Title

Serial Dependence Study in Medical Image Perception via Generative Models

Author

Ren, Zhihang

Publication Date

2024

Peer reviewed|Thesis/dissertation

Serial Dependence Study in Medical Image Perception via Generative Models

By

Zhihang Ren

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Vision Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David Whitney, Co-chair
Professor Stella X. Yu, Co-chair
Professor Bruno Olshausen
Professor Meng Lin

Spring 2024

Serial Dependence Study in Medical Image Perception via Generative Models

Abstract

Serial Dependence Study in Medical Image Perception via Generative Models

by

Zhihang Ren

Doctor of Philosophy in Vision Science

University of California, Berkeley

Professor David Whitney, Co-chair

Professor Stella X. Yu, Co-chair

Medical imaging has been critically important for the health and well-being of millions of patients. Although deep learning has been widely studied in the medical imaging area and the performance of deep learning has exceeded human performance in certain medical diagnostic tasks, detecting and diagnosing lesions still depends on the visual system of human observers (radiologists), who completed years of training to scrutinize anomalies. Routinely, radiologists sequentially read batches of medical images one after the other. A basic underlying assumption of radiologists' precise diagnosis is that their perceptions and decisions on a current medical image are completely independent of the previous reading history of medical images. However, recent research proposed that the human visual system has visual serial dependencies at many levels. Visual serial dependence means that what was seen in the past influences (and captures) what is seen and reported at this moment.

In this dissertation, we first show that visual serial dependence has a disruptive effect on radiological searches that impairs the accurate detection and recognition of tumors or other structures via naive artificial stimuli. However, the naive artificial stimuli have been noted by both untrained observers and expert radiologists to be less authentic, which can not help to reveal the real scenarios of medical image perception. To solve this issue, we propose and build a generative tool via generative adversarial networks (GANs) to generate authentic medical images, replacing the simple stimuli in future experiments. Using the authentic medical images from the GenAI medical image generation tool, we find that the perception of the current simulated medical image was biased towards the previously seen medical images, which strengthens the evidence of the existence of the visual serial dependence effect in medical image perception. Finally, we collaboratively collect real diagnostic data with a data annotation company. Through meticulous data analysis, we find significant serial dependence effects in perceptual discrimination judgments, which negatively impacted

performance measures, including sensitivity, specificity, and error rates. These findings help understand one potential source of systematic bias and errors in medical image perception tasks and hint at useful approaches that could alleviate the errors due to serial dependence.

Dedication

To my parents, advisors, Xushan, and friends who always support me!

Contents

List of Figures

List of Tables

Acknowledgments

Ever since I had the idea of pursuing a Ph.D. degree, I have pondered whom to acknowledge in my dissertation. Because the Ph.D. journey is a multifaceted experience, encompassing not only research papers but also the invaluable aspects of friendship, mentorship, personal fulfillment, and love. It is hard to believe that after five incredible years, I have eventually reached this milestone and am writing the acknowledgments now.

I would like to express my deepest gratitude to my advisors David and Stella. Dave, since I joined your lab, you have always provided us with ample freedom to explore our research directions while consistently offering your support and invaluable hands-on advice. I appreciate that you always have abundant ideas to share and discuss with us while providing insightful feedback. Additionally, the lab culture you established is brilliant. I love exchanging ideas and feedback during lab meetings and enjoy the weekly lab lunch. I met with Stella before joining her lab. I was very impressed by her unique way of thinking about research problems. Having Stella as my advisor is one of the most fortunate aspects of my Ph.D. journey at Berkeley. I appreciate that you set high standards for your students, which at times made me feel disappointed in myself. You always provide us with invaluable suggestions on experiment designs and research directions. Moreover, I am very grateful for you pushing us out of our comfort zones and helping us grow into better researchers. Nevertheless, in the end, I am very respectful of your hardworking spirit which will stimulate me in my future research journey. Wish you all the best in Michigan.

I would like to extend my heartfelt gratitude to my qualifying exam and dissertation committee members: Alexei Efros, Bruno Olshausen, and Meng Lin. Bruno Olshausen and Meng Lin taught my Vision Science fundamental courses. It is you who introduced me to new aspects of understanding computer vision problems. I was very interested in the Vision Science lectures you both provided.

I started the research path in my junior year. I was very lucky to be advised by Prof. Shuaicheng Liu. Prof. Shuaicheng Liu, I truly enjoy working with you on traditional computer vision tasks. You led me into the research area, taught me research skills, and guided me on how to write scientific papers. I can still remember that you patiently answered all the questions of my experiments though some of which were naive, and that we fought for paper deadlines near the Spring festival. Without you, I can not start my Ph.D. and research journey. I would also like to thank my mentors in Shuaicheng's group: Heng Guo and Tong He for your patient guidance in my research. I then moved to San Diego for a Master's Degree, where I fortunately worked with Prof. Nuno Vasconcelos and Prof. Bhaskar D. Rao. I would like to thank Prof. Nuno Vasconcelos for the invaluable statistical visual computing knowledge you have shared with me and your guidance on my research involving deep learning. I would also like to thank Prof. Bhaskar D. Rao. You offered me the opportunity to collaborate with Jacobs Medical Center at UC San Diego Health and discussed ideas in medical imaging applications for signal processing. Lastly, I would also love to thank my friends from Nuno's group: Yi Li, John Ho, Gina Wu, Pei Wang, Yunsheng Li, Bo Liu, Zhaowei Cai, Gautam Nain, and Anwesan Pal.

During my years at Berkeley, I have been so fortunate to meet so many wonderful friends and peers. I would like to thank my whole cohort: Victoria Cynthia Fong, Orneika Flandrin, Iona McLean, John (JT) Pirog, and Julie Self, for their help on my coursework; Friends from Stella's group: Daniel, Sascha Hornauer, Jyh-Jing Hwang, Tsung-Wei Ke, Zhongqi Miao, Ke Wang, Peter Wang, Frank Wang, Nils-Steffen Worzyk, Qian Yu, and Baladitya Yellapragada; Friends from David's group: Teresa Canas-Bajo, Zhimin Chen, Valerie Ekko, Haley Frey, Yifan Fang, Cristina Ghirardo, Jasmine Lopez, Mauro Manassi, Yuki Murai, Jefferson Ortega, Luna Ragot, and Zixuan Wang. I would also like to thank Prof. Ken Nakayama, Prof. William Prinzmetal (Bill), and Prof. Ervin Hafter for their insightful feedback on my research directions and presentations.

To my friend, Prof. Yunhui Guo, we met on the flight to San Diego and ended up in the same lab at Berkeley. I cherish the time we had meals and discussed research ideas together. I have learned a lot from your rigorous attitude towards research. I wish you all the best in your future endeavors in Texas.

Through my Ph.D., I have been fortunate to work and collaborate with numerous outstanding researchers. To my talented research assistants: Ana Hernandez Reyes, Rena (Xinyu) Li, Yifan Wang, Charlie Cheng-Jie Ji, Ethan Shedd, Rina (Qimei) Li, Wish Wang, and Yunqi Li, guiding you to start conducting research is a great experience for me. Your passion and hard work contribute so much to my Ph.D. milestone. Dr. Tapabrata Rohan Chakraborty, collaborating with you on medical imaging applications is a great pleasure to me and expands my vision on clinical usage of AI/ML methods. Dana Pietralla and Julien Vignoud, collaborating was never so smooth before meeting with both of you. We have done amazing work during your short visit to our lab.

During my Ph.D., I completed one internship at Meta Reality Labs. The internship gave me a different experience from academic life. I would like to express my sincere gratitude to Dr. Shahin Aghdam, my mentor at Meta Reality Labs. Working with you has been one of the most memorable industry experiences of my Ph.D. journey.

Shout out to my friends in China: Ruifeng Tang, Fengyi Song, Jianzhi Wang, Wanlun Ma, Xiaoyuan Hu, Song Qing, etc; my friends in the Bay Area: Zhengfeng Shi, Jing Ming, Mulin Yang, Wenyu (Harley) Liu, Jie Yin, Anqi Zhang, Yanjing Liang, Yixin Zou, Feihua Fang, Yifei Liu, Ruby Chen, etc. You are the best! Thank you for your companion whenever I was happy or sad!

To Wentao Wei, Hongli Zhang, Jiayi Li, and Mingxuan Sun, the outdoor adventures with you two couples are incredibly amazing! The scenery we saw while hiking and the mushrooms we collected made my Ph.D. life more colorful.

Special thanks to my two feline companions, Wanda and Crag, whose comforting presence and playful antics brought joy and solace during the long hours of writing. Your unconditional love and companionship made this journey more delightful.

I would like to express my deepest love and gratitude to my parents, Wenhua Zhou and Yongzhong Ren, whose unwavering love, support, and guidance have been the cornerstone of my journey. Your belief in me, even during the most challenging times, has been a source

of strength and inspiration. Your sacrifices, both seen and unseen, have paved the way for my success.

Lastly, I want to dedicate this spot to my wife Xushan. Thank you for your constant support and boundless love throughout the past two years. In the face of countless late nights, weekends sacrificed, and moments missed, you stood by me with unwavering support and encouragement. Your quiet strength sustained me through the challenges and celebrated with me in the triumphs. Thank you for your sacrifices, your understanding, and your unfaltering belief in me. This achievement is as much yours as it is mine. I'm excited and looking forward to our life together!

Zhihang Ren
May 3, 2024
Berkeley

Chapter 1

Introduction

1.1 Background

Medical imaging has transformed modern medicine, allowing clinicians to noninvasively examine and diagnose patients relatively quickly and easily. Cancer diagnosis via medical imaging is critically important for public health, but it is still far from perfect. For example, in breast cancer diagnosis with mammography, false negative and false positive rates have been reported 6% to 46% and approximately 10% respectively. These errors are mostly due to misperceptions and misinterpretations of x-ray images from radiologists [14, 50]. Whereas some sources of errors have been fully identified and characterized (subsequent search misses [24];low prevalence [68], etc.), errors in cancer image interpretation are still largely without explanation [14]. Given the importance of this issue, a great deal of research has been carried out in the last decades to understand how to identify and characterize the source of these mistakes in order to mitigate them as much as possible.

While interpreting mammograms, radiologists are typically asked to detect any present tumors and classify them as well as record their size, location, and type. Typically, radiologists will examine dozens to hundreds of mammograms one after the other in a short time period. A major underlying assumption is that radiologists' perceptions and decisions on a current mammogram are completely independent of previous perceptual history. However, recent theoretical and empirical research from our lab and others raises the possibility that this is not true. Our visual system is characterized by visual serial dependency, a type of sequential effect in which what has previously been seen influences (captures) what is seen and reported at this moment.

Visual serial dependence is a common sequential effect of our visual system and it can manifest in several domains, such as perception [75, 41, 173], decision making [1, 73], and memory [137], and they occur with a variety of features and objects, including orientation[75], position [173, 20], faces [161, 245, 246, 162], attractiveness [246, 282, 141], ambiguous objects [270], ensemble coding of orientation [175], and numerosity [41, 46]. Visual serial dependence is characterized by three main kinds of tuning. First, feature tuning: Visual serial dependence

occurs only between similar features [75, 84, 173, 175], and not between dissimilar ones[84]. Second, temporal tuning: Visual serial dependence gradually decays over time[75, 173, 270]. Third, spatial tuning: Visual serial dependence occurs only within a limited spatial region and it is strongest when previous and current objects are presented at the same location [75, 173, 20]. In addition, attention is a necessary component for Visual serial dependence [75].

The visual serial dependence benefits us because the world we live in and the scenes we experience are usually highly structured, autocorrelated, and stable. However, this situation is not always true. For visual search tasks in mammography, stimuli are not autocorrelated either. Thus, when doing tasks such as tumor localization and classification, serial dependence could introduce a bias in perceptual judgments which would bring in a reduction in sensitivity and increase in errors. The negative impacts of serial dependence in search tasks would be especially prominent in cases where there is low signal, high noise, high uncertainty, or where fine discriminations are required [75, 39, 41, 173, 20, 170]. These are exactly the challenging situations that radiologists routinely face when searching through scans.

This dissertation starts with our initial attempt via naive artificial stimuli. It shows that visual serial dependence has a disruptive effect in radiologic searches that impairs accurate detection and recognition of tumors or other structures[170]. However, this pilot project utilized artificially morphed tumors and simple noisy backgrounds as stimuli which have been noted by both naïve observers and expert radiologists to be less authentic. Recently, Generative Adversarial Networks (GANs) have been well developed to generate authentic images for certain categories, such as faces, cars, and landscapes [127, 130, 198]. It is intuitive to apply GANs to medical images to generate authentic stimuli for our experiments.

Generative Adversarial Networks are special Convolutional Neural Networks (CNNs), which consist of two networks, the generator(G) and the discriminator(D). These two networks are trained iteratively in an adversarial way where the generator(G) generates fake but authentic images to fool the discriminator and the discriminator(D) discriminates the real and fake images [95]. Using this promising computational model, high-quality images with various categories can be generated, such as faces, cars, and landscapes [127, 130, 198]. However, the initial GAN model [95] cannot generate sharp and recognizable images, and the training process is unstable. Later work improved the performance of GAN in different ways. Some papers focus on model architectures [183, 35, 194]. Others focus on improving the loss metrics and training strategies [98, 4, 26]. With these efforts, GAN training stability has improved, and GAN can generate low-resolution images with sufficient quality.

Most recently, several approaches have made high-resolution image generation also possible. PGGAN [130] proposed to train the standard GAN from coarse to fine scale. The parameters for low-resolution blocks are trained first. Then higher-resolution blocks are added on gradually with the corresponding parameters updated accordingly. Based on the same training strategy, StyleGAN [127, 128] proposed to first map the original latent space $\mathcal{Z}$ into the $\mathcal{W}$ space through a non-linear mapping network. Then it is merged into the synthesis network via adaptive instance normalization (AdaIN) at each convolutional block [57, 118]. This improves StyleGAN representations of scenes and details and allows it to produce authentic high-resolution images. In the following project, we adopt StyleGAN as our back-

bone to build a medical image generation tool. Moreover, a controllable approach is also utilized to manipulate the attributes of the generated images. Using the authentic medical images from the GenAI medical image generation tool, we find that the perception of the current simulated medical image was biased towards the previously seen medical images, which strengthens the evidence of the existence of the visual serial dependence effect in medical image perception. Meanwhile, we also utilize this GenAI medical image generation tool to augment the rare case image samples in skin cancer diagnosis and boost the classification performance of self-supervised learning models.

Finally, we collaboratively collect real diagnostic data with a data annotation company. Through meticulous data analysis, we find significant serial dependence effects in perceptual discrimination judgments, which negatively impacted performance measures, including sensitivity, specificity, and error rates. In this data, we also find that medical trainees have image-level idiosyncratic biases when they perform skin cancer diagnosis, and increased diagnostic proficiency is associated with more substantial idiosyncratic biases.

Overall, this dissertation introduces a series of projects aimed at understanding the serial dependence effect in medical image perception with the help of the proposed medical image generation tool via Generative Adversarial Networks (GANs).

1.2 Outline

The organization of this dissertation is presented as follows:

In Chapter 2, we utilize naive artificial medical image stimuli, revealing that visual serial dependence has a disruptive effect in radiologic searches that impairs accurate detection and recognition of tumors or other structures.

In Chapter 3, we introduce the generative tool via Generative Adversarial Networks (GANs). It allows users to controllably generate desired lesions in medical images. We show that this tool can generate authentic simulated medical image stimuli in many modalities.

In Chapter 4, we illustrate a side project where we utilize the proposed generative tool to augment the rare case image samples in skin cancer diagnosis and boost the classification performance of self-supervised learning models.

In Chapter 5, we use the authentic medical images from the GenAI medical image generation tool to design experiments and find that the perception of the current simulated medical image was biased towards the previously seen medical images.

In Chapter 6, we analyze real diagnostic data collaboratively collected with a data annotation company. We find significant serial dependence effects in perceptual discrimination judgments, which negatively impacted performance measures, including sensitivity, specificity, and error rates.

In Chapter 7, we visualize the findings of the same diagnostic data that medical trainees have image-level idiosyncratic biases when they perform skin cancer diagnosis, and increased diagnostic proficiency is associated with more substantial idiosyncratic biases.

In Chapter 8, we provide a summary of the projects introduced throughout this dissertation.

Chapter 2

Serial Dependence in the Perceptual Judgments of Radiologists

2.1 Introduction

Cancer diagnosis in medical images is crucial for the health of millions of people, but it is still far from perfect. For example, within mammography, false negative and false positive rates have been reported to be 0.15% and 9%, respectively [190]. Some of these misdiagnoses are due to misperceptions and misinterpretations of radiographs by clinicians [14, 50]. Interpretive errors in radiology are defined as the discrepancy in interpretation between the radiologist and peer consensus [27, 260], and it has been proposed that perceptual errors account for 60–80% of the total amount [87, 134].

Some sources of interpretive error have been identified and characterized, including search and recognition errors [30, 193], cognitive biases [50, 158], search satisfaction [6, 12], subsequent search misses [16, 24, 105], and low prevalence [279, 280, 218, 181, 68, 114, 149]. However, some other errors in cancer image interpretation are still without explanation [27, 260, 259]. Given the importance of this issue, a great deal of research has been carried out in the last decades to understand how to identify and characterize the source of these mistakes in order to mitigate them as much as possible.

When looking at a radiograph, clinicians are typically asked to localize lesions (if present), and then to classify them by judging their size, class, and so on. Importantly, during this visual search task, radiologists often examine dozens or hundreds of images in batches, sometimes seeing several related images one after the other. During this process, a main underlying assumption is that radiologists' percepts and decisions about a current image are completely independent of prior perceptual events. Recent theoretical and empirical research has raised the possibility that this is not true.

The visual system is characterized by visual serial dependency, a type of sequential effect in which what was previously seen influences (captures) what is seen and reported at this moment [39, 75]. Serial dependencies can manifest in several domains, such as perception [41,

42, 75, 173], decision making [1, 73], and memory [9, 78, 137], and they occur with a variety of features and objects, including orientation, position, faces, attractiveness, ambiguous objects, ensemble coding of orientation, and numerosity [20, 46, 75, 79, 141, 162, 175, 244, 245, 270, 282]. Serial dependence is characterized by three main kinds of tuning. First, feature tuning: serial dependence occurs only between similar features and not between dissimilar ones [75, 84, 175, 173]. Second, temporal tuning: serial dependence gradually decays over time [75, 173, 270]. Third, spatial tuning: serial dependence occurs only within a limited spatial window; it is strongest when previous and current objects are presented at the same location, and it gradually decays as the relative distance increases [20, 44, 75, 173]. In addition, attention is a necessary component for serial dependence [75, 83, 133].

The empirical results above prompted our theoretical suggestion that perception occurs through Continuity Fields—temporally and spatially tuned operators or filters that bias our percepts towards previous stimuli through serial dependence [2, 41, 75, 245, 246]. Continuity Fields are a helpful, beneficial mechanism for promoting perceptual stability because they produce a smoothed percept that better matches the autocorrelation in the world in which we live [75, 161, 175]. In contrast to the highly structured and stable physical world, retinal images are constantly changing due to external and internal sources of noise and discontinuities from eye blinks, occlusions, shadows, camouflage, retinal motion, and other factors. Rather than processing each momentary image or object as being independent of preceding ones, the visual system favors recycling previously perceived features and objects. By incorporating serially dependent perceptual interpretations, the visual system smooths perception (and decision making and memory [137]) over time and helps us perceive a continuous and stable world despite noise and change.

The benefits of serial dependence arise because the world we encounter is usually autocorrelated. But it is not always. In some artificial, human-contrived, situations the world is not autocorrelated. One obvious example of this are visual stimuli attended in laboratory experiments (in visual psychophysics, cognition, psychology, neurophysiology, and many other domains). Often stimuli are randomly ordered, with the assumption that trials are treated independently by the brain [187, 276]. Serial dependence negatively impacts the ability to measure performance in these cases [75, 85, 161].

Visual search in clinical settings, such as reading radiographs or pathology slides, is an even more striking example where stimuli may not be autocorrelated. When seeing and judging lesions under such circumstances, serial dependence could introduce a bias in perceptual judgments that may result in a significant reduction in sensitivity and increase in errors. The negative impacts of serial dependence in search tasks would be especially prominent in cases where there is low signal, high noise, high uncertainty, or where fine discriminations are required [20, 39, 41, 42, 75, 175]. These are exactly the challenging situations that radiologists routinely face when searching scans. We hypothesize that because of serial dependence, radiologists' perceptual decisions on any given current radiograph could be biased towards the previous images they have seen. To preview our results, we measured recognition of simulated tumors in trained clinicians and found that their perceptual judgments were significantly affected by serial dependence.

2.2 Method

Observers and apparatus

All experimental procedures were approved by and conducted in accordance with the guidelines and regulations of the UC Berkeley Institutional Review Board. Participants provided informed consent in accordance with the IRB guidelines of the University of California at Berkeley. All participants had normal or corrected-to-normal vision, and were all naïve to the purpose of the experiment. Fifteen trained radiologists (gender: 4 female, 11 males; qualification: 11 experts, 3 residents, & 1 fellow; age: 27–72 years) participated in Experiment 1. They were recruited at RSNA, Radiological Society of North America Annual Meeting (Chicago, US December 1st–6th, 2019). Of the fifteen, two participants did not complete the study, and their data were excluded. Eleven non-expert observers (7 female; aged 19–21 years) participated in Experiment 2. Sample size was determined based on radiologists' availability at RSNA, and was similar to current studies of serial dependence [42, 171, 199]. Eleven non-expert observers (7 female; aged 19–21 years) participated in Experiment 2. They were recruited from a student pool at UC Berkeley.

Stimuli were generated on a 13.3 inch 2017 MacBook Pro with a 28.7 cm × 18 cm screen with PsychoPy [203, 202]. The refresh rate of the display was 60 Hz and the resolution 1440 × 900 pixels. Stimuli were viewed from a distance of approximately 57 cm. Observers used a laptop keyboard for all responses.

Stimuli and design

To simulate the screening performed by radiologists, we created three objects with random shapes and generated 48 morph shapes in between each pair (147 shapes in total; Fig. 2.1A). We used these shapes as simulated lesions. On each trial, radiologists viewed a random simulated lesion superimposed on a mammogram section and were then asked to adjust a shape to match the simulated lesion they previously saw. The stimuli consisted of light-gray shapes based on 3 original prototype shapes (A/B/C; Fig. 2.1A). A set of 48 shape morph shapes was created between these prototypes, resulting in a morph continuum of 147 shapes. The shapes were approximately 3.7° width and height. Each shape was blurred by using a gaussian blur function in OpenCV with a gaussian kernel size of 1.55°. On each trial, a random shape was presented at a random angular location relative to central fixation (0.35°) in the peripheral visual field (4.4° eccentricity, from center to center). The shape was embedded in a random mammogram (30% transparency level) and was presented for 500 ms (Fig. 2.1B). Mammograms were taken from The Digital Database for Screening Mammography [23] (100 possible alternatives) and enlarged to fit the screen. The mammograms (2000 × 4500 pixels) were enlarged three times and cut at a central position such that about 15% of each x-ray was displayed. This resulted in breast tissue covering the entire screen. Next, we presented a mask composed of random Brownian noise background ($1/f^2$ spatial noise). After the mask, a random shape drawn from the morph continuum (width and height: 3.7°;

color: light-gray) appeared at the fixation point location, and observers were asked to adjust the shape to match the perceived shape using the left/right arrow keys (continuous report, adjustment task; left–right arrow keys to adjust the shape). The starting shape was randomized on each trial. Observers were allowed to take as much time as necessary to respond and pressed the spacebar to confirm the chosen shape. Following the response and a 250 ms delay, the next trial started.

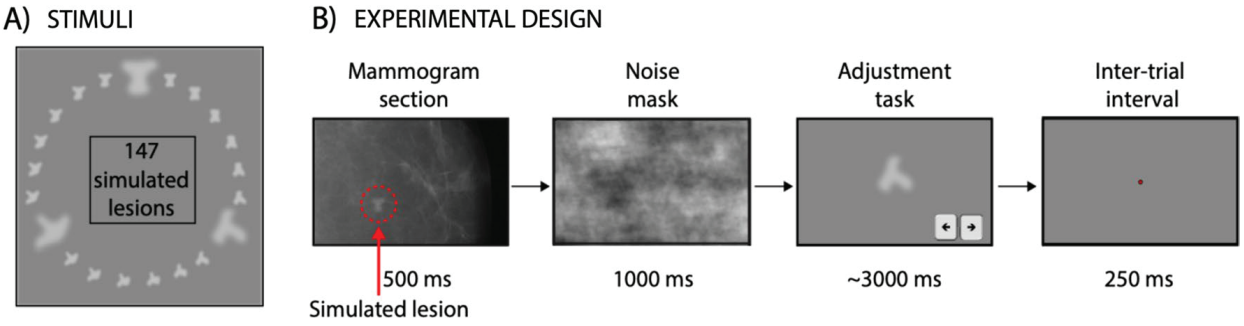


Figure 2.1: Stimuli and design of the Experiments 1 and 2. **A.** We created three objects with random shapes (prototypes A/B/C, shown in a bigger size) and generated 48 morph shapes in between each pair (147 shapes in total). We used these shapes as simulated lesions during radiological screening. **B.** Observers were presented with a random shape (simulated lesion) hidden in a mammogram section, followed by a noise mask. Radiologists were then asked to adjust the shape to match the simulated lesion they previously saw, and pressed spacebar to confirm. During the inter-trial-interval, a red fixation dot appeared in the center. The size of the shape adjustment is identical to the size of the simulated lesion, but it was enlarged for illustrative purposes. After a 250 ms inter-trial interval, the next trial started

During the experiment, observers were asked to continuously fixate a red dot in the center (0.35° radius). On each trial, they were first presented with a shape in a random location at 4.4° eccentricity, followed by a noise mask (Fig. 2.1). Observers were then asked to adjust a shape to match the one they previously saw (adjustment task). Observers performed 3 blocks of 85 trials each (Fig. 2.1B). In a preliminary session, observers completed a practice block of 10 trials. Mean adjustment time was 3240 ± 804 ms in Experiment 1 and 2980 ± 578 ms in Experiment 2. The only difference between Experiment 1 and 2 were the participants. In Experiment 1, we tested trained radiologists, whereas in Experiment 2, we tested students from the UC Berkeley population. Equipment and experimental design were otherwise identical.

2.3 Data analysis

Feature tuning analysis

We measured response errors on the adjustment task to determine whether a subject's judgment of each simulated lesion was influenced by the previously seen lesions. Response error was computed as the shortest distance along the morph wheel between the match morph and the target one (current response – current shape morph). For each participant's data, trials were considered lapses and were excluded if adjustment error exceeded 3 standard deviations from the absolute mean adjustment error or if the response time was longer than 20 s. Less than 2% of data was excluded on average.

Response error was compared to the difference in shape between the current and previous trial, computed as the shortest distance along the morph wheel between the previous target lesion (n-back) and the current target shape (current response – current shape morph). We quantified feature tuning by fitting a von Mises distribution to each subject's data points (see details below). Additionally, for each observer, we computed the running circular average within a 20 morph units window. Figure 2.3A-B shows the average of the moving averages across all the observers, and the corresponding von Mises fit. Figure 2.3E-F shows the half-amplitudes von Mises distribution for individual observers.

Temporal tuning analysis

We quantified temporal tuning by fitting a derivative of von Mises to each subject's data using the following equation:

$$y = -\frac{a\kappa \sin(x-\mu) e^{\kappa \cos(x-\mu)}}{2\pi I_0(\kappa)}$$

where parameter y is response error on each trial, x is the relative orientation of the previous trial, a is the amplitude modulation parameter of the derivative-of-von-Mises, μ indicates the symmetry axis of the von Mises derivative, κ indicates the concentration of the von Mises derivative, and $I_0(\kappa)$ is the modified Bessel function of order 0. In our experiments, μ is set to 0. We fitted the von Mises derivative using constrained nonlinear minimization of the residual sum of squares. As a measure of serial dependence, we reported half the peak-to-trough amplitude of the derivative-of-von-Mises (Figure 2.3E, F). We used the half amplitude of the von Mises, the parameter a in the above equation, to measure the degree to which observers' reports of simulated lesions were pulled in the direction of n-back simulated lesions. For example, if subjects' perception of a lesion was repelled by the 1-back simulated tumor (e.g., because of a negative aftereffect), or not influenced by the 1-back lesion (because of independent, bias-free perception on each trial), then the half-amplitude of the von Mises should be negative or close to zero, respectively.

For each subject's data, we generated confidence intervals by calculating a bootstrapped distribution of the model-fitting parameter values. For each observer, we resampled the data

with replacement 5000 times [61]. The relationship on each trial between response error and relative difference in shape (between the current and previous trial) was maintained. On each iteration, we fitted a new von Mises to obtain a bootstrapped half-amplitude and width for each subject.

Previous research recently showed that individual observers can have idiosyncratic biases in object recognition and localization, which are unrelated to serial dependence. For example, there are individual stable differences in perceived position and size, originating from a heterogeneous spatial resolution that carries across the visual hierarchy [142, 266]. For this reason, we conducted an additional control analysis to remove such potential unrelated biases before fitting the von Mises derivative function. We plotted observer's error values (current response - current shape morph) as a function of the actual stimulus presented (current shape morph), and fit a radial basis function (30 Gaussian Kernels used) to the data. This allowed us to quantify the idiosyncratic bias for each observer. For example, observers may make a consistent error in reporting a simulated lesion of 20 morph units as being 10, thus creating a systematic error of -10 morph units. Conversely, if there was no systematic error, all error would approximate zero. We then regressed out the bias quantified by the radial basis fit by subtracting it from the observer's error. This subtraction left us with residual errors that did not include the idiosyncratic biases unrelated to serial dependence. Importantly, the addition of this control analysis—removing systematic biases unrelated to serial effects—had no significant impact on the serial dependence results. It did not generate or increase the measured serial dependence.

As an additional method to rule out potential unrelated biases on the serial dependence effect, we explored the effect of future trials on the current response [78, 179]. That is, we compared the current trial response error to the difference in shape between the current and following trial (n-forward). Since observers have not seen the future trial shape, their current response in a given trial should not be in any ways related to the shape that will be presented to them next.

Spatial tuning analysis

In order to measure the spatial tuning of serial dependence, we binned trials according to the distance between the current and previous shape angular locations (Fig. 2.4). First, we divided trials from each observer into 3 main relative angular distance groups: $0° - 60°$, $61° - 120°$, and $121° - 180°$ for 1-back trials. For example, a relative angular distance of $0°$ indicates that previous and current lesions were presented at the same location (for example, $45°$ and $45°$ of angular distance in previous and current trials). Similarly, a relative angular distance of $60°$ indicates that previous and current lesions were presented at $30°$ and $90°$ of angular distance. The distance between successive shape locations was computed as $\sqrt{(\S current - \S previous)^2 + (\dagger current - \dagger previous)^2}$. Second, we extracted 60 random trials from each observer for each distance group, and collapsed all the trials from all the observers in three super-subject groups. Third, for each super-subject we fitted a derivative of von Mises and computed the half amplitudes. Fourth, we performed a regression line

analysis across the three half amplitudes of the distance groups. For each super-subject, this analysis yielded a slope of the regression line, which reflects how much serial dependence varies as a function of distance between sequential stimuli. We repeated the procedure 5000 times, by resampling the data with replacement on each iteration.

2.4 Results

We tested whether serial dependence influenced recognition of simulated lesions when viewing consecutive images of mammogram tissues in radiologists and untrained observers. Response error (y-axis) was computed as the shortest distance along the morph wheel between the match shape and the simulated lesion. Average response error was similar across groups; 9.2 ± 1.8 morph units in Experiment 1 (radiologists) and 8.9 ± 1.8 in Experiment 2 (untrained observers; $t(22) = 0.34$, $p = 0.74$).

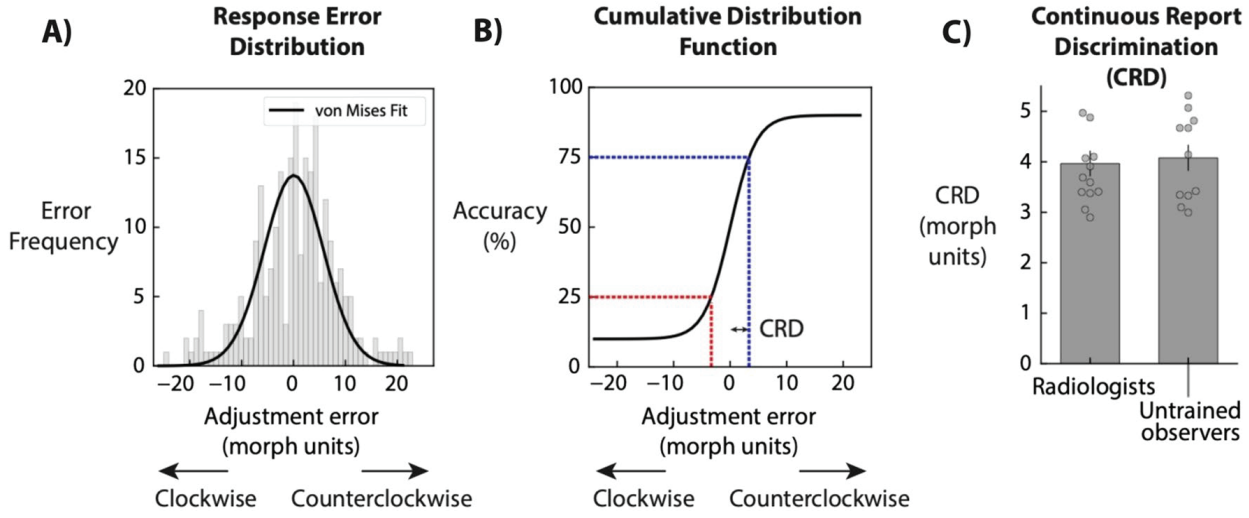


Figure 2.2: Continuous Report Discrimination index (C.R.D). **A.** For each observer, we plotted a frequency histogram of the adjustment errors and fitted a Von Mises to quantify adjustment performance. **B.** We then converted the von Mises fit into a Cumulative Distribution Function. Continuous Report Discrimination index was calculated by taking the half difference between 25 and 75th percentile in terms of adjustment error morph units. **C.** Each dot shows CRD index for individual observers in the two groups. Bars indicate average in Experiment 1 and 2, and error bars indicate standard error

To further quantify discriminability of the simulated lesions, we fit a von Mises function to each observer's response error frequency distribution (Fig. 2.2A) and computed the corresponding Cumulative Distribution Function (CDF; Fig. 2.2B). The CDF was generated with a ceiling and floor parameters of 0.1 and 0.9, respectively, and a free x-axis shift

parameter to allow for any observers' bias to be taken into account. For each observer's individual CDF, a Continuous Report Discrimination index (C.R.D.) was defined as half of the difference between the 25th and 75th percentile of their Cumulative Distribution Function (Fig. 2.2C). This measure can be considered as the equivalent of JND (Just Noticeable Difference) for continuous reports. The mean CRD was 3.97 ± 0.26 morph units for radiologists and 4.08 ± 0.25 morph units for untrained observers.

To test whether radiologists' lesion perception was pulled by lesions in previous mammograms, we plotted the adjustment error on the current trial in relation to the difference in shape between the current and previous trial, computed as the shortest distance along the morph wheel between the previous lesion and the current lesion. A derivative-of-von Mises curve was then fitted to the observers' data (Fig. 2.3A, B, see Feature Tuning analysis). We bootstrapped each subject's data 5000 times and reported the mean bootstrapped half-amplitude as a metric of the sequential dependence (Fig. 2.3E, F).

In Experiment 1, all participants except for one displayed a positive von Mises half-amplitude, indicating that lesion perception on a given trial was significantly pulled in the direction of the lesion presented in the previous trial ($p < 0.001$, group bootstrap, $n = 13$, Fig. 2.3E). Even the lesion two trials in the past influenced current judgments ($p = 0.01$, group bootstrap, Fig. 2.3E). No attraction was found for 3-trials back ($p = 0.09$, group bootstrap, Fig. 2.3E). A similar pattern of results was found in Experiment 2 with untrained observers. Lesion perception on a given trial was significantly pulled in the direction of lesions presented in the previous trial for 1 and 2 trials back ($n = 11$; 1-Back; $p < 0.001$, 2-Back; $p < 0.001$, group bootstrap, Fig. 2.3F) but not for 3-back ($n = 11$; $p = 0.128$, group bootstrap, Fig. 2.3F). There was no statistical difference between radiologists and untrained observers for 1-back and 2-back (Fig. 2.3; 1-back, $p = 0.88$; 2back, $p = 0.19$), whereas there was a statistical difference for 3-back ($p = 0.02$; but no serial dependence was detected in those conditions).

As a control for possible confounds or artifacts, we checked whether lesion perception could have been biased from lesions one, two, or three trials in the future. As expected, lesion perception was not significantly influenced by future stimuli for radiologists (1-forward, group bootstrap half amplitude: 0.27 morph units, $p = 0.50$; 2-forward, group bootstrap half amplitude: 0.35 morph units, $p = 0.5$, 3-forward group bootstrap half amplitude: 0.5 morph units, $p = 0.38$). The same was true for naïve observers (1-forward, group bootstrap half amplitude: -0.83 morph units, $p = 0.16$; 2-forward, group bootstrap half amplitude: 0.22 morph units, $p = 0.72$; 3-forward, group-bootstrap half amplitude: 0.23 morph units, $p = 0.67$).

Average response time was similar across Experiments; 3244 ± 845 ms in Experiment 1 and 2980 ± 578 ms in Experiment 2 ($t(22) = 0.834$, $p = 0.41$). Lesion recognition was therefore strongly attracted toward lesions in previous mammograms seen more than 5s or 10s ago (Fig. 2.3E, F). These results suggest a featural tuning (Fig. 2.3A, B) and temporal tuning of 5–10 s (Fig. 2.3E, F), in accordance with previous literature [75, 84, 173, 185, 245, 270].

In order to further characterize the strength of the serial dependence effect, we computed

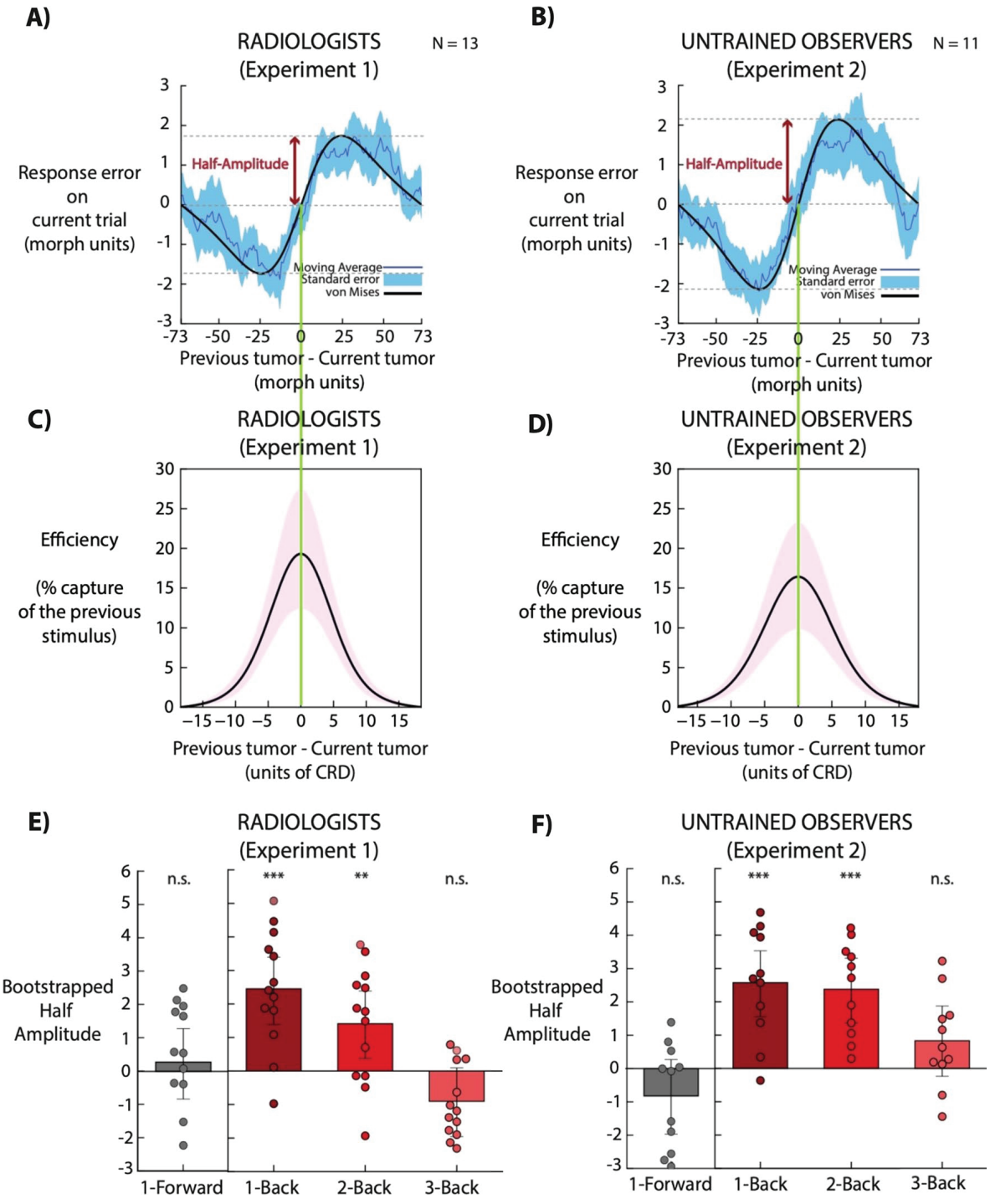


Figure 2.3: (See caption on next page.)

Figure 2.3: (See figure on previous page.) Serial dependence in the perception of simulated lesions by expert radiologists and untrained observers. **A, B.** In units of shape morph steps, the x-axis is the shortest distance along the morph wheel between the current and one-back simulated lesion, and the y-axis is the shortest distance along the morph wheel between the selected match shape and current simulated lesion. Positive x-axis values indicate that the one-back simulated lesion was clockwise on the shape morph wheel relative to the current simulated lesion, and positive y-axis values indicate that the current adjusted shape was also clockwise relative to the current simulated lesion. The average of the running averages across observers (blue line) reveals a clear trend in the data, which followed a derivative-of-von-Mises shape (model fit depicted as black solid line; fit on average of running averages). Light-blue shaded error bars indicate standard error across observers. Lesion perception was attracted toward the morph seen on the previous trial. Importantly, it was tuned for the similarity between the previous and current morph (feature tuning). **C, D.** The derivative-of-von-Mises was converted into its source von Mises function (y-axis), and the relative morph difference was plotted in terms of CRD units (x-axis). Violet-shaded error bars indicate 95% confidence interval. The curve indicates the proportion of change in response predicted by the change in the sequential stimulus. **E, F.** Bootstrapped half amplitudes of the derivative of von Mises fit for 1, 2, and 3 trials back. Half amplitude for 1-forward is shown as a comparison (grey bars). Each filled dot represents the bootstrapped half amplitude (morph units) for a single observer. Bars indicate the group bootstrap and error bars are bootstrapped 95% confidence intervals

how much the current simulated lesion was captured by lesions in the previous trial. We converted the derivative-of-von Mises into its source von Mises function. In order to compare our effect with shape discriminability, we divided the relative morph difference (previous tumor - current tumor; x-axis) by the average CRD index (from Fig. 2.2C). The plots in Fig. 2.3B, C show the proportion of change in response (efficiency) predicted by the change in the sequential stimulus. Serial dependence captured the current (simulated) tumor with peaks of $22-25\%$, and expanded over a large discriminability range (from -10 to $+10$ CRD units).

As an additional analysis, we investigated how much adjustment errors were biased more towards the shape category on the previous trial compared to other previous object categories. Shape categories A/B/C were defined as the prototype $A/B/C \pm 24$ morph units (49 morph units in total). Adjustment responses were coded as indicating category A/B/C. We computed the percentage of mistakes towards the shape category in 1-back trials, and normalized the index by subtracting 33.33% (chance percentage level) from each percentage index (see Fig. 2 in [171] for an in-depth explanation of the analysis). Observers misclassified the simulated lesion on a current trial as the lesion in 1-back trials 8% more often than expected by chance.

In order to further quantify the strength of the 1-back serial dependence effect, we conducted a linear regression analysis on the response error as a function of the relative morph

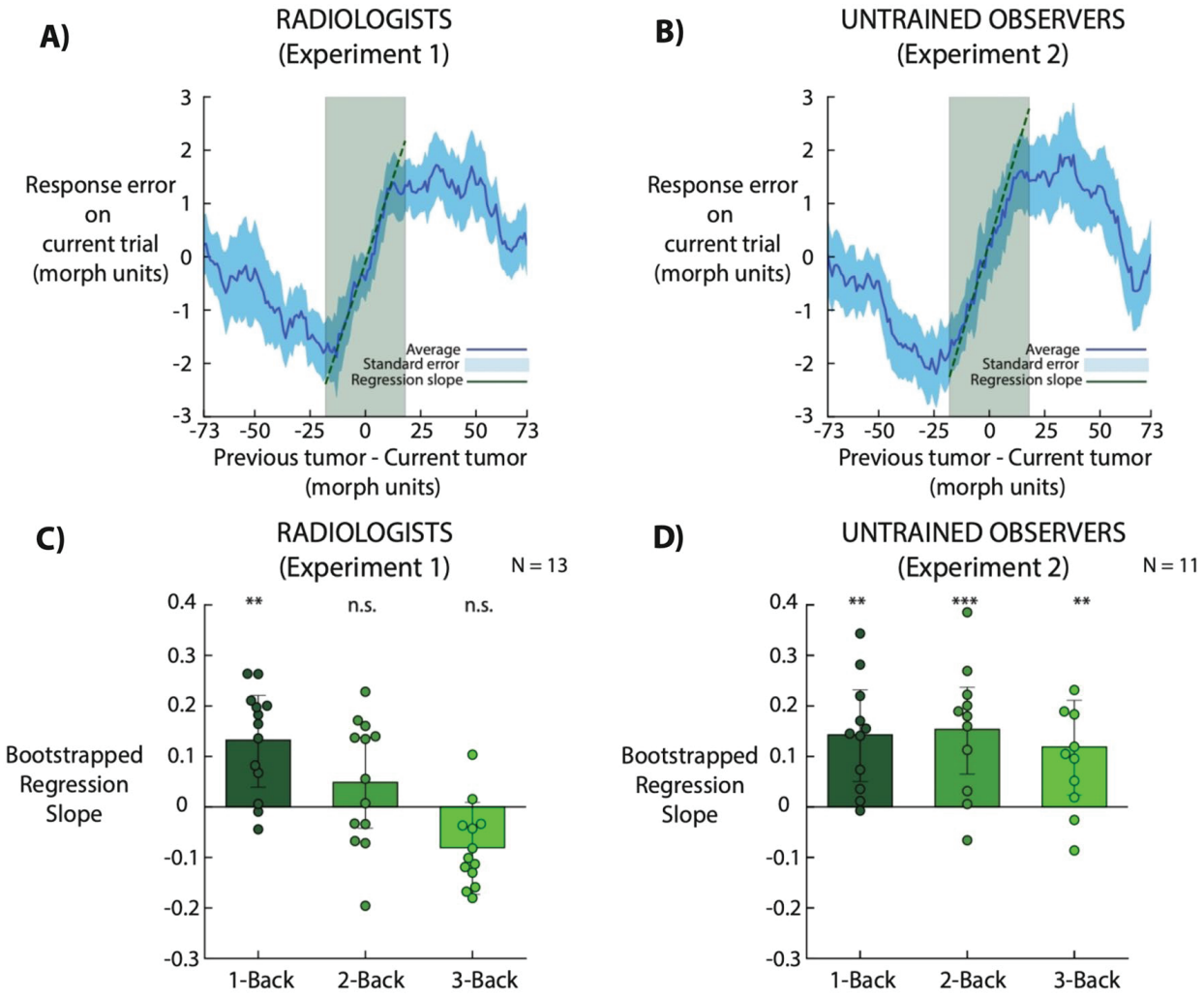


Figure 2.4: Serial dependence effect size estimation. **A, B.** Blue lines indicate the average of the running averages across observers (same data as Fig. 2.2). Light-blue shaded error bars indicate standard error across observers. We fitted a linear regression on the response error as a function of the relative morph difference from -17 to $+17$ morph units (model fit depicted as green dashed line; fit on average of running averages). Dark green shaded areas indicate the morph relative difference considered in the regression analysis. **C, D.** Bootstrapped regression slopes for 1, 2, and 3 trials back. Each filled dot represents the regression slope for a single observer. Bars indicate the group bootstrap slope and error bars are bootstrapped 95% confidence intervals

difference (from −17 to +17 morph units on the x-axis in Fig. 2.3A, B, 25% of the central range). Average slope was 0.132 ± 0.10 in Experiment 1 and 0.143 ± 0.10 in Experiment 2, thus meaning that both radiologists and untrained participants exhibited a perceptual pull of 13% towards simulated lesions viewed 1 trial back (Fig. 2.4, radiologists; 1-back, $p < 0.01$; 2-back, $p = 0.30$; 3-back, $p = 0.09$; naïve observers; 1-back, $p < 0.01$; 2-back, $p < 0.001$; 3-back, $p = 0.01$).

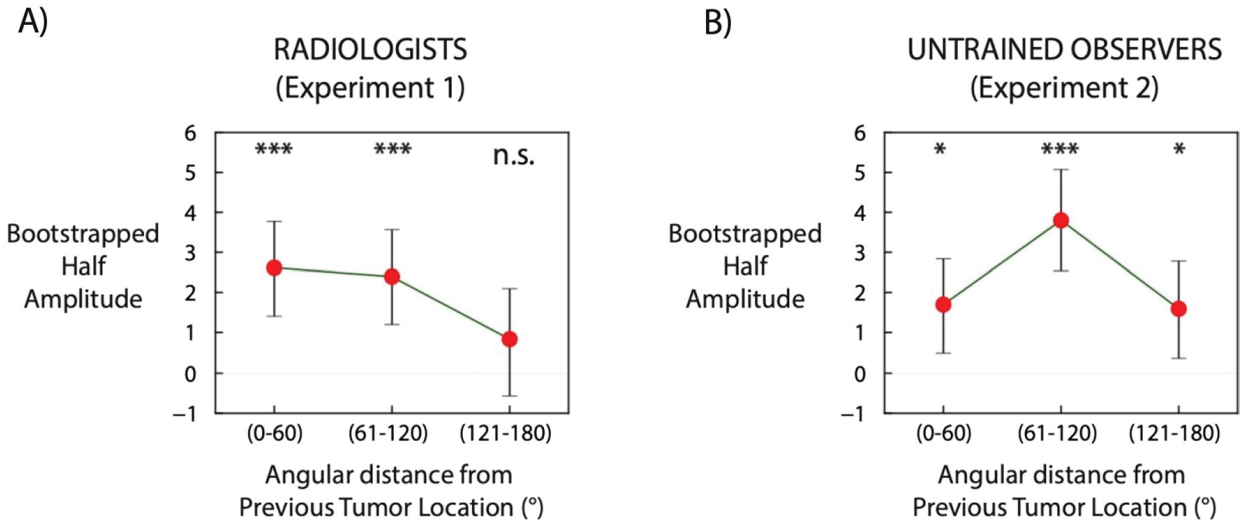


Figure 2.5: Spatial tuning of serial dependence. **A** refers to Experiment 1, whereas **B** refers to Experiment 2. Each red dot refers to a different relative angular distance between current lesion and lesion in the 1-back trial, super-subject bootstrapped mean. For example, a bin distance 0° indicates that current and previous simulated tumor presented at the same location (30° of angular distance, for example). Error bars are bootstrapped 95% confidence intervals. Dashed line indicates half-amplitude zero (no bias)

As previously mentioned, an important property of serial dependence is spatial tuning [20, 41, 75, 79, 173]. We therefore investigated whether serial dependence in simulated radiological screening is affected by the spatial distance between current and previous lesions. On each trial, the simulated lesion was presented at a fixed distance from the center but at random angular distance. Hence, we predicted that serial dependence will be highest when current and previous lesions are presented at a close relative distance, and will gradually decay as relative distance increases. For each participant, we divided the trials into three groups based on the relative distance of the 1-trial back stimulus (Fig. 2.5; See Spatial Tuning analysis section).

In Experiment 1, serial dependence occurred for an angular distance groups of $0° - 60°$ and $61° - 120°$, ($0° - 60°$: $p < 0.001$; $61° - 120°$: $p < 0.001$ group bootstrapped distribution; Fig. 2.5A), whereas no serial dependence occurred for an angular distance group of $121° - 180°$

($121° - 180°$: $p = 0.20$; group bootstrapped distribution; Fig. 2.5A). There was no statistical difference across the two groups for relative distances of $0° - 60°$ ($p = 0.29$), $61° - 120°$ ($p = 0.11$) and $121° - 180°$ ($p = 0.42$). In order to further characterize spatial tuning for 1-trial back, we performed a regression analysis on the three distance groups. Regression slope was significantly different from zero, thus indicating a gradual decay of serial dependence with increased relative distance (slope $= -0.89$; $p = 0.05$; group bootstrapped distribution). These results are consistent with prior findings that serial dependence is modulated by the relative location of the sequential targets. Therefore, in a radiological screening environment, the current lesion may be misperceived as more similar to the previous one if current and previous lesions are presented at similar locations. Interestingly, untrained observers from Experiment 2 did not show the same spatial tuning: serial dependence occurred at all tested angular distance groups ($0° - 60°$: $p < 0.05$; $61° - 120°$: $p < 0.001$; $121° - 180°$: $p < 0.05$; group bootstrapped distribution; Fig. 2.5) with no gradual decay as a function of spatial separation. When performing a regression analysis on the three distance groups, regression slope was not significantly different from zero (slope $= -0.05$; $p = 0.90$; group bootstrapped distribution; Fig. 2.5B). The implications of this result will be discussed in the next section.

Taken together, our results show that simulated tumor recognition is strongly biased towards previously presented simulated lesions up to 10s in the past. Importantly, this sequential effect occurs with expert radiologists and exhibits all the defining properties of traditional serial dependence: feature tuning (Fig. 2.3A, B), temporal tuning (Fig. 2.3E, F) and spatial tuning (Fig. 2.5A).

2.5 Discussion

We found that the perceptual decisions of radiologists were subject to serial dependence. Simulated lesion recognition was biased towards simulated tumors presented up to 10 s in the past (Fig. 2.3A). Importantly, radiologists exhibited a perceptual pull of 13% towards previously seen tumors (Fig. 2.4). Moreover, serial dependence alone resulted in 8% more miscategorizations than were expected by chance or due to noise. This perceptual pull exhibited all three tuning characteristics of Continuity Fields: feature tuning (Fig. 2.3A, B), temporal tuning (Fig. 2.3E, F) and spatial tuning (Fig. 2.5A). In Experiment 2, we found largely similar results with untrained observers, with the exception that less clear spatial tuning was found. Taken together, these results show that radiologists' perceptual judgements are affected by serial dependence.

Our results extend previous work, which investigated the impact of serial dependence in a simulated clinical search task [171]. In untrained observers, it was found that shape classification performance was strongly impaired by recent visual experience, biasing classification judgments toward the previous image content. Whereas those results can be considered as proof of the concept that serial dependence can be detrimental in clinical tasks, the present study extended this in several ways including (1) testing trained radiologists, (2) using actual mammogram textured backgrounds as stimuli and (3) implementing a more thorough con-

tinuous report task instead of a classification judgment. The results thus show that trained radiologists, as well as naïve observers, suffer from serial dependence. Future research will investigate whether this kind of error occurs in a more realistic radiological screening setting.

Interestingly, we did not find spatial tuning in Experiment 2 with untrained observers. Whereas this seems like a somewhat surprising result, it must be considered that the maximum relative distance in our experiments was 8.8° (double the radius), and previous literature has shown that the spatial window where serial dependence occurs is around $10° - 15°$ or even larger [44, 75, 171]. The potentially interesting result, therefore, is the finding of narrower spatial tuning with expert radiologist observers. The reason for this narrowed spatial tuning is unknown, but it does raise questions about the role of familiarity and expertise. Serial dependence is known to scale with uncertainty [41], and it is possible that the spatial tuning of serial dependence varies with familiarity as well.

In addition to differences in expertise and familiarity, an additional difference between the two groups of observers in these experiments could be attentional. Previous literature has shown that serial dependence is gated by attention [75, 79, 163, 208]. In comparison to untrained observers, radiologists may pay more attention to the stimuli or attend to different features of the stimuli; therefore, serial dependence tuning may differ with expertise.

It might be argued that our results can be explained by a mere motor response bias, i.e. the motor response during the adjustment task may be biased towards the previous motor response. However, a large literature has shown that serial dependence still occurs when no adjustment is given in the previous trial, thus ruling out a mere motor effect [75, 175, 173]. In addition, a simple motor bias cannot explain why serial dependence was tuned for the relative spatial location, biasing simulated tumor judgments only when current and previous tumors were presented at a close angular distance (Fig. 2.5A). Neither can it explain relative featural difference, biasing tumor adjustment only when current and previous tumors were similar enough (Fig. 2.3A, B).

Beyond the motor component, there is an intense debate on the underlying mechanism(s) of serial dependence. Among others, serial dependence was proposed to occur on the perception [41, 75, 173], decision [84, 199] and memory level [9, 20]. Our results do not allow us to disentangle on which level(s) serial dependence actually occurs. There is psychophysical evidence that serial dependence acts on perception, thus biasing object appearance towards the past [41, 75, 80]. How serial dependence in perception actually occurs is still a matter of debate; it was recently shown that awareness is required for serial dependence to occur, thus suggesting that a top-down feedback from high level areas is crucial for serial dependence [80, 133].

It may be argued that the duration of the mammogram presentation (500 ms) is too short and radiologists observe mammograms for a much longer period of time. In fact, the average duration of radiograph fixation for hitting the first mass has been reported as $1.8 - 2$s, which is surprisingly brief [147, 193]. Interestingly, sufficiently long mammogram exposure durations may lead to the opposite effect, i.e. negative aftereffect. It was found that when adapting normal observers to image samples of dense or fatty tissues, exposure to fatty images caused an intermediate image to appear denser (and vice versa) [138, 139,

140]. Importantly, mammogram perception was biased away from the past. Future research will establish under which conditions these two biases (perception biased towards or away from the past) arise in radiological screening.

Limitations of current study

Our results show that radiologists suffer from significant serial dependence in their perceptual judgments. Whether these significant serial dependencies are left at the door of the reading room is as-yet untested. However, the results here show that radiologists are not immune from sequential effects in perceptual decisions. This is only a first step, and there are many improvements required to optimize the ecological validity of our findings. Future improvements will be implemented in order to fully address the impact of serial dependence in a clinical setting.

First, the stimuli. Our study tested serial dependence with a generated set of shape stimuli, but actual tumor images will be required to test the role of serial dependence in radiological screening. In addition, within a radiograph, there can be a variety of features which may be interpreted as tumors, from actual masses, to microcalcifications, architectural distortions, and focal asymmetries. Future research will test whether these features, as well as actual lesions, suffer from serial dependence.

Second, the task. We chose a continuous report paradigm in our experiments, as it provides precise trial-wise errors and has proven to be very reliable in measurements of serial dependence in the past [41, 40, 83, 75, 84, 161]. Given the radiologists' time constraints and resulting limited number of trials, we considered this task to be relatively efficient. The untrained observer data provides a useful baseline in this respect. A previous paper that used a 3AFC classification task found a similar amount of serial dependence in untrained observers as that found here [171]. Nevertheless, as the actual task of the radiologist involves classifying lesions and localizing them, implementing more realistic tasks with radiologists will be important in future studies.

Third, mammogram duration. Although radiologists fixate radiographs for slightly longer durations (500 ms in the present and $1.8-2$s reported in the literature [147, 193]), they were shown to perform above chance in detecting abnormalities in chest radiographs with 200 ms duration [151]. It will be interesting to test which biases arise with increasing stimulus duration, whether a positive one (as shown by our results), a negative one [138, 139, 140], or no bias at all.

Finally, whereas our results may indicate that radiological screening is detrimentally affected by serial dependence, they also open avenues to mitigate this bias. Since serial dependence was shown to occur only under restricted featural, spatial, and temporal conditions, some strategies could be implemented to induce perceptual decisions outside of these conditions. For example, mammograms could be presented at different spatial locations. Because of spatial tuning, the relative distance between lesions would be so large that serial dependence would no longer occur. Other strategies may be implemented based on temporal and featural tuning as well.

Chapter 3

Controllable Medical Image Generation via GAN

3.1 Introduction

Medical imaging has transformed modern medicine, allowing clinicians to noninvasively examine and diagnose patients relatively quickly and easily. In recent decades, there have been dramatic advances in the medical imaging technologies themselves, ranging from MRI, CT, PET, photography, ultrasound, among many other techniques. The improvements are astounding, but it is noteworthy that ultimately the data provided by these techniques requires critical human involvement in detection, selection, interpretation, and diagnoses. The imaging techniques themselves are not the only bottleneck for obtaining accurate diagnoses.

Fortunately, along with the technological development, there have also been concomitant advances in the application and use of these technologies. For example, there is a recent surge in computer vision and medical image perception research. These two areas involve, respectively, artificial (algorithmic) and human users. In both machines and humans, there is a great deal of potential to improve the use of medical imaging in clinical practice. In addition to the more ambitious goals of automated diagnoses, filtering, or cuing clinicians [226, 167, 111, 289], there are distinct and more immediately pressing goals of improving clinicians' medical image perception and decisions [259, 273]. For example, in the realms of training, error detection, diagnostic support, among others [260].

To improve both machine and human medical image perception, it is necessary to have sufficient source data. Unfortunately, labeled and de-identified public medical imaging data is scarce. Sometimes researchers resort to collecting their own data from nearby hospitals, usually from local areas that cannot represent the broader population. Second, even if larger datasets are collected, the necessary data processing procedures such as data de-identification, labeling, and categorizing are tedious, time-consuming, and very expensive. For example, in certain medical imaging tasks, such as lesion segmentation, in order to prepare the training data, it requires experts to perform meticulous annotations that are

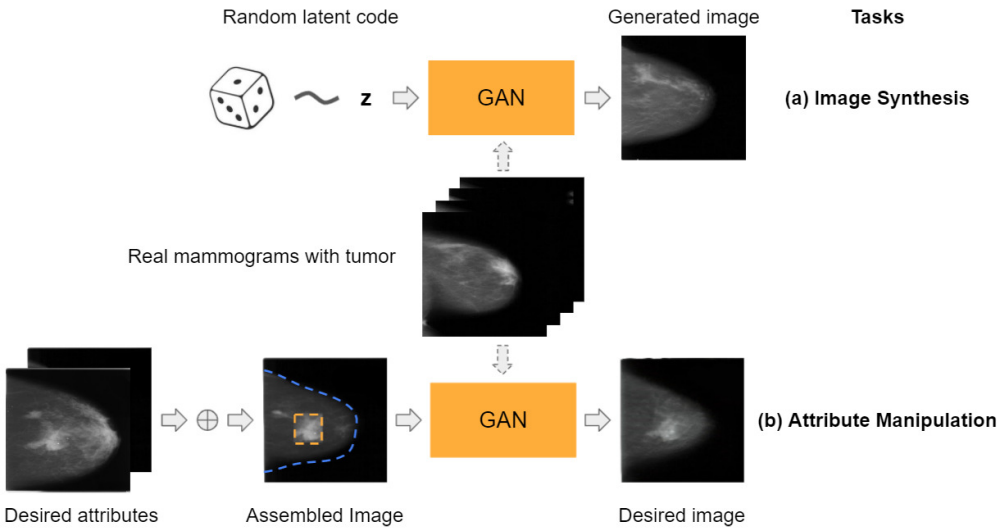


Figure 3.1: **Pipeline.** Controllable medical image generation using the proposed GAN model. (a) Medical image generation: novel and authentic medical images can be generated from random latent codes z. (b) Attribute manipulation: desired attributes can be assembled together to satisfy certain experimental settings. Here, we use mammogram as an example medical modality. Real mammograms with tumor were utilized to train the proposed model. Our proposed model can be easily adapted to other medical modalities, such as MRI, CT, and skin cancer images.

costly and time intensive [272]. Moreover since collected medical images are specific to each individual patient, it can be difficult to find specific images or image properties that satisfy certain desired experimental configurations [171]. Of course, due to intricate tissue structures, manipulating attributes of those collected medical images using traditional image processing methods is difficult or impossible, at least in a realistic manner.

The data scarcity problem has presented a major challenge to research on medical image perception. At a broad level, medical image perception research studies the visual and cognitive processes that clinicians rely on to make decisions. As in other domains of human factors, the goal of understanding those mechanisms is to improve (i.e., guide, cue, facilitate, speed, etc) clinician performance. Recently, in many psychophysical experiments, artificial medical stimuli have been employed [139, 171]. The artificial medical stimuli are often composed of simple shapes or textures with some form of noise background [257, 139, 171]. Related approaches involve using real medical images but superimposing clearly artificial "targets" [139, 171]. An advantage of these approaches is that they are relatively easy to generate and control in a precise manner, which is important for studying the cognitive and perceptual systems of clinicians [139, 171]. For example, the image attributes and "targets" are easy to manipulate such that researchers can perform shape morphing and background

replacement. This level of stimulus control is necessary in perception research to study things like visual search for lesions, visual recognition of lesions, inattentional blindness, cognitive load and interference, etc. However, those artificial medical images are obviously inauthentic, completely unlike what clinicians routinely examine. Thus, the results of these experiments fall invariably within a shadow of a doubt about clinical applicability.

Therefore, generating authentic and easily controllable medical images is critical for the entire field of medical image perception research. Alleviating this limit is only recently realistic, with the impressive development of deep learning in computer vision. For example, Generative Adversarial Network (GAN) is one of the promising models that have achieved great success on image generation tasks. GAN can generate high-quality authentic images with various categories [127, 130, 198], such as faces, cars, landscapes, and so on. Additionally, various methods can be applied to manipulate the attributes of Generative Adversarial Networks' outputs [183, 35, 198].

In this paper, we utilize Generative Adversarial Network (GAN) to generate authentic medical images (Fig 3.1 (a)). We also adopt a controllable approach to manipulate specific attributes of the generated images (Fig 3.1 (b)). The proposed method is tested on various medical image modalities such as mammogram, MRI, CT, and skin cancer images. For example, via controllable generation, we can create authentic mammograms with desired tumor and breast shapes. We also recruited both expert clinicians and untrained participants to discriminate the authenticity of each image (real vs GAN generated) in an objective psychophysical experiment. Finally, we investigate the perceptual loss which is utilized in the controllable generation. Various experiments verify the success of the proposed controllable medical image generation model.

Contributions: We propose a framework for controllable medical image generation with the following contributions.

- We propose to utilize Generative Adversarial Network (GAN) to generate medical images and verify the results on various medical image modalities such as mammogram, MRI, CT, and skin cancer images.
- We adopt a controllable approach to manipulate the attributes of the generated images in order to meet certain experimental configurations.
- We compare traditional similarity measurements with the perceptual metric in medical imaging.

We note that a shorter conference version of this paper appeared in [214]. Our initial conference paper was more limited in scope and did not extend the model to multiple medical image modalities. This paper extends the model to MRI, CT, and skin cancer images. Moreover, this paper compares traditional similarity measurements with the perceptual metric in medical imaging.

3.2 Related work

Convolutional Neural Networks

The idea of Convolutional Neural Networks (CNN) is from the discovery of the edge detector in cat's striate cortex [120]. Based on this finding, Fukushima [86] invented the first simple hierarchical, multilayered artificial neural network. After decades of development, LeCun et. al. [156] leveraged CNN for hand-written ZIP Code numbers recognition and trained the network end-to-end via gradient descent. This fully automatic image recognition model can be applied to many image categories and types. The great success is mainly attributed to the convolution operation, which can reveal the latent semantic information of an image, and the shared hierarchical kernels, which make the convolution shift-invariant. During training, the loss is computed based on specific metrics for certain tasks, updating the model parameters while it back propagates through the whole network.

However, the computation is heavy, which limits the model's capacity and ability for high-resolution images. With the deployment of Graphical Processing Unit (GPU), CNNs [144, 240, 232, 241, 106] have shown promise in computer vision tasks, such as image classification [106], object detection [93, 92, 212, 211], and object segmentation [107]. Recently, many medical imaging tasks have been utilizing CNNs [82, 228, 132]. Compared to traditional image processing methods, CNNs have much better performance with much faster inference speed.

Generative Adversarial Networks

Generative Adversarial Networks are special Convolutional Neural Networks, which consist of two networks, the generator(G) and the discriminator(D). These two networks are trained iteratively in an adversarial way where the generator(G) generates fake but authentic images to fool the discriminator and the discriminator(D) discriminates the real and fake images [95]. Using this promising computational model, high-quality images with various categories can be generated, such as faces, cars, and landscapes [127, 130, 198]. However, the initial GAN model [95] cannot generate sharp and recognizable images, and the training process is unstable. Later work improved the performance of GAN in different ways. Some papers focus on model architectures [183, 35, 194]. Others focus on improving the loss metrics and training strategies [98, 4, 26]. With these efforts, GAN training stability has improved, and GAN can generate low-resolution images with sufficient quality.

Recently, several approaches make high-resolution image generation also possible. PG-GAN [130] proposed to train the standard GAN from coarse to fine scale. The parameters for low-resolution block are trained first. Then higher-resolution blocks are added on gradually with the corresponding parameters updated accordingly. Based on the same training strategy, StyleGAN [127, 128] proposed to first map the original latent space $\mathcal{Z}$ into the $\mathcal{W}$ space through a non-linear mapping network. Then it is merged into the synthesis network via adaptive instance normalization (AdaIN) at each convolutional block [57, 118]. This

improves StyleGAN representations of scenes and details and allows it to produce authentic high-resolution images. In this paper, we adopt StyleGAN as our backbone for medical images generation. Moreover, a controllable approach is also utilized to manipulate the attributes of the generated images.

In medical image applications, [82] utilized DCGAN [207] and ACGAN [194] to generate CT liver lesion patches and boosted the liver lesion classification performance. [103] proposed to use WGAN [98] to generate MR images for data augmentation and physician training. [191] used GAN to predict CT images from MR images. [29] proposed an Auto-GAN to synthesize missing modality for medical images. Moreover, GAN has been widely used for skin cancer image generation and purification [18, 19, 91]. Our approach is different from aforementioned methods. In addition to purely generating new samples as GANs traditionally do, our method can also edit specific images via the encoder of our model.

Perceptual Loss

CNN features have already been utilized for calculating similarity for years. [3] proposed to use pre-trained AlexNet features for image quality measurement. Perceptual loss, which is also based on CNN features, was first proposed in [124] for style transfer [90] and super resolution tasks. Both are ill-posed problems. For style transfer, there is no absolute ground truth image for reference. For image super resolution, one low-resolution image can have many corresponding high-resolution images which can be down-sampled to the same low-resolution image. Thus, per-pixel metric is no longer suitable since semantic similarity matters. Recently, traditional similarity metrics, such as Structural Similarity Index Measure(SSIM) and Peak Signal-to-Noise Ratio (PSNR), are found to be inconsistent with human perception, and a perceptual metric has been utilized to measure the semantic similarity in many papers [157, 293, 119, 290]. In this paper, we use perceptual loss to regularize the encoder training and guide the latent code optimization in the encoding procedure.

3.3 Method

Here, we adapt the Generative Adversarial Network for medical image generation. In order to manipulate the image attributes, an encoder is added to encode certain image attributes into the latent code z which is the input of the GAN generator.

Our proposed model is composed of two parts. The first part is the GAN part which involves the generator(G) and the discriminator(D). The generator(G) will generate authentic(fake) images from the latent codes z, and try to fool the discriminator(D) during training. The discriminator(D) will discriminate whether the image is real (i.e. sampled from real images) or fake (i.e. generated from the generator), and try to beat the generator by distinguishing the fake images from the real ones. The second part of the model is the encoder(E), which can encode image attributes into the latent code z. This latent code can then be utilized to generate the image through the generator. Therefore, it can allow us to

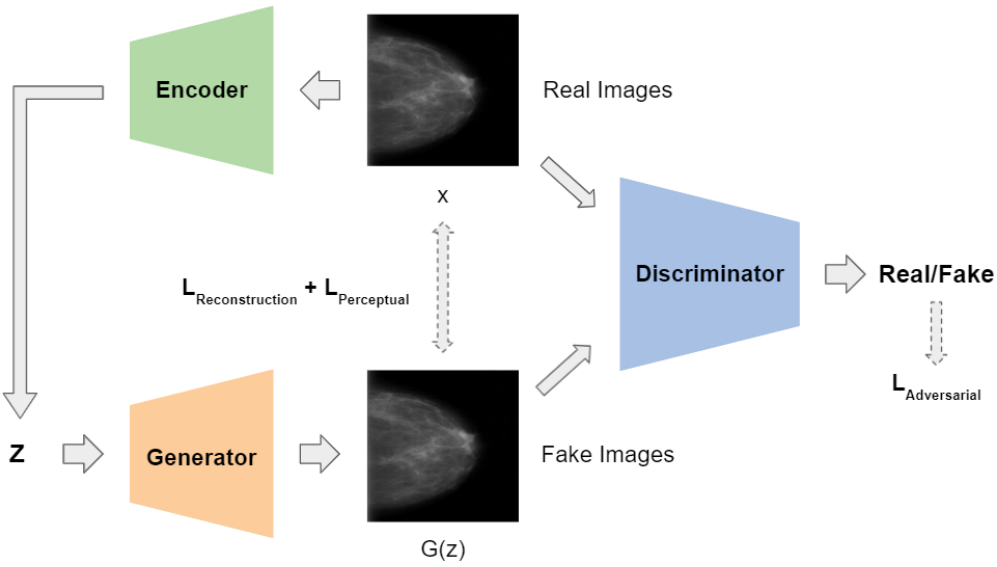


Figure 3.2: **Architecture of proposed method.** The architecture contains three sub-networks, the encoder(E), the generator(G), and the discriminator(D). The training has two phases. In the first phase, the generator and discriminator will be trained first without the encoder (E) via adversarial loss $L_{adversarial}$. In the second phase, the generator (G) will be fixed. The encoder (E) and discriminator (D) will be trained adversarially via the reconstruction loss $L_{reconstruction}$, the perceptual loss $L_{perceptual}$, and the adversarial loss $L_{adversarial}$. The dashed arrows indicate how to compute the corresponding loss metrics.

manipulate the generated image by manipulating the latent code through the encoder. The architecture is shown in Fig 3.2.

While training, the GAN part is first trained progressively [127] via adversarial loss $L_{Adversarial}$. The training process can be formulated as

$$\min_{G} \max_{D} E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim q(z)}[\log(1 - D(G(z)))] \tag{3.1}$$

where $p_{data}(x)$ and $q(z)$ indicate the real data distribution and the latent space distribution respectively, x is the sampled real image, z is the sampled latent code.

Then, we train the encoder part. After training the GAN part, the generator(G) is fixed. While training the encoder network, traditional methods [10] regularize the encoder on the latent space, encouraging the encoder to encode the same latent codes for the corresponding generated images regardless of the reconstructed images. This method can degrade the reconstruction quality. Instead, we adopt the idea from In-domain GAN inversion [292], where the regularization of the encoder is on the image space. In particular, the encoded vector is passed into the generator(G) again and the regularization is on the reconstructed image. The L2 reconstruction loss $L_{Reconstruction}$ and the perceptual loss [124] $L_{Perceptual}$ are utilized

for the regularization. Additionally, adversarial loss $L_{Adversarial}$ is also utilized to guarantee that the reconstructed image looks authentic. The whole process can be summarized as follows

$$\begin{aligned}\min_{E} ||x - G(E(x))||_2 + \lambda_1 ||F(x) - F(G(E(x)))||_2 \\ - \lambda_2 E_{x\sim p_{data}(x)}[\log D(G(E(x)))]\end{aligned} \tag{3.2}$$

$$\begin{aligned}\min_{D} E_{x\sim p_{data}(x)}[\log D(G(E(x)))] - E_{x\sim p_{data}(x)}[\log D(x)] \\ + \frac{\gamma}{2} E_{x\sim p_{data}(x)}[||\nabla_x D(x)||_2^2]\end{aligned} \tag{3.3}$$

where $p_{data}(x)$ indicates the real data distribution, x is the real image, E represents the encoder, F represents the VGG feature extraction [232], and λ_1, λ_2 and γ are weights for the perceptual loss, the adversarial loss, and the gradient penalty [98].

Since the inverse mapping via the encoder(E) will not always be perfect, in order to get the optimal inverse latent code, we apply another optimization on the latent code. This optimization will update the latent code based on the reconstruction loss and the perceptual loss within the neighborhood of the original encoded vector (the encoder regularization). The optimization process can be described as below

$$\begin{aligned}z^{inv} = \min_{z} ||x - G(E(x))||_2 + \lambda_3 ||F(x) - F(G(z))||_2 \\ + \lambda_4 ||z - E(G(z))||_2\end{aligned} \tag{3.4}$$

where z^{inv} is the optimized inverse code, λ_3 and λ_4 are weights for the perceptual loss, and the code reconstruction loss (i.e., the encoder regularization). This optimization metric can be computed using the whole image region (for image reconstruction) or the region of interest (for image manipulation).

Medical image synthesis

In general, informative images lie on a manifold. Through the GAN training, the generator(G) learns a transformation from the latent space to the image space, imitating the real image manifold of the training dataset. Thus, we can utilize this learned transformation to generate images authentic to the real images. First, the latent code z will be sampled from the latent space. Then, the generated image $x = G(z)$ is produced by the generator.

Using the learned transformation, we can also generate similar medical images. As a manifold, the nearby images on the manifold are similar to each other. Therefore, we can sample a series of latent codes z_i on a closed path C, then passing these latent codes into the generator(G), we can obtain a series of gradually and continuously morphing images x_i.

$$x_i = G(z_i), z_i \sim C \tag{3.5}$$

Attribute manipulation

While training the encoder(E), without the discriminator(D), the encoder and the generator form an autoencoder [210, 178]. The training encourages the encoder to embed useful image attributes into the latent code. Since the generator is pretrained under the GAN, the generator has learned how to reconstruct the embedded image attributes with proper tissue context.

In order to manipulate the image attributes, we first need to combine the desired image attributes into one assembled image x'. The combination can be achieved by merging image patches P_i which contain the desired image attributes.

$$x' = \bigcup_{i=1}^{n} P_i \tag{3.6}$$

Then this assembled image x' will be encoded by the encoder, $z' = E(x')$, obtaining the corresponding image attributes latent code z'. The generator will finally reconstruct those image attributes with proper tissue texture, $x_{reconstruct} = G(z')$.

Since the image with all desired attributes may not exist on the image manifold, the reconstructed image may not have the exact desired attributes as we designed. The final optimization (shown in Equation 3.4) can be conducted on the region where the attributes need to be accurate. The pipeline for attribute manipulation is shown in Fig 3.3.

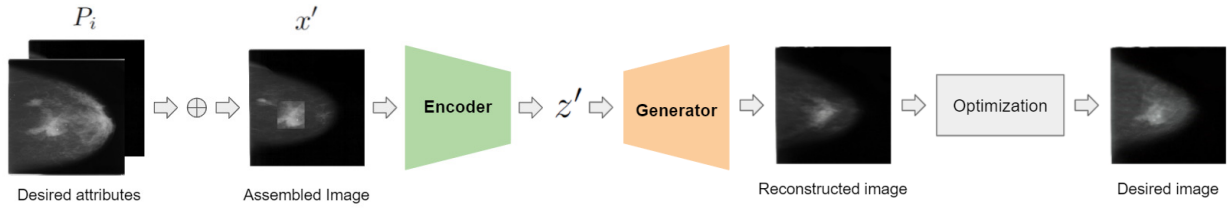


Figure 3.3: **Attribute manipulation pipeline.** First, desired image attributes are combined by merging image patches that contain those attributes. Then, the corresponding latent code is produced by the encoder. The generator reconstructs the image with desired attributes. At last, the desired image can be obtained after the final optimization.

3.4 Experiments and Results

Implementation details

For the Generative Adversarial Network (GAN), we adopt StyleGAN [127]. The training is progressive. Starting from 8×8, the latter resolution blocks are added progressively after the previous blocks finish training. The output image resolution is 256×256. While training

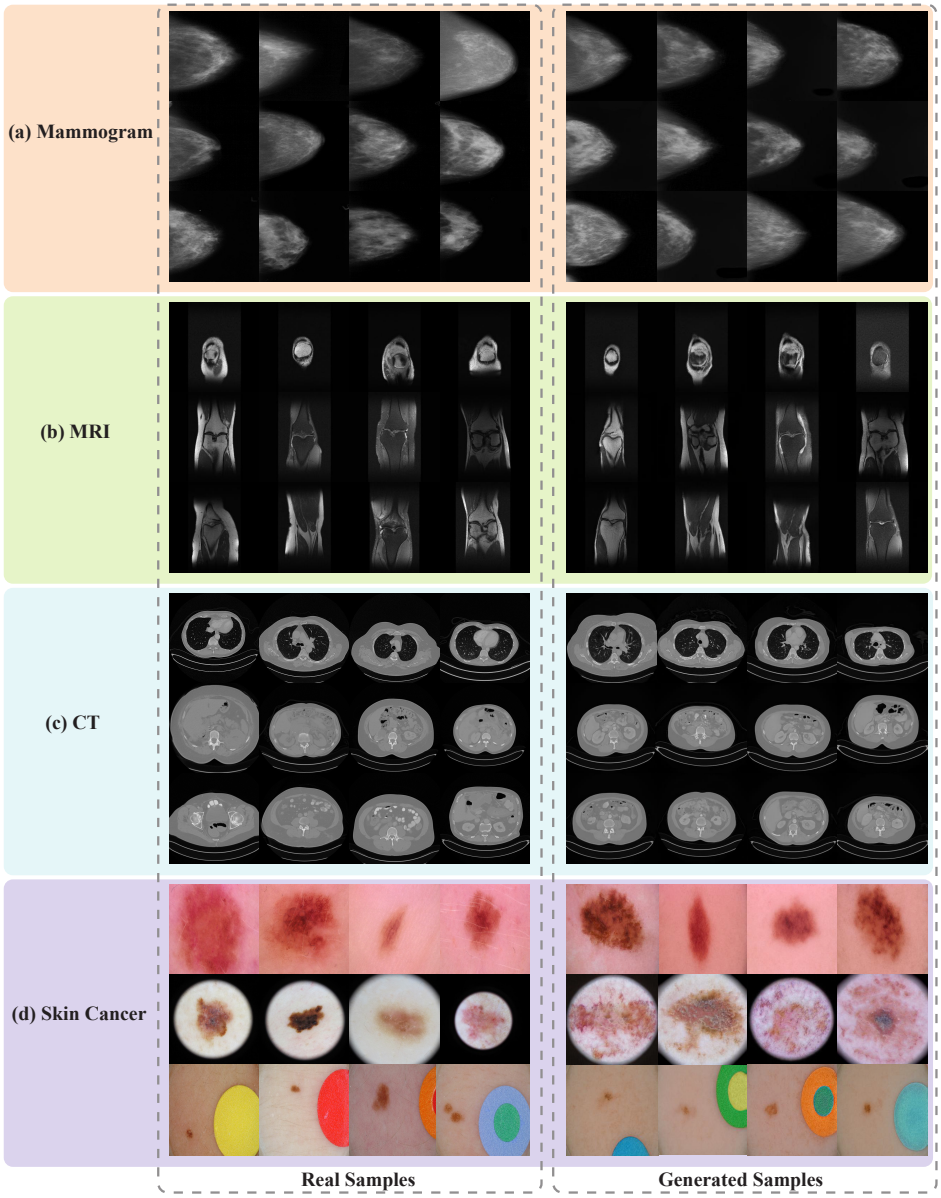


Figure 3.4: **GAN generated results.** The generated results for different medical image modalities. Comparing the real samples to the generated samples, it is clear that the generator has learned how to imitate tissue texture, tissue distribution, tissue shapes, and color distribution. Thus, it appears to generate authentic images (see below for psychophysical results confirming this).

the encoder, the generator is fixed. Only the encoder and discriminator parameters are updated. For the perceptual loss, VGG [232] $conv4\_3$ feature layer is utilized. As for the hyperparameters, $\lambda_1 = 0.00005$, $\lambda_2 = 0.1$, $\lambda_3 = 0.00005$, $\lambda_4 = 2$, and $\gamma = 10$. We use the Adam optimizer [135] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate is set to 0.0001. Pytorch is utilized for coding.

For mammogram images, we use DDSM [23] dataset which contains $2,620$ normal, benign, and malignant cases. Only the benign and malignant cases are utilized for training. For MRI images, we utilize fastMRI [287] multi-coil dataset which contains 7135 images. For CT images, DeepLesion [283] dataset is used. We utilize the abdomen image dataset which contains 14601 images. For skin cancer images, we use images from ISIC Archive [1] which contains 69445 images in total.

GAN generated results

For different medical image modalities, we train the whole network separately using corresponding datasets. After the GAN part has been trained, we randomly sample latent codes z and pass them to the generator. The generated results for Mammogram, MRI, CT, and Skin Cancer are shown in Fig 3.4. Compared to the real samples on the left, the generated samples on the right appear very similar, and this holds across different medical image modalities. It is clear that the generator has learned the semantic statistics of the training dataset for different medical image modalities. The generator can generate authentic tissue texture, tissue distribution, tissue shapes, and color distributions. Moreover, not only can the generator reconstruct the original medical images, but also it can produce novel and authentic medical images which do not actually exist in the real world.

Since the GAN training in general learns the manifold of the training dataset, we can also generate gradually and continuously morphing medical images for certain experiments. First, the latent codes need to be sampled from a closed path in the latent space. To do so, we randomly pick three anchor points in the latent space and calculate the interpolations between each pair of them. Then, passing those codes to the generator, we can obtain the gradually and continuously morphing medical images. The result is shown in Fig 3.5. Due to the space limit, we only show three interpolations between each pair; arbitrarily fine grained interpolations can be created between any number of pairs.

Attribute manipulation

Our proposed model can generate desired medical images by manipulating the image attributes. For illustration, we show how we generate mammograms with the desired lesion patch and desired breast shapes. The results are shown in Fig 3.6.

First, we combine the desired image attributes, i.e. the lesion patch (Fig 3.6A) and shape templates (Fig 3.6B), by merging the lesion patch and shape templates directly. Then we en-

[1] https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main

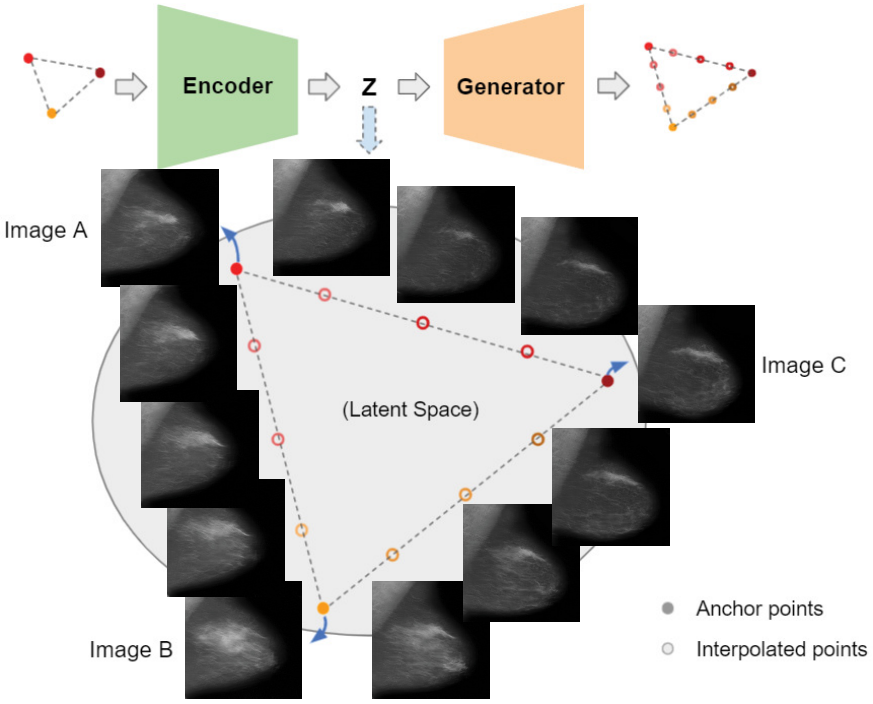


Figure 3.5: **Interpolation results.** Here, we show a mammogram loop gradually changing among three anchor images. The mammograms between two of the anchor images are generated by passing the interpolated codes of those two anchor images to the trained generator. Any number of interpolated images between any pair of anchors can be created.

code these intermediate combined images (Fig 3.6C) using the encoder and pass the codes to the generator. The reconstructed images from the generator are shown in Fig 3.6E (without optimization). It is clear that the shapes are already the same as the shape templates and the overall texture is authentic. But the desired lesion texture is not maintained. After the last step of optimization over the lesion patch, as it is shown in Fig 3.6F, the lesion texture is recovered. We also compare the results with the ones produced by a traditional image blending method. As it is shown in Fig 3.6D, the transition region between the lesion texture and the shape template background is not natural. Our proposed method can maintain both the breast shape and the lesion texture while generating authentic tissue texture.

Human evaluation

To verify the authenticity of the generated images for different medical image modalities, we conducted an online psychophysical experiment, recruiting both untrained participants (i.e. no knowledge of medical imaging) and experts (e.g. radiologists or practicing clinicians who routinely read radiographs).

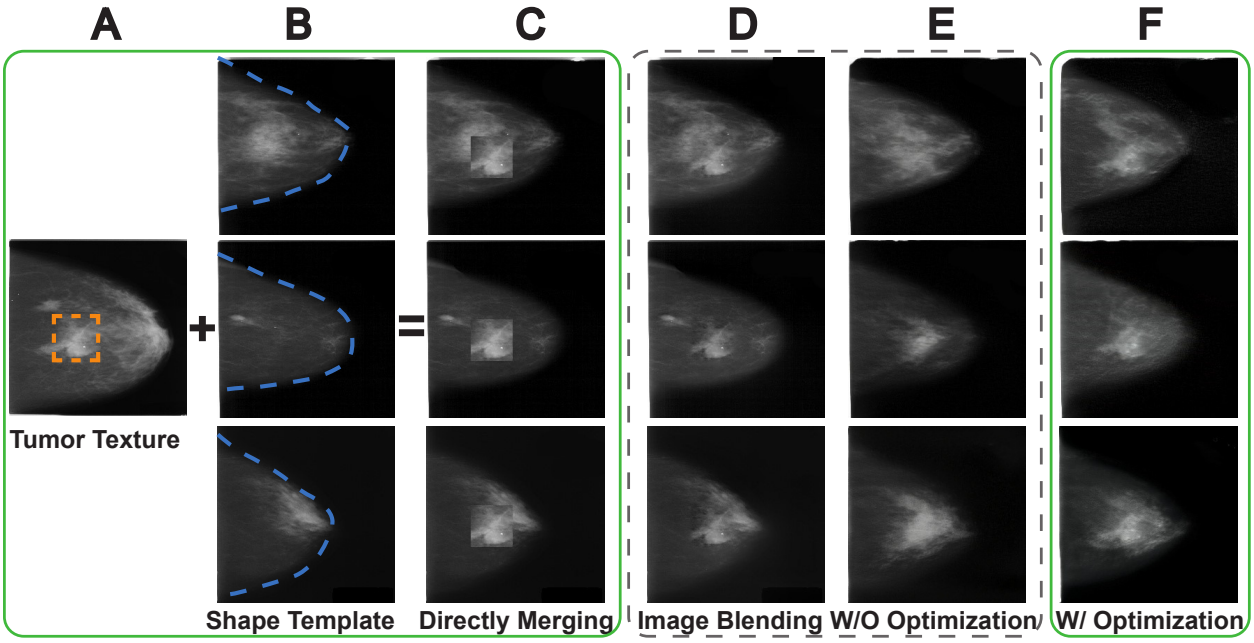


Figure 3.6: **Attribute manipulation results.** The desired image attributes are combined by merging the corresponding image patches (in Column A and B) directly. Then, the encoder will encode the manipulated image attributes, and the generator will produce the output correspondingly. After the final optimization, it is clear that the proposed method can generate the mammograms with the desired lesion texture and breast shape (Column F), compared to the results from the traditional image blending method (Column D) and the proposed method without the final optimization (Column E).

Participants

Six untrained observers (3 females, age range: 22-25) and seven experts (3 females, age range: 32-39) participated in the mammogram online survey. Two experts were excluded from the mammogram online survey (one dropped out and the other gave the same response on every trial). Five untrained observers (3 females, age range: 23-25) and seven experts (3 females, age range: 28-40) participated in the CT online survey.

All subjects reported to have normal or corrected-to-normal vision. Participants voluntarily participated and were offered $15 per hour as optional compensation. In our experience, radiologists typically refuse this modest compensation. The experiments were approved by the Institutional Review Board at the University of California, Berkeley. Participants provided informed consent.

Stimuli

For the mammogram online survey, 50 real mammograms and 50 fake (model generated) images were included. For the CT online survey, 50 real CT images and 50 fake CT images were presented. All the images were randomly selected from the corresponding data pools.

Procedure

The task was to rate each image from 0 (fake/generated image) to 10 (real image) in the data pool. Each individual image was shown for 5 seconds, and observers were asked to respond as quickly as possible. The experiment was self-paced, so observers viewed the stimuli as long as they wanted (up to 5 sec), and they did not have time limit for giving responses. To ensure that participants did not randomly guess (or lapse), a small number of repetitive trials were also included in the online survey to establish a baseline test-retest reliability estimate. We compute the similarity among those repetitive trials.

Results

The results for mammogram and CT images in terms of the Receiver Operating Characteristic (ROC) curves are shown in Fig 3.7. For both untrained participants and radiologists, and for both mammogram and CT images, their performance curves are near the diagonal (i.e. the chance level performance region), indicating that the generated medical images appeared authentic. The area under the curve (AUC) can also confirm the chance-level performance. The mean AUCs are 0.52 (p=0.395, permutation test) and 0.60 (p=0.126, permutation test) for untrained participants and radiologists respectively in mammogram online survey. The mean AUCs are 0.42 (p=0.888, permutation test) and 0.42 (p=0.844, permutation test) for untrained participants and radiologists respectively in CT online survey. As shown in the permutation tests, the large p-values indicate that performance is not statistically different from random performance.

Although the observers were not able to accurately discriminate real from fake images, this does not mean that observers randomly responded or failed to pay attention to the task. To confirm this, we calculated the test-retest reliability of each observers responses for repeated images. From the small number of repeated trials, the average test-retest similarity is 0.65, indicating "good" consistency. For a near-threshold task, the noise ceiling is not 1, and 0.65 is "good" in the sense that it is statistically reliable and significant [76, 77, 38]. The similarity is computed using Sokal-Michene metric [288]. It is noteworthy that observers can have high test-retest reliability despite low sensitivity (low AUC). The test-retest reliability indicates that observers tended to make the same judgments in repeated trials: they consistently confused some real (fake) images as being fake (real). This resulted in low sensitivity (low AUC) but consistent responses ("good" test-retest reliability).

We also provide the results of MRI and Skin Cancer images in the Appendix 3.6 to avoid redundancy. Results indicate the same conclusion that the generated medical images appeared authentic.

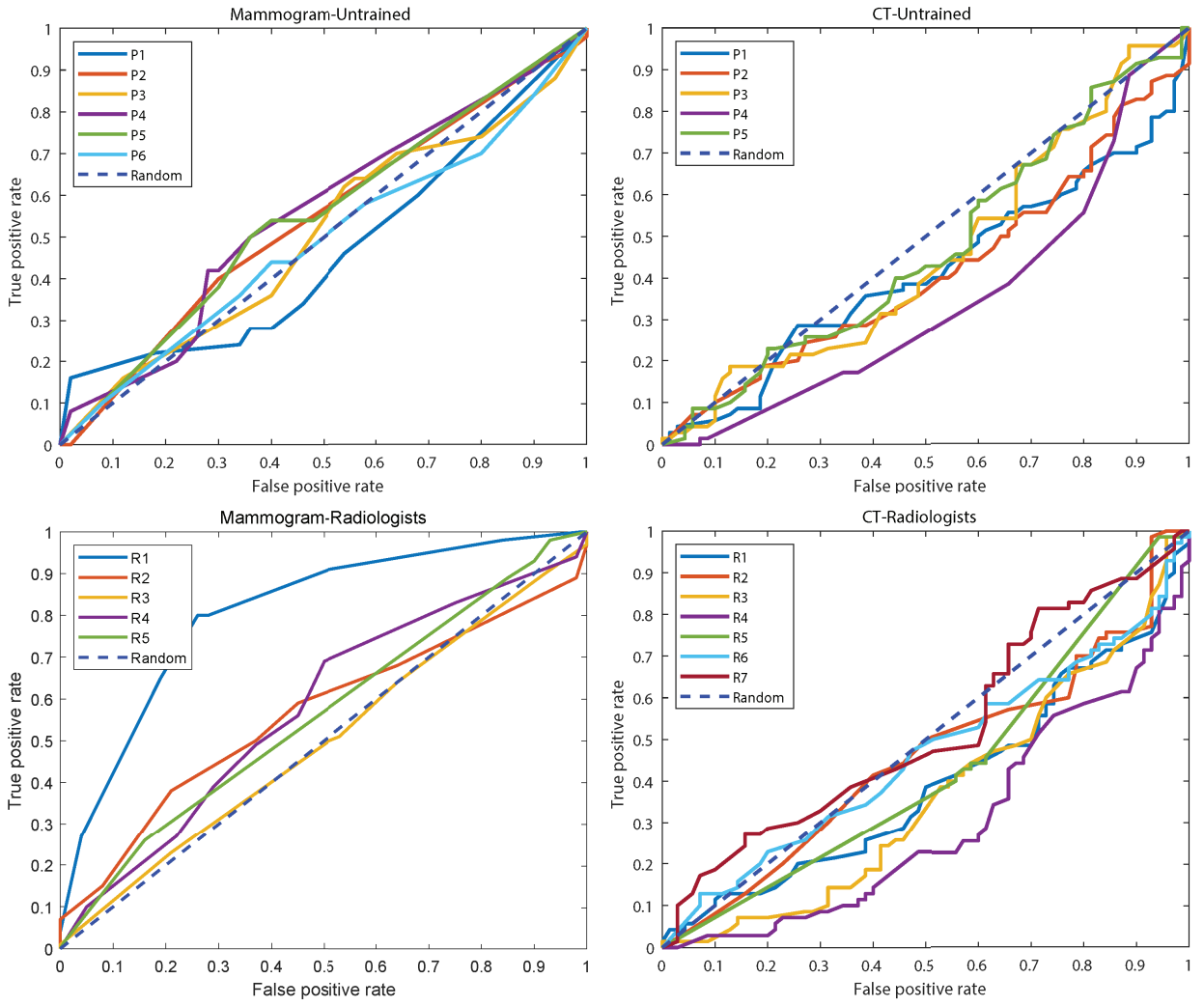


Figure 3.7: **Human evaluation results.** Participant performance is shown in the Receiver Operating Characteristic (ROC) curves. It is clear that their performance is near chance level (curves near the diagonal region), indicating that the generated medical images are authentic. Here, $P_1 - P_N$ and $R_1 - R_N$ represent different untrained observers and experts in corresponding experiments.

Limitations

Online studies have a range of potential limitations [17]. However, it has been well documented in the literature that online studies can reveal even very subtle psychophysical phenomena reliably, and the methods are established [225, 209, 51]. In our online experiment, variations in the environment or monitor settings that might occur could add noise to the data, but they wouldn't generate the high test retest reliability we found, or the consistent pattern of results. The growing literature on internet-based psychophysics is consistent with this [225]. Moreover, we believe that our data adds a unique perspective on this issue: the advantages of online experiments are pronounced in cases where subjects are rare and/or very expensive to recruit, as is the case with the experienced and highly trained radiologist observers reported here. Future studies should consider online data collection for medical image perception tasks, in order to broaden representation, diversity, and improve sample sizes.

Another consideration with the experiments here is the images were viewed for at most 5 seconds. The experiment was self-paced, and the participants could view the images as long as needed to make a choice, but this was limited to 5 seconds maximum viewing. There are both theoretical and empirical reasons that 5 seconds is likely to be sufficient for the task (see Appendix C), but it is conceivable that performance could change if observers were forced to view the images for prolonged periods of time. Future experiments should therefore examine the temporal integration of the visual processes that contribute to discrimination of near-metameric medical images.

Perceptual loss

Currently, perceptual loss has been utilized as a similarity metric in many computer vision tasks [157, 293, 119, 290]. In this section, we investigate the perceptual loss as a similarity metric in medical imaging domain. We compare its results with the results of Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR), which are two common similarity metrics.

In the experiment, we utilize random samples from mammogram, MRI, CT, and skin cancer images as reference images (256×256). First, we apply traditional image distortions on those reference images, such as Gaussian blur, contrast distortion, geometric distortion, spatial shifting, and spatial rotation. Then, we calculate the similarity measurements for different outputs from traditional image distortions with respect to the reference images. Detailed computation algorithms can be found in Appendix 3.6.

For quantitative comparison, we show the similarity measurement results in the following tables. For the SSIM and PSNR metrics, the larger the measurement is, the more similar it is between the measured image and the reference image (indicating by $\uparrow$). For perceptual metric, the smaller the measurement is, the more similar it is between the measured image and the reference image (indicating by $\downarrow$). Table 3.1, Table 3.2, Table 3.3, Table 3.4 show the similarity measurements for mammogram, MRI, CT, and skin cancer images respectively.

Table 3.1: Similarity Measurements for Mammogram Images

	Image 1	Image 2
SSIM↑	**0.93**	0.89
PSNR(dB)↑	**38.99**	30.61
Perceptual↓	0.96	**0.64**

Table 3.2: Similarity Measurements for MRI Images

	Image 1	Image 2
SSIM↑	**0.84**	0.68
PSNR(dB)↑	**34.42**	33.01
Perceptual↓	3.89	**2.96**

Table 3.3: Similarity Measurements for CT Images

	Image 1	Image 2
SSIM↑	**0.54**	0.24
PSNR(dB)↑	**31.04**	29.91
Perceptual↓	27.77	**7.42**

Table 3.4: Similarity Measurements for Skin Cancer Images

	Image 1	Image 2
SSIM↑	**0.87**	0.74
PSNR(dB)↑	**36.26**	31.58
Perceptual↓	3.15	**1.99**

For qualitative comparison, we compare the similarity measurement between Gaussian blur outputs (Fig 3.8 Image 1 Column) and the outputs from the rest of the traditional image distortions (Fig 3.8 Image 2 Column). We first asked human participants to give their judgements of which image was more similar to the reference image. The results are labeled with green check marks as shown in Fig 3.8. Then, according to the similarity measurements, we select the images which are preferred by SSIM/PSNR or perceptual loss metric. It is clear that SSIM and PSNR disagree with human judgements. However, the similarity decisions from the perceptual loss metric are consistent with human judgements. Thus, the perceptual metric is more suitable for the similarity measurement in medical imaging area.

3.5 Discussion

In this paper, we utilize Generative Adversarial Networks for medical image generation. Our results demonstrate generalizability of the proposed approach across different modalities,

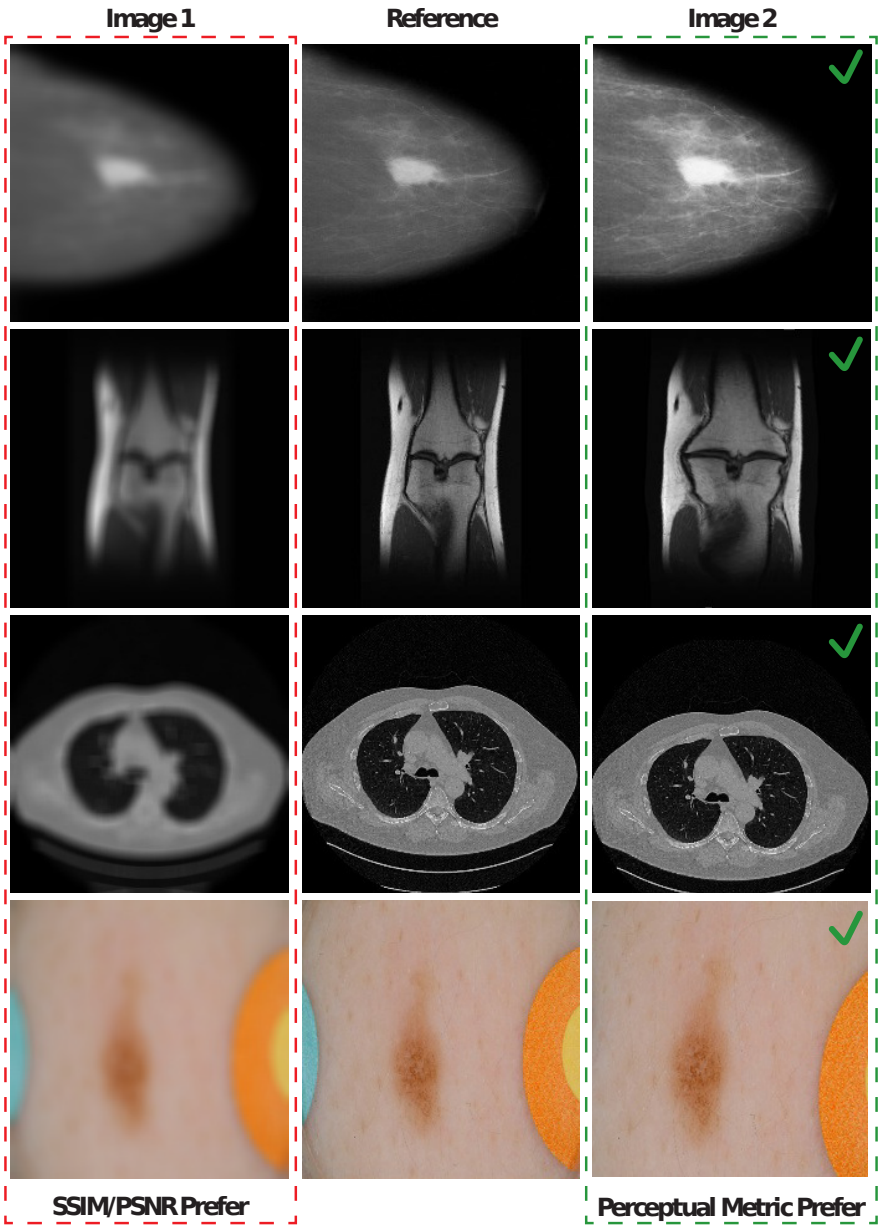


Figure 3.8: **Which image is more similar to the reference?** Image 1 Column shows the distortion by Gaussian blur. Image 2 Column shows the distortions by contrast distortion, geometric distortion, spatial shifting, and spatial rotation respectively. The human judgements are marked using green ticks. It is clear that SSIM/PSNR results disagree with human judgements while perceptual metric agree well with human judgements.

such as mammogram, MRI, CT, and skin cancer images. We also manipulate the generated images such that they contain desired attributes. Compared to traditional image blending methods which mainly edit locally, our proposed method not only embeds the desired image attributes but also edits the surrounding tissue texture accordingly to make the overall tissue texture distribution reasonable. Through adversarial training, the GAN model here learns an estimated manifold which is similar to the image manifold of the training dataset. This estimated manifold well-characterizes the semantic statistics of the training dataset, such as the tissue texture, tissue distribution, tissue shapes, and color distribution. Thus, once the contents of certain regions is altered, the GAN knows how to edit the surrounding region to match the semantic statistics of the training dataset, producing authentic manipulated images.

Our model can generate a vast range of possible stimuli that accomplish a range of specific and controllable goals. For example, the model can output specific body part shapes, lesion types and locations, background and tissue textures, etc. Additionally, our model is capable of generating gradually and continuously morphing medical images. In certain medical image perception tasks, such as visual search [55, 278], visual detection and recognition [189], and decision making [250, 249], this kind of controllable medical image stimuli can be very useful. The intrinsic problem using real medical image data is that individual differences are substantial: it is not realistic to collect gradually morphing medical images from real medical image data (e.g., finding a sequence of naturally occurring tumors that smoothly morph between shapes or textures is highly unlikely). Using our proposed method, we can generate any number of authentic medical image stimuli that gradually morph. Moreover, all the images are generated via interpolation, which allows us to control the grain of the morphing.

For the perceptual loss metric, researchers [290] have found that traditional similarity metrics, such as SSIM and PSNR, are not consistent with human perception of typical natural images. But deep neural network based perceptual metrics can, surprisingly, agree with human judgement. Through experiments, we find the same conclusion in medical imaging domain as well that perceptual metrics preferred medical images are more perceptually similar to the reference images compared to traditional similarity metrics. Thus, perceptual loss metric provides an important measurement for what counts as similar in medical imaging. Using perceptual loss metric as the similarity measurement, we can also generate metamers for any specific medical image. The metamers are a cluster of perceptually similar images which have been widely used in perception researches.

Medical image perception research is growing fast. Typical approaches directly or indirectly assume that computer vision will be an alternative to clinical practice. Our study introduces an additional but very different perspective, which is to use computer vision to improve research on medical image perception. Clinicians will not be replaced anytime soon (if ever). To help clinicians make better judgments, we need to understand clinician perception, cognition, and decision. That requires having stimuli (datasets) that are simultaneously realistic (from the perspective of clinicians) and also controllable. Without this, it will be impossible to make the connection between the cognitive mechanisms that clinicians possess,

and their diagnostic success in their practice.

Interestingly, the model and morphing approach we present here could be readily extended to three-dimensional volumetric images. Volumetric medical imaging is increasingly standard practice in clinical setttings. The GAN model and morphing approach can be combined in future work to flexibly create volumetric data sets. Moreover, the GAN model is currently unconditioned. We can also change it to conditional GAN model such that changing certain part of the latent code (not through the encoder) can directly modify corresponding attributes of the output image.

3.6 Conclusion

In this paper, we propose to use Generative Adversarial Network (GAN) for medical image generation. We test our method on various medical image modalities such as mammogram, MRI, CT, and skin cancer images. Human evaluations verify the success of our method. We also adopt a controllable approach to manipulate the attributes of the generated images in order to meet certain experimental configurations. In the experiments, we successfully generate mammograms with the desired lesion texture and breast shape. The same approach can also be applied to MRI, CT, skin cancer images, and other medical imaging modalities. Finally, we compare traditional similarity measurements with the perceptual metric in medical imaging. We find that the perceptual metric performs better than the traditional similarity metrics such as SSIM and PSNR.

Appendix A. Similarity Measurements

A.1 SSIM

The Structural Similarity Index Measure (SSIM) is computed over various patches of an image. The measure between two patches x and y of the same size is:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3.7}$$

where μ_x is the average of x, μ_y is the average of y, σ_x^2 is the variance of x, σ_y^2 is the variance of y, σ_{xy} is the covariance of x and y, $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are two variables to stabilize the division with weak denominator with $L = 2^{\#bits\ per\ pixel} - 1$, $k_1 = 0.01$, and $k_2 = 0.03$.

A.2 PSNR

Given a $m \times n$ reference image I and its distorted version K, the PSNR is defined as:

$$PSNR = 20\log_{10}(MAX_I) - 10\log_{10}(MSE) \tag{3.8}$$

where MAX_I is 255 for 8-bit images, and the MSE is computed as:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \tag{3.9}$$

A.3 Perceptual loss

We utilize the same perceptual loss as [124]. The loss network is VGG-16 [232]. For the reference image r and the distorted image x, the perceptual loss is defined as:

$$L(x,r) = \lambda_c l_{feat}^{\phi,j}(x,r) + \lambda_s l_{style}^{\phi,J}(x,r) \tag{3.10}$$

where λ_c and λ_s are scalars. In the experiment, we set $\lambda_c = 1$ and $\lambda_s = 1 \times 10^5$. ϕ represents the VGG network. $l_{feat}^{\phi,j}(x,r)$ is the feature reconstruction loss. Let $\phi_j(x)$ be the activation of the jth layer of the network ϕ with a shape of $C_j \times H_j \times W_j$. The feature reconstruction loss is defined as:

$$l_{feat}^{\phi,j}(x,r) = \frac{1}{C_j H_j W_j}||\phi_j(x) - \phi_j(r)||_2^2 \tag{3.11}$$

The style reconstruction loss is defined as:

$$l_{style}^{\phi,J}(x,r) = ||G_j^\phi(x) - G_j^\phi(r)||_F^2 \tag{3.12}$$

where $G_j^{\phi}(x)$ is the Gram matrix with a shape of $C_j \times C_j$. The elements of the Gram matrix can be computed as:

$$G_j^{\phi}(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'} \tag{3.13}$$

Appendix B. Human Evaluation of Generated MRI and Skin Cancer Images

We also collected human evaluation experiment data for MRI and Skin Cancer images. For the MRI experiment, four observers (1 expert, age range: 25-39) participated. For the Skin Cancer experiment, five observers (1 expert, age range: 20-39) participated. Unlike CT and mammogram image experiments, we could not recruit sufficient experts for MRI and Skin Cancer online surveys. All experiments were approved by the Institutional Review Board at UC Berkeley and the participants provided informed consent. Stimuli were 50 real and 50 fake corresponding images. All participants followed the same experimental procedures as described in Sec 3.4.

The results for MRI and Skin Cancer images in terms of the Receiver Operating Characteristic (ROC) curves are shown in Fig 3.9. The mean area under the curves (AUCs) are 0.57 (p=0.241, permutation test) and 0.62 (p=0.123, permutation test) for MRI and Skin Cancer respectively. Notably, although we did not have experts for these MRI and Dermatology tests, we did have one trained radiologist participate and their data echoes the untrained observers, all of which are consistent with the CT and mammogram data.

Appendix C. Stimulus Duration Considerations

There are both empirical and theoretical reasons for limiting the display to 5 seconds, and the empirical results confirm that 5 seconds was more than sufficient for observers to reach a reliable decision.

First, previous research has demonstrated that radiologists can reliably discriminate radiographs in well under 1 second [151, 30, 31, 195, 88, 186, 70, 69, 122, 116]. In our experiment, we provide far more time than 1 second. Moreover, in self paced studies with static radiographs, radiologists often spend less than 5 seconds [227].

Second, our results show that accuracy does not vary with decision time. The decision time is reported as the time from the first viewing of the page to the final "submit" click by the observer. This is a conservative estimate of the decision duration. The relation between the error and decision time is shown in Fig 3.10. The fitted line reveals that error and decision time are not correlated; more time did not make observers more accurate. Moreover, in this experiment, 60.0% of the decisions were made before stimuli disappeared. Notably, the peak of response density does not occur at the 5 seconds boundary. It occurs around 2 seconds,

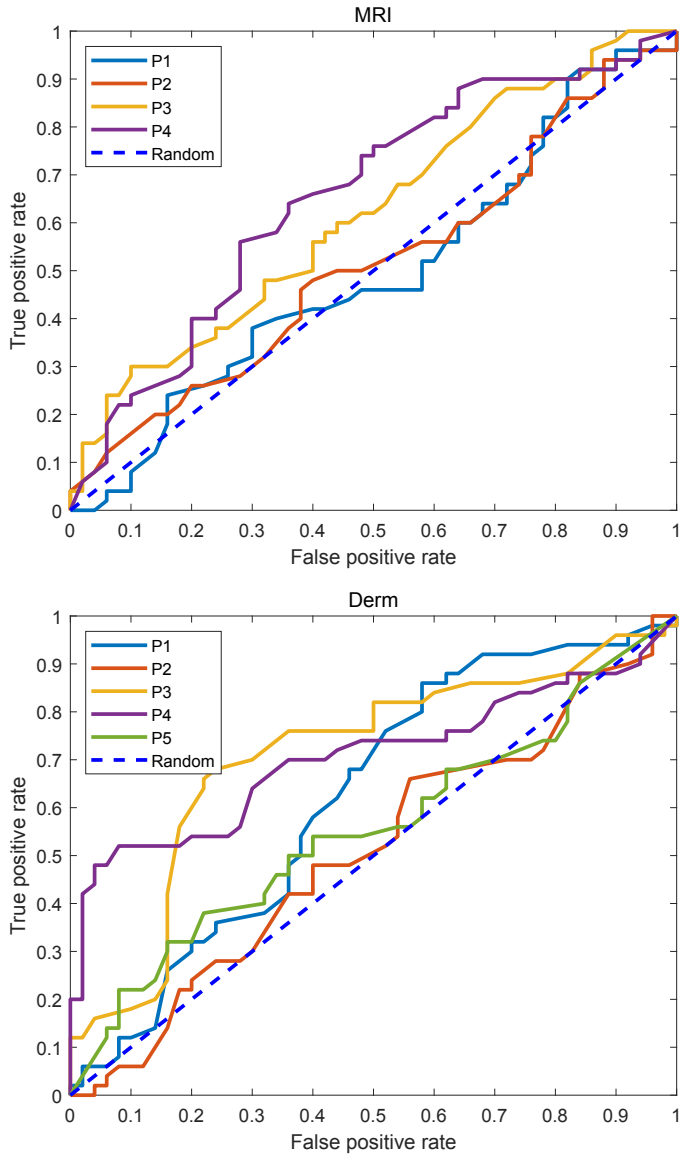


Figure 3.9: **Human evaluation results for MRI and Skin Cancer images.** Participant performance is shown in the Receiver Operating Characteristic (ROC) curves. It is clear that their performance is near chance level (curves near the diagonal region), indicating that the generated medical images were authentic. Here, $P_1 - P_N$ represent different untrained observers and experts in corresponding experiments.

which indicates that the 5 seconds stimulus duration limit does not pressure participants' decision.

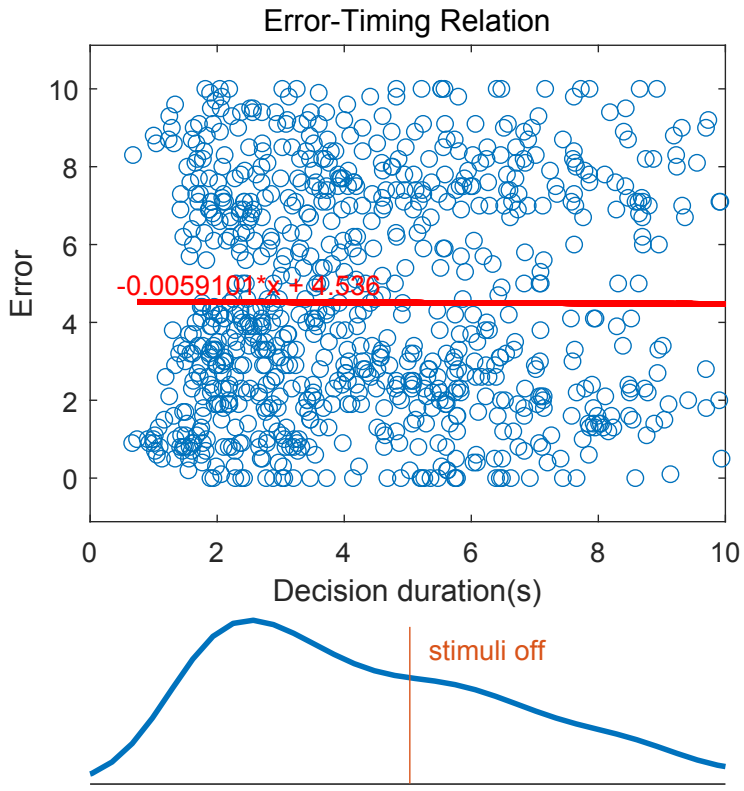


Figure 3.10: **Error-Timing Relation.** The scatter plot shows the raw data of participants' error and their decision duration. We fit a linear function to reveal the relation between them. It is clear that the error and their decision time are not correlated. The bottom density distribution represents the distribution of participants' decision time. The orange line indicates the time point when stimuli disappeared. In this experiment, 60.0% of the decisions were made before stimuli disappeared.

Third, the significant test-retest reliability demonstrates that observers were consistent in their responses. If exposure duration limited performance, it would add noise and that test-retest reliability would be low [70].

Together, all of these considerations suggest that the duration of the image was probably not the limiting factor. From the examples here, it seems visually clear that scrutinizing the real and generated images for more than a few seconds does not make them appear more or less similar. This hints that the metameric quality of the images is not due to a time constraint. Nevertheless, we did not force observers to scrutinize the images for

more than 5 seconds, and it is conceivable that forcing an extended viewing of the images could improve performance. For this reason, it will be valuable in future studies to examine the temporal integration of the visual processes that contribute to discrimination of near-metameric medical images.

Chapter 4

Improve Image-based Skin Cancer Diagnosis with Generative Self-Supervised Learning

4.1 Introduction

Skin cancer is increasingly becoming a public health concern. It is the most common cancer in many countries including the United States [100, 99]. In certain developing countries, e.g. Brazil, the situation is even worse [230]. Currently, the best method for early detection of skin cancer is to track the changes in skin lesions. But it is hard to implement in developing countries due to the scarcity of experts and their availability in remote areas. Thus, in Brazil, certain types of skin cancer, such as basal cell cancer (BCC) and squamous cell cancer (SCC), are usually diagnosed at advanced stages [231].

Teledermatology provides a promising technology for monitoring skin cancer [177, 74, 37, 33]. This is mainly attributed to the accessibility and ubiquity of smartphones. App users can be diagnosed remotely by a group of dermatology experts without meeting the dermatologist in-person. Dermatologists can therefore serve not only their local patients but also patients far from their working site. Most available mobile health apps for skin cancer detection utilize machine learning algorithms which heavily rely on handcrafted features [33]. Currently, deep learning has been widely utilized in medical image analysis, where features are learned automatically by neural networks.

Deep learning has achieved great success in general image recognition tasks [106, 117, 243], and researchers have also applied deep learning methods to medical image analysis, in particular; For example, in lesion classification [284, 169], lesion detection [152, 188], lesion segmentation [286, 123], and so on. While the models become deeper and deeper, a data scarcity issue emerges. Supervised models need ground truth labels along with image data. For traditional computer vision tasks, researchers have created publicly labeled datasets, such as ImageNet [52], Microsoft COCO [165], and CIFAR-10 [143]. However, collecting

and labeling those data is tedious and expensive. In medical image arenas, researchers usually collect data from hospitals. But the requisite data processing procedures, such as file categorization, annotation, and data de-identification, are time-consuming. Moreover, it requires experts (e.g., extensively trained and highly paid radiologists or pathologists) to perform meticulous annotations that is even harder or impossible [272]. For teledermatology and other mobile health technologies, the data scarcity issue may be even worse due to the lack of users at the early stage.

With limited labeled data, few-shot learning [234, 239, 236] has become popular. In few-shot learning, we are given some categories where each category only has a limited number of images. Few-shot learning methods [234, 239, 236] utilize the knowledge learned from some base categories which are different from the given categories. Hence, it is possible to learn a novel category by showing one or several images [71]. Few-shot learning has already been successful in handwritten characters, birds, dogs, and other natural images [234, 112]. However, it is still difficult to apply the few-shot learning techniques to medical images because of the lack of base category data.

Unsupervised learning is a promising approach to reduce the labelling cost for training deep neural networks [281, 196, 34, 108, 96]. Unsupervised learning methods [281, 196, 34, 108] aim to learn a useful representation directly from unlabeled data. Then, the learned representation can be reused for supervised learning with limited labeled data [34], thus reducing the cost of data labeling. For traditional computer vision tasks, there are usually sufficient unlabeled data for self-supervised learning. However, in clinical practice, medical images with lesions are anomalies, and are therefore rare and naturally hard to find and collect. In particular, certain skin cancers, such as Merkel cell carcinoma, are so rare that it is impossible to find sufficient data [109]. Therefore, it is still difficult for medical image tasks to obtain adequate data for self-supervised learning. Notably, the effectiveness of the learned representation with self-supervised learning depends on the size of the unlabeled dataset [108]. It is thus critical to augment the unlabeled data when the amount of unlabeled data itself is limited in medical image analysis.

To improve the performance of self-supervised learning on skin cancer images, we propose to use a Generative Adversarial Network (GAN) to augment the unlabeled dataset. Generative Adversarial Networks have been utilized to create a range of authentic images [198, 129, 127, 128], including faces, cars, landscapes, and so on. Trained on real image datasets, a GAN can learn to estimate the manifold that represents the training images. Through training, the learned manifold and the real image manifold can be practically aligned. In doing this, GANs can learn both local and global statistics of the real images from the training dataset, and the generated images can have similar semantic content to that of real images. However, it is unclear whether GAN generated images can be utilized to boost the performance of self-supervised learning on skin cancer images.

In this paper, we investigate how to leverage GAN generated skin cancer images to improve the self-supervised learning performance on skin cancer classification task for teledermatology (Fig. 4.1). We first train StyleGAN [127] on unlabeled data to generate high quality skin cancer images which are semantically similar to the unlabeled training dataset.

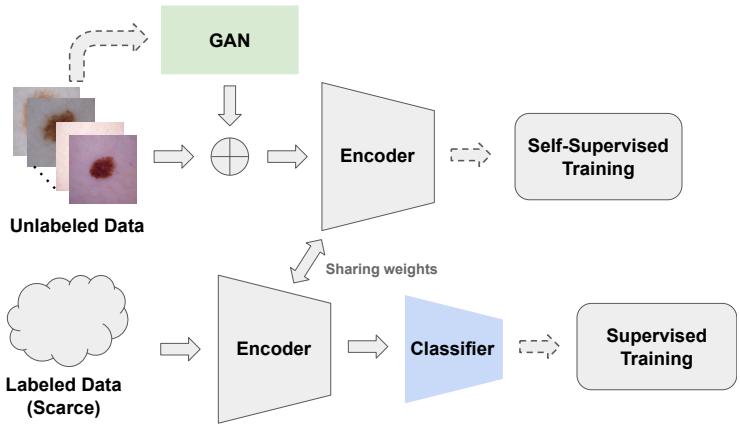


Figure 4.1: Proposed method. We first train StyleGAN [127] on unlabeled data to generate high quality skin cancer images which are semantically similar to the unlabeled training dataset. Then, we train a feature encoder via self-supervised learning. At last, a linear classifier is attached to the feature encoder to test the performance of skin cancer classification on the scarce labeled data.

Then, we train a feature encoder via self-supervised learning using the augmented training dataset which includes the StyleGAN generated images and the labeled training images. At last, a linear classifier is attached to the feature encoder to test the performance of skin cancer classification on the scarce labeled data.

Contributions

In this paper, we propose to use a Generative Adversarial Network (GAN) to augment training data for self-supervised learning on skin cancer images. The contributions of this work can be summarised as follows:

- We propose to use StyleGAN [127] for data augmentation to boost the self-supervised skin cancer classification accuracy. To the best of our knowledge, it is the first time that GAN-based data augmentation is applied to self-supervised learning algorithms for skin cancer image classification tasks.
- The self-supervised skin cancer classification accuracy can be boosted by 11.17% on BCN20000 [45] and 3.07% on HAM10000 [251] after StyleGAN-based data augmentation.

4.2 Related work

Generative Adversarial Networks: The GAN is a promising image synthesis model. The model consists of two networks, a generator network and a discriminator network. Inspired by game theory, those two networks are trained in an adversarial process where the generator generates fake but authentic images to fool the discriminator and the discriminator discriminates between the real and fake images repeatedly [95]. Conceptually, the training process can be described as a minmax game, which is formulated as follows:

$$\min_G \max_D E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{4.1}$$

where G represents the generator, D represents the discriminator, $p_{data}(x)$ indicates the real data distribution, and $p_z(z)$ indicates the noise vector distribution. While generating new images, the generator takes in noise vectors z sampled from distribution $p_z(z)$ and maps onto the estimated image manifold. The training process guarantees that the estimated image manifold is aligned with the training image manifold by optimizing this minmax loss, i.e. the adversarial loss. Ideally, this minmax game has a global optimum at $p_g = p_{data}(x)$, where p_g is implicitly defined by the generator G while $G(z)$ is the sample when $z \sim p_z(z)$.

Originally, the training process of Generative Adversarial Networks (GAN) is highly unstable. This makes the optimum point of the training hard to reach. Hence, the generated images from this pioneering work are blurry and hard to recognize. Later work [98, 4, 26] focuses on improving the loss metrics and training strategies, which improves the generated image quality. A modified approach, the PGGAN [129] proposed to train the GAN in a coarse to fine manner. Starting with low resolution, high resolution layers will be added and trained after the lower layers. Upon the same training strategy, StyleGAN [127] added another mapping from original latent space $\mathcal{Z}$ into the $\mathcal{W}$ space through a non-linear mapping network and then merged into the synthesis network via adaptive instance normalization (AdaIN) at each convolutional layer [57, 118]. This potentially improves the representational ability of StyleGAN and allows it to generate stunningly high resolution images.

In medical image applications, Frid-Adar et al. [82] utilized DCGAN [207] and ACGAN [194] to generate CT liver lesion patches and boosted the liver lesion classification performance. Han et al. [103] proposed to use WGAN [98] to generate MR images for data augmentation and physician training. Nie et al. [191] used GAN to predict CT images from MR images. And, Cao et al. [29] proposed an Auto-GAN to synthesize missing modality for medical images. In particular, GAN has been widely used for skin cancer image generation and purification [19, 91, 18].

Unsupervised Learning: Unsupervised learning aims at learning useful representations from unlabeled data. In [281], Wu et al. try to learn an embedding function by enforcing the features to be discriminative among individual instances. In unsupervised contrastive learning, the goal is to learn a good representation by pulling together positive sample pairs and pushing apart negative sample pairs. The idea of unsupervised contrastive learning is instantiated via different self-supervised learning methods [196, 113, 32, 34, 108,

96] in which the positive sample pairs are crafted by applying different data augmentations on the same image. In self-supervised learning, the augmentations of the same image are attracted and the augmentations of different images are repulsed in the embedding space. In particular, SimCLR [34] leverages the composition of data augmentations and large batch sizes to improve the effectiveness of the representation. MoCo [108] uses a momentum encoder to improve the consistency of the queue of negative samples. From an augmented view of an image, BYOL [96] trains an online network to predict a target network representation of the same image under a different augmented view.

In addition to supervised learning approaches, several groups have applied unsupervised learning to medical image registration and classification tasks [8, 291, 148]. For example, Armanious et al. [5] proposed an unsupervised translation framework for PET-CT translation and MR motion correction. Li et al. [160] utilized multi-modal data for retinal disease diagnosis via self-supervised learning. In particular for skin cancer images, [22, 262, 7] used self-supervised learning for skin cancer classification tasks.

Traditional Data Augmentation: Data augmentation is a traditional approach to improve model generality. Common methods include cropping, rotation, occlusion, flipping, shearing, zooming in/out, image blurring, and changing brightness or contrast. In supervised learning, traditional augmentation methods have been widely utilized [144]. But the performance improvement is limited since those elementary image operations do not introduce much variety to the training data. Recently, GAN-based data augmentation methods have been widely utilized. Shin et al. [229] used GAN-based data augmentation to improve the performance of tumor segmentation in brain MRI. Lim et al. [164] proposed an adversarial autoencoder to augment the data for unsupervised anomaly detection. Waheed et al. [258] proposed CovidGAN to enhance the performance of CNN for COVID-19 detection.

Our proposed method aims to utilize GAN generated skin cancer images to augment the training data for self-supervised learning. Unlike [19, 91, 18], which mainly aim at skin cancer image generation, and [22, 262, 7], which mainly focus on the unsupervised learning for skin cancer images, our proposed method leverages the advantages of both methods and improves the performance of the self-supervised learning.

4.3 GAN Augmentation for Self-Supervised Learning on Skin Cancer Images

In this paper, we propose to utilize Generative Adversarial Networks (GANs) to generate synthetic unlabeled data, which is then used for self-supervised learning of skin cancer images. For traditional computer vision tasks, unlabeled data is easy to collect. However, for medical image analysis, even unlabeled data is scarce, particularly for some rare diseases. Our proposed approach allows self-supervised learning on a limited number of unlabeled data. The self-supervised pre-trained model can be further utilized to boost performance on skin

cancer image classification.

Self-supervised learning on skin cancer images

It is infeasible to train deep neural networks with a limited number of labeled skin cancer images, so we employed self-supervised learning to pretrain the model on unlabeled images. In self-supervised learning, the goal is to learn a useful representation directly from unlabeled data. Several factors influence the success of self-supervised learning: the amount of unlabeled data [108], the training batch size [34], and the composition of data augmentation operations [34]. In contrast to natural images, unlabeled medical images are also expensive to collect. Therefore, we augment the unlabeled data with GAN generated synthetic images to increase the size of the unlabeled dataset. We employ two recently proposed self-supervised learning methods, SimCLR [34] and BYOL [96], to pretrain the model on unlabeled images.

SimCLR

SimCLR [34] is a recently proposed contrastive self-supervised learning method. The goal of SimCLR is to learn representations by attracting differently augmented views of the same data example in the latent space. For each image x in a given set of N images, SimCLR generates two augmented views of x via a stochastic data augmentation module, resulting in a total of $2N$ images. The two differently augmented views of the same image form a positive pair and the other $2(N-1)$ images are negative samples. SimCLR applies a neural network base encoder and a projection head to embed each image in a latent space. The embedded vector is denoted as $\mathbf{z}$. For a positive pair $\{\mathbf{z}_i, \mathbf{z}_j\}$, the contrastive loss is written as,

$$c_{i,j} = -\log \frac{\exp(\mathbf{z}_i^T \mathbf{z}_j/\tau)}{\sum_{k=1,k\neq i}^{2N} \exp(\mathbf{z}_i^T \mathbf{z}_k/\tau)} \tag{4.2}$$

where τ is a temperature parameter. Positive pairs will be attracted in the latent space by minimizing Equation 4.2.

BYOL

More recently, BYOL [96] was proposed as a self-supervised learning method which does not rely on negative samples. BYOL learns the representation by iteratively predicting one augmented view of a given image via a differently augmented view of the same image.

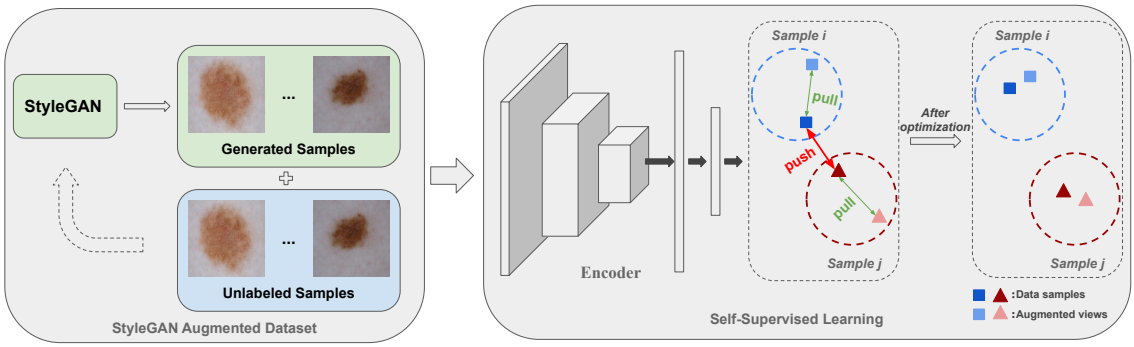
Formally, given an image x, BOYL applies stochastic data augmentation to generate two augmented views x' and x''. The online network with parameter θ generates a representation z'_θ based on x' and the target network with parameter ξ generates a representation z''_ξ based on x''. Then the target network outputs a prediction $c_\theta(z'_\theta)$ of z''_ξ with a classifier c_θ. The prediction $c_\theta(z'_\theta)$ and z''_ξ are both L2-normalized and are optimized to be close via a mean squared error $\mathcal{L}'_{\theta,\xi}$. The loss function is further symmetrized by feeding x' to the target

network and x'' to the online network to obtain $\mathcal{L}''_{\theta,\xi}$. The final loss function is written as $\mathcal{L}_{\theta,\xi} = \mathcal{L}'_{\theta,\xi} + \mathcal{L}''_{\theta,\xi}$.

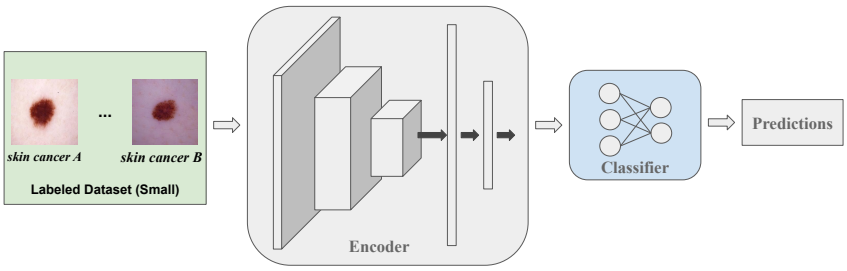
The parameter θ of the online network is optimized via stochastic gradient descent and the parameter ξ of the target network is updated with a moving average,

$$\theta \leftarrow OPT(\theta, \nabla\mathcal{L}_{\theta,\xi}) \qquad \xi \leftarrow \gamma\xi + (1-\gamma)\theta \tag{4.3}$$

where OPT is an optimizer and $\nabla\mathcal{L}_{\theta,\xi}$ is the gradient of the loss function. γ is a hyperparameter which controls the smoothness of the moving average.



(a) Self-Supervised Learning Pipeline



(b) Classification Pipeline

Figure 4.2: Proposed Pipeline. (a) Self-supervised learning pipeline: StyleGAN is first trained using the unlabeled samples and generates authentic skin cancer samples to augment the original training dataset. Then we use self-supervised learning to train a feature encoder. We generate augmented views for each sample in the augmented dataset. The augmented views are treated as positive pairs that are trained to pull towards each other. The augmented views from other samples form negative pairs that are pushed away from each other. (b) Classification pipeline: we leverage the self-supervised trained feature encoder on the skin cancer image classification with limited labeled data. During training, we attach a fully connected layer as the classifier. Only the parameters of the classifier are updated.

Method

In this section, we describe the proposed pipelines of data augmentation with GAN for self-supervised learning on skin cancer images. The pipeline of our proposed method is shown in Figure.4.2. First, StyleGAN is trained on the unlabeled skin cancer images and generates authentic samples for self-supervised learning. Then we train the feature encoder on the augmented dataset including the scarce labeled images and the generated samples from StyleGAN. At last, we leverage the self-supervised learned feature encoder on the skin cancer image classification task on scarce labeled data.

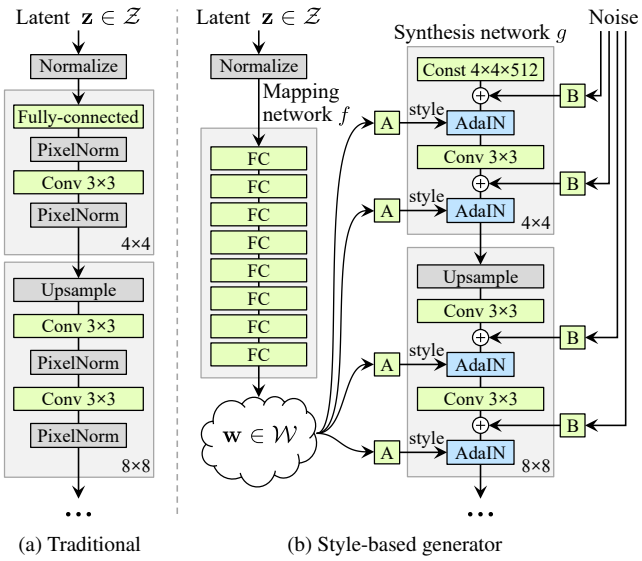


Figure 4.3: StyleGAN Architecture. Compared to traditional GAN models, whose generator directly takes in the latent code only from the input layer, the generator of StyleGAN first maps the latent space to an intermediate latent space $\mathcal{W}$ using a 8-layer Multilayer Perceptron (MLP). Then it will be merged into each convolutional layer via adaptive instance normalization (AdaIN). Gaussian noise will be added after each convolution before the activation layer. "A" represents a learned affine transform and "B" represents learned per-channel scaling factors to the noise input. (Figure is reprinted from [127])

StyleGAN-based Data Augmentation

StyleGAN is the state-of-the-art high resolution image synthesis model. The architecture is shown in Figure.4.3. Unlike a traditional generator, the latent code z will first be mapped to w in an intermediate latent space through a non-linear mapping network, i.e. a 8-layer Multilayer Perceptron (MLP). Then the learned affine transformations specialize w to styles

$\mathbf{y} = (\mathbf{y_s}, \mathbf{y_b})$ that control adaptive instance normalization (AdaIN) [57, 118] operations after each convolution layer of the synthesis network. The AdaIN operation is defined as

$$AdaIN(\mathbf{x_i}, \mathbf{y}) = \mathbf{y_{s,i}} \frac{\mathbf{x_i} - \mu(\mathbf{x_i})}{\sigma(\mathbf{x_i})} + \mathbf{y_{b,i}} \tag{4.4}$$

where $\mathbf{x_i}$ is the feature map at each layer. It will be normalized separately, then scaled and biased according to the scalar components from styles $\mathbf{y}$.

StyleGAN is trained in a progressive manner similar to PGGAN [129]. The training starts from 4×4 resolution. Then after previous resolution layers finish training, layers for the next resolution will be attached for training. In this paper, the generator network consist of 14 layers – two for each resolution (4^2–256^2). The final resolution for the generated image is 256×256.

Using the same training dataset as the one for self-supervised learning, we train a StyleGAN. Then, we sample vectors z in the latent space and pass them into the StyleGAN generator to generate extra skin cancer images for data augmentation. Finally, the GAN-generated images and original training data are combined together for self-supervised learning. In total, $20,000$ skin cancer images are generated for data augmentation, augmenting the training dataset size to $25,000$.

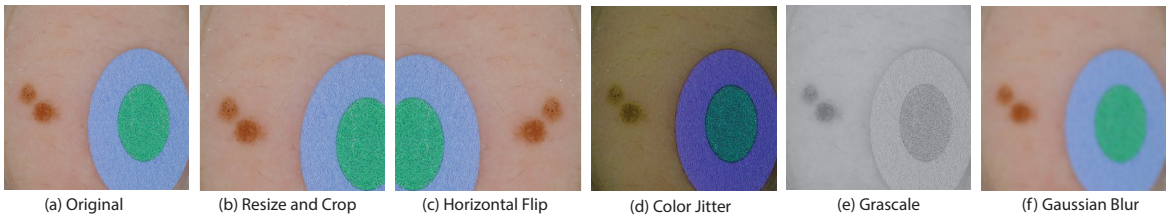


Figure 4.4: Illustration of the operations for SimCLR augmented views. Here, we show all elementary operations. During training, each augmented view is generated by randomly combining those operations. In this paper, we generated two augmented views for self-supervised training.

Self-supervised learning on skin cancer images

For SimCLR [34], we use a Resnet18 [106] backbone for feature encoding. During training, we generate 2 augmented views for each image via random cropping, random horizontal flipping, random color jittering, and random grayscaling. Augmented views from the same image are treated as positive pairs. In SimCLR, the positive pairs are attracted in the latent space. While for augmented views from different images, they are negative pairs, which will be repelled from each other. Positive pairs augmented from an example skin cancer image are shown in Figure.4.4. For BYOL [96], we generate 2 augmented views by using the same

image operations as SimCLR. And the online network and target network are optimized iteratively. After self-supervised learning, the feature encoder is fixed.

For both methods, we use $5,000$ images randomly subsampled from the training dataset (BCN20000 [45] or HAM10000 [251]) to train the feature encoder. We do not use all the image from the training dataset because we aim to simulate the data scarcity problem which widely exists in medical image arenas. Moreover, in this data scarcity setting, we can test whether the GAN-based data augmentation is able to boost the self-supervised learning performance. While using StyleGAN-based data augmentation, we add the aforementioned $20,000$ generated samples together with the original training images for self-supervised learning.

Classification via StyleGAN-boosted feature encoder

For the baseline model, we use the Resnet18 [106] feature encoder trained by self-supervised learning methods as mentioned in the previous section using subsampled $5,000$ skin cancer images. We add one fully connected layer attached to the feature encoder to classify the skin cancer images. Then, we finetune this classifier. Samples from this dataset are shown in Figure.4.5.

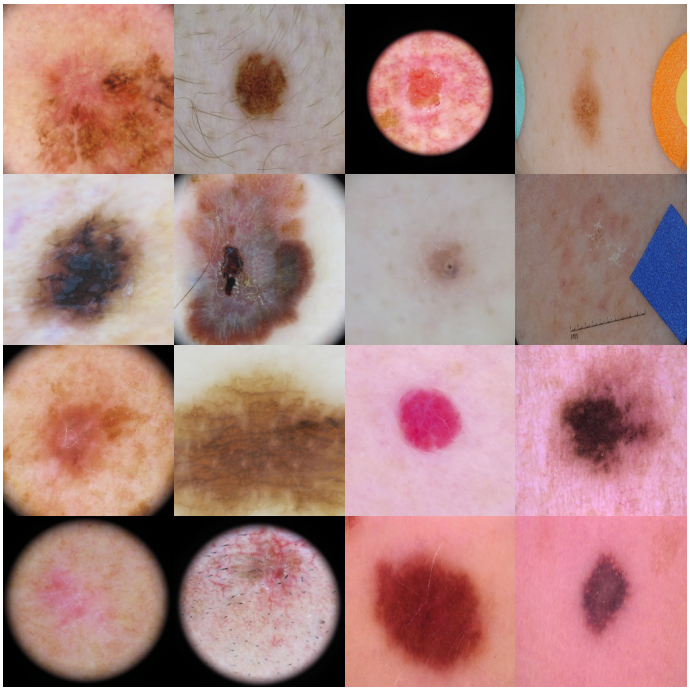


Figure 4.5: Training samples extracted from BCN20000 [45]. It is clear that the variety of the dataset is large. The images have various skin tones, dark corners, hairs, and color patches, which makes the classification extremely hard without a good feature encoder.

During training, the feature encoder parameters are fixed, and only the parameters of the last fully connected layer will be optimized. The training utilizes 80% of the scarce labeled dataset (5000 images subsampled from BCN20000 [45] or HAM10000 [251]). The remaining 20% of the dataset is used for testing. For all the experiments, we repeatedly train the last fully connected layer for 5 times with different random seeds and record the mean test accuracy and its standard deviation.

4.4 Experiments

In the experiments, we investigate the following questions:

- Does the self-supervised pretraining improve the accuracy of the skin cancer classification?
- Does the StyleGAN-based data augmentation improve the quality of the representation learned by self-supervised learning?
- Does the quantity of augmented images influence the improvement of skin cancer classification performance?

Evaluation datasets

BCN20000 [45] is the dataset from the International Skin Imaging Collaboration (ISIC) 2019 Challenge. It contains $25,331$ labeled but unbalanced skin cancer images. 8 skin cancer types are included: nevus, melanoma, basal cell carcinoma, seborrheic keratosis, actinic keratosis, squamos cell carcinoma, dermatofibroma, vascular lesion.

HAM10000 [251] is the dataset from the ISIC 2018 Challenge. It contains $10,000$ skin cancer images, including actinic keratosis, basal cell carcinoma, benign keratosis, dermatofibroma, melanocytic nevi, melanoma, and vascular lesion.

Both datasets are highly imbalanced. The quantity of each skin cancer category varies a lot. Compared to HAM10000 [251], BCN20000 [45] is a more challenging dataset. BCN20000 contains lesions found in hard to diagnose locations (nails and mucosa) [45]. Most of the images would be considered hard-to-diagnose [45].

Model training and implementation details

The latent vector for StyleGAN generator has the dimension of 512. The generator consists of 14 layers – two for each resolution (4^2–256^2). The discriminator has the mirrored structure of the generator – also two for each resolution (256^2–4^2). We use the Adam optimizer [135] with $\beta_1 = 0.0$ and $\beta_2 = 0.99$. The learning rate is set to 0.002. During training, the images are reshaped to 256×256.

While training via SimCLR [34], we augment 2 views for each training image. We train the Resnet18 [106] backbone for 200 epochs with a batchsize of 256. The learning rate is set to 0.0003. During training, the images are reshaped to 96×96.

For the fine tuning, we attach 1 fully connected layer to the Resnet feature encoder for skin cancer image classification. We fix the parameters for the feature encoder and train the attached fully connected layer 200 epochs. The Adam optimizer [135] is used with default parameters. The learning rate is 0.0001.

Quantitative Results

With vs. Without Self-supervised Pretraining

In order to investigate whether self-supervised learning would improve the accuracy of the skin cancer classification, we compare the classification results using the feature encoder with and without self-supervised pretraining on both BCN20000 [45] and HAM10000 [251]. For the result without self-supervised pretraining, we randomly initialize the Resnet18 feature encoder parameters. While for self-supervised pretraining, we train a Resnet18 feature encoder using SimCLR and BYOL. During testing, we attach a fully connect layer as the classifier and train its parameters with the feature encoder parameters fixed. Here, both SimCLR and BYOL are trained using the $5,000$ images subsampled from BCN20000 [45] or HAM10000 [251]. The comparison of the skin cancer classification accuracy is shown in Table.4.1.

	BCN20000	HAM10000
w/o pretraining	26.05±1.24%	67.87±0.38%
SimCLR	34.73±1.07%	71.84±0.23%
BYOL	35.71±2.04%	71.37±0.36%

Table 4.1: Classification accuracy w/o vs. w/ self-supervised pretraining on BCN20000[45] and HAM10000[251]

On BCN20000 [45], the classification accuracy is $26.05 \pm 1.24\%$ without self-supervised pretraining. SimCLR and BYOL achieve $34.73 \pm 1.07\%$ and $35.71 \pm 2.04\%$ respectively. On HAM10000 [251], the classification accuracy is $67.87 \pm 0.38\%$ without self-supervised pretraining. SimCLR and BYOL achieve $71.84 \pm 0.23\%$ and $71.37 \pm 0.36\%$ respectively.

Clearly, with SimCLR and BYOL self-supervised pre-training, we can improve the skin cancer classification accuracy compared to a random feature encoder (without self-supervised pretraining). This indicates that self-supervised learning methods can learn useful representations directly from unlabeled skin cancer images. It further reveals that it is possible to utilize the knowledge from unlabeled images to improve the medical image classification.

With vs. Without StyleGAN-based Data Augmentation

We then investigate whether the StyleGAN-based data augmentation would improve the self-supervised learning performance on the skin cancer classification. First, we train a Resnet18 network as the feature encoder via self-supervised learning using $5,000$ skin cancer images subsampled from BCN20000 [45] or HAM10000 [251]. We utilize both SimCLR and BYOL. For the StyleGAN-based data augmentation, $20,000$ generated skin cancer images are added into the training dataset for self-supervised learning, augmenting the training dataset size to $25,000$. The comparison of the skin cancer classification accuracy is shown in Table.4.2.

	BCN20000	HAM10000
SimCLR w/o DA	34.73±1.07%	71.84±0.23%
SimCLR w/ DA	38.55±0.44%	72.52±0.25%
BYOL w/o DA	35.71±2.04%	71.37±0.36%
BYOL w/ DA	46.88±0.48%	74.44±0.28%

Table 4.2: Classification accuracy w/o vs. w/ GAN-based Data Augmentation (DA) on BCN20000[45] and HAM10000[251]

For SimCLR, the classification performance is boosted from $34.73 \pm 1.07\%$ to $38.55 \pm 0.44\%$ on BCN20000[45], and the classification performance is boosted from $71.84 \pm 0.23\%$ to $72.52 \pm 0.25\%$ on HAM10000[251]. For BYOL, the classification performance is boosted from $35.71 \pm 2.04\%$ to $46.88 \pm 0.48\%$ on BCN20000[45], and the classification performance is boosted from $71.37 \pm 0.36\%$ to $74.44 \pm 0.28\%$ on HAM10000[251].

Clearly, the StyleGAN-based data augmentation can improve the self-supervised learning performance on the skin cancer classification. Since we are using the feature encoder trained via self-supervised learning methods, it further indicates that the StyleGAN-based data augmentation can help self-supervised learning methods learn more useful representation from unlabeled skin cancer images.

Influence of the StyleGAN Augmented Sample Quantity

In this experiment, we vary the quantity of both raw unlabeled training images and StyleGAN augmented samples to train self-supervised classification via SimCLR. The experiment is conducted on BCN20000[45]. The quantity of raw unlabeled images is set at $1k$, $3k$, $5k$ and $7k$. The quantity of StyleGAN augmented samples is set at 0, $10k$ and $20k$. We investigate the classification accuracy under different combinations of those two parameters, i.e. at different augmentation ratio. The augmentation ratio is defined as follow.

$$ratio = \frac{Q_{raw}}{Q_{augmentation}} \tag{4.5}$$

where Q_{raw} is the quantity of raw unlabeled images and $Q_{augmentation}$ is the quantity of StyleGAN augmented samples. The skin cancer classification accuracy at different augmentation ratios is shown in Fig.4.6.

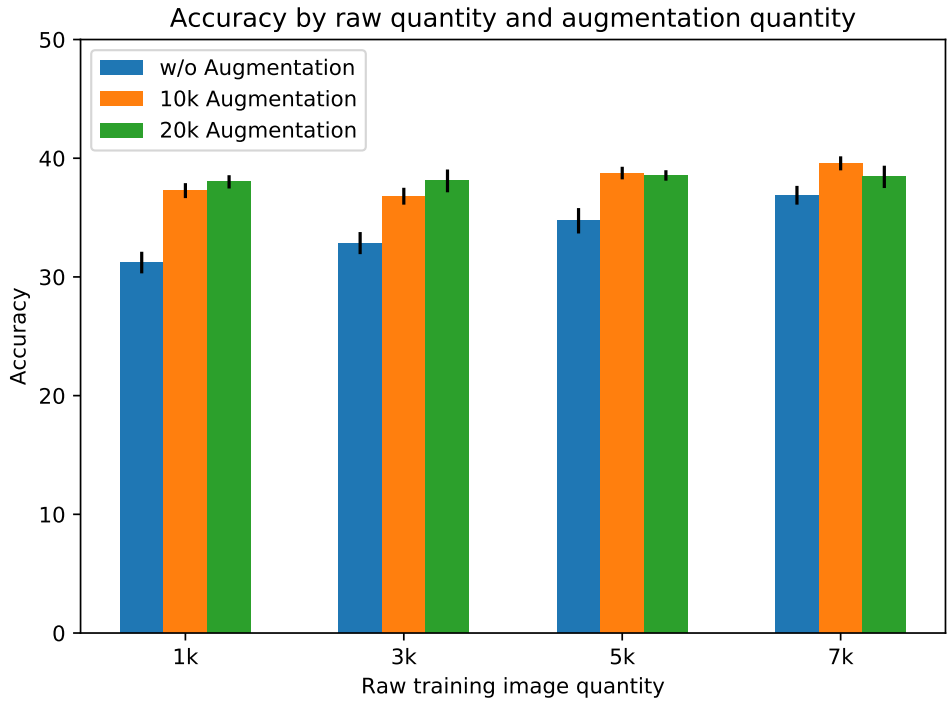


Figure 4.6: Classification accuracy on BCN20000[45] at different StyleGAN augmented sample quantities.

From the bar chart, it is clear that without StyleGAN-based data augmentation, increasing the raw unlabeled image quantity can help to improve the self-supervised classification result. While applying StyleGAN-based data augmentation, for small raw unlabeled images quantities where the augmentation ratio is large, such as $1k$ and $3k$, the skin cancer classification accuracy can gain a lot. However, for larger raw unlabeled images quantities where the augmentation ratio is small, such as $5k$ and $7k$, the accuracy boost is reduced.

Qualitative Results

StyleGAN generated results

After StyleGAN has been trained, we randomly sample latent codes z, then pass them to the generator. The generated images are shown in Figure.4.7 and Figure.4.8 for BCN20000[45] and HAM10000[251] respectively. From the generated results, it is clear that the StyleGAN

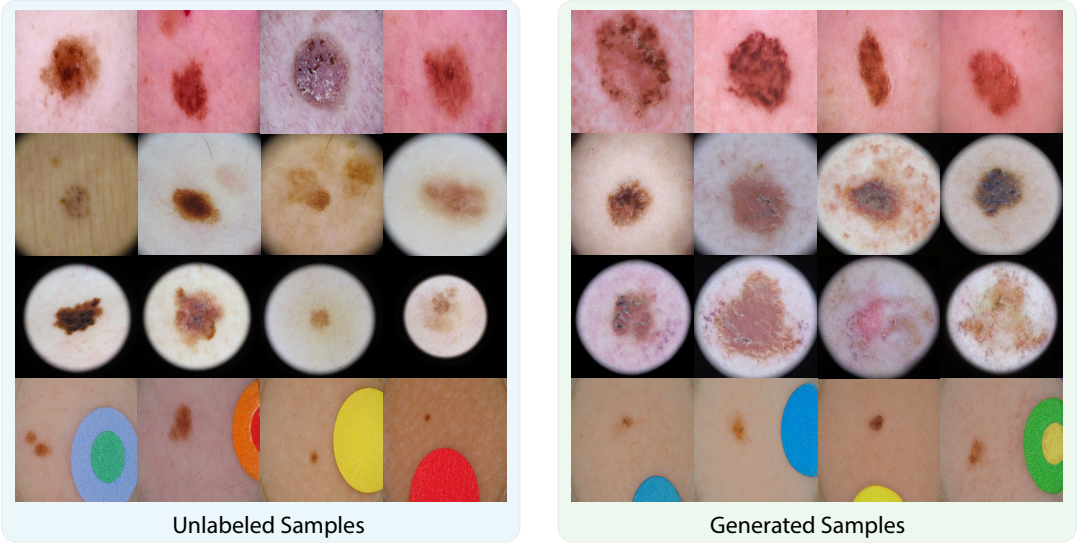


Figure 4.7: Uncurated set of novel images produced by StyleGAN on BCN20000[45]. Compared to the images from unlabeled training dataset, the generated samples well maintained the semantic statistics, such as the skin tone, the dark corner of the image, and some color patches. The generated skin cancer image resolution is 256×256.

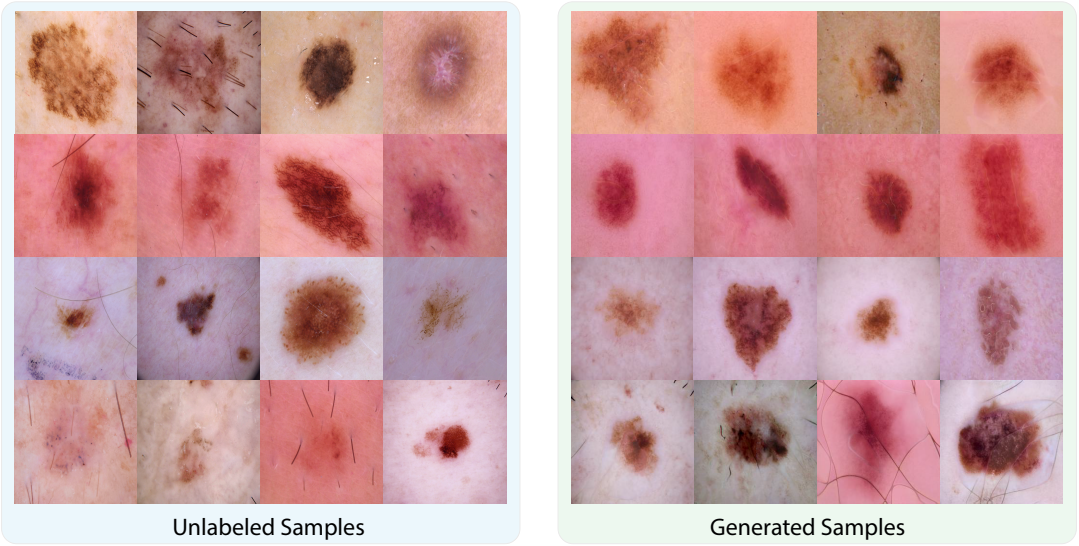


Figure 4.8: Uncurated set of novel images produced by StyleGAN on HAM10000[251]. The generated skin cancer images are semantically similar to the unlabeled training samples. It is clear that compared to the images in BCN20000[45], HAM10000[251] has less diverse image texture.

generator has learned the semantic statistics of the training dataset, such as the skin tone, the dark corner of the image, and some color patches. Moreover, the generator can utilize the learned statistics to produce novel images which do not exist in the real world.

Additionally, it is clear that BCN20000[45] is a more challenging dataset since it has more diverse image texture compared to HAM10000[251]. Intuitively, this indicates that compared to HAM10000, images in BCN20000 scatter in a sparser way on the image manifold such that StyleGAN based data augmentation can efficiently interpolate between the image samples. On the contrary, HAM10000 is less diverse, i.e., the image samples are locally denser on the image manifold. Therefore, the performance boost from StyleGAN based data augmentation is limited on HAM10000.

Comparison between PGGAN and StyleGAN

We also train PGGAN to perform the skin cancer image generation on BCN20000[45]. We compare PGGAN generation quality with StyleGAN generation quality because they both share the same progressive training manner and have similar network structures. During training, we use the same number of epochs with the same optimizer setting and learning rate. The generation results are randomly picked for both models and are arranged based on certain semantics. The comparison is shown in Figure.4.9.

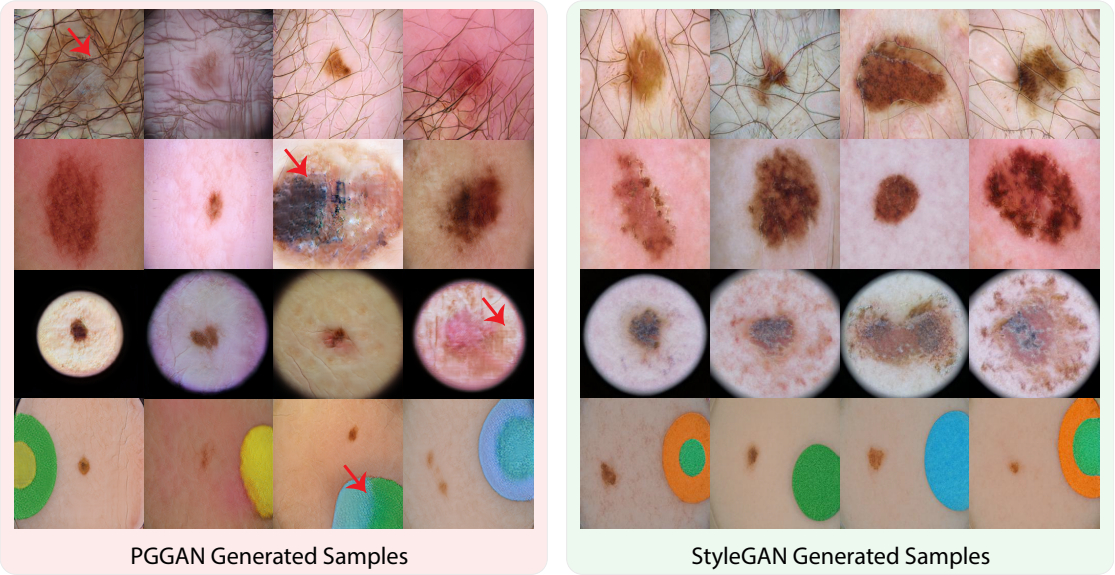


Figure 4.9: PGGAN and StyleGAN skin cancer image generation quality comparison. It is clear that overall StyleGAN generated skin cancer images have higher visual quality compared to those generated by PGGAN. As indicated by the red arrows, the skin cancer image details, such as hair, lesion texture, surrounding skin texture and color patches, are maintained sharper and more meaningful in StyleGAN generated samples.

In general, StyleGAN generation is better than the PGGAN generation visually. As indicated by the red arrows, StyleGAN can generate sharper details for hair, lesion texture, surrounding skin texture and color patches, while those details are blurry and unreasonable in PGGAN generated skin cancer images. This is attributed to the non-linear mapping network from original latent space $\mathcal{Z}$ into the $\mathcal{W}$ space and the merging branch via adaptive instance normalization (AdaIN) at each convolutional layer [57, 118] in StyleGAN.

4.5 Discussion

In this paper, we showed that StyleGAN is capable of synthesising authentic skin cancer images. This is valuable because the ability to generate images like those presented here helps ameliorate the significant problem of data scarcity. In particular, for rare skin cancer cases, the data augmentation benefit is even larger. Thus, the proposed approach can reduce the cost and human effort required for teledermatology.

Moreover, other mobile health apps will also suffer data scarcity issue at the early stage. We can apply the proposed method to other medical modalities as well. The generated images can also be used in other domains, such as medical image perception and medical image analysis.

A second goal of this paper was to test whether StyleGAN generated samples can be utilized for data augmentation for self-supervised learning. We found that StyleGAN-based data augmentation significantly boosted the performance of self-supervised skin cancer classification. Essentially, using the generated images helped the classifier better discriminate skin lesions. In a followup experiment, we found that the classification performance improvement was most significant in cases when there were fewer labeled training data. That is, the benefit of augmenting data is most pronounced when labeled data are scarce.

Compared to supervised learning, self-supervised learning only requires a small quantity of labeled data at the final training stage. Thus, with the gradually growing unlabeled training data from users, self-supervised learning system is easier to scale. Moreover, it only requires experts to label a small amount of key data. Therefore, it is more suitable for mobile health app systems. With our proposed generative self-supervised learning, the performance of mobile health apps could be much improved.

4.6 Conclusion

In this paper, we trained StyleGAN to augment the training dataset for self-supervised learning of skin cancer images for teledermatology. Our model was able to generate authentic skin cancer images, and those images were effective as a source of augmentation for self-supervised learning. The benefit of augmenting real datasets with StyleGAN-based generated data was most prominent when the original dataset was limited in size. Therefore, when real data are scarce (for example, in several types of skin cancer including Merkel cell carcinoma),

the augmentation approach presented here could be highly beneficial. This, in turn, could be very helpful for mobile health applications.

Chapter 5

Serial dependence in perception across naturalistic GAN-generated mammograms

5.1 Introduction

Clinical diagnosis based on radiographs is not always perfect because of misperceptions and misinterpretations[14, 50]. Some sources of interpretive error have been identified and characterized, including search and recognition errors[30, 193], cognitive biases[50, 158], search satisfaction[6, 12], subsequent search misses[16, 24, 105], and low prevalence[279, 280, 218, 181, 68, 114, 149]. However, some other errors in cancer image interpretation are still without explanation[27, 260, 259]. Thus, a great deal of research has been carried out in the last several decades to identify and characterize the sources of these errors in order to mitigate them.

Radiologists often read dozens or hundreds of radiographs in batches [180], sometimes looking at several related images one after the other. Their job is to localize the lesions (if present), and then to recognize them by judging their size, class, and so on. A main underlying assumption here is that radiologists' perceptual decisions about the current radiograph are independent of prior perceptual experience.

Recent theoretical and empirical research suggests that this assumption is not true. For example, the human visual system is characterized by visual serial dependency, a type of sequential effect in which what was previously experienced influences (captures) what is seen and reported at this moment[39, 75]. Serial dependencies can manifest in several domains, such as perception[41, 42, 75, 173], decision making[1, 73], and memory[9, 78, 137], and they occur with a variety of features and objects, including orientation[75, 175], position[20, 173], faces[161, 245], attractiveness[246, 282, 141], ambiguous objects[270], ensemble coding of orientation[175], and numerosity[41, 46]. Serial dependence is characterized by three main kinds of tuning. First, feature tuning: serial dependence occurs only between similar features and

not between dissimilar ones[75, 84, 175, 173]. Second, temporal tuning: serial dependence gradually decays over time[75, 173, 270]. Third, spatial tuning: serial dependence occurs only within a limited spatial window; it is strongest when previous and current objects are presented at the same location, and it gradually decays as the relative distance increases[20, 44, 75, 173]. In addition, attention is a necessary component for serial dependence[75, 83, 133].

Because our visual world is stable–objects that were present a moment ago tend to still be present at this moment–we benefit from serial dependence most of the time. This is because it is more efficient to simply recycle perceptual history[75, 41, 173], using the past to predict the present. However, this recycling is not always beneficial. When stimuli are randomly ordered or in unnatural situations–such as when the visual world is not autocorrelated or stable–serial dependence can negatively impact perceptual decisions[75, 85, 161]. For example, visual search in clinical settings, such as reading randomly ordered radiographs or pathology slides, is a striking example where stimuli may not be autocorrelated. In this case, the past may not be a good predictor of the present, and serial dependence in perceptual decisions would be problematic. In fact, empirical experiments have found that clinicians' perceptual decisions can be biased towards the previous images they have seen[171, 174].

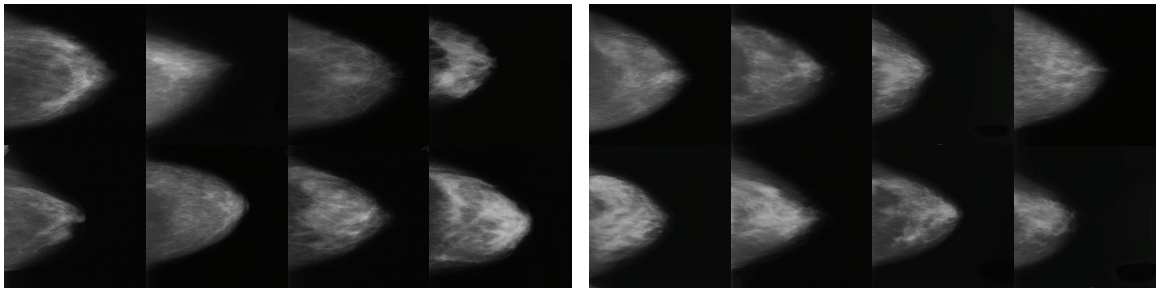
A drawback of previous work[171, 174] is that serial dependence was measured with unrealistic stimuli, such as random geometric shapes superimposed on a mammogram section (Figure 5.2A). Although well-controlled, these images are clearly inauthentic and are therefore far from naturalistic mammograms[171, 174]. Unfortunately, because serial dependence has only been measured with unrealistic stimuli, it remains unclear whether serial dependence in perceptual judgments would even occur for truly realistic radiographs.

In this study, we aim to measure the presence of sequential effects in the perceptual decisions of observers who view controlled, realistic GAN-generated radiographs. To accomplish this, we created authentic-looking medical images generated by a computer vision model. The model allows precise control over the stimulus space, while simultaneously ensuring that the simulated radiographs are realistic. In fact, a previous study found that these images are indistinguishable from (i.e., metameric with) down-sampled real radiographs, even to many professional clinicians[214, 213]. We hypothesize that even with authentic-looking simulated mammograms, perceptual decisions about any given current image will be biased toward the previously seen images, due to serial dependence.

5.2 Methods

Mammogram Generation

In computer vision, generative models[136, 95] have been utilized for authentic image generation for years. In particular, Generative Adversarial Networks (GANs) are a promising method to create authentic images in different modalities such as human faces, places, animals, cars, and so on[98, 129, 127]. Similar approaches have also been applied for medical



Real Samples **Generated Samples**

Figure 5.1: Generated samples via GAN. Here, we show a comparison between the real sample (down-sampled mammograms from DDSM Dataset which are collected from the hospital) and GAN-generated samples. After training, GAN learns the image manifold of down-sampled real samples and then samples on the learned manifold to generate novel simulated samples. Additionally, since the manifold has been learned, interpolation can be applied to generate quantifiably similar images. The resolution of the real and generated samples is equated.

image generation[103, 82, 214, 215]. In this study, we adopted a controllable medical image generation method[214, 213] to create all stimuli used in our experiments. Because of the GAN generation paradigm, the generated samples share the similar data distribution as the real samples maintaining the variety of the real ones. A comparison between down-sampled real samples and generated samples is shown in Figure 5.1.

Once the GAN was pretrained, we randomly picked three anchor points in the latent space and generated interpolations between each of the three anchor points. Each anchor points and the corresponding interpolations are latent vectors with the size of 512. [1] Then, we passed the anchors as well as the interpolations through the pretrained generator to generate the corresponding images, forming a circular continuum (Figure 5.2B). One hundred and forty-seven images (48 between each anchor) were generated on the circular continuum with size 256×256 [2]. In the experiment, 20 circular continuums like this were generated by creating 20 sets of anchors and passing them along with the corresponding interpolations to the pretrained generator. In total, we generated 2940 images. Four example continua are shown in Figure 5.3.

Since the images within a given continuum were generated based on interpolations in the latent space, nearby images on the circular continuum are similar, while distant images on the circular continuum differ from each other. With random picking, the generated image

[1] Please see the publications[214, 213] for thorough details about the model and latent space.

[2] The reason we used 256×256 was for proof of concept and because the training takes exponentially longer with higher-resolution images.

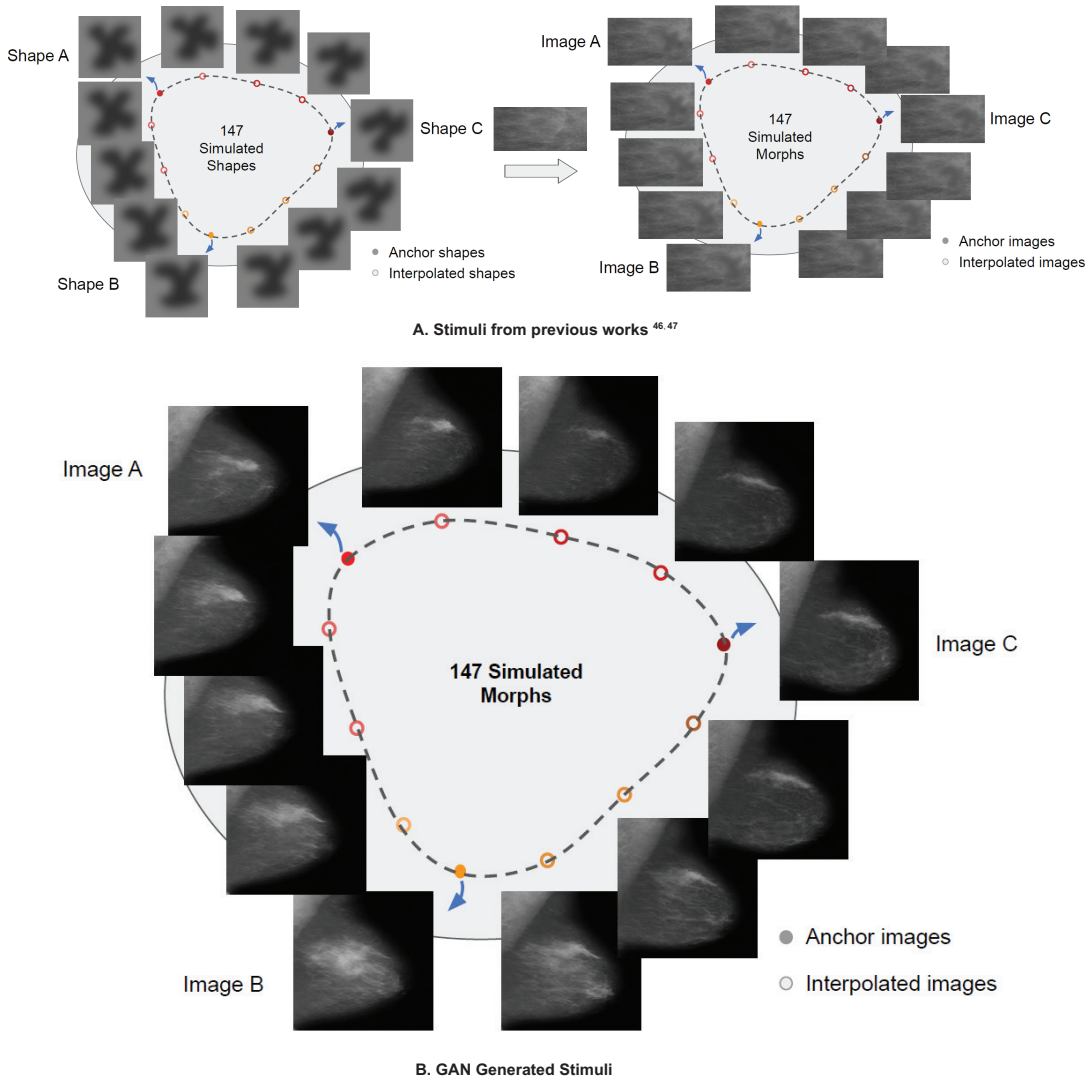


Figure 5.2: Comparison between stimuli used in previous experiments and current GAN-generated stimuli. (A) Stimuli from previous works [171, 174]. A circular continuum of simple shapes is generated first, then each shape is fused onto a mammogram tissue background section to form the experiment stimuli. (B) We randomly picked three anchor points in the latent space (Image A, B and C shown with solid dots) and generated 48 interpolated morphs in between each pair (shown with hollow dots) via GAN (147 morphs in total) to form a circular morph continuum. In total, 20 circular continuums were generated. Here, we show 1 continuum as an example. More continuum examples can be found in Figure 5.3.

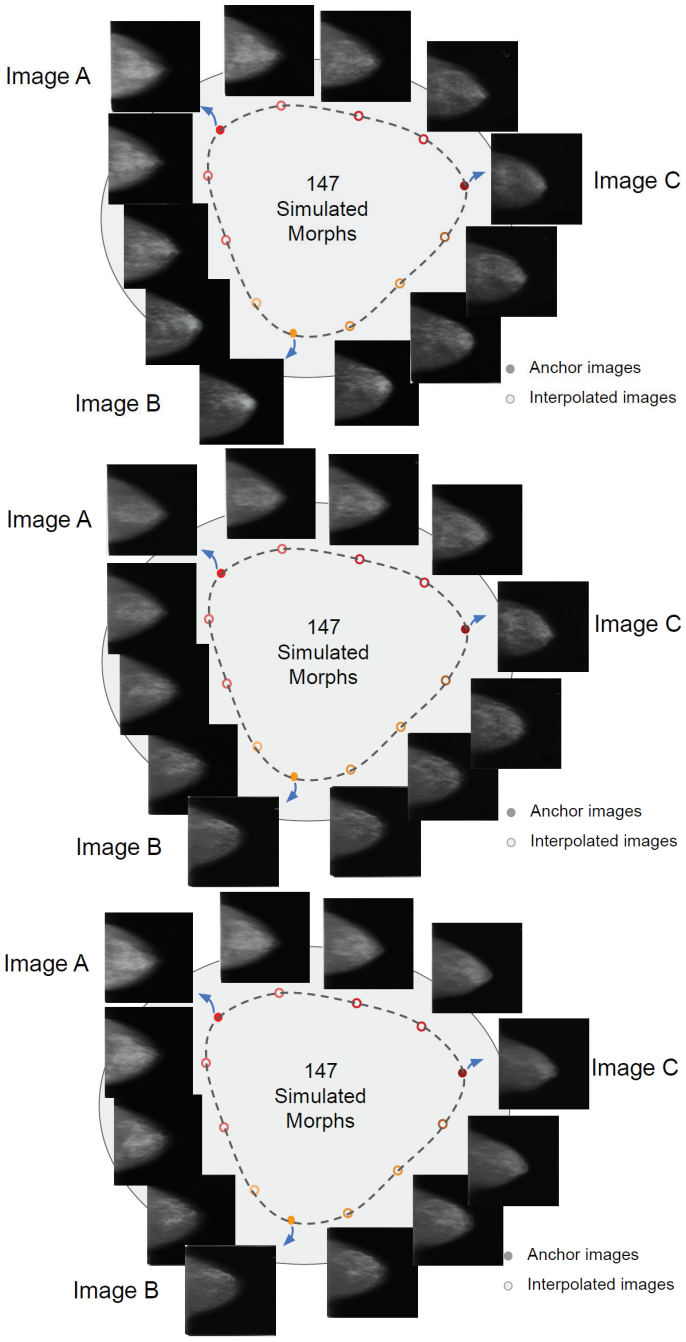


Figure 5.3: Three extra example continua. Each shows a circular morph continuum generated from different anchor sets. Here, we only show 3 out of 48 interpolations between anchor points. The actual similarity steps between sequential interpolations are much closer.

sequence can represent a certain variety of the real samples. Moreover, moving around the circular continuum, the tissue texture, tumor size, tumor location, and other semantic properties gradually change and return to the same place when looping through all the GAN-generated mammograms on this circular continuum.

Dataset

Training data for the Generative Adversarial Network (GAN) is from the Digital Database for Screening Mammography (DDSM[23]). It contains $2,620$ normal, benign, and malignant cases with verified pathology information. The images were first center cropped then resized to 256×256 for training. In order to generate stimuli containing tumors for the visual search task, only benign and malignant cases were utilized for training. Several down-sampled real samples are shown in Figure 5.1.

Participants and Apparatus

All experimental procedures were approved by and conducted in accordance with the guidelines and regulations of the UC Berkeley Institutional Review Board. Participants provided informed consent in accordance with the IRB guidelines of the University of California at Berkeley. All participants had normal or corrected-to-normal vision, and were all naïve to the purpose of the experiment. 80 non-expert participants (28 males, aged 18-72; 52 females, aged 18-62) participated in the experiment. They were students and affiliates at UC Berkeley.

Experiments were coded with PsychoPy and published on Pavlovia. Participants were able to access the experiment by themselves through the Internet. Sets of 4 participants were assigned to the same circular continuum, and there were 20 circular continuums in total (for a total of 80 observers). Participants used a keyboard for all responses.

Experiment Design

The 20 circular morph continua mentioned in Sec. 5.2 were used to test the perceptual decisions of the participants. Each simulated mammogram of any continuum contains a particular pattern of lesions and texture, and these characteristics gradually change along the circular continuum. On each trial, participants viewed a random simulated mammogram, which was randomly extracted from one of the 20 circular continua mimicing the randomness in real diagnostic scenarios. The simulated mammogram was presented for 500 ms. Next, we presented a mask composed of random Gaussian noise for 1000 ms (to avoid the possibility of afterimages). After the mask, a random simulated mammogram drawn from the same morph continuum appeared at the fixation point location, and participants were asked to adjust the simulated mammogram to match the perceived simulated mammogram using the left/right arrow keys (continuous report, adjustment task; left–right arrow keys to adjust the simulated mammogram along the circular morph continuum). The starting simulated

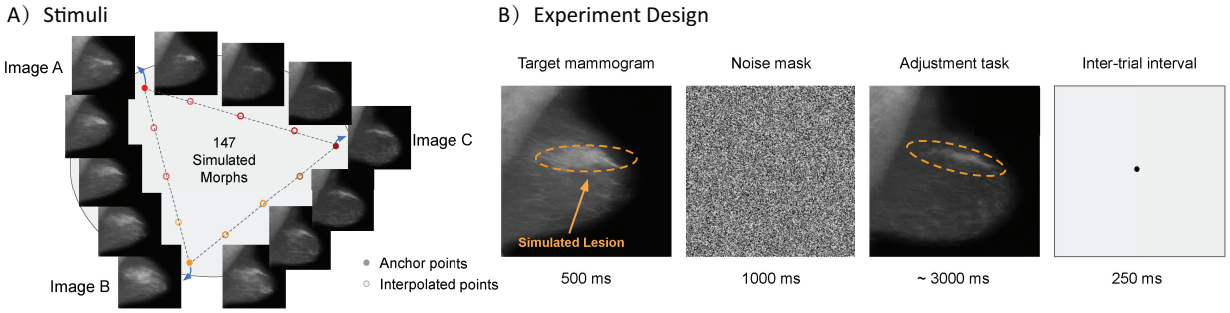


Figure 5.4: Stimuli and experiment design. A) An example circular continuum generated via GAN. B) Observers were presented with a random morph on a specific morph continuum, followed by a noise mask. They were then asked to adjust the morph (the start point is randomly picked along the same morph continuum.) to match the target morph they previously saw, and pressed space bar to confirm. During the inter-trial interval, a black fixation dot appeared in the center. After a 250 ms inter-trial interval, the next trial started.

mammogram was randomized on each trial. Participants were allowed to take as much time as necessary to respond and pressed the space bar to confirm the chosen simulated mammogram was the correct match. Following the response and a 250 ms delay, the next trial started. A concise experiment pipeline can be found in Figure 5.4B.

During the experiment, participants were asked to continuously fixate a black dot in the center. In total, each participant performed 3 blocks of 85 trials. Between each block, participants were allowed to take a break. In a preliminary session, observers completed a practice block of 10 trials to familiarize themselves with this experiment. Among the 80 participants, 3 participants were removed from data analysis because they hit the space bar all the time during the experiment without any adjustments.

Data Analysis

Response error was computed as the smallest difference along the morph continuum between the match morph and the target morph (current match morph - current target morph). For each participant's data, trials were removed if the response error was 3 standard deviations away from the mean response error or if the response time was longer than 20s. The average reaction time was 3.42 ± 2.47 seconds.

Previous research shows that individual observers can have idiosyncratic biases in object recognition and localization, which are unrelated to serial dependence [266, 142]. For example, observers may make a consistent error in reporting a simulated lesion of 20 morph units as being 10, thus creating a systematic error of -10 morph units. Conversely, if there was no systematic error, all error would approximate zero. For this reason, we conducted an additional data processing strategy to remove such potential unrelated biases before further

analyses. We modeled observers' response error as a function of the target morph presented by fitting a Radial Basis Function (RBF) where 30 Gaussian kernels are utilized. This allowed us to quantify the idiosyncratic bias for each observer. We then regressed out the bias quantified by the radial basis fit by subtracting it from the observer's error. This subtraction left us with residual errors that did not include the idiosyncratic biases unrelated to serial dependence.

Feature Tuning Analysis

The difference in morphs between the current and previous trial is computed as the smallest difference along the morph continuum between the previous target morph (n-back) and the current target morph (previous target morph - current target morph). In order to quantify the feature tuning characteristic of serial dependence, we fit a derivative of Von Mises distribution to each observer's data points. The derivative of Von Mises distribution can be expressed by the following equation:

$$y = -\frac{a\kappa \sin(x-\mu)e^{\kappa\cos(x-\mu)}}{2\pi I_0(\kappa)} \tag{5.1}$$

where parameter y is response error on each trial, x is the relative orientation of the previous trial, a is the amplitude modulation parameter of the derivative-of-von-Mises, μ indicates the symmetry axis of the von-Mises derivative, κ indicates the concentration of the von-Mises derivative, and $I_0(\kappa)$ is the modified Bessel function of order 0. In our experiments, μ is set to 0. We fitted the von-Mises derivative using constrained nonlinear minimization of the residual sum of squares. As a measure of serial dependence, we reported half the peak-to-trough amplitude of the derivative-of-von-Mises.

Additionally, for each observer, we computed the running circular average within a 20 morph units window. Figure 5.5 (blue line) shows the average of the moving averages across all the observers, and the corresponding von Mises derivative fit.

Temporal Tuning Analysis

In this study, we report half the peak-to-trough amplitude of the derivative-of-von-Mises as a measure of serial dependence (Figure 5.5). Sequentially, we can get the strength of 1-back, 2-back and 3-back serial dependence effects by fitting the derivative of Von Mises distribution on the data points where the difference in morphs between the current and previous trial is computed as the smallest difference along the morph continuum between the 1-, 2-, and 3-trial back target morph and the current target morph.

Additionally, as a control analysis, we explored the effect of future trials on the current response to check for potential unrelated biases and artifacts that might be lurking in the data [78, 179]. In particular, we calculated whether the current trial response error depended in some fashion on the difference in stimuli between the current and 1-forward (following) trials. Since observers have not seen the future trial stimulus, their current response in a

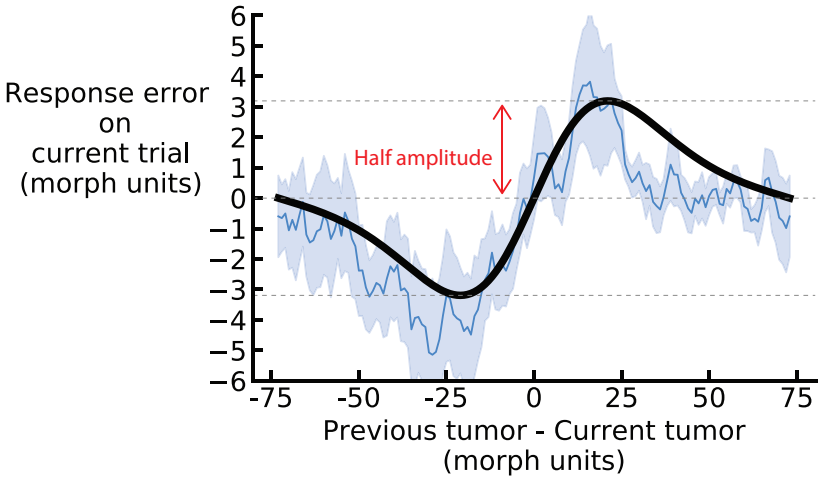


Figure 5.5: Derivative-of-von Mises curve fit for a representative continuum (one of the twenty different morph continuums. In units of shape morph steps, the x-axis is the shortest distance along the morph continuum between the current and one-back simulated lesion, and the y-axis is the shortest distance along the morph continuum between the selected match shape and current simulated lesion. Positive x axis values indicate that the one-back simulated lesion was clockwise on the shape morph continuum relative to the current simulated lesion, and positive y axis values indicate that the current adjusted shape was also clockwise relative to the current simulated lesion. The average of the running averages across observers (blue line) reveals a clear trend in the data, which followed a derivative-of-von-Mises shape (model fit depicted as black solid line; fit on average of running averages). Light-blue shaded error bars indicate standard error across observers. We operationalized the strength of pull towards the previous observed stimuli as the half amplitude of the derivate-of-von-Mises curve, as noted in red.

given trial should not be influenced by the future morph stimuli. If there are artifacts in the data, however, (for example observers perseverate on a particular response from trial to trial), there might appear to be an effect of future stimuli on the current response. This analysis reveals and serves as a control for such artifacts [75, 174].

Bootstrapping: For each result we obtained, we resampled the data with replacement, processed the sampled data recursively for 5000 times, and reported the mean result with 95% confidence intervals.

Permutation Test: Significance testing was done through permutation tests. Data was randomly shuffled and processed 5000 times. The 97.5% upper bound of the permuted null distribution was compared with the error bar from bootstrapping to confirm the significance of the result.

In an additional analysis, to more intuitively convey the magnitude of the serial depen-

dence effect, we analyzed the percentage difference between pro-SD (pulling effect due to serial dependence) and anti-SD (repelling effect against serial dependence) for 1, 2, and 3 trials back. Stimuli on the circular continuum were categorized into 3 types according to the nearest anchor images. Trials in which the response image was not within the same category as the target image were considered classification errors, which are misjudgments of the image category. Classification errors that are in a direction consistent with the previously seen stimulus are Pro-SD errors, and those that are in a direction opposite the previously seen stimulus are Anti-SD errors. In principle, classification errors should be randomly distributed, not biased in either direction. As a sanity check, we also analyzed the percentage difference between pro-SD and anti-SD for 1 trial forward. Because future trials naturally are not correlated with current trials.

5.3 Results

The goal of this experiment was to test whether perceptual decisions on consecutive realistic GAN-generated images of mammograms were biased towards the previously seen images. Here, observers' response error in a particular trial was computed as the shortest distance along the morph continuum between the actual observed shape and the chosen answer shape. Average response error was 17.26 ± 5 morph units, and average reaction time was 3.43 ± 1.50 seconds.

To test whether there are sequential effects in observers' judgments of realistic GAN-generated mammograms, we first analyzed the response error in relation to the difference in stimulus shape between the current and previous trials for each continuum separately. Then, we fitted a derivative-of-Von-Mises (DoVM) function to this data (Figure 5.5).

We operationalized serial dependence, the pull towards previous stimuli, as the half amplitude of the DoVM curve of each continuum. We bootstrapped the half amplitude and reported the average bootstrapped half amplitude for each continuum: all continua showed a positive half amplitude (Figure 5.6A). Importantly, the average half amplitude across all continua was significant (average bootstrapped 1-back half amplitude = 2.77 morph units, $p < 0.001$, permutation analysis), which suggests an influence of the simulated radiograph in the previous trial on the current response. The influence of previous stimuli extended to two trials back (average bootstrapped 2-back half amplitude = 1.38 morph units, $p < 0.01$, permutation analysis). In contrast, the stimuli presented three trials prior had no influence on the current response (average bootstrapped 3-back half amplitude = 0.09 morph units, $p > 0.05$, permutation analysis). To control for artifacts, we calculated the influence of the stimuli presented in the next trial on the current response. We found a modest bias, as found in previous studies of sequential effects [179, 36], but, importantly, the 1-back and 2-back effects were significantly larger than this 1-forward baseline (1-back versus 1-forward: $p < 0.05$; 2-back versus 1-forward: $p < 0.05$. This confirms that there are sequential effects in perceptual decisions about realistic GAN-generated mammograms.

To quantify the serial dependence effect in an alternative manner, we also analyzed the

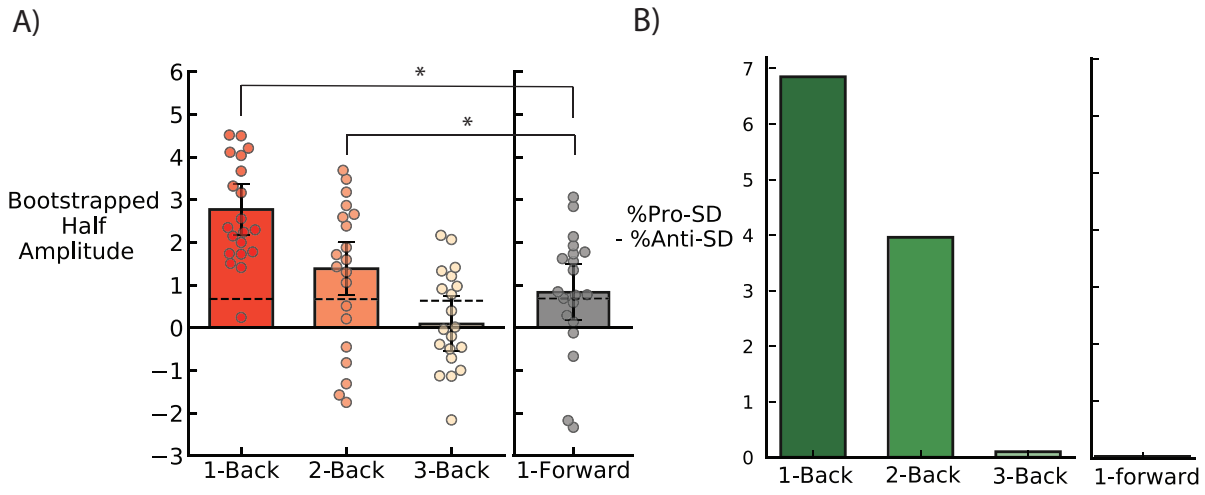


Figure 5.6: A) Bootstrapped half amplitudes of derivative of von Mises fit for 1, 2, and 3 trials back. Half amplitude for 1-forward is shown as a comparison (grey bars). Each filled dot represents the bootstrapped half amplitude for a single circular morph continuum. Bars indicate the group bootstrap and error bars are bootstrapped 95% confidence intervals. B) Classification error analysis. Stimuli on the circular continuum are categorized into 3 types according to the nearest anchor images. Classification errors are categorized based on distance to the three anchors. Pro-SD means the classification error on the current trial is attracted towards the previous stimuli, while anti-SD means the current classification error is repelled from (opposite) the previous stimulus. The differences in these two types of error are computed for 1, 2, 3 trials back and for 1 trial forward as a control.

percentage difference between pro-SD and anti-SD classification errors (Figure 5.6B). Overall, the classification error rate is 28.35%. The 1-back and 2-back percentage differences were 6.85% and 3.96% respectively, indicating the dominance of serial dependence in the sequential effects in perceptual decisions of participants. Essentially, when there are classification errors, these are much more likely to be in the direction of previous stimuli. The 3-back percentage difference was 0.1%. Overall, serial dependence dominated the sequential effects for 1 and 2 trials back. In addition, the sanity check of 1 trial forward, 0.03%, shows no influence of future trials on classification errors in the current trial. This is expected and confirms that there were no artifacts masquerading as serial dependence.

5.4 Discussion

Serial dependence in medical image perception has been studied for years[171, 174]. However, none of the previous research used realistic medical images. In previous studies, the stimuli

incorporated simple geometric shapes and artificial backgrounds consisting of either healthy tissue texture or simple noise patterns. Although the prior empirical results indicate the existence of serial dependence in the perception of those unrealistic stimuli, whether serial dependence extends to and occurs for realistic medical images remained unknown.

In this study, we tested whether there is serial dependence in perceptual judgments of more realistic GAN-generated radiographs. We utilized authentic-looking simulated medical image stimuli created with a Generative Adversarial Network[214, 213]. The magnitude of serial dependence found in the current study was similar to that found in previous studies. Prior studies found that perceptual judgments were pulled towards the stimulus presented in the previous trial, and the pull effect was around 15% for 1-back trials. Moreover, this effect lasted up to 10 seconds or more in the past[174]. The results in the current study were comparable. For example, the half amplitudes of the DoVM curve in Figure 5.6 show a similar effect size as that previously reported. This indicates that serial dependence affects untrained observers' judgments of the simulated radiographs. The fact that clinicians show serial dependence in other domains [171, 174], and the fact that serial dependence can increase with expertise [252] hints at the possibility that clinicians may not be immune from serial dependence. Nevertheless, whether serial dependence influences clinician judgments of the more realistic GAN-generated radiographs here remains an important question for future research.

In addition to replicating and extending the presence of serial dependencies in perceptual judgments of realistic medical images, our study also highlights the broader point that computer vision tools can be used in concert with psychophysical experiments to isolate and shed light on human performance limits. Computer vision models, in this approach, are not employed with the goal of replacing human readers. Rather, computer vision is used to create controlled stimuli that allow human performance to be more accurately assessed, controlled, and potentially enhanced. Computer vision models are in the service of human behavior.

There are several caveats and concerns that readers will have noted. It may be argued, for example, that the presentation duration of the simulated mammogram was too short (500 ms) or too low resolution (256x256) in our study, whereas clinicians typically have longer periods of time to process higher-resolution radiographs. In fact, the average fixation duration when targeting the first mass has been reported as 1.8–2s, which is surprisingly brief [147, 193]. Moreover, when scrolling through volumetric images, the viewing time in any given slice can be a fraction of a second. In addition, peripheral viewing and effectively lower resolution images can be sufficient to detect abnormalities [68, 70, 25]. Conversely, images viewed for a sufficiently long exposure duration can lead to negative aftereffects. For example, it was found that adapting normal observers to image samples of dense or fatty tissues caused a subsequent image to appear less dense (and vice versa; a type of negative aftereffect) [138, 139, 140]. Sequential effects (either repulsive or attractive) can therefore emerge across many different exposure durations.

In addition to the fixed duration of the stimuli in this experiment, this study has some additional limitations. First, we chose a continuous report matching task in our experiments,

as it provides precise trial-wise errors and has proven to be very reliable in measurements of serial dependence in the past [41, 40, 83, 75, 84, 161]. However, the actual task of the typical radiologist is far more complicated, and involves detecting, locating and classifying the lesions. Future studies should therefore implement more realistic tasks. Second, we only tested untrained observers in this study. Future studies should also recruit clinician observers. Third, the simulated mammograms were only presented briefly in our experiment, to mimic the brevity of images viewed in quick succession. To generalize the results here, it will be necessary to test which biases arise with longer presentation durations. Fourth, even though we utilized both benign and malignant images for training, we did not consider the malignancy of the stimuli in the GAN model and experiments. Future studies can investigate how malignancy can be disentangled in the GAN model and how malignancy may influence the diagnostic tasks. Our goal in this study was to test the presence of sequential effects in judgments of more realistic and controlled GAN-generated medical images and we found evidence for this. However, the caveats and concerns described here prevent us from concluding that serial dependence impacts clinical image interpretation in real clinical practice. The results raise the possibility, though, and, if there are serial dependencies in clinical interpretations, then the consecutive similarity between images from one or more patients could matter. Future work is needed to test this.

5.5 Conclusion

In this study, we utilized a Generative Adversarial Network(GAN) to produce authentic-looking GAN-generated mammograms. These realistic stimuli were used in a psychophysical experiment that tested for serial dependence in perceptual judgments. We found that the perception of the current simulated mammogram was biased towards the previously seen mammograms. On average, perceptual judgments of naturalistic GAN-generated mammograms had 7% categorization errors that were pulled in a direction consistent with serial dependence, and this pulling effect lasted up to 10 seconds in the past. Our study provides evidence that serial dependence may contribute to decision errors in the perception of realistic-looking medical images.

Chapter 6

Serial Dependence in Dermatological Judgments

6.1 Introduction

The natural visual world is autocorrelated: objects do not spontaneously pop into or out of existence in the typical visual experience, and what was present a moment ago tends to still be present at this moment. The human visual system has developed adaptive mechanisms that take advantage of these natural autocorrelations by introducing serial dependence in perceptual interpretations. Due to this mechanism, objects recognized at one moment appear more like similar objects seen in the last several seconds. The result of this serial dependence is that perceptual experience seems smoother and more stable than it should be. This is beneficial because without it, the visual world would look jittery and unstable; object identities would appear to fluctuate due to changes in lighting, viewpoint, blinks, and myriad sources of internal and external noise [75, 42].

It is intuitive that human vision benefits from recycling visual history, smoothing and stabilizing perceptual experience in the natural world. However, the benefit of serial dependence has limits because the visual world is not always natural. In certain artificial, human-designed visual tasks, such as medical image perception or randomized laboratory experiments, visual stimuli are no longer naturally autocorrelated. Visual images in these situations can vary randomly from one moment to the next. If the visual system imposes serial dependence, smoothing or reusing previous visual history, this could introduce systematic errors by attracting current perception towards previous visual history.

Studies have shown that this is exactly what happens. Serial dependence systematically biases current perception toward visual history in many tasks, such as perception of orientation [75], attractiveness [246, 244], and emotional expression [253]. Serial dependence also introduces systematic perceptual errors in medical image perception tasks [174]. However, these studies were conducted under lab conditions with highly artificial stimuli and experimental designs that are not typical in clinical practice [174]. More recently, progress in

Generative Adversarial Networks (GANs) affords the opportunity to generate more realistic simulated medical images as stimuli [213, 216]. However, even in these studies, the relatively complex psychophysical tasks were not comparable to realistic, clinically relevant scenarios.

In this study, we address several of the shortcomings in prior work and we test whether serial dependence occurs in a teledermatology setting, one of the most important and commonly employed subsets of telemedicine [204, 201]. Remote store-and-forward teledermatology, which involves sequential judgments of static images, is an especially fast growing area of telemedicine [60, 271, 269, 159], and it requires the involvement of clinicians because automated systems are not sufficient to make accurate diagnostic classifications [115, 131, 81]. The question in the present study is whether sequential judgments of dermatological lesions in a remote store-and-forward setting result in serial dependence.

We analyzed 758,139 skin cancer diagnostic judgments from 1137 participants collected from an app developed by Centaur Labs, a US medical Artificial Intelligence (AI) company based in Boston. The task was a straightforward 2AFC (two-alternative forced choice) (yes/no) discrimination, with a goal of diagnosing whether an actual skin cancer image was nevus (benign) or melanoma (malignant). This is comparable to a realistic, remote store-and-forward teledermatology task, with a more natural two-alternative forced choice (yes/no) design.

We found that there was statistically significant serial dependence in discrimination judgments that was tuned to the sequential similarity in the malignancy of the lesions. The consequence of the serial dependence was a statistically significant reduction in metrics of sensitivity and specificity, including reduced d-prime (d') and increased error rates. Additionally, using a recent Learned Perceptual Image Patch Similarity (LPIPS) computer vision model, we quantified serial dependence as a function of the semantic similarity between sequential images and found that serial dependence varied as a function of the patchwise similarity between sequential images.

Together, our results suggest that serial dependence in perceptual decisions may impact realistic dermatological judgments, at least under certain circumstances akin to those in remote store-and-forward teledermatology [247, 285].

6.2 Materials and Methods

Experiment Stimuli

All skin cancer images utilized in the trials on the app were subsampled from ISIC 2019 Challenge Datasets [251, 43, 45]. This set of images contains two types of lesion, i.e., nevus and melanoma, indicating benign and malignant cases. The images were dermoscopy images after manual correction of color hue, luminance, and alignment and were taken by different devices using polarized and non-polarized dermatoscopy. Samples of skin cancer image stimuli are shown in Figure 6.1. In summary, for all the skin cancer images that were shown, 57.3% were benign and 42.7% were malignant.

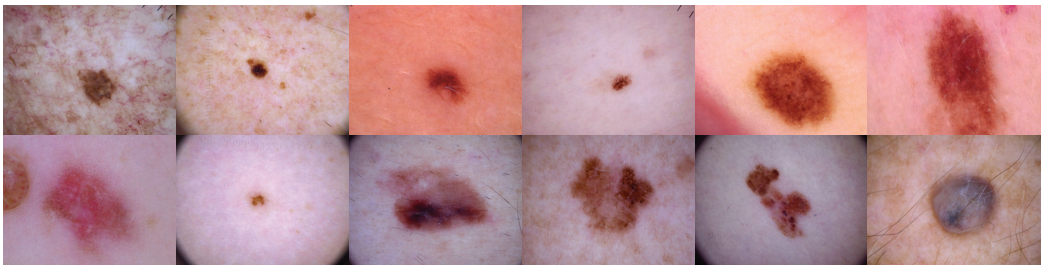


Figure 6.1: Samples of skin cancer image stimuli. A total of 7798 images were drawn from the ISIC 2019 Challenge Datasets [251, 43, 45], which contain various nevus and melanoma lesions. In each trial, a single random sample image was selected and presented to the participant. Observers judged whether the image was nevus (benign) or malignant (yes/no forced choice design). Feedback was provided after each trial.

Participants

The users of the app are predominantly medical students, with some medical residents. Individual participant information such as age, sex, and demographics that are typically gathered in scientific experiments are not known for this group of observers because this information is saved in the user profile of the app and was not available to us. However, it is known that all users had normal or corrected-to-normal vision. Since the use of the app does not work outside of the United States, users must be located in the U.S. at the time of app usage. Before using the app, users gave consent to have Centaur Labs use the data they provide through app usage. Users received earnings from a predefined money pool (around US$ 50) for each task they participated in.

Experiment Design

For the dermatological classification task that was investigated in this study, users first completed a training session of 10 trials with 10 separate stimuli. This training explained the procedure of the task and prepared users for the actual classification task, which was identical to the training.

In each trial, a random skin cancer image was selected and presented to the participant. Below the image, they were prompted to choose one of the two possible responses, “benign” or “malignant”. Feedback was provided after every trial to inform users if their response was correct or incorrect. Afterward, users voluntarily moved on to the next trial at their own pace. Users were told they could end the task at any time.

We were provided with 758,139 data points across 13 variables, which were collected between 4 September 2020 and 21 June 2021. Each data point corresponded to one decision of a user, classifying a dermatological image as either benign or malignant. After pre-processing,

756,001 data points from 1137 users were used for further analyses (pre-processing steps and exclusion criteria are illustrated in Appendix 6.5).

Serial Dependence

Serial Dependence has three main kinds of tuning. First, feature tuning: serial dependence occurs most strongly between relatively similar features and not between identical ones or highly dissimilar ones [75, 84]. For example, when two identical images are seen in succession, serial dependence does not bias judgments in any direction because the images are identical; likewise, if the two successive images are extremely different from each other (e.g., apples and oranges), then serial dependence does not bias judgments either. Only when two successive images are moderately similar is there a serial dependence in perceptual judgments. Serial dependence is also temporally tuned: the magnitude of serial dependence gradually decays over time or with intervening visual information [75, 270]. Third, spatial tuning: serial dependence occurs only within a limited spatial region, and it is strongest when previous and current objects are presented at the same location [75, 20]. In general, we can utilize feature and temporal tuning as the most important metrics to probe the serial dependence effect and to rule out other artifacts, such as simply repeating the same response or lapsing.

To measure the presence of feature tuning, we measured serial dependence as a function of the similarity in sequential stimuli. In this study, we adopt two metrics of similarity. One is malignancy similarity, where malignancy is estimated based on a popularity vote. The "similarity" in this respect is an abstracted concept based on behavioral judgments of independent observers. What counts as similar is not necessarily in the image or pixel domain but in the degree of malignancy (Figure 6.2). The second form of "similarity" that we quantified is semantic similarity, using a popular Learned Perceptual Image Patch Similarity (LPIPS) metric [290] approach borrowed from computer vision.

Malignancy Similarity

The malignancy of each stimulus was estimated based on a popularity vote: -100 means all users classified the lesion as benign; 100 means all users classified the lesion as malignant. Figure 6.2 shows the distribution of malignancy over all stimuli. "Malignancy similarity," used in subsequent analyses of serial dependence, was computed as the malignancy difference between any two sequential stimuli. Any two adjacent stimuli on the abscissa of Figure 6.2 have high similarity; conversely, any two distantly separated stimuli have low similarity.

Semantic Similarity

The semantic similarity is computed via the Learned Perceptual Image Patch Similarity (LPIPS) metric [290]. This is a popular nonlinear similarity metric utilized in computer vision. For deep learning models, there are deep features after each convolutional layer [145, 232, 106]. The semantic similarity is computed as a sum of weighted differences

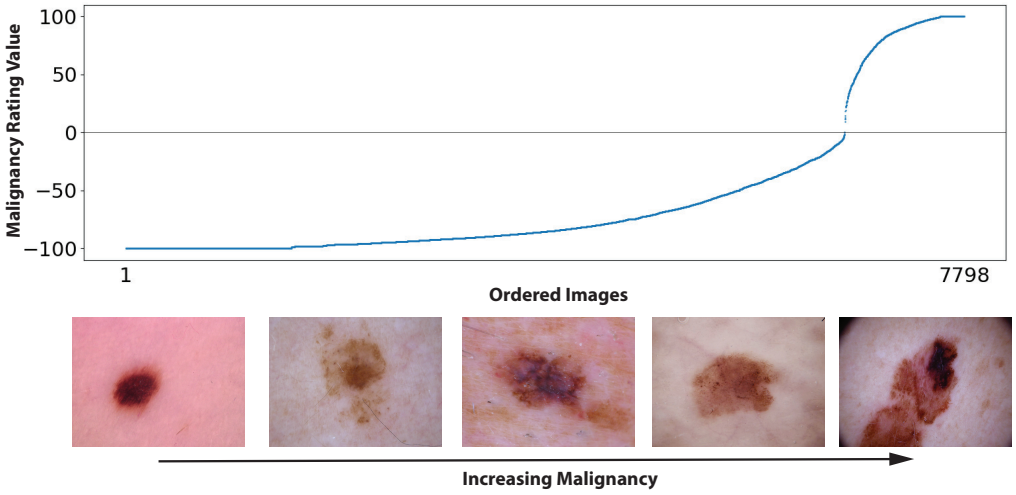


Figure 6.2: Overview of all 7798 (6688 benign, 1110 malignant) unique images used, sorted by the consensus malignancy rating value (−100: classified as benign by all users, 100: classified as malignant by all users). The five sample images below the abscissa show a sequence of example images that had varying degrees of agreement, from benign to malignant.

between the corresponding deep features at different layers. If the semantic similarity is small, two images would share more patch-wise similarity in the pixel domain, with 0 representing identical. In particular, we utilized AlexNet [145] as the backbone of the LPIPS metric. Figure 6.3 shows two groups of similar and dissimilar skin cancer images based on LPIPS metric. A similar pair is defined as a pair of images whose similarity is less than the mean similarity of all image pairs, and vice versa.

Diagnostic Performance Evaluation

To measure the presence of serial dependence, we analyzed users' performance in the dermatological classification task. Multiple metrics from signal detection theory were utilized, including *Sensitivity or Hit Rate (HR)* $= TP/(TP + FN)$, *Specificity* $= TN/(TN + FP)$, and *Error Rate* $= (FN + FP)/(TP + FN + FP + TN)$, where "Positive" (P) represents the malignant case, and "Negative" (N) represents the benign case. Then, TP (True Positive), FN (False Negative), FP (False Positive), and TN (True Negative) can be defined accordingly. We also utilized d-prime (d') and the criterion (c) to evaluate observers' discrimination and bias. These can be computed as follows:

$$d' = z(HR) - z(FAR)$$

$$c = -0.5 * (z(HR) + z(FAR))$$

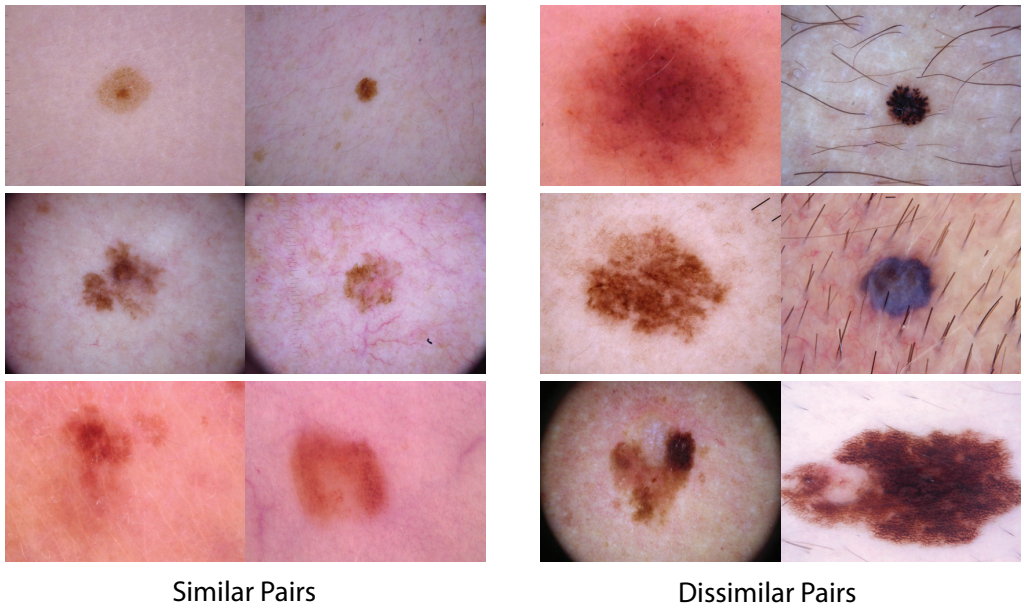


Figure 6.3: LPIPS semantic similarity [290] example image pairs. Based on this semantic metric, we can group images into similar pairs vs. dissimilar pairs. Note the patch-wise similarity that similar image pairs have.

where $z(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution, and *False Alarm Rate (FAR)* $= FP/(FP + TN)$.

Feature Tuning Analysis

We evaluated the diagnostic performance metrics described above while taking into account the sequential similarity between successive images that each observer saw. There were two types of similarity that we evaluated. In the first one, the malignancy similarity, we computed the n-back similarity as $|M_{t-n} - M_t|$, where M_t represents the malignancy of the current trial image and M_{t-n} represents the malignancy of the n-back trial image. We used the absolute value of the difference because the sign of the malignancy does not matter. Then, we grouped the malignancy similarities with a group range of 10, resulting in a total of 20 similarity groups. Performance metrics were computed within each group. In the end, we obtained the sensitivity, specificity, d', c, and error rate in relation to the n-back malignancy similarity.

The n-back semantic similarity can be obtained directly from the LPIPS metric [290], $f(I_{t-n}, I_t)$, where I_t represents the current trial image, I_{t-n} represents the n-back trial image, and $f(\cdot)$ is the LPIPS model. Then, the semantic similarities were grouped with a group range of 0.02, with groups that have insufficient trials excluded. We analyzed groups in the

semantic similarity range of $[0.3, 0.68]$. Performance metrics were also computed within each group. In the end, we obtained the performance metrics in relation to the n-back semantic similarity.

In order to probe the impact of serial dependence on diagnostic performance, we measured the net change of those metrics relative to what is expected by chance. To conservatively estimate this "chance" baseline, we used the future trial (N + 1) stimulus because this stimulus is not predictable and cannot influence the past. Essentially, because the stimuli are randomly ordered, the current response is only predictive of the future stimulus about half of the time, which gives a baseline estimate of chance performance. If the current judgment is pulled toward the previous stimulus (serial dependence), then the current trial accuracy will decrease relative to that chance performance. By using the future (N + 1) accuracy as a baseline, we control for any systematic response biases that observers might have [179, 200]. For example, simply pressing the same button on every trial results in a response bias, but this will not show up as measured serial dependence because the serial dependence is normalized relative to the N + 1 trial.

Finally, we computed the net change in sensitivity, specificity, d', criterion, and error rate as a function of the sequential similarity between successive images. As serial dependence only occurs for relatively similar features, we expected the serial dependence effect, if present, to be maximal when sequential stimuli are moderately similar.

Temporal Tuning Analysis

After checking the feature tuning characteristics, we fit Gaussian curves (Equation (6.1)) on top of the net change graphs to quantify the magnitude of the serial dependence effect (as shown in Section 6.3).

$$f(x|\mu, \sigma^2) = \frac{a}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.1}$$

where x is the data variable, μ and σ are the mean and standard deviation of the Gaussian distribution, and a is an amplitude modulation parameter. Here, a, μ, and σ will be optimized during curve fitting. After fitting, we report the peak value of the fitted Gaussian curve as the amplitude of the serial dependence effect.

We analyzed the serial dependence effect magnitude of 1-back (N-1), 2-back (N-2), 3-back (N-3), and 4-back (N-4) trials. Then, we obtained the relation between the serial dependence effect magnitude and intervening time between trials.

6.3 Results

Overall summary statistics revealed that observers were highly sensitive to the malignancy discrimination task. Across the user population, sensitivity was 78.72%, specificity was

74.74%, d' was 1.46, c was −0.065, and the error rate was 18.6%. These metrics indicate that observers were able to perform the dermatological judgment task, consistent with the observers having some degree of expertise. These overall metrics, however, do not reveal whether dermatological judgments on a given image are impacted by sequential dependencies.

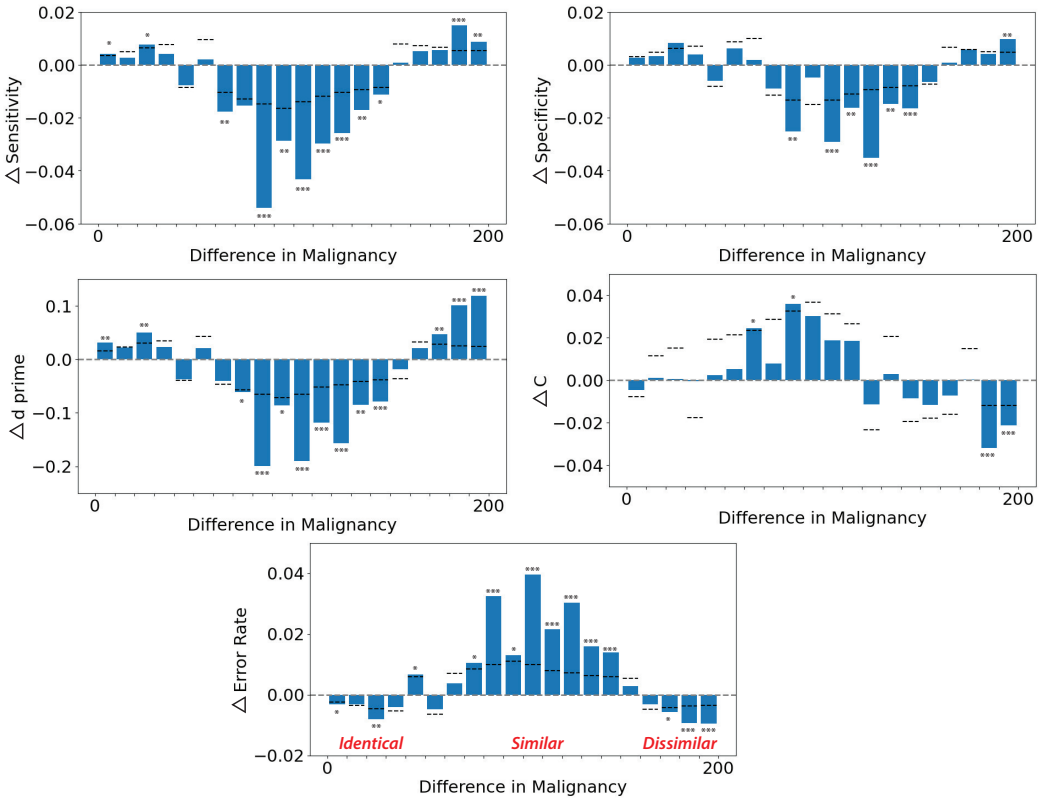


Figure 6.4: Serial dependence in dermatological classification judgments negatively impacts performance. Performance in the discrimination task was assessed with metrics of sensitivity, specificity, d-prime (d'), criterion (c), and error rate. The abscissa of each graph shows the similarity in the rated malignancy (Figure 6.2) of successive pairs of images; 0 represents identical successive images, and 200 represents very different sequential images. The ordinate of each graph shows the net change in performance metric (e.g., sensitivity or d') on the current trial as a function of the similarity of the previous stimulus (N-1 trial) seen by the observer. When the previous stimulus was moderately similar (central regions on the abscissa), all performance metrics dropped, indicating worse performance. For example, when the sequential images were moderately similar, there was an increase in error rates of up to 4.1% on the current trial. Horizontal dashed lines indicate the upper 95% boundary of the permuted null distribution for each bar.Asterisks indicate statistical significance ($*$: $p < 0.05; ** : p < 0.01; *** : p < 0.001$).

Our primary goal was to measure whether serial dependence was present in dermatological judgments. To do this, we calculated the performance metrics above on a trial-wise basis, as a function of the sequential similarity between successive images, as illustrated in Section 6.2.

Figure 6.4 shows the net change in sensitivity, specificity, d', criterion (c), and error rate as a function of the malignancy similarity between current and previous images (1-back or N-1 trial image). The abscissa of each graph shows the similarity in the rated malignancy (Figure 6.2) of successive pairs of images; 0 represents identical successive images, 200 represents very different sequential images, and the middle range represents similar images. When the previous stimulus was moderately similar (central regions on the abscissa), all performance metrics dropped, indicating worse performance. The worst case occurred when the uncertainty reached the maximum. This is consistent with the findings in previous studies [75, 174]. In summary, sensitivity decreased up to 5.4% on the current trial, specificity was decreased up to 3.5% on the current trial, d' was decreased up to 0.20 on the current trial, criterion (c) was biased up to 0.036 on the current trial, and the error rate was increased up to 4.1% on the current trial. Horizontal dashed lines indicate the upper 95% boundary of the permuted null distribution for each bar. Asterisks indicate statistical significance ($p < 0.05, 0.01, 0.001$).

As the semantic similarity via the LPIPS metric is nonlinear, we clustered the performance metrics within small groups into two super-groups, i.e., groups of similar and dissimilar images. The 1-back (N-1) net change in performance for similar and dissimilar sequential images is shown in Figure 6.5. When similar sequential images were viewed by participants ("similar" on the abscissa), participants had higher error rates, lower specificity, and biased criterion. In particular, the net change in the error rate from similar to dissimilar groups was up to 3.38%, the net change in the specificity from similar to dissimilar groups was up to 7.53%, and the net change in the criterion from similar to dissimilar groups was up to 0.185. The was not a significant change in d' or sensitivity between similar and dissimilar groups. Overall, there was a negative impact of serial dependence on performance measured by most metrics, including, crucially, the error rate.

After analyzing 1-back (N-1) serial dependence via malignancy similarity, we conducted the same analysis for 2-back (N-2), 3-back (N-3), and 4-back (N-4) trials. Then, Gaussian curves (as described in Equation (6.1)) were fitted onto the intermediate results of feature tuning as shown in Figure 6.6A,B. The amplitude was taken as a measure of the impact of serial dependence on error rates and d'. As shown in Figure 6.6C, the amplitude of the Gaussian was the strongest for the N-1 stimulus and weaker for the following N-2, N-3, and N-4 stimuli, indicating that serial dependence is temporally tuned—stronger for more recent similar stimuli. In particular, the serial dependence (SD) amplitude for error rates decreased from 3.14% to 0.63%, and the SD amplitude for d' decreased from 0.17 to 0.038.

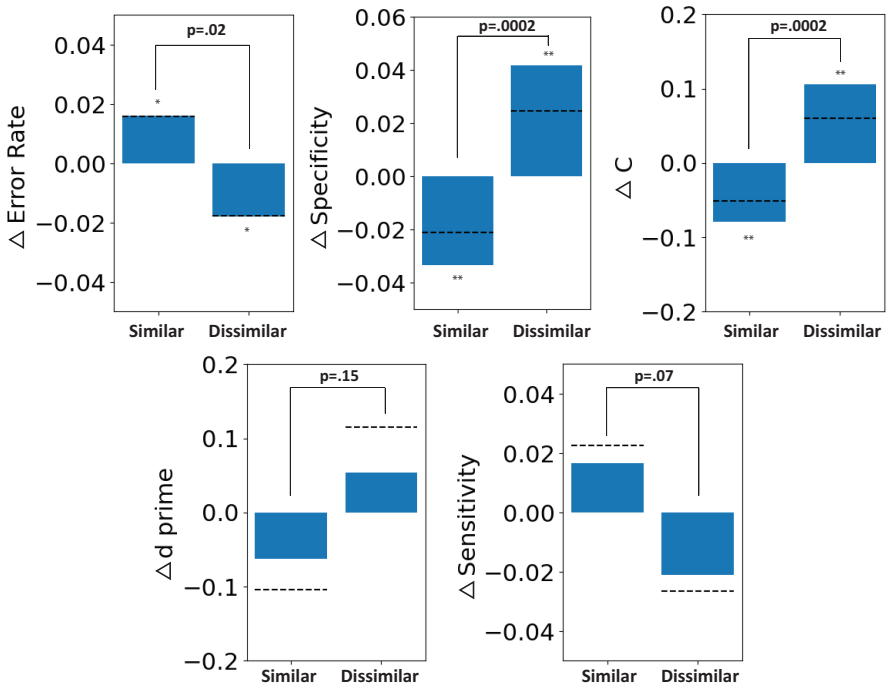


Figure 6.5: Serial dependence in dermatological discrimination judgments impacts performance. Asterisks indicate statistical significance ($* : p < 0.05; ** : p < 0.01$). Here, the similarity between sequential images was measured using the LPIPS metric [290]. When similar sequential images were viewed by participants ("similar" on the abscissa), participants had higher error rates, lower specificity, and biased criterion. Sensitivity was not negatively impacted, interestingly, but this was not significant and did not counteract the negative impacts found in all other metrics.

6.4 Discussion

The goal of this study was to test if there is serial dependence in the perceptual judgments of real skin lesions in a relatively realistic situation akin to remote store-and-forward teledermatology [60, 269, 74, 255]. We found that there was significant serial dependence in observer judgments of malignancy, and this effect was tuned to the similarity in the sequential images. Moreover, the effects were temporally tuned, strongest for more recent similar stimuli, consistent with the diagnostic criteria of serial dependence.

Serial dependence is a specific process in which the brain smooths perceptual interpretations over time to improve efficiency and accuracy and stabilizes the appearance of the natural world [75, 42]. Serial dependence has been found in many perceptual tasks ranging from low-level [94] to high-level cognition [172]. It has also been reported in some clinically relevant domains but with less realistic stimuli and tasks [174]. Serial dependence is not

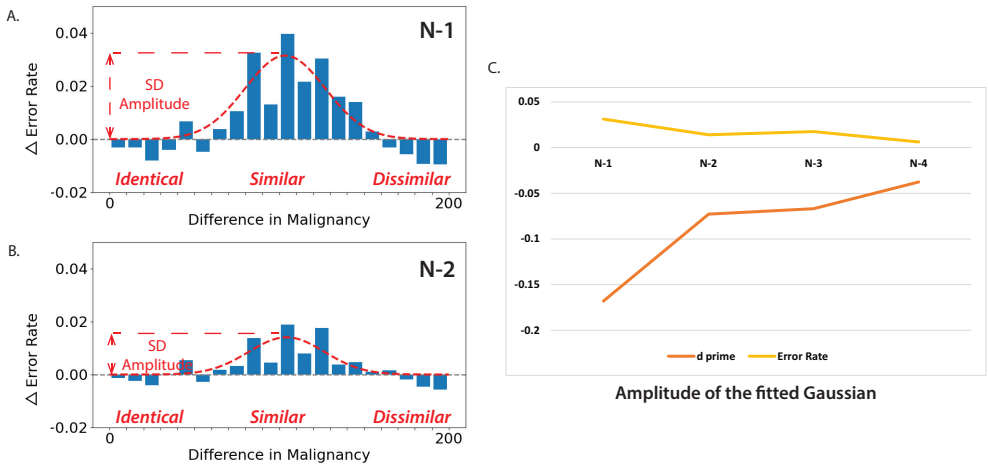


Figure 6.6: Serial dependence in dermatological discrimination judgments is temporally tuned. (**A**) Error rates such as those in Figure 6.4 were computed for 1-back trials (just as in Figure 6.4) and (**B**) for 2-back trials. The increased error rate near the central part of the abscissa indicates that the similarity in the image presented 2 trials before the current trial impacted performance, but less so than the impact of the 1-back stimulus. Gaussian curves were fit to the change in error rates as well as in d', and the amplitude was taken as a measure of the impact of serial dependence (SD) on error rates and d'. (**C**) The amplitude of the Gaussian—the strength of serial dependence (SD)—was the strongest for the N-1 stimulus and weaker for the following N-2, N-3, and N-4 stimuli, indicating that serial dependence is temporally tuned—stronger for more recent similar stimuli.

a generalized repetition of responses, and it is not just lapsing, central tendency biases, or other artitfacts [75, 42, 200].

The serial dependence effect we found here is not due to artifacts such as lapsing, central tendency, repeated button presses, or perseverating on the same response. Those kinds of artifacts are problematic, and they can have a serious detrimental influence on dermatological judgments, but they are not serial dependence, per se. As in previous studies [75, 84, 200], here, we dissociated serial dependence from these other artifacts using three approaches. First, we confirmed that the measured serial dependence effect here was tuned to the sequential similarity between images. A perseverating or stereotyped response (e.g., pressing the same button over and over again for any number of reasons) does not result in biases that are tuned to the similarity between sequential images. Instead, it simply results in a uniform and stable shift in criterion. Second, we dissociated serial dependence from lapsing, stereotyping, and other artifacts by controlling for any biases that seem to depend on the future. Serial dependence is mainly a bias of the current perceptual decision toward past experience. The future stimulus is unpredictable and random, and therefore cannot influence the current decision. However, if there are stereotyped responses (e.g., simply re-

peating the same button press or central tendency biases), this will result in what seems like the future being predictive of the present. By subtracting out this future bias, we isolated the 1-trial back effect. This approach—measuring and controlling artifacts by using the future—is a common control in studies of serial dependence [179, 200, 174, 216]. Finally, in a third control, we created permuted and shuffled null distributions. These control for overall biases, lapsing, and stereotyped responses among other potential artifacts as well. All of these controls together demonstrate that serial dependence genuinely impacted performance in dermatology judgments.

Previous studies have tried to measure criterion and d' in dermatological judgments over time [268, 153, 263], but they did not examine trial-wise effects. Serial dependence is a trial-by-trial effect [75, 42, 174, 216]: sometimes it happens in random sequences (when sequential stimuli are coincidentally similar) and sometimes it does not happen (when sequential stimuli happen to be different). In typical vision science experiments, stimuli are random and their sequential similarity is not measured, considered, or controlled. Serial dependence will therefore not show up in typical analyses because (1) responses are pooled or collapsed across blocks of trials and (2) sequential similarity is unknown or ignored. So, it is not surprising that serial dependence was not found in a previous study [277] because that study did not measure sequential stimulus similarity and it pooled trials together in blocks, washing out any serial dependence that may have been present. The results of the large data set here confirm that serial dependence is likely to be present in other similar data sets, such as [277]. Serial dependence does not show up in simple signal detection metrics such as d' and criterion unless one takes into account the trial-wise nature of the effect. Serial dependence is not just a shift in the criterion, and it is not just a change in d'. It can result in both shifts in criterion and d', as we found here, but these are dynamic over time–they fluctuate from trial to trial. We were able to measure changes in SDT (Signal Detection Theory) metrics including d' and criterion because we analyzed the data in a trial-wise manner and, more importantly, conditioned the analysis on the sequential similarity between stimuli. We found that d' decreases for similar (non-identical) stimuli. However, if the stimuli are nearly identical or are very different, then there is no decrease in d'. Likewise, we found that criterion changed depending on the sequential similarity between successive stimuli. Both of these results are important: they indicate that standard SDT metrics including d' and criterion should not be treated as rigid and fixed over time but should be considered as dynamic features that can reflect the fluctuations of stimuli in the world. Future studies of clinician perception and performance should consider the dynamic nature of signal detection metrics.

Serial dependence is a phenomenon that has been observed in many domains, from low-level perception to high-level cognition [75, 137, 42, 94, 172]. An outstanding question in the literature on the basic mechanisms of serial dependence is whether feedback might modulate it. For example, one might speculate that trial-wise feedback could reduce or eliminate serial dependence. The results here speak to this question because observers did receive feedback during the task. Despite that feedback, there was still a significant serial dependence that was tuned to both feature similarity and time. This suggests that feedback (even where it is possible) is not a panacea to eliminate serial dependence. Pragmatically, of course,

feedback is not possible in clinically relevant settings because there is no prior ground truth in medical image perception. Nevertheless, it is theoretically and practically valuable to know that feedback is not enough to overcome the visual system's built-in smoothing operations that cause serial dependence.

There are several limitations in this study. First, this study only investigated one source of perceptual bias–serial dependence. Of course, there are other sources of bias, individual observer differences, attentional differences, lapsing, and myriad other sources of error. We controlled these because our goal was to isolate one particular operationally defined source of perceptual bias: serial dependence. Whether there are interactions among serial dependence and other types of perceptual bias is an open question for future research. A second limitation of this study is that the skin cancer images utilized in the experiment contained only two types of lesion, i.e., nevus (benign) and melanoma (malignant). Though the dermatological classification task is similar to realistic skin cancer diagnostic scenarios in some teledermatology settings, it does not fully capture the range or variety of various skin cancer disease types. Moreover, for the images presented to participants, 57.3% were benign and 42.7% were malignant. This deviates from a realistic distribution, where malignant cases are typically much rarer than benign cases. That said, serial dependence does not hinge on the rate of malignancy—it impacts d' independent of target frequency, and it is, therefore, likely to occur even for rare target situations. However, the issue of disease prevalence remains a very important and open question for future research.

Another limitation is that this study is restricted to store-and-forward teledermatology, which is naturally different from office-based dermatology clinics in several ways, such as available resources and diagnostic procedures. For example, office-based clinicians have multi-modal information about the lesion available, not just photographs, and clinical decisions are more complex than binary ones as in our study. However, during the COVID-19 pandemic, we witnessed a rapid shift from office-based dermatology clinics into teledermatology [205, 206]. In line with these recent developments, the teledermatology market size is forecasted to be $67.43 billion in 2030 [121, 217]. Accordingly, we chose to investigate remote store-and-forward teledermatology, as it is a highly scalable and increasingly employed form of telemedicine. Finally, it is important to mention that most participants recruited in this study were medical students rather than experts. However, clinicians are not always more accurate than medical students or residents [277]. The reasons for this difference in performance might be the recency of training, attention, or other factors. The simple assumption that trained (older) clinicians are better than less trained (younger) ones is not clear for remote store-and-forward teledermatology, in particular. Future research is needed to explore how expertise might interact with remote teledermatology [277].

There are several additional important avenues of future investigation. Future work should test whether the serial dependence found here is spatially tuned. For example, if sequential images were viewed on different screens (rather than a single mobile device), would there be a reduction in serial dependence? Moreover, how does attention to the task modulate the serial dependence in dermatological judgments? Future studies should address these questions, along with designs that incorporate a larger variety of lesions and a more

realistic distribution of malignancy. Finally, future studies should also focus on how to utilize serial dependence tuning functions, i.e., feature tuning and temporal tuning that we found here to potentially alleviate the biases reported here.

6.5 Conclusions

In this study, we analyzed 758,139 skin cancer diagnostic records from an online app in which participants made a series of malignancy discrimination judgments. We quantified sequential malignancy similarity and sequential semantic similarity between successively viewed images, and we investigated classification performance as a function of these similarity metrics. We found significant serial dependence effects in perceptual discrimination judgments, which negatively impacted performance measures, including sensitivity, specificity, and error rates. Moreover, we showed that the serial dependence was tuned to the similarity in the images, and it decayed over time. These findings help understand one potential source of systematic bias and errors in medical image perception tasks and hint at useful approaches that could alleviate the errors due to serial dependence.

Appendix A. Data Preprocessing

In total, 7 of the 13 variables of the data provided by the app are of interest to this research paper and define each data point. They are defined as: User ID (defining a unique ID for each user of the app), score (defining if the answer given has been correct (100) or incorrect (0)), response submitted at (defining at what particular time the response of the user was given), problem appeared at (defining at what particular time the image appeared on the device of the user), origin (defining the image name of the particular image shown), current correct answer (defining if the correct answer is either malignant or benign), and chosen answer (defining if the answer given is either malignant or benign).

Prior to analyses, the following steps were conducted to include only valid data points in the analyses: first, all data points with a larger response time than 3600 s (1 h) were excluded. As data were collected on a smartphone app, it is assumed that for responses over 1 h, the app was running without users paying attention to it. Second, all remaining data points with a longer response time than three standard deviations of the raw data were excluded, which is a common method to exclude outliers [182]. Third, all users with less than 10 trials were excluded to achieve reliable data for the calculation of n-back accuracy. In total, 1083 data points were excluded due to invalidity. The exclusion of these data points did not qualitatively change the pattern of results.

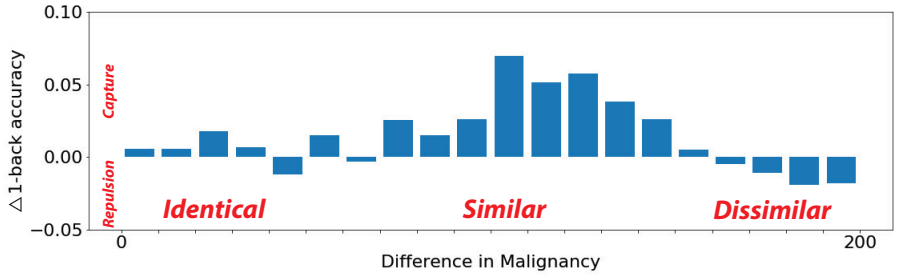


Figure 6.7: Relationship between difference in malignancy and the 1-back accuracy. The abscissa shows the similarity in the rated malignancy (Figure 6.2) of successive pairs of images; 0 represents identical successive images, and 200 represents very different sequential images. The ordinate shows the net change in 1-back accuracy on the current trial as a function of the similarity of the previous stimulus (N-1 trial) seen by the observer. When the previous stimulus was moderately similar (central regions on the abscissa), responses were consistently attracted towards the previous stimulus. This pulling effect was up to 7%. The dynamic change of the 1-back accuracy is consistent with performance metrics' change in Figure 6.4.

Appendix B. Evidence of Attracting Effect

Overall, we found significant serial dependence effects in dermatological discrimination judgments. One of the important properties of serial dependence is the attracting effect. Here, we show evidence of attraction in perceptual discrimination judgments as well. We defined the 1-back accuracy as

$$\textit{1-back accuracy} = \frac{\#\textit{current response} == \textit{previous stimulus label}}{\#\textit{trials}}$$

Next, we measured the net change of the 1-back accuracy relative to what is expected by chance. Similarly, we used the future trial (N + 1) stimulus as the "chance" baseline. If there is an attracting effect, the 1-back accuracy will be greater than 0. In reverse, if a repulsion effect occurs, the 1-back accuracy will be smaller than 0. In summary, we found evidence of an attracting effect when previous stimuli were moderately similar, thus aligning with the serial dependence property (Figure 6.7).

Chapter 7

Idiosyncratic biases in the perception of medical images

7.1 Background

Clinicians routinely depend on visual processing to make diagnostic decisions, and their experience with medical image stimuli and tasks is understood to be critical. Indeed, proficiency largely determines whether clinicians can perform well in clinical diagnostic tasks [150, 146, 223]. However, studies have repeatedly demonstrated that individual clinicians vary significantly in their diagnostic performance [66, 72, 11, 63, 242, 64, 155, 65]. It is of crucial importance to understand the nature of these individual variations, because this could help optimize training and selection criteria to improve clinical diagnostic accuracy.

There are two axes that characterize clinicians' proficiency in medical image perception. The first one—the one that has been most intensively studied in the past—is the visual sensitivity of clinicians [233, 47, 15, 154, 238, 237]. Visual sensitivity here refers to the clinicians' visuospatial and object recognition skills, which contribute to the individual variations in diagnostic accuracy. Sensitivity differences could originate from genetic variations that affect basic visual perceptual abilities of human observers [275, 294, 265, 295], as well as variability in clinician experience and training [166, 63, 13, 176, 184, 220].

The second axis that characterizes proficiency is also the under-explored one: the visual biases of individual clinicians. In the past decade, accumulating research has revealed that untrained observers can have many visual biases[248, 75, 197] and these biases can vary strongly from individual to individual [275, 125, 264, 224, 270, 97, 274, 28, 49, 266, 48, 267]. These idiosyncratic biases exist at every level of human visual perception, from the lowest level such as localization, motion, and color perception[224, 270, 142, 126, 67, 266], to higher-level object and face perception[275, 219, 28, 49, 48]. For instance, despite the extensive exposure to faces, human observers vary dramatically in their face recognition abilities[56, 222, 221, 264, 21]. Recent studies have started to shed light on this topic and revealed that clinicians, as human observers, also have their own visual biases towards medical images

[174, 216]. These biases could serve as a non-exclusive, alternative origin of the substantial individual differences in clinician diagnostic performance.

But, how, exactly, are individual biases related to diagnostic performance? This relationship—between perceptual biases and proficiency—remains unanswered in the previous literature. There are two opposing predictions about their association: 1) Biases could be reduced among skilled readers. This might be expected, given the numerous studies revealing that training can reduce visual biases [256, 102, 110, 104, 192, 54, 89]. 2) An opposing prediction is that biases will not vanish with proficiency [174, 267, 216]. Instead, biases may be exaggerated or even more consistent within more skilled individual readers. On its face, this is counterintuitive. However, a recent study hinted this possibility in untrained observers performing non-diagnostic tasks [267]. Nevertheless, it remains unclear whether diagnostic performance in skilled observers could be directly associated with visual biases.

To address this possibility, we analyzed a relatively large dataset of dermatological judgments collected through a digital medical training application, DiagnosUs, containing $758,139$ diagnoses from $1,173$ participants, using $7,818$ images. Dermatological judgments are ideal for addressing the possible association between proficiency and perceptual biases because images of skin lesions are naturally limited to two-dimensions (non volumetric) within the visual modality, they are increasingly used in remote store-and-forward applications, and they are available at a large scale.

To isolate the precise nature of idiosyncratic biases, we characterized the individual stimulus-level effects, by breaking down clinicians' biases based on image content, using a deep computer vision model. It is likely that medical images vary in their ambiguity, and thus vary in difficulty and uncertainty [235]. Because visual biases can be exaggerated when uncertainty increases [75, 142], it is conceivable that perceptual biases manifest under more difficult circumstances. We employed a novel image clustering technique to perform content-based image analysis and further investigated whether individual differences in visual biases remain homogeneous across different lesion images. To foreshadow, our results revealed that these biases are more prominent among images that have higher uncertainty (difficulty), echoing our hypothesis about increasing biases under ambiguous circumstances.

Together, our findings indicate that medical trainees have unique, idiosyncratic biases in their perception of skin lesion images. These biases do not vanish even in highly skilled observers. Instead, surprisingly, they turn out to play an essential role in distinguishing skilled observers (high performers) from less skilled observers (low performers), especially when ambiguous and difficult medical images are involved. Our study provides a new perspective to potentially improve clinicians' diagnostic performance by further understanding the individual biases that help skilled observers stand out.

7.2 Methods

Datasets and participants

The data used in this research comprises 7,818 pigmented skin lesion images along with user melanoma diagnoses. The images are curated by the International Skin Imaging Collaboration (ISIC) [251, 43, 45], which is the largest publicly available collection of quality-controlled dermoscopic images of skin lesions. The dataset contains two types of lesion, nevus and melanoma, corresponding to benign and malignant cases. The dermoscopy images underwent manual correction of color hue, luminance, and alignment and were captured via different devices using polarized and non-polarized dermoscopy. Samples of skin lesion image stimuli are shown in Fig. 7.1.

Skin lesion diagnoses were collected through DiagnosUs, an app developed by Centaur Labs, a US medical Artificial Intelligence (AI) company based in Boston, MA. The diagnosis dataset contains the skin lesion image ID reference to the ISIC Archive, the participant anonymous ID, the participant diagnosis, and the response time. The original diagnostic dataset contained 758,139 diagnoses from 1,173 participants.

The participants were mostly composed of medical students, with some medical residents. Individual subject information such as age or sex is not known. All participants have normal or corrected-to-normal vision. Users receive earnings from a predefined money pool (around 50 USD) for each task they complete.

Task

After downloading the DiagnosUs app and giving consent to have Centaur Labs use the data they provide through app usage, users can choose between different tasks. For the dermatological classification task that was investigated in this study, users first completed a training session of 10 trials with 10 separate stimuli. This training explained the procedure of the task and prepared users for the actual classification task, which was identical to the training. In each trial, a random skin lesion image was selected and presented to the participant. Below the image, they were prompted to choose one of the two possible responses, "benign" or "malignant". Feedback was provided after every trial to inform users if their response was correct or incorrect. Afterward, users voluntarily moved on to the next trial at their own pace. Users were told they could end the task at any time.

Feature extraction and clustering

We used the top-performance model of SIIM-ISIC Melanoma Classification Challenge to extract image features [101]. This deep learning model is an ensemble of convolutional neural network (CNN) models with different architectures. In this study, we only used the best-performing model within the ensemble for embedding. The model was pretrained by predicting the diagnosis for each skin lesion image. In the end, the model would yield an

embedding of a 2048 dimension, i.e., the feature vector, for each skin lesion image. Then, we utilized t-SNE[254] to map the 2048-dimensional embeddings into a two-dimensional map. Moreover, we clustered the $7,818$ mapped image embeddings into 100 clusters using the K-means algorithm [168]. On average, image clusters contained 78 images.

Data analysis

We first filtered out diagnoses with negative response time (only one diagnosis). Given that response times spanned up to multiple hours, we identified outlier data points using the interquartile range defined as $IQR = Q_3 - Q_1$ [53, 261], where Q_1 is the 25th percentile of the response times, and Q_3 the 75th percentile. Outlier diagnoses were identified as the data points with response time lower than $Q_1 - 1.5 * IQR$ or higher than $Q_3 + 1.5 * IQR$. In the end, it shrinks the analysis size to $76,051$ diagnoses.

While computing accuracy standard deviations of image clusters and estimating participants' proficiency, we randomly sampled 25% of each participant's diagnoses. With the remaining 75% of data, we filtered out participants without at least 2 trials in each of the image clusters. In other words, we only used diagnoses of participants with at least 2 diagnoses in each of the 100 clusters, leaving 81 participants and $333,600$ diagnoses remaining for analysis.

Within-subject correlation in diagnostic accuracy was calculated with a split-half correlation for each participant with respect to image clusters. We first assigned every diagnosis to their associated image clusters. We then used a non-parametric bootstrap method to estimate split-half correlations [62]. On each iteration, for each participant and each image cluster, we randomly split the diagnoses into two halves and calculated the mean accuracy for each half. Next, the two halves were correlated and then the Pearson's r value was transformed into a Fisher z value. We averaged the z values across all participants for the first analysis, and then within each performance group for the second analysis. We repeated this procedure $1,000$ times to estimate the mean within-subject correlations and 95% bootstrapped confidence intervals.

Between-subject consistency was calculated similarly. After splitting every participant's data into two random halves (i.e., by randomly selecting with replacement 50% of the data on each iteration), we correlated one half from one participant with one half from another participant. At each iteration, 200 random pairs of participants were sampled, and the pairwise correlations were then averaged to estimate the between-subject consistency. By repeating the procedure $1,000$ times, we obtained the mean between-subject correlations and 95% bootstrapped confidence intervals. The first analysis relied on all participants, while the second analysis computed between-subject consistency for each performance group separately.

Next, we estimated the expected chance level within and between-subject correlations by calculating permuted null distributions. On each iteration, and for each participant and image cluster, we again split the diagnoses into two halves, as we did in the bootstrap procedure. We then randomly shuffled the diagnostic accuracy values across clusters. The

resulting correlations from individual participants (within-subject) or different pairs of participants (between-subject) were averaged together to get the permuted within-subject or between-subject correlations. This permutation method allowed us to estimate the null correlations by correlating the response accuracy of different stimuli with each other while at the same time preserving the relationship between similar stimuli[58, 59]. This permutation procedure was repeated $1,000$ times to estimate permuted null distributions for within-subject and between-subject consistency. We performed this test successively with all participants and then separately for each performance group. The mean empirical bootstrapped correlations were then compared to their corresponding permuted null distributions to estimate the statistical significance of the mean bootstrapped within and between-subject correlations.

7.3 Results

All our analyses of participants' diagnostic performance were conducted on a fine scale of 100 skin cancer "image clusters" to reveal more detail about individual differences. Image clusters were formed by grouping embeddings from the EfficientNet computer vision model[243, 101], trained to diagnose skin lesion images. Fig. 7.1A shows the skin lesion image samples and the corresponding color-coded embeddings by malignancy. Visualization is implemented via t-SNE[254]. Each dot in the figure represents a skin lesion image. Fig. 7.1B shows the 100 clusters produced by the computer vision model[101] via the K-Means clustering algorithm [168].

By grouping neighboring embeddings into clusters, we expected images within one cluster to be "semantically similar". One aspect of similarity seems to be the malignancy of images, as detailed in Fig. 7.1A. Another aspect of this similarity may be captured by Fig. 7.2 where patterns of diagnostic accuracy and standard deviation seem to arise from the image clusters. The diagnostic test accuracy of each image cluster is defined as $Accuracy = \frac{\text{Hits + Correct Rejections}}{\text{Number of diagnoses}}$. The gold standard test is used as ground truth, with melanoma diagnoses defined as positive instances and nevus diagnoses as negative instances.

To better analyze individual differences, we computed participants' diagnostic accuracy relative to the average acurracy within each of the 100 clusters. Fig. 7.3 illustrates the relative accuracy of 5 representative participants. In addition to the obvious deviations from the group performance, and unique patterns between the individual observers, it is also clear that there are many individual differences, particularly in the high standard deviation regions (Fig. 7.2B) between the benign and malignant groups of images (Fig. 7.1A). Fig. 7.12 shows a broader overview of individual differences through the relative accuracy of 30 participants. Additional performance metrics, such as sensitivity, specificity, d prime (d'), and criterion (c) are depicted in Fig. 7.8, 7.9, 7.10, and 7.11 and the average of each metric across all participants can be found in Fig. 7.7.

To investigate the individual differences, we calculated the within-subject correlation and between-subject correlation based on participants' diagnostic accuracy over corresponding skin lesion image clusters. We obtained significant within-participant correlation (Fig. 7.4,

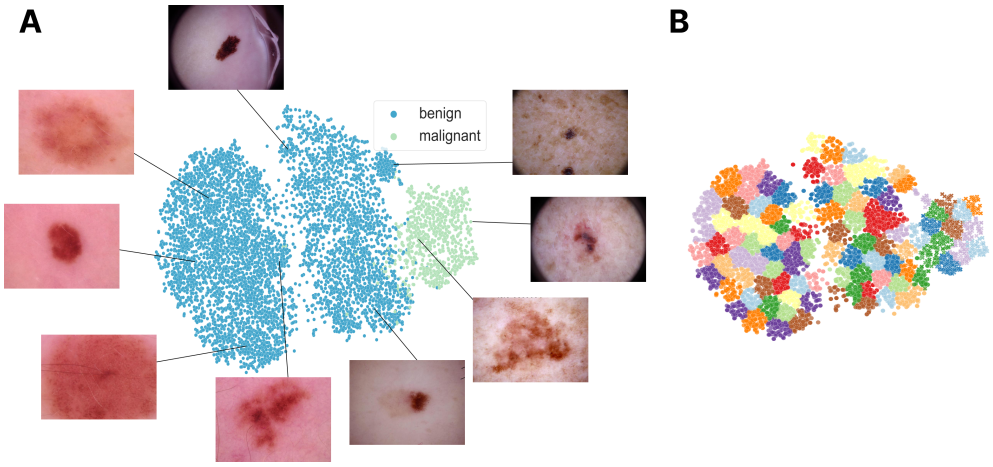


Figure 7.1: **(A)** Skin cancer samples and their corresponding embeddings. Each dot represents one of the 7,818 skin lesion images. The position of each dot is defined by the internal image representation of the computer vision model. The model has been trained to diagnose skin lesion images and reaches an almost perfect accuracy [101]. It is therefore expected that benign and malignant images are spatially separated. This is one aspect of semantic similarity captured by these embeddings, images seem to be spatially located according to malignancy. **(B)** The 100 image clusters, represented with different colors, each cluster containing dozens of similar skin lesion images. Due to the large number of clusters, some colors occur multiple times. Participants' skin lesion diagnostic performance metrics were evaluated on those clusters.

orange bar; mean $r = 0.72$, permutation test, $p < 0.001$) and between-participant correlation (Fig. 7.4, blue bar; mean $r = 0.62$, permutation test, $p < 0.001$). Importantly, the within-subject correlation is significantly higher than the between-subject correlation (permutation test, $p < 0.001$). This indicates that observers agree with themselves more than they agree with each other. As expected, this aligns with previous individual difference findings on medical image perception tasks [267].

We also investigated the individual differences as a function of proficiency. To measure participants' proficiency, we randomly sampled a fifth of each participant's response diagnoses to estimate their individual diagnostic accuracy. Participants were then split into two halves, a "high-performance group" and a "low-performance group". We conducted the same individual difference analysis as performed in Fig. 7.4 for each proficiency group and found that results similar to Fig. 7.4 also hold for both groups. In other words, for both low- and high-performance groups, within-participant and between-participant correlations were significant (Fig. 7.5A, permutation tests, all $p < 0.001$; high-performance group within-subject mean Pearson's $r = 0.81$ and between-subject mean $r = 0.73$; low-performance group within-subject mean $r = 0.61$ and between-subject mean $r = 0.53$), and the within-

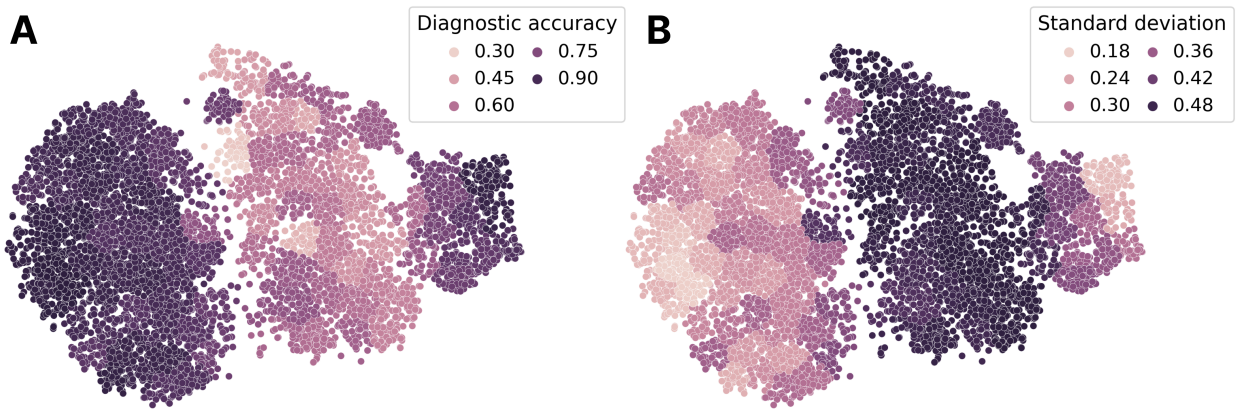


Figure 7.2: **(A)** Diagnostic test accuracy per cluster across all participants. **(B)** Standard deviation of response diagnoses per cluster across all participants. Visually, it seems that the three main groups of images (dots) are associated with different accuracy and standard deviations. Note that the location of the dots are solely defined using the computer vision model, while the color-coding is independently based on the participants' diagnoses. Given the differences in standard deviation across clusters, the model may group ambiguous images and easily classified images separately.

subject correlation was significantly higher than the between-subject correlation for each group (permutation tests, $p < 0.001$).

Among all participants' responses, the majority of the diagnoses were accurate, as illustrated by the average cluster accuracy in Fig. 7.2A. Because some disagreement is necessary for potential individual differences to arise, we analyzed idiosyncratic biases as a function of participant disagreement. To measure participant disagreement, we relied on the standard deviation of participant accuracy of each cluster. We successively subsampled clusters to measure individual biases within increasing disagreement levels.

Starting with all clusters, we successively removed clusters with the lowest disagreement levels by using a lower bound threshold on the standard deviation of participant accuracy. The first batch of clusters filtered out were the image clusters containing skin lesions that were virtually perfectly diagnosed (i.e., easy diagnoses for most observers), and the last remaining clusters contained the most contentious skin lesion images, associated with lower diagnostic accuracy. Fig. 7.5 shows the first and last cases of the 14 different disagreement levels tested. Fig. 7.5A analyzes all 100 clusters, while Fig. 7.5B analyzes the most contentious image clusters with the highest disagreement levels. The remaining thresholds will be discussed with Fig. 7.6.

For both Fig. 7.5A and 7.5B, we found that the within-participant and between-participant correlations were significant (permutation tests, $p < 0.001$), and within-subject correlations were significantly higher than the between-subject correlations (permutation tests, $p < 0.05$).

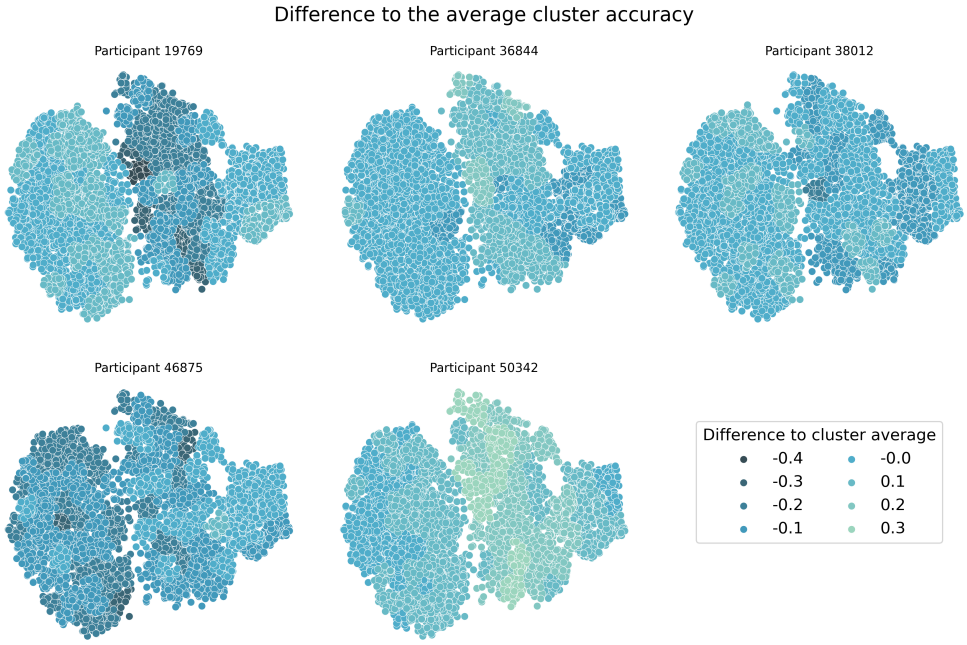


Figure 7.3: Relative diagnostic accuracy per cluster for 5 participants. Each dot represents a skin lesion image. Color coding illustrates participants' diagnostic accuracy compared to the mean performance of all participants within each cluster. The cluster average accuracy across all participants is represented in Fig. 7.2A.

This confirms that there was group-wide agreement and also significant individual differences (more consistency in the within-observer diagnoses than the between-observer diagnoses).

Moreover, for almost all disagreement levels, we found that the high-performance group showed higher within-participant and between-participant correlations than the low-performance group (permutation tests, $p < 0.001$). That means the high-performance participants had significantly more agreement both within and between themselves. When considering the most contentious clusters (Fig. 7.5B) we found that high-performance participants had a significantly lower between-subject correlation than the low-performance group (Fig. 7.5B, highest horizontal square bracket; permutation test, $p < 0.01$). That is, the high-performance group showed more disagreement than the low-performance group. This is a potential sign of higher idiosyncratic biases in the high-performance group. For less ambiguous clusters (smaller thresholds), we found that the high-performance group showed higher within-participant and between-participant correlations than the low-performance group (permutation tests, $p < 0.001$).

In a detailed analysis, we measured the average difference between the within-participant correlation and between-participant correlation, i.e., the idiosyncratic bias magnitude, at different disagreement levels for both low- and high-performance groups. We found an in-

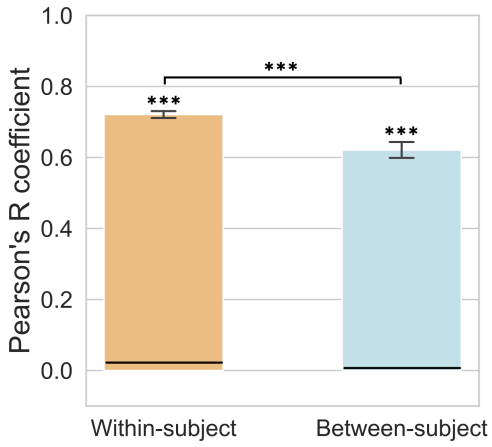


Figure 7.4: Individual differences analysis across all participants. Within-subject correlation and between-subject correlation were averaged across all participants. The within-subject correlation was significantly higher than the between-subject correlation, represented by the horizontal square bracket. Error bars represent the 95% bootstrapped confidence intervals, and the 97.5% upper bounds of the permuted null distributions for the within-subject and between-subject correlations are shown as horizontal black lines. $***p < 0.001$.

creasing idiosyncratic bias magnitude difference between the two groups with respect to the standard deviation threshold (Fig. 7.6). In other words, as the images got more difficult to diagnose, the high-performance group showed more idiosyncratic bias than the low-performance group. High-performing individuals therefore have more stable individual perceptual biases. For the cluster subsets with a standard deviation higher than 0.375, we found that the high-performance group showed a significantly higher idiosyncratic bias magnitude than the low-performance group. Thus, proficiency is associated with higher idiosyncratic biases.

7.4 Discussion

In this study, we used a large dataset of teledermatology records to isolate and identify the nature of individual observer-specific biases in the perception of skin lesions and to determine how these are related to proficiency. Our results demonstrated that, counterintuitively, proficiency is associated with increased idiosyncracy—increased and more consistent biases within individual observers. Rather than becoming more alike and homogeneous, skilled observers tend to have more unique patterns of perceptual bias. Skilled observers are not worse performers, they are simply more unique. The results confirm the importance of proficiency, and, more importantly, they reveal the growing importance of individual differences that arise with proficiency. The results have consequential implications for individualized

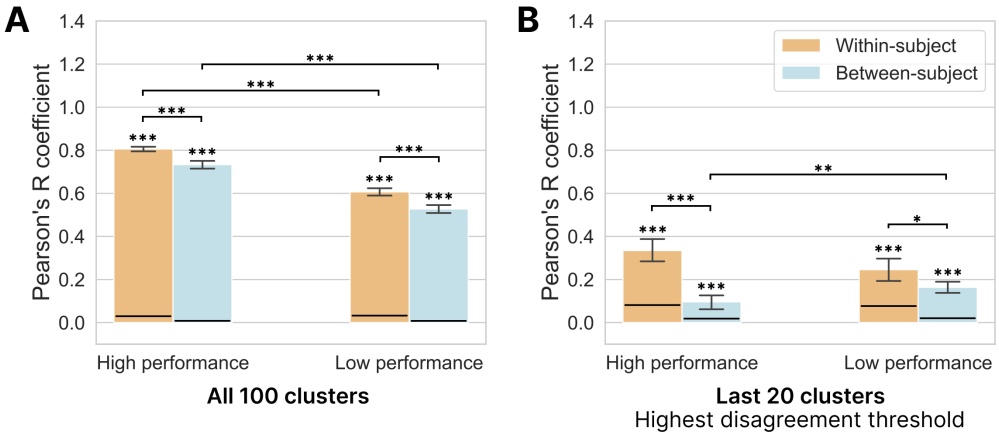


Figure 7.5: **(A)** Within-subject and between-subject correlations of the low- and high-performance groups. Correlation coefficients were significant (permutation tests, $p < 0.001$). The horizontal black lines mark the 97.5% upper bounds of the permuted null distributions. For both groups, within-subject correlations were also significantly higher than between-subject correlations, denoting that both low and high performers exhibited idiosyncratic biases. **(B)** Given a disagreement threshold, we filtered out image clusters with lower levels of participant disagreement. Using the remaining clusters, we computed the within-subject and between-subject correlations of each group. Here, we showed the 0th (A) and 100th (B) percentiles, respectively the lowest and highest disagreement threshold used. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

clinician training, paired-reader performance and optimization, bias-mitigation strategies, and the use of computer vision in assessing and assisting clinicians.

To isolate and measure individual differences in observer performance, we harnessed a computer vision model in conjunction with 758,139 skin cancer diagnostic judgments collected from 1,173 medical trainees. This computer vision model is important and necessary: it clustered skin lesion images based on pixel-level similarity, and thus provided the anchors for our analysis to unravel the counterintuitive relationship between proficiency and bias. Previous papers could not have addressed the association between proficiency and bias because they did not have a way of clustering or analyzing image properties. Therefore, previous analyses are limited to a global-level of description. Our study also uniquely harnessed the power of large scale behavioral measures: it provided solid evidence of image-specific individual differences. Indeed, it is the novel combination of massive behavioral measures along with computer vision modelling that reveals the counterintuitive relationship between proficiency and bias.

Our results may raise a number of questions that we address in the following discussion. First, it might be argued that stronger idiosyncratic biases exhibited by skilled observers could simply result from the high-performance group being more attentive to the task or

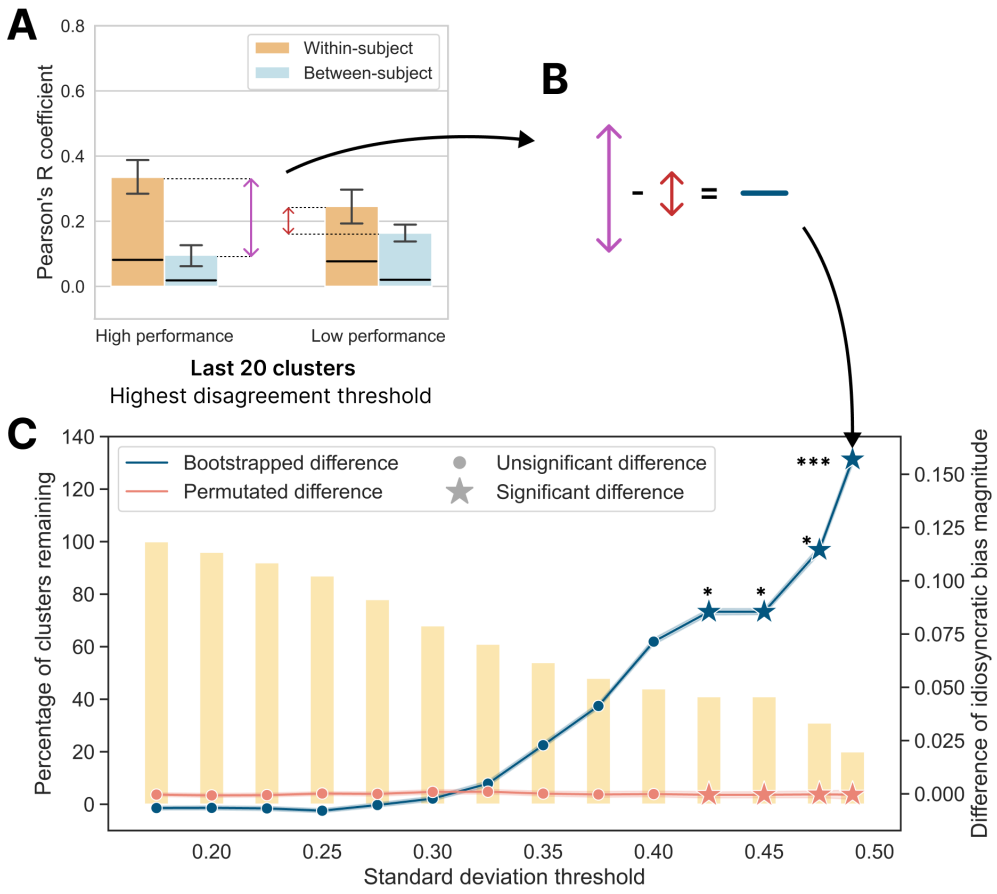


Figure 7.6: Idiosyncratic bias magnitude difference between the two groups with respect to cluster standard deviation thresholds (participant disagreement thresholds). Using the remaining clusters, we computed the idiosyncratic bias magnitude difference. **(A)** Given one subset of clusters (i.e. disagreement threshold) we measured the difference between within-participant correlation and between-participant correlation (idiosyncratic bias magnitude) for each group. **(B)** We then computed the difference of magnitude between the two performance group. Note that **(A)** is the same figure as Fig. 7.5B with a different y-axis range. **(C)** We repeated this procedure for increasing disagreement thresholds. The bootstrapped idiosyncratic bias magnitude and the permutation test values are represented by the dark blue line and the pink line respectively. The yellow columns represent the percentage of remaining image clusters after apply thresholds. Star markers denote where the permutation test is statistically significant, i.e., when the difference between the blue line and pink line is significant. Asterisks represent Bonferroni-adjusted p-value significance with $^{*}p < 0.05$ and $^{***}p < 0.001$.

lapsing less frequently. By comparing participant reaction times, we found that the time taken to submit diagnoses between the two groups was comparable and not significantly different (t-test, $p > 0.05$). Hence, it is unlikely that the stronger individual differences within the high-performers simply arose due to difference of attentiveness. Furthermore, while higher levels of attention could account for the higher within- and between-correlations of high-performers in some settings (Fig 7.5A), it may not explain altogether the increased difference of self-consistency and group-agreement displayed by skilled observers.

One might be concerned about the internal consistency of these idiosyncratic biases, that these biases are not systematic. Using the split-half Pearson's correlation, participants had a significant internal reliability of 0.68 (permutation test, $p < 0.001$). When measuring the internal reliability of each group, we found that high-performers reached a correlation coefficient of 0.77 (permutation test, $p < 0.001$) and the low-performers 0.58 (permutation test, $p < 0.001$). Furthermore, we measured a Cronbach's alpha of 0.95, underscoring the high internal consistency of participants' answers.

Leveraging the fact that images were sometimes diagnosed multiple times by one participant, we evaluated participants' test-retest reliability. That is, we assessed whether participants reported similar diagnoses when assessing the same image at different times. We found that participants showed significant reliability (permutation tests, $p < 0.001$), with Pearson's r coefficients of 0.44 for all participants, 0.46 when considering only the high-performance group, and 0.40 for low-performers.

One may worry that the skin lesion images utilized in the experiment contained only two types of lesion, i.e., nevus (benign) and melanoma (malignant). This does not capture the full range or variety of various skin cancer disease types. Moreover, for the images presented to participants, 57.3% were benign, and 42.7% were malignant. Although the skin cancer types and prevalence of the skin lesion images utilized in the experiment are somewhat different from typical scenarios, the skin lesion images are extracted directly from real diagnostic records. They contain a variety of different lesions, and textures, and include different skin cancer subtypes. This diversity allows the computer vision model to cover more territory. Thus, the size and scope of the dataset is a strength, which helps to reflect the biases that clinicians may actually have in routine store-and-forward diagnosis. Future studies can expand the lesion categories and investigate the effect of more lesion types in typical skin cancer diagnostic scenarios. Whether disease prevalence may influence individual differences is another interesting question to investigate in future studies.

All the data in this study were collected online because our goal was to investigate remote store-and-forward teledermatology. Whether in-person dermatologists might exhibit the same sorts of idiosyncratic biases as a function of proficiency remains unclear and should be investigated in future work. Because teledermatology has attracted a great deal of attention and increased in popularity recently, our results are valuable in explaining diagnostic errors and understanding the relationship between bias and proficiency in remote medicine. However, we acknowledge that the results here should not be extrapolated to in-person settings because telemedicine is not directly comparable to in-person diagnosis. For example, in the clinic, dermatologists have access to much more information, including tactile cues, a

larger field of view, and a variety of other sources of information.

Another question that might arise is whether time pressure or constraints may have imposed a burden on observers that led to the biases. This is not a likely explanation because participants had unlimited time to respond to each image. So the biases are not due to time pressure or a speed-accuracy trade-off.

The findings reported here hint at several promising avenues to improve remote store-and-forward diagnostic performance. Our visualization directly revealed the idiosyncratic "fingerprints" of perceptual bias in medical image judgments. Taking into account these fingerprints could improve individual clinician training by selectively shaping and mitigating the idiosyncratic biases. Existing training programs and approaches [256, 102, 110, 104, 192, 54, 89], which treat observers as having uniform biases, or as a uniform group, cannot address the heterogeneity of the biases. Highlighting this, the observers here received feedback about their performance on every trial, and yet that feedback did not mitigate the biases. Only by characterizing the individual clinician-level fingerprint of biases can observer-specific training programs be developed that might correct these biases.

The idiosyncratic perceptual biases reported here also have very important implications for multiple reader approaches in medical image diagnosis. Multi-reader improvement hinges on independence between clinicians, but our results here demonstrate that what counts as independent requires knowing very precisely the idiosyncratic image-level biases of individual clinicians. Knowing these fingerprints for individual observers means that independent observers could be strategically paired, thus improving performance above and beyond even the most skilled individual observer.

7.5 Conclusions

In summary, we found that medical trainees have image-level idiosyncratic biases when they perform skin cancer diagnosis, and increased diagnostic proficiency is associated with more substantial idiosyncratic biases. Isolating these fingerprints of perception could be valuable in the future to improve individualized training, computer assisted diagnosis, paired-reader approaches, and bias mitigation strategies.

Appendix A. Additional performance metrics

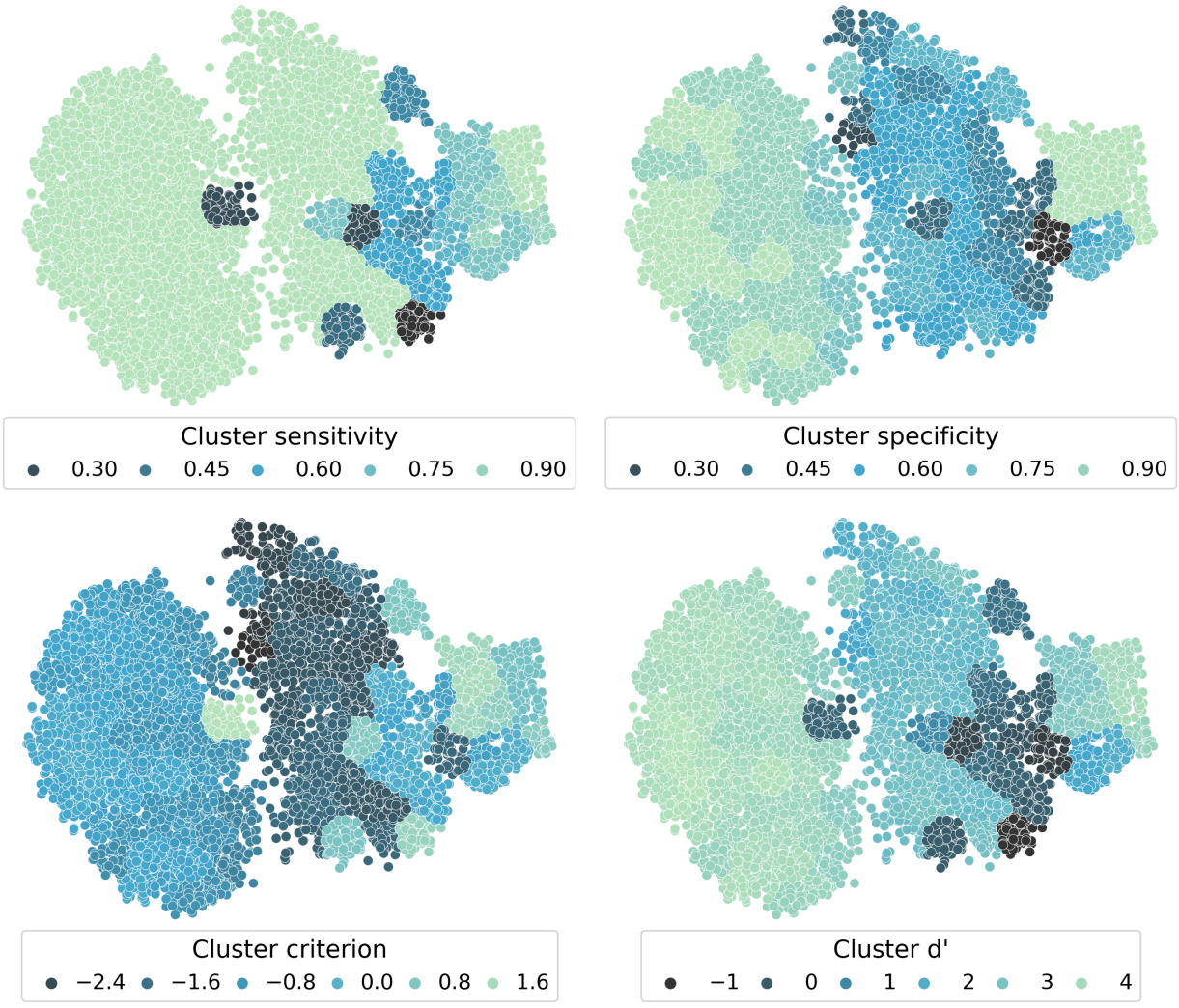


Figure 7.7: Cluster evaluation metrics averaged across all participants. Each dot represents a skin lesion image. Colors encode diagnostic metrics evaluated at the cluster-level, when considering all participants' diagnoses.

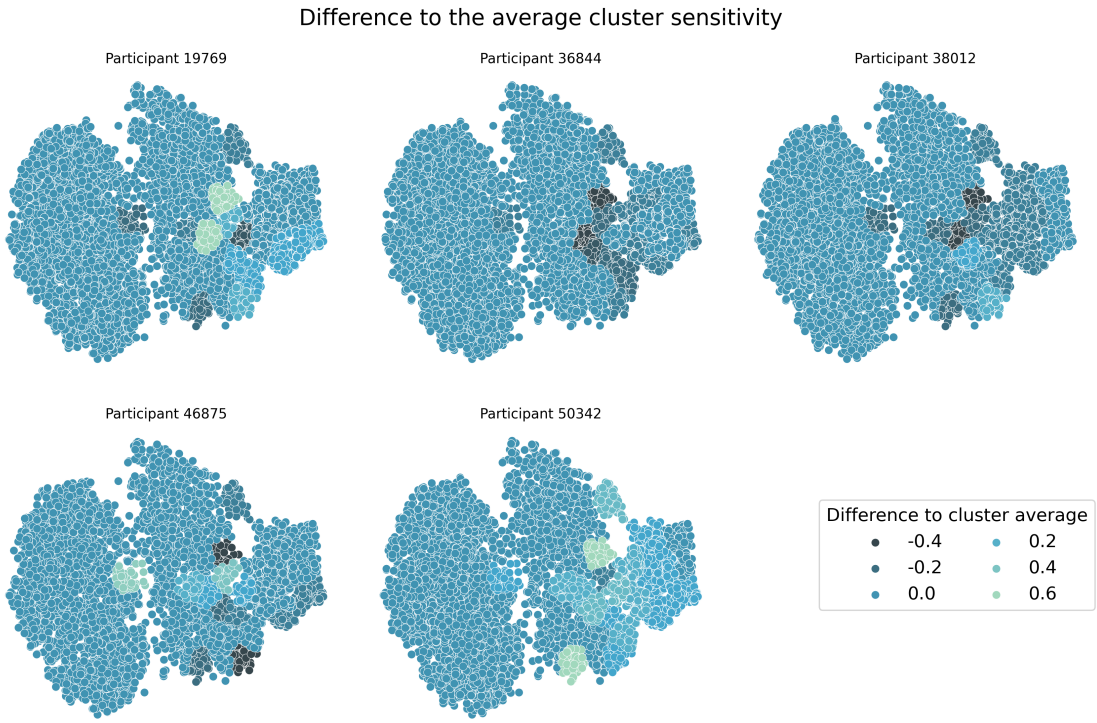


Figure 7.8: Diagnostic sensitivity per cluster for 5 participants compared to the cluster average across all participants. Because we interested in individual differences, after computing diagnostic metrics for one participant at the cluster-level, we compute the difference between this participant and the average of all participants within each cluster.

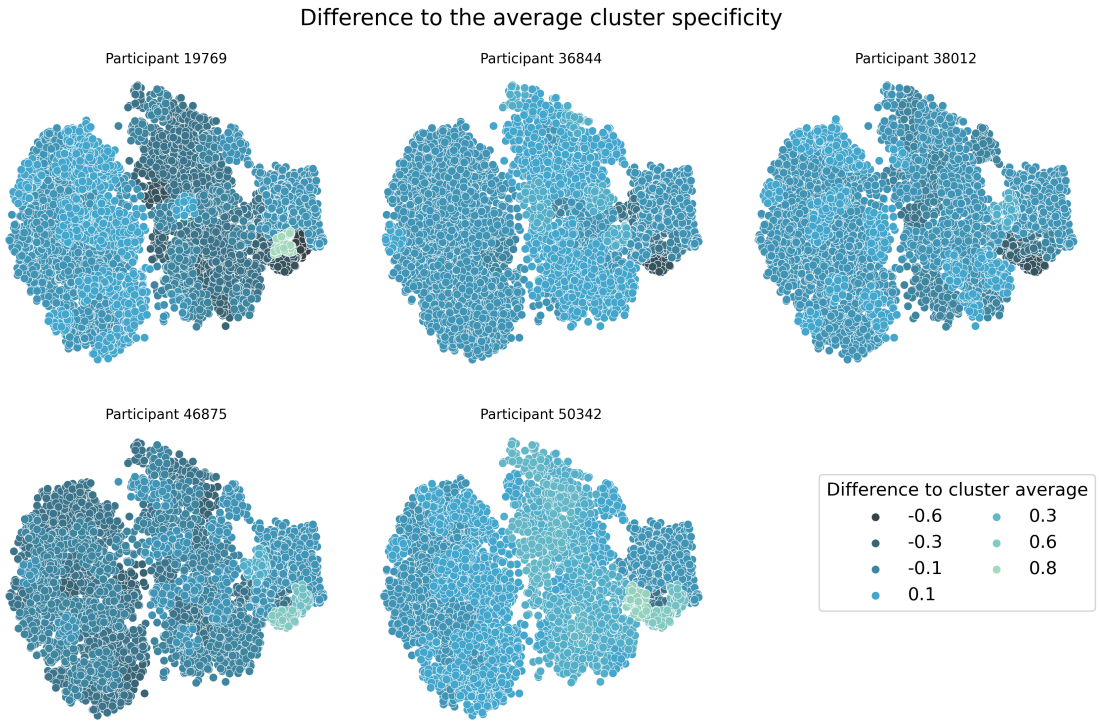


Figure 7.9: Relative diagnostic specificity per cluster for 5 participants

Figure 7.10: Relative diagnostic d' per cluster for 5 participants

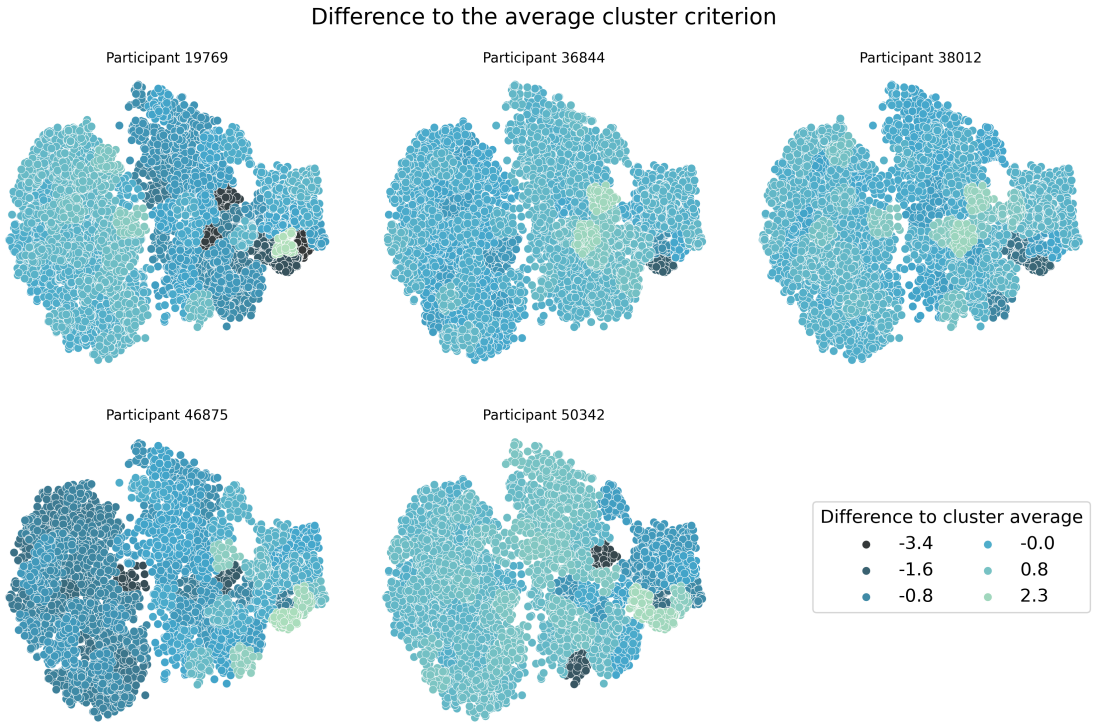


Figure 7.11: Relative diagnostic criterion per cluster for 5 participants

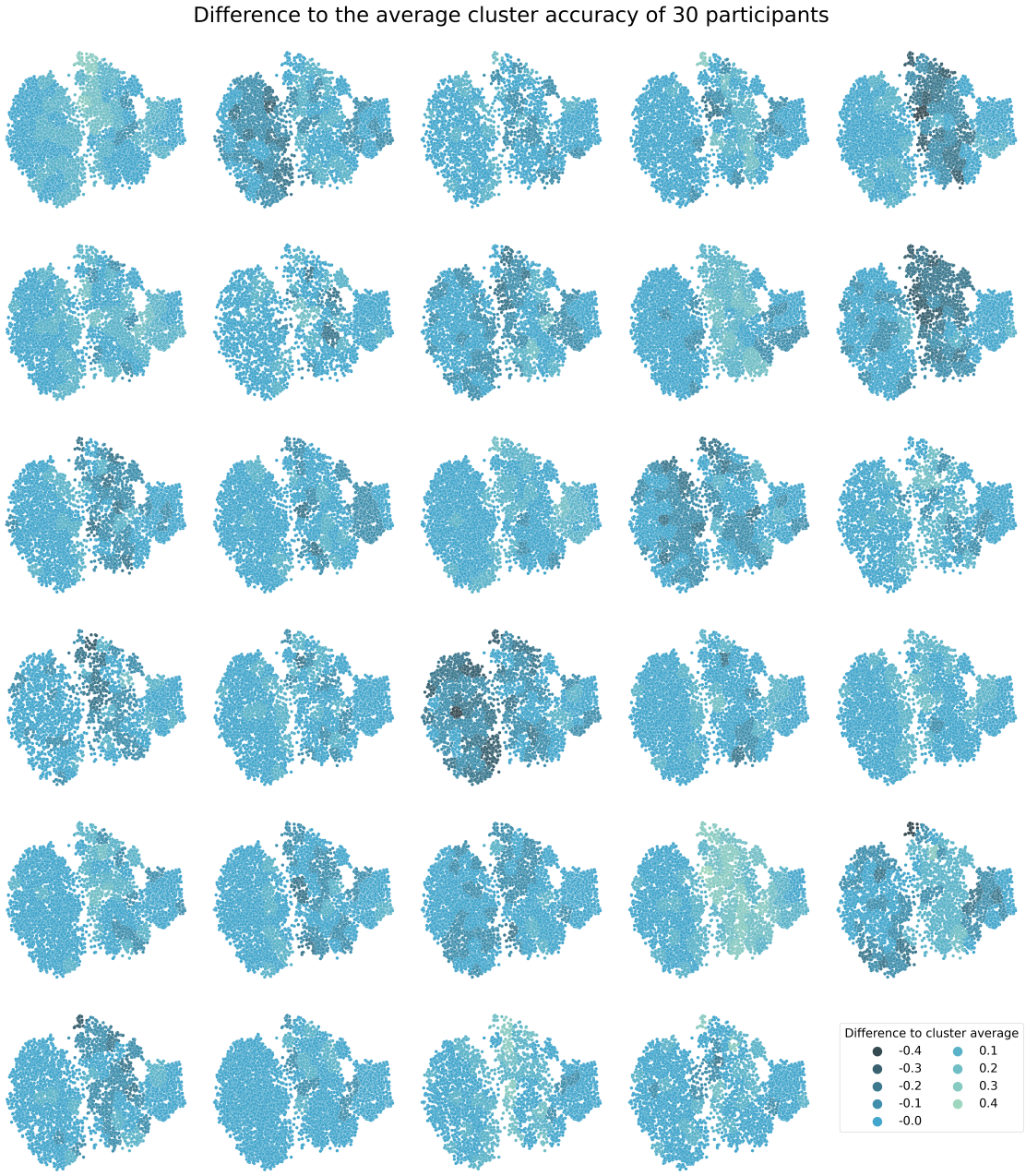


Figure 7.12: Relative diagnostic accuracy per cluster for 30 participants. Each dot represents a skin lesion image. Given that not all participants submitted a diagnosis for each image, some fingerprints contain fewer images (dots) than others. The accuracy (color) is computed at the cluster level.

Chapter 8

Conclusion

In the research discussed in this dissertation, we extended the visual serial dependence study in the medical image perception area. Chapter 2 illustrates our initial attempt to verify the existence of the visual serial dependence effect in medical image diagnosis via naive artificial medical image stimuli. We found that visual serial dependence has a disruptive effect in radiologic searches that impairs accurate detection and recognition of tumors or other structures. However, due to the limitation of the stimuli generation methods, the naive artificial stimuli have been noted by both untrained observers and expert radiologists to be less authentic, which can not help to reveal the real scenarios of medical image perception. To solve this obstacle, we proposed and built a generative tool via generative adversarial networks (GANs) to generate authentic medical images, replacing the simple stimuli in future experiments. The GenAI tool is introduced in Chapter 3. We also elaborate on another usage of our proposed GenAI tool in Chapter 4, i.e., to augment the rare case image samples in skin cancer diagnosis and boost the classification performance of self-supervised learning models. Using the authentic medical images from the GenAI medical image generation tool, in Chapter 5, we find that the perception of the current simulated medical image was biased towards the previously seen medical images, which strengthens the evidence of the existence of the visual serial dependence effect in medical image perception. Finally, in Chapter 6, we analyzed real diagnostic data collaboratively collected with a data annotation company. We found significant serial dependence effects in perceptual discrimination judgments, which negatively impacted performance measures. We further visualized new findings of the same diagnostic data in Chapter 7. In particular, we found that medical trainees have image-level idiosyncratic biases when they perform skin cancer diagnosis, and increased diagnostic proficiency is associated with more substantial idiosyncratic biases.

In closing, the research discussed in this dissertation reveals that visual serial dependence exists in medical image perception. These findings help understand one potential source of systematic bias and errors in medical image perception tasks and hint at useful approaches that could alleviate the errors due to serial dependence. The research also sheds light on the interdisciplinary area of human vision and computer vision, showing potential success in combining both fields.

Bibliography

[1] Arman Abrahamyan et al. "Adaptable history biases in human perceptual decisions". In: *Proceedings of the National Academy of Sciences* 113.25 (2016), E3548–E3557.

[2] David Alais, Johahn Leung, and Erik Van der Burg. "Linear summation of repulsive and attractive serial dependencies: Orientation and motion dependencies sum in motion perception". In: *Journal of Neuroscience* 37.16 (2017), pp. 4381–4390.

[3] Seyed Ali Amirshahi, Marius Pedersen, and Stella X Yu. "Image quality assessment by comparing CNN features between images". In: *Journal of Imaging Science and Technology* 60.6 (2016), pp. 60410–1.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.

[5] Karim Armanious et al. "Unsupervised medical image translation using `Cycle-MedGAN`". In: *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE. 2019, pp. 1–5.

[6] Carol J Ashman, Joseph S Yu, and Darcy Wolfman. "Satisfaction of search in osteoradiology". In: *American Journal of Roentgenology* 175.2 (2000), pp. 541–544.

[7] Shekoofeh Azizi et al. "Big self-supervised models advance medical image classification". In: *arXiv preprint arXiv:2101.05224* (2021).

[8] Guha Balakrishnan et al. "An unsupervised learning model for deformable medical image registration". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9252–9260.

[9] João Barbosa and Albert Compte. "Build-up of serial dependence in color working memory". In: *Scientific reports* 10.1 (2020), pp. 1–7.

[10] David Bau et al. "Seeing what a gan cannot generate". In: *Proceedings of the `IEEE/CVF` International Conference on Computer Vision*. 2019, pp. 4502–4511.

[11] Craig A Beam, Peter M Layde, and Daniel C Sullivan. "Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample". In: *Archives of internal medicine* 156.2 (1996), pp. 209–213.

[12] Kevin S Berbaum and Edmund A Franken Jr. "Satisfaction of search in radiographic modalities". In: *Radiology* 261.3 (2011), pp. 1000–1001.

[13] Wendie A Berg et al. "Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography?" In: *Radiology* 224.3 (2002), pp. 871–880.

[14] Leonard Berlin et al. "Accuracy of diagnostic procedures: has it improved over the past five decades". In: *AJR Am J Roentgenol* 188.5 (2007), pp. 1173–1178.

[15] D Birchall. "Spatial ability in radiologists: a necessary prerequisite?" In: *The British journal of radiology* 88.1049 (2015), p. 20140511.

[16] Robyn L Birdwell et al. "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection". In: *Radiology* 219.1 (2001), pp. 192–202.

[17] Michael H Birnbaum. "Human research and data collection via the Internet". In: *Annu. Rev. Psychol.* 55 (2004), pp. 803–832.

[18] Alceu Bissoto, Eduardo Valle, and Sandra Avila. "GAN-Based Data Augmentation and Anonymization for Skin-Lesion Analysis: A Critical Review". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1847–1856.

[19] Alceu Bissoto et al. "Skin lesion synthesis with generative adversarial networks". In: *OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis*. Springer, 2018, pp. 294–302.

[20] Daniel P Bliss, Jerome J Sun, and Mark D'Esposito. "Serial dependence is absent at the time of perception but increases in visual working memory". In: *Scientific reports* 7.1 (2017), pp. 1–13.

[21] Anna K Bobak, Peter JB Hancock, and Sarah Bate. "Super-recognisers in action: Evidence from face-matching and face memory tasks". In: *Applied Cognitive Psychology* 30.1 (2016), pp. 81–91.

[22] Sayedali Shetab Boushehri et al. "Annotation-efficient classification combining active learning, pre-training and semi-supervised learning for biomedical images". In: *bioRxiv* (2020).

[23] K Bowyer et al. "The digital database for screening mammography". In: *Third international workshop on digital mammography*. Vol. 58. 1996, p. 27.

[24] B Boyer et al. "Retrospectively detectable carcinomas: review of the literature". In: *Journal de radiologie* 85.12 Pt 2 (2004), pp. 2071–2078.

[25] Patrick C Brennan et al. "Radiologists can detect the 'gist'of breast cancer before any overt signs of cancer appear". In: *Scientific reports* 8.1 (2018), pp. 1–12.

[26] Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large scale GAN training for high fidelity natural image synthesis". In: *arXiv preprint arXiv:1809.11096* (2018).

[27] Michael A Bruno, Eric A Walker, and Hani H Abujudeh. "Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction". In: *Radiographics* 35.6 (2015), pp. 1668–1676.

[28] Teresa Canas-Bajo and David Whitney. "Stimulus-specific individual differences in holistic perception of Mooney faces". In: *Frontiers in Psychology* 11 (2020), p. 585921.

[29] Bing Cao et al. "Auto-GAN: self-supervised collaborative learning for medical image synthesis". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 10486–10493.

[30] Dennis P Carmody, Calvin F Nodine, and Harold L Kundel. "An analysis of perceptual and cognitive factors in radiographic interpretation". In: *Perception* 9.3 (1980), pp. 339–344.

[31] Dennis P Carmody, Calvin F Nodine, and Harold L Kundel. "Finding lung nodules with and without comparative visual scanning". In: *Perception & psychophysics* 29.6 (1981), pp. 594–598.

[32] Mathilde Caron et al. "Unsupervised learning of visual features by contrasting cluster assignments". In: *arXiv preprint arXiv:2006.09882* (2020).

[33] Tiago M de Carvalho et al. "Development of smartphone apps for skin cancer risk assessment: progress and promise". In: *JMIR Dermatology* 2.1 (2019), e13376.

[34] Ting Chen et al. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[35] Xi Chen et al. "InfoGAN: interpretable representation learning by information maximizing generative adversarial nets". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016, pp. 2180–2188.

[36] Adrien Chopin and Pascal Mamassian. "Predictive properties of visual adaptation". In: *Current biology* 22.7 (2012), pp. 622–626.

[37] Naomi Chuchu et al. "Teledermatology for diagnosing skin cancer in adults". In: *Cochrane Database of Systematic Reviews* 12 (2018).

[38] Domenic V Cicchetti. "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology." In: *Psychological assessment* 6.4 (1994), p. 284.

[39] Guido Marco Cicchini, Giovanni Anobile, and David C Burr. "Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform". In: *Proceedings of the National Academy of Sciences* 111.21 (2014), pp. 7867–7872.

[40] Guido Marco Cicchini, Alessandro Benedetto, and David C Burr. "Perceptual history propagates down to early levels of sensory analysis". In: *Current Biology* 31.6 (2021), pp. 1245–1250.

[41] Guido Marco Cicchini, Kyriaki Mikellidou, and David Burr. "Serial dependencies act directly on perception". In: *Journal of vision* 17.14 (2017), pp. 6–6.

[42] Guido Marco Cicchini, Kyriaki Mikellidou, and David C Burr. "The functional role of serial dependence". In: *Proceedings of the Royal Society B* 285.1890 (2018), p. 20181722.

[43] Noel CF Codella et al. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)". In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 168–172.

[44] Thérèse Collins. "The perceptual continuity field is retinotopic". In: *Scientific Reports* 9.1 (2019), pp. 1–6.

[45] Marc Combalia et al. "BCN20000: Dermoscopic lesions in the wild". In: *arXiv preprint arXiv:1908.02288* (2019).

[46] Jennifer E Corbett, Jason Fischer, and David Whitney. "Facilitating stable representations: Serial dependence in vision". In: *PLoS One* 6.1 (2011).

[47] CA Corry. "The future of recruitment and selection in radiology. Is there a role for assessment of basic visuospatial skills?" In: *Clinical radiology* 66.5 (2011), pp. 481–483.

[48] Aline F Cretenoud et al. "Individual differences in the perception of visual illusions are stable across eyes, time, and measurement methods". In: *Journal of Vision* 21.5 (2021), pp. 26–26.

[49] Aline F. Cretenoud et al. "Individual differences in the Müller-Lyer and Ponzo illusions are stable across different contexts". In: *Journal of Vision* 20.6 (June 2020), pp. 4–4. ISSN: 1534-7362. DOI: 10.1167/jov.20.6.4. eprint: https://arvojournals.org/arvo/content_public/journal/jov/938476/i0035-8711-453-1-07044.pdf. URL: https://doi.org/10.1167/jov.20.6.4.

[50] Pat Croskerry. "The importance of cognitive errors in diagnosis and strategies to minimize them". In: *Academic medicine* 78.8 (2003), pp. 775–780.

[51] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. "Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research". In: *PloS one* 8.3 (2013), e57410.

[52] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[53] Yadolah Dodge. *The concise encyclopedia of statistics*. Springer Science & Business Media, 2008.

[54] Barbara Dosher and Zhong-Lin Lu. "Visual perceptual learning and models". In: *Annual review of vision science* 3 (2017), pp. 343–363.

[55] Trafton Drew et al. "Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images?" In: *Radiographics* 33.1 (2013), pp. 263–274.

[56] Brad Duchaine and Ken Nakayama. "The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants". In: *Neuropsychologia* 44.4 (2006), pp. 576–585.

[57] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. "A learned representation for artistic style". In: *arXiv preprint arXiv:1610.07629* (2016).

[58] Meyer Dwass. "Modified randomization tests for nonparametric hypotheses". In: *The Annals of Mathematical Statistics* (1957), pp. 181–187.

[59] Eugene Edgington and Patrick Onghena. *Randomization tests*. CRC press, 2007.

[60] DJ Eedy and R Wootton. "Teledermatology: a review". In: *British Journal of Dermatology* 144.4 (2001), pp. 696–707.

[61] Bradley Efron and Robert Tibshirani. "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy". In: *Statistical science* (1986), pp. 54–75.

[62] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[63] Joann G Elmore, Carolyn K Wells, and Debra H Howard. "Does diagnostic accuracy in mammography depend on radiologists' experience?" In: *Journal of Women's Health* 7.4 (1998), pp. 443–449.

[64] Joann G Elmore et al. "Screening mammograms by community radiologists: variability in false-positive rates". In: *Journal of the National Cancer Institute* 94.18 (2002), pp. 1373–1380.

[65] Joann G Elmore et al. "Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy". In: *Radiology* 253.3 (2009), pp. 641–651.

[66] Joann G Elmore et al. "Variability in radiologists' interpretations of mammograms". In: *New England Journal of Medicine* 331.22 (1994), pp. 1493–1499.

[67] Kara Emery et al. "Color vs. motion: decoding perceptual representations from individual differences". In: *Journal of Vision* 19.8 (2019), pp. 8–8.

[68] Karla K Evans, Robyn L Birdwell, and Jeremy M Wolfe. "If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening". In: *PloS one* 8.5 (2013), e64366.

[69] Karla K Evans et al. "A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast". In: *Proceedings of the National Academy of Sciences* 113.37 (2016), pp. 10292–10297.

[70] Karla K Evans et al. "The gist of the abnormal: Above-chance medical decision making in the blink of an eye". In: *Psychonomic bulletin & review* 20 (2013), pp. 1170–1175.

[71] Li Fei-Fei, Rob Fergus, and Pietro Perona. "One-shot learning of object categories". In: *IEEE transactions on pattern analysis and machine intelligence* 28.4 (2006), pp. 594–611.

[72] Judith Feldman et al. "Peer review of mammography interpretations in a breast cancer screening program." In: *American journal of public health* 85.6 (1995), pp. 837–839.

[73] Samuel W Fernberger. "Interdependence of judgments within the series for the method of constant stimuli." In: *Journal of Experimental Psychology* 3.2 (1920), p. 126.

[74] Anna Finnane et al. "Teledermatology for the diagnosis and management of skin cancer: a systematic review". In: *JAMA dermatology* 153.3 (2017), pp. 319–327.

[75] Jason Fischer and David Whitney. "Serial dependence in visual perception". In: *Nature neuroscience* 17.5 (2014), pp. 738–743.

[76] Joseph L Fleiss. *Design and analysis of clinical experiments*. Vol. 73. John Wiley & Sons, 2011.

[77] Joseph L Fleiss. "Reliability of measurement". In: *The design and analysis of clinical experiments* (1986).

[78] Michele Fornaciai and Joonkoo Park. "Attractive serial dependence between memorized stimuli". In: *Cognition* 200 (2020), p. 104250.

[79] Michele Fornaciai and Joonkoo Park. "Serial dependence in numerosity perception". In: *Journal of Vision* 18.9 (2018), pp. 15–15.

[80] Michele Fornaciai and Joonkoo Park. "Spontaneous repulsive adaptation in the absence of attractive serial dependence". In: *Journal of vision* 19.5 (2019), pp. 21–21.

[81] Mohammad Fraiwan and Esraa Faouri. "On the Automatic Detection and Classification of Skin Cancer Using Deep Transfer Learning". In: *Sensors* 22.13 (2022), p. 4963.

[82] Maayan Frid-Adar et al. "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification". In: *Neurocomputing* 321 (2018), pp. 321–331.

[83] Matthias Fritsche and Floris P de Lange. "The role of feature-based attention in visual serial dependence". In: *Journal of Vision* 19.13 (2019), pp. 21–21.

[84] Matthias Fritsche, Pim Mostert, and Floris P de Lange. "Opposite effects of recent history on perception and decision". In: *Current Biology* 27.4 (2017), pp. 590–595.

[85] Ingo Fründ, Felix A Wichmann, and Jakob H Macke. "Quantifying the effect of intertrial dependence on perceptual decisions". In: *Journal of vision* 14.7 (2014), pp. 9–9.

[86] Kunihiko Fukushima. "Neural network model for a mechanism of pattern recognition unaffected by shift in position-Neocognitron". In: *IEICE Technical Report, A* 62.10 (1979), pp. 658–665.

[87] Brian Funaki, George X Szymski, and Jordan D Rosenblum. "Significant on-call misses by radiology residents interpreting computed tomographic studies: perception versus cognition". In: *Emergency Radiology* 4 (1997), pp. 290–294.

[88] AG Gale et al. "Reporting in a flash". In: *Br J Radiol* 63.S (1990), p. 71.

[89] Roberto Gammeri et al. "Effects of prism adaptation and visual scanning training on perceptual and response bias in unilateral spatial neglect". In: *Neuropsychological Rehabilitation* (2023), pp. 1–26.

[90] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. "Texture synthesis using convolutional neural networks". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*. 2015, pp. 262–270.

[91] Amirata Ghorbani et al. "Dermgan: Synthetic generation of clinical skin images with pathology". In: *Machine Learning for Health Workshop*. PMLR. 2020, pp. 155–170.

[92] Ross Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[93] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[94] Alexander Goettker and Emma EM Stewart. "Serial dependence for oculomotor control depends on early sensory signals". In: *Current Biology* 32.13 (2022), pp. 2956–2961.

[95] Ian J Goodfellow et al. "Generative adversarial networks". In: *arXivpreprintarXiv:1406.2661* (2014).

[96] Jean-Bastien Grill et al. "Bootstrap your own latent: A new approach to self-supervised learning". In: *arXiv preprint arXiv:2006.07733* (2020).

[97] Lukasz Grzeczkowski et al. "About individual differences in vision". In: *Vision research* 141 (2017), pp. 282–292.

[98] Ishaan Gulrajani et al. "Improved training of wasserstein gans". In: *arXiv preprint arXiv:1704.00028* (2017).

[99] Gery P Guy Jr et al. "Prevalence and costs of skin cancer treatment in the US, 2002- 2006 and 2007- 2011". In: *American journal of preventive medicine* 48.2 (2015), pp. 183–187.

[100] Gery P Guy Jr et al. "Vital signs: melanoma incidence and mortality trends and projections—United States, 1982–2030". In: *MMWR. Morbidity and mortality weekly report* 64.21 (2015), p. 591.

[101] Qishen Ha, Bo Liu, and Fuxu Liu. "Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge". In: *arXiv preprint arXiv:2010.05351* (2020).

[102] Qi Haijiang et al. "Demonstration of cue recruitment: Change in visual appearance by means of Pavlovian conditioning". In: *Proceedings of the National Academy of Sciences* 103.2 (2006), pp. 483–488.

[103] Changhee Han et al. "GAN-based synthetic brain MR image generation". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 734–738.

[104] Sarah J Harrison, Benjamin T Backus, and Anshul Jain. "Disambiguation of Necker cube rotation by monocular and binocular depth cues: Relative effectiveness for establishing long-term bias". In: *Vision research* 51.9 (2011), pp. 978–986.

[105] Jennifer A Harvey, Laurie L Fajardo, and CA8249720 Innis. "Previous mammograms in patients with impalpable breast carcinoma: retrospective vs blinded interpretation. 1993 ARRS President's Award." In: *AJR. American journal of roentgenology* 161.6 (1993), pp. 1167–1172.

[106] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[107] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[108] Kaiming He et al. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9729–9738.

[109] US Department of Health, Human Services, et al. "The Surgeon General's call to action to prevent skin cancer". In: (2014).

[110] Michael H Herzog et al. "Reverse feedback induces position and orientation specific changes". In: *Vision research* 46.22 (2006), pp. 3761–3770.

[111] Mohammad Hesam Hesamian et al. "Deep learning techniques for medical image segmentation: achievements and challenges". In: *Journal of digital imaging* 32.4 (2019), pp. 582–596.

[112] Nathan Hilliard et al. "Few-shot learning with metric-agnostic conditional embeddings". In: *arXiv preprint arXiv:1802.04376* (2018).

[113] R Devon Hjelm et al. "Learning deep representations by mutual information estimation and maximization". In: *arXiv preprint arXiv:1808.06670* (2018).

[114] Todd S Horowitz. "Prevalence in visual search: From the clinic to the lab and back again". In: *Japanese Psychological Research* 59.2 (2017), pp. 65–108.

[115] Khalid M Hosny, Mohamed A Kassem, and Mohamed M Foaud. "Skin cancer classification using deep learning and transfer learning". In: *2018 9th Cairo international biomedical engineering conference (CIBEC)*. IEEE. 2018, pp. 90–93.

[116] Joseph P Houghton et al. "Diagnostic performance on briefly presented digital pathology images". In: *Journal of pathology informatics* 6 (2015).

[117] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[118] Xun Huang and Serge Belongie. "Arbitrary style transfer in real-time with adaptive instance normalization". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1501–1510.

[119] Xun Huang et al. "Multimodal unsupervised image-to-image translation". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 172–189.

[120] DH Hubel and TN Wiesel. "Receptive fields of single neurones in the cat's striate cortex". In: *The Journal of Physiology* 148.3 (1959), p. 574.

[121] Fortune Business Insights. *Teledermatology Market Size, Share & COVID-19 Impact Analysis, and Regional Forecast 2021-2028*. Online report. 2020. URL: `https://www.fortunebusinessinsights.com/teledermatology-market-103491`.

[122] Thomas Jaarsma et al. "Expertise under the microscope: Processing histopathological slides". In: *Medical education* 48.3 (2014), pp. 292–300.

[123] Mohammad H Jafari et al. "Skin lesion segmentation in clinical images using deep learning". In: *2016 23rd International conference on pattern recognition (ICPR)*. IEEE. 2016, pp. 337–342.

[124] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". In: *European conference on computer vision*. Springer. 2016, pp. 694–711.

[125] Ryota Kanai and Geraint Rees. "The structural basis of inter-individual differences in human behaviour and cognition". In: *Nature Reviews Neuroscience* 12.4 (2011), pp. 231–242.

[126] Sae Kaneko et al. "Individual variability in simultaneous contrast for color and brightness: Small sample factor analyses reveal separate induction processes for short and long flashes". In: *i-Perception* 9.5 (2018), p. 2041669518800507.

[127] Tero Karras, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4401–4410.

[128] Tero Karras et al. "Analyzing and improving the image quality of stylegan". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8110–8119.

[129] Tero Karras et al. "Progressive Growing of GANs for Improved Quality, Stability, and Variation". In: *International Conference on Learning Representations*. 2018.

[130] Tero Karras et al. "Progressive growing of gans for improved quality, stability, and variation". In: *arXiv preprint arXiv:1710.10196* (2017).

[131] Mohamed A Kassem, Khalid M Hosny, and Mohamed M Fouad. "Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning". In: *IEEE Access* 8 (2020), pp. 114822–114832.

[132] Baris Kayalibay, Grady Jensen, and Patrick van der Smagt. "CNN-based segmentation of medical imaging data". In: *arXiv preprint arXiv:1701.03056* (2017).

[133] Sujin Kim et al. "Serial dependence in perception requires conscious awareness". In: *Current Biology* 30.6 (2020), R257–R258.

[134] Young W Kim and Liem T Mansfield. "Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors". In: *American journal of roentgenology* 202.3 (2014), pp. 465–470.

[135] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[136] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *stat* 1050 (2014), p. 1.

[137] Anastasia Kiyonaga et al. "Serial dependence across perception, attention, and memory". In: *Trends in Cognitive Sciences* 21.7 (2017), pp. 493–497.

[138] Elysse Kompaniez et al. "Adaptation aftereffects in the perception of radiological images". In: *PloS one* 8.10 (2013), e76175.

[139] Elysse Kompaniez-Dunigan et al. "Adaptation and visual search in mammographic images". In: *Attention, Perception, & Psychophysics* 77.4 (2015), pp. 1081–1087.

[140] Elysse Kompaniez-Dunigan et al. "Visual adaptation and the amplitude spectra of radiological images". In: *Cognitive research: principles and implications* 3.1 (2018), pp. 1–12.

[141] Aki Kondo, Kohske Takahashi, and Katsumi Watanabe. "Sequential effects in face-attractiveness judgment". In: *Perception* 41.1 (2012), pp. 43–49.

[142] Anna Kosovicheva and David Whitney. "Stable individual signatures in object localization". In: *Current Biology* 27.14 (2017), R700–R701.

[143] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009).

[144] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105.

[145] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[146] Elizabeth A Krupinski. "Current perspectives in medical image perception". In: *Attention, Perception, & Psychophysics* 72.5 (2010), pp. 1205–1217.

[147] Elizabeth A Krupinski. "Visual scanning patterns of radiologists searching mammograms". In: *Academic radiology* 3.2 (1996), pp. 137–144.

[148] Dongyang Kuang and Tanya Schmah. "Faim–a convnet method for unsupervised 3d medical image registration". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2019, pp. 646–654.

[149] Melina A Kunar et al. "Low prevalence search for cancers in mammograms: Evidence using laboratory experiments and computer aided detection." In: *Journal of Experimental Psychology: Applied* 23.4 (2017), p. 369.

[150] Harold L Kundel. "History of research in medical image perception". In: *Journal of the American college of radiology* 3.6 (2006), pp. 402–408.

[151] Harold L Kundel and Calvin F Nodine. "Interpreting chest radiographs without visual search". In: *Radiology* 116.3 (1975), pp. 527–532.

[152] Carson Lam et al. "Retinal lesion detection with deep learning using image patches". In: *Investigative ophthalmology & visual science* 59.1 (2018), pp. 590–596.

[153] Sonia A Lamel et al. "Application of mobile teledermatology for skin cancer screening". In: *Journal of the American Academy of Dermatology* 67.4 (2012), pp. 576–581.

[154] Jean Langlois et al. "Spatial abilities of medical graduates and choice of residency programs". In: *Anatomical Sciences Education* 8.2 (2015), pp. 111–119.

[155] Elizabeth Lazarus et al. "BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value". In: *Radiology* 239.2 (2006), pp. 385–391.

[156] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[157] Christian Ledig et al. "Photo-realistic single image super-resolution using a generative adversarial network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4681–4690.

[158] Cindy S Lee et al. "Cognitive and system factors contributing to diagnostic errors in radiology". In: *American Journal of Roentgenology* 201.3 (2013), pp. 611–617.

[159] Jonathan J Lee and Joseph C English. "Teledermatology: a review and update". In: *American journal of clinical dermatology* 19 (2018), pp. 253–260.

[160] Xiaomeng Li et al. "Self-supervised Feature Learning via Exploiting Multi-modal Data for Retinal Disease Diagnosis". In: *IEEE Transactions on Medical Imaging* 39.12 (2020), pp. 4023–4033.

[161] Alina Liberman, Jason Fischer, and David Whitney. "Serial dependence in the perception of faces". In: *Current biology* 24.21 (2014), pp. 2569–2574.

[162] Alina Liberman, Mauro Manassi, and David Whitney. "Serial dependence promotes the stability of perceived emotional expression depending on face similarity". In: *Attention, Perception, & Psychophysics* 80 (2018), pp. 1461–1473.

[163] Alina Liberman, Kathy Zhang, and David Whitney. "Serial dependence promotes object stability during occlusion". In: *Journal of vision* 16.15 (2016), pp. 16–16.

[164] Swee Kiat Lim et al. "Doping: Generative data augmentation for unsupervised anomaly detection with gan". In: *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2018, pp. 1122–1127.

[165] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[166] MN Linver et al. "Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases." In: *Radiology* 184.1 (1992), pp. 39–43.

[167] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[168] Stuart Lloyd. "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[169] Amirreza Mahbod et al. "Skin lesion classification using hybrid deep neural networks". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 1229–1233.

[170] Mauro Manassi, Árni Kristjánsson, and David Whitney. "Serial dependence determines object classification in visual search". In: *Journal of Vision* 17.10 (2017), pp. 221–221.

[171] Mauro Manassi, Árni Kristjánsson, and David Whitney. "Serial dependence in a simulated clinical visual search task". In: *Scientific reports* 9.1 (2019), pp. 1–10.

[172] Mauro Manassi and David Whitney. "Illusion of visual stability through active perceptual serial dependence". In: *Science advances* 8.2 (2022), eabk2480.

[173] Mauro Manassi et al. "Serial dependence in position occurs at the time of perception". In: *Psychonomic Bulletin & Review* 25.6 (2018), pp. 2245–2253.

[174] Mauro Manassi et al. "Serial dependence in the perceptual judgments of radiologists". In: *Cognitive research: principles and implications* 6.1 (2021), pp. 1–13.

[175] Mauro Manassi et al. "The perceived stability of scenes: serial dependence in ensemble representations". In: *Scientific reports* 7.1 (2017), pp. 1–9.

[176] David Manning et al. "How do radiologists do it? The influence of experience and training on searching for chest nodules". In: *Radiography* 12.2 (2006), pp. 134–142.

[177] Stefan Markun et al. "Mobile teledermatology for skin cancer screening: a diagnostic accuracy study". In: *Medicine* 96.10 (2017).

[178] Jonathan Masci et al. "Stacked convolutional auto-encoders for hierarchical feature extraction". In: *International conference on artificial neural networks*. Springer. 2011, pp. 52–59.

[179] Gerrit W Maus et al. "The challenge of measuring long-term positive aftereffects". In: *Current Biology* 23.10 (2013), R438–R439.

[180] Robert J McDonald et al. "The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload". In: *Academic radiology* 22.9 (2015), pp. 1191–1198.

[181] Tamaryn Menneer et al. "High or low target prevalence increases the dual-target cost in visual search." In: *Journal of Experimental Psychology: Applied* 16.2 (2010), p. 133.

[182] Jeff Miller. "Reaction time analysis with outlier exclusion: Bias varies with sample size". In: *The quarterly journal of experimental psychology* 43.4 (1991), pp. 907–912.

[183] Mehdi Mirza and Simon Osindero. "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784* (2014).

[184] Eduard Molins et al. "Association between radiologists' experience and accuracy in interpreting screening mammograms". In: *BMC health services research* 8.1 (2008), pp. 1–10.

[185] Pieter Moors et al. "Serial correlations in continuous flash suppression". In: *Neuroscience of Consciousness* 2015.1 (2015), niv010.

[186] Mark D Mugglestone et al. "Diagnostic performance on briefly presented mammographic images". In: *Medical Imaging 1995: Image Perception*. Vol. 2436. International Society for Optics and Photonics. 1995, pp. 106–115.

[187] Martijn J Mulder et al. "Bias in the brain: a diffusion model analysis of prior probability and potential payoff". In: *Journal of Neuroscience* 32.7 (2012), pp. 2335–2343.

[188] Tanya Nair et al. "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation". In: *Medical image analysis* 59 (2020), p. 101557.

[189] Ryoichi Nakashima et al. "Visual search of experts in medical image reading: the effect of training, target prevalence, and expert knowledge". In: *Frontiers in psychology* 4 (2013), p. 166.

[190] Heidi D Nelson et al. "Factors associated with rates of false-positive and false-negative results from digital mammography screening: an analysis of registry data". In: *Annals of internal medicine* 164.4 (2016), pp. 226–235.

[191] Dong Nie et al. "Medical image synthesis with context-aware generative adversarial networks". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2017, pp. 417–425.

[192] Andrey R Nikolaev, Sergei Gepshtein, and Cees van Leeuwen. "Intermittent regime of brain activity at the early, bias-guided stage of perceptual learning". In: *Journal of Vision* 16.14 (2016), pp. 11–11.

[193] Calvin F Nodine et al. "Nature of expertise in searching mammograms for breast masses". In: *Academic radiology* 3.12 (1996), pp. 1000–1006.

[194] Augustus Odena, Christopher Olah, and Jonathon Shlens. "Conditional image synthesis with auxiliary classifier gans". In: *International conference on machine learning*. PMLR. 2017, pp. 2642–2651.

[195] JW Oestmann et al. "Lung lesions: correlation between viewing time and detection." In: *Radiology* 166.2 (1988), pp. 451–453.

[196] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[197] Jacob L Orquin, Sonja Perkovic, and Klaus G Grunert. "Visual biases in decision making". In: *Applied Economic Perspectives and Policy* 40.4 (2018), pp. 523–537.

[198] Taesung Park et al. "Semantic image synthesis with spatially-adaptive normalization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2337–2346.

[199] D Pascucci et al. "Laws of concatenated perception: Vision goes for novelty". In: *Deci-sions for perseverance. bioRxiv* 15 (2017), p. 929.

[200] David Pascucci et al. "Laws of concatenated perception: Vision goes for novelty, decisions for perseverance". In: *PLoS biology* 17.3 (2019), e3000144.

[201] Paola Pasquali et al. "Teledermatology and its current perspective". In: *Indian dermatology online journal* 11.1 (2020), p. 12.

[202] Jonathan W Peirce. "Generating stimuli for neuroscience using PsychoPy". In: *Frontiers in neuroinformatics* 2 (2009), p. 343.

[203] Jonathan W Peirce. "PsychoPy—psychophysics software in Python". In: *Journal of neuroscience methods* 162.1-2 (2007), pp. 8–13.

[204] Douglas A Perednia and NA Brown. "Teledermatology: one application of telemedicine." In: *Bulletin of the Medical Library Association* 83.1 (1995), p. 42.

[205] Sara Perkins et al. "Teledermatology in the era of COVID-19: experience of an academic department of dermatology". In: *Journal of the American Academy of Dermatology* 83.1 (2020), e43–e44.

[206] Kyla N Price et al. "Strategic dermatology clinical operations during the coronavirus disease 2019 (COVID-19) pandemic". In: *Journal of the American Academy of Dermatology* 82.6 (2020), e207–e209.

[207] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *arXivpreprintarXiv:1511.06434* (2015).

[208] Mohsen Rafiei et al. "Optimizing perception: Attended and ignored stimuli create opposing perceptual biases". In: *Attention, Perception, & Psychophysics* 83 (2021), pp. 1230–1239.

[209] Sivananda Rajananda et al. "Visual psychophysics on the web: open-access tools, experiments, and results using online platforms". In: *Journal of Vision* 18.10 (2018), pp. 299–299.

[210] Marc'Aurelio Ranzato et al. "Unsupervised learning of invariant feature hierarchies with applications to object recognition". In: *2007 IEEE conference on computer vision and pattern recognition*. IEEE. 2007, pp. 1–8.

[211] Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[212] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *arXiv preprint arXiv:1506.01497* (2015).

[213] Zhihang Ren, X Yu Stella, and David Whitney. "Controllable Medical Image Generation via GAN". In: *Journal of Perceptual Imaging* 5 (2022), pp. 1–15.

[214] Zhihang Ren, Stella X. Yu, and David Whitney. "Controllable medical image generation via Generative Adversarial Networks". In: *Electronic Imaging* 2021 (2021), pp. 112-1–112-5. DOI: https://doi.org/10.2352/ISSN.2470-1173.2021.11.HVEI-112.

[215] Zhihang Ren et al. "Improve Image-based Skin Cancer Diagnosis with Generative Self-Supervised Learning". In: *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE. 2021, pp. 23–34.

[216] Zhihang Ren et al. "Serial dependence in perception across naturalistic GAN-generated mammograms". in revision.

[217] Precedence Research. *Teledermatology Market - Global Industry Analysis, Size, Share, Growth, Trends, Regional Outlook, and Forecast 2021 – 2030*. Online report. 2021. URL: https://www.precedenceresearch.com/teledermatology-market.

[218] Anina N Rich et al. "Why do we miss rare targets? Exploring the boundaries of the low prevalence effect". In: *Journal of vision* 8.15 (2008), pp. 15–15.

[219] Jennifer J Richler et al. "Individual differences in object recognition." In: *Psychological Review* 126.2 (2019), p. 226.

[220] Tony Rosen et al. "Radiologists' training, experience, and attitudes about elder abuse detection". In: *AJR. American journal of roentgenology* 207.6 (2016), p. 1210.

[221] Richard Russell, Garga Chatterjee, and Ken Nakayama. "Developmental prosopagnosia and super-recognition: No special role for surface reflectance processing". In: *Neuropsychologia* 50.2 (2012), pp. 334–340.

[222] Richard Russell, Brad Duchaine, and Ken Nakayama. "Super-recognizers: People with extraordinary face recognition ability". In: *Psychonomic bulletin & review* 16.2 (2009), pp. 252–257.

[223] Ehsan Samei and Elizabeth A Krupinski. *The handbook of medical image perception and techniques*. Cambridge University Press, 2018.

[224] Alexander C Schütz. "Interindividual differences in preferred directions of perceptual and motor decisions". In: *Journal of vision* 14.12 (2014), pp. 16–16.

[225] Kilian Semmelmann and Sarah Weigelt. "Online psychophysics: Reaction time effects in cognitive experiments". In: *Behavior Research Methods* 49.4 (2017), pp. 1241–1260.

[226] Dinggang Shen, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis". In: *Annual review of biomedical engineering* 19 (2017), pp. 221–248.

[227] Heather Sheridan and Eyal M Reingold. "The holistic processing account of visual expertise in medical image perception: A review". In: *Frontiers in psychology* 8 (2017), p. 1620.

[228] Hoo-Chang Shin et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1285–1298.

[229] Hoo-Chang Shin et al. "Medical image synthesis for data augmentation and anonymization using generative adversarial networks". In: *International workshop on simulation and synthesis in medical imaging*. Springer. 2018, pp. 1–11.

[230] Rebecca Siegel et al. "Cancer statistics, 2011: the impact of eliminating socioeconomic and racial disparities on premature cancer deaths." In: *CA: a cancer journal for clinicians* 61.4 (2011), pp. 212–236.

[231] Carlos Eduardo Goulart Silveira et al. "Digital photography in skin cancer screening by mobile units in remote areas of Brazil". In: *BMC dermatology* 14.1 (2014), pp. 1–5.

[232] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[233] WR Smoker et al. "Spatial perception testing in diagnostic radiology". In: *American journal of roentgenology* 143.5 (1984), pp. 1105–1109.

[234] Jake Snell, Kevin Swersky, and Richard S Zemel. "Prototypical networks for few-shot learning". In: *arXiv preprint arXiv:1703.05175* (2017).

[235] Yang Song et al. "Large margin local estimate with applications to medical image classification". In: *IEEE transactions on medical imaging* 34.6 (2015), pp. 1362–1377.

[236] Qianru Sun et al. "Meta-transfer learning for few-shot learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 403–412.

[237] Mackenzie A Sunday, Edwin Donnelly, and Isabel Gauthier. "Both fluid intelligence and visual object recognition ability relate to nodule detection in chest radiographs". In: *Applied Cognitive Psychology* 32.6 (2018), pp. 755–762.

[238] Mackenzie A Sunday, Edwin Donnelly, and Isabel Gauthier. "Individual differences in perceptual abilities in medical imaging: the Vanderbilt Chest Radiograph Test". In: *Cognitive Research: Principles and Implications* 2.1 (2017), pp. 1–10.

[239] Flood Sung et al. "Learning to compare: Relation network for few-shot learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1199–1208.

[240] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[241] Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[242] Alai Tan et al. "Variation in false-positive rates of mammography reading among 1067 radiologists: a population-based assessment". In: *Breast cancer research and treatment* 100 (2006), pp. 309–318.

[243] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114.

[244] Jessica Taubert and David Alais. "Serial dependence in face attractiveness judgements tolerates rotations around the yaw axis but not the roll axis". In: *Visual Cognition* 24.2 (2016), pp. 103–114.

[245] Jessica Taubert, David Alais, and David Burr. "Different coding strategies for the perception of stable and changeable facial attributes". In: *Scientific reports* 6.1 (2016), pp. 1–7.

[246] Jessica Taubert, Erik Van der Burg, and David Alais. "Love at second sight: Sequential dependence of facial attractiveness in an on-line dating paradigm". In: *Scientific reports* 6.1 (2016), pp. 1–5.

[247] E Tensen et al. "Two decades of teledermatology: current status and integration in national healthcare systems". In: *Current dermatology reports* 5 (2016), pp. 96–104.

[248] Steven P Tipper. "The negative priming effect: Inhibitory priming by ignored objects". In: *The quarterly journal of experimental psychology* 37.4 (1985), pp. 571–590.

[249] Jennifer S Trueblood et al. "Disentangling prevalence induced biases in medical image decision-making". In: *Cognition* 212 (2021), p. 104713.

[250] Jennifer S Trueblood et al. "The impact of speed and bias on the cognitive processes of experts and novices in medical image decision-making". In: *Cognitive Research: Principles and Implications* 3.1 (2018), pp. 1–14.

[251] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Nature Scientific Data* 5 (2018), p. 180161.

[252] Kaitlyn Turbett et al. "Individual differences in serial dependence of facial identity are associated with face recognition abilities". In: *Scientific reports* 9.1 (2019), pp. 1–12.

[253] Erik Van der Burg et al. "Serial dependence of emotion within and between stimulus sensory modalities". In: *Multisensory research* 35.2 (2021), pp. 151–172.

[254] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[255] Alessia Villani, Massimiliano Scalvenzi, and Gabriella Fabbrocini. "Teledermatology: a useful tool to fight COVID-19". In: *Journal of Dermatological Treatment* 31.4 (2020), pp. 325–325.

[256] Rufin Vogels and Guy A Orban. "The effect of practice on the oblique effect in line orientation judgments". In: *Vision research* 25.11 (1985), pp. 1679–1687.

[257] Robert F Wagner and David G Brown. "Unified SNR analysis of medical imaging systems". In: *Physics in Medicine & Biology* 30.6 (1985), p. 489.

[258] Abdul Waheed et al. "Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection". In: *Ieee Access* 8 (2020), pp. 91916–91923.

[259] Stephen Waite et al. "Analysis of perceptual expertise in radiology–Current knowledge and a new perspective". In: *Frontiers in human neuroscience* 13 (2019), p. 213.

[260] Stephen Waite et al. "Interpretive error in radiology". In: *American Journal of R oentgenology* 208.4 (2017), pp. 739–749.

[261] Xiang Wan et al. "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range". In: *BMC medical research methodology* 14 (2014), pp. 1–13.

[262] Dan Wang et al. "Unlabeled skin lesion classification by self-supervised topology clustering network". In: *Biomedical Signal Processing and Control* 66 (2021), p. 102428.

[263] Robin H Wang et al. "Clinical effectiveness and cost-effectiveness of teledermatology: Where are we now, and what are the barriers to adoption?" In: *Journal of the American Academy of Dermatology* 83.1 (2020), pp. 299–307.

[264] Ruosi Wang et al. "Individual differences in holistic processing predict face recognition ability". In: *Psychological science* 23.2 (2012), pp. 169–177.

[265] Ying Wang et al. "Heritable aspects of biological motion perception and its covariation with autistic traits". In: *Proceedings of the National Academy of Sciences* 115.8 (2018), pp. 1937–1942.

[266] Zixuan Wang, Yuki Murai, and David Whitney. "Idiosyncratic perception: A link between acuity, perceived position and apparent size". In: *Proceedings of the Royal Society B* 287.1930 (2020), p. 20200825.

[267] Zixuan Wang et al. "Idiosyncratic biases in the perception of medical images". In: *Frontiers in Psychology* (2022).

[268] Erin M Warshaw et al. "Accuracy of teledermatology for pigmented neoplasms". In: *Journal of the American Academy of Dermatology* 61.5 (2009), pp. 753–765.

[269] Erin M Warshaw et al. "Teledermatology for diagnosis and management of skin conditions: a systematic review". In: *Journal of the American Academy of Dermatology* 64.4 (2011), pp. 759–772.

[270] Mark Wexler, Marianne Duyck, and Pascal Mamassian. "Persistent states in vision break universality and time invariance". In: *Proceedings of the National Academy of Sciences* 112.48 (2015), pp. 14990–14995.

[271] John D Whited. "Teledermatology research review". In: *International journal of dermatology* 45.3 (2006), pp. 220–229.

[272] Martin J Willemink et al. "Preparing medical imaging data for machine learning". In: *Radiology* 295.1 (2020), pp. 4–15.

[273] Lauren H Williams and Trafton Drew. "What do we know about volumetric medical image interpretation?: A review of the basic science and medical image perception literatures". In: *Cognitive research: principles and implications* 4.1 (2019), pp. 1–24.

[274] Jeremy B Wilmer. "Individual differences in face recognition: A decade of discovery". In: *Current Directions in Psychological Science* 26.3 (2017), pp. 225–230.

[275] Jeremy B Wilmer et al. "Human face recognition ability is specific and highly heritable". In: *Proceedings of the National Academy of sciences* 107.11 (2010), pp. 5238–5241.

[276] Jasper Winkel et al. "Early evidence affects later decisions: Why evidence accumulation is required to explain response time data". In: *Psychonomic bulletin & review* 21 (2014), pp. 777–784.

[277] Jeremy M Wolfe. "How one block of trials influences the next: persistent effects of disease prevalence and feedback on decisions about images of skin lesions in a large online study". In: *Cognitive Research: Principles and Implications* 7.1 (2022), p. 10.

[278] Jeremy M Wolfe and Todd S Horowitz. "Five factors that guide attention in visual search". In: *Nature Human Behaviour* 1.3 (2017), pp. 1–8.

[279] Jeremy M Wolfe, Todd S Horowitz, and Naomi M Kenner. "Rare items often missed in visual searches". In: *Nature* 435.7041 (2005), pp. 439–440.

[280] Jeremy M Wolfe et al. "Low target prevalence is a stubborn source of errors in visual search tasks." In: *Journal of experimental psychology: General* 136.4 (2007), p. 623.

[281] Zhirong Wu et al. "Unsupervised feature learning via non-parametric instance discrimination". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3733–3742.

[282] Ye Xia, Allison Yamanashi Leib, and David Whitney. "Serial dependence in the perception of attractiveness". In: *Journal of vision* 16.15 (2016), pp. 28–28.

[283] Ke Yan et al. "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning". In: *Journal of medical imaging* 5.3 (2018), p. 036501.

[284] Jordan Yap, William Yolland, and Philipp Tschandl. "Multimodal skin lesion classification using deep learning". In: *Experimental dermatology* 27.11 (2018), pp. 1261–1267.

[285] Kaitlyn M Yim et al. "Teledermatology in the United States: an update in a dynamic era". In: *Telemedicine and e-Health* 24.9 (2018), pp. 691–697.

[286] Yading Yuan, Ming Chao, and Yeh-Chi Lo. "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance". In: *IEEE transactions on medical imaging* 36.9 (2017), pp. 1876–1886.

[287] Jure Zbontar et al. "fastMRI: An open dataset and benchmarks for accelerated MRI". In: *arXiv preprint arXiv:1811.08839* (2018).

[288] Bin Zhang and Sargur N Srihari. "Properties of binary vector dissimilarity measures". In: *Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing*. Vol. 1. 2003.

[289] Jianpeng Zhang et al. "Medical image classification using synergic deep learning". In: *Medical image analysis* 54 (2019), pp. 10–19.

[290] Richard Zhang et al. "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[291] Shengyu Zhao et al. "Unsupervised 3D end-to-end medical image registration with volume tweening network". In: *IEEE journal of biomedical and health informatics* 24.5 (2019), pp. 1394–1404.

[292] Jiapeng Zhu et al. "In-domain gan inversion for real image editing". In: *European Conference on Computer Vision*. Springer. 2020, pp. 592–608.

[293] Jun-Yan Zhu et al. "Generative visual manipulation on the natural image manifold". In: *European conference on computer vision*. Springer. 2016, pp. 597–613.

[294] Qi Zhu et al. “Heritability of the specific cognitive ability of face perception”. In: *Current Biology* 20.2 (2010), pp. 137–142.

[295] Zijian Zhu et al. “A genome-wide association study reveals a substantial genetic basis underlying the Ebbinghaus illusion”. In: *Journal of Human Genetics* 66.3 (2021), pp. 261–271.