Learning to Compare Longitudinal Images

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Longitudinal studies, where a series of images from the same set of individuals are acquired at different time points, represent a popular technique for studying and characterizing temporal dynamics in biomedical applications, including in the healthcare setting. The classical approach for longitudinal comparison involves normalizing for nuisance variations, such as image orientation or contrast differences, via pre-processing. A statistical analysis is, in turn, conducted to detect changes of interest, either at the individual or population level. This classical approach can suffer from pre-processing issues and limitations of the statistical analysis. For example, normalizing for nuisance variation might be hard in settings where there is a lot of idiosyncratic changes. Population-level statistical analyses often yield a single summary of longitudinal changes that might not capture the full diversity across individuals. Individual-level statistical analyses, on the other hand, do not typically leverage the population-level data. In this paper, we present a simple machine learning-based approach that can alleviate these issues. In our approach, we train a deep learning model (called PaIRNet, for Pairwise Image Ranking Network) to compare pairs of longitudinal images, with or without supervision. In the self-supervised setup, for instance, the model is trained to temporally order the images, which requires learning to recognize time-irreversible changes Our results from four datasets demonstrate that PaIRNet can be very effective in localizing and quantifying meaningful longitudinal changes, while discounting nuisance variation.

Keywords: Self-supervised Learning, Longitudinal Analysis, Pairwise Comparison

1. Introduction

Whether it is in the context of development, aging, or progressive diseases, characterizing and quantifying change over time is a fundamental goal in bio-medicine. The longitudinal paradigm, where individuals are observed over multiple time-points, is the primary approach for studying temporal dynamics. In medical imaging, longitudinal studies are widely adopted for capturing anatomical and physiological changes.

There are two scenarios where longitudinal imaging is mainly utilized. The first scenario is to localize and quantify change at an individual level, for example, in a brain cancer patient where the objective is to monitor tumor growth and spread, as in (Konukoglu et al., 2007). In the second scenario, researchers are interested in characterizing population-level longitudinal changes that are shared across individuals, with the objective of gaining insights into biological processes. Common examples of the population study include Alzheimer's disease and aging, as described in (Fox et al., 1996; Fjell et al., 2009).

In a longitudinal imaging study, whether it is at the individual or group-level, the core challenge is disentangling nuisance variation (e.g., due to differences in imaging parameters or artifacts) from meaningful (e.g. clinically significant) temporal changes (e.g., atrophy or

© 2023 Anonymous.

tumor growth). Virtually all existing algorithmic approaches rely on pre-processing the data to help with this core challenge. In pre-processing, the objective is to minimize nuisance variation. A common step in pre-processing, for instance, is image registration, where the longitudinal images are spatially normalized to some standard reference coordinate frame. The image registration step can account for orientation differences and/or other geometric changes that can be confounding - such as distortion. Even after pre-processing, there is likely residual nuisance variation that will need to be discounted and this is where statistical or machine learning-based techniques are often used. As we discuss below, these analyses can be constrained by modeling choices. For example, population-level statistical methods are not designed for characterizing longitudinal changes at the individual-level.

In this paper, we present a simple learning-based approach as a novel way for analyzing longitudinal imaging data. Crucially, our approach does not require pre-processing, is model-free, and leverages population-level data to capture time-irreversible changes that are shared across individuals, while offering the ability to visualize individual-level changes. Our core assumption is that nuisance variation, such as changes in orientation or image contrast, are independent of time, when examined across the population. We train a deep learning model, called Pairwise Image Ranking Network (PaIRNet), to compare pairs of longitudinal images. We consider two possible tasks. In the first task, which we refer to as *supervised*, PaIRNet is trained to predict a target variable that captures meaningful temporal change - such as tumor volume difference. In the second task, which we call *self-supervised*, the model is trained to predict the temporal ordering of the input longitudinal image pair. In the self-supervised setting, the only additional information we need is the timing of the images, which is often readily available. In our experiments, we demonstrate that the proposed strategy can localize and quantify individual-level longitudinal changes, even in the self-supervised setting.

2. Related Works

2.1. Individual-level Longitudinal Change Detection

A core objective of a longitudinal study, particularly in the healthcare setting, is to localize and quantify clinically significant, non-spurious changes over time at the patient level. Classical approaches, such as (Smith et al., 2001) rely on pre-processing steps such as image registration, as in (Reuter et al., 2012), intensity normalization, and/or segmentation, as in (Smith, 2002). After these pre-processing steps, a simple strategy might be just to overlay the images to reveal pixel-level changes or study differences in morpholigcal measurements, such as size of an anatomical structure. However, these differences can still contain confounding changes, due to image artifacts or failures in pre-processing steps. One can rely on statistical analyses to separate out spurious and clinically-significant changes, as demonstrated in (Nguyen et al., 2018). An important weakness in these individual-level analyses is that they often do not leverage population data to better characterize and detect longitudinal change. Model-based methods mitigate this weakness by adopting mechanisms that capture population and individual-level dynamics (Sophocleous et al., 2022), yet these are limited by the model's capacity to capture the signal.

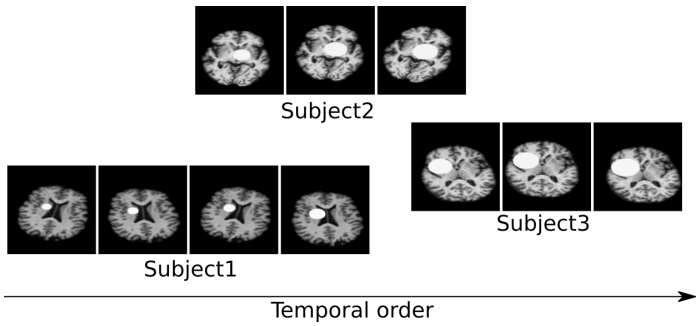

Figure 1: How would you characterize clinically meaningful longitudinal change that is common across these cases? Note the nuisance variation due to changes in orientation and contrast. There is also significant variability across individual brains that can be confounding.

2.2. Statistical Analysis of Population-level Longitudinal Change

Another objective in population health is to characterize temporal changes due to processes like development, aging, or disease progression, that is shared across individuals. These population-level longitudinal studies, by construction, pool data across a group of individuals to reveal a shared map of temporal change. These studies also demand pre-processing steps to suppress nuisance variation across the images. The classic approach, in turn, relies on statistical techniques like linear mixed effects models (Bernal-Rusiel et al., 2013) to reveal non-spurious temporal associations, as demonstrated in (Huang et al., 2022; Wang and Guo, 2019; Peters et al., 2019; Durrleman et al., 2013).

Although widely adopted, population-level statistical analyses for studying longitudinal change have at least two challenges. Pre-processing and extracting measurements often require human input (annotations, quality control, etc), which can be costly and hard to achieve at scale. It can also be difficult to decide what time-varying variable to examine associations with. For example, chronological age is a common variable for studying aging-associated changes. Yet, the aging process can vary significantly across individuals, compromising the interpretation of the results.

2.3. Machine Learning Methods for Longitudinal Images

In recent years, machine learning based methods have been used for longitudinal medical imaging studies. One approach leverages the supervised paradigm and trains a model to predict clinical trajectories from pairs of images, as described in (Bhagwat et al., 2018). To derive the target label to train on, these techniques require meta-data, such as disease severity scores, which can be unreliable or unavailable.

An alternative learning technique is the so-called self-supervised paradigm, which does not require ground-truth labels. This approach has recently been applied to longitudinal imaging studies (Zhao et al., 2021; Ouyang et al., 2022; Couronné et al., 2021), where the temporal ordering is used to regularize the learning of representations extracted from the

image data, which is trained to capture inter-subject variability and within-subject temporal changes. These techniques rely on a reconstruction loss for regularization, which can make them sensitive to nuisance variation in the images, such as orientation or contrast changes over time. Thus these methods require pre-processing of the image data and cannot handle misaligned images. In contrast, the learning paradigm we propose in this paper is trained on a single task (supervised or self-supervised) and does not depend on any pre-processing. The task, which we implement a neural network to solve, involves comparing pairs of images from the same individual. In the self-supervised implementation, the model learns to rank the input image pair based on temporal ordering.

2.4. Pairwise Image Ranking

Pairwise image ranking is an important problem in computer vision with many applications, such as assessing aesthetic preference (Lee and Kim, 2019), skill (Doughty et al., 2018), perceptual image-error (Prashnani et al., 2018), and street safety (Dubey et al., 2016). Most of these prior works use a shared feature extraction back-bone (also called a Siamese architecture), In this work, we apply a similar model, for the first time, to longitudinal medical images and demonstrate how this approach can yield individual-level detection and quantification of temporal change.

3. Detecting Changes in Longitudinal Images

3.1. Problem Definition

Ranking (or ordering) is a system where items are ordered and classified as either higher or lower or equal to the second item for any two items. Motivated by longitudinal self-supervised learning problems, the attribute is temporal order in this paper(Agrawal et al., 2022; Zhao et al., 2021; Couronné et al., 2021).

Longitudinal study design involves repeated measures of the same variables at different time points. Each variable ($s = 1, ..., n$), e.g., subject, has multiple images I_{si}, $i = 1, ..., m$, acquired at different time t_{si}. Given a set of image pairs (I_{si}, I_{sj}), $i, j = 1, ..., m$, and temporal ordering of the pairs, our goal is to train the model to detect longitudinal changes while learning to find the correct order of time. We formulate this temporal ordering problem as a ranking problem based on the seminar work of RankNet (Burges et al., 2005) and its variant, DirectRanker (Köppel et al., 2019).

$$R(I_{si}, I_{sj}) = \begin{cases} 1 & \text{if } t_{si} < t_{sj} \\ 0 & \text{if } t_{si} = t_{sj} \\ -1 & \text{if } t_{si} > t_{sj} \end{cases} \tag{1}$$

3.2. Pairwise Image Ranking Network

Pairwise Image Ranking Network (PaIRNet, R) consists of a feature extracting network (f), feature fusion (subtraction), and ranking network (g) as depicted in Figure 2. Given an image pair with corresponding labels that can be any quantitative values or preference, PaIRNet is trained to predict the order. A ranking function R satisfies the three properties:

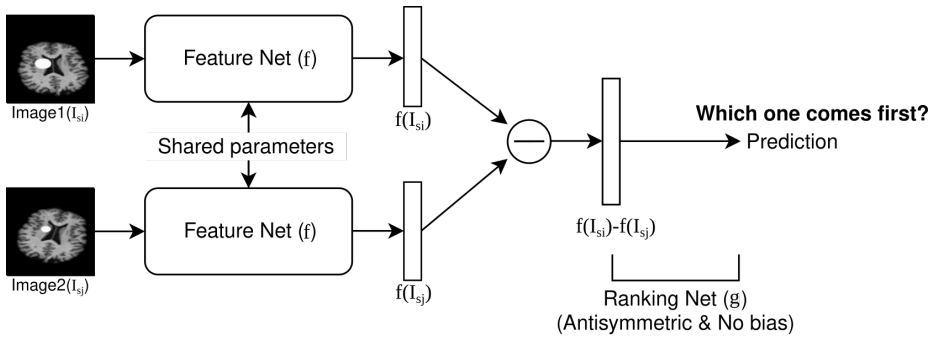

Figure 2: Pairwise Image Ranking Network (PaIRNet) Structure. PaIRNet consists of a feature network, subtraction, and ranking network. Given an image pair, PaIRNet predicts the rank order following Equation (1).

(a) Reflexivity: if $t_{si} = t_{sj}$, then $R(I_{si}, I_{sj}) = 0$

(b) Antisymmetry: $R(I_{si}, I_{sj}) = -R(I_{sj}, I_{si})$

(c) Transitivity: if $R(I_{si}, I_{sj}) \geq 0$ and $R(I_{sj}, I_{sk}) \geq 0$, then $R(I_{si}, I_{sk}) \geq 0$

where I_{si} is an input image with the temporal rank t_{si} of subject s.

We can show that PaIRNet satisfies the above properties. (a) is guaranteed by construction (Equation (1)). As shown in (Köppel et al., 2019; Rigutini et al., 2008), the feature differences of two images from the same feature network guarantee the antisymmetry when biases are removed and antisymmetric activation functions are used. (b) is easily satisfied when g is a linear neural network without biases and with antisymmetric activation functions, like tanh and softshrink. If (b) is satisfied, (a) is also fulfilled as $r(I_1, I_1) = -r(I_1, I_1)$, then $r(I_1, I_1) = 0$. For (c), as g is an odd linear function where the sign of input is preserved, if $r(I1, I2) = g(f(I1) - f(I2)) > 0$ and $r(I2, I3) = g(f(I2) - f(I3)) \geq 0$, then $r(I1, I2) + r(I2, I3) = g(f(I1) - f(I2) + f(I2) - f(I3)) = r(I1, I3) > 0$.

3.3. Localization of longitudinal change

Class activation maps (CAMs) and improved methods like Grad-CAMs calculated from classification CNNs have been used for weakly-supervised object localization. We modified CAMs, which visualize CNN predictions using the global average pooling (GAP) and weights from the GAP output to output classes (Zhou et al., 2016), to localize the difference between a pair of images. The CAM from the original paper is a weighted sum of the feature maps of the last convolutional layer using the weights predicting the class. As PaIRNet learns to compare two image inputs by negating a pair of features, we additionally used the absolute value of feature difference and the absolute weights of the last layer together and calculated the weighted sum of activation maps as follows:

$$\sum f_{activation} \times |w_g(f(I_{si}) - f(I_{sj})| \tag{2}$$

where w_g is a weight of the ranking layer and $f_{activation}$ is activation maps before GAP.

4. Experiments

4.1. Experimental setup

4.1.1. Dataset

Two synthetic datasets and two real-world datasets are used to validate the PaIRNet method. Two series of samples for each dataset are illustrated in Figure 1. All four datasets have sets of images that temporal change is monotonic and irreversible. We used 0.6/0.2/0.2 ratio for the train/validation/test split except for the Starmen. The split is done at the subject level.

Starmen. We used publicly available Starmen datasets that have been used for evaluating longitudinal frameworks (Bône et al., 2018; Couronné et al., 2021). Raising the left arm is the only temporal change within a subject in the original dataset. We randomly rotated (-6.8,6.8) degrees and translated (-10,10) pixels for subject variability. Out of 1000 sets of 10 longitudinal images, 400 sets of images are used for training, 100 sets for validation, and the rest for testing. The ground-truth value related to the progression of the left arm is ranged in -0.12 ± 4.25.

Tumor. Synthetic tumor growth dataset is generated based on T1 images of OASIS-2 dataset collected to study dementia (Marcus et al., 2010). To avoid possible capture of common degeneration from demented cohorts, we only selected first-visit images from 72 healthy cohorts. The images are preprocessed (brain extraction and intensity normalization) using the in-house software. We generated 3 to 5-time points of growing tumor placed within the brain tissue on the middle slice. The color, the initial size of the tumor, and the speed of the tumor growth rate are randomly generated. In total, 292 synthetic images with an average tumor size of 14.74 ± 6.86 are generated. All images are randomly rotated and translated for [-10, 10) degrees/pixels.

Embryo. Embryo development dataset is images of morphokinetics In Vitro Fertilization (IVF) collected using time-lapse imaging incubators (TLI)(Gomez et al., 2022). The images of developing embryos are fully annotated with 16 development phases. Embryos with more than one phase are used. We only selected one image in the middle if there are multiple images per phase. In total, 7784 images from 698 embryos are used. The embryos have on average 12.23 ± 2.41 phases.

Aging. OASIS-3 is a longitudinal dataset collected for normal aging and Alzheimer's disease (LaMontagne et al., 2019). We excluded Alzheimer's patients to focus on the aging effect. In total, 754 T1 images from 272 healthy cohorts ranging in age from $42 - 86$ (64.34 ± 8.53 on average) years at the initial scan are used. The subjects have 2.77 ± 0.96 repeated scans on average. All images are linearly aligned before extracting the middle slice.

4.1.2. Baseline methods

We compared the proposed method with the state-of-the-art longitudinal self-supervised learning methods. LSSL with cosine loss (LSSL-CL) (Zhao et al., 2021) and longitudinal variational autoencoder (LVAE) (Couronné et al., 2021) methods showed successful results

on learning latent representation related to changes in time for synthetic data and brain aging task. They both use image reconstruction loss of an autoencoder style architecture and combine with additional loss to learn temporal change. LSSL-CL method optimizes the colinearity between temporal trajectories obtained from latent vectors representing feature differences of input pairs and a global direction using cosine similarity. LVAE method optimizes two encoders, one for subject-wise variability with DeepSet method to attain permutation invariance (Zaheer et al., 2017) and one for longitudinal changes with soft-ranking loss (Blondel et al., 2020). We will use the same notation of their paper for latent features (z for LSSL-CL and Z^{Ψ} for LVAE).

When the ground-truth temporal attribute for the longitudinal change is given, a weakly-supervised object localization (WSOL) method can be another solution that can avoid learning subject-wise nuisance factors. For a single instance, the CAM-based method that visualizes the activation map for the class label has been a popular method (Zhou et al., 2016). Two supervised methods are used for the comparison: WSOL trained with supervised regression (SR) without considering the longitudinal components and PaIRNet with supervision.

4.1.3. Implementation details

For the feature network of the PaIRNet, we used the ResNet-18 (He et al., 2016). PaIRNet is trained with cross-entropy loss, Adam optimizer (Kingma and Ba, 2014), and the early stopping criterion that is 5 epochs without a decrease of validation loss. We used a grid search for a learning rate. The details are in the Appx. A.

We used the same architecture and visualization method for the WSOL methods. For LSSL-CL (Zhao et al., 2021) and LVAE (Couronné et al., 2021), we followed the original paper and parameter settings.

4.2. Longitudinal Change Detection

The activation maps of PaIRNet and baseline methods are illustrated in Figure 3 for each dataset. The images are a subset of the longitudinal series and are represented in the temporal order (t0, t1, t2). Among the SOTA LSSL methods, we could only visualize LVAE as the individual image-level visualization for longitudinal change is not available for the LSSL-CL. We followed the original LVAE visualization method but we used the image-wise feature Z^{Ψ} instead of averaged feature. The color red and blue of the LVAE activation map represent positive and negative gradients. For the supervised methods, we used the same weighted CAM method as PaIRNet.

Figure 3 shows the successful performance of the PaIRNet method compared to the baseline methods. Although the WSOL using regression tasks can localize the changes well in general, the method struggles to generalize. For example, the method fails to localize when the tumor sizes are smaller than the average (Figure 3B t0 and t1 in Subject1, t0 in Subject 2), when the embryo images have a noisy background (Figure 3C t0 and t1 in Subject1 and t2 in Subject 2), when the embryo images are translated (Figure 3C t2 in Subject 1 and t2 in Subject2) and when learning the task is difficult (Figure 3D).

The LVAE method is designed to separate subject-wise variability from longitudinal changes by learning the distribution of the subset of longitudinal images while learning the

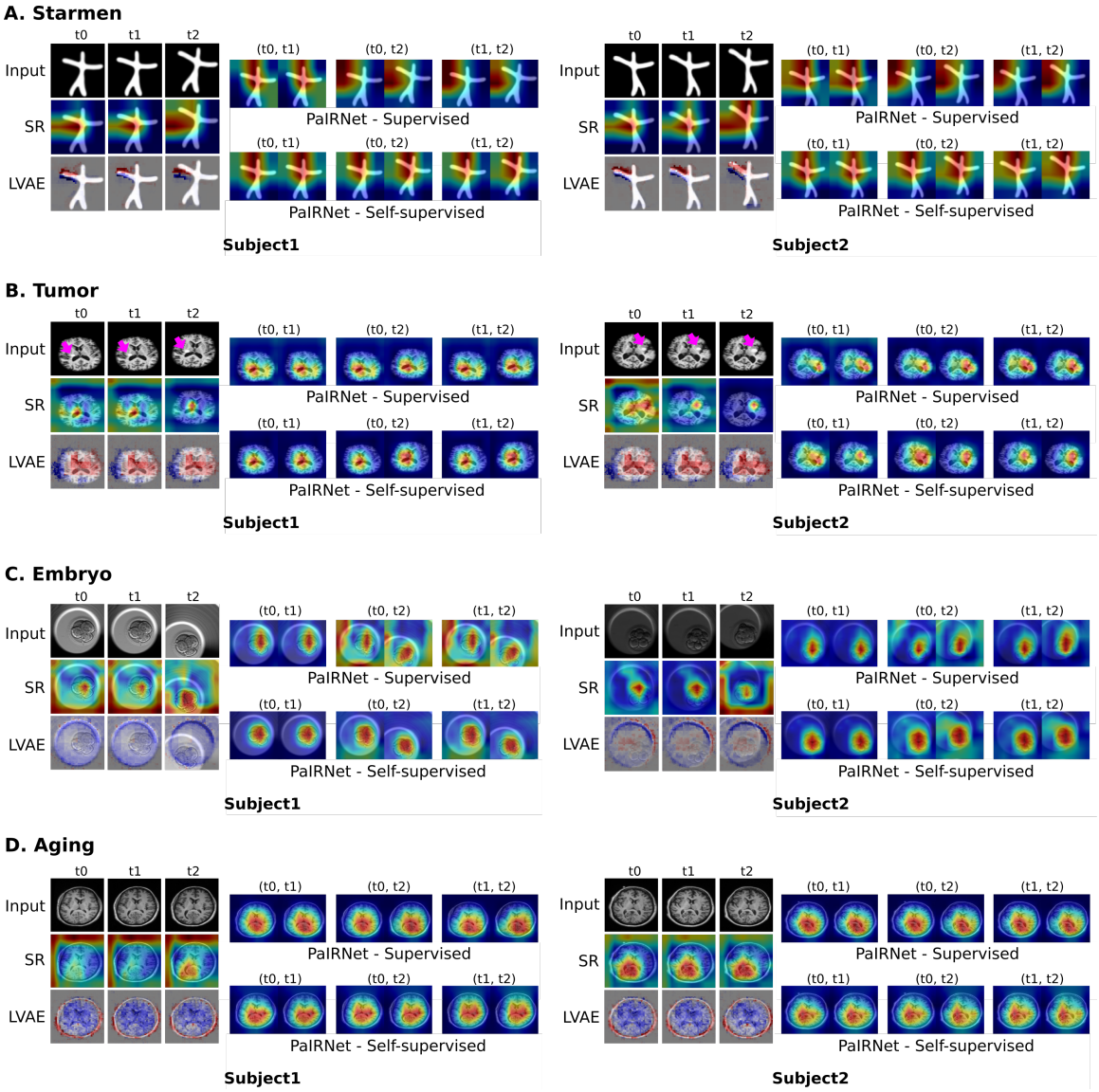

Figure 3: Activation map results. PaIRNet finds the similar activation patterns to supervised regression (SR) without using the ground-truth. Compared to the gradient maps of longitudinal variational autoencoder (LVAE), PaIRNet shows robust results in finding changes when the images are rotated or translated.

SoftRank of the images. Although successful in simple datasets, the method fails to separate and localize the longitudinal changes when the image has nuisance factors. For example, the factors in our experiments include image transformation in Starmen (Figure 3A t2) and Embryo (Figure 3C t2), image rotation in Tumor (Figure 3B t2) and brain with background, which includes CSF and skull, (Figure 3D). On the other hand, the PaIRNet methods could

learn to ignore the noisy factors by focusing on finding the difference between longitudinal pairs.

4.3. Correlation with ground-truth target label

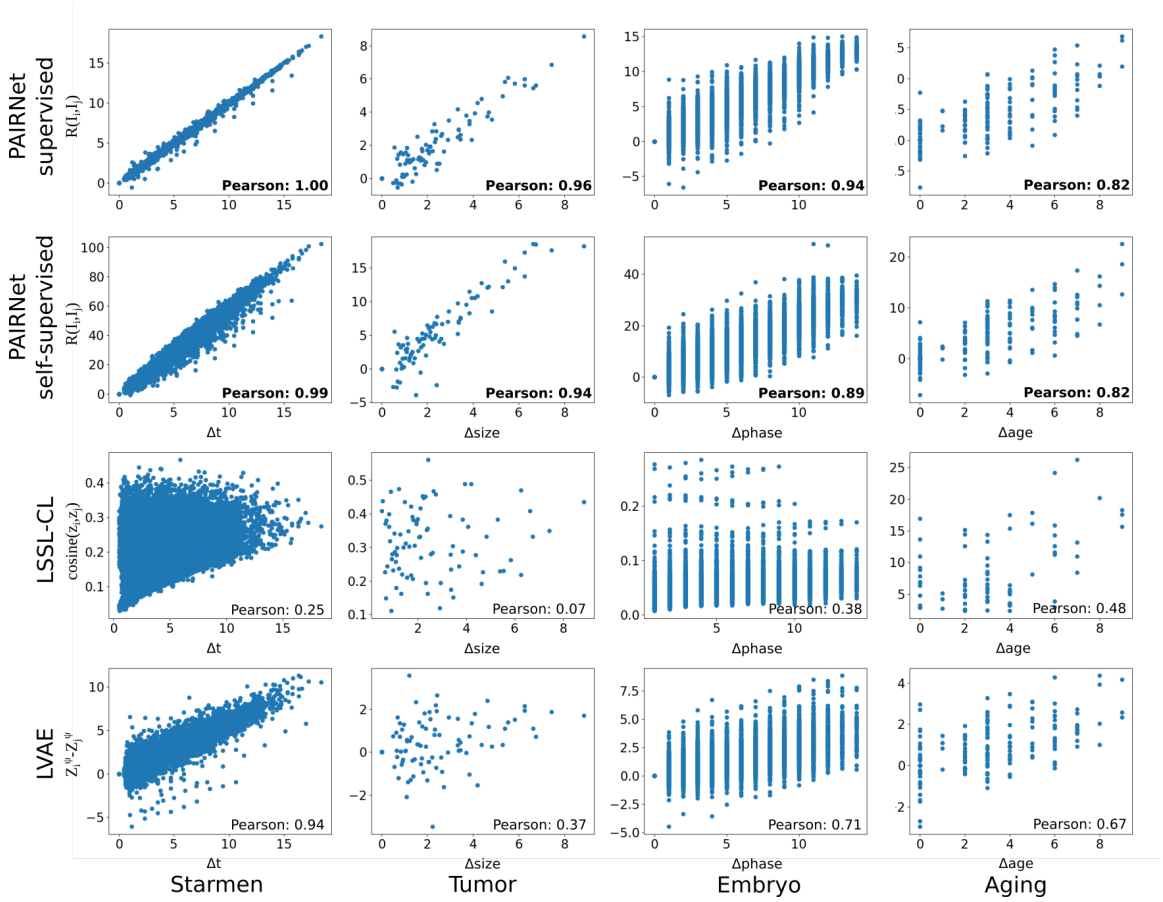

Figure 4: Correlation between ground-truth label and predicted values.

We evaluate the performance of detecting the change rate quantitatively. We calculated Pearson correlation coefficient(Benesty et al., 2009) to measure the linear correlation between the ground-truth temporal change and the estimated value. Only the non-negative pairs are used to avoid duplicates, e.g., $t_{si} <= t_{sj}$. As our model outputs the scalar value representing the difference of a given input pair, we used the network output before activation. For the SOTA LSSL baselines, we calculated the cosine similarity of a given pair, $cosine(z_i, z_j)$, for LSSL-CL and the difference, $Z_i^{\Psi} - Z_j^{\Psi}$, for LVAE.

In all four datasets, the proposed method shows a significantly higher correlation with the ground truth than LSSL-CL and LVAE. As the learning representation for longitudinal changes underperformed due to the nuisance factors, measuring the longitudinal difference between images would be affected. Self-supervised PaIRNet, contrariwise, learned to localize

and quantify the longitudinal change well. The self-supervised PaIRNet results are even comparable to the PaIRNet trained with supervision.

5. Conclusion

We have shown that pairwise image ranking network (PaIRNet) is an effective self-supervised method to capture important temporal changes by ranking the ordinal labels of longitudinal data. The proposed method find better representation to localize changes and the change rates while ignoring irrelevant subject-wise variations compared to other self-supervised longitudinal learning methods. However, some challenges exist. First, longitudinal changes are often not monotonic. Second, assuming a single temporal factor is unrealistic in the real-world scenarios e.g., numerous factors are involved in brain aging (Mattson and Arumugam, 2018). Third, class activation map (CAM)-based methods are known to have low-resolution visualization due to multiple convolution and pooling layers. Applying full resolution CAM (F-CAM) method (Belharbi et al., 2022) that adds an additional parametric decoder can improve the visualization. Tackling these challenges and improving the proposed method will be our future direction. We believe the simplicity of PaIRNet will make it practical for various longitudinal imaging analysis tasks.

References

Monica N Agrawal, Hunter Lang, Michael Offin, Lior Gazit, and David Sontag. Leveraging time irreversibility with order-contrastive pre-training. In *International Conference on Artificial Intelligence and Statistics*, pages 2330–2353. PMLR, 2022.

Soufiane Belharbi, Aydin Sarraf, Marco Pedersoli, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. F-cam: Full resolution class activation maps via guided parametric upscaling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3490–3499, 2022.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

Jorge L Bernal-Rusiel, Douglas N Greve, Martin Reuter, Bruce Fischl, Mert R Sabuncu, Alzheimer's Disease Neuroimaging Initiative, et al. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. *Neuroimage*, 66:249–260, 2013.

Nikhil Bhagwat, Joseph D Viviano, Aristotle N Voineskos, M Mallar Chakravarty, Alzheimer's Disease Neuroimaging Initiative, et al. Modeling and prediction of clinical symptom trajectories in alzheimer's disease using longitudinal data. *PLoS computational biology*, 14(9):e1006376, 2018.

Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959. PMLR, 2020.

Alexandre Bône, Olivier Colliot, and Stanley Durrleman. Learning distributions of shape trajectories from longitudinal datasets: a hierarchical model on a manifold of diffeomorphisms. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9271–9280, 2018.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

Raphaël Couronné, Paul Vernhet, and Stanley Durrleman. Longitudinal self-supervision to disentangle inter-patient variability from disease progression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 231–241. Springer, 2021.

Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6057–6066, 2018.

Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *European conference on computer vision*, pages 196–212. Springer, 2016.

Stanley Durrleman, Xavier Pennec, Alain Trouvé, José Braga, Guido Gerig, and Nicholas Ayache. Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. *International journal of computer vision*, 103(1):22–59, 2013.

Anders M Fjell, Lars T Westlye, Inge Amlien, Thomas Espeseth, Ivar Reinvang, Naftali Raz, Ingrid Agartz, David H Salat, Doug N Greve, Bruce Fischl, et al. High consistency of regional cortical thinning in aging across multiple samples. *Cerebral cortex*, 19(9): 2001–2012, 2009.

NC Fox, EK Warrington, PA Freeborough, P Hartikainen, AM Kennedy, JM Stevens, and Martin N Rossor. Presymptomatic hippocampal atrophy in alzheimer's disease: a longitudinal mri study. *Brain*, 119(6):2001–2007, 1996.

Tristan Gomez, Magalie Feyeux, Justine Boulant, Nicolas Normand, Laurent David, Perrine Paul-Gilloteaux, Thomas Fréour, and Harold Mouchère. A time-lapse embryo dataset for morphokinetic parameter prediction. *Data in Brief*, page 108258, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Chao Huang, Zhenlin Xu, Zhengyang Shen, Tianyou Luo, Tengfei Li, Daniel Nissman, Amanda Nelson, Yvonne Golightly, Marc Niethammer, and Hongtu Zhu. Dadp: dynamic abnormality detection and progression for longitudinal knee magnetic resonance images from the osteoarthritis initiative. *Medical Image Analysis*, 77:102343, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ender Konukoglu, Maxime Sermesant, Olivier Clatz, Jean-Marc Peyrat, Hervé Delingette, and Nicholas Ayache. A recursive anisotropic fast marching approach to reaction diffusion equation: Application to tumor growth modeling. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 687–699. Springer, 2007.

Marius Köppel, Alexander Segner, Martin Wagener, Lukas Pensel, Andreas Karwath, and Stefan Kramer. Pairwise learning to rank by neural networks revisited: Reconstruction, theoretical analysis and practical performance. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 237–252. Springer, 2019.

Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, 2019.

Jun-Tae Lee and Chang-Su Kim. Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1191–1200, 2019.

Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *Journal of cognitive neuroscience*, 22(12):2677–2684, 2010.

Mark P Mattson and Thiruma V Arumugam. Hallmarks of brain aging: adaptive and pathological modification by metabolic states. *Cell metabolism*, 27(6):1176–1199, 2018.

Thanh D Nguyen, Shun Zhang, Ajay Gupta, Yize Zhao, Susan A Gauthier, and Yi Wang. Fast and robust unsupervised identification of ms lesion change using the statistical detection of changes algorithm. *American Journal of Neuroradiology*, 39(5):830–833, 2018.

Jiahong Ouyang, Qingyu Zhao, Ehsan Adeli, Greg Zaharchuk, and Kilian M Pohl. Disentangling normal aging from severity of disease via weak supervision on longitudinal mri. *IEEE Transactions on Medical Imaging*, 2022.

Jurriaan M Peters, Anna Prohl, Kush Kapur, Audrey Nath, Benoit Scherrer, Sean Clancy, Sanjay P Prabhu, Mustafa Sahin, David Neal Franz, Simon K Warfield, et al. Longitudinal effects of everolimus on white matter diffusion in tuberous sclerosis complex. *Pediatric neurology*, 90:24–30, 2019.

Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018.

Martin Reuter, Nicholas J Schmansky, H Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4):1402–1418, 2012.

Leonardo Rigutini, Tiziano Papini, Marco Maggini, and Monica Bianchini. A neural network approach for learning object ranking. In *International Conference on Artificial Neural Networks*, pages 899–908. Springer, 2008.

Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3): 143–155, 2002.

Stephen M Smith, Nicola De Stefano, Mark Jenkinson, and Paul M Matthews. Normalized accurate measurement of longitudinal brain change. *Journal of computer assisted tomography*, 25(3):466–475, 2001.

Froso Sophocleous, Alexandre Bône, Andrew IU Shearn, Mari Nieves Velasco Forte, Jan L Bruse, Massimo Caputo, and Giovanni Biglino. Feasibility of a longitudinal statistical atlas model to study aortic growth in congenital heart disease. *Computers in Biology and Medicine*, 144:105326, 2022.

Yikai Wang and Ying Guo. A hierarchical independent component analysis model for longitudinal neuroimaging studies. *NeuroImage*, 189:380–400, 2019.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Qingyu Zhao, Zixuan Liu, Ehsan Adeli, and Kilian M Pohl. Longitudinal self-supervised learning. *Medical Image Analysis*, 71:102051, 2021.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Appendix A. Parameters?

For learning rate, we used grid search.

Starmen Augmentation diff-resnet18-1layer: 0.001 diff-delta-supervised: 0.01 Supervised: 0.01

Synth Tumor Diff-resnet18-1layer: 0.0001 diff-delta-supervised: 0.01 Supervised: 0.01

Embryo Diff-resnet18-1layer: 0.001 diff-delta-supervised: 0.001 Supervised: 0.001

Brain aging: Diff-resnet18-1layer: 0.001 diff-delta-supervised: 0.01 Supervised: 0.01

Appendix B. PaIRNet classification performance

To show that PaIRNet is an effective method for finding the temporal differences without ground-truth label, we compared the PaIRNet's ranking performance with supervised regression (SR). At inference of SR, the difference of estimated values are used to decide the ranking of a given pair. We did not use the same ranking pairs at testing time, which becomes a binary classification task. The AUC result in Table 1 shows the PaIRNet we can learn the order properly with high accuracy (AUC$>$ 0.95). Although mean square error

(MSE) of regression task (SR-MSE in Table 1 is reasonably low considering standard deviation of ground-truth labels, the ranking using the SR results in Tumor, Embryo and Aging datasets underperform compared to PaIRNet results. This demonstrates that pairwise comparison may be more suitable task for learning difference between images than than SR. The results from Aging dataset highlights this. From the high SR MSE, we can determine predicting age from images is more difficult tasks than others. However, PaIRNet AUC was 0.985, which is comparable to other datasets.

Table 1: AUC result of binary classification from testing sets. Pairwise Imaging Ranking Network (PaIRNet) outperforms supervised regression (SR) results. Although mean square error (MSE) loss of SR is low, the AUC was not as high as PaIRNet.

	Starmen	Tumor	Embryo	Aging
PaIRNet	**1.0**	**0.964**	**0.994**	**0.985**
SR	**1.0**	0.929	0.965	0.893
SR-MSE	0.017	1.239	1.355	40.215