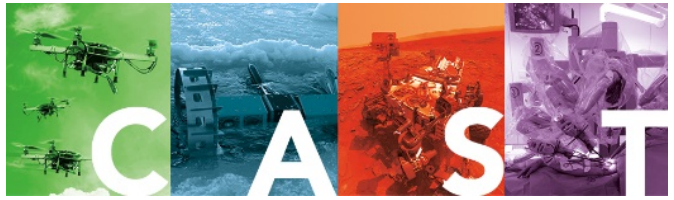


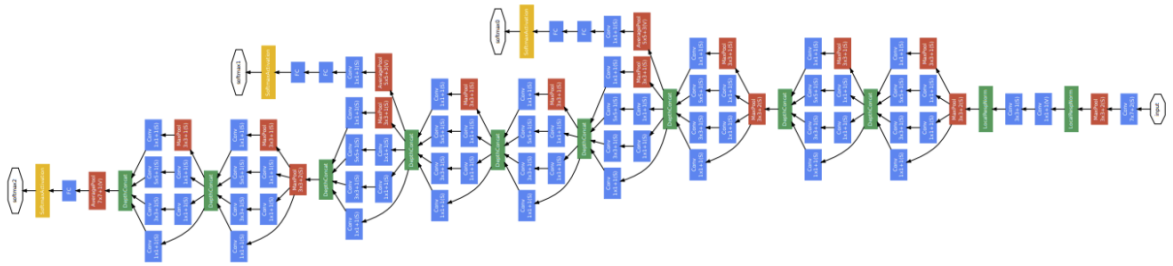
dolcit Caltech

Two Vignettes in Robust Detection & Adversarial Analysis in Control

Yisong Yue

Deep learning in the lab

$$f\left(\right.$$  $$\left.\right)$$



$$= \begin{pmatrix} 0.01 \\ \vdots \\ 0.15 \\ \vdots \\ 0.02 \end{pmatrix} \in \mathbb{R}^n$$

$$x \in \mathbb{R}^d$$

fawn

ImageNet Large Scale Visual Recognition Challenge, Russakovsky et al, 2012

Deep learning in the wild

“ I want to use deep learning to optimize the design, manufacturing and operation of our aircrafts. But I need some guarantees. ” -- Aerospace Director

Robustness is Essential to Real-World Systems

- Many different concrete formal definitions.

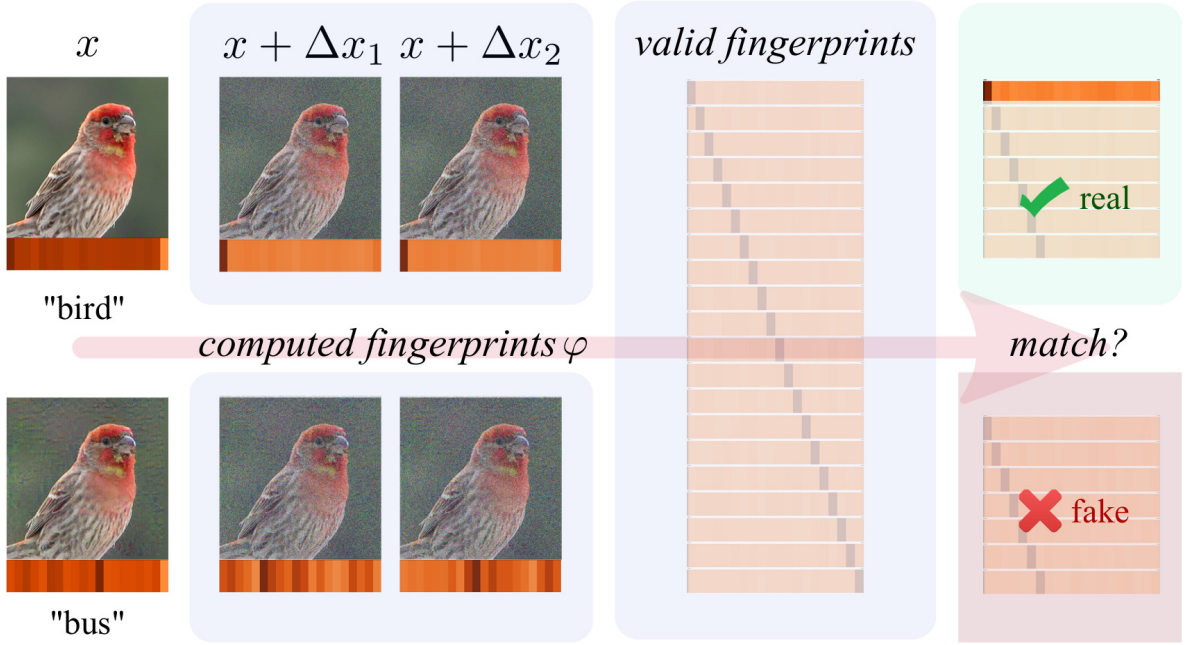
Input/output or dynamical

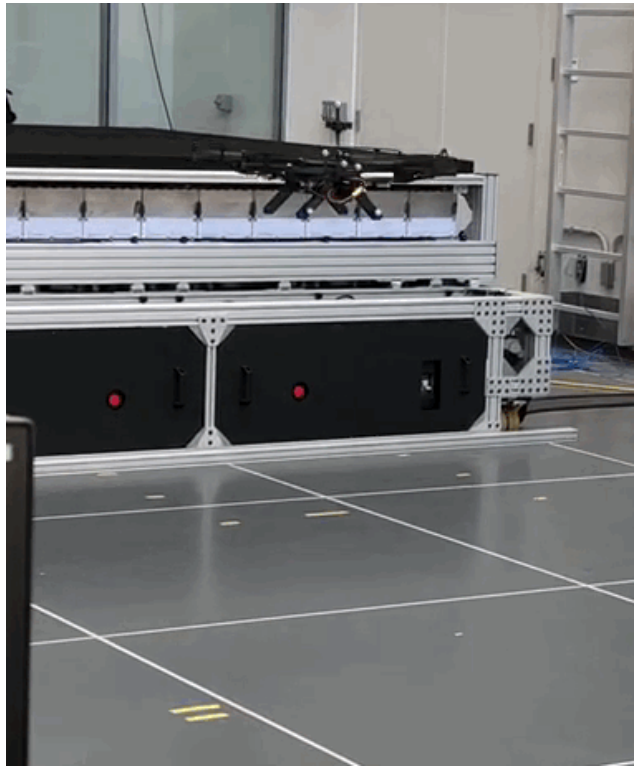
Smoothness or invariance

The behavior of system should be stable under disturbances to the environment the system is operating in.

Adversarial perturbations, domain shift, etc.

Two Vignettes in Robustness





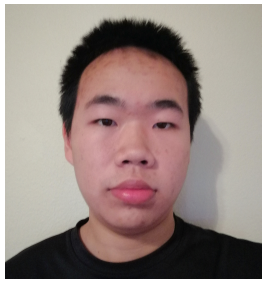
- High dimensional inputs (images)
- Robust input/output behavior
- Detect fake images

- Low(-er) dim. description
- Robust dynamical behavior
- Certify landing won't crash

Detecting Adversarial Examples via Neural Fingerprinting

Sumanth Dathathri

Stephan Zheng



Tianwei Yin

Richard Murray

Detecting Adversarial Examples via Neural Fingerprinting,
Sumanth Dathathri, Stephan Zheng, Tianwei Yin, Yisong Yue, Richard M. Murray, arXiv.

Adversarial Examples

Given: data *(x, y),* loss function *L,* and model parameters θ, an **attacker** tries to find *x' = x + dx* such that:

$$\max_{x':||x-x'||_2<\delta} L(x', f(x'), y^*; \theta)$$

A **defender** tries to find a θ, mechanism, …, to ensure no solutions *x'* exist within distance δ of *x*.

2014 - … : ongoing "arms race".

- Many attacks and defenses have been proposed in recent years.
- Many defenses have been broken by stronger attacks.
- Hard to (theoretically) guarantee robustness.

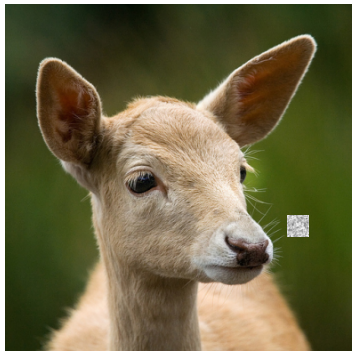
Landscape of Adversarial Robustness Regimes

Robust Detection vs Prediction

$$f\left(\text{  }\right) = \begin{pmatrix} 0.01 \\ \vdots \\ 0.15 \\ \vdots \\ 0.02 \end{pmatrix}$$

fawn

non-adversarial :)

$$f\left(\text{  }\right) = \begin{pmatrix} 0.03 \\ \vdots \\ 0.08 \\ \vdots \\ 0.01 \end{pmatrix}$$

~~monkey~~

adversarial!

Black-Box/Grey-Box/White-Box

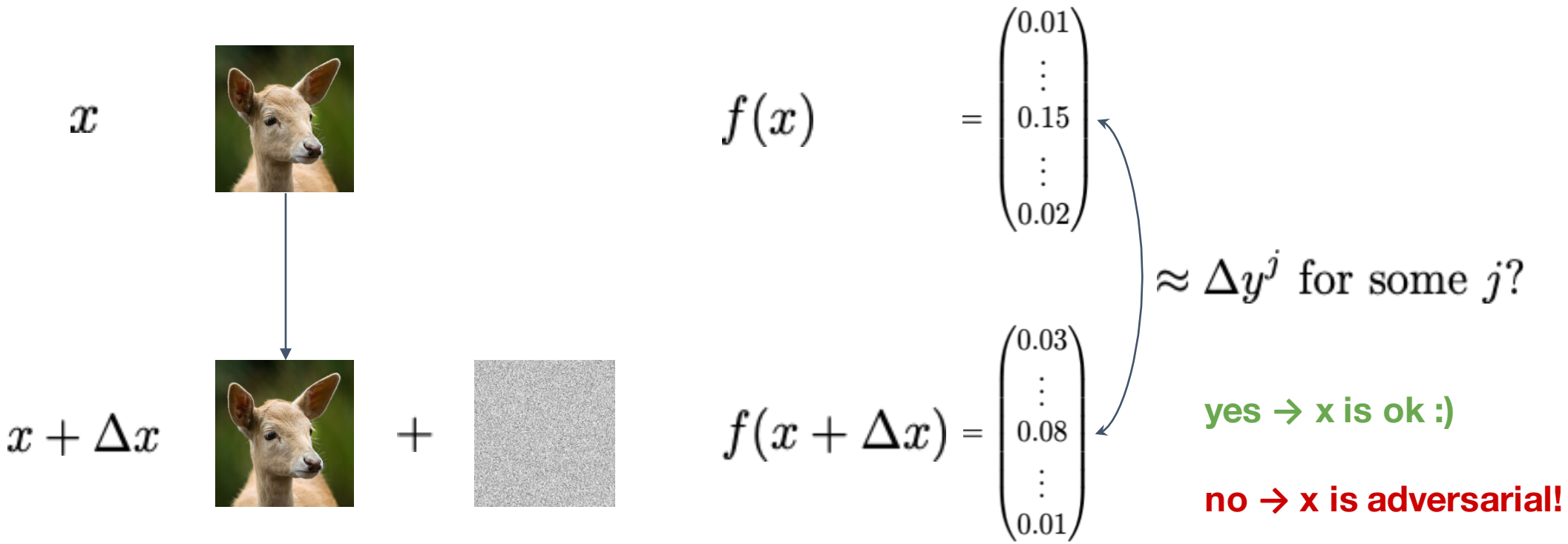
Black-Box: Attacker only has black-box access to f.

Grey-Box: Attacker has access to f, training data, but not to a secret key.

White-Box: Attacker has access to everything.

Neural Fingerprinting
Robust Detection under Grey-Box Adversarial Attacks

- ***Given:*** *fingerprint* Δx, $\{\Delta y^j\}$, and a trained model f.
- NFP does "local consistency check" around input x.



Neural Fingerprinting
Robust Detection under Grey-Box Adversarial Attacks

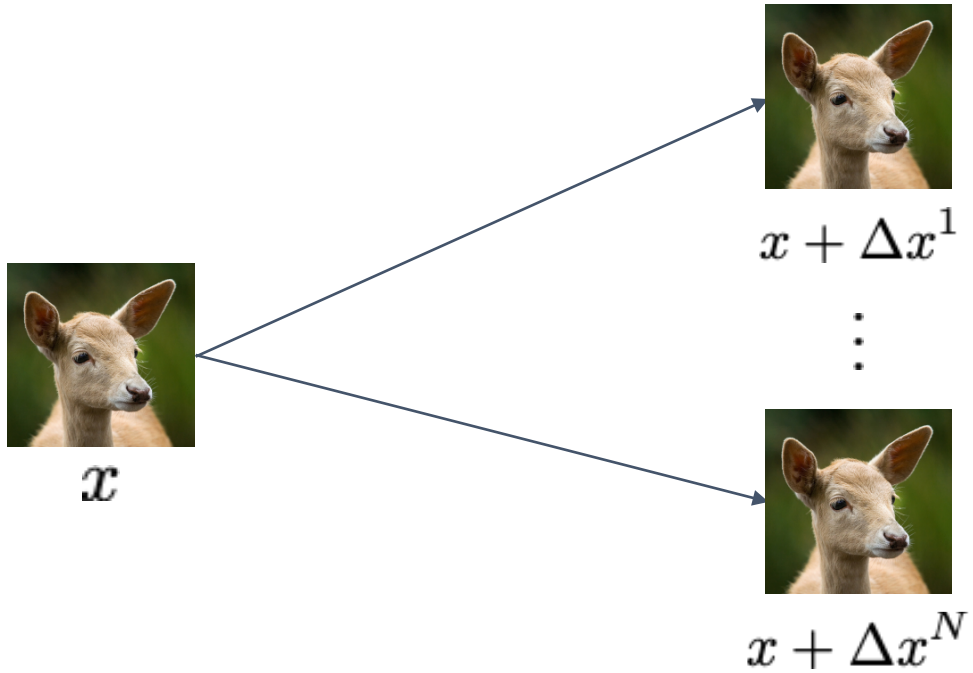
- **Intuition:** increasingly hard to find perturbation dx that conforms with a collection of (secret) fingerprints!

Fingerprints:

$$\chi^{i,j} = (\Delta x^i, \Delta y^{i,j})$$
$$i = 1 \dots N, j = 1 \dots J$$

Robust Detection:

$$?\exists j : \frac{1}{N} \sum_{i=1}^{N} ||F(x, \Delta x) - \Delta y^{i,j}||_2^2 < \tau$$
$$F(x, \Delta x) = f(x + \Delta x) - f(x)$$



Training with Neural Fingerprinting

- (Assume fingerprints already chosen)

Original Training Loss

Fingerprint Training Loss

$$\min_{\theta} \sum_{(x,y)} (L_0(x, y; \theta) + \alpha L_{fp}(x, y, \chi; \theta))$$

$$L_{fp}(x, y, \chi; \theta) = \sum_{i=1}^{N} ||F(x, \Delta x^i) - \Delta y^{i,k}||_2^2$$

$$F(x, \Delta x) = f(x + \Delta x) - f(x)$$

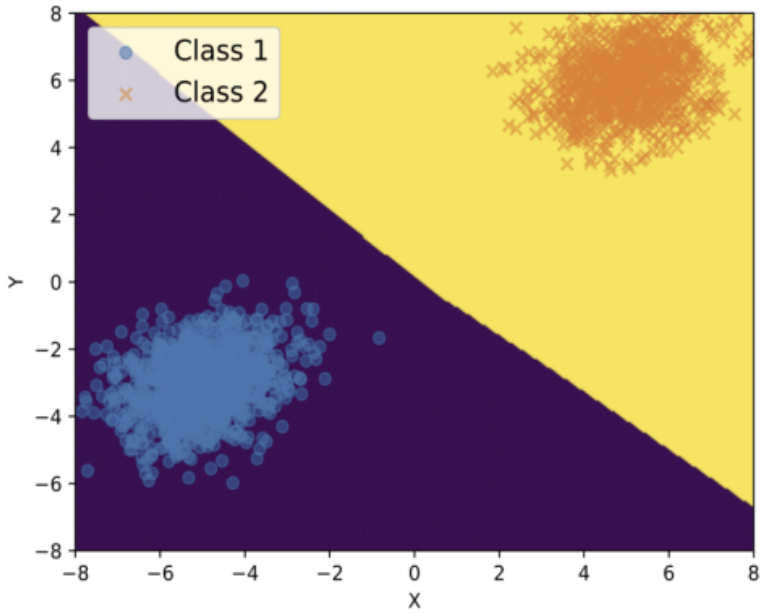
k is ground truth class

Visualizing Fingerprint Loss

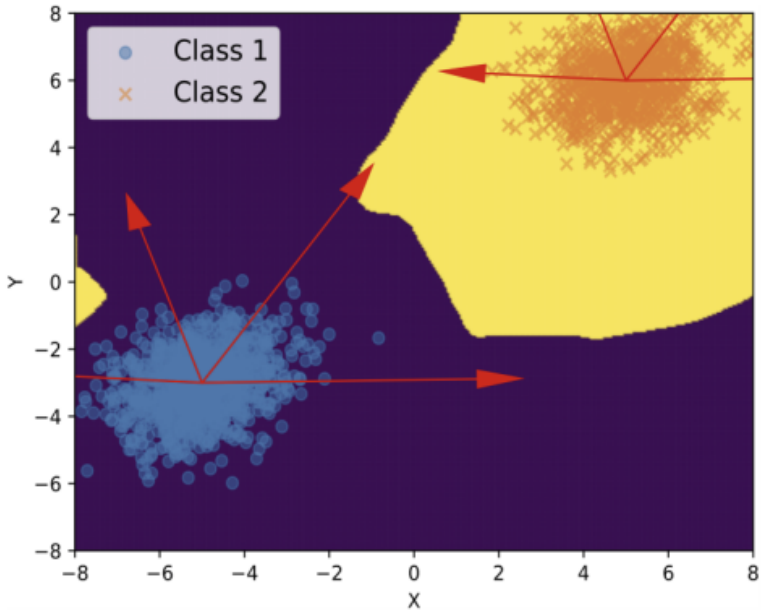
- Toy example, 2 classes, 4 fingerprints

$$L_{fp}(x, y, \chi; \theta) = \sum_{i=1}^{N} ||F(x, \Delta x^i) - \Delta y^{i,k}||_2^2$$

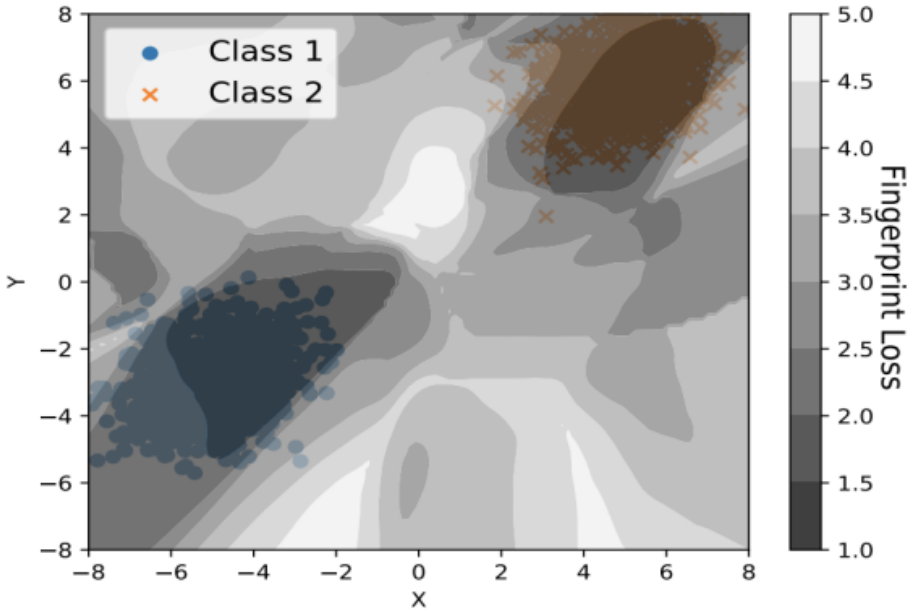
$$F(x, \Delta x) = f(x + \Delta x) - f(x)$$



Without NFP



With NFP



NFP Loss

(Theoretical characterization for linear models in paper)

Choosing Fingerprints

- Choose Δx, $\{\Delta y^j\}$ by random sampling.

$$\Delta x^i \sim \mathbb{N}(0, \sigma^2)$$

$$\Delta y^{i,k}_{l \neq k} = -\alpha(2p-1)$$

$$\Delta y^{i,k}_{l=k} = \beta(2p-1)$$

$$p \sim \text{Bern}\left(\frac{1}{2}\right)$$

p resampled for each i

$\alpha = 0.25$, $\beta = 0.75$ (method is not sensitive)

Grey-Box: fingerprints not known to attacker
White-box: fingerprints known to attacker

Whole Recipe

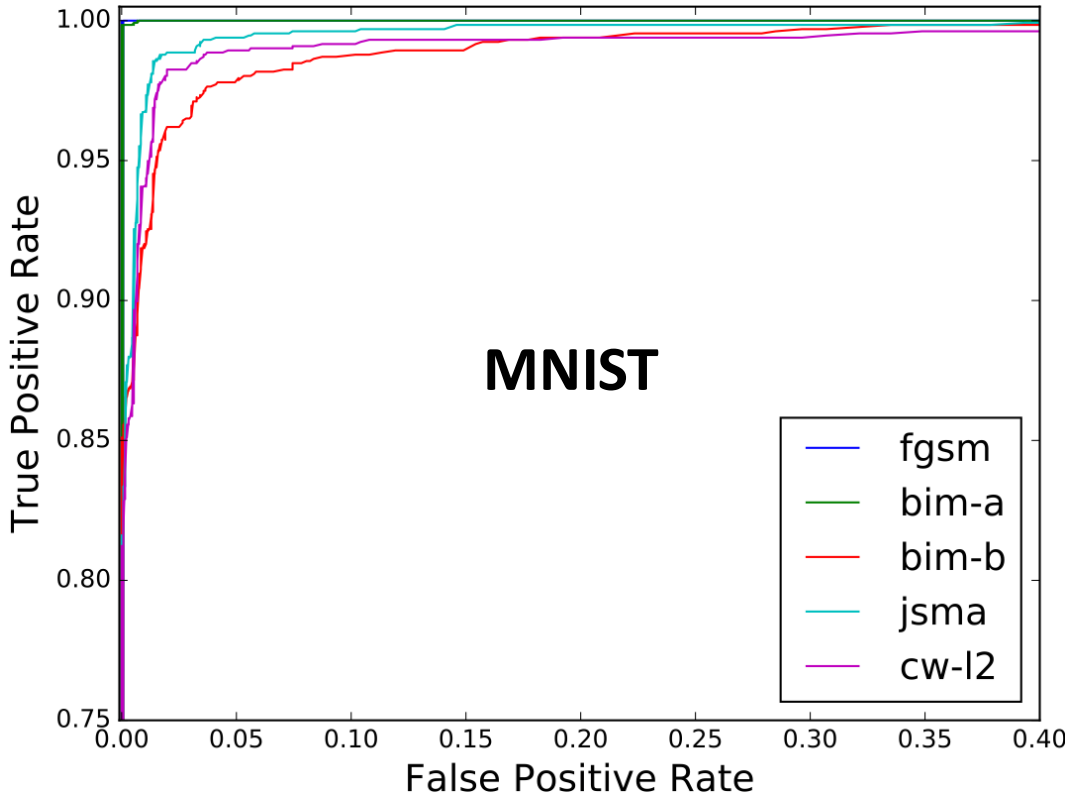
- Choose Fingerprints via Random Sampling
- Train NN model w/ Fingerprint Loss
- Deploy NN model
  - Detects fake examples while also doing prediction

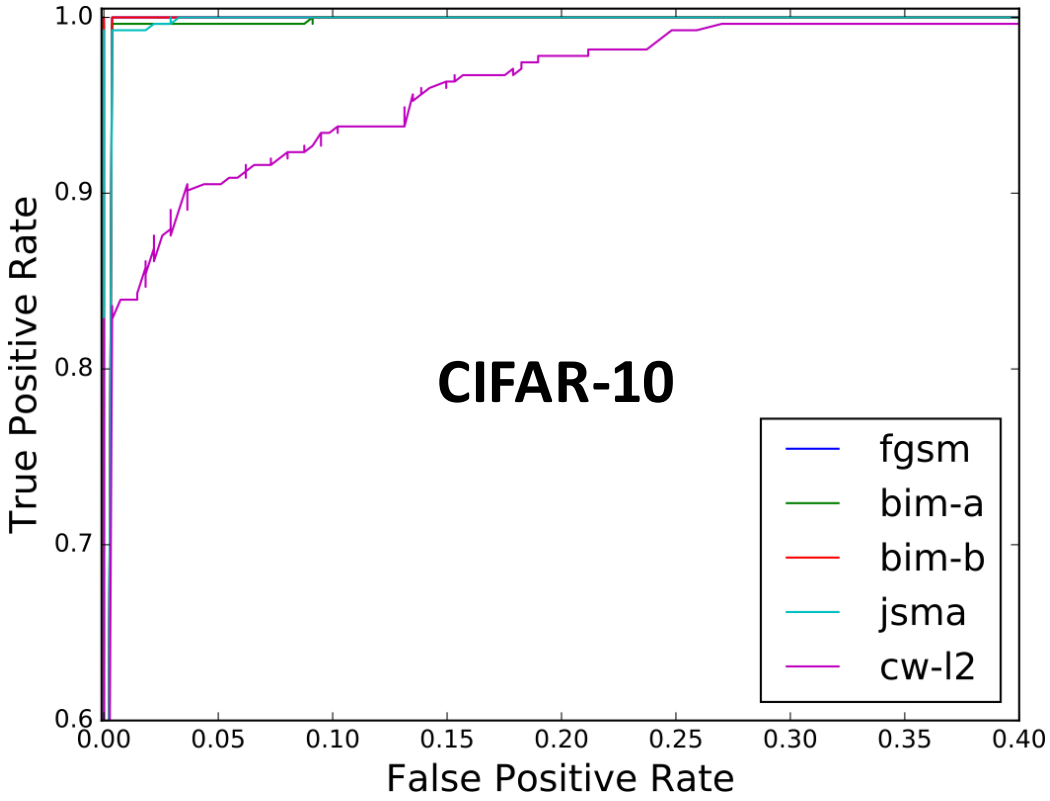
Grey-Box: Near-Perfect Detection of SotA Attacks

Data	Method	FGSM	JSMA	BIM-a	BIM-b	CW-L_2
MNIST	LID	99.68	96.36	99.05	99.72	98.66
	KD	81.84	66.69	99.39	99.84	96.94
	BU	27.21	12.27	6.55	23.30	19.09
	KD+BU	82.93	47.33	95.98	99.82	85.68
	NeuralFP	**100.0**	**99.97**	**99.94**	**99.98**	**99.74**
CIFAR-10	LID	82.38	89.93	82.51	91.61	93.32
	KD	62.76	84.54	69.08	89.66	90.77
	BU	71.73	84.95	82.23	3.26	89.89
	KD+BU	71.40	84.49	82.07	1.1	89.30
	NeuralFP	**99.96**	**99.91**	**99.91**	**99.95**	**98.87**

Data	FGSM	BIM-b
MiniImagenet-20	99.96	99.68

Grey-Box: Near-Perfect Detection of SotA Attacks

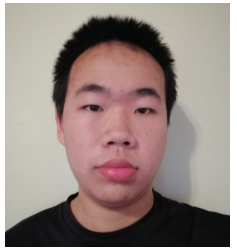




White-Box: Robust Against Existing Attacks

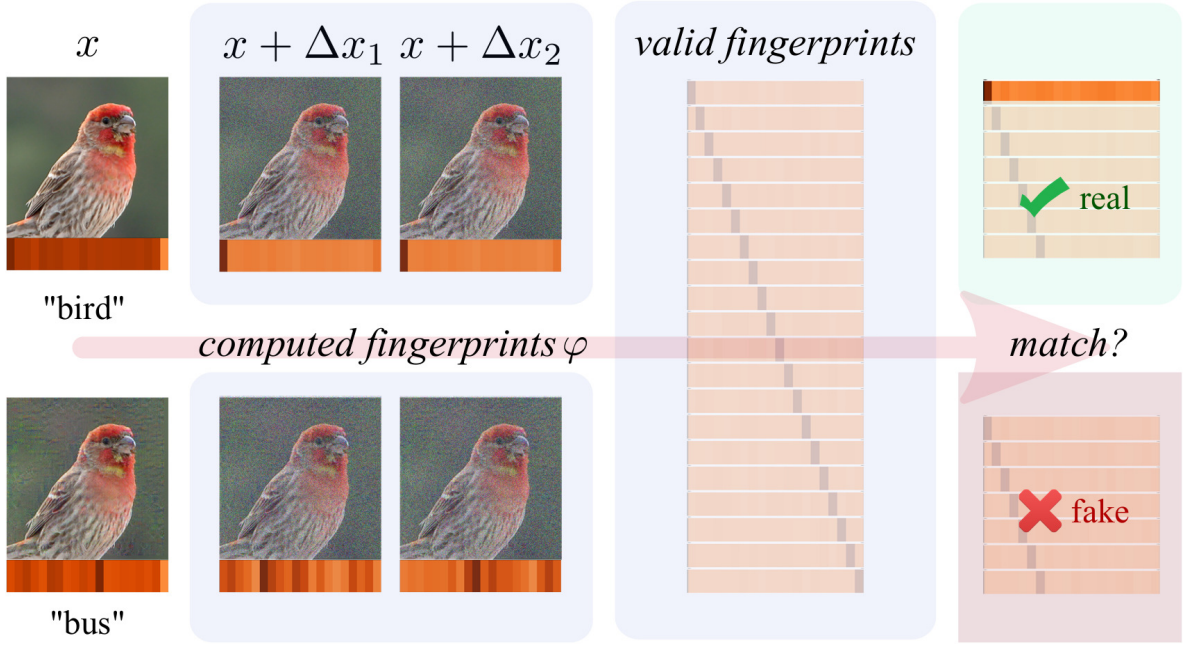
Data	Method	Adaptive-FGSM	Adaptive-BIM-b	Adaptive-CW-L_2	Adaptive-CW-L_2 ($\gamma_2 = 1$)	Adaptive-SPSA
MNIST	*NeuralFP*	99.91	99.37	95.04	99.17	99.94
CIFAR-10	*NeuralFP*	99.99	99.92	97.19	97.56	99.99

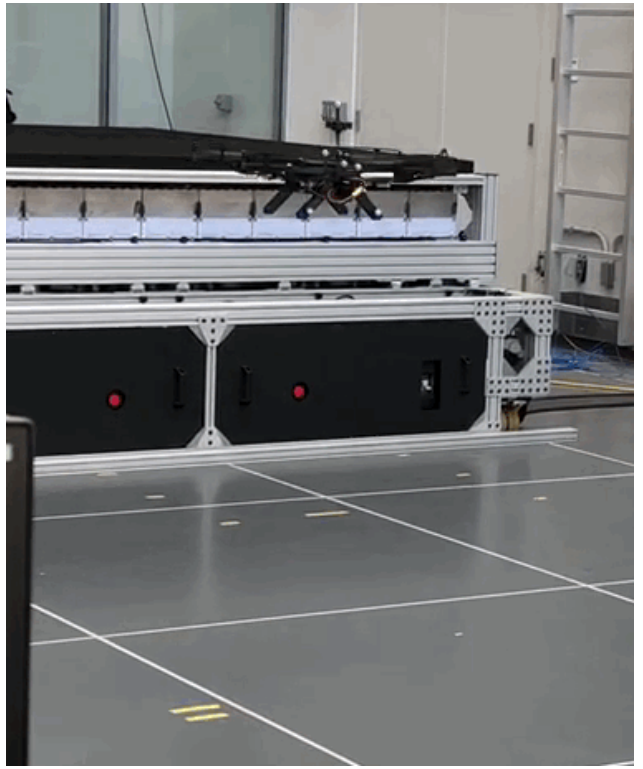
- We subsequently designed new attack to break Neural Fingerprinting in Adaptive White-Box Setting.



Tianwei Yin

Two Vignettes in Robustness

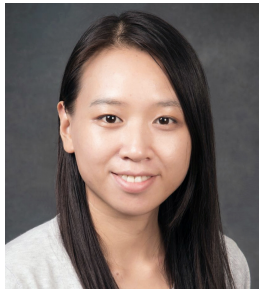




- High dimensional inputs (images)
- Robust input/output behavior
- Detects fake images

- Low(-er) dim. description
- Robust dynamical behavior
- Certify landing won't crash

Robust Regression for Safe Exploration in Control



Angie Liu

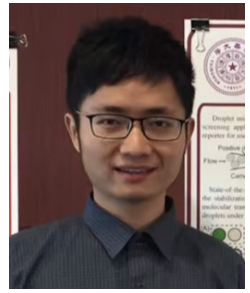
Guanya Shi

Anima Anandkumar

Soon-Jo Chung

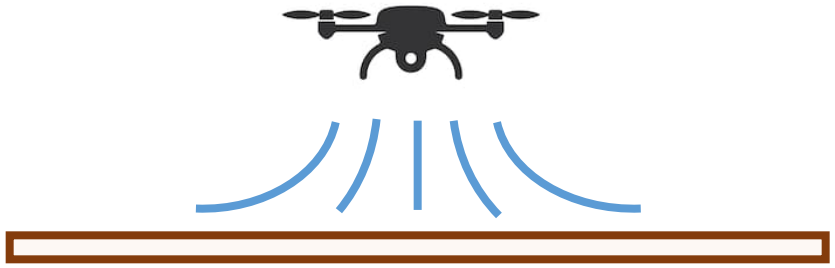
Robust Regression for Safe Exploration in Control
Anqi Liu, Guanya Shi, Soon-Jo Chung, Anima Anandkumar, Yisong Yue, arXiv

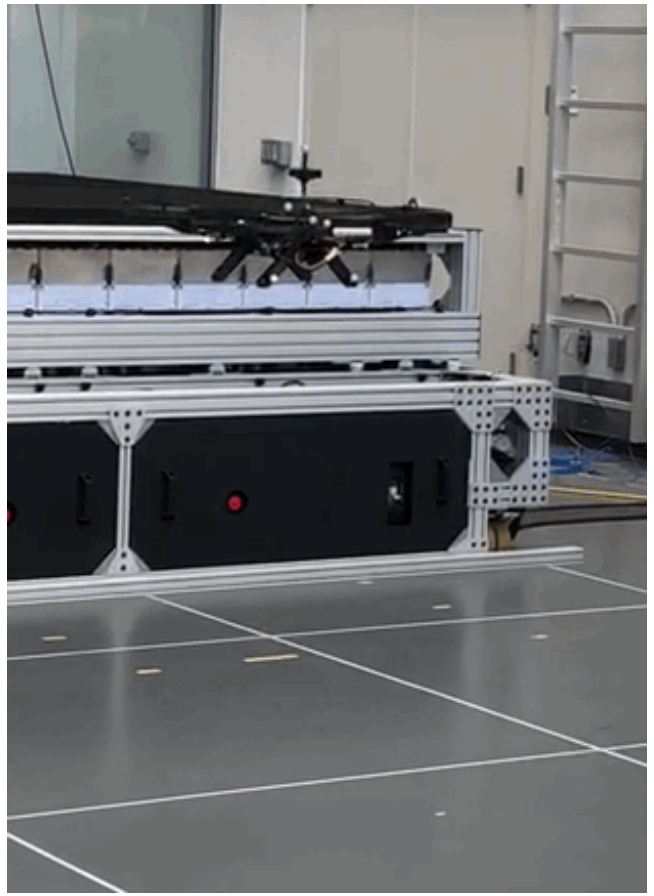
Stable Drone Landing



Guanya Shi

Ground effect

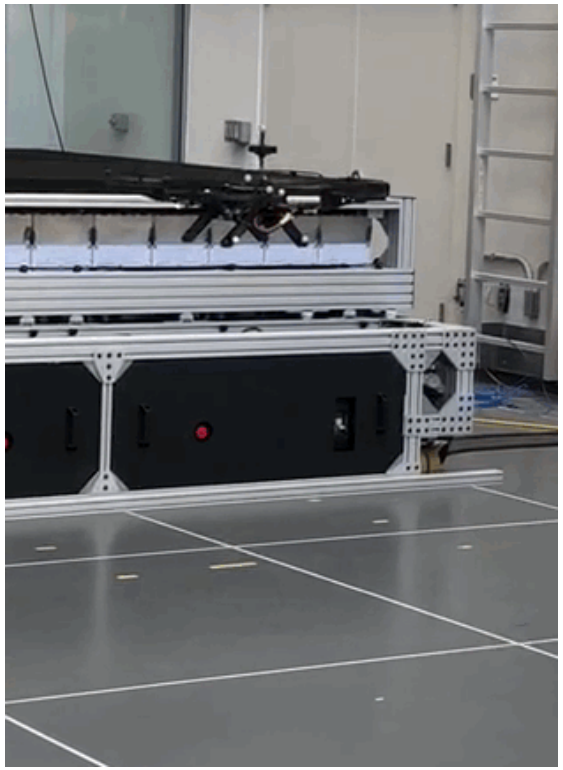




Neural Lander: Stable Drone Landing Control using Learned Dynamics

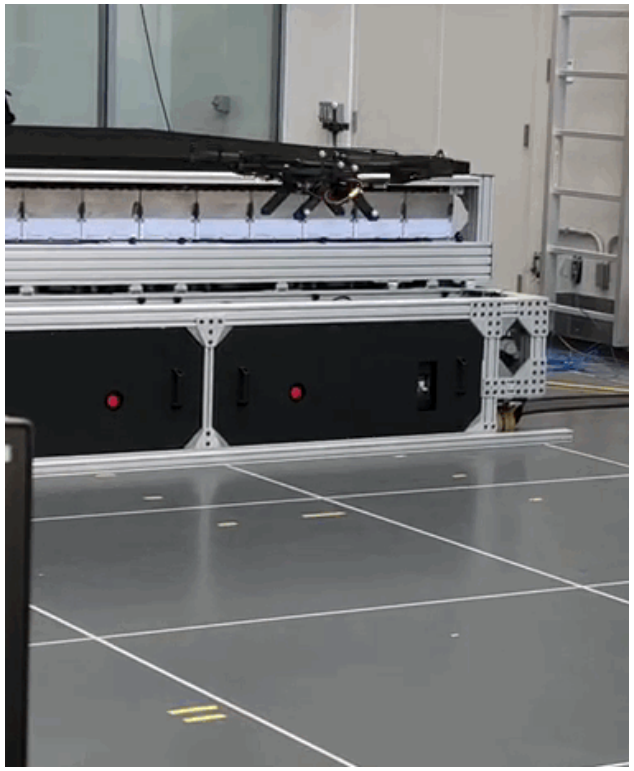
Guanya Shi, Xichen Shi, Michael O'Connell, Rose Yu, Kamyar Azizzadenesheli, Anima Anandkumar, Yisong Yue, Soon-Jo Chung. ICRA 2019

Robust Landing Control (with pre-collected data)



PD

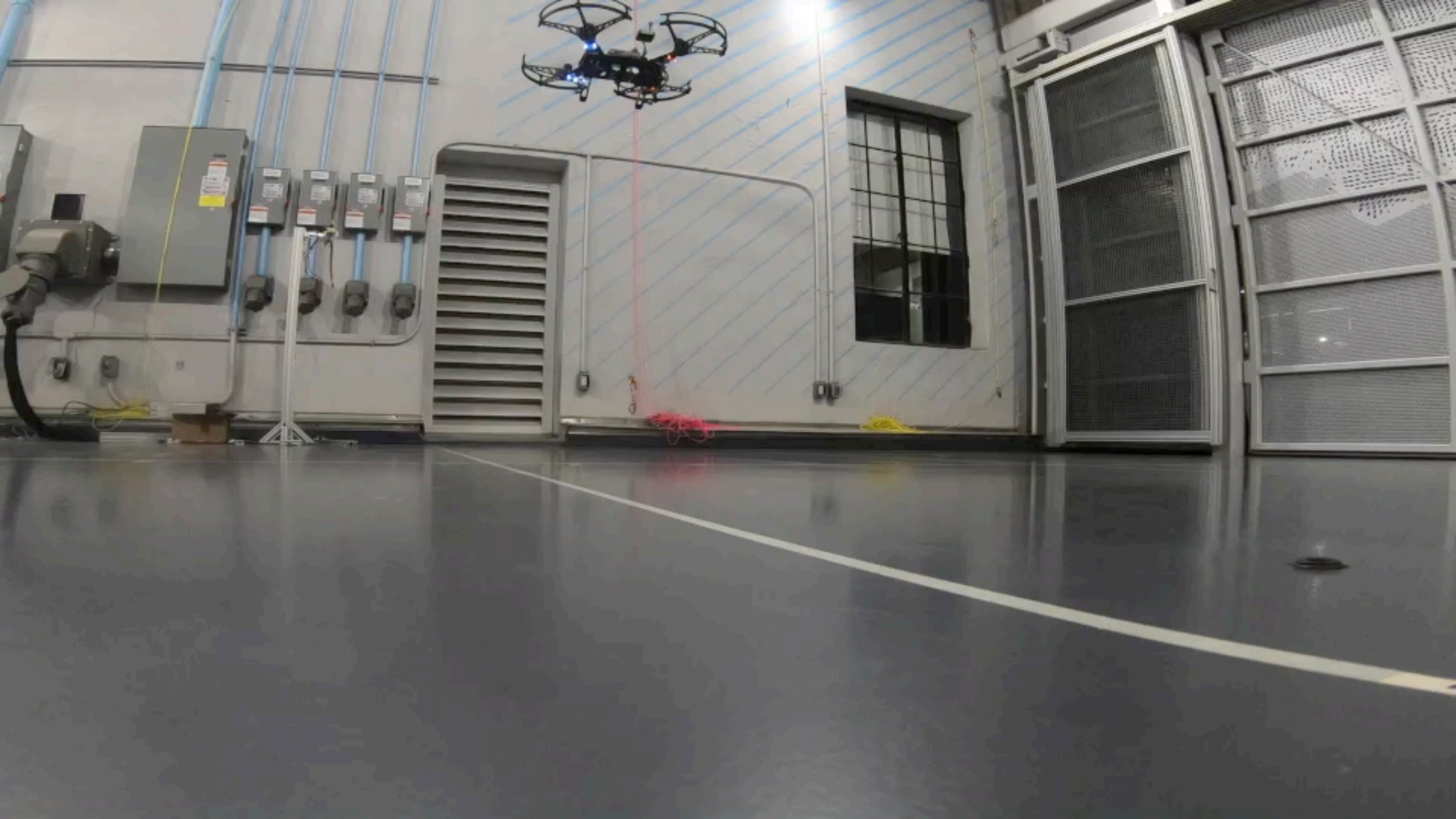
PID



Neural-Lander (PD+Fa)

https://www.youtube.com/watch?v=C_K8MkC_SSQ

Neural Lander: Stable Drone Landing Control using Learned Dynamics Guanya Shi, Xichen Shi, Michael O'Connell, Rose Yu, Kamyar Azizzadenesheli, Anima Anandkumar, Yisong Yue, Soon-Jo Chung. ICRA 2019



Controller Design (simplified)

Guanya Shi

- Nonlinear Feedback Linearization:

$$u_{nominal} = K_s \eta \qquad \eta = \begin{bmatrix} p - p^* \\ v - v^* \end{bmatrix}$$

Desired Trajectory (tracking error)

Stability Guarantee: (simplified)

Time delay

Unmodeled disturbance

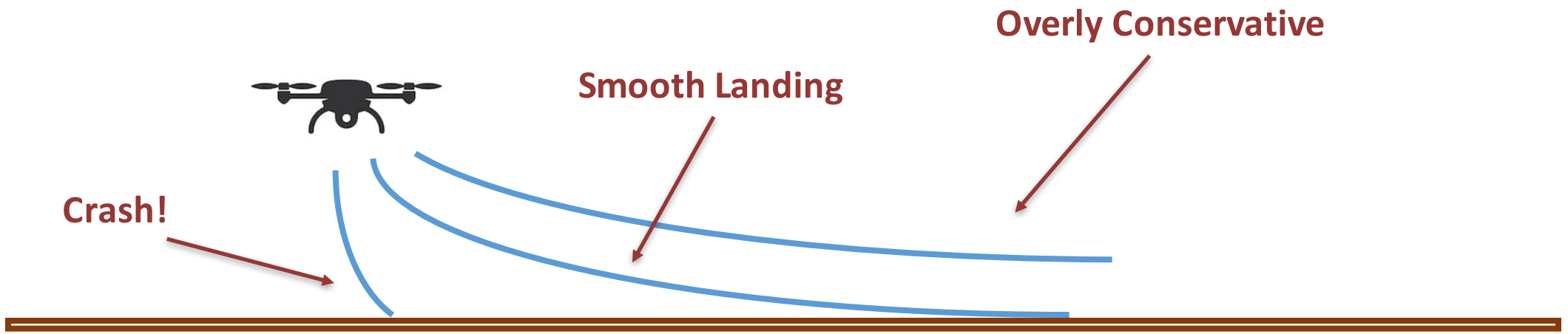
$$\|\eta(t)\| \leq \|\eta(0)\| \exp\left\{\frac{\lambda_{min}(K) - \tilde{L}\rho}{C} t\right\} + \frac{\epsilon}{\lambda_{min}(K) - \tilde{L}\rho}$$

Lipschitz of NN

$$\Longrightarrow \|\eta(t)\| \to \frac{\epsilon}{\lambda_{min}(K) - \tilde{L}\rho}$$

Exponentially fast

Uncertain Dynamics => Uncertain Control



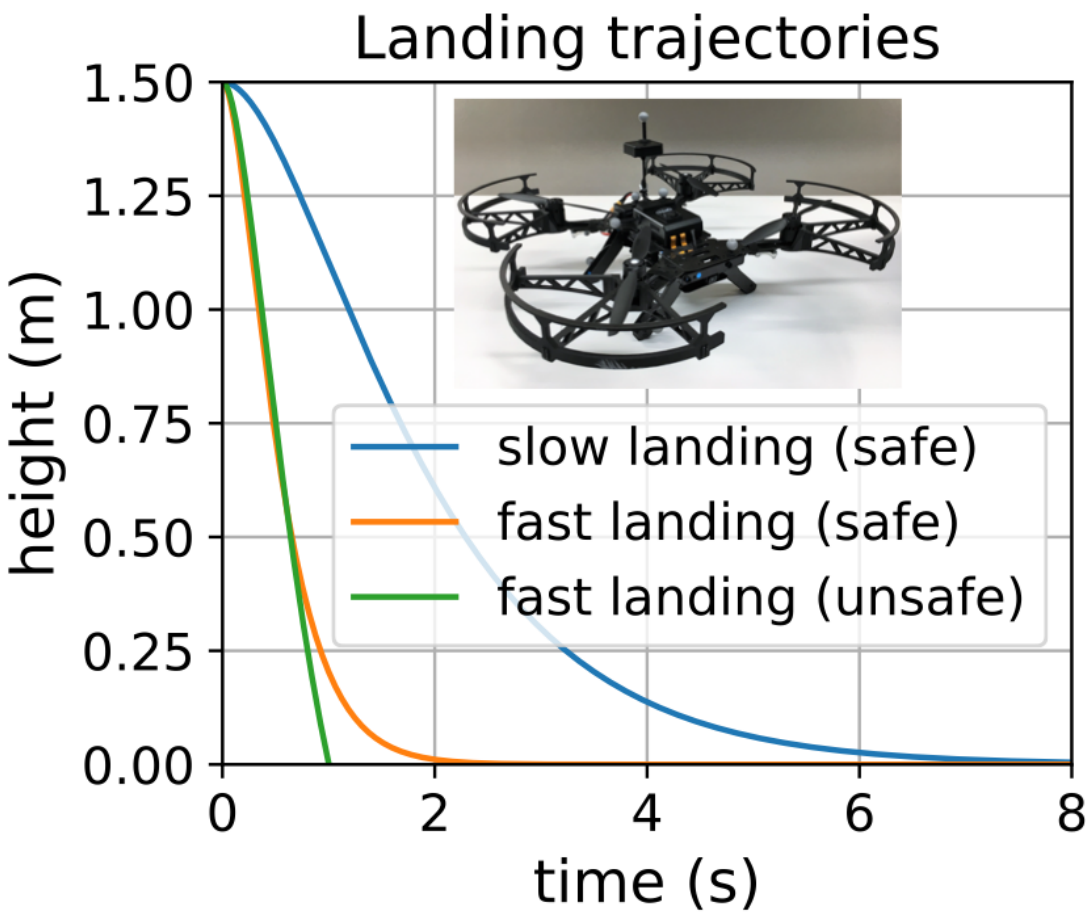
Worst Case Analysis: **we might crash!**

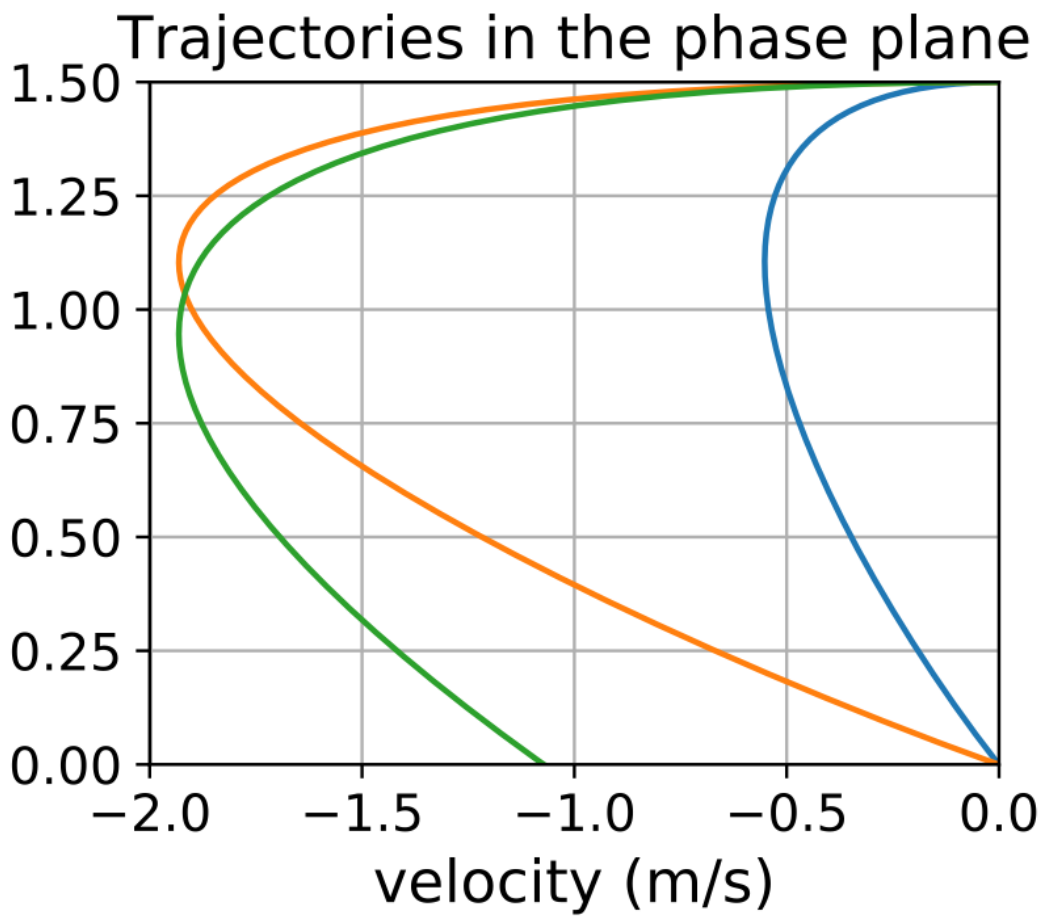
Note: analysis requires propagating uncertainty over time

Robust Regression for Safe Exploration in Control
Anqi Liu, Guanya Shi, Soon-Jo Chung, Anima Anandkumar, Yisong Yue, arXiv

Certify Safety of Landing Trajectories

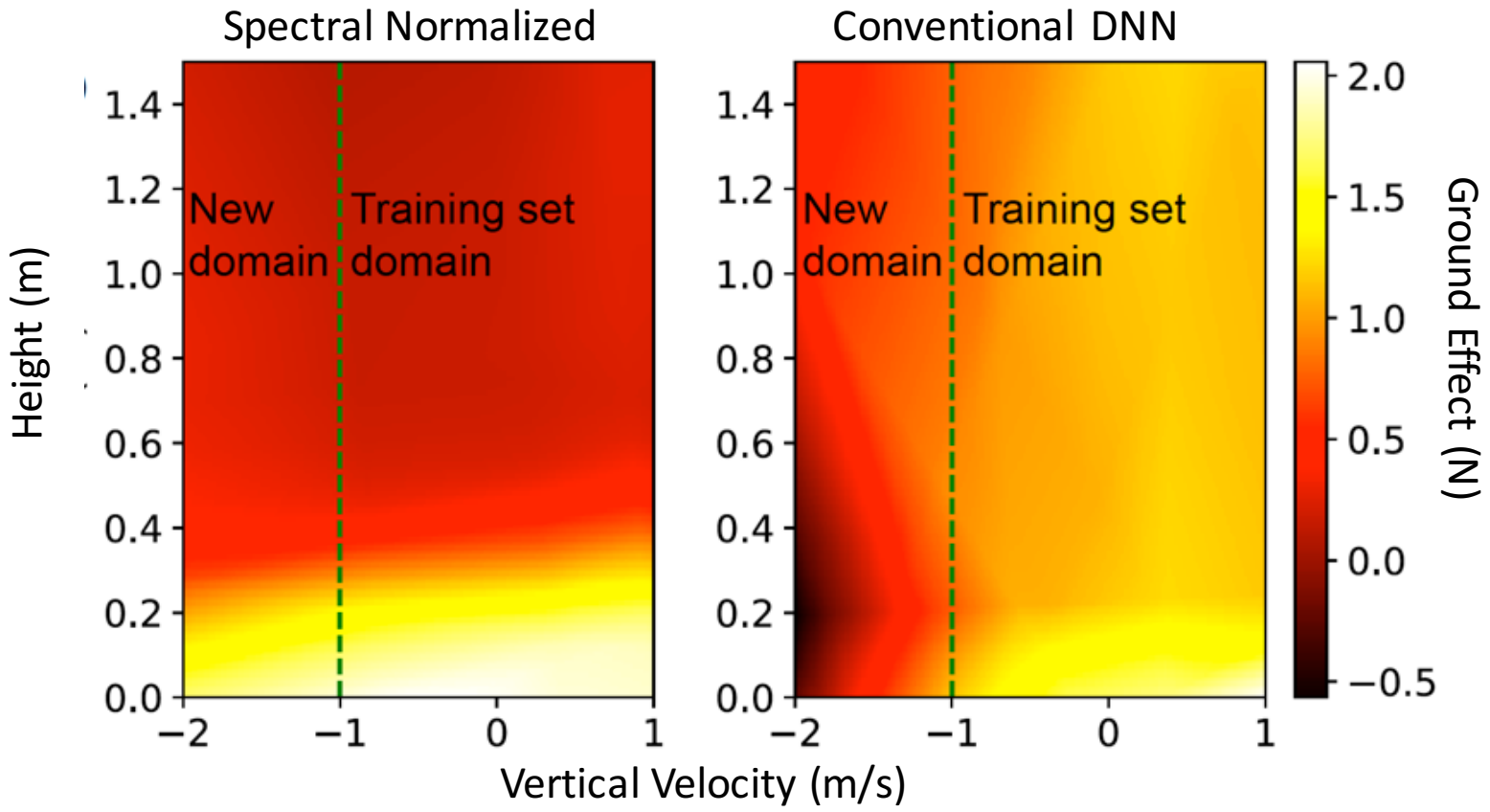
(Adversarial Analysis Based on Current Training Data)





Empirical Visualization

(Pre-Collected Training Data, via Spectral Normalization)



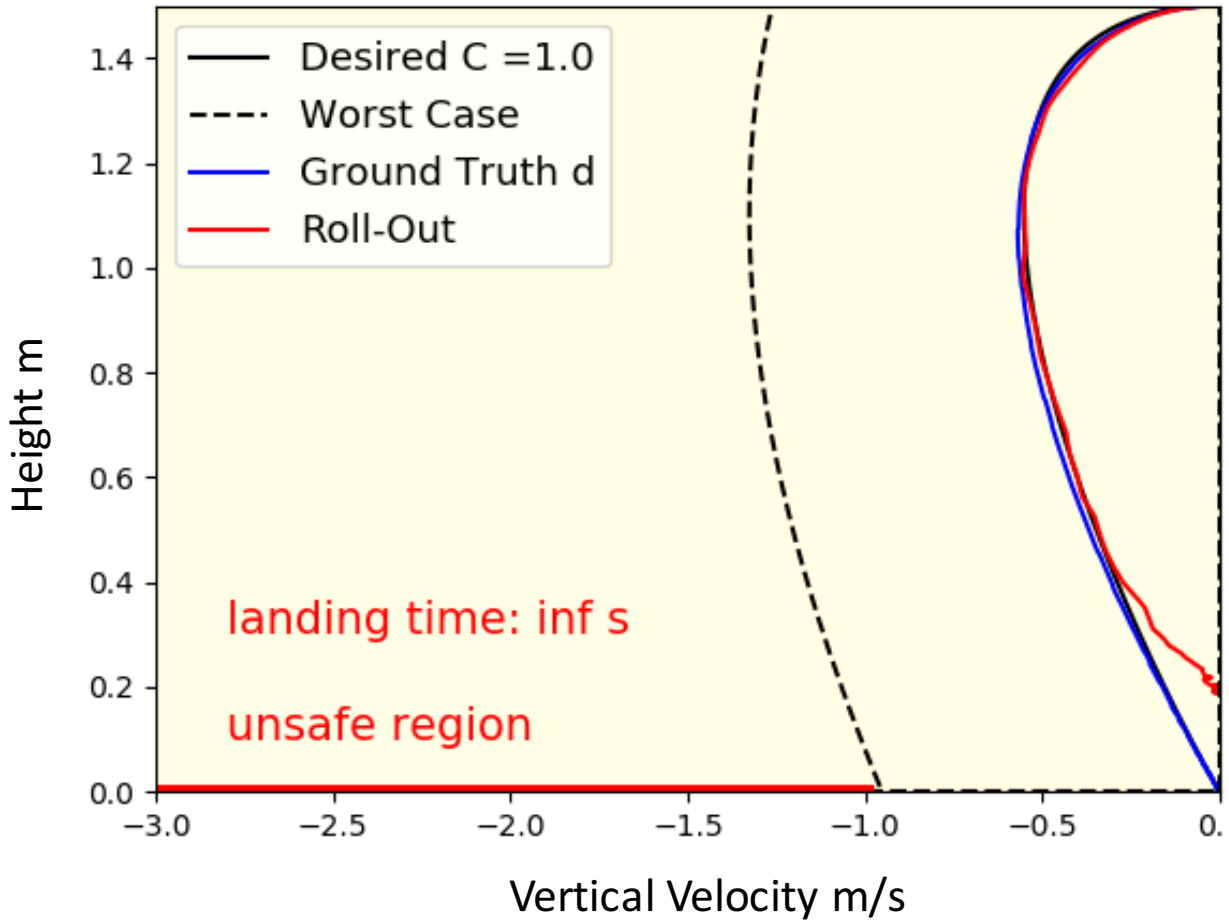
Neural Lander: Stable Drone Landing Control using Learned Dynamics

Guanya Shi, Xichen Shi, Michael O'Connell, Rose Yu, Kamyar Azizzadenesheli, Anima Anandkumar, Yisong Yue, Soon-Jo Chung. ICRA 2019

Adversarial Analysis of Tracking Error

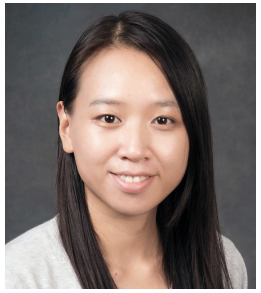
Tracking Error: $\dfrac{\epsilon}{\lambda_{min}(K) - \tilde{L}\rho}$

Need worst-case analysis
on test trajectory!



Robust Regression for Safe Exploration in Control
Anqi Liu, Guanya Shi, Soon-Jo Chung, Anima Anandkumar, Yisong Yue, arXiv

Key Tool: Robust Regression



Angie Liu

- Goal: learn P(y|x) that is robust to target distribution

Chosen Adversarially

$$argmin_\theta \mathbb{E}_{P_{target}(x)} \mathbb{E}_{P(y|x)} L(y, \hat{P}_\theta(y|x))$$

Minimize "surprise"

s.t.

$\hat{P}_\theta(y|x)$ fits training data

Robust Regression for Safe Exploration in Control
Anqi Liu, Guanya Shi, Soon-Jo Chung, Anima Anandkumar, Yisong Yue, arXiv

Robust Regression Guarantee

Corollary 2. [The neural network case] *Let neural networks* $\phi(x) = F_{\mathcal{A}}(x)$ *use* L *fixed nonlinearities* $(\sigma_1, ..., \sigma_L)$*, which are* ρ_i*-Lipschitz and* $\sigma_i(0) = 0$*. Let reference matrices* $(C_1, .., C_L)$ *be given, as well as spectral norm bounds* $(s_i)_{i=1}^L$*, and* l_1 *norm bounds* $(b_i)_{i=1}^L$*. If* $\sqrt{\sum_i ||x_i||_2^2} \leq I$*, for robust regression using network* $F_{\mathcal{A}}$ *with weight matrices* $\mathcal{A} = (A_1, ..., A_L)$ *and maximum dimension of each layer is at most* D *obey* $||A_i||_\sigma \leq s_i$*,* $||A_i^T - C_i^T||_{2,1} \leq b_i$*, and* $||F_{\mathcal{A}}(x)||_2 \leq \mathfrak{X}$*, the following holds with probability at least* $1 - \delta$*,*

$$\mathbb{E}_{P_{trg}(x,y)}[(y - \hat{f}(x))^2]$$

$$\leq W \left[(2RB + \sigma_0^{-2})^{-1} + \lambda + \frac{32A\mathfrak{X}}{Bn^{\frac{3}{2}}} + \frac{288A^2\mathfrak{X}}{nB^2I} \ln n \sqrt{\mathcal{R}_{\mathcal{A}} \ln(2D^2)} + \frac{3A^2\mathfrak{X}^2}{B^2} \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \right], \quad (7)$$

where $\mathcal{R}_{\mathcal{A}}$ *is the spectral complexity of networks* $F_{\mathcal{A}}(x)$*,* $\mathcal{R}_{\mathcal{A}} := \left(\prod_{j=1}^L s_j^2 \rho_j^2 \right) \left(\sum_{i=1}^L (\frac{b_i}{s_i})^{\frac{2}{3}} \right)^3$*; The corresponding perturbation bounds for spectral normalized deep neural networks is,*

Robustness under perturbation →

$$\sup_{x \in \mathbb{B}(\epsilon), y \sim f(x)} [(y - \hat{f}(x))^2] \leq ((2RB + \sigma_0^{-2})^{-1/2} + \sqrt{\lambda}) + \left(L + \frac{A}{B} \mathcal{R}_{\mathcal{A}} \right) ||\epsilon||)^2 \quad (8)$$

Robust Regression for Safe Exploration in Control

Anqi Liu, Guanya Shi, Soon-Jo Chung, Anima Anandkumar, Yisong Yue, arXiv

Integration with Control

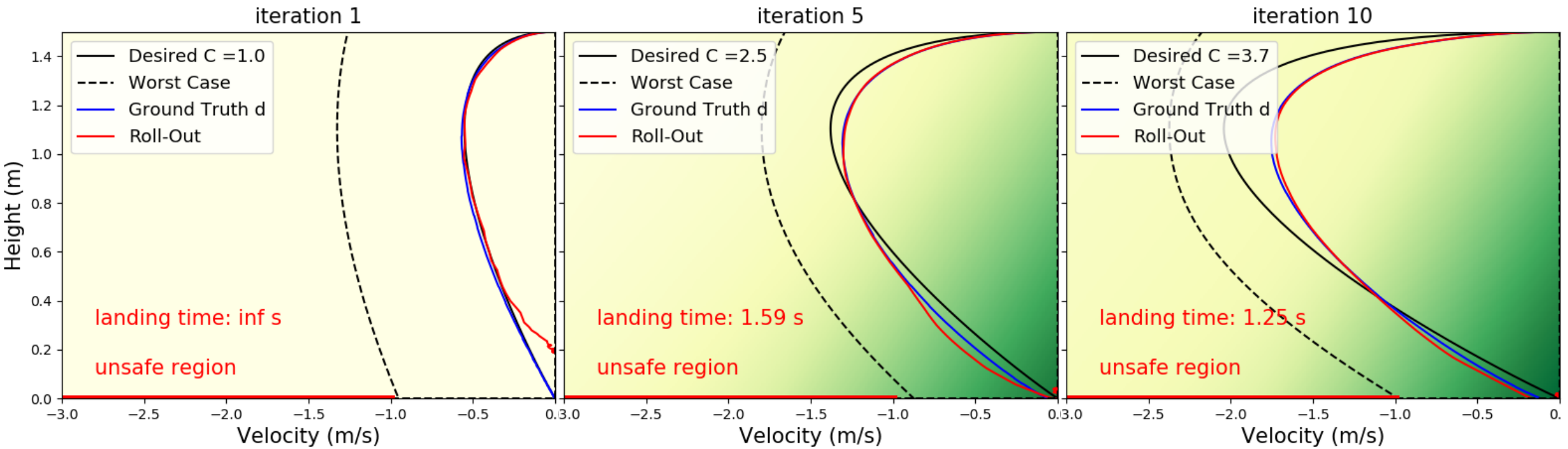
Theorem 2. *Suppose x is in some compact set $\mathcal{X}$, and $\epsilon_m = \sup_{x\in\mathcal{X}} \|\epsilon\|$. Then $\tilde{x}$ will exponentially converge to the following ball: $\lim_{t\to\infty} \|\tilde{x}(t)\| = \gamma \cdot \epsilon_m$, where*

$$\gamma = \frac{\lambda_{\max}(M)}{\lambda_{\min}(K)\lambda_{\min}(M)} \sqrt{\left(\frac{1}{\lambda_{\min}(\Lambda)}\right)^2 + \left(1 + \frac{\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Lambda)}\right)^2}. \tag{12}$$

Worst-case uncertainty in realized trajectory

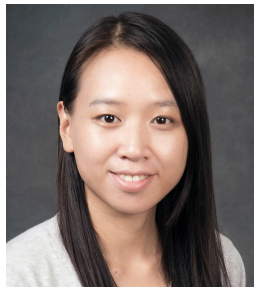
Robust Regression for Safe Exploration in Control
Anqi Liu, Guanya Shi, Soon-Jo Chung, Anima Anandkumar, Yisong Yue, arXiv

Results



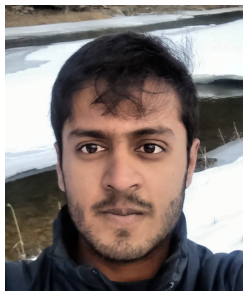
Robust Regression for Safe Exploration in Control

Anqi Liu, Guanya Shi, Soon-Jo Chung, Anima Anandkumar, Yisong Yue, arXiv



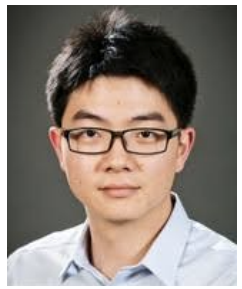
Angie Liu

Guanya Shi

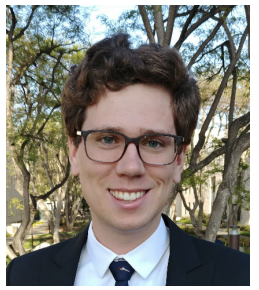


Sumanth Dathathri

Stephan Zheng

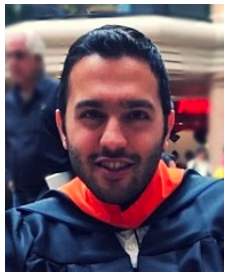


Xichen Shi

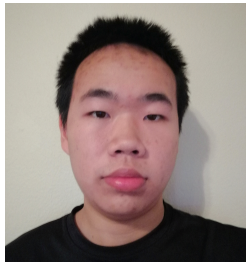


Michael O'Connell

Rose Yu



Kamyar Azizzadenesheli



Tianwei Yin

Anima Anandkumar

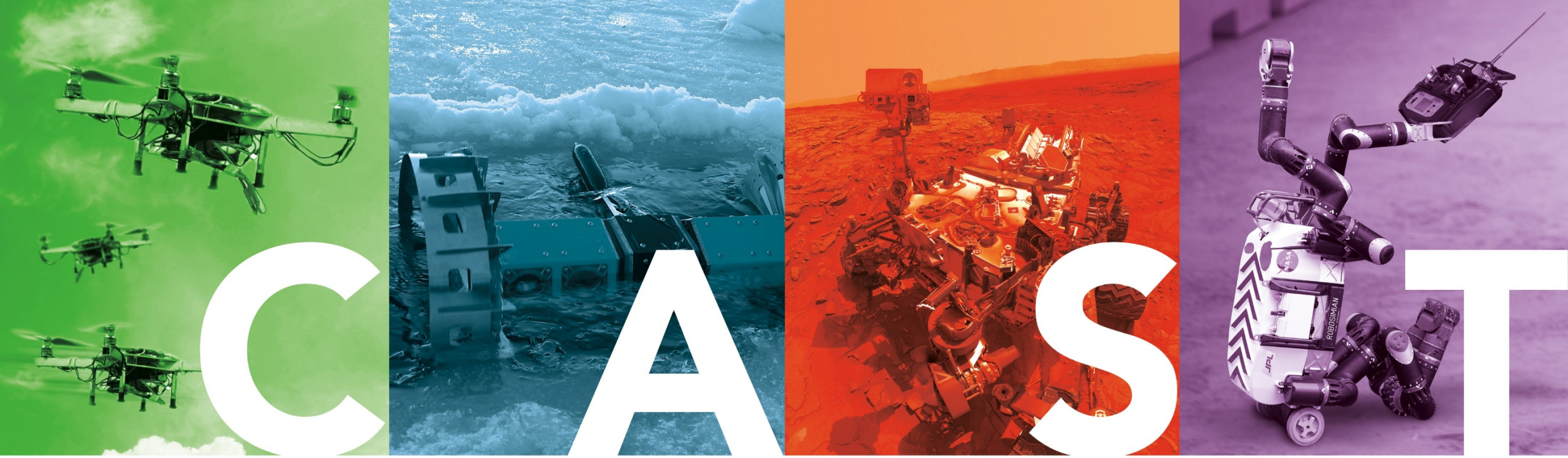
Soon-Jo Chung

Richard Murray

Detecting Adversarial Examples via Neural Fingerprinting,
Sumanth Dathathri, Stephan Zheng, Tianwei Yin, Yisong Yue, Richard M. Murray, arXiv.

Neural Lander: Stable Drone Landing Control using Learned Dynamics
Guanya Shi, Xichen Shi, Michael O'Connell, Rose Yu, Kamyar Azizzadenesheli, Anima Anandkumar, Yisong Yue, Soon-Jo Chung. ICRA 2019

Robust Regression for Safe Exploration in Control
Anqi Liu, Guanya Shi, Soon-Jo Chung, Anima Anandkumar, Yisong Yue, arXiv



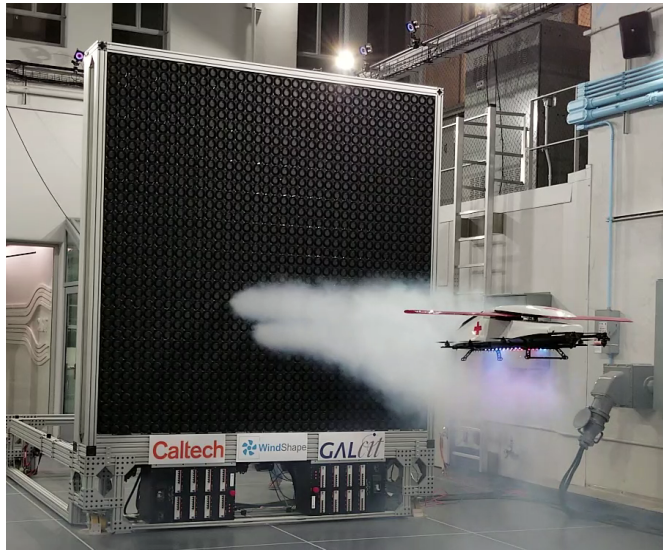
Center for Autonomous Systems and Technologies

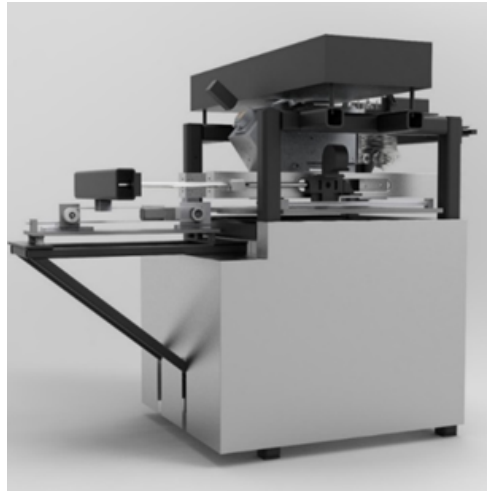
A New Vision for Autonomy

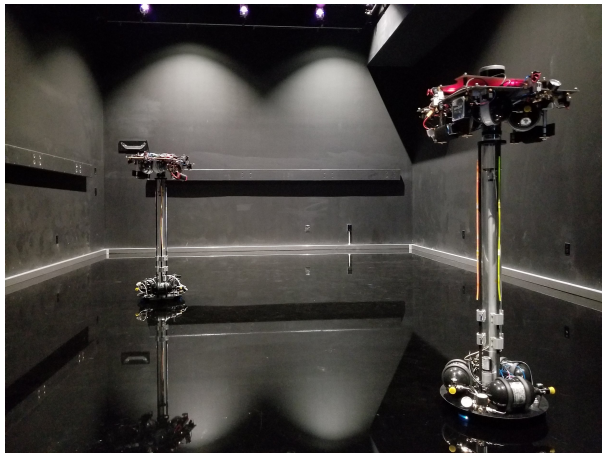
Caltech

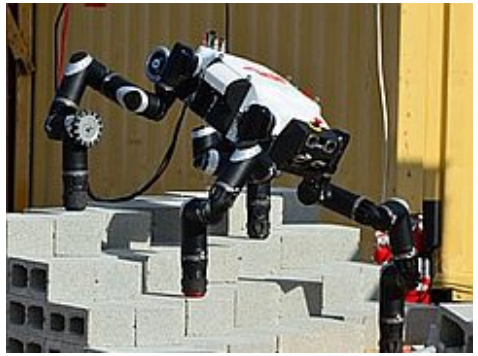
http://cast.caltech.edu

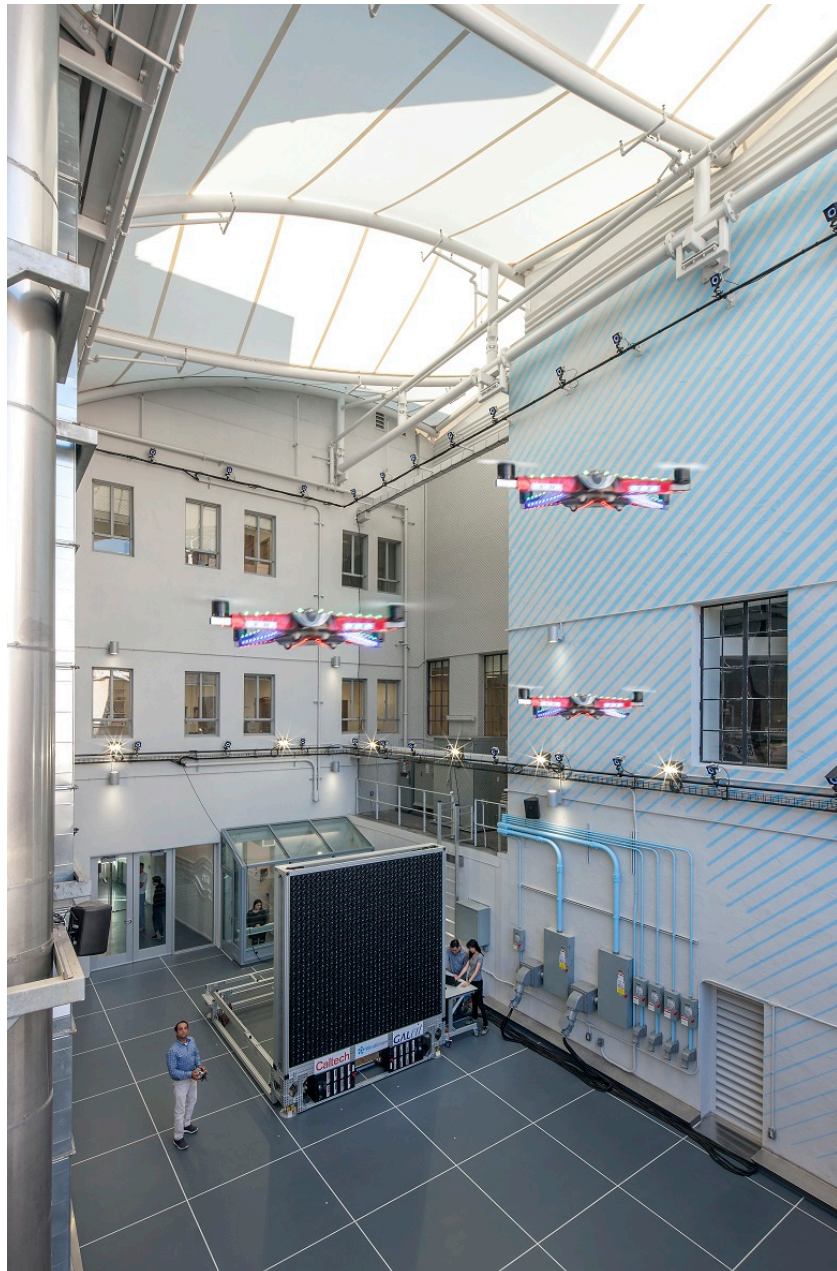
Autonomous Dynamic Robots

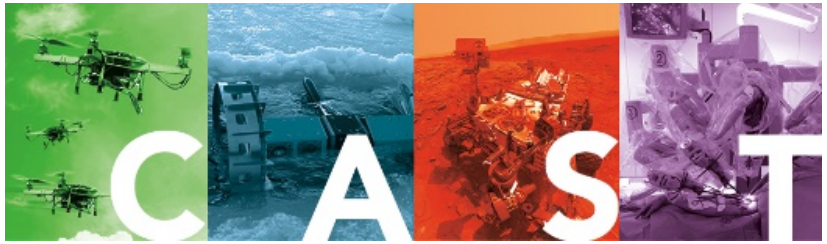










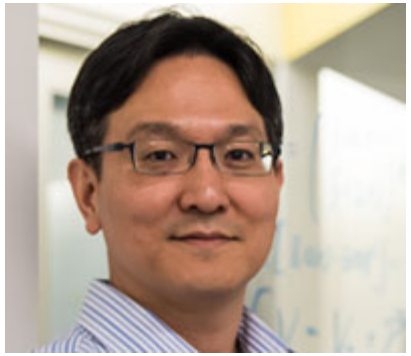


http://cast.caltech.edu

Postdoc Openings!

(applications considered starting January)

Mory Gharib



Soon-Jo Chung

Aaron Ames

Anima Anandkumar

Yisong Yue

Joel Burdick

Katie Bouman

Pietro Perona