On GAN-based Data Integrity Attacks Against Robotic Spatial Sensing

Tamim Khatib, Patrick Kreidl,
Ayan Dutta and Swapnoneel Roy
School of Computing
University of North Florida
Jacksonville, Florida

Ladislau Bölöni
Department of Computer Science
University of Central Florida
Orlando, Florida

Abstract

Communication is arguably the most important way to enable cooperation among multiple robots. In numerous such settings, robots exchange local sensor measurements to form a global perception of the environment. One example of this setting is adaptive multi-robot informative path planning, where robots' local measurements are "fused" using probabilistic techniques (e.g., Gaussian process models) for more accurate prediction of the underlying ambient phenomena. In an adversarial setting, in which we assume a malicious entity–the adversary–can modify data exchanged during inter-robot communications, these cooperating robots become vulnerable to data integrity attacks. Such attacks on a multi-robot informative path planning system may, for example, replace the original sensor measurements with fake measurements to negatively affect achievable prediction accuracy. In this paper, we study how such an adversary may design data integrity attacks using a Generative Adversarial Network (GAN). Results show the GAN-based techniques learning spatial patterns in training data to produce fake measurements that are relatively undetectable yet significantly degrade prediction accuracy.

1 Introduction

In a generic multi-robot coordination setting, robots share local perceptions and sensor measurements with each other to achieve a common global objective. Examples of this generic framework can be found across most multi-robot applications, including multi-robot SLAM (Atanasov et al. 2015; Thrun and Liu 2005), multi-robot path planning (Dutta and Dasgupta 2017; Yu and LaValle 2013), multi-robot manipulation (Feng et al. 2020; Kaiser et al. 2022), among others. We consider a variant of traditional multi-robot path planning, namely the multi-robot information sampling problem—instead of finding n paths for n robots which are optimal in the joint space, we find n length-B paths such that the global information collection metric (e.g., entropy, mutual information) is optimized. This problem is well-known to be NP-hard, and optimization becomes computationally intractable even with a small number of robots involved. No matter how the problem is tackled, however, increased prediction accuracy becomes achievable if robots share collected information (e.g., images, temperature/humidity measurements, etc.) with each other (Dutta et al. 2020; Dutta, Kreidl, and O'Kane 2021).

Unfortunately, as with any cyber-physical system, multi-robot systems are vulnerable to cyber-attacks. One or more malicious entities can alter the sensor measurements, for example, and such data integrity attacks can create havoc in the real-world (e.g., an autonomous robot can spread pesticides on healthy crops instead of the weeds and kill them due to fake reported measurements (Gupta et al. 2020)). As can be understood, this can have significant economic and societal impacts. Recently, blockchain-based secure coordination protocols have been proposed as a countermeasure for such data integrity attacks (Samman et al. 2021; 2022; Strobel, Castelló Ferrer, and Dorigo 2020), albeit the studies consider relatively simplistic rules for falsifying measurements. In this paper, we position ourselves into the adversary's seat under more sophisticated attacker assumptions. The specific research question we ask is "*how can we generate fake-yet-plausible data, or falsify measurements yet still resemble properties of the true process?*" For example, it is commonly the case that measurements over a geographical area will be spatially correlated. An adversary who falsifies data that too often appears spurious relative to those correlations may find that fake data rejected as outliers, self-mitigating the intended harm, or even risk detection in settings that anticipate adversarial influences.

To this end, we propose a Generative Adversarial Network (GAN)-based technique for falsifying sensor measurements in an ambient phenomena. More specifically, we adopt a popular GAN model, namely Deep Convolutional GAN (DCGAN) (Radford, Metz, and Chintala 2016) and assume that the adversary has access only to a subset of training data (e.g., through offline exfiltration) from which the multi-robot system is itself designed. That is, our GAN-based generative technique is designed to falsify measurements without

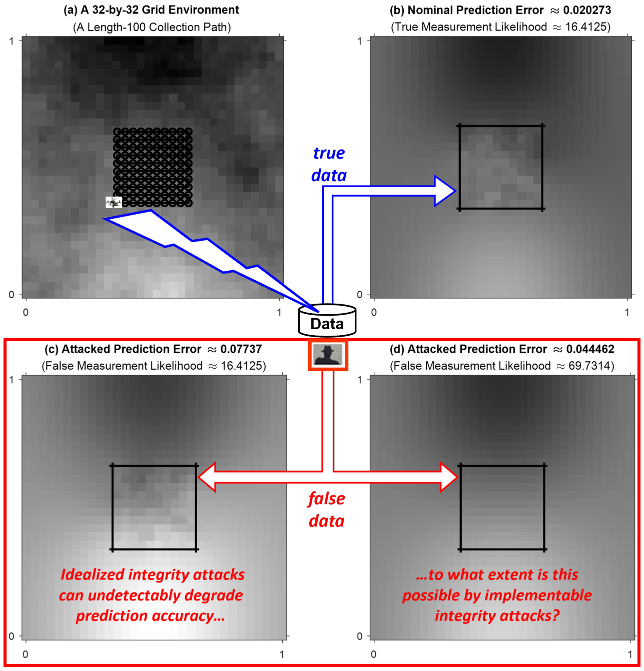


Figure 1: Illustration of our primary research question: "how can an attacker falsify measurements yet still resemble the properties of the true data?" (a) A specific information state over a 32-by-32 grid environment after a sensing robot visits only the central 10-by-10 subregion. (b) The case of no attacker, rendering a nominal prediction for all unvisited cells conditioned on the true data from the visited cells. (c) The case of an idealized attacker who, through access to the true measurements, can falsify without altering the likelihood and is thus able to undetectably increase prediction error (by 282% in this example). (d) The case of an attacker who cannot access the true measurements and rather replaces them by the mean field of the visited cells, seen to still increase prediction error (by 119% in this example) but with risk of eventual detection by virtue of the likelihood discrepancy between true and false measurements.

knowledge of the original measurements they will replace. We implement such a sophisticated attacker in the context of a typical informative path planning formulation, using a Gaussian process to represent the environment's spatially-correlated information (Cao, Low, and Dolan 2013; Dutta, Ghosh, and Kreidl 2019; Krause, Singh, and Guestrin 2008; Viseras et al. 2016). As Figure 1 illustrates, we assume each robot's budget falls well-short of achieving full area coverage on its own, and thus prediction accuracy across the unvisited region (conditioned on the data collected within the visited region) can serve as a surrogate for whether multi-robot coordination remains successful. The attacker's efficacy is empirically assessed along two competing metrics: the first quantifies "harmfulness" in terms of degraded prediction accuracy relative to that achieved in the absence of attacks, while the second quantifies "detectability" in terms of likelihood discrepencies between the generated fake measurements and the replaced original measurements. Both metrics employ well-understood Gaussian process techniques (Rasmussen 2003), where the prior statistics that enable nominal prediction methods (e.g., optimal filtering) also play a fundamental role for attack detection methods (e.g., significance testing). In particular, the idealized likelihood-preserving strategy depicted by Figure 1(c) certainly renders a statistically undetectable attacker, but our GAN-based attacker falls under the class of "implementable" attacks depicted by Figure 1(d). More generally, because attack detection typically involves a batch of samples, even implementable strategies that render likelihood discrepancies on a per-attack basis can remain essentially undetectable, depending on the extent that the false measurement likelihoods remain "batch-wise" equivalent to the true counterparts.

Altogether, in the multi-robot informative path planning context, this paper's primary contributions are:

1. The first to propose an automated GAN-based generative technique for falsifying measurements, providing the basis by which to better protect future spatial sensor systems against anticipated AI-enabled attacks on data integrity.

2. The first to apply statistical testing concepts to quantify the detectability of falsified measurements relative to the nominal information model.

Our experimental results demonstrate the implemented GAN-based attacker, when evaluated on the spectrum between the two baseline attackers depicted in Figure 1, achieves significant harm with relatively low detectability. While this evaluation rests upon the validity of our Gaussian information assumptions, the proposed GAN-based technique itself is applicable to any corpus of training data. We thus believe this paper's proposed attack technique has general relevance to cybersecurity and robotics research, certainly towards preparing future automation for increased resilience against data integrity attacks by AI-enabled adversaries (Lu, Issaranon, and Forsyth 2017; Xia and Liu 2016).

2 Related Work

Informative robotic path planning has gained considerable traction in recent years due to its immense practical relevance. Most of the proposed studies use Gaussian Processes (GPs) (Rasmussen 2003) to model the underlying ambient phenomena (Dutta, Ghosh, and Kreidl 2019; Kemna et al. 2017; Ma, Liu, and Sukhatme 2016). In an online adaptive version of the problem, robots' current sensor measurements drive their future path planning (Dutta, Kreidl, and O'Kane 2021; Ma et al. 2018; Luo and Sycara 2018). None of these above-mentioned studies on informative path planning consider potential data integrity attacks. Such was considered only recently, proposing to preserve data integrity by integrating a Blockchain-based consensus protocol within the informative path planning modules (Samman et al. 2021; 2022). These studies needed only simple attack strategies to demonstrate the blockchain's value towards preserving integrity, focusing rather on characterizing the communication and computational overhead to sustain the blockchain whether robots communicate continuously, periodically, or opportunistically. No blockchain is present in the scope of this paper, however, so the spatial sensing system remains vulnerable to data integrity attacks and our focus turns to prospects of a more sophisticated (e.g., AI-enabled) adversary. To this end, we propose a formal deep convolutional generative adversarial network (DCGAN) model (Goodfellow et al. 2014; Radford, Metz, and Chintala 2016) and experimentally demonstrate its efficacy for realistic-yet-fake data generation. Synthetic data generation using such GANs is a challenging task in many domains (Dewi et al. 2022; Dupont et al. 2018). The literature on generative adversarial networks (GANS) is vast—the interested reader is referred to (Creswell et al. 2018) for an overview.

3 Attacks via DCGAN

Background: The collection of technologies referred to as deep learning evolved from neural network technologies developed over several decades. While the concept of training a neural network using backpropagation of the errors was well established in the 1980s, initial consensus viewed backpropagating across multiple layers as computationally infeasible. Thus, early neural networks were shallow, typically consisting of a single hidden layer. While convolutional neural networks with multiple layers was demonstrated by LeNet (LeCun et al. 1989) and opened the possibility of efficiently processing image data, the wide use of deep neural networks took off after 2012 with the spectacular performance of a deep network on the ImageNet competition (Krizhevsky, Sutskever, and Hinton 2012). The success of deep networks was due to advances in GPU-based parallel computing as well as a number of innovations in the training process, including a standardized stochastic gradient descent formulation of the training problem, dropout regularization (Srivastava et al. 2014), ReLU non-linearity and batch normalization (Ioffe and Szegedy 2015).

Most early successes of deep learning were for discriminative models that learn a distribution $P(c|X)$ to infer a class c based on an input X. In contrast, in certain applications we want to train generative models that allow us to sample a distribution $P(X)$, often in conditional form $P(X|c)$. Generative adversarial networks (GANs) (Goodfellow et al. 2014) and variational autoencoders (VAEs) (Kingma and Welling 2013) are two of the most successful techniques to train generative models. VAEs combine an encoder with a generator that aims to reproduce the encoded image. In contrast, a GAN combines a generator with a discriminator that learns to distinguish between "real" samples coming from the original distribution (the training samples) and "fake" samples produced by the generator. The joint training of generator and discriminator promotes the generator to better approximate the real samples and, in turn, the discriminator to better distinguish fake ones.

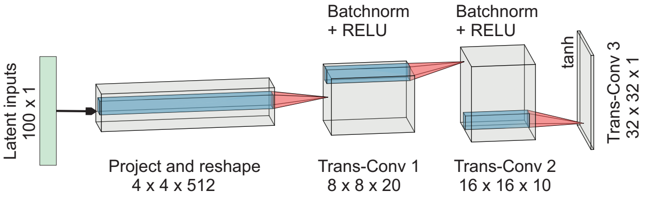


Figure 2: The generator network of our DCGAN architecture, using a 5×5 filter and stride of size 2, where the number of filters in the three Transposed Convolutional layers are 64, 32 and 1, respectively.

Architecture: While the initial GAN formulation (Goodfellow et al. 2014) already experimented with various convolutional models in the generator and discriminator components, successful training of GANs was foud to be highly sensitive to architecture, parameter choices and even the random initialization. This lack of stability introduced a significant amount of human intervention in the learning process. This makes the initial GAN formulation less practical for applications, such as an AI-enabled adversary synthesizing data integrity attacks, where human judgment of the training process is frequently infeasible.

The Deep Convolutional GAN formulation (Radford, Metz, and Chintala 2016), or DCGAN for short, is a collection of architectural choices that has been found to consistently train in a stable manner. These choices include being fully convolutional, eliminating both (i) the mass pooling layers that in previous architectures had been interspersed with convolutional layers in the discriminator as well as (ii) the fully connected layers except a very simple layer at the output of the discriminator (D_θ) and at the input of the generator (G_ϕ). DCGAN also consistently uses ReLU activation in the generator (except for a tanh function at the output layer) and Leaky ReLU activation in all layers of the discriminator. Finally, DCGAN uses the batch normalization algorithm in both components of the GAN. Within the context of these architectural ideas, DCGAN models can be customized to the needs of the application. Figure 2 describes the specific generator model selected to design our DCGAN attacker.

4 Gaussian Process Models

Gaussian Process (GP) models of environmental uncertainties (Guestrin, Krause, and Singh 2005; Krause, Singh, and Guestrin 2008; Rasmussen and Williams 2006) assume that all the collection locations generate information according to Gaussian random vector $\mathbf{X}$ with known (prior) mean vector μ and covariance matrix Σ. When navigation planning constraints do not permit full coverage of the environment, the set of all collection locations can be decomposed into two disjoint subsets, U and V, corresponding to the unvisited and visited locations, respectively. Under negligible sensor noise assumptions, the Gaussian random vector $\mathbf{X}_U$ characterizing the uncollected information has (posterior) mean vector and covariance matrix given by

$$\begin{aligned} \mu_{U|\mathbf{x}_V} &= \mu_U + \Sigma_{UV}\Sigma_{VV}^{-1}(\mathbf{x}_V - \mu_V) \\ \Sigma_{UU|\mathbf{x}_V} &= \Sigma_{UU} - \Sigma_{UV}\Sigma_{VV}^{-1}\Sigma_{VU} \end{aligned} \tag{1}$$

with $\mathbf{X}_V = \mathbf{x}_V$ denoting the values measured in the visited locations and the prior statistics organized into block forms corresponding to sets U and V i.e.,

$$\mu = \left[\begin{array}{c} \mu_U \\ \mu_V \end{array} \right] \quad \text{and} \quad \Sigma = \left[\begin{array}{cc} \Sigma_{UU} & \Sigma_{UV} \\ \Sigma_{VU} & \Sigma_{VV} \end{array} \right].$$

The experiments to be discussed in Section 5 will leverage the following facts for GP models. Firstly, the posterior mean vector is the minimum-mean-square-error (MMSE) predictor of the process $\mathbf{X}_U$ given measurements $\mathbf{x}_V$ i.e., in the context of (1),

$$\mu_{U|\mathbf{x}_V} = \arg \min_{\hat{x} \in \mathbb{R}^{|U|}} \mathbf{E}\left[||\mathbf{X}_U - \hat{x}||^2 \mid \mathbf{X}_V = \mathbf{x}_V \right] \tag{2}$$

with $|\cdot|$ denoting set cardinality, $\mathbf{E}[\cdot]$ denoting the expectation operator, and $||\cdot||$ denoting the vector 2-norm (so $||\cdot||^2$ is the sum-square-error between process and estimate). The associated posterior covariance characterizes the MMSE predictor's achievable accuracy e.g., the expected minimum sum-square-error in (2) is given by $\mathrm{Tr}\left(\Sigma_{UU|\mathbf{x}_V}\right)$, where $\mathrm{Tr}(\cdot)$ denotes the matrix trace. Secondly, because the process $\mathbf{X}_V$ in the visited region is itself Gaussian with (prior) statistics μ_V and Σ_{VV}, the likelihood of a given measurement $\mathbf{x}_V$ is

$$L\left(\mathbf{x}_V\right) = \frac{\exp\left(-\frac{1}{2}\left(\mathbf{x}_V - \mu_V\right)^T \Sigma_{VV}^{-1}\left(\mathbf{x}_V - \mu_V\right)\right)}{(2\pi)^{\frac{|V|}{2}}|\Sigma_{VV}|^{\frac{1}{2}}}$$

and its maximum is $L(\mu_V)$ i.e., the mean field μ_V is the most-likely measurement. Our simulation experiments involve sample-based approximations of the GP, for each sample $\mathbf{X} = \mathbf{x}$ computing specifically the root-mean-square-error and the log-likelihood,

$$\mathcal{E}(\mathbf{x}) = \left|\left| \frac{\mathbf{x}_U - \mu_{U|\mathbf{x}_V}}{|U|} \right|\right| \text{ and } \mathcal{L}(\mathbf{x}_V) = \log L\left(\mathbf{x}_V\right), \tag{3}$$

to quantify harmfulness and detectability, respectively.

In real-world environments, the practical challenge of Gaussian prediction is to obtain accurate prior statistics. Such priors are typically derived from training data via statistical learning methods (e.g., maximum-likelihood (Rasmussen and Williams 2006)) and, for spatially-distributed Gaussian processes, usually also leverage domain-specific environmental considerations. A length-p Gaussian process has $d = 2p + p(p-1)/2$ degrees-of-freedom, in general, where requirements that the number of training samples $n \gg d$ are often formidable in robotics applications. In such situations, it is common to assume a reduced-order structure for the Gaussian process. For example, the so-called "homogeneous and isotropic Gaussian Markov random field using an exponential kernel" defines the covariance matrix using just two hyper-parameters: given any pair of locations i and j at spatial positions $\mathbf{q}_i$ and $\mathbf{q}_j$, respectively, the kernel function is given by

$$\sigma_{ij}^2 = \beta^2 \exp\left(-||\mathbf{q}_i - \mathbf{q}_j||/\ell\right)$$

where $\beta > 0$ is the local standard deviation and $\ell > 0$ is the exponential rate of diminishing covariance between increasingly-distant locations. Observe that such a process, when $\beta = 0$, will deterministically render only the mean field μ, whose p degrees-of-freedom can analogously exhibit a reduced-order structure. Figure 3 illustrates a realization from such a structured length-1024 Gaussian process—as discussed in Section 5, this exact process is sampled to synthesize experimental data sets (for both DCGAN training and subsequent evaluation).

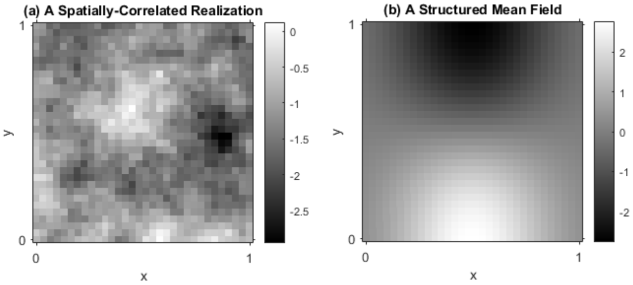


Figure 3: (a) A length-1024 realization from a zero-mean Gaussian process, mapped onto locations over a 32x32 uniform grid in the unit square, with a spatially-correlated covariance structure induced by exponential kernel parameters $\beta = \ell = 1$. (b) A mean-field that within the spatial region is proportional to the sum of an upper "valley" (with a peak depth of -2.5) and a lower "hill" (with a peak height of +2.5), each a Gaussian-shaped surface centered by its length-2 mean vector (taking values $[0.50; 0.75]$ and $[0.50; 0.25]$, respectively) with elliptical contours defined by its 2-by-2 diagonal covariance matrix (taking values 0.0625 and 0.2500 in x and y, respectively).

Figure 1 is also based upon a realization from the structured Gaussian process described in Figure 3. Figure 1(a) shows the true values over the full 32-by-32 grid as well as the length-100 path by which a robot has visited the central 10-by-10 sub-grid. Figure 1(b) shows the prediction of the unvisited cells given the true measurements $\mathbf{x}_V$, while the lower-two images show predictions given false measurements $\mathbf{z}_V \neq \mathbf{x}_V$. Specifically, Figure 1(d) implements the strategy $\mathbf{z}_V := \mu_V$, which assumes only that the attacker has knowledge of the prior mean (e.g., via exfiltration of historical data, as assumed in Section 3 for the DCGAN attacker) but, of course, exhibits no variation over successive attacks and is thus trivially detectable. Figure 1(c) implements the strategy $\mathbf{z}_V := 2\mu_V - \mathbf{x}_V$, which is likelihood-preserving (and thus undetectable) but also idealized (relative to the DCGAN attacker) in the sense that the attacker can intercept the true measurements $\mathbf{x}_V$ beforehand.

5 Experimental Setup and Results

Training: The training data available to our DCGAN consists of 3000 samples of the 32 × 32 spatial information field, each a realization of the structured GP described in Figure 3. Training of the DCGAN is implemented in MATLAB, using 1000 epochs with the learning rate set to 0.0002. We have introduced a noise of 0.3 in the discriminator network to provide enough learning opportunities for the generator. In each epoch, we select a random set of minibatches of size 128. For each such minibatch of training data, the loss is calculated using the negative log-likelihood function. We have used Adam optimization for both networks. Training performance of the GAN uses the following scoring mechanism: the generator's score is calculated by averaging the probabilities of the discriminator's output for the generated samples Y, i.e., $score(G_\phi) = mean(D_\theta(Y))$, while the discriminator's score is the sum of the averages of the discriminator's real and fake classification probabilities, i.e., letting Y' denote the real samples, $score(D_\theta) = 0.5(mean(D_\theta(Y'))+mean(D_\theta(1-Y)))$. In an ideal setting both $score(G_\phi)$ and $score(D_\theta)$ would converge to 0.5, indicating that the discriminator can identify a generated sample correctly in one iteration and the generator can fool the discriminator in the next. These scores for our trained GAN are shown in Figure 4, where the scores follow this ideal pattern closely. Finally, Figure 5 presents a visual comparison between a set of true information fields, actually within the training data, and a set of fake information fields, generated by the trained DCGAN–our attacker appears successful at producing fake yet plausible measurements.

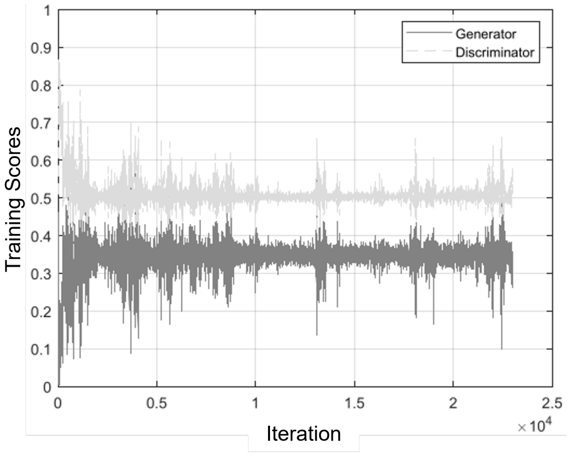
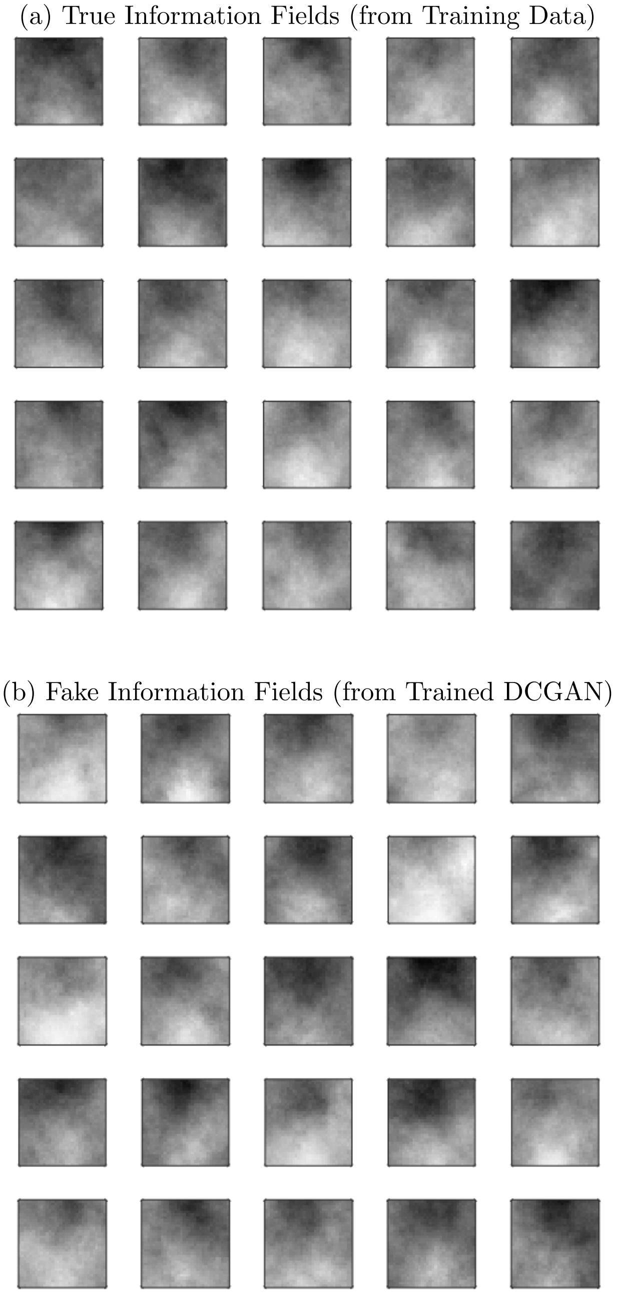


Figure 4: DCGAN training results based on 3000 training samples from the GP described in Figure 3. The converged scores of discriminator and generator networks suggest our DCGAN trains in a stable manner.


(a) True Information Fields (from Training Data)

(b) Fake Information Fields (from Trained DCGAN)

Figure 5: A comparison between (a) true information fields and (b) fake information fields, providing visual evidence that our DCGAN is successfully trained.

Note that the training data is sufficient for traditional Gaussian learning methods to well-approximate the governing prior statistics. Thus, in the scope of our experiments, falsification strategies that depend upon knowledge of prior statistics (e.g., the mean-field attacker) are considered as implementable as our DCGAN attacker. We again emphasize that the choice to synthesize Gaussian training data is strictly for evaluation purposes, permitting the concepts and formulas of Section 4 to be employed when comparing attack strategies in terms of harmfulness and detectability. Contrary to strategies enabled by Gaussian learning, however, our DCGAN attacker is designed to learn its strategy without a-priori assumptions on structure or distribution.

Evaluation: The evaluation of our DCGAN attack strategy is in terms of both harmfulness and detectability relative to the two baseline attack strategies depicted in Figure 1. As Figure 1 suggests, the (trivially-detectable) mean-field strategy will typically cause less harm than the (undetectable) idealized strategy, but the latter's access to true measurements is a privilege our DCGAN attacker does not possess. Thus, DCGAN success is the extent that both (i) harmfulness surpasses that of the mean-field baseline and (ii) detectability remains near that of the idealized baseline.

The testing data consists of 1000 samples, each again a realization $\mathbf{X} = \mathbf{x}$ of the structured GP described in Figure 3 and subject to the computations in (3). Harmfulness of each attack $\mathbf{z}_V$ is quantified by the increased prediction error relative to that associated with the prediction using the true measurements $\mathbf{x}_V$; more precisely, the metric on a per-sample basis is

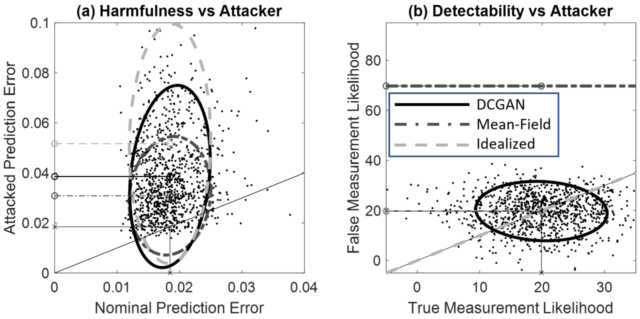
$$\mathcal{H}\left(\mathbf{x}, \mathbf{z}_V\right) = \mathcal{E}\left(\left[\mathbf{x}_U \quad \mathbf{z}_V\right]^T\right) / \mathcal{E}(\mathbf{x}) - 1,$$

which is zero under no attack but otherwise can be positive or negative to reflect whether the attacked prediction error is greater than or less than, respectively, the nominal counterpart. Detectability of each attack is quantified by the likelihood discrepancy between true and false measurements; more precisely, the metric on a per-sample basis is

$$\mathcal{D}\left(\mathbf{x}, \mathbf{z}_V\right) = \left[\mathcal{L}\left(\mathbf{z}_V\right) - \mathcal{L}\left(\mathbf{x}_V\right)\right] / \left[\mathcal{L}\left(\mu_V\right) - \mathcal{L}\left(\mathbf{x}_V\right)\right],$$

which is zero under no attack, one under the mean-field attack and otherwise positive (yet upper bounded by one) or negative to reflect whether the false likelihood is greater or less than, respectively, the true counterpart.

Figure 6 quantifies the extent that our DCGAN attacker (i) surpasses the mean-field baseline with respect to harmfulness and (ii) remains near the idealized baseline with respect to detectability. The harmfulness metric sees DCGAN increasing prediction error by 116% on average, compared to 72% and 190% for mean-field and idealized. The detectability metric of -2% on average implies DCGAN is nearly undetectable by significance tests based on likelihood means, but the 21% standard deviation implies moderate risk of detection by tests based on likelihood volatility. Of course, whether such increase in detectability is worth the increase in harmfulness depends upon the broader adversarial posture.



Attacker Strategy	Prediction Error $\mathcal{E}$	Harmfulness Metric $\mathcal{H}$	Measurement Likelihood $\mathcal{L}$	Detectability Metric $\mathcal{D}$
None	0.018 ± 0.004	0.00 ± 0.00	19.87 ± 6.88	0.00 ± 0.00
DCGAN	0.039 ± 0.018	1.16 ± 1.05	19.74 ± 7.30	-0.02 ± 0.21
Mean-Field	0.031 ± 0.013	0.72 ± 0.75	69.73 ± 0.00	1.00 ± 0.00
Idealized	0.052 ± 0.026	1.90 ± 1.61	19.87 ± 6.88	0.00 ± 0.00

Figure 6: Attacker evaluation results based on 1000 test samples from the GP described in Figure 3. The scatter plots show how the primary features, namely prediction error $\mathcal{E}$ in (a) and measurement likelihood $\mathcal{L}$ in (b), depend upon the considered attacks. Both scatter plots use (i) axes with a 2:5 aspect ratio, (ii) the horizontal axis to score the case of no attack, (iii) the vertical axis to score the cases of attack, (iv) a diagonal line to show where the attack score equals the non-attack score and (v) points to explicitly show the 1000 sample outcomes only under the DCGAN attacker, from which the solid-lined ellipse is rendered via a least-squares fit; the rendered dash-dotted and dashed ellipses in (a) are analogously obtained under, respectively, the mean-field and idealized attackers who in (b) render, respectfully, as horizontal and diagonal lines. The table lists the individual sample statistics (mean±stdv) of not only each primary feature but also of each associated metric.

6 Conclusion and Future Work

Motivated by a future prevalence of multi-robot spatial sensing systems along with increasing prospects that they fall prey to data integrity attacks, this paper studies the adversarial question of how to generate fake yet plausible measurements. Assuming a sophisticated AI-enabled adversary, we specifically implemented a GAN-based technique for falsifying measurements in an ambient phenomena. Our study discussed GAN architectural considerations to promote stability in training and identified key statistical metrics to quantify harmfulness and detectability in evaluation. Experimental results show the GAN-based technique learning spatial patterns in training data to produce fake measurements that are relatively undetectable yet still cause significant harm. Future work should extend the study towards increasingly general relevance for secure robotics, such as other cooperative decision scenarios (e.g., capture-the-flag), modified attacker assumptions (e.g., multi-agent collusion) or proposing AI-enabled defenders that demonstrably enhance resilience.

Acknowledgments

This work was supported in part by the NSF Cyber-Physical Systems Grants #1932300 and #1931767.

References

Atanasov, N.; Le Ny, J.; Daniilidis, K.; and Pappas, G. J. 2015. Decentralized active information acquisition: Theory and application to multi-robot slam. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 4775–4782. IEEE.

Cao, N.; Low, K. H.; and Dolan, J. M. 2013. Multi-robot informative path planning for active sensing of environmental phenomena: a tale of two algorithms. In Gini, M. L.; Shehory, O.; Ito, T.; and Jonker, C. M., eds., *International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS '13, Saint Paul, MN, USA, May 6-10, 2013*, 7–14. IFAAMAS.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine* 35(1):53–65.

Dewi, C.; Chen, R.-C.; Liu, Y.-T.; and Tai, S.-K. 2022. Synthetic data generation using dcgan for improved traffic sign recognition. *Neural Computing and Applications* 34(24):21465–21480.

Dupont, E.; Zhang, T.; Tilke, P.; Liang, L.; and Bailey, W. 2018. Generating realistic geology conditioned on physical measurements with generative adversarial networks. *arXiv preprint arXiv:1802.03065*.

Dutta, A., and Dasgupta, P. 2017. Bipartite graph matching-based coordination mechanism for multi-robot path planning under communication constraints. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 857–862. IEEE.

Dutta, A.; Bhattacharya, A.; Kreidl, O. P.; Ghosh, A.; and Dasgupta, P. 2020. Multi-robot informative path planning in unknown environments through continuous region partitioning. *International Journal of Advanced Robotic Systems* 17(6):1729881420970461.

Dutta, A.; Ghosh, A.; and Kreidl, O. P. 2019. Multi-robot informative path planning with continuous connectivity constraints. In *2019 International Conference on Robotics and Automation (ICRA)*, 3245–3251. IEEE.

Dutta, A.; Kreidl, O. P.; and O'Kane, J. M. 2021. Opportunistic multi-robot environmental sampling via decentralized markov decision processes. In *International Symposium Distributed Autonomous Robotic Systems*, 163–175. Springer.

Feng, Z.; Hu, G.; Sun, Y.; and Soon, J. 2020. An overview of collaborative robotic manipulation in multi-robot systems. *Annual Reviews in Control* 49:113–127.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27.

Guestrin, C.; Krause, A.; and Singh, A. P. 2005. Near-optimal sensor placements in gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, 265–272. ACM.

Gupta, M.; Abdelsalam, M.; Khorsandroo, S.; and Mittal, S. 2020. Security and privacy in smart farming: Challenges and opportunities. *IEEE Access* 8:34564–34584.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456.

Kaiser, T. K.; Lang, C.; Marwitz, F. A.; Charles, C.; Dreier, S.; Petzold, J.; Hannawald, M. F.; Begemann, M. J.; and Hamann, H. 2022. An innate motivation to tidy your room: Online onboard evolution of manipulation behaviors in a robot swarm. In Matsuno, F.; Azuma, S.-i.; and Yamamoto, M., eds., *Distributed Autonomous Robotic Systems*, 190–201. Cham: Springer International Publishing.

Kemna, S.; Rogers, J. G.; Nieto-Granda, C.; Young, S.; and Sukhatme, G. S. 2017. Multi-robot coordination through dynamic voronoi partitioning for informative adaptive sampling in communication-constrained environments. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2124–2130. IEEE.

Kingma, D. P., and Welling, M. 2013. Auto-encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*.

Krause, A.; Singh, A.; and Guestrin, C. 2008. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research* 9(Feb):235–284.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1(4):541–551.

Lu, J.; Issaranon, T.; and Forsyth, D. 2017. Safetynet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE international conference on computer vision*, 446–454.

Luo, W., and Sycara, K. 2018. Adaptive sampling and online learning in multi-robot sensor coverage with mixture of gaussian processes. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 6359–6364. IEEE.

Ma, K.-C.; Ma, Z.; Liu, L.; and Sukhatme, G. S. 2018. Multi-robot informative and adaptive planning for persistent environmental monitoring. In *Distributed Autonomous Robotic Systems*. Springer. 285–298.

Ma, K.-C.; Liu, L.; and Sukhatme, G. S. 2016. An information-driven and disturbance-aware planning method for long-term ocean monitoring. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2102–2108. IEEE.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In Bengio, Y., and LeCun, Y., eds., *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings.*

Rasmussen, C. E., and Williams, C. K. 2006. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge.

Rasmussen, C. E. 2003. Gaussian processes in machine learning. In *Summer School on Machine Learning*, 63–71. Springer.

Samman, T.; Spearman, J.; Dutta, A.; Kreidl, O. P.; Roy, S.; and Bölöni, L. 2021. Secure multi-robot adaptive information sampling. In *2021 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, 125–131.

Samman, T.; Dutta, A.; Kreidl, O. P.; Roy, S.; and Bölöni, L. 2022. Secure multi-robot information sampling with periodic and opportunistic connectivity. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*, 1–7.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1):1929–1958.

Strobel, V.; Castelló Ferrer, E.; and Dorigo, M. 2020. Blockchain technology secures robot swarms: a comparison of consensus protocols and their resilience to byzantine robots. *Frontiers in Robotics and AI* 7:54.

Thrun, S., and Liu, Y. 2005. Multi-robot slam with sparse extended information filers. In *Robotics Research. The Eleventh International Symposium*, 254–266. Springer.

Viseras, A.; Wiedemann, T.; Manss, C.; Magel, L.; Mueller, J.; Shutin, D.; and Merino, L. 2016. Decentralized multi-agent exploration with online-learning of gaussian processes. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 4222–4229. IEEE.

Xia, F., and Liu, R. 2016. Adversarial examples generation and defense based on generative adversarial network. *arXiv preprint* 1712.

Yu, J., and LaValle, S. M. 2013. Structure and intractability of optimal multi-robot path planning on graphs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence.*