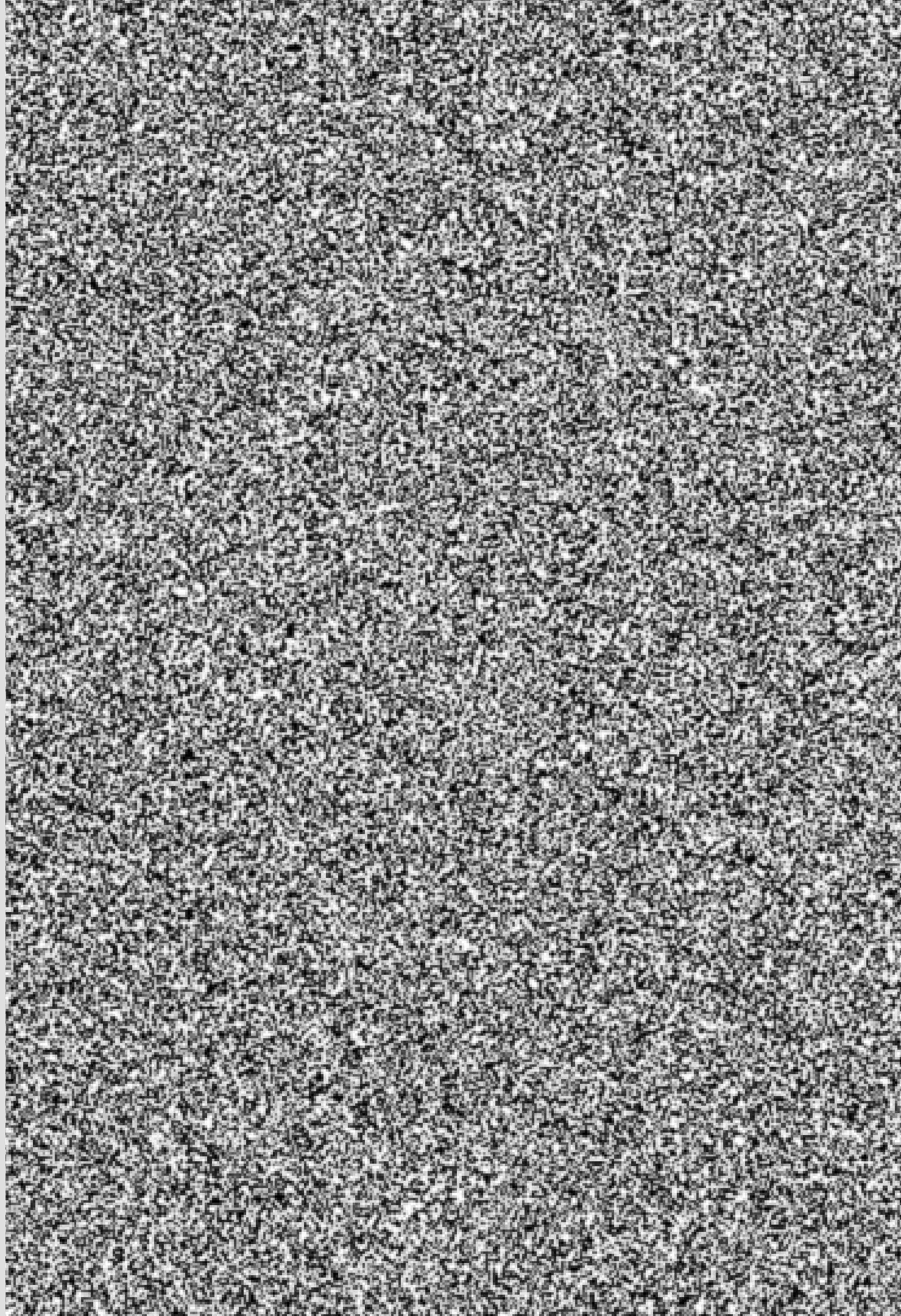


Projet 1 :

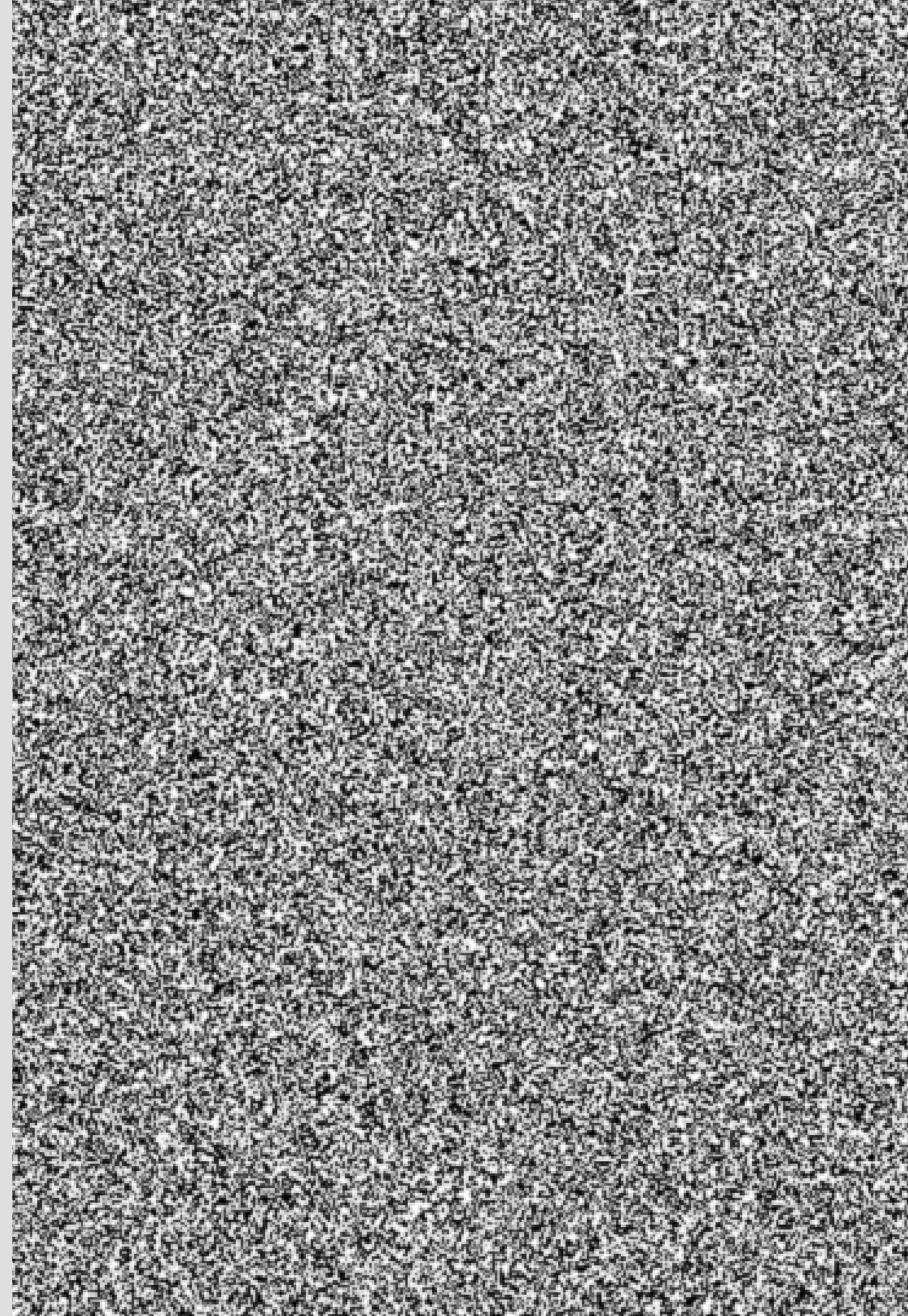
Construire la meilleure défense contre les attaques adverses



Plan

1. Motivations et intuition
2. Présentation du papier illustrant notre intuition
3. Description de la méthode
4. Difficultés d'implémentation et résultats

1. Motivations et intuition



Choix de la méthode

- L'entraînement adverse est une des meilleures méthodes de défense
- Comment améliorer encore l'entraînement adverse ?
- Intuition : pour entraîner une bonne défense il faut savoir simuler une bonne attaque
- Data augmentation ?

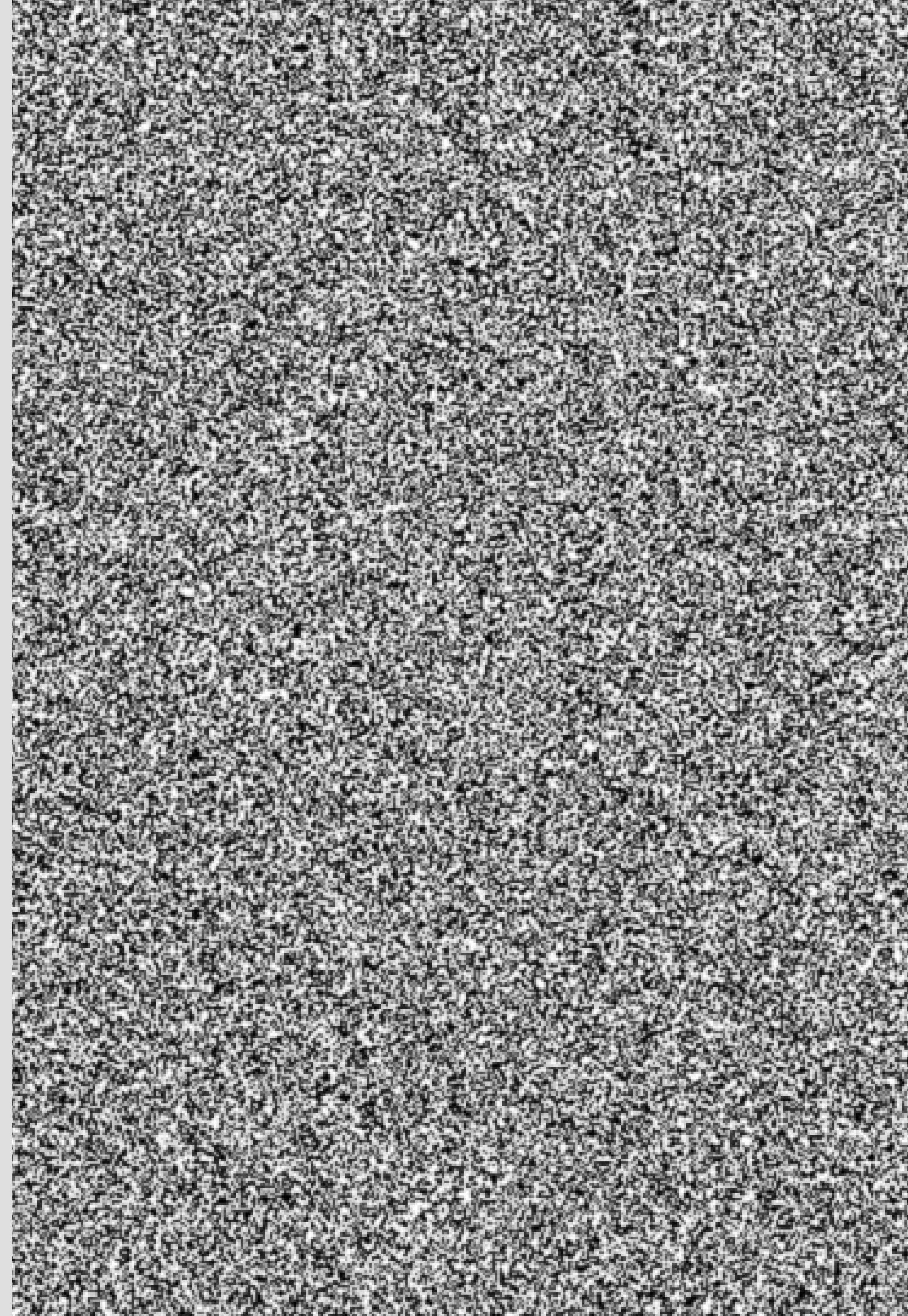
→ Generative Adversarial Network (GAN)

Intuition

- GAN couramment utilisés pour génération d'image
- Entraîne simultanément un générateur et un discriminateur
- Le générateur et le discriminateur se renforcent mutuellement
- Le générateur essaie de tromper le discriminateur

Le discriminateur apprend à ne pas se tromper sur des images attaquées, et sur des fausses images

2. Présentation du papier illustrant notre intuition



Rob-GAN: Generator, Discriminator, and Adversarial Attacker

Xuanqing Liu and Cho-Jui Hsieh
University of California, Los Angeles

- Le papier se concentre autant sur l'aspect générateur, que sur l'aspect défense
- On veut combiner l'adversarial training, et l'entraînement sur de fausses images
- Grâce aux fausses images, on obtient un classifieur qui généralise mieux à des images jamais vues

**"a novel framework called Rob-GAN,
which integrates generator,
discriminator and adversarial attacker
as a three-player game"**

Autres avantages :

- Meilleure vitesse de convergence du générateur
- Meilleur générateur d'images fake

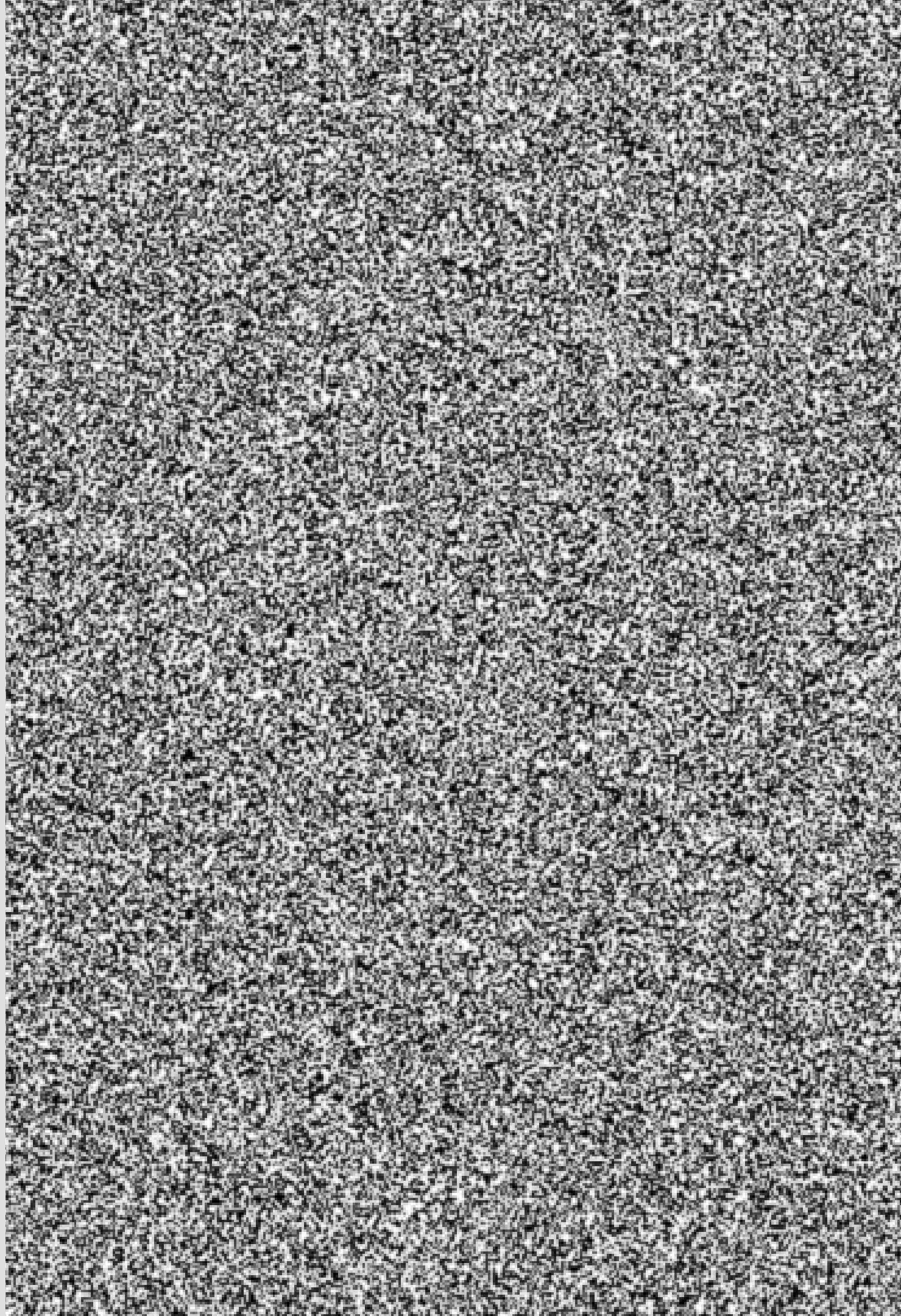
Rob-GAN: Generator, Discriminator, and Adversarial Attacker

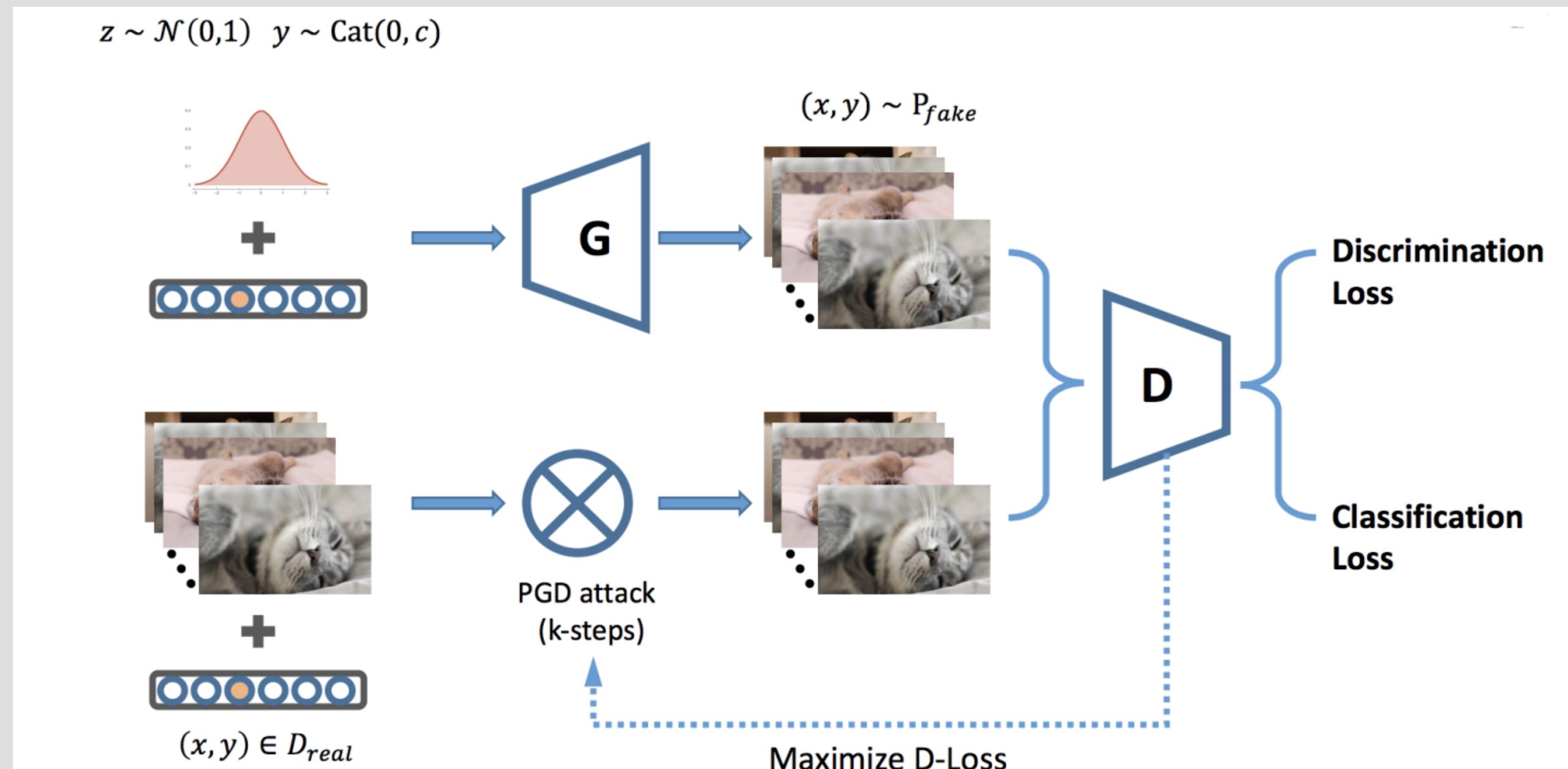
Xuanyang Liu and Cho-Jui Hsieh
University of California, Los Angeles

KEY INSIGHTS (extrait du papier)

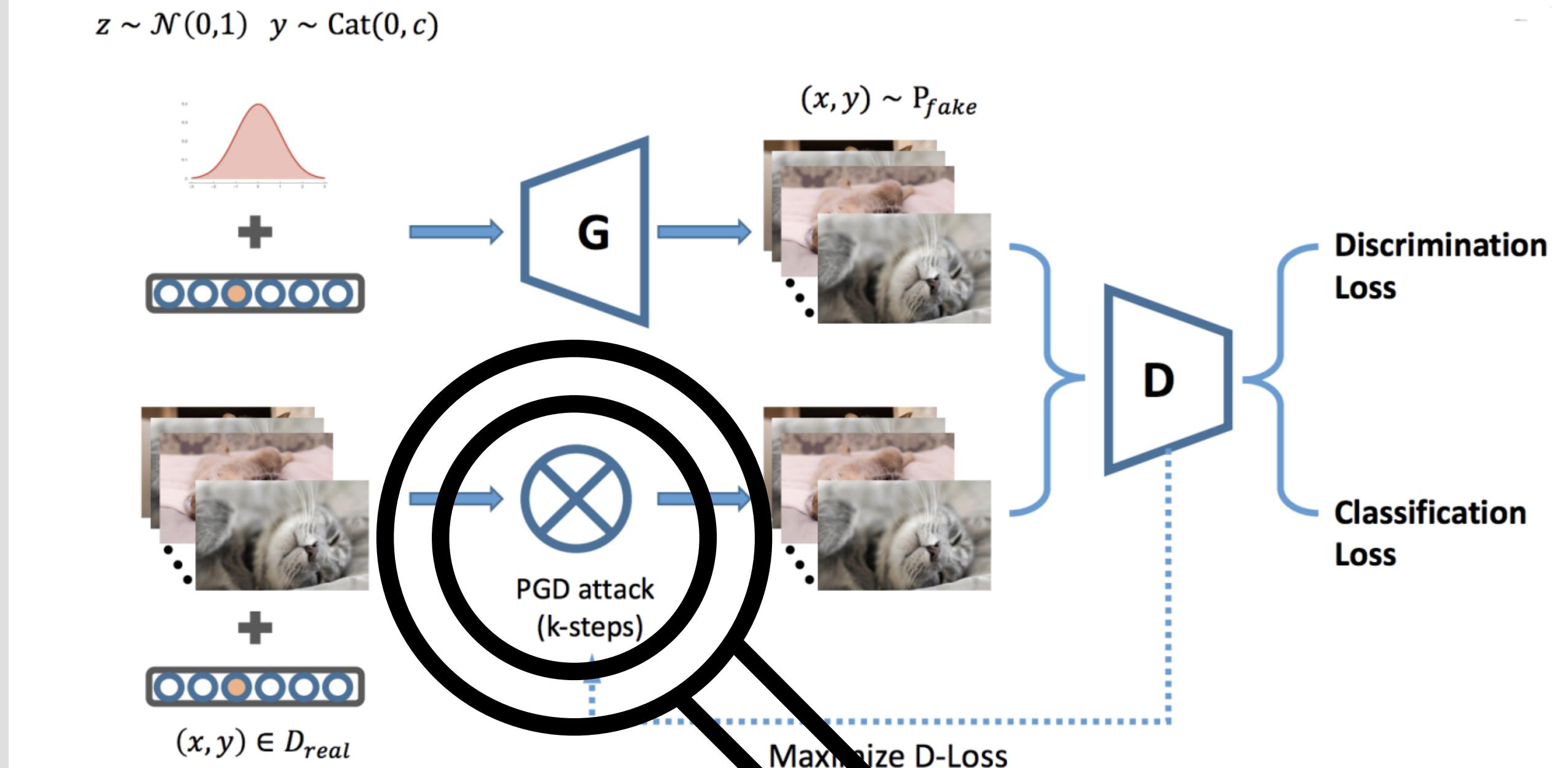
1. The robustness of adversarial trained classifier can be improved if we have a deeper understanding of the image distribution. Therefore a **generator can improve the adversarial training process.**
2. GAN training can be very slow to reach the equilibrium if the discriminator has a large curvature on the image manifold. Therefore an **adversarial trained discriminator can accelerate GAN training.**

3. Description de la méthode





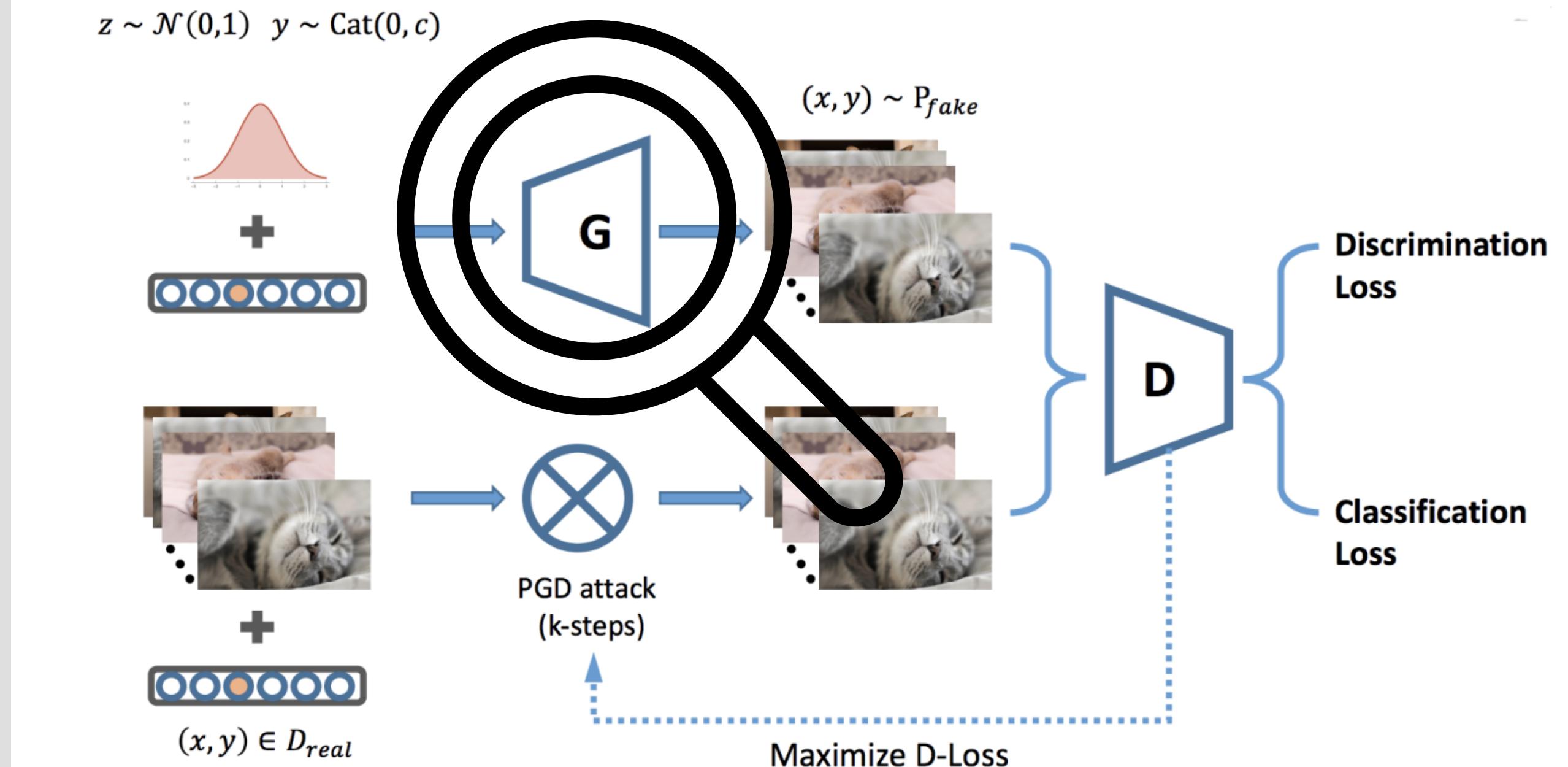
$$z \sim \mathcal{N}(0,1) \quad y \sim \text{Cat}(0, c)$$



Les attaques adverses

- Perturbe des images réelles pour tromper le discriminateur avec un PGD attack
- Sans les attaques adverses, le Rob-GAN est un GAN classique

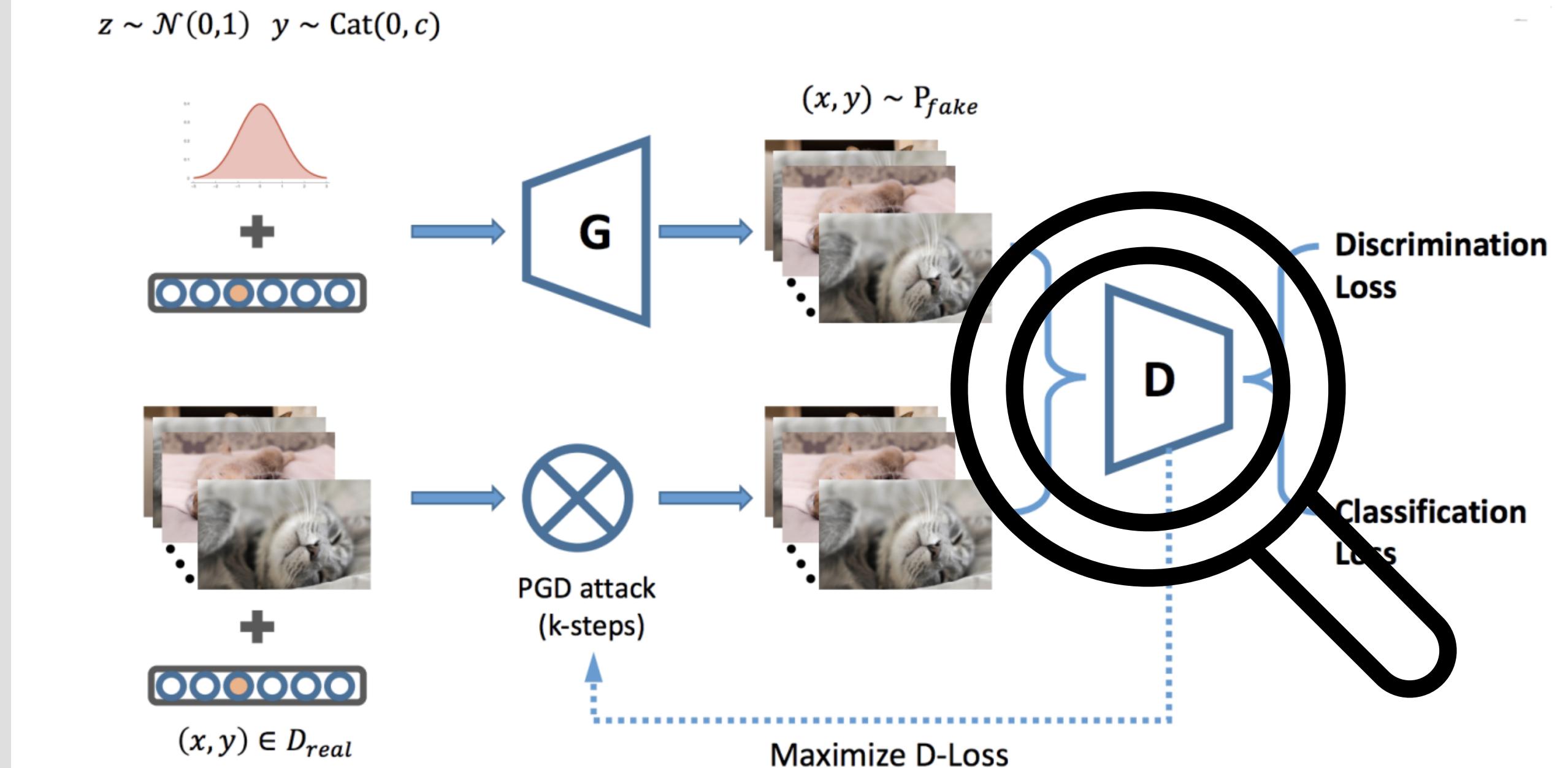
$$z \sim \mathcal{N}(0,1) \quad y \sim \text{Cat}(0, c)$$



Le générateur

- Génère de fausses images pour tromper le discriminateur
- Il est mis à jour de manière régulière pour apprendre à imiter la distribution de données réelles mais perturbées
- Sans le générateur, le Rob-GAN est une méthode d'attaque adverses classique

$$z \sim \mathcal{N}(0,1) \quad y \sim \text{Cat}(0, c)$$



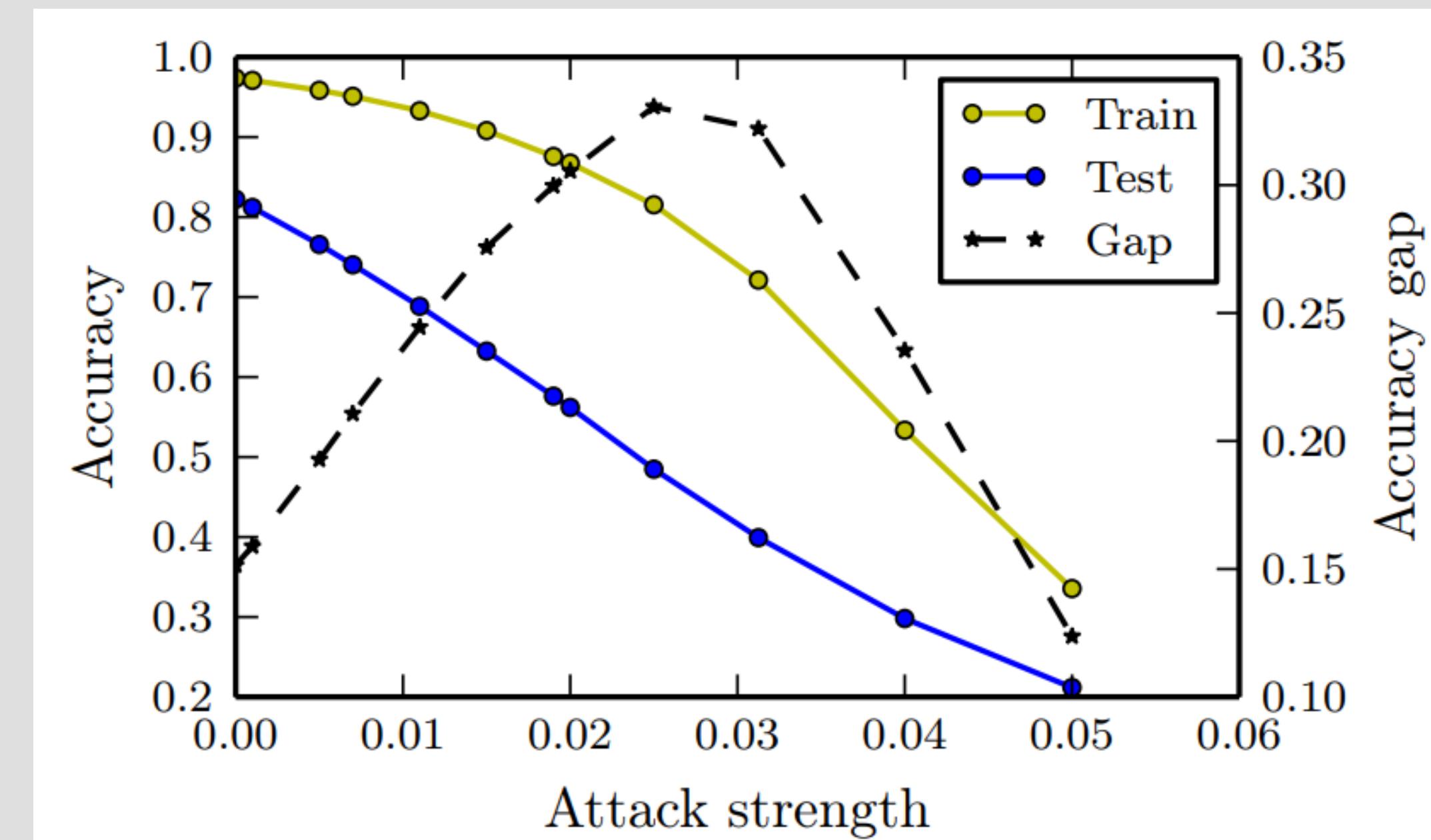
Le discriminateur

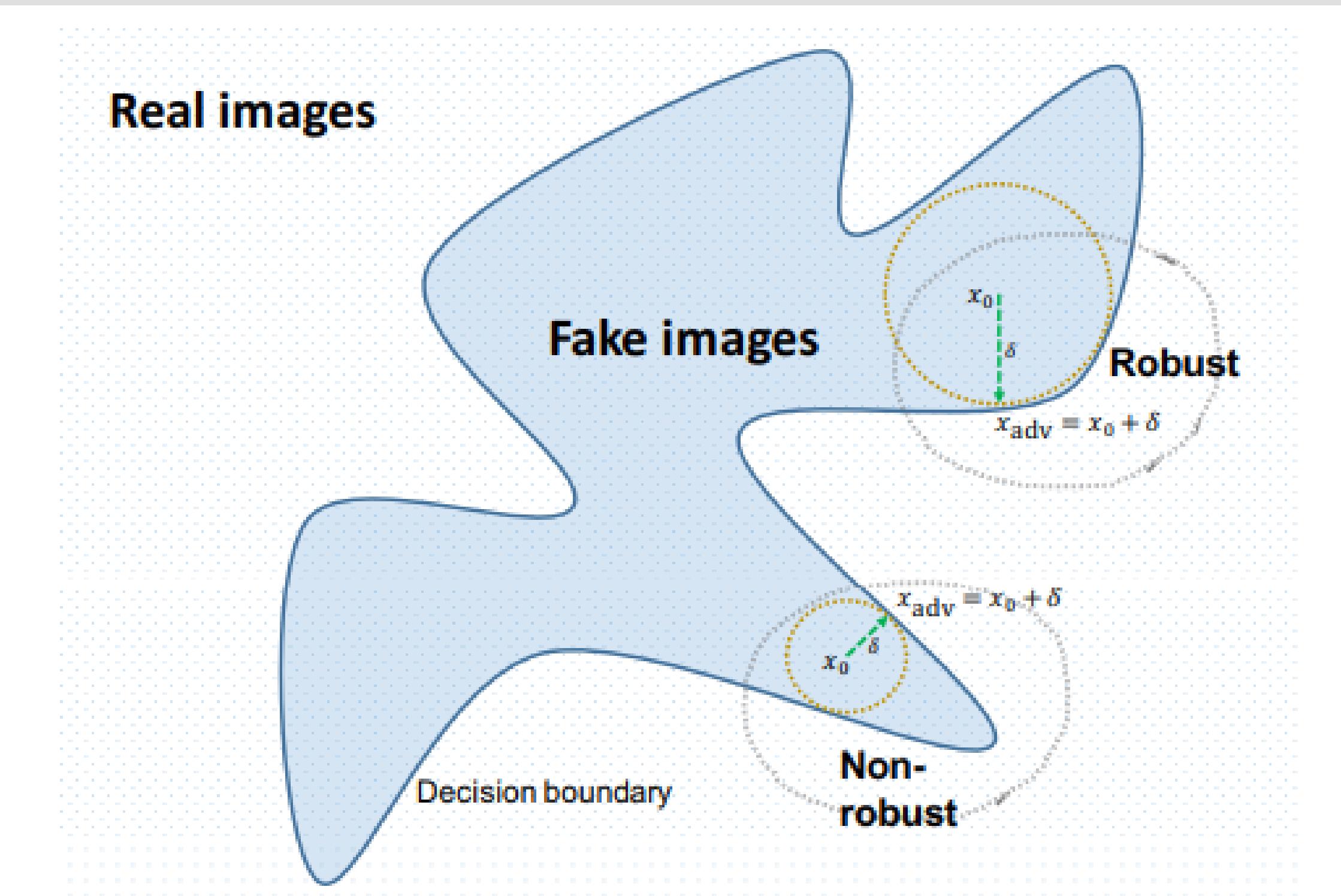
- Deux tâches :
 - Classification des images
 - Discrimination entre les fausses images et les images perturbées

Comment les GAN améliorent-ils la robustesse des réseaux avec un entraînement adversarial ?

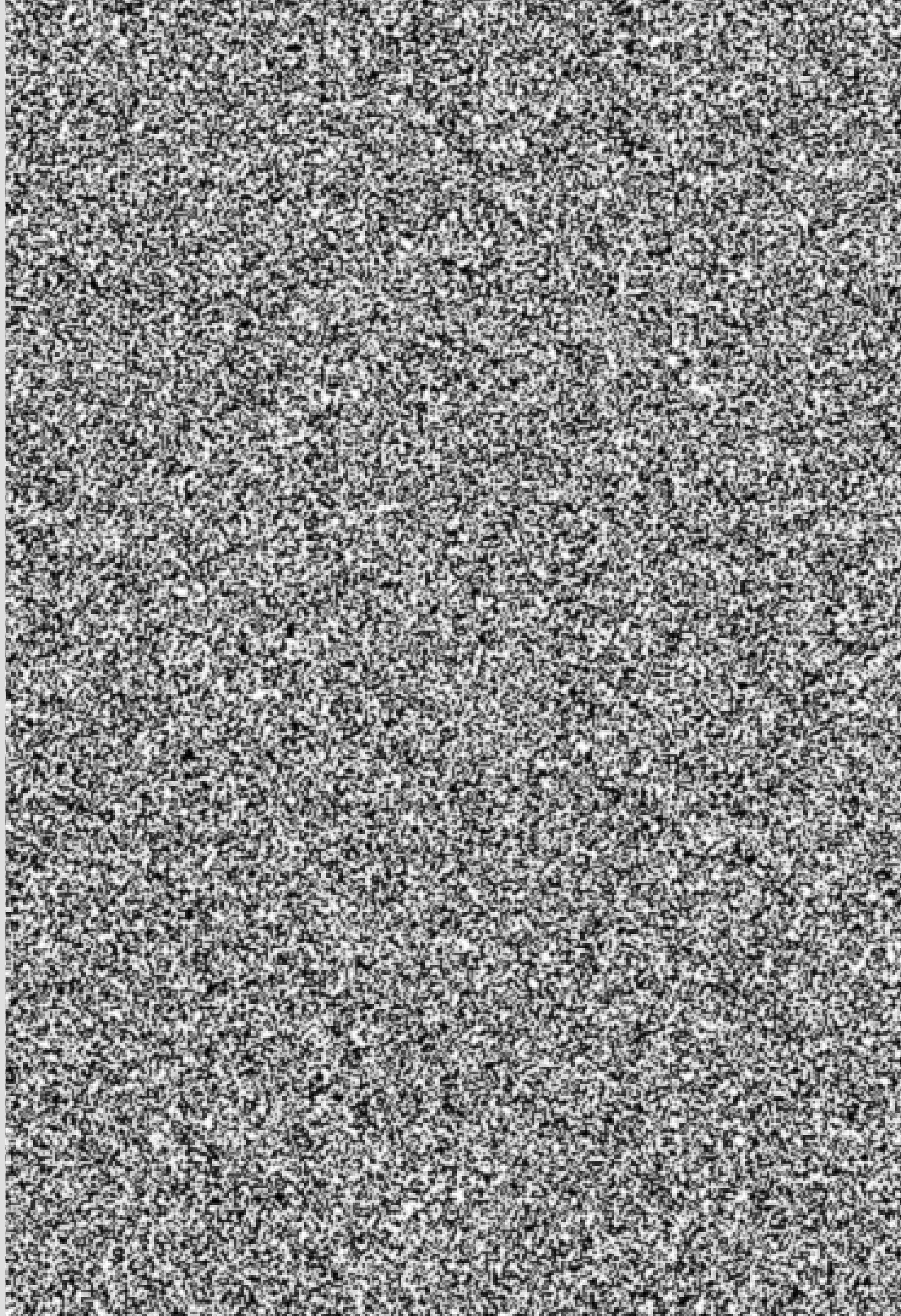
Adversarial training = très robuste sur le training set, mais grosse baisse de performance sur le test set.

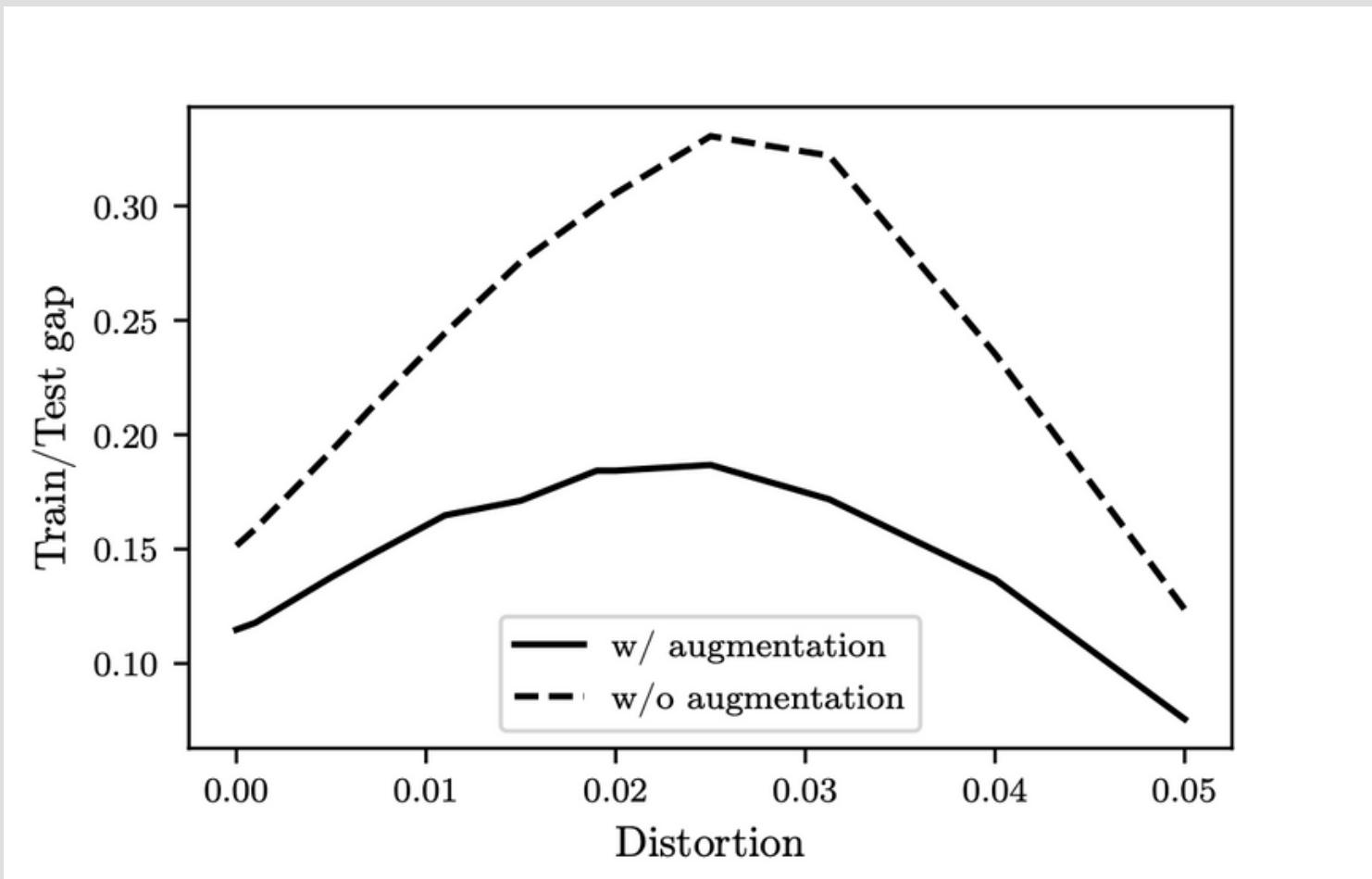
Problème de généralisation





4. Difficultés d'implémentation et résultats

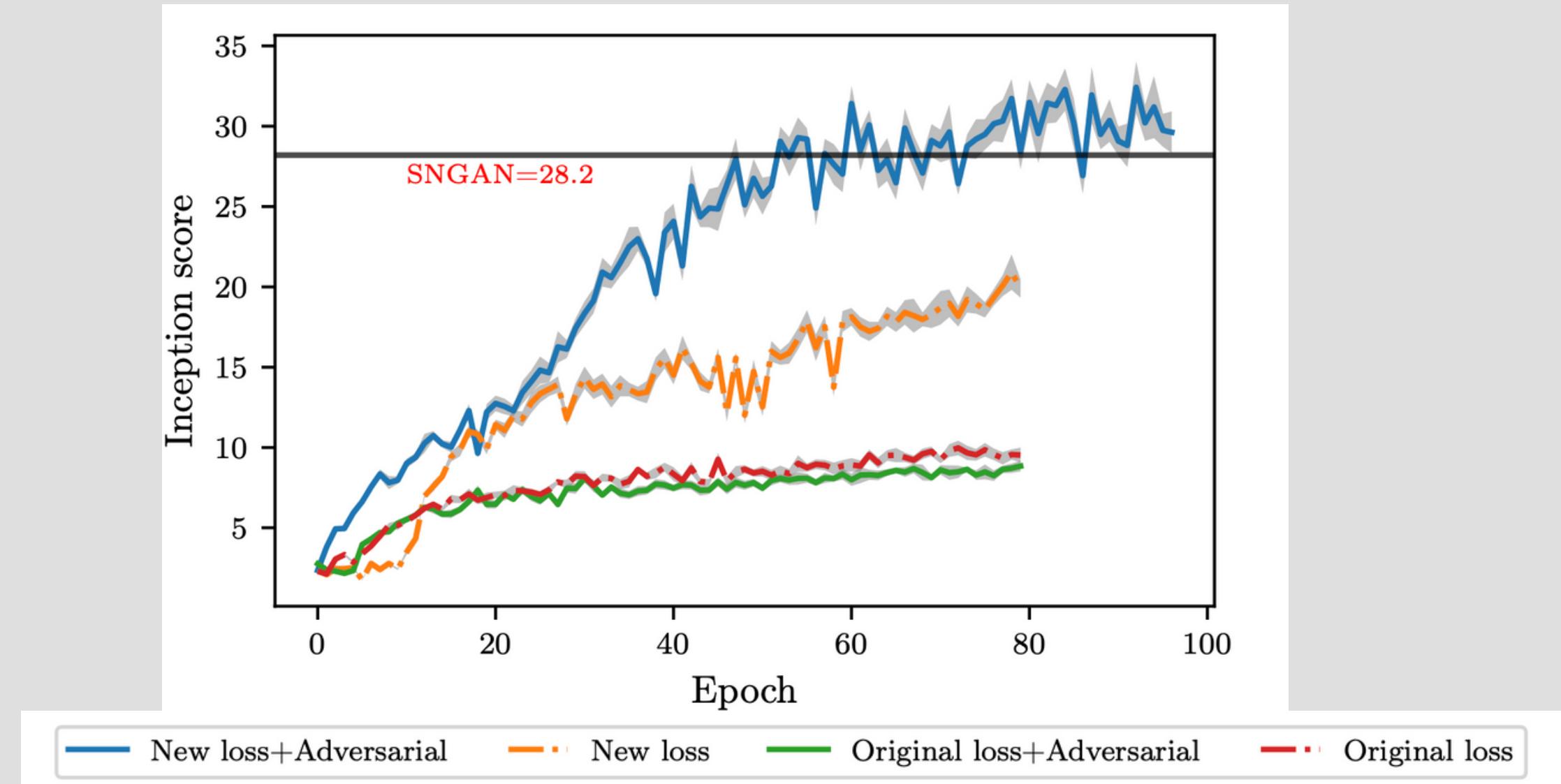




Dataset	Defense	δ_{\max} of ℓ_∞ attacks			
		0	0.02	0.04	0.08
CIFAR10	Adv. training	81.45%	69.15%	53.74%	23.58%
	Rob-GAN (w/ FT)	81.1%	70.41%	57.43%	30.25%

Résultats du papier

- Le réseau rob-gan est bien plus résistant aux attaques avec une distortion moyenne



Résultats du papier

- Le générateur génère des images plus crédibles que les autres réseaux

Difficultés

- Implémentations vieillissantes
- Puissance de calcul et mémoire
- Choisir un Générateur et un Discriminateur
- Loss qui divergent

