HAT: Toward Robustness to Adversarial Silly Hat Attacks

Bjorn Lutjens
MIT
lutjens@mit.edu

Justin Kay
MIT
kayj@mit.edu

Abstract

Computer vision systems are increasingly being deployed for real-time detection of invasive species, such as sea urchins. After a period of initial success, detection rates have been dropping, with the reasons still being unknown. Here, we demonstrate that a trend in sea urchins wearing silly hats can decrease detection rates in commonly used urchin recognition software. To overcome the fragility of the underlying neural networks to this distribution shift, we propose silly **h**at **a**ugmenta**t**ions. HAT is a universal augmentation strategy adding generative AI-based hats onto sea urchin imagery, and we show that using HAT increases adversarial robustness to silly hat attacks over 82.8%.

1. Introduction

In this paper we target a real-world problem where computer vision stands to have an outsized impact [14]. As oceans are warming, sea urchin populations have expanded, creating underwater deserts devoid of biodiversity [17]. Sea otters and other natural predators are being reintroduced to counteract the invasive species [16]. Recent advances in computer vision algorithms have made it possible to identify the areas of most efficient sea otter translocation by deploying automated underwater monitoring systems that geolocate urchin populations in real time.

While effective, this work suffers from a key limitation: the urchins know. In fact, biodiversity monitoring efforts around the globe are consistently thwarted by animals gaining awareness of human observation [4–6, 10]. These animals have devised an ingenious attack to counteract human computer vision systems: *simply don a silly hat* (see Fig. 1). Computer vision algorithms are surprisingly brittle in the presence of adversarial animals wearing silly hats, often miscategorizing animals as other animals, trendsetters, A. Lincoln, iPods, etc. This poses a serious threat with real consequences. For instance, obfuscating urchin populations as harmless sea anemones diverts key resources from invasive species monitoring efforts.

Figure 1. A recent surge in topical disguises has been observed in sea urchin populations, aiming to dodge automated detection by AI algorithms. We propose silly hat augmentations that increase robustness in computer vision algorithms to this adversarial behavior.

Here, we propose a novel strategy for combating adversarial silly hat attacks, deemed **HAT**: silly **h**at **a**ugmen**t**ations. Wait, sorry, we came up with another method thing. Our method, deemed **HAT: h**at augmentations and **a**daptive op**t**imization, aims to—okay, one more thing, sorry. **HAT: h**at augmentations, **a**daptive optimization, and **t**uquel inversion: a novel three-pronged approach for combating adversarial hat attacks. Yeah that sounds cool. Inspired by the hat-invariance of a sea otter's natural visual system [1, 9], HAT uses generative AI [12] to create hat-invariance in the input, latent, and optimization space concurrently.

We show that a computer vision algorithm that nominally detects urchins with 86% accuracy drops in accuracy to 0.1% under realistic hat attacks, and fine-tuning the algorithm with HAT augmentation increases the algorithm's accuracy back to 83%. All numbers will be bolded.

2. Method

2.1. HAT

We first introduce a novel data augmentation strategy for combating adversarial silly hat attacks, deemed **HAT**: silly

Equal contribution, same hat
†Presented at SigTBD April Fools' Day 2025

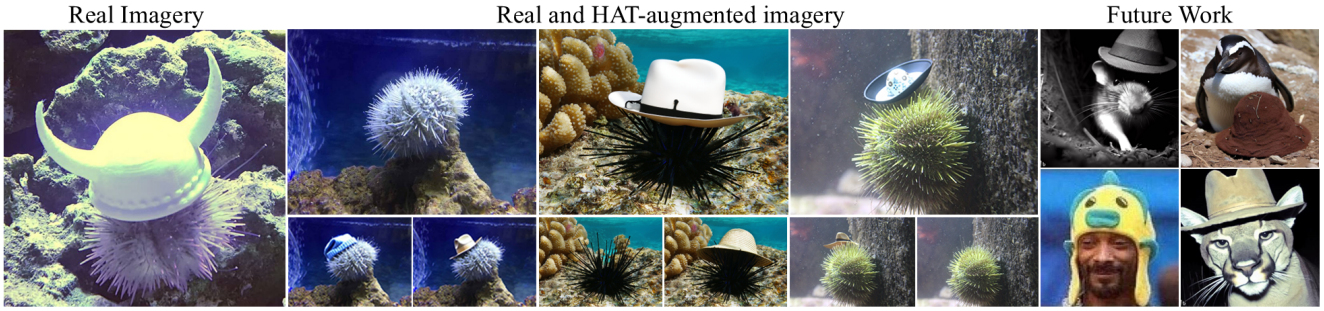

Figure 2. An adversarial urchin wears a hat (left; this is real [2]). HAT augments imagery of hat-less sea urchins with hats (middle; our contribution). Future work includes robustness to rats disguised as Al Pacino, Calvin Cordozar Broadus Jr. disguised as a fish, etc. (right).

Algorithm 1 The HAT algorithm

Require: $X_1, ...X_n$; $\hat{}_1, ..., \hat{}_k$
  while True **do**
    $X_i \leftarrow X_i + \hat{}_j$
    Train on X_i
  end while

hat **a**ugmen**t**ations. Our method introduces a new type of neurologically-inspired invariance into visual recognitions networks: namely, invariance in the presence of silly hats. To do so, we utilize generative AI [11] to superimpose plausible hats into the input space. We do so using Alg. 1.

2.2. $\widehat{SGD}$

Next we derive a novel optimizer for training neural networks, $\widehat{SGD}$ ("hat SGD"). Our optimizer is agnostic to any loss function $\widehat{L}$, where the $\hat{}$ may be a fedora, beanie, beret, etc.

Let $\widehat{w}_{\widehat{0}}$ be the initial parameters of our model. Then, each timestep $\widehat{t} \in \{\widehat{0}, \widehat{1}, \widehat{2}, \dots\}$, we compute the gradient $\widehat{L}$ with respect to the parameters $\widehat{w}_{\widehat{t}}$. We denote this gradient by:

$$\hat{\nabla}\widehat{L}(\widehat{w}_{\widehat{t}}).$$

This gradient is the same as $\nabla L(w_t)$ from SGD, but novel because of the $\hat{}$. Then, let $\widehat{\eta}$ denote the *learning rate*. Then,

$$\widehat{w}_{\widehat{t}+\hat{1}} \;=\; \widehat{w}_{\widehat{t}} \;-\; \widehat{\eta}\,\nabla\widehat{L}(\widehat{w}_{\widehat{t}}).$$

Note how $\widehat{t} + \hat{1}$ is also wearing a hat.

2.3. Tuquel Inversion

Next we propose a method for introducing a similar hat invariance in the latent space, deemed Tuquel[1] Inversion. We use $\widehat{SGD}$ on the image tokens to do whatever textual inversion does [8]. It definitely sounds like it could probably work.

[1] https://www.thecanadianencyclopedia.ca/en/article/tuque

Method	Numbers
Schmidhuber (1987) [11]	**32, 65, 77, 801, 1, 920**
Schmidhuber (1989) [12]	**5, 67, 64, 386, 100, 100**
Schmidhuber (1990) [13]	**0.4, 50, 61, 0, 784, 88**
Ours	**66, 87, 405, 32, 1, 10**

Table 1. Comparison of our method, in terms of boldness, with prior work. While previous methods are pretty bold, ours clearly generates the boldest numbers. Because this is published in a top-tier venue [3], we conclude that our method, HAT, is now state-of-the-art. Numbers have no units or meaning.

3. Results

The results of our approach are depicted in Fig. 1 with additional examples in Fig. 2. Quantitative results are shown in Tab. 1.

4. Conclusions

There are several interesting, legitimately laboursome but yieldful directions for future work, e.g., how best to construct the **HAT** acronym. In our work, we propose a method that upweights the contribution of the first letter of each word: **h**at augmentations, **a**daptive optimization, and **t**uquel inversion. However other approaches are possible. For example, **ha**t augmen**t**ations, **h**at augmentations and **a**dap**t**ive optimization, or even recursive methods such as **hat** [15].

Figure 2 illustrates possible extensions of our algorithm. Notably, one interesting avenue is to tackle the issue of penguins that avoid satellite-based colony tracking algorithms [7] by disguising their guano with hats. Other extensions include celebrities wearing urchin-hats to avoid paparazzi detection or predator species looking for a normal life by blending in through popular headwear.

Acknowledgements

We take our hats off to the urchins without whom this study would not have been possible.

References

[1] Sea Otters Hold Hands While Sleeping and They Even Cuddle — discovermagazine.com. https://www.discovermagazine.com/planet-earth/sea-otters-hold-hands-while-sleeping-and-they-even-cuddle. [Accessed 12-03-2025]. 1

[2] Urchin imagery, 2025. Raw urchin imagery sourced from Tumblr, aquariumadvice.com, natural history museum, and new england aquarium. AI imagery generated using bing AI and adobe photoshop. 2

[3] Anonymous authors. Hat: Toward robustness to adversarial silly hat attacks. *Desk reject from CVPR; Under review SigTBD*, 2025. 2

[4] Margaret Wise Brown, Clement Hurd, Si Kahn, Cathy Fink, and Marcy Marxer. *Goodnight moon*. HarperTrophy New York, 1947. 1

[5] Donna Lynn Browning. Clifford, the big red dog. *The Reading Teacher*, 40(3):369–369, 1986.

[6] Eric Carle. The very hungry caterpillar. *Early Years Educator*, 2(3):38–41, 2000. 1

[7] P.T. Fretwell, R.A. Phillips, M. de L. Brooke, A.H. Fleming, and A. McArthur. Using the unique spectral signature of guano to identify unknown seabird colonies. *Remote Sensing of Environment*, 156:448–456, 2015. 2

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[9] RL Gentry and RS Peterson. Underwater vision of the sea otter. *Nature*, 216(5114):435–436, 1967. 1

[10] Marcus Pfister, Detlev Jöcker, and Armin Nufer. *Der Regenbogenfisch*. Nord-Süd, 1995. 1

[11] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987. 2

[12] Jurgen Schmidhuber. A local learning algorithm for dynamic feedforward and recurrent networks. *Connection Science*, 1 (4):403–412, 1989. 1, 2

[13] Jürgen Schmidhuber. Reinforcement learning in markovian and non-markovian environments. *Advances in neural information processing systems*, 3, 1990. 2

[14] DR SEUSS'S. The lorax. 2021. 1

[15] Richard Stallman et al. The gnu manifesto. 1985. 2

[16] Nathan L. Stewart and Brenda Konar. Kelp forests versus urchin barrens: Alternate stable states and their effect on sea otter prey quality in the aleutian islands. *Journal of Marine Sciences*, 2012(1):492308, 2012. 1

[17] Adriana Vergés, Christopher Doropoulos, Hamish A. Malcolm, Mathew Skye, Marina Garcia-Pizá, Ezequiel M. Marzinelli, Alexandra H. Campbell, Enric Ballesteros, Andrew S. Hoey, Ana Vila-Concejo, Yves-Marie Bozec, and Peter D. Steinberg. Long-term empirical evidence of ocean warming leading to tropicalization of fish communities, increased herbivory, and loss of kelp. *Proceedings of the National Academy of Sciences*, 113(48):13791–13796, 2016. 1