# Alleviating Adversarial Attacks on Variational Autoencoders with MCMC

Anna Kuzina, Max Welling, Jakub M. Tomczak

VRIJE UNIVERSITEIT AMSTERDAM
UNIVERSITY OF AMSTERDAM
Hybrid Intelligence

Paper (pre-print)

## Summary

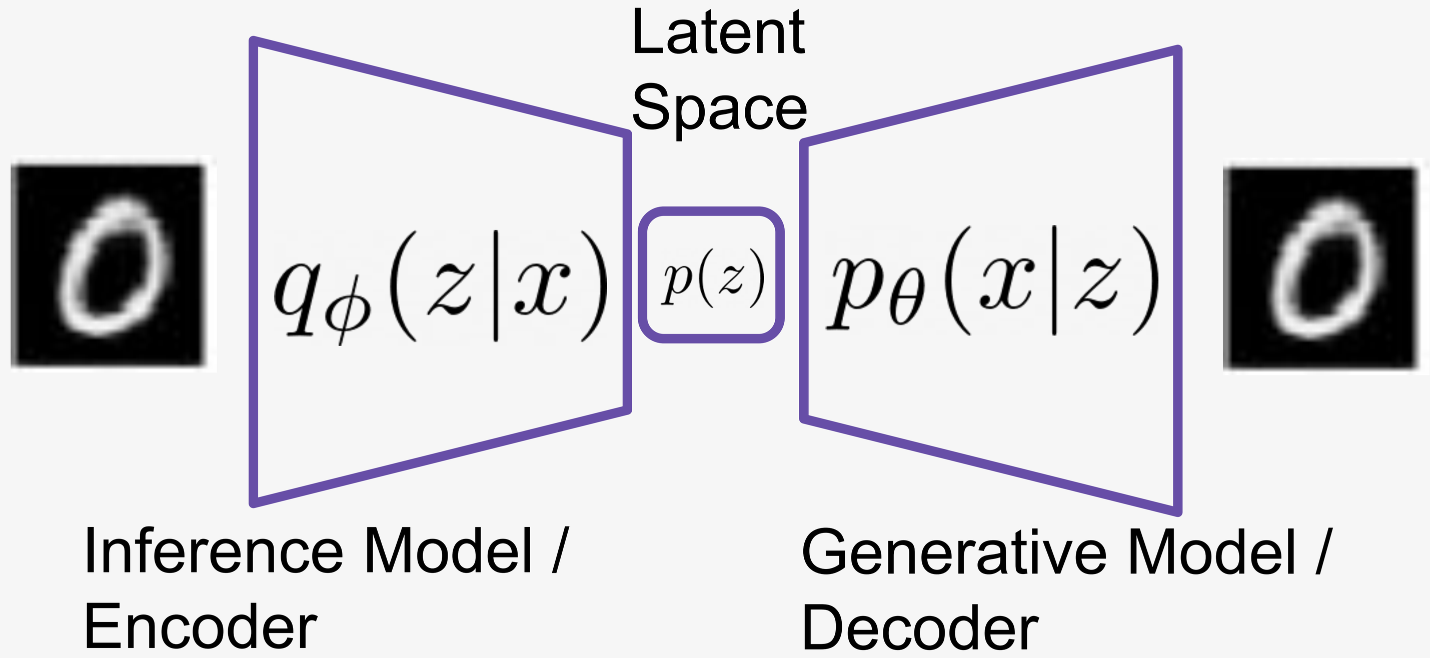
VAE are not robust to adversarial attacks

We propose the way to alleviate the effect

Method does not require changing the training procedure

We have theoretical justification why it works

## Variational Auto-Encoder

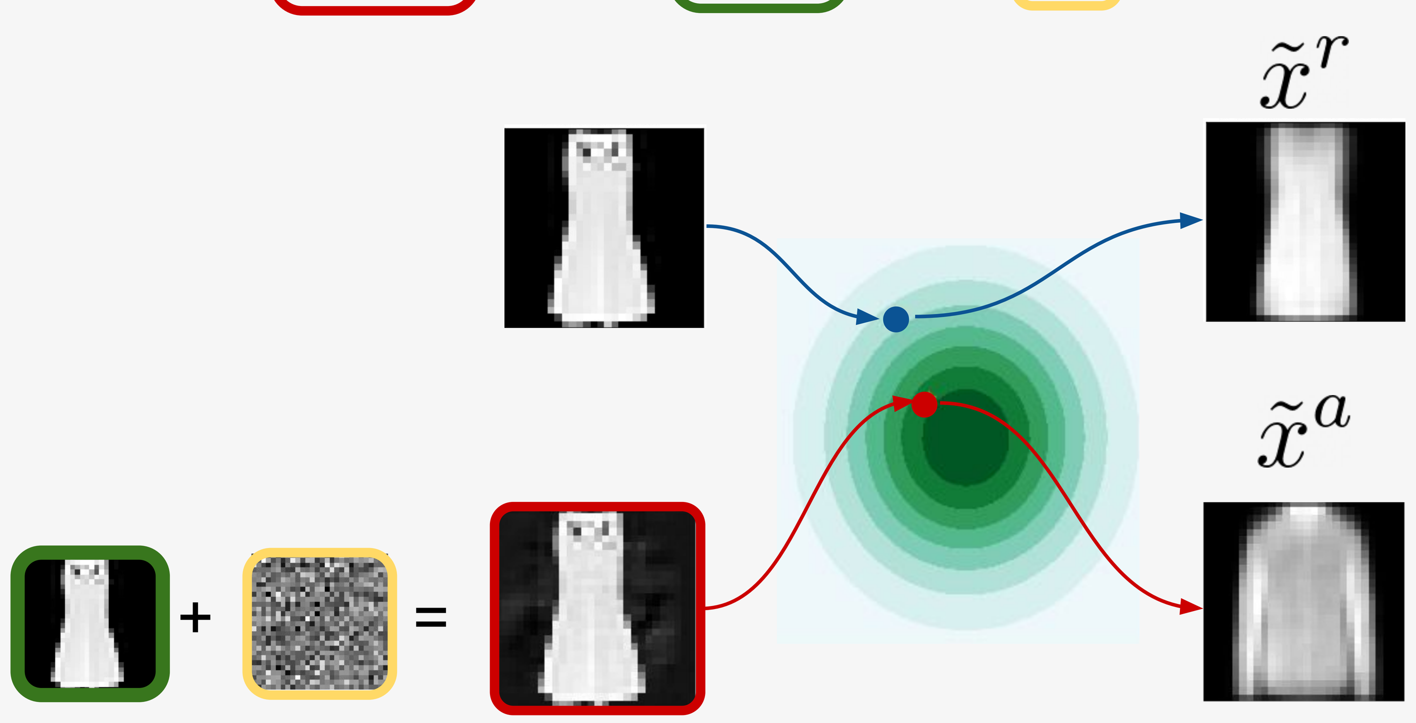



## Hierarchical VAE

L latent variables $\mathbf{z} = (z_1, \dots, z_L)$

## Adversarial Input

$$\boxed{x^a} = \boxed{x^r} + \boxed{\varepsilon}$$



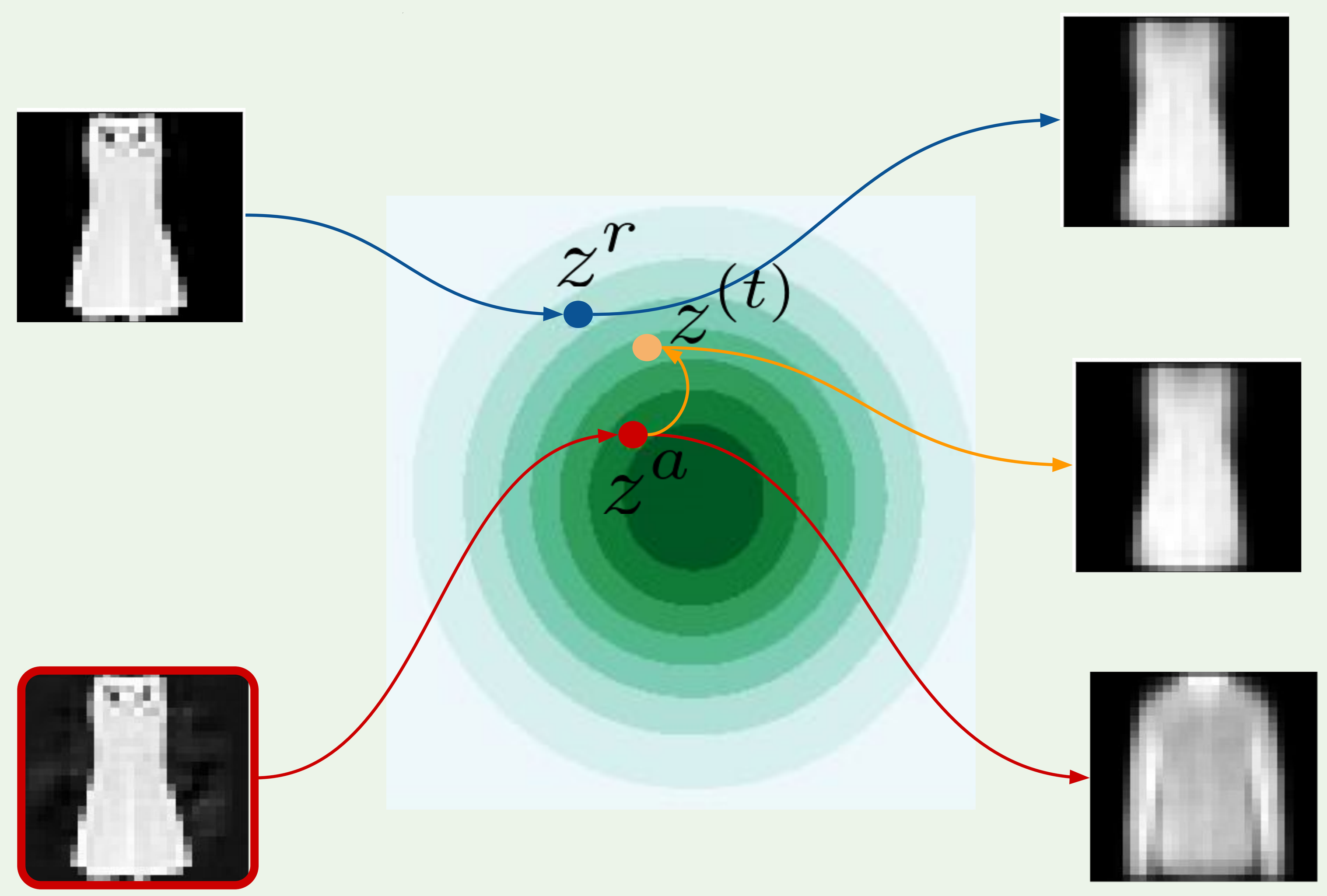

## Can we reduce the effect of attack?

$z^r \sim q_\phi(z|x^r)$ is what we want

$z^a \sim q_\phi(z|x^a)$ is what we get instead

Let's sample from the true posterior $p_\theta(z|x^a) \propto p(z)p_\theta(x^a|z)$

$$z^{(t)} \sim q^{(t)}(z|x^a) = \int q_\phi(z_0|x^a)Q^{(t)}(z|z_0)dz_0$$



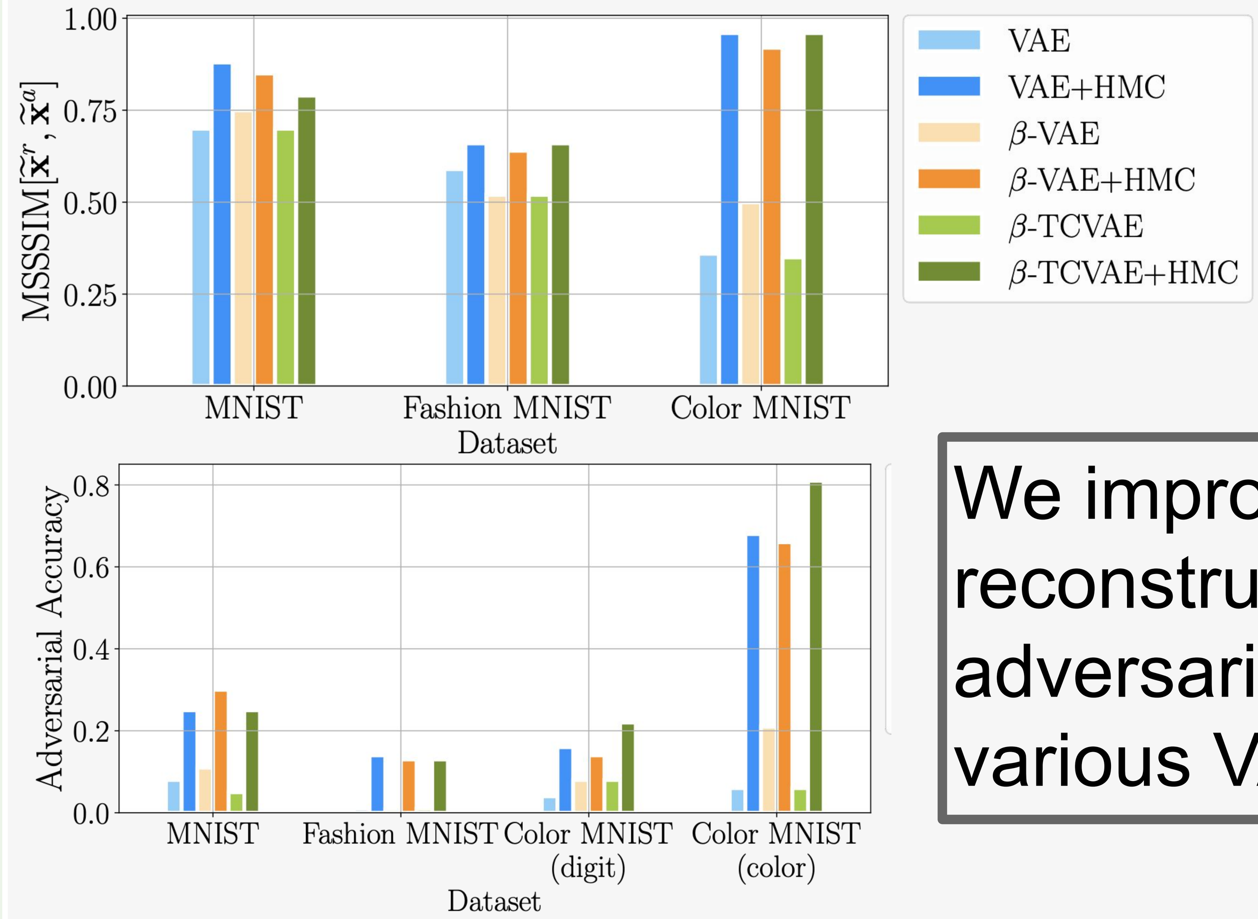

## Why does it work?

$$\mathrm{TV}[q^{(t)}(z|x^a)\|q_\phi(z|x^r)] \leq \underbrace{\sqrt{\tfrac{1}{2}\mathrm{KL}\left[q^{(t)}(z|x^a)\|p_\theta(z|x^a)\right]}}_{\text{Gets smaller with each MCMC step}}$$

$$+ \underbrace{\sqrt{\tfrac{1}{2}\mathrm{KL}\left[q_\phi(z|x^r)\|p_\theta(z|x^r)\right]}}_{\text{VAE amortization gap}}$$

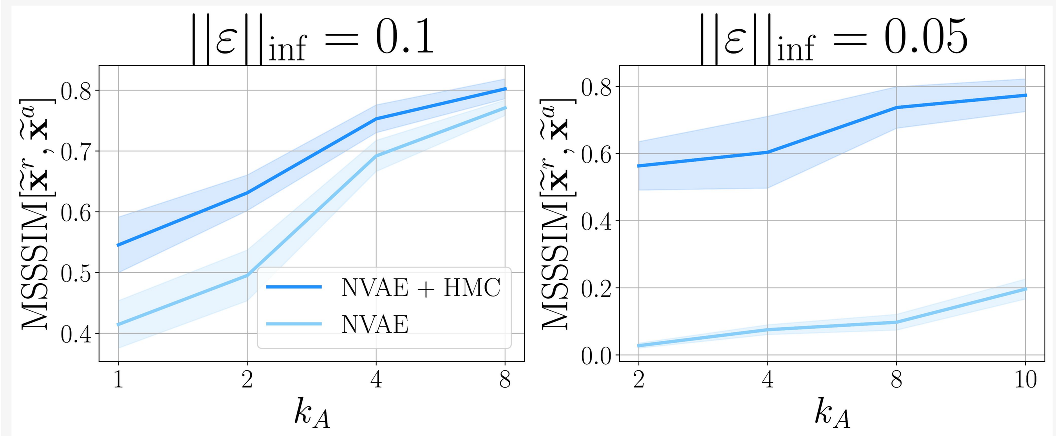
$$+ \underbrace{\boxed{o(\sqrt{\|\varepsilon\|})}}_{\text{Attack radius}}$$

## Results: VAE




We improve both reconstruction quality and adversarial accuracy for various VAE modificationd

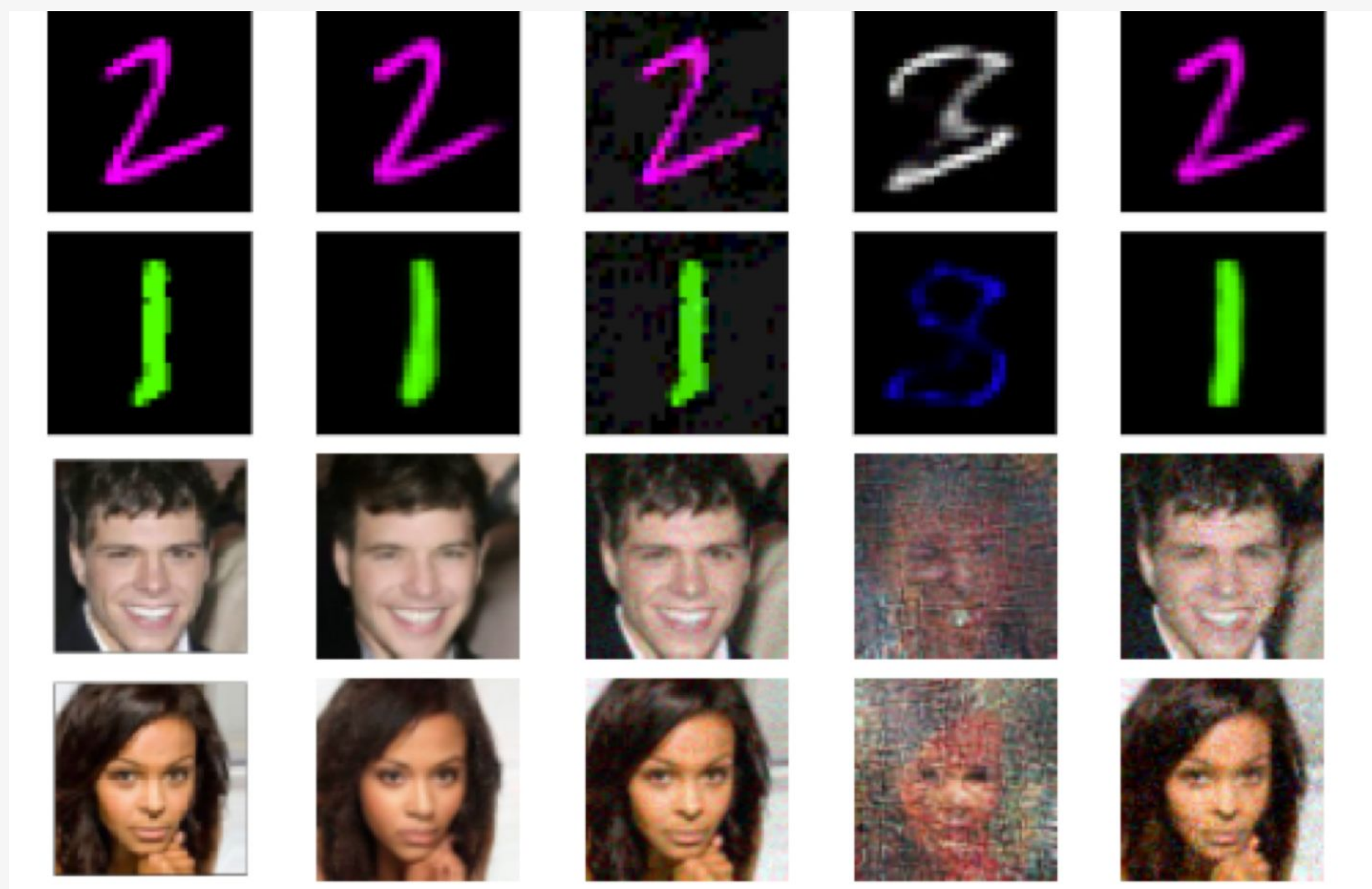
## Results: NVAE




MNIST CelebA

Consider only last $k_A$ latent variables for attack construction

We observe that reconstructions and more similar to the reference when we use the proposed method



(a) $\mathbf{x}^r$ (b) $\tilde{\mathbf{x}}^r$ (c) $\mathbf{x}^a$ (d) $\tilde{\mathbf{x}}^a$ (e) $\tilde{\mathbf{x}}^a_{\mathrm{HMC}}$