My research interests are in the intersection of machine learning, computer security, and software engineering, with a focus on building *certifiably trustworthy* deep learning (DL) systems. Specifically, my research aims at answering this core question: *how to build large-scale deep learning systems with certified trustworthiness?*

1 Background

In the past decade, we have witnessed deep learning and AI revolution in every aspect of our life. Deep learning has reshaped many fields such as image recognition and generation, language understanding and reasoning, program synthesis and testing [13], and decision-making.

Despite such success, the wide deployment of DL has exposed serious safety and social threats due to their lack of trustworthiness. For example, an adversary can rotate the camera [14] or attach tiny stickers on road signs [28] to fool DL-based autonomous driving systems, possibly leading to fatal traffic accidents. Such an adversary has been made practical to attack commercial DL systems [6, 7]. As noted in the recent *AI bill of rights* from the White House [32], such lack of trustworthiness in current DL-based tools threatens the rights of the public and challenges democracy.

Existing literature mainly evaluates the trustworthiness via detection approaches (i.e., "attacks" in computer security literature). These approaches find trustworthy vulnerabilities by searching trustworthiness-violating examples, e.g., specific rotation angles that fool the DL model into making wrong predictions. Then, some approaches enhance the DL system so that detection approaches cannot find trustworthiness issues. However, when a detection method fails to find trustworthy vulnerabilities, it could be because the detection method is not powerful enough, rather than the DL system is truly trustworthy. Indeed, many enhanced DL systems are shown not trustworthy by stronger detection approaches [23], which brings "a false sense of trustworthiness." Hence, *certified methods* are in urgent demand since they generate rigorous guarantees of the trustworthiness of DL systems.

2 Research Overview

My research brings certified trustworthiness to large-scale deep learning systems by proposing certified methods for various trustworthy properties and applications. As shown in Figure 1, my work certifies trustworthy properties such as robustness against ℓ_p-bounded perturbations, robustness against semantic transformations, distributional robustness, distributional fairness, and numerical reliability. These properties cover the top two aspects in the blueprint of trustworthy AI [32]: safe & effective systems and algorithmic discrimination protections. Besides, my work brings certified methods to important DL applications, including deep reinforcement learning [3, 2], point cloud recognition [19], and robotics perception [4]. Our certified methods can be divided into two categories: *certification approaches* and *certified-enhancing approaches*. Certification approaches provide rigorous trustworthiness guarantees given a DL system, a data instance or distribution, and a specification. Certified-enhancing approaches build DL systems that are easier to be certified.

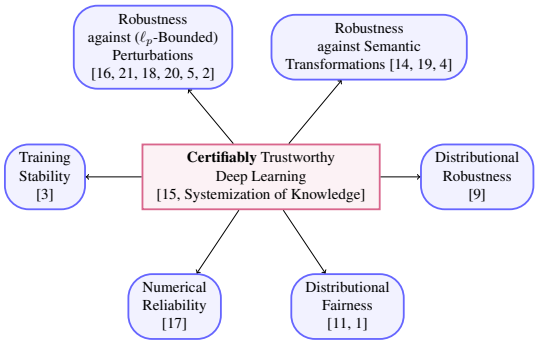

Figure 1: My research enables and enhances the certified trustworthiness of various properties for large-scale DL systems.

Highlights. I am recognized as the **2022 Adversarial Machine Learning Rising Star** [26](two awardees annually) and the **2022 Rising Star in Data Science** [27]. My research leads the frontier of certification approaches: I proposed DSRS [16], which provides asymptotically optimal robustness certification under mild conditions for bounded input domains (such as image domains) and is proven to **bypass the widely believed certification tightness barrier** on high-dimensional data [37, 24, 30, 29, 35] for the first time. I am the key contributor of α-β-CROWN verifier [5][1]which **won the latest neural network verification competition VNNCOMP'22** [34] and has received over 100 stars on GitHub in two months. Furthermore, my work enables rigorous and scalable **certification for several trustworthy properties for the first time**, such as robustness against semantic transformations [14], numerical reliability [17], and distributional fairness [11].

My research also pioneers certified-enhancing approaches: I proposed DRT [21] and TSS [14], which train DL systems to achieve state-of-the-art robustness against ℓ_p-bounded perturbation and semantic transformation on large-scale datasets such as ImageNet. Besides, I published the **first systematization of knowledge** paper in certifiably trustworthy deep learning [15], along with the **first standardized toolkit and benchmark leaderboard**[2] to accelerate the deployment of certified methods.

3 Unified Design of Certified Deep Learning Methods

This section introduces core principles and methodologies that guide my design of certified methods for deep learning systems, including both certification approaches and certified-enhancing approaches.

[1] https://github.com/Verified-Intelligence/alpha-beta-CROWN
[2] https://sokcertifiedrobustness.github.io/.

3.1 Certification Approaches A certification approach should rigorously guarantee some trustworthy properties of the given DL systems. Taking the robustness against ℓ_p-bounded perturbation property as an example, we require that tiny perturbations δ (whose ℓ_p norm is bounded by a small threshold ϵ, i.e., $\|\delta\|_p \leq \epsilon$), when added to a normal input x, should not alter the system's prediction, i.e., $f(x) = f(x')$ [33]. A certification approach $\mathcal{A}$ will take the DL system f, a data instance x, and perturbation magnitude ϵ as the inputs; when it outputs true, the DL system should always be robust. Formally, $\mathcal{A}(f, x, \epsilon) = \text{True} \Rightarrow \forall \delta, \|\delta\|_p \leq \epsilon, f(x+\delta) = f(x)$.

We design scalable certification approaches by casting certification as bounding worst-case system performance on a (possibly infinite) perturbation set. We propose a unified view of representative trustworthy properties as summarized in Table 1. Within our view, the goal of certification approaches is to upper bound an optimization problem where the trustworthy property determines the input domain constraints, and the DL system determines the objective to certify. Still, take the robustness against perturbation property as the example, the input domain is $\{x' : \|x' - x\|_p \leq \epsilon\}$, and the optimization objective of the certification approach is to compute an upper bound of $\max |f(x') - f(x)|$.

Table 1: Unified view for certification approaches—bounding worst-case performance for various trustworthy properties. "Dist." means distribution.

Property	Perturbation Set	Objective to Upper Bound	Our Work
Robustness against (ℓ_p-Bounded) Perturbation	$x' : \|x' - x\|_p \leq \epsilon$	$\max \|f(x') - f(x)\|$	[16, 5, 2]
Robustness against Semantic Transformation	Domain of transformation (e.g., rotation, scaling) function	Same as above	[14, 19]
Distributional Robustness	Set of dist. surrounding the training dist.	Expected test loss on new dist.	[9]
Training Stability	Same as above	Expected test loss on original dist., for DL system trained on new dist.	[3]
Fairness	Same as above, with additional base rate fairness constraint	Expected test loss on new dist.	[11]
Numerical Reliability	Set of all valid input and valid weights	Absolute value distance to `FLOAT_MAX` and `FLOAT_MIN`	[17]

3.1.1 White-Box Certification White-box certification approaches analyze the architecture of deep neural networks (DNNs) to compute an upper bound for the corresponding optimization. Since optimization involving a DNN is generally highly non-convex, these approaches compute a rigorous upper bound by propagating linear relaxation through layers and non-linear activations of the DNN model, as shown in Figure 2.

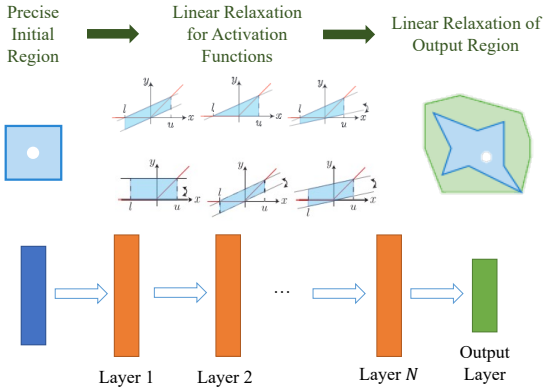

Figure 2: Our white-box certification propagates linear relaxation through DNN to get upper bounds.

Two major drawbacks of white-box certification are its *looseness* (especially when models have many layers since the relaxation region is amplified layer by layer) and *expensive computational cost* (since the relaxation set needs to be maintained through layers). To this end, I co-developed the α-β-CROWN verifier that achieves the best trade-off between tightness and scalability so far and won the DNN verification competition VNNCOMP'22 [34] this year. In our verifier, we proposed batched branch-and-bound algorithm for DNN certification, where we divide the input domains for some non-linear relaxations and generate tighter sub-problems to solve. For each generated sub-problem, we dynamically query tighter linear constraints from strong solvers and apply such constraints to tighten current relaxations in an online manner [5]. Thanks to our unified view of certification, I further enable the certification of new trustworthy properties, such as numerical reliability [17].

3.1.2 Black-Box Certification Black-box certification does not analyze the DNN architecture. Instead, it requires only oracle access to the model's prediction to compute the certification. Therefore, they are more scalable and support larger DNN models than white-box certification. Black-box certification approaches usually alter the inference protocol of DL systems: instead of taking one DNN forward pass to get inference result, they apply random noise (called "smoothing") to the input and take the majority vote or mean among sampled DNN outputs for noised inputs as the final system output. In this way, the output of the DL system is based on distribution statistics instead of a single instance. For example, if the random noise is Gaussian $\mathcal{N}(0, \sigma^2 I)$, for input x, the DL system's output is based on Gaussian distribution $\mathcal{N}(x, \sigma^2 I)$. As a result, the statistics at one input instance can be used to bound the DL system's outcome in its neighborhood since distributions can overlap. Hence, certification can be derived for any input within the perturbation set.

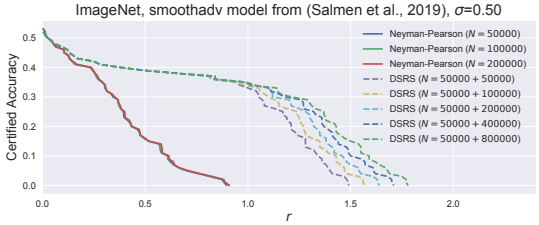

Figure 3: ([16, Fig. 2(b)]) our DSRS approach brings much tighter certification for robust large-scale ImageNet classification.

Two major drawbacks of black-box certification are their *limited tightness* and *limited applicability*. My research significantly mitigates these two drawbacks. (1) **Limited tightness**: Existing black-box approaches rely on the voting probability or mean value under the smoothed input distribution to compute the certification for the perturbation set. Therefore, the tightness of the certification is limited by the overlap ratio between the original smoothed distribution and

the perturbed smoothed distribution. Several theoretical analyses reveal that, for high-dimensional data, such overlap ratio is diminishing and poses an intrinsic barrier named "ℓ_∞ barrier" or "curse of dimensionality" for black-box certification [37, 24, 30, 29, 35]. In my recent work [16], for the first time, I proved that the theoretical barrier could be bypassed by using additional statistics other than those under the original smoothed distribution. *This result refutes the common belief ("black-box certification is intrinsically loose") in the field and establishes the theoretical foundation for applying black-box certification to the large-scale high-dimensional data domain.* In practice, as shown in Figure 3, our DSRS approach, which leverages additional statistics, is significantly tighter than existing certification. (2) **Limited applicability**: Existing black-box approaches rely on the overlap between distributions to derive the certification. However, for some trustworthy properties, the defined perturbation set (see Table 1) may not incur such overlap. For example, in the physical world, the natural environment usually incurs semantic transformations to image input, such as rotation, scaling, and brightness change. When additive Gaussian noise is directly added to input, the noised original input and noised transformed input can share little overlap, which makes existing black-box certification inapplicable. We propose the TSS approach [14, 4]. TSS first applies semantic-specific smoothing by imposing random noise in the parameter space of the latent transformation function to enlarge the overlap between two distributions and then computes the Lipschitz constant for generic transformation functions to lower bound the overlap ratio. *For the first time*, TSS provides strong certified robustness for large-scale DL systems against semantic transformations in the physical world. We also propose the *first* suites of black-box certification methods for distributional robustness [9] and fairness [11]. For DL applications, we develop the *first* family of practical black-box deep reinforcement learning certification methods [2, 3].

3.2 Certified-Enhancing Approaches Certified-enhancing approaches build certification-friendly DL systems. For both white-box and black-box certification, we build suitable DL systems, respectively.

3.2.1 For White-Box Certification: Subspace Optimization, Pruning, and Lipschitz Architecture. The white-box certification computes the relaxation for DNNs to derive the certification. To build certification-friendly DL systems, the common strategy is to plug the relaxation bound into the training objective. However, such relaxation can be loose for large models, making such optimization challenging.

We propose three strategies to mitigate this fundamental challenge: (1) **Subspace optimization**: In our Robustra approach [18], instead of applying over-approximation to the whole perturbation set, we use a surrogate DNN model to select a subset that is more likely to violate our trustworthy property to apply the relaxation. Therefore, the training procedure aims to minimize the risk of property violation in a smaller set to ease the optimization burden and boost the certified trustworthiness of the trained DL system. (2) **Pruning**: The relaxation of white-box certification becomes more significant for larger models. Thus, reducing the model size could tighten the certification and achieve better-certified trustworthiness. I co-developed FaShapley [10], which leverages Shapley values stemming from game theory to identify important neurons and prune other neurons. FaShapley achieves both tighter certification and faster certification time. (3) **Lipschitz architecture**: If we can design DNN architectures to satisfy some nice properties by design, we can get rid of expensive relaxation propagation. I co-developed the LOT approach [20], which computes orthogonal convolutional layers by applying Newton's iteration in the Fourier domain. In this way, the constructed DNN model has an overall Lipschitz constant 1, so we can bypass the expensive propagation and tightly and efficiently certify the system's trustworthiness with this Lipschitz constant.

3.2.2 For Black-Box Certification: Gradient Diversity and Knowledge-Based Correction. The black-box certification applies random noise to input and uses output statistics to derive certification. We propose two principled views to design certified-enhancing approaches for black-box certification: (1) **View 1**: *black-box certification brings intrinsic higher-order smoothness to the DL system*. I proved the first known higher-order smoothness bound for the noise smoothing procedure in black-box certification [21]. This intrinsic bound depends only on the added noise variance. Thus, when black-box approaches are used and noise variance is fixed, their performance can be guaranteed with only zeroth- and first-order quantities. We propose DRT [21], which trains a DNN model to enlarge zeroth-order quantity (by encouraging confidence margins) and to shrink first-order quantity (by promoting gradient diversity). DRT achieves state-of-the-art certified robustness on the large-scale ImageNet dataset. (2) **View 2**: *black-box certification needs a powerful denoising ability to achieve certified trustworthiness*. Since the black-box certification imposes random noise on the input, the DL system should be powerful in handling noised inputs to achieve high performance. When domain knowledge of the task is at hand, we propose the CARE approach [8], which integrates predictions from hierarchical and attribute classification models by graph neural networks to apply error correction when the main model is corrupted by input noise. Our CARE approach achieves drastic improvements in black-box certified robustness for tasks with domain knowledge, and it is the first-of-its-kind approach that effectively leverages domain knowledge to enhance certified trustworthiness.

4 Research Plan

My future research will cover five topics: strong, scalable, and practical certified methods for trustworthy DL; theoretical understanding of trustworthy DL; software engineering for DL; DL for software engineering; and DL for system security.

4.1 Strong, Scalable, and Practical Certified Methods for Trustworthy DL

Continuing from my past and ongoing research, the primary direction of my future research is to answer this ultimate question: *can we build AI systems that are as certifiably trustworthy as our humans on real-world, large-scale tasks?* To answer this question, I plan to focus on the following three fields.

Field 1: Tight and Scalable Certification for Trustworthiness. Though my research has enabled and significantly improved certified trustworthiness for DL systems, there is still a significant gap between the current trustworthiness level and the level of desire. For example, for robustness against perturbations, on ImageNet, the best accuracy lower bound against ℓ_p-bounded perturbations is merely around 30%. To close this gap, I plan to design tighter and more scalable certification methods to bring trustworthy guarantees for powerful large models and to design effective certified-enhancing methods for large models. I propose two principles to guide the design: (1) *Integration of domain-specific or task-specific knowledge*: Pure data-driven learning may be intrinsically limited toward human-level trustworthiness for complex tasks [25]. Hence, domain-specific or task-specific knowledge can be leveraged. As a future step, I believe that automated knowledge collection, knowledge distillation, and knowledge integration, for both certification approaches and certification-enhancing approaches, could be the key to large-scale certified trustworthiness. (2) *Efficient abstraction of knowledge-enhanced DL system*: Certification approaches can be viewed as constructing abstractions for the DL system to characterize its behavior on the perturbation set. The abstraction of existing certification is usually too generic, which handles the model in a way agnostic of its training or constructing method. I will develop such a next-generation approach for DL certification, where efficient abstraction of the DL system, especially the integration of knowledge and reasoning components, is exploited to derive a much tighter and more scalable certification.

Field 2: Generic Certified Trustworthiness. Besides certification for existing trustworthy properties, certification is also in need against more foreseen or unforeseen trustworthy threats. In this aspect, I propose a roadmap containing three stages: *Stage 1: Discover, define, and certify more practical trustworthy properties*. Toward trustworthy AI, we need to identify more trustworthy properties that incur profound social impacts. After properties are defined, we need to define them formally and propose corresponding certification approaches. To this end, one of my future research questions is "Can we formally construct a DL-based system, such as a DL-based autonomous vehicle or robotic agent, that is guaranteed to be universally safe when deployed in the physical world?" *Stage 2: Certify multiple trustworthy properties at the same time*. Existing certified methods mainly provide certified trustworthiness for a single property. Can we achieve certified trustworthiness under multiple properties at the same time? If so, what would be the efficient method? If not, is there any inherent trade-off between different properties? What else (e.g., more data, more structured knowledge, more human supervision, or better algorithms) do we need to achieve so? *Stage 3: Develop certifiably trustworthy DL systems holding multiple foreseen and unforeseen properties*. Instead of defining trustworthy properties and developing methods to certify them, is it possible to achieve "meta-trustworthiness"? In other words, can we develop DL systems in a property-agnostic way to achieve human-level trustworthiness under all existing notions (robustness, fairness, privacy, etc.) and possible future notions?

Field 3: Deployment of Certified Trustworthy Systems and Study of Social Impacts. There is no free lunch—the certified methods for DL systems come at a cost. Costs include inference overhead, training overhead, and normal performance degradation. These negative costs are understudied but impose great barriers for deploying certified DL systems in the real-world. To deploy certified trustworthy systems in practice, I plan to mitigate these practical challenges. We also need to understand the practical implications and social impacts of certifiably trustworthy systems—if certifiably trustworthy DL systems are deployed, to what extent they can benefit social good, equality, democracy, and inclusion.

Funding Plan. AI trustworthiness is of core interest in large firms and the government, where *AI bill of rights* highlights the importance of AI trustworthiness by the White House [32]. During my PhD study, I helped my advisor with several proposals, sources of which range from large companies such as Meta and Amazon to government agencies such as NSF and DARPA. For example, we got the 2020 AWS Amazon Research Award and the 2021 Facebook Research Award since our research topic improves AI safety and reliability. As the individual PI, my proposal on autonomous driving safety verification is the finalist of the 2022 Qualcomm Innovation Fellowship. In the future, I will actively seek funding from these large firms and governmental agencies, and I will actively seek collaborations to jointly apply for fundings from more sources.

4.2 Theoretical Understanding of Trustworthy DL

I plan to advance our theoretical understanding of deep learning toward better trustworthiness. This line of research is motivated by both problem space and solution space. From the problem space, to develop more effective certified methods for trustworthy DL, we need to leverage more intrinsic properties of DL, which motivates us to theoretically understand DL better. For example, the study of learning dynamics, convergence, and generalization for generic training in DL can be done, and such a study can provide a principled way of designing new certified-enhancing approaches or new generic DL paradigms; the expression power analysis such as the universal approximation ability of the abstraction domains of certification can be done, and such analysis can guide the design of better certification methods. From the solution space, many techniques proposed for certified methods can be migrated to solve much broader classes of

problems in DL and theory, including Bayesian learning, neuro-symbolic methods, and privacy accounting.

4.3 Software Engineering for DL The wide deployment of DL calls for effective and efficient practices of developing, maintaining, and testing DL-based systems. To this end, I believe that leveraging and extending the rich experience from traditional software engineering is an important direction. Our group led by my co-advisor Prof. Tao Xie has a rich body of knowledge and infrastructure in traditional software engineering, such as the symbolic-execution-based testing tool Pex [36]. I believe that future generations of DL testing can come from combining my certified methods with these traditional software testing tools, including but not limited to the topics of counter-example-guided testing, symbolic execution, interpretable DL testing, and fuzzing for DL systems. More broadly, I plan to systematically investigate the whole development cycle of DL systems and leverage my interdisciplinary expertise to solve practical challenges and ease the human cost towards a higher level of automation in the development of DL-based software.

4.4 DL for Software Engineering I have gained abundant knowledge, hands-on experiences, and research ideas in DL methods for software engineering (c.f. [13, 12]) through industry internships, ranging from large language models (LLMs) for comment modeling, LLMs for program synthesis, and automatic DL pipeline generation (AutoML). There is a synergy between certifiably trustworthy DL and traditional software engineering: On the one hand, certifiably trustworthy DL systems should be easy to deploy in software engineering scenarios since their certified properties are directly compatible with traditional program analysis. Specifically, a certified DL system rigorously satisfies corresponding preconditions and assertions and can be plugged into off-the-shelf static analysis and symbolic execution pipelines to enable end-to-end system-level certification. On the other hand, traditional programs can augment the certified trustworthiness of DL systems. Since traditional programs are usually interpretable and easier to be certified, I plan to investigate approaches that leverage such advantages from traditional programs to boost the certified trustworthiness of the end-to-end DL system. I believe trustworthy DL and the rich body of research in software engineering can be combined towards a reliable, interpretable, and certifiably trustworthy intelligent system that revolutionizes computing technology.

4.5 DL for System Security The deep learning methods can benefit the research in system security. Many tasks in traditional system security require the detection of malicious instances from huge amounts of features, such as intrusion detection, malware detection, and phishing detection. In these scenarios, though a direct application of DL methods may achieve decent end-to-end precision or accuracy, the intrinsic property of security applications further requires DL methods to be trustworthy and interpretable. For example, a non-trustworthy DL method may easily be circumvented by targeted attackers, and a non-interpretable DL method may fail the audit by authorities and cannot be fixed when drawbacks are discovered. Hence, I plan to extend my certifiably trustworthy pipelines and DL interpretation methods [31] to build trustworthy, interpretable, and efficient DL systems for system security by integrating domain knowledge from system security into the pipeline.

References

[1] Bhaskar Ray Chaudhury, **Linyi Li**, Mintong Kang, Bo Li, and Ruta Mehta. Fairness in federated learning via core-stability. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[2] Fan Wu, **Linyi Li**, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li. CROP: Certifying robust policies for reinforcement learning through functional smoothing. In *International Conference on Learning Representations (ICLR)*, 2022.

[3] Fan Wu*, **Linyi Li***, Chejian Xu, Huan Zhang, Bhavya Kailkhura, Krishnaram Kenthapadi, Ding Zhao, and Bo Li. COPA: Certifying robust policies for offline reinforcement learning against poisoning attacks. In *International Conference on Learning Representations (ICLR)*, 2022.

[4] Hanjiang Hu, Zuxin Liu, **Linyi Li**, Jiacheng Zhu, and Ding Zhao. Robustness certification of visual perception models via camera motion smoothing. In *6th Annual Conference on Robot Learning (CoRL)*, 2022.

[5] Huan Zhang*, Shiqi Wang*, Kaidi Xu*, **Linyi Li**, Bo Li, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. General cutting planes for bound-propagation-based neural network verification. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

[6] Huichen Li*, **Linyi Li***, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. Nonlinear projection based gradient estimation for query efficient blackbox attacks. In *24th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2021.

[7] Jiawei Zhang*, **Linyi Li***, Huichen Li, Xiaolu Zhang, Shuang Yang, and Bo Li. Progressive-scale boundary blackbox attack via projective gradient estimation. In *38th International Conference on Machine Learning (ICML)*. PMLR, 2021.

[8] Jiawei Zhang, **Linyi Li**, Ce Zhang, and Bo Li. CARE: Certifiably robust learning with reasoning via variational inference. In *First IEEE Conference on Secure and Trustworthy Machine Learning (SatML)*, 2023.

[9] Maurice Weber, **Linyi Li**, Boxin Wang, Zhikuan Zhao, Bo Li, and Ce Zhang. Certifying out-of-domain generalization for blackbox functions. In *39th International Conference on Machine Learning (ICML)*, 2022.

[10] Mintong Kang, **Linyi Li**, and Bo Li. Fashapley: Fast and approximated shapley based model pruning towards certifiably robust DNNs. In *First IEEE Conference on Secure and Trustworthy Machine Learning (SatML)*, 2023.

[11] Mintong Kang*, **Linyi Li***, Maurice Weber, Yang Liu, Ce Zhang, and Bo Li. Certifying some distributional fairness with subpopulation decomposition. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

[12] Ripon K Saha, Akira Ura, Sonal Mahajan, Chenguang Zhu, **Linyi Li**, Yang Hu, Hiroaki Yoshida, Sarfraz Khurshid, and Mukul R Prasad. SAPIENTML: Synthesizing machine learning pipelines by learning from human-written solutions. In *IEEE/ACM 44th International Conference on Software Engineering (ICSE)*, 2022.

[13] **Linyi Li**, Zhenwen Li, Weijie Zhang, Jun Zhou, Pengcheng Wang, Jing Wu, Guanghua He, Xia Zeng, Yuetang Deng, and Tao Xie. Clustering test steps in natural language toward automating test automation. In *28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Industry Track (ESEC/FSE Industry)*, 2020.

[14] **Linyi Li***, Maurice Weber*, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. TSS: Transformation-specific smoothing for robustness certification. In *2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2021.

[15] **Linyi Li**, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks. In *44th IEEE Symposium on Security and Privacy (SP)*, 2023.

[16] **Linyi Li**, Jiawei Zhang, Tao Xie, and Bo Li. Double sampling randomized smoothing. In *39th International Conference on Machine Learning (ICML)*, 2022.

[17] **Linyi Li**, Yuhao Zhang, Luyao Ren, Yingfei Xiong, and Tao Xie. Reliability assurance for deep neural network architectures against numerical defects. In *IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 2023.

[18] **Linyi Li***, Zexuan Zhong*, Bo Li, and Tao Xie. Robustra: Training provable robust neural networks over reference adversarial space. In *28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4711–4717, 2019.

[19] Wenda Chu, **Linyi Li**, and Bo Li. TPC: Transformation-specific smoothing for point cloud models. In *39th International Conference on Machine Learning (ICML)*, 2022.

[20] Xiaojun Xu, **Linyi Li**, and Bo Li. LOT: Layer-wise orthogonal training on improving l2 certified robustness. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

[21] Zhuolin Yang*, **Linyi Li***, Xiaojun Xu, Bhavya Kailkhura, Tao Xie, and Bo Li. On the certified robustness for ensemble models and beyond. In *International Conference on Learning Representations (ICLR)*, 2022.

[22] Zhuolin Yang*, **Linyi Li***, Xiaojun Xu*, Shiliang Zuo, Qian Chen, Pan Zhou, Benjamin I. P. Rubinstein, Ce Zhang, and Bo Li. Trs: Transferability reduced ensemble via promoting gradient diversity and model smoothness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[23] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *35th International Conference on Machine Learning (ICML)*, 2018.

[24] Avrim Blum, Travis Dick, Naren Manoj, and Hongyang Zhang. Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images. *Journal of Machine Learning Research*, 21:211:1–211:21, 2020.

[25] Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[26] AdvML Rising Star Award Committee. 2022 advml rising star award. `https://sites.google.com/view/advml/advml-rising-star-award`.

[27] Data Science Institute, the University of Chicago. Rising stars in data science, DSI. `https://datascience.uchicago.edu/rising-stars/#rising-stars`.

[28] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[29] Jamie Hayes. Extensions and limitations of randomized smoothing for robustness guarantees. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Workshops (CVPRW)*, 2020.

[30] Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In *37th International Conference on Machine Learning (ICML)*, 2020.

[31] Klas Leino, Shayak Sen, Anupam Datta, Matt Fredrikson, and **Linyi Li**. Influence-directed explanations for deep convolutional networks. In *2018 IEEE International Test Conference (ITC)*, 2018.

[32] White House Office of Science and Technology Policy. *Blueprint for an AI bill of rights*. The White House, Washington, D.C., 2022.

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations (ICLR)*, 2014.

[34] VNN-COMP'22. 3rd international verification of neural networks competition (VNN-COMP'22). `https://sites.google.com/view/vnn2022`.

[35] Yihan Wu, Aleksandar Bojchevski, Aleksei Kuvshinov, and Stephan Günnemann. Completing the picture: Randomized smoothing suffers from the curse of dimensionality for a large family of distributions. In *24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[36] Tao Xie, Nikolai Tillmann, and Pratap Lakshman. Advances in unit testing: theory and practice. In *Proceedings of the 38th international conference on software engineering companion*, pages 904–905, 2016.

[37] Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *37th International Conference on Machine Learning (ICML)*, 2020.