$$\partial_\tau P + \nabla \cdot (Pf) = 0, \text{ with } \tau \in [\tau_i, \tau_j] \text{ and } z \in \mathbb{R}^n, \text{where } P(\cdot, \tau_i) = \mu,\ P(\cdot, \tau_j) = \nu$$

Smoothing the dynamics The velocity fields $f(z, \tau)$ follows the Lipschitz condition $|f(z_1, \tau) - f(z_2, \tau)| \leq L|z_1 - z_2|$ where $L > 0$, which ensures a smooth and controlled transport process. With the following setup, Benamou & Brenier [23] show that the Wasserstein distance with order 2 (W_2) is:

$$W(\mu, \nu)_2^2 = \inf_{(p,f)} \int_{\mathbb{R}^n} \int_{\tau_i}^{\tau_j} P(z, \tau) \|f(z, \tau)\|^2 d\tau dz$$

Minimizing the transport cost Based on the main theorems in [53, 54], this problem aims to find the trajectory f that minimizes the transport cost on the path space $\mathbb{R}^n$, we define the ground distance in the latent joint embedding space to be the Euclidean distance:

$$W(\mu, \nu)_2^2 = \inf_f \mathbb{E}\left[\int_{\tau_i}^{\tau_j} \|f(z_\tau, \tau)\|^2 d\tau \right] \text{ s.t. } \frac{\partial z(\tau)}{\partial \tau} = f_\theta(z_\tau, \tau),\ z_{\tau_i} \sim \mu,\ z_{\tau_j} \sim \nu$$

Here, f_θ follows the ODE (3a) or the SDE (7).

With the above setups, ImageFlowNet is equivalent to a dynamic optimal transport problem trying to match the density at different states.

□

C Additional Background on Longitudinal Image Data

Longitudinal image datasets, including but not limited to retinal images or even medical images, often come with several challenges: (1) high dimensionality, (2) temporal sparsity, (3) sampling irregularity, and (4) spatial misalignment.

(1) **High Dimensionality** is intrinsic to image data. For images with height of H pixels, width of W pixels and C image channels, the dimensionality of the data is $\mathbb{R}^{H \times W \times C}$, which can easily go beyond a hundred thousand dimensions: a small image of $256 \times 256 \times 3$ has 196.6 thousand dimensions. Such high dimensionality is rarely encountered by most methods in time series prediction and temporal dynamics modeling [54, 56–60].

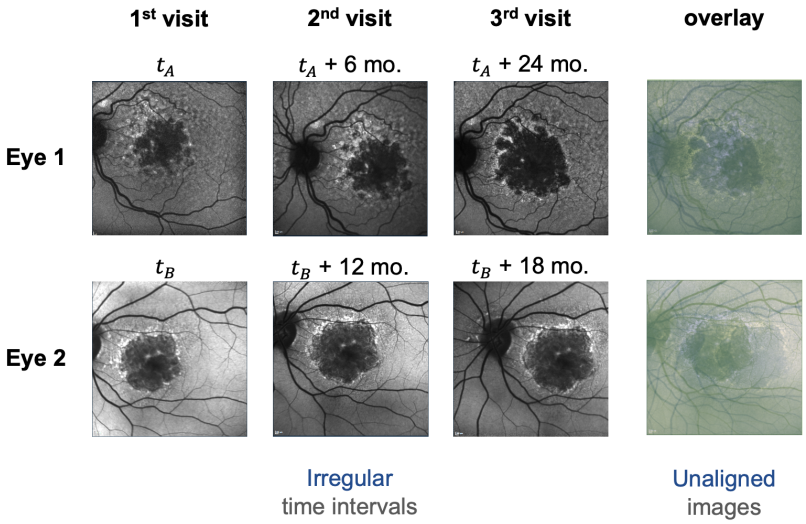

Figure S1: Temporal sparsity, sampling irregularity and spatial misalignment in longitudinal images.

(2) **Temporal Sparsity** is especially common in longitudinal images in healthcare, as images are usually acquired at separate visits of the patient, where the time gap can be several months or years. In contrast, a relatively well-studied adjacent field is video data [61–63], where the frame rate can easily be 60 Hz or higher. This renders our data of interest easily 10^8 times sparser compared to the better studied video data.

(3) **Sampling Irregularity** is also ubiquitous in clinical practice, both *within* and *among* longitudinal image series. *Within-series* irregularity means that the visits are not necessarily evenly distributed for the same patient over time. *Among-series* irregularity means that different patients do not follow the same readmission schedule either — *in terms of both time intervals and number of visits*. Times for visits can significantly vary based on doctors' evaluation of the condition, the availability of doctors and/or imaging facilities, and the patient's own preferences, among others. This feature defies the assumptions of most methods that require regular sampling or common sampling [54, 64].

(4) **Spatial Misalignment** is often seen in longitudinal medical images too. Indeed, it is almost impossible to enforce pixel-perfect alignment of images acquired at different visits. Luckily, this problem can be addressed by image registration without any compounding effect with the temporal sparsity or sampling irregularity issues. See Appendix E for an illustration of image registration.

Temporal sparsity, sampling irregularity, and spatial misalignment are illustrated in Figure S1. These properties and challenges listed above lead to **a fairly unique area of research** that is **largely underexplored but highly interesting** to healthcare professionals.

Consider retinal imaging as an example. Most existing approaches to estimate disease progression in retinal imaging data do not operate in the image space, but rather in a vector space of a few clinical features extracted from the images. Examples of these derived statistics include the area of geographic atrophy lesions [65], the number of lesions [66], the lesion perimeter [65], its prior

observed growth rate [66], the presence and pattern of hyperfluorescence around the border of a lesion [67]. Although these approaches have been effective, they compress the rich context in the images to just a few metrics, and the output is an oversimplified representation of the disease states. This simplification overlooks the nuanced variations and complexities that are discarded during the feature extraction process and limits the interpretability of the output to a few preselected scalar-valued features.

In contrast, our proposed ImageFlowNet capitalizes on the extensive information available in the image to provide a nuanced representation of future conditions and also addresses the limitations of traditional metrics-based methodologies by offering a more dynamic and detailed visualization of disease progression. This method gives healthcare professionals an intuitive understanding of the expected progression of the disease and allows them to provide patients with a visual forecast that goes beyond mere numerical data.

We hope that our method can establish a new standard in the discipline and potentially transform clinical practices in areas including but not limited to ophthalmology or neurology, with the help of the latest imaging and measurement techniques [68, 69] as well as computational tools for disease diagnosis [70–72], risk prediction [73, 74], uncertainty quantification [75, 76], planning [77–80], and patient care [81–83].

D Additional Background on Why Time-Awareness is Important

Solving our problem outlined in Section 2 with deep learning requires designing and optimizing a model $\mathcal{F} : (\mathbb{R}^{H\times W\times C}, \mathbb{R}, \mathbb{R}) \to \mathbb{R}^{H\times W\times C}$, such that $\widehat{x_j} = \mathcal{F}(x_i, t_i, t_j)$ and $\widehat{x_j} \approx x_j$.

In most existing image-to-image tasks, the mapping between each pair of input x_i and output x_j obeys the same transformation rules, and hence their models are designed to be time-agnostic. For example, in denoising [84, 85], x_j is the noise-free version of x_i; in super-resolution [86–89], x_j is higher in resolution than x_i by a fixed factor; in reconstruction [90–95], x_j is the transformed version of x_i through a fixed set of rules guided by physics; in contrast mapping [96–100], x_j represents the effect of staining or contrast agents when applied to x_i; and in segmentation [101–107], x_j returns a label map describing the anatomical or functional segments in x_i. For these purposes, time-agnostic models, such as UNet or most diffusion models[1] remain competitive.

However, in our scenario, the output image is a function of both the input image and time. Given the same input image x_i, it will not end up at the same output image if the time interval changes. An image showing a disease 2 years after onset may look very different compared to 2 days after onset. In such cases, attempting to solve this problem using a model without time-modeling capabilities would be fundamentally ill-posed. In short, the spatial-temporal problem requires a spatial-temporal solution, which inspired our development of ImageFlowNet.

[1] While diffusion models have modules that can encode time, many variants are used in a time-agnostic manner for tasks like denoising or super-resolution, where "time" is no different from "iteration".

E Image Registration

E.1 Retinal Images

For all images, we extracted descriptive keypoints with SuperRetina [108], a high-quality keypoint detector trained on retinal images. Then we identified the keypoint correspondences for each image pair in each longitudinal series with a k-nearest-neighbor matcher and considered any image pair that has at least 15 keypoint correspondences a successful match. Next, we selected the image that produced the most successful matches as the "anchor image". Finally, we aligned all images in the longitudinal series towards the "anchor image" using perspective transformation so that the degree of freedom is constrained to the adjustment of camera angle or position. As a post-processing step, for each longitudinal series, we cropped all images with the biggest common foreground square so that no image contained any background pixel outside the retina region.

The image registration process for a pair of images from the same longitudinal series is illustrated in Figure S2. It can be seen that all veins are aligned in the resulting images while atrophy borders are not. This is expected from perspective transformation and is exactly desirable for our task.

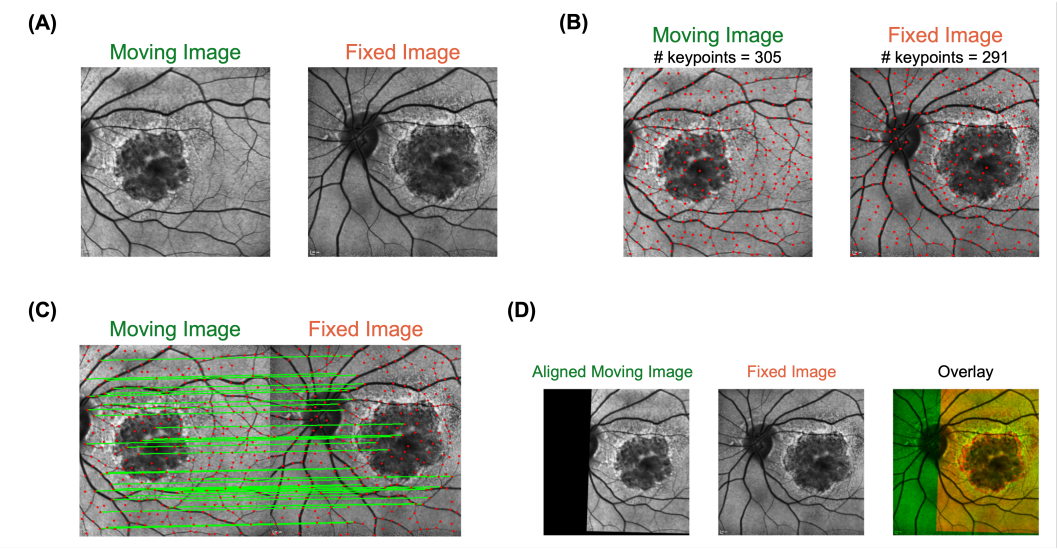

Figure S2: Our image registration pipeline. (A) Moving and fixed images come from the same eye at different time points. (B) SuperRetina is used to detect consistent and descriptive keypoints. (C) Keypoints are matched by descriptor similarity and filtered by distance heuristics. (D) The moving image is aligned under the constraint of a perspective transformation.

E.2 Brain Multiple Sclerosis Images

These images were already registered. No additional work was done.

E.3 Brain Glioblastoma Images

We used the scans in the "DeepBraTumIA" folders, which were registered to a common atlas, but the registration did not adequately align the scans in each longitudinal series. We used the Python tool from ANTS [109] to perform *Affine* followed by *Diffeomorphic* registration with [4, 2, 1] iterations to align each scan towards the first scan in series.

F Implementation Details

Architectures The proposed ImageFlowNet combines UNet and Neural ODEs. The UNet model follows the time-conditional UNet implementation in Guided Diffusion [33]. Neural ODEs are implemented with torchdiffeq [111].

Data Augmentation We used the albumentations package [112] to perform flipping, shifting, scaling, rotation, random brightness, and random contrast. We also make the UNet training a denoising process by adding random Gaussian noise to the input.

Hyperparamters and training details All experiments were performed on a SLURM server, where each job was allocated either an NVIDIA A100 GPU, an NVIDIA A5000 GPU, or an NVIDIA RTX 3090 GPU. All jobs can be completed within 2-5 days on a single GPU with 8 CPU cores. T-Diffusion usually takes the longest to train. ImageFlowNet$_{\text{SDE}}$ variant may require a 40-GB GPU (sometime that will still hit an OOM error if running too many function evaluations in the SDE) while all other methods can be trained on a 20-GB GPU. Experiments shared the same set of hyperparameters: learning rate = 0.0001, batch size = 64, number of epochs = 120. Adam with decoupled weight decay (AdamW) [113] optimizer was used, along with a cosine annealing learning rate scheduler with linear warmup.

To accommodate the GPU VRAM limits, we used gradient aggregation to trade efficiency for space while achieving the desired effective batch size — we used an actual batch size of 1, scaled the loss by $\frac{1}{64}$, and updated the weights every 64 batches.

The code has been uploaded to GitHub and we will release it once the paper is accepted.

Training of the segmentation networks are described in the next section (Evaluation Metrics).

G Evaluation Metrics

The evaluation metrics cover image similarity, residual magnitude, and atrophy similarity.

Image similarity We measure the image similarity between the real future image x_j and the predicted future image $\widehat{x_j}$ using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). These two metrics are widely used in image-to-image tasks such as super-resolution, denoising, inpainting, etc.

PSNR is a normalized version of the mean squared error between two images that takes into account the dynamic range of the image data. The formula is given by Eqn (8).

$$\begin{aligned} \mathrm{PSNR}(x_a, x_b) &= 10 \log_{10} \left(\frac{R}{\mathrm{MSE}(x_a, x_b)} \right), \text{ where} \\ R &\text{ is the common dynamic range of the images} \\ \mathrm{MSE}(x_a, x_b) &= \frac{1}{H \times W} \sum_{h \in H, w \in W} ||x_a^{(h,w)} - x_b^{(h,w)}||^2 \end{aligned} \tag{8}$$

SSIM measures the similarity between two images by describing the perceived change in structural information. The formula is given by Eqn (9). We used the implementation in Scikit-image [114].

$$\begin{aligned} \mathrm{SSIM}(x_a, x_b) &= \frac{(2\mu_{x_a}\mu_{x_b} + c_1)(2\sigma_{x_a x_b} + c_2)}{(\mu_{x_a}^2 + \mu_{x_b}^2 + c_1)(\sigma_{x_a}^2 + \sigma_{x_b}^2 + c_2)}, \text{ where} \\ \mu_{x_a} &\text{ is the pixel sample mean of } x_a \\ \mu_{x_b} &\text{ is the pixel sample mean of } x_b \\ \sigma_{x_b}^2 &\text{ is the variance of } x_b \\ \sigma_{x_b}^2 &\text{ is the variance of } x_b \\ \sigma_{x_b x_b} &\text{ is the covariance of } x_a \text{ and } x_b \\ c_1 = (0.01R)^2&, c_2 = (0.03R)^2 \\ R &\text{ is the common dynamic range of the images} \end{aligned} \tag{9}$$

Residual magnitude We evaluated the magnitude of the residual maps $\widehat{x_j} - x_j$ using the mean average error (MAE) and the mean squared error (MSE).

Atrophy similarity We also want to emphasize the precise representation of the atrophy region. To this end, the simplest metric is the dice similarity coefficient (DSC) and Hausdorff distance (HD) of the binarized atrophy regions. DSC and HD between two binary masks X and Y are given by Eqn (10) and Eqn (11), respectively. For HD, we used the implementation in Scikit-image [114].

$$\mathrm{DSC}(X, Y) = \frac{|X \cap Y|}{|X| + |Y|} \tag{10}$$

$$\mathrm{HD}(X, Y) = \max \left\{ \sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y) \right\} \tag{11}$$

To perform atrophy segmentation, we separately trained three auxiliary image segmentation network on all images, one for each dataset. All retinal images have their atrophy regions labeled by ophthalmologists. All brain images have associated segmentation maps from the dataset providers. These segmentators that we trained have an nn-UNet [115] architecture and were trained with an AdamW [113] optimizer at an initial learning rate of 0.001 for 120 epochs. With these networks, we can segment the atrophy regions in both the real future image x_j and the predicted future image $\widehat{x_j}$. DSC and HD can be computed on the segmentation masks between each pair of interest.

Supplementary References

[55] Guillaume Huguet, Daniel Sumner Magruder, Alexander Tong, Oluwadamilola Fasina, Manik Kuchroo, Guy Wolf, and Smita Krishnaswamy. Manifold interpolating optimal-transport flows for trajectory inference. *Advances in neural information processing systems*, 35:29705–29718, 2022.

[56] Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free schrödinger bridges via score and flow matching. *arXiv preprint arXiv:2307.03672*, 2023.

[57] Trang Nguyen, Alexander Tong, Kanika Madan, Yoshua Bengio, and Dianbo Liu. Causal discovery in gene regulatory networks with gflownet: Towards scalability in large systems. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

[58] María Ramos Zapatero, Alexander Tong, James W Opzoomer, Rhianna O'Sullivan, Ferran Cardoso Rodriguez, Jahangir Sufi, Petra Vlckova, Callum Nattress, Xiao Qin, Jeroen Claus, et al. Trellis tree-based analysis reveals stromal regulation of patient-derived organoid drug responses. *Cell*, 186(25):5606–5619, 2023.

[59] Lazar Atanackovic, Alexander Tong, Bo Wang, Leo J Lee, Yoshua Bengio, and Jason S Hartford. Dyngfn: Towards bayesian inference of gene regulatory networks with gflownets. *Advances in Neural Information Processing Systems*, 36, 2024.

[60] Jialin Chen, Jan Eric Lenssen, Aosong Feng, Weihua Hu, Matthias Fey, Leandros Tassiulas, Jure Leskovec, and Rex Ying. From similarity to superiority: Channel clustering for time series forecasting. *arXiv preprint arXiv:2404.01340*, 2024.

[61] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[62] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[63] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022.

[64] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*, 2023.

[65] Maximilian Pfau, Moritz Lindner, Lukas Goerdt, Sarah Thiele, Jennifer Nadal, Matthias Schmid, Steffen Schmitz-Valckenberg, Srinivas R Sadda, Frank G Holz, Monika Fleckenstein, et al. Prognostic value of shape-descriptive factors for the progression of geographic atrophy secondary to age-related macular degeneration. *Retina*, 39(8):1527–1540, 2019.

[66] Liangbo L Shen, Mengyuan Sun, Aneesha Ahluwalia, Benjamin K Young, Michael M Park, and Lucian V Del Priore. Geographic atrophy growth is strongly related to lesion perimeter: unifying effects of lesion area, number, and circularity on growth. *Ophthalmology Retina*, 5(9):868–878, 2021.

[67] Liangbo L Shen, Mengyuan Sun, Aneesha Ahluwalia, Michael M Park, Benjamin K Young, and Lucian V Del Priore. Local progression kinetics of geographic atrophy depends upon the border location. *Investigative Ophthalmology & Visual Science*, 62(13):28–28, 2021.

[68] Shah Hussain, Iqra Mubeen, Niamat Ullah, Syed Shahab Ud Din Shah, Bakhtawar Abduljalil Khan, Muhammad Zahoor, Riaz Ullah, Farhat Ali Khan, and Mujeeb A Sultan. Modern diagnostic imaging technique applications and risk factors in the medical field: a review. *BioMed research international*, 2022(1):5164970, 2022.

[69] Chuqin Huang, Yanda Cheng, Wenhan Zheng, Robert W Bing, Huijuan Zhang, Isabel Komornicki, Linda M Harris, Praveen R Arany, Saptarshi Chakraborty, Qifa Zhou, et al. Dual-scan photoacoustic tomography for the imaging of vascular structure on foot. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2023.

[70] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[71] Yuanzhou Wei, Dan Zhang, Meiyan Gao, Yuanhao Tian, Ya He, Bolin Huang, and Changyang Zheng. Breast cancer prediction based on machine learning. *Journal of Software Engineering and Applications*, 16(8):348–360, 2023.

[72] Yuanzhou Wei, Dan Zhang, Meiyan Gao, Aliya Mulati, Changyang Zheng, and Bolin Huang. Skin cancer detection based on machine learning. *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, 3(2):72–86, 2024.

[73] Xinyu Dong, Rachel Wong, Weimin Lyu, Kayley Abell-Hart, Jianyuan Deng, Yinan Liu, Janos G Hajagos, Richard N Rosenthal, Chao Chen, and Fusheng Wang. An integrated lstm-heterorgnn model for interpretable opioid overdose risk prediction. *Artificial intelligence in medicine*, 135:102439, 2023.

[74] Davide Placido, Bo Yuan, Jessica X Hjaltelin, Chunlei Zheng, Amalie D Haue, Piotr J Chmura, Chen Yuan, Jihye Kim, Renato Umeton, Gregory Antell, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature medicine*, 29(5):1113–1122, 2023.

[75] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

[76] Danqi Liao, Chen Liu, Benjamin W Christensen, Alexander Tong, Guillaume Huguet, Guy Wolf, Maximilian Nickel, Ian Adelstein, and Smita Krishnaswamy. Assessing neural network representations during training using noise-resilient diffusion spectral entropy. In *2024 58th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2024.

[77] Karin T Kirchhoff, Bernard J Hammes, Karen A Kehl, Linda A Briggs, and Roger L Brown. Effect of a disease-specific planning intervention on surrogate understanding of patient goals for future medical treatment. *Journal of the American Geriatrics Society*, 58(7):1233–1240, 2010.

[78] Jingdi Chen, Tian Lan, and Carlee Joe-Wong. Rgmcomm: Return gap minimization via discrete communications in multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17327–17336, 2024.

[79] Caitao Zhan, Himanshu Gupta, and Mark Hillery. Optimizing initial state of detector sensors in quantum sensor networks. *ACM Transactions on Quantum Computing*, 5(2):1–25, 2024.

[80] Jingdi Chen and Tian Lan. Minimizing return gaps with discrete communications in decentralized pomdp. *arXiv preprint arXiv:2308.03358*, 2023.

[81] Cheng Jin, Heng Yu, Jia Ke, Peirong Ding, Yongju Yi, Xiaofeng Jiang, Xin Duan, Jinghua Tang, Daniel T Chang, Xiaojian Wu, et al. Predicting treatment response from longitudinal images using multi-task deep learning. *Nature communications*, 12(1):1851, 2021.

[82] Zongxing Xie, Hanrui Wang, Song Han, Elinor Schoenfeld, and Fan Ye. Deepvs: A deep learning approach for rf-based vital signs sensing. In *Proceedings of the 13th ACM international conference on bioinformatics, computational biology and health informatics*, pages 1–5, 2022.

[83] Zongxing Xie, Bing Zhou, Xi Cheng, Elinor Schoenfeld, and Fan Ye. Vitalhub: Robust, non-touch multi-user vital signs monitoring using depth camera-aided uwb. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 320–329. IEEE, 2021.

[84] Linwei Fan, Fan Zhang, Hui Fan, and Caiming Zhang. Brief review of image denoising techniques. *Visual Computing for Industry, Biomedicine, and Art*, 2(1):7, 2019.

[85] Chunwei Tian, Lunke Fei, Wenxian Zheng, Yong Xu, Wangmeng Zuo, and Chia-Wen Lin. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, 2020.

[86] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[87] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014.

[88] Boyang Wang, Fengyu Yang, Xihang Yu, Chao Zhang, and Hanbin Zhao. Apisr: Anime production inspired real-world anime super-resolution. *arXiv preprint arXiv:2403.01598*, 2024.

[89] Boyang Wang, Bowen Liu, Shiyu Liu, and Fengyu Yang. Vcisr: Blind single image super-resolution with video compression synthetic data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4302–4312, 2024.

[90] Yunmei Chen, Hongcheng Liu, Xiaojing Ye, and Qingchao Zhang. Learnable descent algorithm for nonsmooth nonconvex image reconstruction. *SIAM Journal on Imaging Sciences*, 14(4):1532–1564, 2021.

[91] Wanyu Bian, Albert Jang, and Fang Liu. Multi-task magnetic resonance imaging reconstruction using meta-learning. *arXiv preprint arXiv:2403.19966*, 2024.

[92] Matthew J Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, et al. Results of the 2020 fastmri challenge for machine learning mr image reconstruction. *IEEE transactions on medical imaging*, 40(9):2306–2317, 2021.

[93] Wanyu Bian, Albert Jang, and Fang Liu. Improving quantitative mri using self-supervised deep learning with model reinforcement: Demonstration for rapid t1 mapping. *Magnetic Resonance in Medicine*, 2024.

[94] Chi Ding, Qingchao Zhang, Ge Wang, Xiaojing Ye, and Yunmei Chen. Learned alternating minimization algorithm for dual-domain sparse-view ct reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 173–183. Springer, 2023.

[95] Wanyu Bian, Qingchao Zhang, Xiaojing Ye, and Yunmei Chen. A learnable variational model for joint multimodal mri reconstruction and synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 354–364. Springer, 2022.

[96] Chen Liu, Nanyan Zhu, Haoran Sun, Junhao Zhang, Xinyang Feng, Sabrina Gjerswold-Selleck, Dipika Sikka, Xuemin Zhu, Xueqing Liu, Tal Nuriel, et al. Deep learning of mri contrast enhancement for mapping cerebral blood volume from single-modal non-contrast scans of aging and alzheimer's disease brains. *Frontiers in Aging Neuroscience*, 14:923673, 2022.

[97] Jens Kleesiek, Jan Nikolas Morshuis, Fabian Isensee, Katerina Deike-Hofmann, Daniel Paech, Philipp Kickingereder, Ullrich Köthe, Carsten Rother, Michael Forsting, Wolfgang Wick, et al. Can virtual contrast enhancement in brain mri replace gadolinium?: a feasibility study. *Investigative radiology*, 54(10):653–660, 2019.

[98] Xueshen Li, Hongshan Liu, Xiaoyu Song, Brigitta C Brott, Silvio H Litovsky, and Yu Gan. Generating virtual histology staining of human coronary oct images using transformer-based neural network. In *Diagnostic and Therapeutic Applications of Light in Cardiology 2024*, volume 12819, page 1281903. SPIE, 2024.

[99] Nanyan Zhu, Chen Liu, Xinyang Feng, Dipika Sikka, Sabrina Gjerswold-Selleck, Scott A Small, and Jia Guo. Deep learning identifies neuroimaging signatures of alzheimer's disease using structural and synthesized functional mri data. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 216–220. IEEE, 2021.

[100] Haoran Sun, Xueqing Liu, Xinyang Feng, Chen Liu, Nanyan Zhu, Sabrina J Gjerswold-Selleck, Hong-Jian Wei, Pavan S Upadhyayula, Angeliki Mela, Cheng-Chia Wu, et al. Substituting gadolinium in brain mri using deepcontrast. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 908–912. IEEE, 2020.

[101] Wenchao Zhang, Chong Fu, Yu Zheng, Fangyuan Zhang, Yanli Zhao, and Chiu-Wing Sham. Hsnet: A hybrid semantic network for polyp segmentation. *Computers in biology and medicine*, 150:106173, 2022.

[102] Nanyan Zhu, Chen Liu, Britney Forsyth, Zakary S Singer, Andrew F Laine, Tal Danino, and Jia Guo. Segmentation with residual attention u-net and an edge-enhancement approach preserves cell shape features. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2115–2118. IEEE, 2022.

[103] Zishun Feng, Dong Nie, Li Wang, and Dinggang Shen. Semi-supervised learning for pelvic mr image segmentation based on multi-task residual fully convolutional networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 885–888. IEEE, 2018.

[104] Haoyu Xie, Chong Fu, Xu Zheng, Yu Zheng, Chiu-Wing Sham, and Xingwei Wang. Adversarial co-training for semantic segmentation over medical images. *Computers in biology and medicine*, 157:106736, 2023.

[105] Ziyan Li, Jianjiang Feng, Zishun Feng, Yunqiang An, Yang Gao, Bin Lu, and Jie Zhou. Lumen segmentation of aortic dissection with cascaded convolutional network. In *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9*, pages 122–130. Springer, 2019.

[106] Chen Liu, Matthew Amodio, Liangbo L Shen, Feng Gao, Arman Avesta, Sanjay Aneja, Jay C Wang, Lucian V Del Priore, and Smita Krishnaswamy. Cuts: A deep learning and topological framework for multigranular unsupervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024.

[107] Honghui Liu, Jianjiang Feng, Zishun Feng, Jiwen Lu, and Jie Zhou. Left atrium segmentation in ct volumes with fully convolutional networks. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 39–46. Springer, 2017.

[108] Jiazhen Liu, Xirong Li, Qijie Wei, Jie Xu, and Dayong Ding. Semi-supervised keypoint detector and descriptor for retinal image matching. In *European Conference on Computer Vision*, pages 593–609. Springer, 2022.

[109] Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ants). *Insight j*, 2(365):1–35, 2009.

[110] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[111] Ricky TQ Chen. torchdiffeq. *URL https://github. com/rtqichen/torchdiffeq*, 124, 2018.

[112] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[113] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[114] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[115] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.