Towards Certifiably Robust Face Recognition

2024 European Conference on Computer Vision

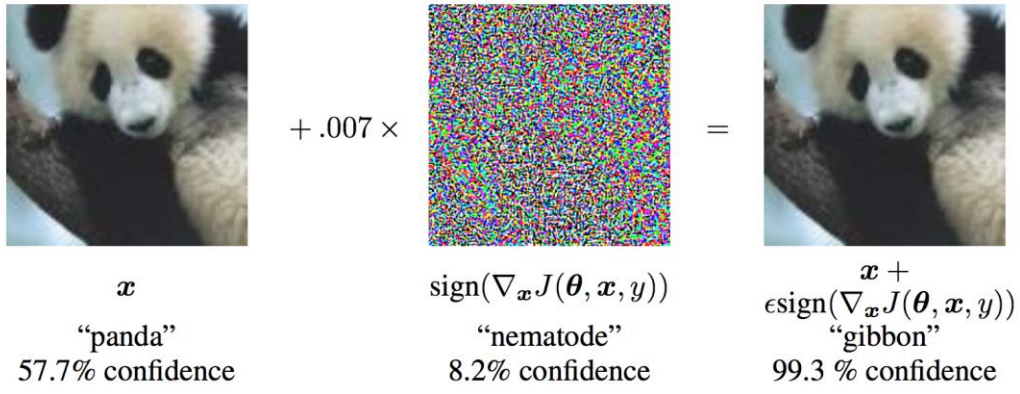
Seunghun Paik
Chanwoo Hwang
Jae Hong Seo*
Dongsoo Kim
Sunpill Kim

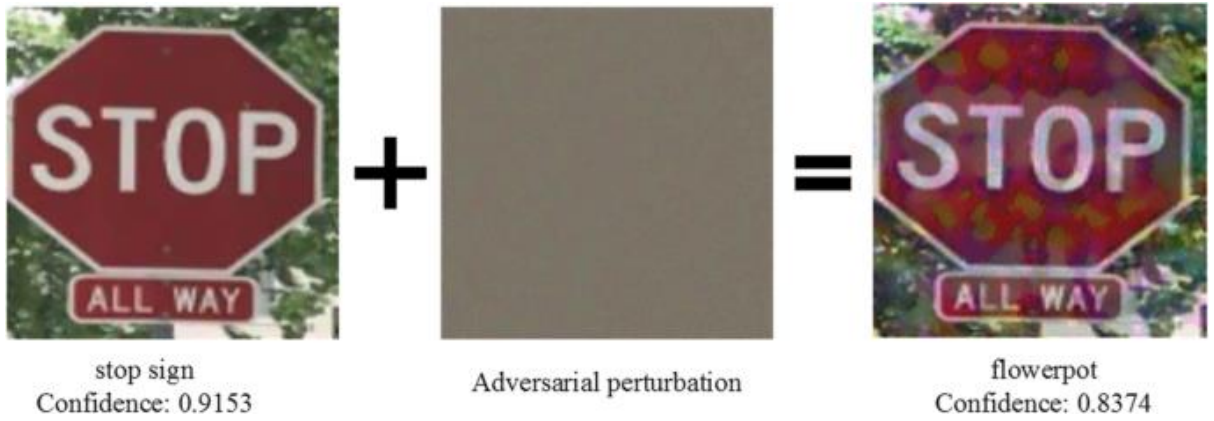
Department of Mathematics
Hanyang University

*Corresponding Author.
**This work was supported by Korea Creative Content Agency (RS-2024-00332210)

Adversarial Examples

- A subtle, intended noise that makes the target neural network totally malfunctioning!



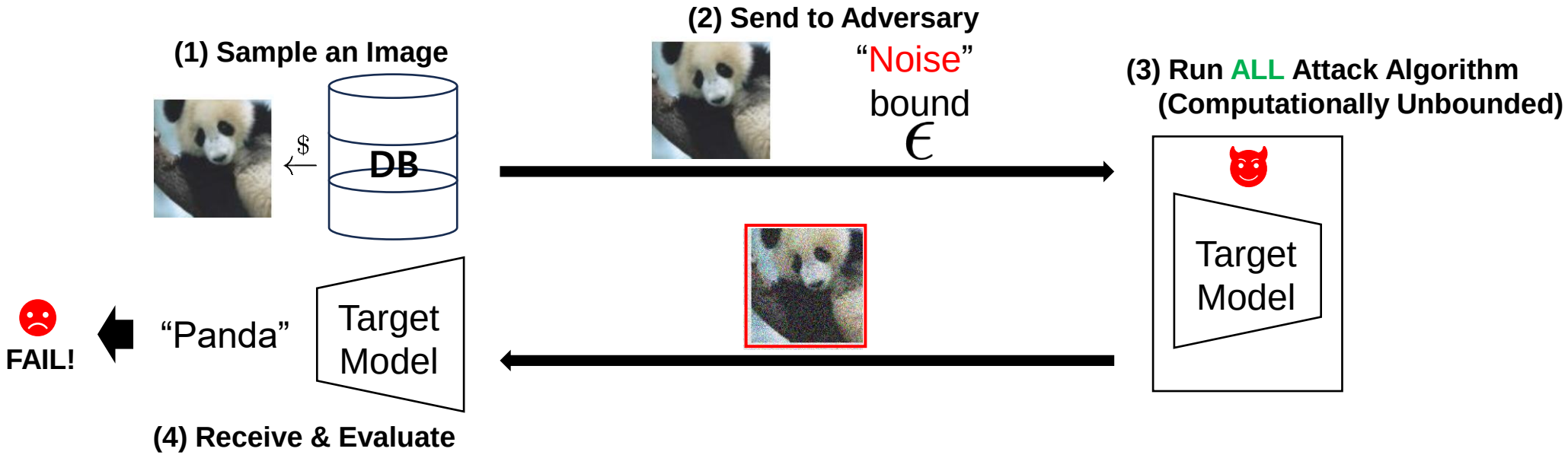


- Its prevalence is well-known, even in security/safety critical applications.
  - ➢ Deep learning-based face recognition.
  - ➢ Autonomous driving.

- In such applications, "**provable**" defense of them is necessary!

Certifiable Robustness

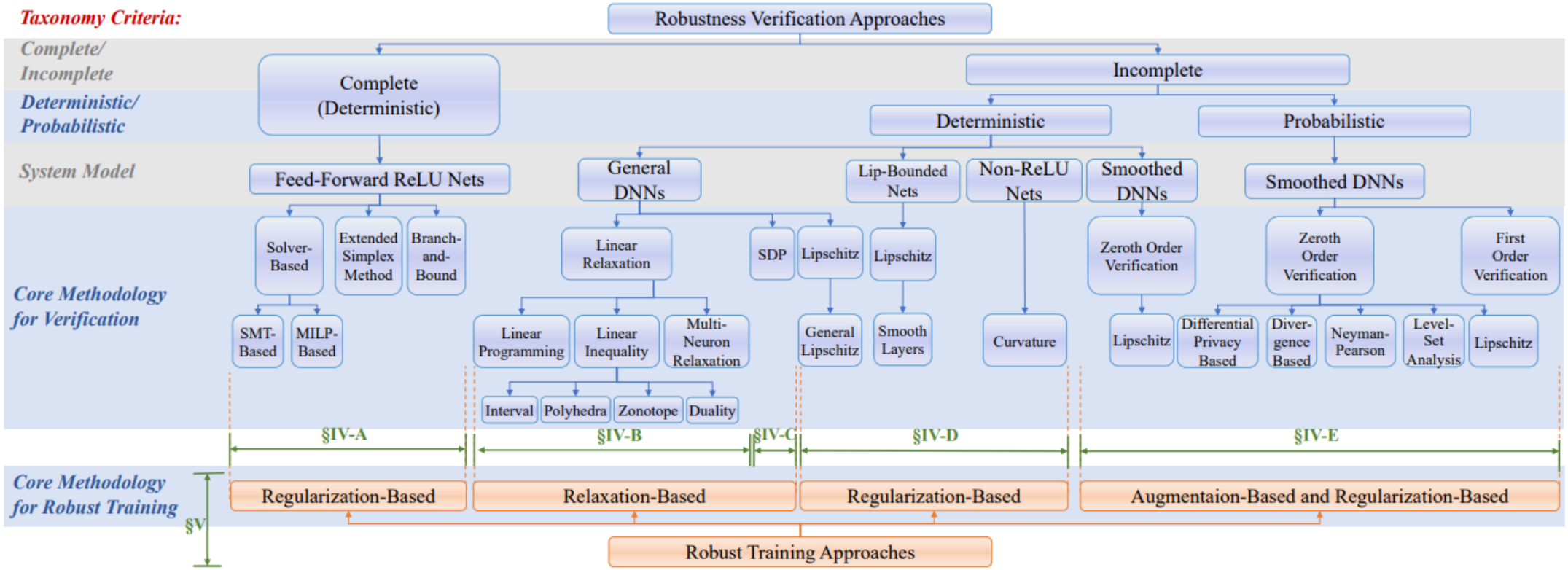
- **Goal**: Find a condition that there is no adversarial examples within a noise bound ϵ.
  - ➢ There is no (even computationally unbounded) adversarial attacks for the given input.



- A typical approach? Analyze the range of logit value of the adversarial example.

Methodologies for Certifiable Robustness

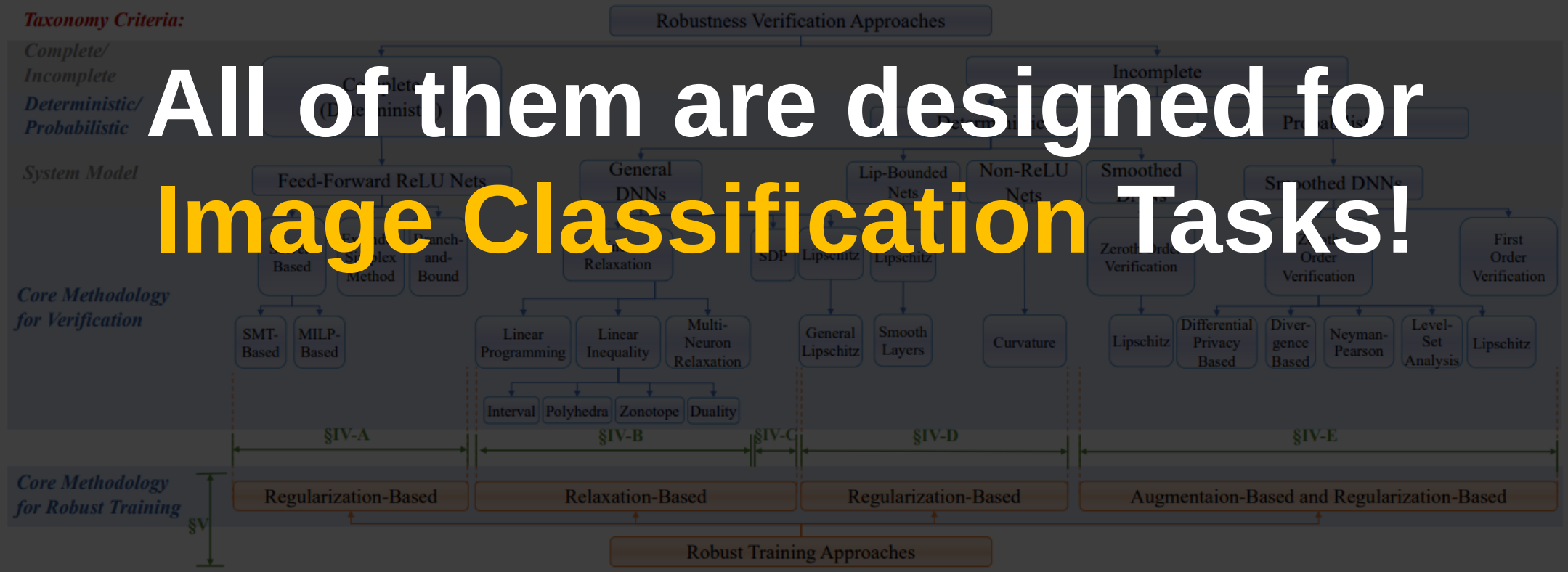
- Due to its attractive feature, several efforts have been made for achieving this.
  - ➢ Figure from a SoK paper on certifiable robustness [LXL23].



[LXL23] Li, L., Xie, T., & Li, B. (2023, May). Sok: Certified robustness for deep neural networks. In *2023 IEEE symposium on security and privacy (SP)* (pp. 1289-1310). IEEE.

Methodologies for Certifiable Robustness

- Due to its attractive feature, several efforts have been made for achieving this.
  - ➢ Figure from a SoK paper on certifiable robustness [LXL23].

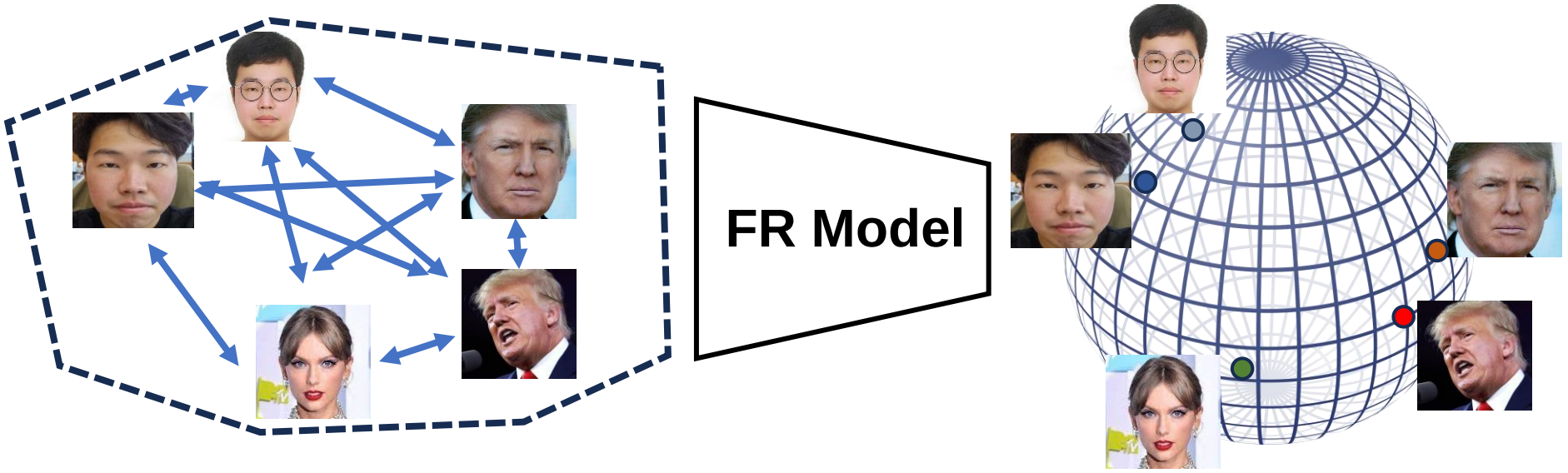


All of them are designed for Image Classification Tasks!

[LXL23] Li, L., Xie, T., & Li, B. (2023, May). Sok: Certified robustness for deep neural networks. In *2023 IEEE symposium on security and privacy (SP)* (pp. 1289-1310). IEEE.

Why It is Hard for Face Recognition?

- FR model utilizes **metric learning** & It is deployed in **open-set setting**.
  - ➢ Let the FR model catch up “implicit” distance relationships between faces
  - ➢ Feature vectors are represented as unit vectors; cosine similarity (or, angular distance) is used.



- There are no predetermined “classes” or “logit” values.
  - ➢ Previous certifiable robustness techniques for image classification is no longer available… ☹

Our Contribution

- **First** certifiable robustness result for "open-set" face recognition scenarios

- **Main Theorem**: If the FR model is 1-Lipschitz in ℓ_2, then it is certifiably robust.
  - ➢ Same condition as image classification tasks [TSS18, SSF21].
  - ➢ Novel proof technique tailored for dealing with angular distance.

- Careful analysis on the certified radius (upper bound of the size of noise)
  - ➢ We found that the certified radius is proportional to the norm of the feature vector.
  - ➢ We also derived the upper bound of the achievable certified radius.

- Proof of concept implementation & empirical verification

[TSS18] Tsuzuku, Y., Sato, I., & Sugiyama, M. (2018). Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in neural information processing systems*, *31*.
[SSF21] Singla, S., Singla, S., & Feizi, S. (2021). Improved deterministic l2 robustness on CIFAR-10 and CIFAR-100. *arXiv preprint arXiv:2108.04062*.

Thank You ☺

If you are interested, feel free to visit our poster!
Section 7, Poster No. #196, Fri Oct 4, 10:30 – 12:00

Our Poster

Github Repository