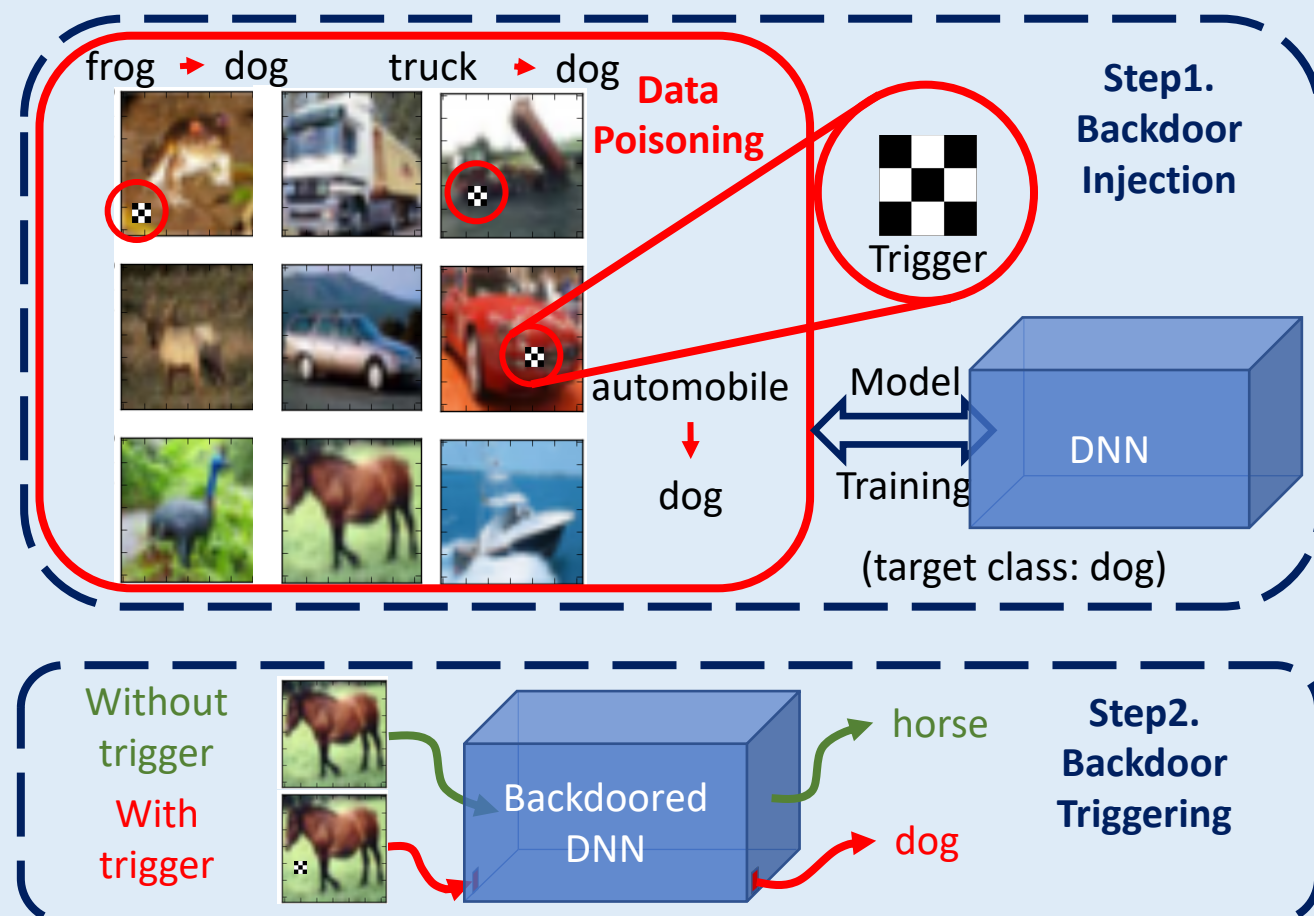


# Defending Neural Backdoors via Generative Distribution Modeling

## Neural Backdoor Background

### Neural Backdoor Attacks

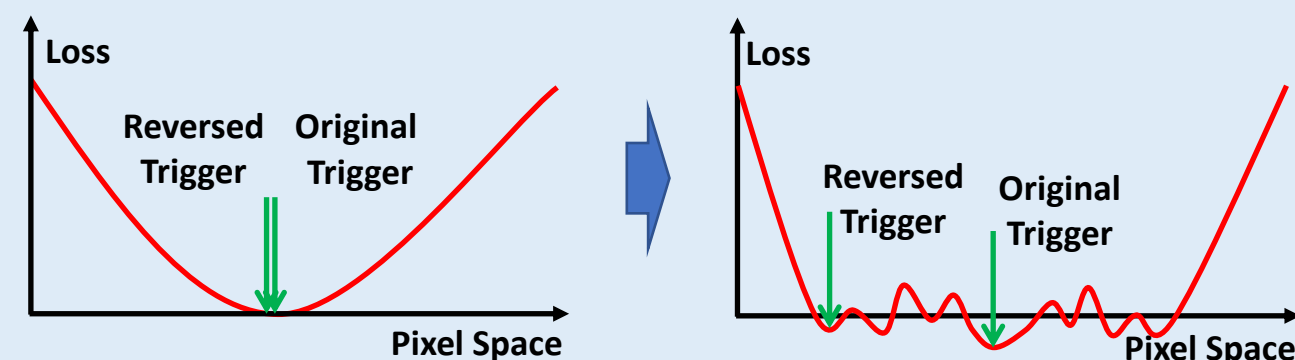


### Neural Backdoor Defenses

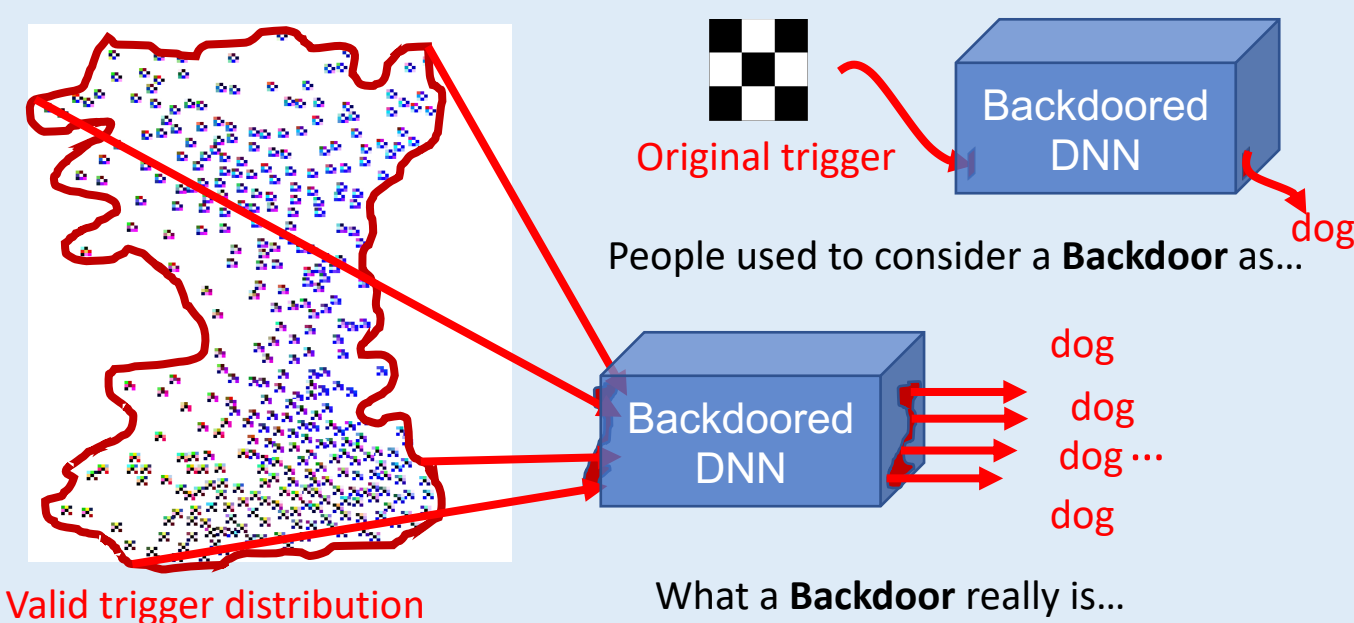
- Reversed engineer the backdoor trigger
- Apply the trigger to training data
- Retrain the model with correct label

## Backdoor Distribution Hypothesis

### Deficiency of existing defenses

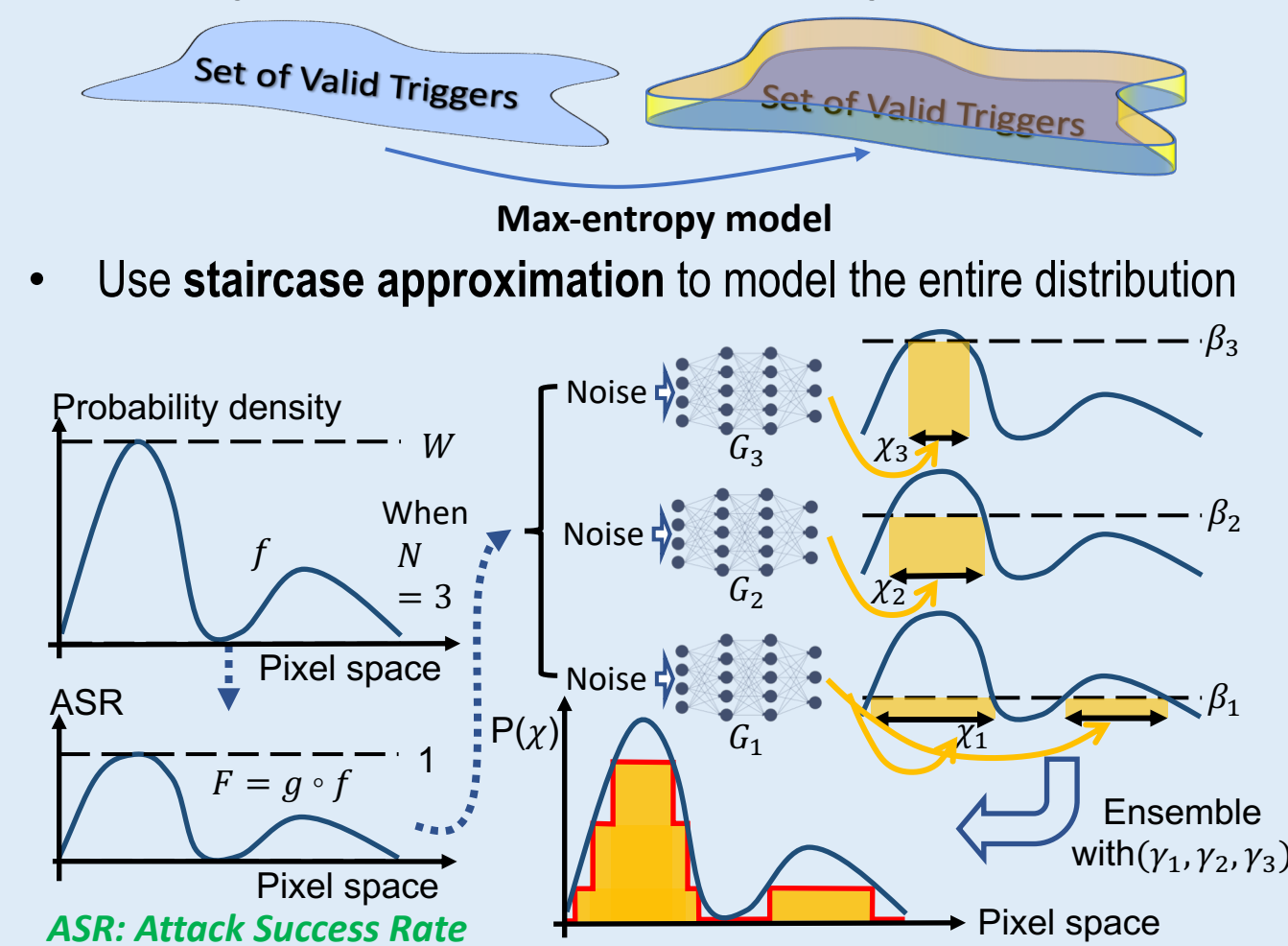


### Hypothesis: backdoors as distributions



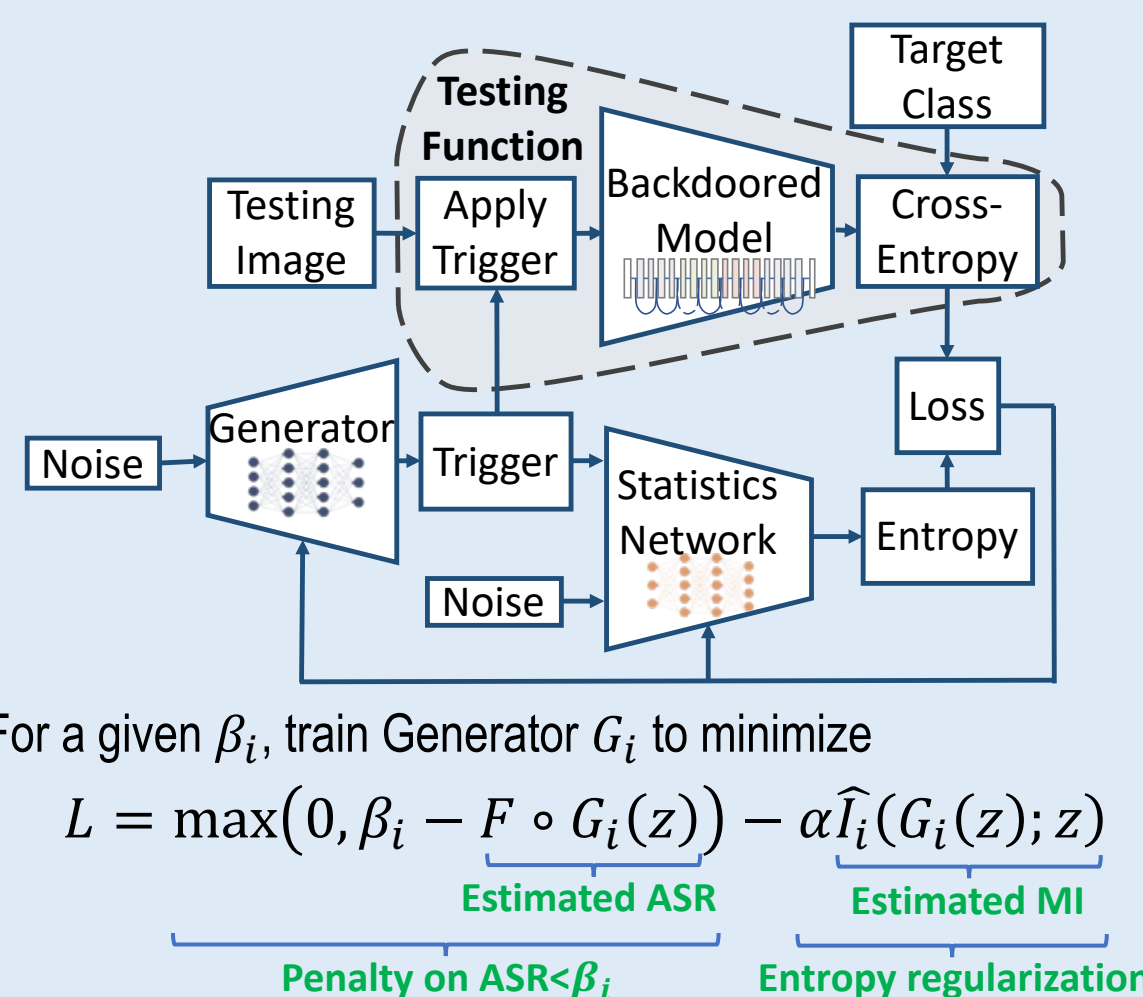
## Max-Entropy Staircase Approximator

- Typical distribution methods (GAN, VAE) fails due to the unobtainable groundtruth dataset
- Entropy maximization** can explore a single constraint set



## Implementation

- Estimate ASR by cross-entropy loss (Testing Function)
- Estimate entropy by [1] (Statistics Network)
- Generate backdoor trigger by a neural network (Generator)

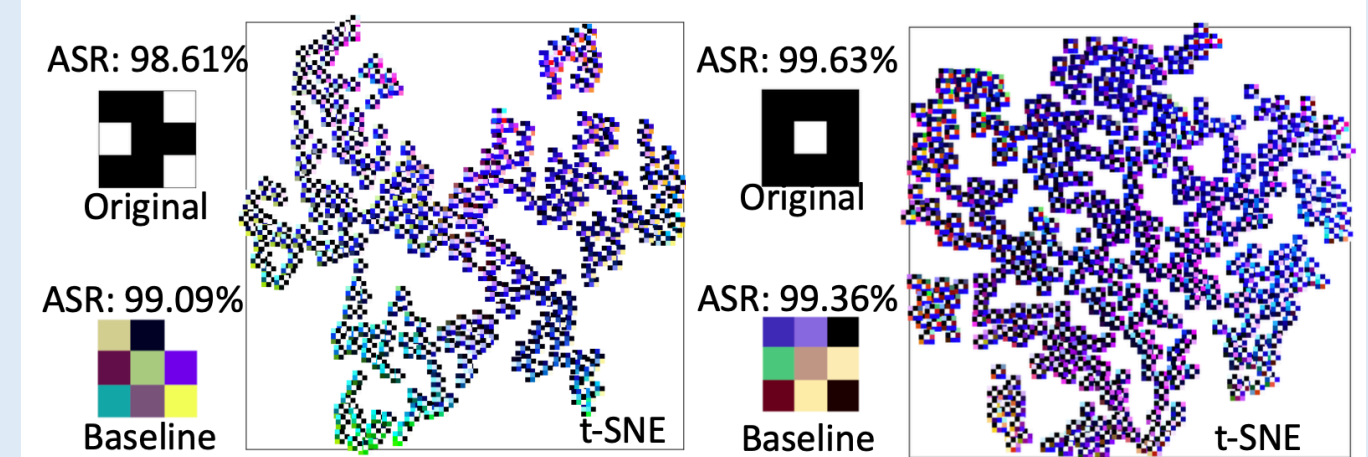


- For a given  $\beta_i$ , train Generator  $G_i$  to minimize  $L = \max(0, \beta_i - F \circ G_i(z)) - \alpha \hat{I}_i(G_i(z); z)$
- $F$ : testing function.  $\hat{I}_i$ : estimated mutual information (MI)
- $z$ : Gaussian input noise.  $\alpha$ : regularization strength

[1] Belghazi et al., Mine: mutual information neural estimation, ICML 2018

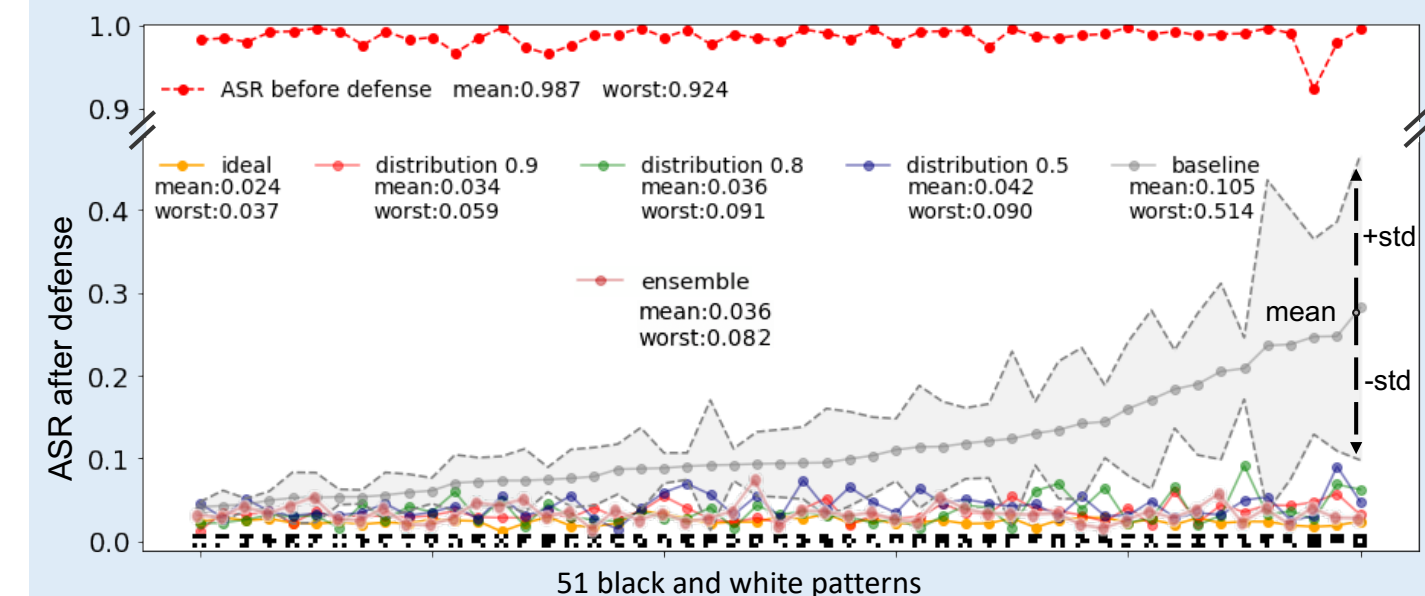
## Backdoor Visualization

- Dataset: CIFAR10
- Trigger size: 3×3
- Trigger type: black-white
- Baseline: pixel space SGD
- Visualization: t-SNE



## Backdoor Defense

- Retrain the model with the reversed trigger distribution gives more robust defense (baseline: single reversed trigger)
- CIFAR10 with black-white triggers



- CIFAR10/100 with random color triggers

