

# GAN Dissection: Visualizing and Understanding Generative Adversarial Networks

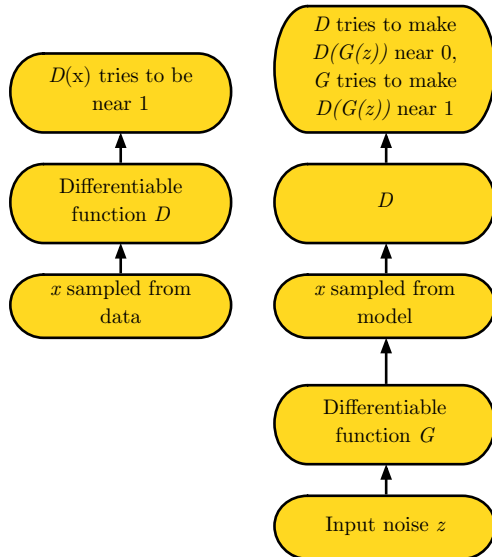
Martin Šafránek

Faculty of Information Technology  
Czech Technical University in Prague



- GAN overview
- GAN dissection (paper)

# Adversarial Nets Framework



$$J^{(D)} = -\frac{1}{2}\mathbb{E}_{x \sim p_{\text{data}}} \log D(x) - \frac{1}{2}\mathbb{E}_z \log (1 - D(G(z)))$$

$$J^G = -J^D$$

- Generator minimizes the log-probability of the discriminator being correct
- Equilibrium if the discriminator is unable to differentiate between real and generated input

- presents method for visualizing and understanding GAN
- learned GAN contains variables for doors, trees, ...
- can interactively manipulate objects in a scene



(a) Generate images of churches



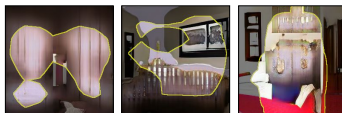
(b) Identify GAN units that match trees



(c) Ablating units removes trees



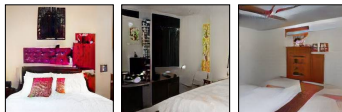
(d) Activating units adds trees



(e) Identify GAN units that cause artifacts



(f) Bedroom images with artifacts



(g) Ablating "artifact" units improves results