# Efficient Global Robustness Certification of Neural Networks via Interleaving Twin-Network Encoding (Extended Abstract) *

**Zhilu Wang**[1] , **Chao Huang**[2] , **Qi Zhu**[1]

[1]Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA
[2]Department of Computer Science, University of Liverpool, Liverpool, UK
zhilu.wang@u.northwestern.edu, chao.huang2@liverpool.ac.uk, qzhu@northwestern.edu

## Abstract

The robustness of deep neural networks in safety-critical systems has received significant interest recently, which measures how sensitive the model output is under input perturbations. While most previous works focused on the *local robustness* property, the studies of the *global robustness* property, i.e., the robustness in the entire input space, are still lacking. In this work, we formulate the global robustness certification problem for ReLU neural networks and present an efficient approach to address it. Our approach includes a novel interleaving twin-network encoding scheme and an over-approximation algorithm leveraging relaxation and refinement techniques. Its timing efficiency and effectiveness are evaluated and compared with other state-of-the-art global robustness certification methods, and demonstrated via case studies on practical applications.

## 1 Introduction

Deep neural networks (DNNs) could be vulnerable to small adversarial perturbations on their inputs [Biggio *et al.*, 2013]. The formally-defined *robustness* metric of a DNN tries to bound such uncertain behavior, measuring how much the network's output may deviate when its input has a bounded perturbation. The *local robustness* problem has been extensively studied, with formal methods developed to bound the output range for a bounded disturbance around *a given input* [Katz *et al.*, 2017; Singh *et al.*, 2019; Huang *et al.*, 2020b; Zhang *et al.*, 2018; Wang *et al.*, 2021a]. However, it is hard to apply these techniques in safety verification of a dynamic system (e.g., an autonomous vehicle) [Zhu *et al.*, 2020; Wang *et al.*, 2021d; Liu *et al.*, 2022] as we will need to conduct local robustness analysis *during runtime* for each input sample that the system encounters or may encounter, and they are typically too computationally expensive for that.

This challenge motivates us to address the safety of DNN-enabled dynamic systems by considering the problem of *global robustness*, which measures the worst-case DNN output deviation against bounded perturbation for *all possible input values*. We can conduct such worst-case analysis offline, decoupling it from the safety verification [Huang *et al.*, 2019; Fan *et al.*, 2020; Huang *et al.*, 2022] by applying the maximum deviation for all possible inputs. However, the previous MILP (mixed-integer linear programming) or SMT-based techniques for global robustness analysis [Katz *et al.*, 2017; Chen *et al.*, 2021] are still too complex for DNNs in practical systems, even in offline computation. Various approaches were proposed to tackle the complexity challenge, but they are either still computationally expensive, e.g., with region-based robustness analyses [Gopinath *et al.*, 2018; Mangal *et al.*, 2019], or lack the deterministic guarantees, e.g., with sampling-based techniques [Ruan *et al.*, 2019; Bastani *et al.*, 2016; Mangal *et al.*, 2019].

In this work, we propose an efficient certification approach to *over-approximate* the global robustness. Our approach introduces a novel network encoding structure, namely interleaving twin-network encoding, to compare two copies of the neural network side-by-side under different inputs, with extra interleaving dependencies added between them to improve efficiency. Our approach also includes over-approximation techniques based on network decomposition and LP (linear programming) relaxation, to further reduce the computation complexity. To the best of our knowledge, our approach is the *first global robustness over-approximation method that certifies the robustness among the entire input domain with sound and deterministic guarantee*. Experiments show that our approach is much more efficient and scalable than the exact global robustness methods such as Reluplex [Katz *et al.*, 2017], with tight over-approximation. A case study of close-loop control system safety verification with perception DNN component further demonstrates the potential of our approach in practical systems.

## 2 Global Robustness Certification

### 2.1 Problem Formulation

An $n$-layer neural network $F : \mathbb{R}^{m_0} \to \mathbb{R}^{m_n}$ maps input $x^{(0)} \in \mathbb{R}^{m_0}$ into output $x^{(n)} \in \mathbb{R}^{m_n}$. The output of layer $i$ is denoted as $x^{(i)} \in \mathbb{R}^{m_i}$. The mapping between two consecutive layers $x^{(i-1)}$ and $x^i$ is composed with a linear transformation $y^{(i)} = W^{(i)}x^{(i-1)} + b^{(i)}$ and (optionally) a ReLU

[*] This is an extended abstract of a paper that won the Best Paper Award at *Design, Automation and Test in Europe Conference (DATE 2022).*

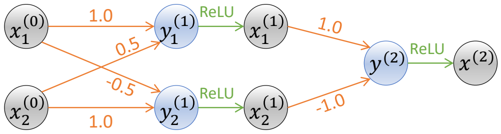


Figure 1: An example neural network.

activation function, where the linear transformation result is denoted as variable $y^{(i)} \in \mathbb{R}^{m_i}$. An illustrating example of a 2-layer neural network is shown in Fig. 1. For simplicity, all bias terms $b^{(i)}$ are 0.

As defined in [Katz *et al.*, 2017; Wang *et al.*, 2021d], the neural network global robustness measures the worst-case output variation when there is a small input perturbation for any possible input sample in the entire input domain $X$.

**Definition 1** (Global Robustness)**.** *The $j$-th output of a neural network $F$ is $(\delta, \varepsilon)$-globally robust in the input domain $X$ iff*

$$\forall x^{(0)}, \hat{x}^{(0)} \in X, \|\hat{x}^{(0)} - x^{(0)}\|_\infty \leq \delta \implies |\hat{x}_j^{(n)} - x_j^{(n)}| \leq \varepsilon,$$

*where $x^{(n)} = F(x^{(0)})$ and $\hat{x}^{(n)} = F(\hat{x}^{(0)})$.*

In this work, we tackle the problem of measuring *how robust a neural network is*, as formally defined in Problem 1:

**Problem 1.** *For a neural network $F$, given an input perturbation bound $\delta$, determine the minimal output variation bound $\varepsilon$ such that $F$ is guaranteed to be $(\delta, \varepsilon)$-globally robust.*

[Katz *et al.*, 2017] proposes to solve this problem by encoding two copies of the neural network side by side, as illustrated in the left part of Fig. 2. $\hat{x}^{(0)}$ represents a perturbed input of $x^{(0)}$, by the bounded perturbation $\Delta x^{(0)}$. And $\varepsilon$ will be the bound of output distance $\Delta x^{(n)}$. Under this encoding, Problem 1 can be formulated as an optimization problem:

$$\begin{aligned} \varepsilon := \max \quad & |\hat{x}^{(n)} - x^{(n)}|, \\ \text{s.t.} \quad & \hat{x}^{(n)} = F(\hat{x}^{(0)}),\ x^{(n)} = F(x^{(0)}), \\ & \hat{x}^{(0)}, x^{(0)} \in X,\ \|\hat{x}^{(0)} - x^{(0)}\|_\infty < \delta. \end{aligned} \tag{1}$$

Note that Eq. (1) can be solved with MILP by introducing a binary variable for each ReLU activation [Cheng *et al.*, 2017], but the complexity is too high to be scalable. To overcome this, we present a new interleaving twin-network encoding (ITNE) scheme with two approximation techniques to efficiently find an over-approximated solution $\bar{\varepsilon} \geq \varepsilon$.

## 2.2 Interleaving Twin-Network Encoding

In this work, we design the interleaving twin-network encoding (ITNE) as shown in the right side of Fig. 2. Compared with the basic twin-network encoding (BTNE) [Katz *et al.*, 2017] (the left side of Fig. 2), besides the connections between the input and output layers, interleaving connections are added for all hidden neurons between the two network copies. Specifically, for each neuron $x = relu(y)$, two variables, $\Delta y = \hat{y} - y$ and $\Delta x = \hat{x} - x$, are added to encode the distance of $y$ and $x$ between the two copies. These distance variables reflect the hidden neuron variation caused by the perturbation. These changes enable the usage of the over-approximation techniques introduced below.

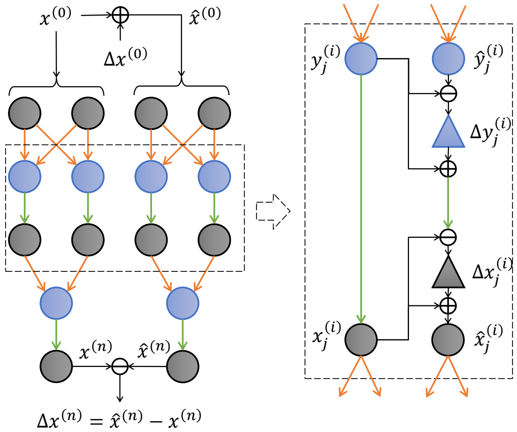


Figure 2: *Left:* The basic twin-network encoding (BTNE) for global robustness certification. *Right:* the neuron-level interleaving twin-network encoding (ITNE) built upon the basic structure, where the hidden layer neurons are connected between the two copies with distance variables $\Delta y_j^{(i)}$ and $\Delta x_j^{(i)}$.

## 2.3 Over-Approximation Techniques

Leveraging ITNE, we design two over-approximation techniques, network decomposition (ND) and LP relaxation (LPR), to improve the global robustness certification efficiency. Inspired by the local robustness certification work in [Huang *et al.*, 2020b], our ND and LPR techniques are specifically designed for global robustness.

### ITNE-Based Network Decomposition (ND)

The main idea of ND is to divide a neural network into sub-networks and decompose the entire optimization problem into smaller problems to significantly reduce the optimization complexity. In the ITNE schema, instead of finding the output range of two copies of each sub-network, we look for the range of the original sub-network and the range of the output distance. For a decomposed network $F_w(x_j^{(i)})$ with input $x^{(i-w)}$ and output $x_j^{(i)}$, given the input bounds $\underline{\overline{x}}^{(i-w)}$ and $\Delta\underline{\overline{x}}^{(i-w)}$[1], optimization problem in Eq. (1) is formulated on $F_w(x_j^{(i)})$ to derive output ranges $\underline{\overline{x}}_j^{(i)}$ and $\Delta\underline{\overline{x}}_j^{(i)}$.

### ITNE-Based LP Relaxation (LPR)

The idea of LPR is to relax the ReLU relation $x = \max(0, y)$, $y \in [\underline{y}, \overline{y}]$ into linear constraints. When $\underline{y} \leq 0 \leq \overline{y}$, ReLU relation can be relaxed by three linear inequations:

$$x \geq 0, \quad x \geq y, \quad (\overline{y} - \underline{y})x \leq \overline{y}(y - \underline{y}). \tag{2}$$

In this work, we relax ReLUs in the original network $x^{(n)} = F(x^{(0)})$ by Eq. (2), and relax the ReLU distance $\Delta x = relu(y + \Delta y) - relu(y)$, as shown in Fig. 3. For $\forall y \in \mathbb{R}$, the $(\Delta x, \Delta y)$ mapping always falls in the shadowed area. Given $\Delta y \in [\Delta\underline{y}, \Delta\overline{y}]$, the relation between $\Delta x$ and $\Delta y$ can be bounded by a linear lower and upper bound:

$$\frac{l(u - \Delta y)}{u - l} \leq \Delta x \leq \frac{u(\Delta y - l)}{u - l}, \tag{3}$$

where $l = \min(0, \Delta\underline{y})$ and $u = \max(0, \Delta\overline{y})$.

[1] We denote $\underline{\overline{v}} = [\underline{v}, \overline{v}]$ as the range of variable $v$.

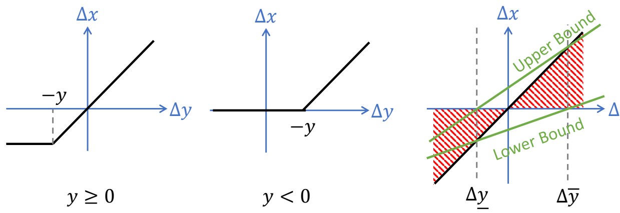
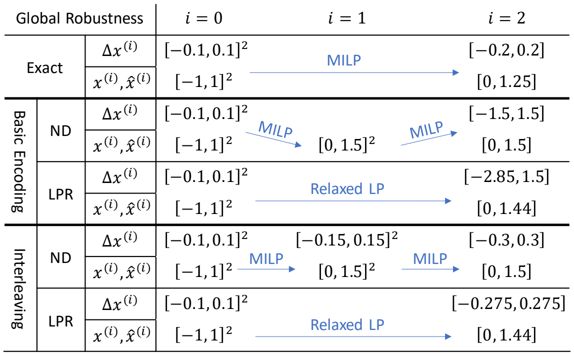


Figure 3: *Left:* ReLU distance relation when $y \geq 0$; *Middle:* ReLU distance relation when $y < 0$; *Right:* LP-relaxation of ReLU distance relation. The ReLU distance relation for $\forall y \in \mathbb{R}$ lays in the shadowed area. Within the distance range $\Delta y \in [\Delta \underline{y}, \Delta \overline{y}]$, $\Delta x$ is bounded by the lower and upper bounds.



| Global Robustness | | | $i = 0$ | | $i = 1$ | | $i = 2$ |
|---|---|---|---|---|---|---|---|
| Exact | | $\Delta x^{(i)}$ | $[-0.1, 0.1]^2$ | | MILP | | $[-0.2, 0.2]$ |
| | | $x^{(i)}, \hat{x}^{(i)}$ | $[-1, 1]^2$ | | $\longrightarrow$ | | $[0, 1.25]$ |
| Basic Encoding | ND | $\Delta x^{(i)}$ | $[-0.1, 0.1]^2$ | | | | $[-1.5, 1.5]$ |
| | | $x^{(i)}, \hat{x}^{(i)}$ | $[-1, 1]^2$ | $\xrightarrow{\text{MILP}}$ | $[0, 1.5]^2$ | $\xrightarrow{\text{MILP}}$ | $[0, 1.5]$ |
| | LPR | $\Delta x^{(i)}$ | $[-0.1, 0.1]^2$ | | Relaxed LP | | $[-2.85, 1.5]$ |
| | | $x^{(i)}, \hat{x}^{(i)}$ | $[-1, 1]^2$ | | $\longrightarrow$ | | $[0, 1.44]$ |
| Interleaving | ND | $\Delta x^{(i)}$ | $[-0.1, 0.1]^2$ | | $[-0.15, 0.15]^2$ | | $[-0.3, 0.3]$ |
| | | $x^{(i)}, \hat{x}^{(i)}$ | $[-1, 1]^2$ | $\xrightarrow{\text{MILP}}$ | $[0, 1.5]^2$ | $\xrightarrow{\text{MILP}}$ | $[0, 1.5]$ |
| | LPR | $\Delta x^{(i)}$ | $[-0.1, 0.1]^2$ | | Relaxed LP | | $[-0.275, 0.275]$ |
| | | $x^{(i)}, \hat{x}^{(i)}$ | $[-1, 1]^2$ | | $\longrightarrow$ | | $[0, 1.44]$ |

Figure 4: Illustrating example: Global robustness certification processes of exact MILP, network decomposition (ND) and LP relaxation (LPR) for each network encoding schema (BTNE and ITNE).

## 2.4 Illustrating Example

We consider the example neural network in Fig. 1, and set the input perturbation bound as $\delta = 0.1$ and the input domain as $x^{(0)} \in [-1, 1]^2$. The example neural network can be decomposed into three sub-networks:

$$x_1^{(1)} = relu(y_1^{(1)}) = relu(x_1^{(0)} + 0.5x_2^{(0)}),$$

$$x_2^{(1)} = relu(y_2^{(1)}) = relu(-0.5x_1^{(0)} + x_2^{(0)}),$$

$$x^{(2)} = relu(y^{(2)}) = relu(x_1^{(1)} - x_2^{(1)}).$$

In Fig. 4, we demonstrate the global robustness certification processes for different techniques. The exact MILP derives the exact output variation range $[-0.2, 0.2]$. When applying ND or LPR under BTNE, they are applied to each individual network copy. After ND, The distance information between $x^{(1)}$ and $\hat{x}^{(1)}$ is not encoded and is lost, resulting in a 7.5x over-approximation of the range of $\Delta x^{(2)}$. The LPR is based on the bounds $y^{(1)}, \hat{y}^{(1)} \in [-1.5, 1.5]^2$, and $y^{(2)}, \hat{y}^{(2)} \in [-1.5, 1.5]$, resulting in a 10.9x over-approximation. On the other hand, under ITNE, the ITNE-based ND derives the range of $\Delta x^{(2)}$ based on the range of $x^{(1)}$ and $\Delta x^{(1)}$, which is only 1.5x of the exact one. Given $\Delta y^{(1)} \in [-0.15, 0.15]^2$ and $\Delta y^{(2)} \in [-0.3, 0.3]$, the ITNE-based LPR derives a tight 1.38x over-approximation, significantly improving over BTNE.

Table 1: Neural network setting and experimental results.

| ID | Neurons | $t_R$ | $t_M$ | $t_{our}$ | $\varepsilon$ | $\overline{\varepsilon}_{our}$ |
|---|---|---|---|---|---|---|
| 1 | 8 | 2s | 0.1s | 0.3s | 0.0583 | 0.0657 |
| 2 | 12 | 130s | 0.2s | 0.4s | 0.0527 | 0.0722 |
| 3 | 16 | 8h | 0.8s | 1s | 0.0496 | 0.0653 |
| 4 | 32 | >24h | 74s | 5s | 0.0481 | 0.0673 |
| ID | Neurons | Layers | | $t_{our}$ | $\underline{\varepsilon}$ | $\overline{\varepsilon}_{our}$ |
| 5 | 64 | FC:3 | | 50s | 0.0452 | 0.0731 |
| 6 | 1416 | Conv:1<br>FC:2 | | 4.8h | 0.347 | 0.578 |
| | | | | | 0.300 | 0.572 |
| 7 | 3872 | Conv:2<br>FC:2 | | 3.3h | 0.453 | 0.874 |
| | | | | | 0.420 | 0.723 |
| 8 | 5824 | Conv:3<br>FC:2 | | 3.5h | 0.519 | 1.521 |
| | | | | | 0.407 | 1.175 |

## 2.5 Efficient Over-Approximation Algorithm

Finally, our global robustness certification algorithm is designed by combining the ITNE-based ND and LPR techniques with Selective Refinement (SR)[2].

For layer $i$, ND constructs sub-networks $F_w(x_j^{(i)}), \forall j$. MILP problems are built by LPR and SR to get the output variation bounds of layer $i$. The over-approximated output variation bound $\overline{\varepsilon}$ is derived by iteratively evaluating the output bounds (both the $\Delta \underline{y}^i$ and $\Delta \underline{x}^i$) of each hidden layer.

# 3 Evaluation and Applications

We first evaluate our algorithm on various DNNs and compare its results with exact global robustness (when available) and an under-approximated global robustness. Then, we demonstrate the application of our approach in a case study of safety verification for a vision-based robotic control system, and show the importance of efficient global robustness certification for safety-critical systems that involve neural nets. Finally, we discuss its comparison with adversarial training.

## 3.1 Performance Evaluation

We compare our approach with other methods on a set of DNNs, as shown in Table 1. DNNs 1 to 5 are 3-layer fully-connected (FC) networks trained on the Auto MPG dataset [Quinlan, 1993]. DNNs 6 to 8 are convolutional networks trained on the MNIST dataset [Lecun *et al.*, 1998][3]. The input perturbation bound $\delta = 0.001$ for DNNs 1 to 5 and $\delta = 2/255$ for DNNs 6 to 8.[4]

We compare our over-approximated output variation bound $\overline{\varepsilon}_{our}$ with the exact bound $\varepsilon$ solved by Reluplex [Katz *et al.*, 2017; Katz *et al.*, 2019] and the MILP encoding in Eq. (1).

[2] While LPR can remove all integer variables in the MILP formulation to reduce the complexity, such extreme over-approximation may be too inaccurate. Thus, we try to selectively refine a limited number of neurons, by *not* relaxing their ReLU relations. This is similar to the layer-level refinement idea in [Huang *et al.*, 2020b], but with a focus on global robustness.

[3] Due to limited space, we only present 2 outputs of MNIST (out of 10) in Table 1. The rest show similar trends.

[4] More detailed experiment settings can be found in the full paper [Wang *et al.*, 2022b].

The runtime of Reluplex $t_R$ and MILP $t_M$ quickly increases with respect to neural network size. None of them can address 64-neuron DNN-5 within 24 hours. From DNNs 1 to 4, whose exact bounds $\varepsilon$ are available, our algorithm can finish in seconds with only about 13% to 40% over-approximation. Starting from DNN 5, there is no other work in the literature that can derive a sound and deterministic global robustness in a reasonable time. To assess our over-approximated results for larger networks, we leverage adversarial examples from the Projected Gradient Descent (PGD) [Madry *et al.*, 2018] among the entire dataset to derive an *under-approximated* output variation bound $\underline{\varepsilon}$, inspired by [Ruan *et al.*, 2019]. The experiments of DNNs 6 to 8 demonstrate that **our method can provide meaningful over-approximation (less than 3x of the under-approximation) for DNNs with more than 5000 hidden neurons within 5 hours**.

In our recent work [Wang *et al.*, 2022c], we further improved the efficiency and tightness of our approach, where we leverage novel symbolic propagation technique inspired by $\beta$-CROWN [Wang *et al.*, 2021a] to replace the MILP solver in this work. The symbolic propagation technique takes advantage of GPU acceleration and can certify DNNs 6 to 8 within 3 hours and reduce the $(\overline{\varepsilon}_{our} - \underline{\varepsilon})$ gap by $9\% - 60\%$.

### 3.2 Case Study on Control Safety Verification

For control systems that use neural networks for perception, a critical and yet challenging question is whether the system can remain safe under perturbation of network inputs. In [Wang *et al.*, 2021d], we formulate this as a design-time safety assurance problem based on global robustness. Here, leveraging our global robustness certification technique, we demonstrate a solution for this safety assurance problem.

In particular, we consider an advanced cruise control (ACC) case study, where an ego vehicle, equipped with a camera, is following a reference vehicle. The captured images may be slightly perturbed. The distance from the reference vehicle is inferred from the images by a DNN. A feedback controller controls the ego vehicle based on the estimated distance. We model this example in the tool Webots [Michel, 2004] (Fig. 5). The ego vehicle is safe if distance $d \in [0.5, 1.9]$ and speed $v_e \in [0.1, 0.7]$. The reference vehicle speed $v_r$ is randomly adjusted within $[0.2, 0.6]$. The camera takes RGB images with resolution $24 \times 48$. A 5-layer convolutional network is trained with 100k pre-captured images. We model the entire dynamic system as an LTI system with external disturbance terms[5] [Wang *et al.*, 2022b], where the system states include distance and vehicle speed. The control input follows the feedback control law $u = K\hat{x}$, where $\hat{x}$ is the estimated system state. According to the invariant set based verification [Huang *et al.*, 2020a; Wang *et al.*, 2020; Wang *et al.*, 2021c], the vehicle control safety can be verified if the distance estimation error $\Delta d$ is within $[-0.14, 0.14]$.

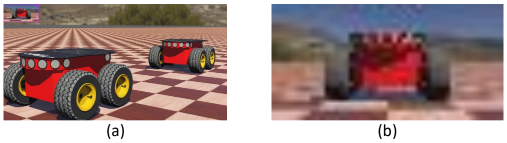


Figure 5: (a) Simulated ACC in Webots: The ego car (left) follows the reference car (right); (b) An example camera image.

The distance estimation error $\Delta d = \Delta d_1 + \Delta d_2$ contains the DNN model inaccuracy $\Delta d_1$ and the output variation $\Delta d_2$ caused by input perturbation. While $|\Delta d_1| \leq 0.0730$ is the worst-case model inaccuracy among the dataset, $\Delta d_2$ is the focus of this study and can be bounded by our global robustness certification algorithm. Assuming the input perturbation is bounded by $\delta = 2/255$, the certified output variation bound becomes $|\Delta d_2| \leq \overline{\epsilon} = 0.0568$. Combined with $\Delta d_1$, we have $|\Delta d| \leq 0.1298$. Therefore, under the assumed perturbation bound, we can assert that the DNN in this ACC system is safe.

This is validated in Webots simulations, where adversarial perturbations are added by the Fast Gradient Sign Method (FGSM) [Goodfellow *et al.*, 2015]. During simulations, when the perturbation bound $\delta = 2/255$, we always have $|\Delta d| \leq 0.14$ and a safe system. If we further increase input perturbation, $|\Delta d| > 0.14$ is observed when $\delta = 5/255$ and unsafe states are observed when $\delta = 10/255$. This shows the impact of input perturbation on system safety and the importance of our global robustness analysis.

### 3.3 Comparison with Adversarial Training

Neural network adversarial attack techniques and adversarial training algorithms are often jointly developed. Adversarial training performance is usually measured under the existing attack techniques, which may be less effective for more advanced attacks in the future. Instead, network robustness is a metric independent of attack techniques and can provide a deterministic guarantee. Compare to local robustness, where the guarantee is only on a finite set of data, global robustness can provide a universal guarantee for all possible inputs. Besides being a reliable metric, global robustness can also provide guidance to improve network robustness [Wang *et al.*, 2022c; Fu *et al.*, 2022] under arbitrary adversarial attacks.

## 4 Conclusion and Future Work

We present an efficient certification algorithm to provide sound and deterministic global robustness analysis for ReLU neural networks. Experiments demonstrate that our approach is much more efficient and scalable than the exact certification approaches while providing tight over-approximation, and a case study further demonstrates its potential for practical systems. We believe that the approach has the potential to be applied in a variety of domains such as autonomous driving [Jiao *et al.*, 2021; Zhu *et al.*, 2021; Wang *et al.*, 2021e] and smart building control [Xu *et al.*, 2021], where disturbances and noises to sensor inputs are common. We will explore these applications in future work. We also plan to develop methods that can improve the global robustness of neural networks during their design and training, and investigate methods for joint design and verification, inspired by our recent works in this area [Wang *et al.*, 2021b; Wang *et al.*, 2022a; Wang *et al.*, 2023a; Wang *et al.*, 2023b].

[5] External disturbances are caused by the randomness of reference vehicle speed and the inaccuracy of the linear model.

## Ethical Statement

There are no ethical issues with the proposed technique.

## Acknowledgments

We gratefully acknowledge funding support by National Science Foundation (NSF) grants 1834701, 1724341, 2038853, Department of Energy (DOE) award DE-EE0009150, and Office of Naval Research grant N00014-19-1-2496.

## References

[Bastani *et al.*, 2016] Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, et al. Measuring neural net robustness with constraints. In *NeurIPS*, volume 29, 2016.

[Biggio *et al.*, 2013] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases*, pages 387–402, Berlin, Heidelberg, 2013.

[Chen *et al.*, 2021] Yizheng Chen, Shiqi Wang, Yue Qin, Xiaojing Liao, Suman Jana, and David Wagner. Learning security classifiers with verified global robustness properties. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, page 477–494, New York, NY, USA, 2021.

[Cheng *et al.*, 2017] Chih-Hong Cheng, Georg Nührenberg, and Harald Ruess. Maximum resilience of artificial neural networks. In *Automated Technology for Verification and Analysis*, pages 251–268, 2017.

[Fan *et al.*, 2020] Jiameng Fan, Chao Huang, Xin Chen, Wenchao Li, and Qi Zhu. Reachnn*: A tool for reachability analysis of neural-network controlled systems. In Dang Van Hung and Oleg Sokolsky, editors, *Automated Technology for Verification and Analysis*, pages 537–542, Cham, 2020. Springer International Publishing.

[Fu *et al.*, 2022] Feisi Fu, Zhilu Wang, Jiameng Fan, Yixuan Wang, Chao Huang, Xin Chen, Qi Zhu, and Wenchao Li. REGLO: Provable neural network repair for global robustness properties. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022.

[Goodfellow *et al.*, 2015] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[Gopinath *et al.*, 2018] Divya Gopinath, Guy Katz, Corina S. Păsăreanu, and Clark Barrett. Deepsafe: A data-driven approach for assessing robustness of neural networks. In *Automated Technology for Verification and Analysis*, pages 3–19, Cham, 2018.

[Huang *et al.*, 2019] Chao Huang, Jiameng Fan, Wenchao Li, Xin Chen, and Qi Zhu. Reachnn: Reachability analysis of neural-network controlled systems. *ACM Transactions on Embedded Computing Systems (TECS)*, 18(5s):1–22, 2019.

[Huang *et al.*, 2020a] C. Huang, S. Xu, Z. Wang, et al. Opportunistic intermittent control with safety guarantees for autonomous systems. In *DAC*, pages 1–6, 2020.

[Huang *et al.*, 2020b] Chao Huang, Jiameng Fan, Xin Chen, et al. Divide and slide: Layer-wise refinement for output range analysis of deep neural networks. *TCAD*, 39, 2020.

[Huang *et al.*, 2022] Chao Huang, Jiameng Fan, Xin Chen, Wenchao Li, and Qi Zhu. Polar: A polynomial arithmetic framework for verifying neural-network controlled systems. In *Automated Technology for Verification and Analysis*. Springer International Publishing, 2022.

[Jiao *et al.*, 2021] Ruochen Jiao, Hengyi Liang, Takami Sato, Junjie Shen, Qi Alfred Chen, and Qi Zhu. End-to-end uncertainty-based mitigation of adversarial attacks to automated lane centering. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 266–273, 2021.

[Katz *et al.*, 2017] Guy Katz, Clark Barrett, David L. Dill, et al. Reluplex: An efficient smt solver for verifying deep neural networks. In *CAV*, pages 97–117, 2017.

[Katz *et al.*, 2019] Guy Katz, Derek A. Huang, Duligur Ibeling, et al. The marabou framework for verification and analysis of deep neural networks. In *CAV*, 2019.

[Lecun *et al.*, 1998] Y. Lecun, L. Bottou, Y. Bengio, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 1998.

[Liu *et al.*, 2022] Xiangguo Liu, Chao Huang, Yixuan Wang, Bowen Zheng, and Qi Zhu. Physics-aware safety-assured design of hierarchical neural network based planner. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*, pages 137–146. IEEE, 2022.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[Mangal *et al.*, 2019] Ravi Mangal, Aditya V. Nori, and Alessandro Orso. Robustness of neural networks: A probabilistic and practical approach. In *2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*, pages 93–96, 2019.

[Michel, 2004] O. Michel. Webots: Professional mobile robot simulation. *Journal of Advanced Robotics Systems*, 1(1):39–42, 2004.

[Quinlan, 1993] R. Quinlan. Auto MPG. UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5859H.

[Ruan *et al.*, 2019] Wenjie Ruan, Min Wu, Youcheng Sun, et al. Global robustness evaluation of deep neural networks with provable guarantees for the hamming distance. In *IJCAI*, 2019.

[Singh *et al.*, 2019] Gagandeep Singh, Timon Gehr, Markus Püschel, et al. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3, January 2019.

[Wang *et al.*, 2020] Yixuan Wang, Chao Huang, and Qi Zhu. Energy-efficient control adaptation with safety guarantees for learning-enabled cyber-physical systems. In *Proceedings of the 39th International Conference on Computer-Aided Design*, pages 1–9, 2020.

[Wang *et al.*, 2021a] Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In *Advances in Neural Information Processing Systems*, volume 34, pages 29909–29921, 2021.

[Wang *et al.*, 2021b] Yixuan Wang, Chao Huang, Zhilu Wang, Shichao Xu, Zhaoran Wang, and Qi Zhu. Cocktail: Learn a better neural network controller from multiple experts via adaptive mixing and robust distillation. In *58th ACM/IEEE Design Automation Conference, DAC 2021, San Francisco, CA, USA, December 5-9, 2021*, pages 397–402. IEEE, 2021.

[Wang *et al.*, 2021c] Zhilu Wang, Chao Huang, Hyoseung Kim, Wenchao Li, and Qi Zhu. Cross-layer adaptation with safety-assured proactive task job skipping. *ACM Trans. Embed. Comput. Syst.*, 20(5s), sep 2021.

[Wang *et al.*, 2021d] Zhilu Wang, Chao Huang, Yixuan Wang, et al. Bounding perception neural network uncertainty for safe control of autonomous systems. In *DATE*, 2021.

[Wang *et al.*, 2021e] Zhilu Wang, Hengyi Liang, Chao Huang, and Qi Zhu. Cross-layer design of automotive systems. *IEEE Design Test*, 38(5):8–16, 2021.

[Wang *et al.*, 2022a] Yixuan Wang, Chao Huang, Zhilu Wang, Zhaoran Wang, and Qi Zhu. Design-while-verify: Correct-by-construction control learning with verification in the loop. In *59th ACM/IEEE Design Automation Conference, DAC 2022, San Francisco, CA, USA, July 10-14, 2022*, 2022.

[Wang *et al.*, 2022b] Zhilu Wang, Chao Huang, and Qi Zhu. Efficient global robustness certification of neural networks via interleaving twin-network encoding. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1087–1092, 2022.

[Wang *et al.*, 2022c] Zhilu Wang, Yixuan Wang, Feisi Fu, Ruochen Jiao, Chao Huang, Wenchao Li, and Qi Zhu. A tool for neural network global robustness certification and training. *arXiv preprint arXiv:2208.07289*, 2022.

[Wang *et al.*, 2023a] Yixuan Wang, Simon Zhan, Zhilu Wang, Chao Huang, Zhaoran Wang, Zhuoran Yang, and Qi Zhu. Joint differentiable optimization and verification for certified reinforcement learning. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)*, ICCPS '23, page 132–141, New York, NY, USA, 2023. Association for Computing Machinery.

[Wang *et al.*, 2023b] Yixuan Wang, Simon Sinong Zhan, Ruochen Jiao, Zhilu Wang, Wanxin Jin, Zhuoran Yang, Zhaoran Wang, Chao Huang, and Qi Zhu. Enforcing hard constraints with soft barriers: Safe-driven reinforcement learning in unknown stochastic environments. In *ICML'23: Proceedings of the International Conference on Machine Learning*, 2023.

[Xu *et al.*, 2021] Shichao Xu, Yangyang Fu, Yixuan Wang, Zheng O'Neill, and Qi Zhu. Learning-based framework for sensor fault-tolerant building hvac control with model-assisted learning. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, BuildSys '21, page 1–10, New York, NY, USA, 2021. Association for Computing Machinery.

[Zhang *et al.*, 2018] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[Zhu *et al.*, 2020] Qi Zhu, Wenchao Li, Hyoseung Kim, Yecheng Xiang, Kacper Wardega, Zhilu Wang, Yixuan Wang, Hengyi Liang, Chao Huang, Jiameng Fan, and Hyunjong Choi. Know the unknowns: Addressing disturbances and uncertainties in autonomous systems. In *Proceedings of the 39th International Conference on Computer-Aided Design*, ICCAD '20, New York, NY, USA, 2020. Association for Computing Machinery.

[Zhu *et al.*, 2021] Qi Zhu, Chao Huang, Ruochen Jiao, Shuyue Lan, Hengyi Liang, Xiangguo Liu, Yixuan Wang, Zhilu Wang, and Shichao Xu. Safety-assured design and adaptation of learning-enabled autonomous systems. In *Proceedings of the 26th Asia and South Pacific Design Automation Conference*, ASPDAC '21, page 753–760, New York, NY, USA, 2021. Association for Computing Machinery.