Human-in-the-loop synthesis of behaviorally salient “super-distractors”

Debadrita Sen (debadritas20@iiserbpr.ac.in)
Department of Biological Sciences,
IISER Berhampur, Berhampur-760010, Odisha, India

Achin Parashar (achinp@iisc.ac.in)
Centre for Neuroscience,
Indian Institute of Science, Bangalore-560012, Karnataka, India

Abhimanyu Ray (abhimanyuray@iisc.ac.in)
Centre for Neuroscience,
Indian Institute of Science, Bangalore-560012, Karnataka, India

Shristi Chourasiya (shristic20@iiserbpr.ac.in)
Department of Biological Sciences,
IISER Berhampur, Berhampur-760010, Odisha, India

Devarajan Sridharan (sridhar@iisc.ac.in)
Centre for Neuroscience and Computer Science and Automation,
Indian Institute of Science, Bangalore-560012, Karnataka, India

Abstract

Working memory (WM) is the ability to maintain and manipulate information that is no longer present in the environment. The resilience of WM to distraction is largely tested by studies employing simple stimuli (e.g., gratings, shapes, isolated objects). Hence, what kinds of complex, naturalistic images make for potent WM distractors remains unknown. Here we leverage recent advances in deep generative models to synthesize naturalistic images that powerfully disrupt WM. Our approach generates synthetic images with a class-conditional generative adversarial network (GAN), while concurrently testing the efficacy of these images as distractors on participants (n=16) performing a spatial WM task (human-in-the-loop). With a genetic algorithm for optimization, we identify the most salient feature combinations and refine them over generations to produce powerful "super-distractor" images. Our study demonstrates the feasibility of generating novel kinds of images optimized for specific behaviors, with a human-in-the-loop paradigm.

Keywords: distraction; working memory; deep learning; generative models; genetic algorithm

Introduction

Working memory (WM) is the ability to temporarily encode, retain, and manipulate goal-relevant information. Recent research suggests that despite its remarkable robustness, WM can be disrupted by specialized distractors (Lorenc et al., 2021). For example, distractors that resemble targets can strongly interfere with the contents of WM (Yoon et al., 2006). Distractor mechanisms have been largely studied with simple, target stimuli, like oriented gratings, with only simple features. Designing effective distractors for more naturalistic stimuli – with complex features – remains an open challenge.

It is likely that "optimal distractors" for naturalistic target stimuli must likewise be complex in terms of their features and colors. To generate such naturalistic distractor images, we leverage deep-generative networks (DGNs). Previous studies employed DGNs to generate unique images whose features could drive unnaturally high responses in specific visual areas (Gu et al., 2022; Ponce et al., 2019). These algorithms typically involve iterative optimization of images by measuring (or predicting) neural activity for novel, artificially-generated images, followed by selecting, retaining and refining image features that drive the strongest brain responses. Other studies have designed naturalistic images based on optimizing human fMRI activations in particular brain regions (Luo et al., 2023).

Yet, to our knowledge, no previous study has sought to directly optimize images based on their behavioral salience. Here we generate novel images optimized to disrupt WM by directly measuring behavior, with a human-in-the-loop. We demonstrate the efficacy of these "super-distractors" with degrading performance accuracy in a spatial WM task.

Methods

Behavior-based image optimization We leverage the XDream framework (Ponce et al., 2019; Xiao & Kreiman, 2020) for iterative image optimization. Our modified XDream framework consists of four key components: 1) a BigGAN-deep network (Brock et al., 2018) that generates class-conditional images from vector codes and a class label, 2) a classifier that assigns class probabilities to the generated images and ensures only class-relevant images are selected (Mukherjee et al., 2024), 3) a genetic algorithm (GA) that selects the most salient distractors and optimizes their image codes based on behavioral outcomes and 4) an image filter based on class membership and similarity. The last step (#4) is an amalgamation of two sub-steps. First, images generated by BigGAN-deep were passed to a classifier and images outside the target class were rejected. Next, pairwise cosine similarity was computed using features extracted from AlexNet, and duplicates of similar looking images (similarity score $>$0.65) were excluded to ensure feature diversity in the presented images. After each generation, the GA selects images for optimization, probabilistically, based on participant's accuracies for each image. Here, we present results with images generated from an exemplar "ladybugs" class, although other classes of images were also explored.

Spatial working memory task Behavior-based image optimization was tested in a spatial WM task (n=16 participants; 8 experimental + 8 controls) (Fig. 1A). In each trial, after a 500ms fixation window, four colored images were shown concurrently along the cardinal axis for 250ms, equidistant from the center. Following a 750ms noise mask, participants were shown the previously displayed images in a random sequence and asked to indicate the presented location of each image using arrow keys. Each trial, therefore, involved four responses. The experiment was performed over 4 blocks, comprising a total of 400 trials (100 trials each).

On each trial, three out of the four images presented were non-optimized images and the fourth was a distractor image whose features were optimized with XDream. 30 non-optimized images were generated prior to the start of the experiment from which 10 unique triplets of images were constructed for every block. These images remained invariant and were the same for all participants. 10 participant-specific distractor images were synthesized and optimized over 4 generations. Distractor and non-optimized image triplets were combined in a counterbalanced fashion across trials.

To quantify the effectiveness of each distractor across generations of the GA, we implemented a scoring procedure based on the participants localization accuracy. On each trial, correctly localized, non-optimized images were scored +1 and were scored 0, otherwise. We hypothesized that stronger distractors would impair spatial WM, leading to poorer behavioral localization accuracy. Accordingly, each optimized distractor was assigned a GA score that was inversely proportional to the average localization accuracy for non-optimized images, across the trials in which the respective distractor was pre-

sented (n=8; experimental). Thus, distractors that consistently disrupted localization performance were afforded higher GA scores and were more likely to be selected for the next generation of images.

To validate the specificity of this optimization procedure, we included a control group (n=8), in which the scoring logic was inverted: distractors that induced better behavioral localization performance were assigned higher GA scores. Thus, over generations, the most ineffective distractors would be selected. This manipulation served as a control to evaluate whether the observed optimization effects in the experimental group were genuinely sensitive to participants' behavior, and to ensure that the GA was effectively tuning image features based on their real impact on localization performance.

Results

Synthesizing behavioral super-distractors For the experimental group (n=8 participants), distractor images optimized using the behavior-based XDream framework produced a progressive decline in spatial WM performance over successive generations (Fig. 1B). This indicates that the optimization procedure was successful in producing images that increasingly disrupted spatial WM performance. WM accuracies were statistically significantly lower in the last, compared to the first generation (p=0.003, n=8). In contrast, the control group of participants (n=8) – in which image generation was not actively guided to select the most effective distractors – showed no significant decrease in WM accuracy across generations (Fig. 1C). Accuracy remained unchanged between the first and the last generations (p=0.574, n=8). The results also indicate that the decrease in accuracy in the experimental group was not due to the effects of fatigue or tedium.

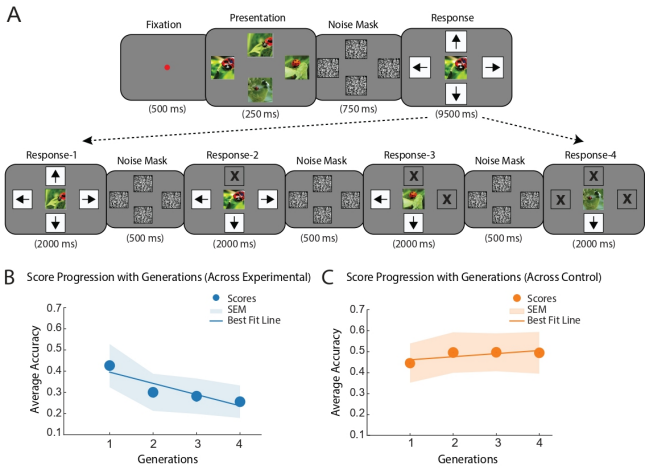

Figure 1. **(A)** Schematic of spatial WM task. **(B)** Decline in WM accuracy (y-axis) across GA generations (x-axis) for the experimental group (blue). **(C)** No significant change in accuracy across generations for the control group (orange).

Optimization yields feature convergence Distractor images generated in the final generation appeared visually more homogeneous compared to those from the inital generation (Fig. 2A). To quantify this image convergence, we computed the pairwise cosine similarity between image features extracted from the last convolutional layer of the VGG16 neural network, across generations; numerically, experimental group images (Fig. 2B, blue) were more converged than control group images (Fig. 2B, orange). Additionally, we computed the change in similarity from the first to the last generation (i.e., final generation similarity minus initial generation similarity) to capture the degree of convergence over time; however, this metric was not statistically significantly different between the experimental and control groups (p=0.195) (Fig. 2C).

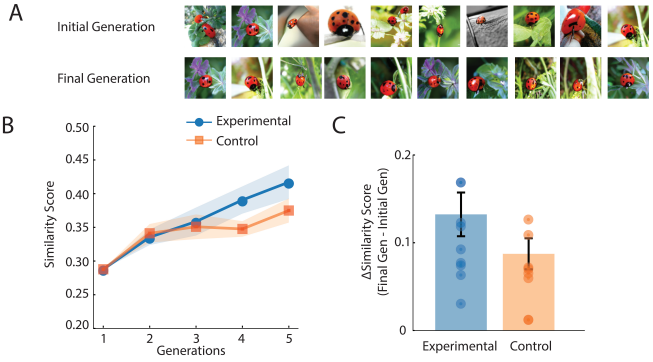

Figure 2. **(A)** Super-distractors from the initial (top) and final (bottom) generations (exemplar participant). **(B)** Similarity scores across generations and **(C)** their difference (final-initial) for experiment (blue) and control (orange) groups.

Conclusion

We propose a novel approach, leveraging deep generative models, to synthesize naturalistic "super-distractor" images for spatial WM. These images may be relevant for understanding brain mechanisms of distractor interference in visual WM.

While images from two other categories (ImageNet classes: "cars" and "plates") were also explored in pilot experiments, "ladybugs" class images exhibited better convergence visually. These latter images were the only ones employed in the behavioral experiments reported here. Future work will explore ways to more systematically select image categories for such distractor optimization. It is also possible that certain super-distractor features are common across different categories of images, and this needs to be tested in future experiments. A key advantage of our method is that it can generate participant-specific super-distractors. Nevertheless, it is possible that some super-distractor features are common across participants, a hypothesis that, again, requires further study. More generally, the images that we produce could be relevant for neuromarketing, as well as to better understand mechanisms of distractibility in neurodevelopmental disorders like attention-deficit/hyperactivity disorder (ADHD).

Acknowledgements

This work was supported by a Department of Science and Technology SwarnaJayanti fellowship, a Gore Subraya Bhat Chair Associate Professorship in Digital Health, an India-Trento Programme for Advanced Research grant and a Pratiksha Trust-Indian Institute of Science Intramural grant (all to D.S.).

References

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. Retrieved from https://tfhub.dev/s?q=biggan

Gu, Z., Jamison, K. W., Khosla, M., Allen, E. J., Wu, Y., St-Yves, G., ... Kuceyeski, A. (2022, 2). Neurogen: Activation optimized image synthesis for discovery neuroscience. *NeuroImage*, *247*, 118812. doi: 10.1016/j.neuroimage.2021.118812

Lorenc, E. S., Mallett, R., & Lewis-Peacock, J. A. (2021, 3). *Distraction in visual working memory: Resistance is not futile* (Vol. 25). Elsevier Ltd. doi: 10.1016/j.tics.2020.12.004

Luo, A. F., Henderson, M. M., Wehbe, L., & Tarr, M. J. (2023, 6). Brain diffusion for visual exploration: Cortical discovery using large scale generative models. *Advances in Neural Information Processing Systems*, *36*. Retrieved from https://arxiv.org/abs/2306.03089v2

Mukherjee, S., Parashar, A., Murthy, C. R., & Sridharan, D. (2024). Designing salient, naturalistic "super-stimuli" with deep generative models. In *Cognitive computational neuroscience (ccn).* Boston, MA. (Conference presentation)

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019, 5). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, *177*, 999-1009.e10. doi: 10.1016/j.cell.2019.04.005

Xiao, W., & Kreiman, G. (2020, 6). Xdream: Finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLOS Computational Biology*, *16*, e1007973. doi: 10.1371/journal.pcbi.1007973

Yoon, J. H., Curtis, C. E., & D'Esposito, M. (2006, 2). Differential effects of distraction during working memory on delay-period activity in the prefrontal cortex and the visual association cortex. *NeuroImage*, *29*, 1117-1126. doi: 10.1016/J.NEUROIMAGE.2005.08.024