Arizona State University

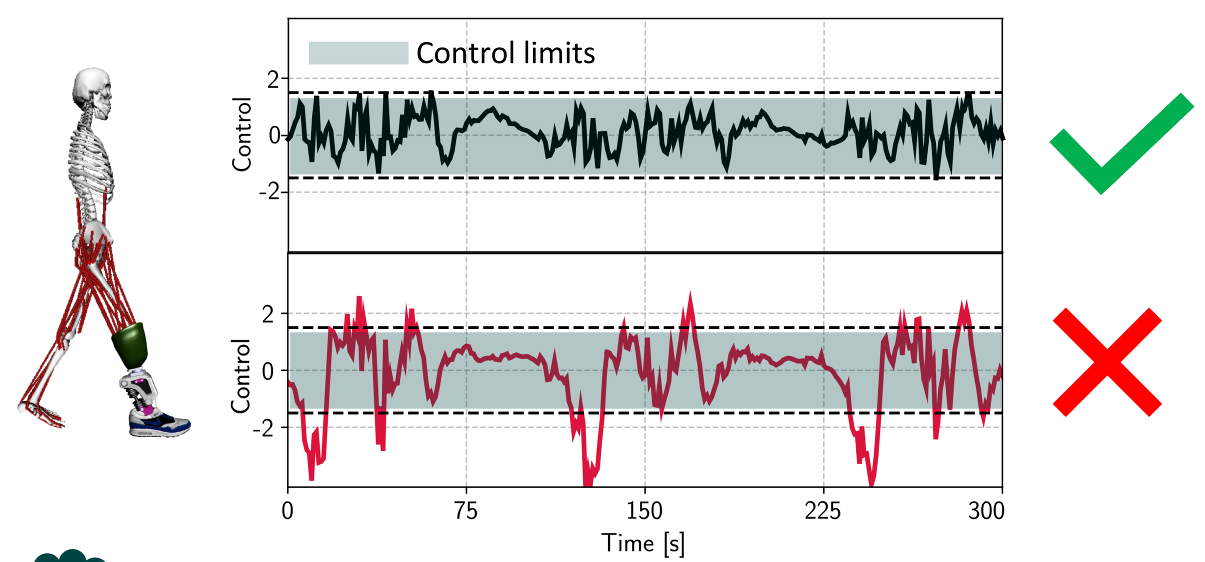
Enhancing Neural Networks for Safety in Robot Learning: Direct Weight Modification and Expansion

Keyvan Majd, Georgios Fainekos, Heni Ben Amor

2023 IROS
OCTOBER 1 - 5, 2023
IEEE/RSJ International Conference on Intelligent Robots and Systems
Workshop on Formal methods techniques in robotics systems: Design and control

DNNs in Assistive Robotics

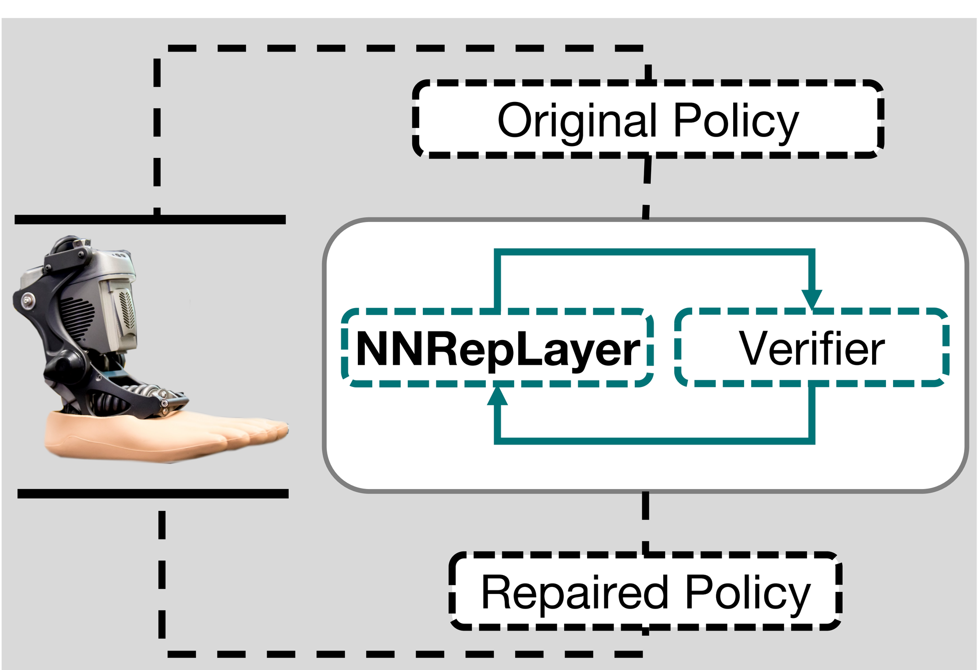
DNN policies may demonstrate unsafe behavior in formerly unseen scenarios



NN Repair in Assistive Robotics

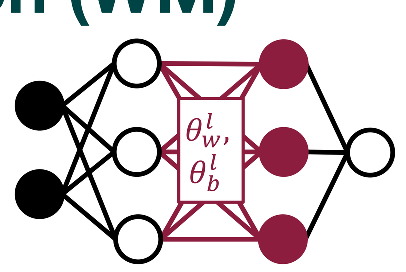
- Model the safety constraints as hard constraints on the networks' output
- Ensure the satisfaction of safety properties through global optimization as gradient descent approaches cannot provide any guarantees

Overview of Our Framework



Weight Modification (WM)

Given
- The repair set of samples $\mathcal{X}_r$
- Fixed weights $\{\theta^i\}_{i=l+1}^{L+1}$



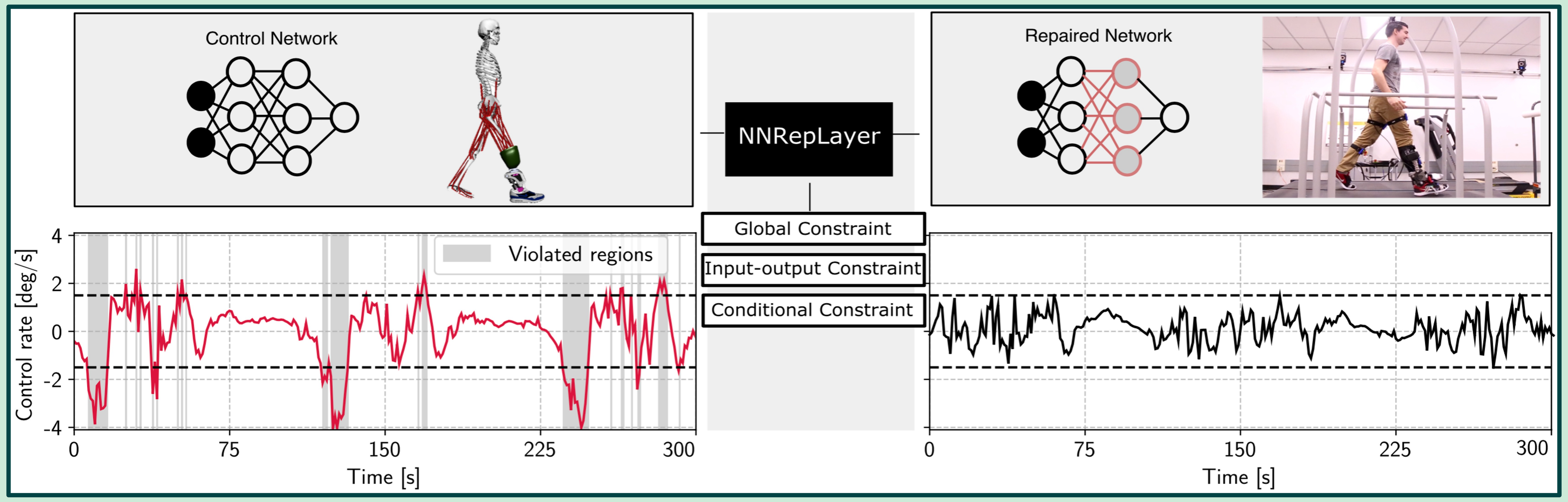
$$\min_{\delta,\theta_w^l,\theta_b^l,\{x^i\}_{i=l}^{L+1}} E(\theta_w^l,\theta_b^l)+\delta, \quad \triangleright \text{ Quadratic loss function}$$

s.t.

$$\Psi(y,x^0), \text{ for } x^0\in\mathcal{X}_r \quad \triangleright \text{ Safety predicate}$$

$$x^i = R(\theta_w^i x^{i-1}+\theta_b^i), \text{ for } \{i\}_{i=l}^{L+1} \quad \triangleright \text{ Forward pass}$$

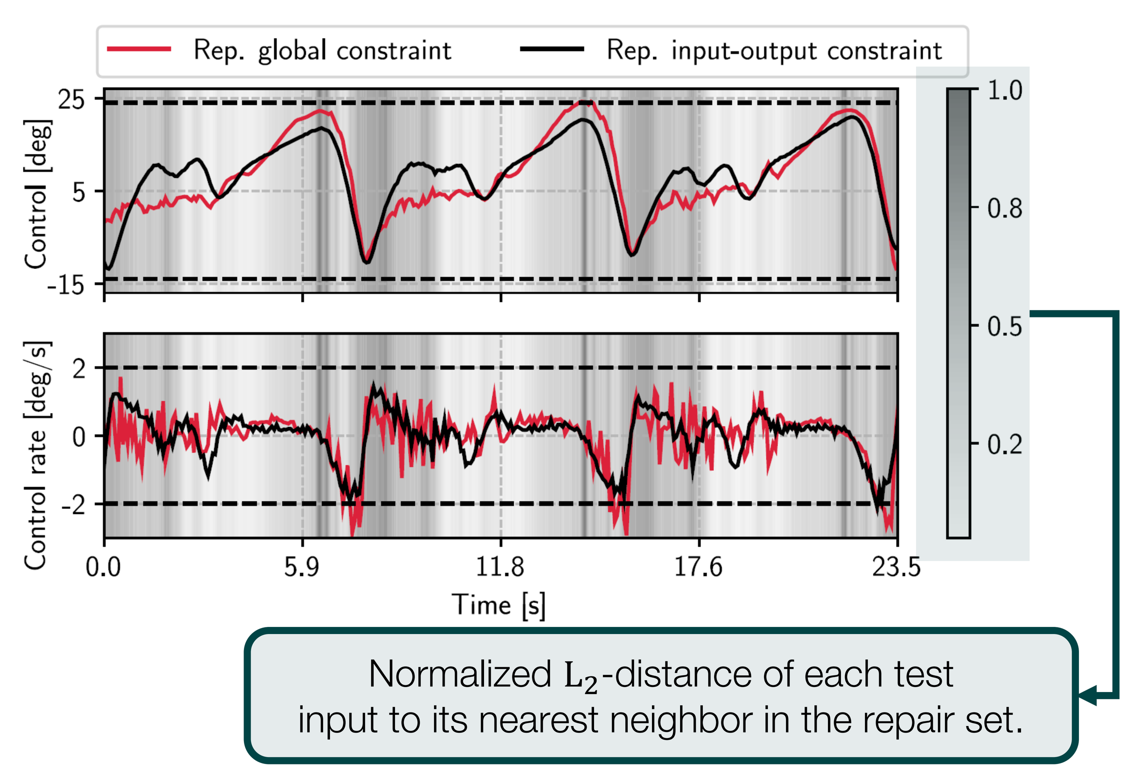
$$\delta \ge \|\theta^l-\theta^{l,init}\|_\infty \ge 0. \quad \triangleright \text{ Bounded weight error}$$



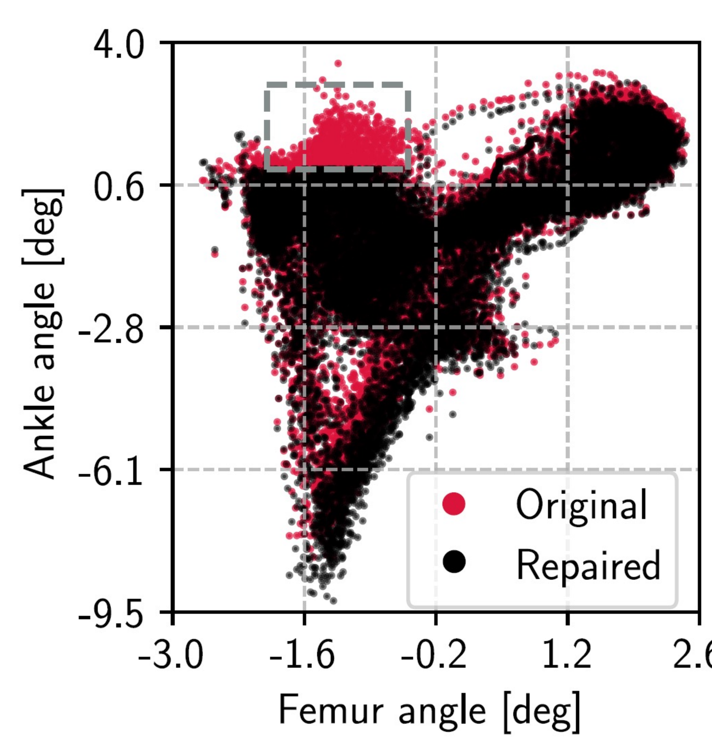
Unified repair and verification formally **guarantees the safety** of policy networks, **minimally deviates** from the original performance of network, and **does not require the modification of the training data** to learn safety.

Main Results (WM)

Real walking tests

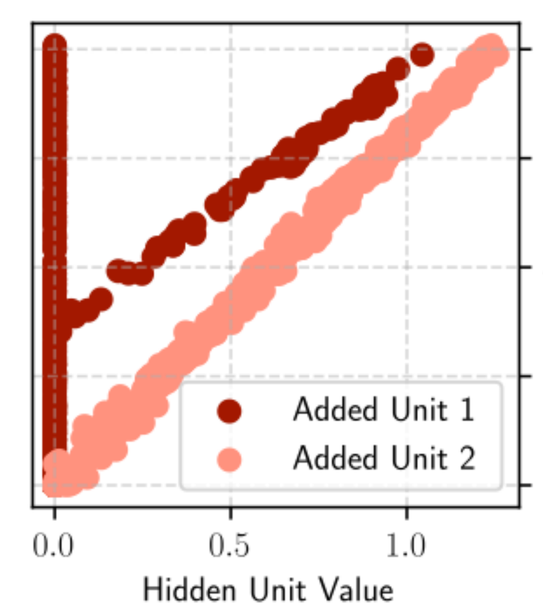


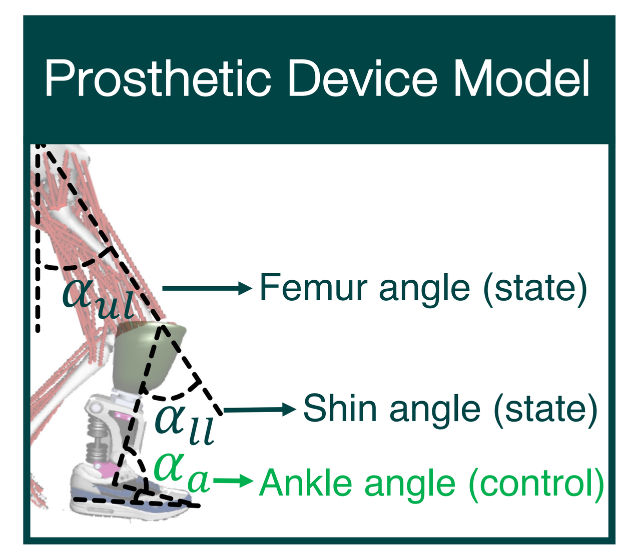
Conditional Constraint:



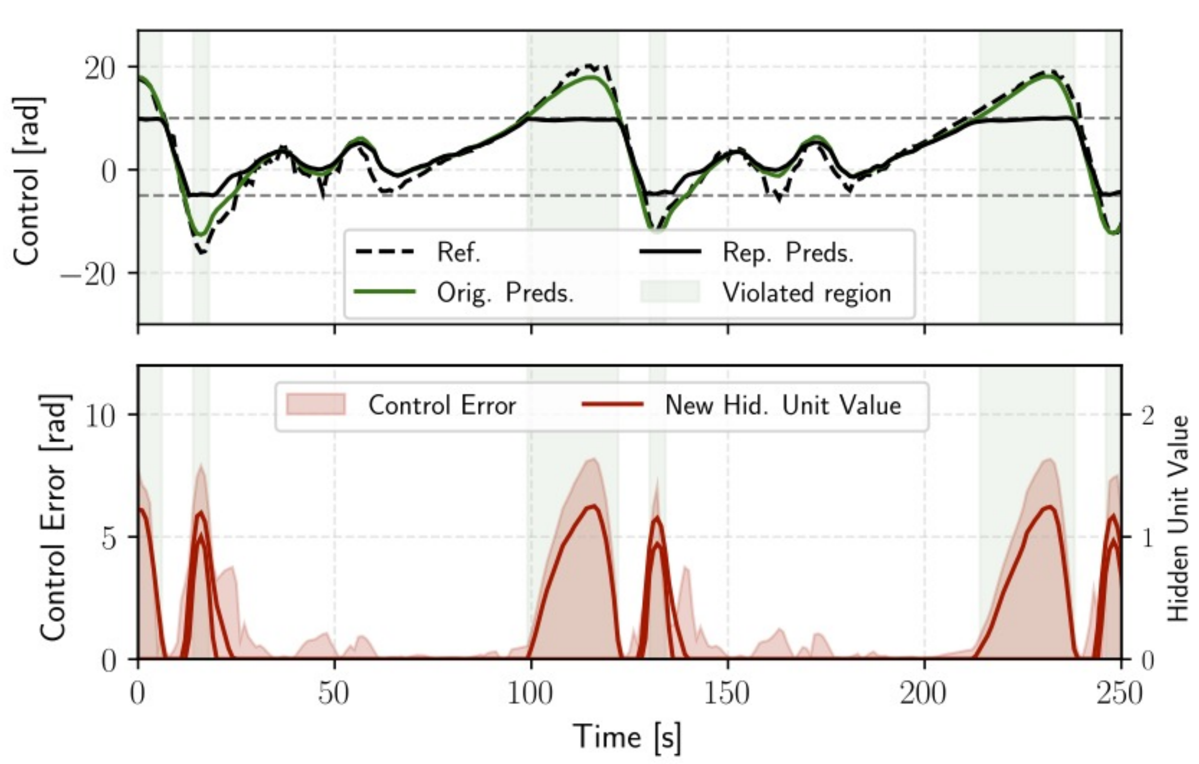
Main Results (DE)

Newly Added Nodes Activations:





Bounding Constraint:

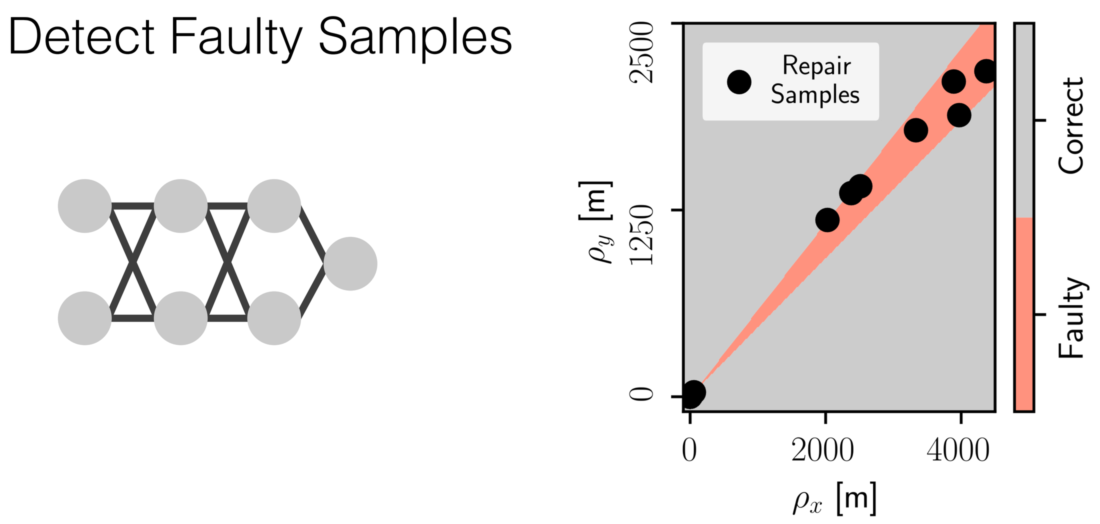


Scalability Challenges

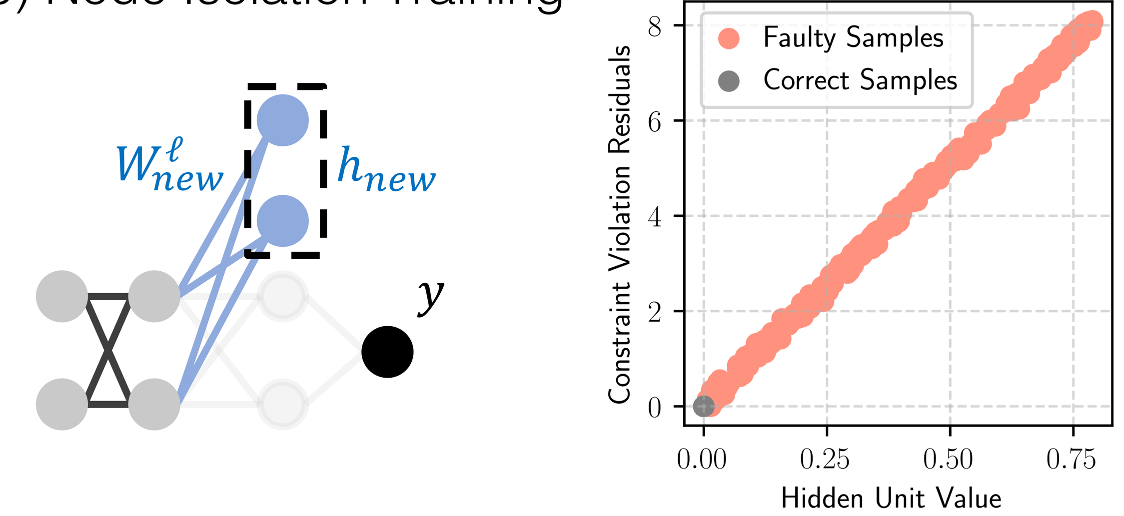
- Solving MIQP is a demanding process that scales with the size of network.
- Provable repair only restricted to DNNs with linear piece-wise continuous activations

DNN Expansion (DE)

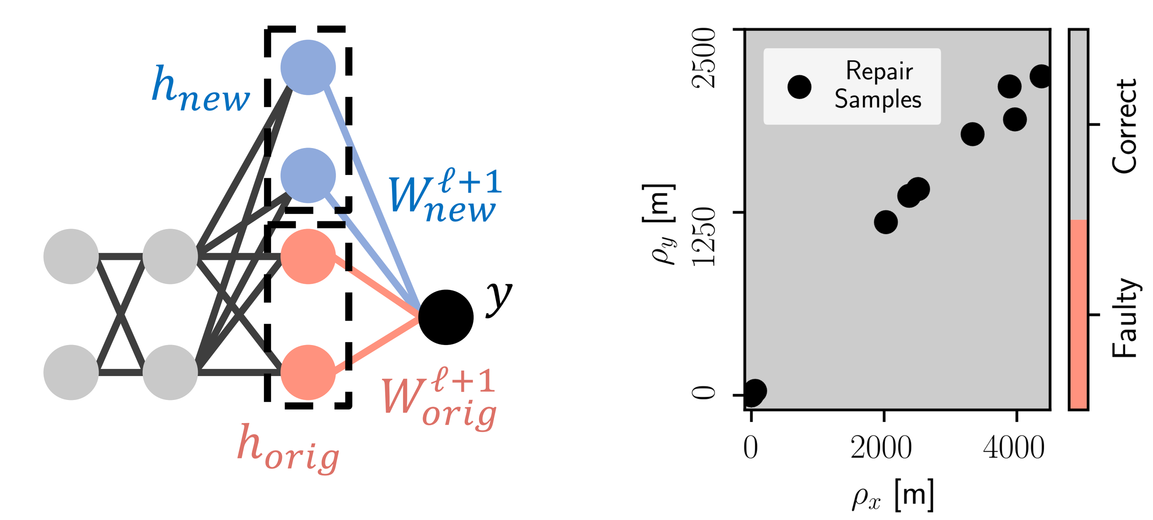
(a) Detect Faulty Samples



(b) Node Isolation Training



(c) QP-tuned Output Layer



Main Findings

Our method,
- Guarantees the satisfaction of constraints for the adversarial samples
- Offers a natural way to extend the network using the same activation functions as the other neurons
- Activates within the faulty region while preserving the original performance in the unaffected regions