

Introduction to Generative Adversarial Networks

Ian Goodfellow, OpenAI Research Scientist
NIPS 2016 Workshop on Adversarial Training
Barcelona, 2016-12-9

OpenAI

Adversarial Training

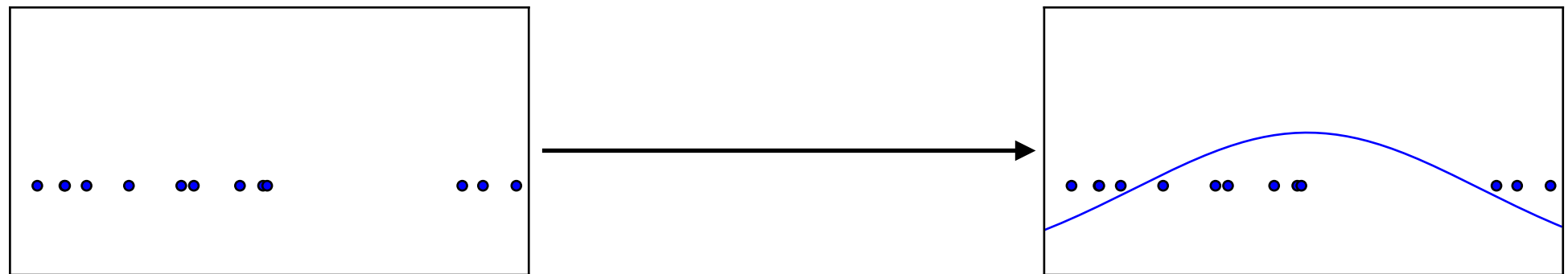
- A phrase whose usage is in flux; a new term that applies to both new and old ideas
- My current usage: “Training a model in a worst-case scenario, with inputs chosen by an adversary”
- Examples:
 - An agent playing against a copy of itself in a board game (Samuel, 1959)
 - Robust optimization / robust control (e.g. Rustem and Howe 2002)
 - Training neural networks on adversarial examples (Szegedy et al 2013, Goodfellow et al 2014)

Generative Adversarial Networks

- Both players are neural networks
- Worst case input for one network is produced by another network

Generative Modeling

- Density estimation



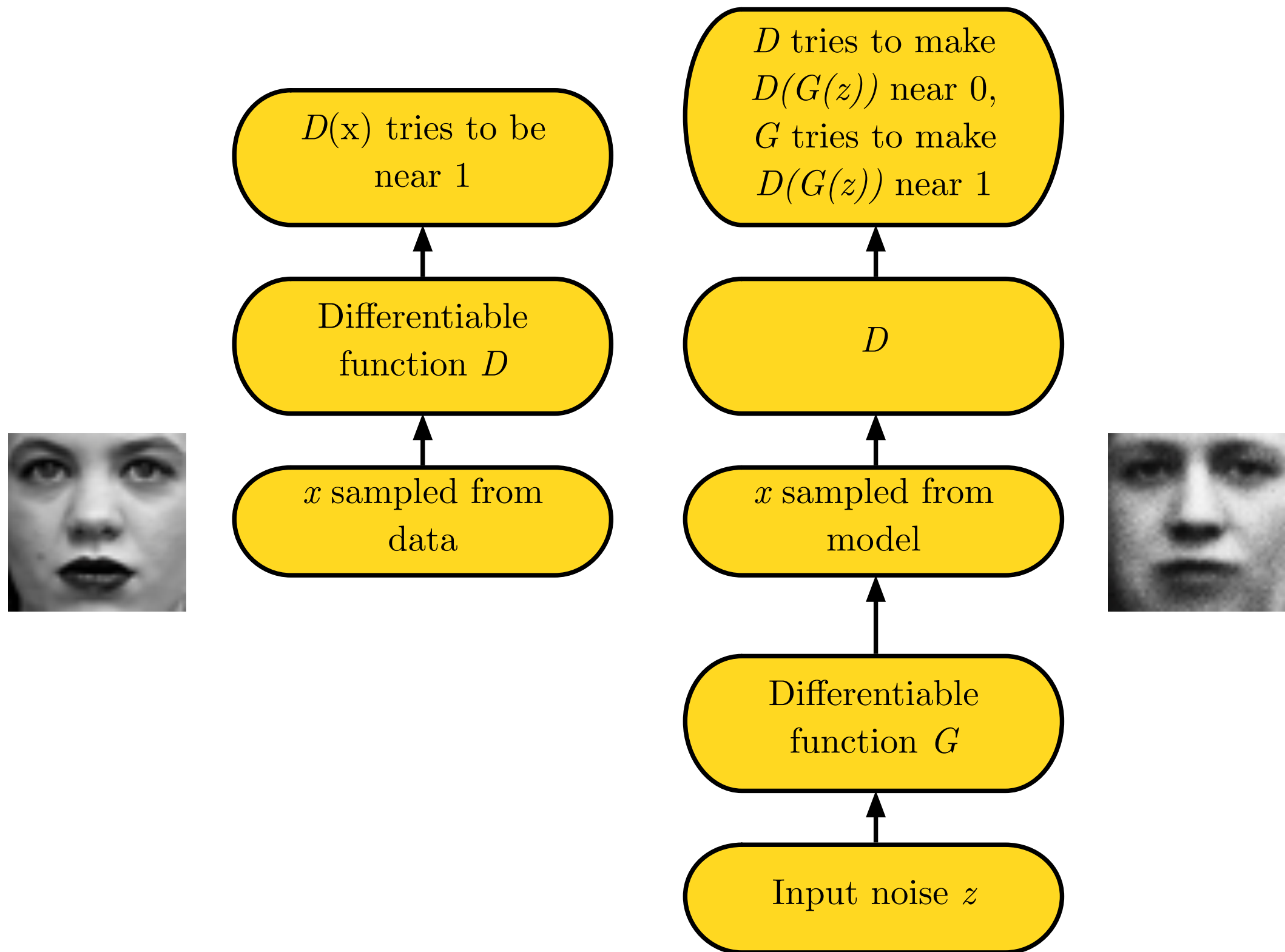
- Sample generation



Training examples

Model samples

Adversarial Nets Framework



Minimax Game

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

$$J^{(G)} = -J^{(D)}$$

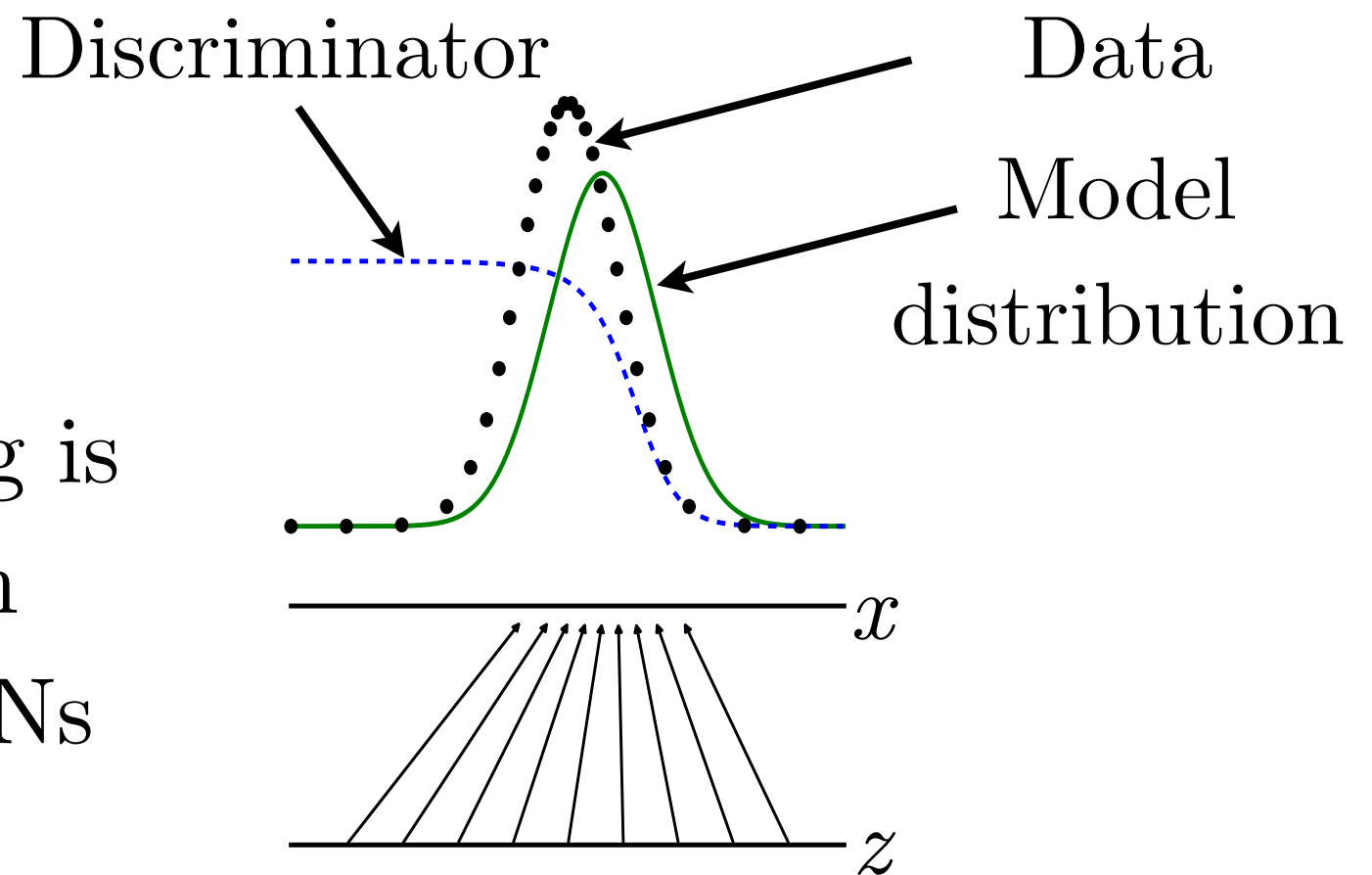
- Equilibrium is a saddle point of the discriminator loss
- Resembles Jensen-Shannon divergence
- Generator minimizes the log-probability of the discriminator being correct

Discriminator Strategy

Optimal $D(\mathbf{x})$ for any $p_{\text{data}}(\mathbf{x})$ and $p_{\text{model}}(\mathbf{x})$ is always

$$D(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_{\text{model}}(x)}$$

Estimating this ratio
using supervised learning is
the key approximation
mechanism used by GANs



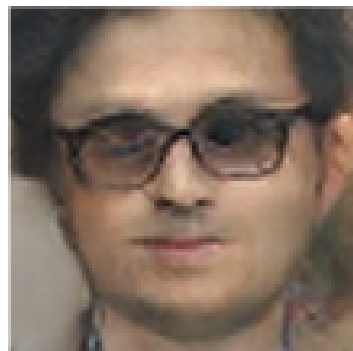
Non-Saturating Game

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

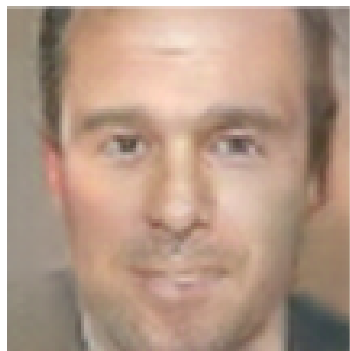
$$J^{(G)} = -\frac{1}{2}\mathbb{E}_{\mathbf{z}} \log D(G(\mathbf{z}))$$

- Equilibrium no longer describable with a single loss
- Generator maximizes the log-probability of the discriminator being mistaken
- Heuristically motivated; generator can still learn even when discriminator successfully rejects all generator samples

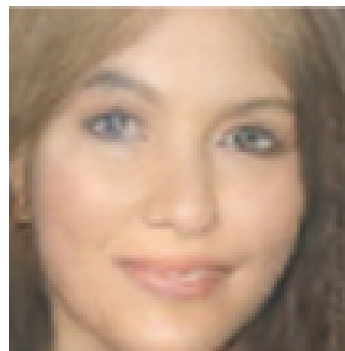
Vector Space Arithmetic



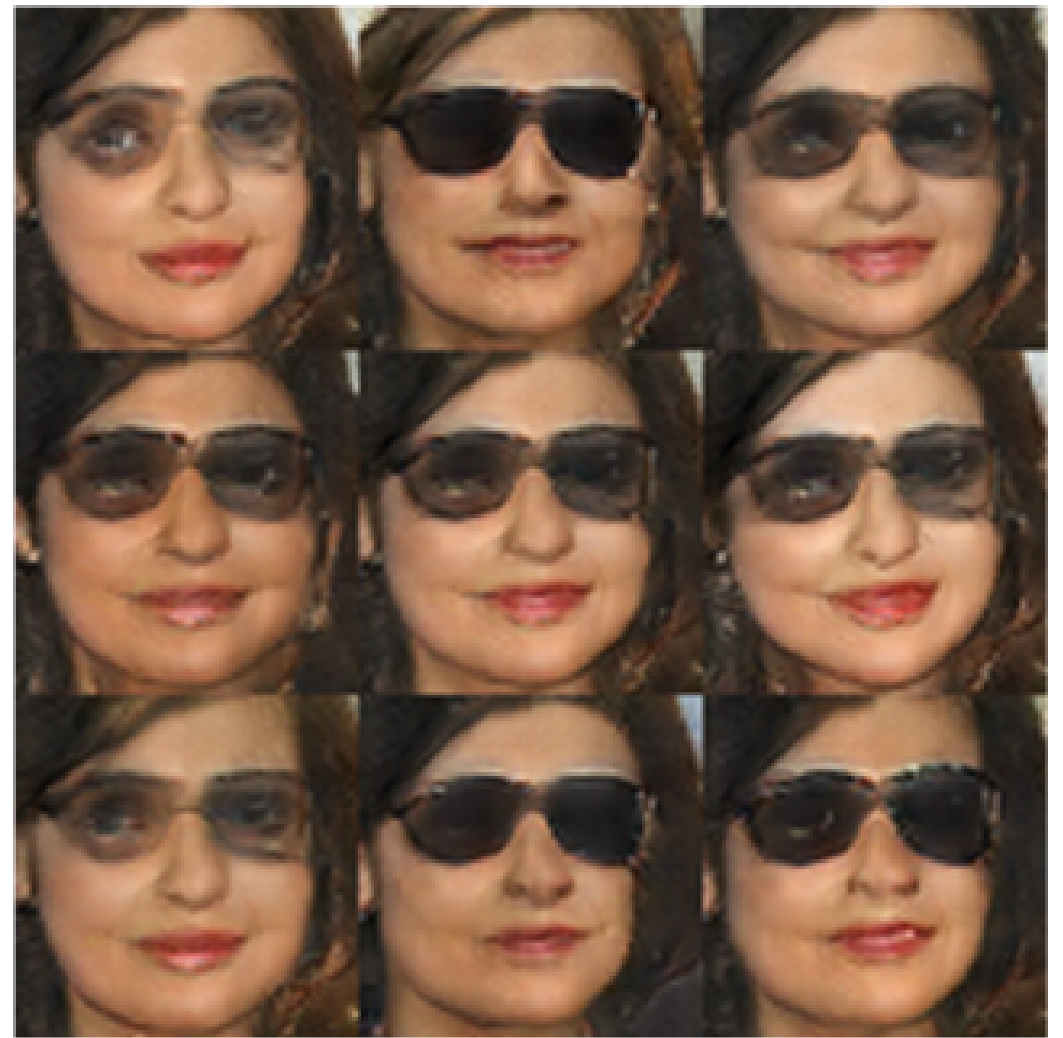
-



+



=



Man
with glasses

Man

Woman

Woman with Glasses

(Radford et al, 2015)

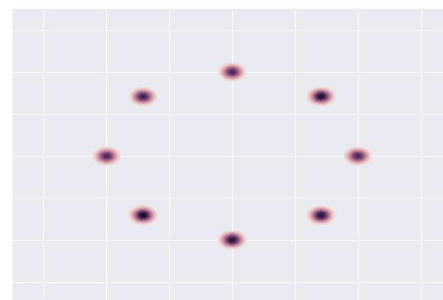
Non-convergence

- Optimization algorithms often approach a saddle point or local minimum rather than a global minimum
- Game solving algorithms may not approach an equilibrium at all

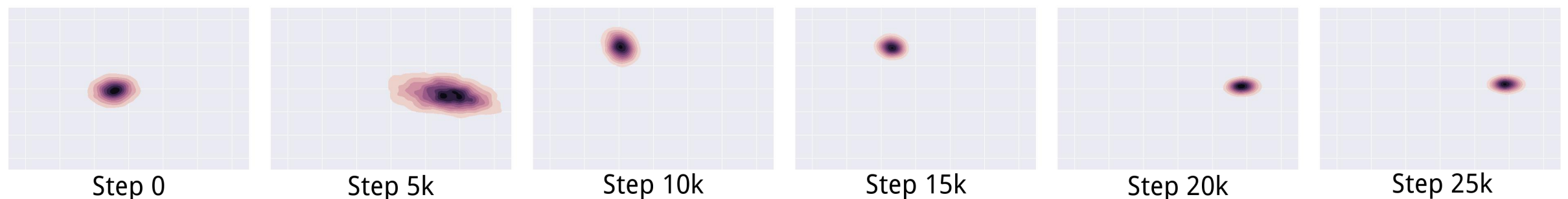
Mode Collapse

$$\min_G \max_D V(G, D) \neq \max_D \min_G V(G, D)$$

- D in inner loop: convergence to correct distribution
- G in inner loop: place all mass on most likely point



Target

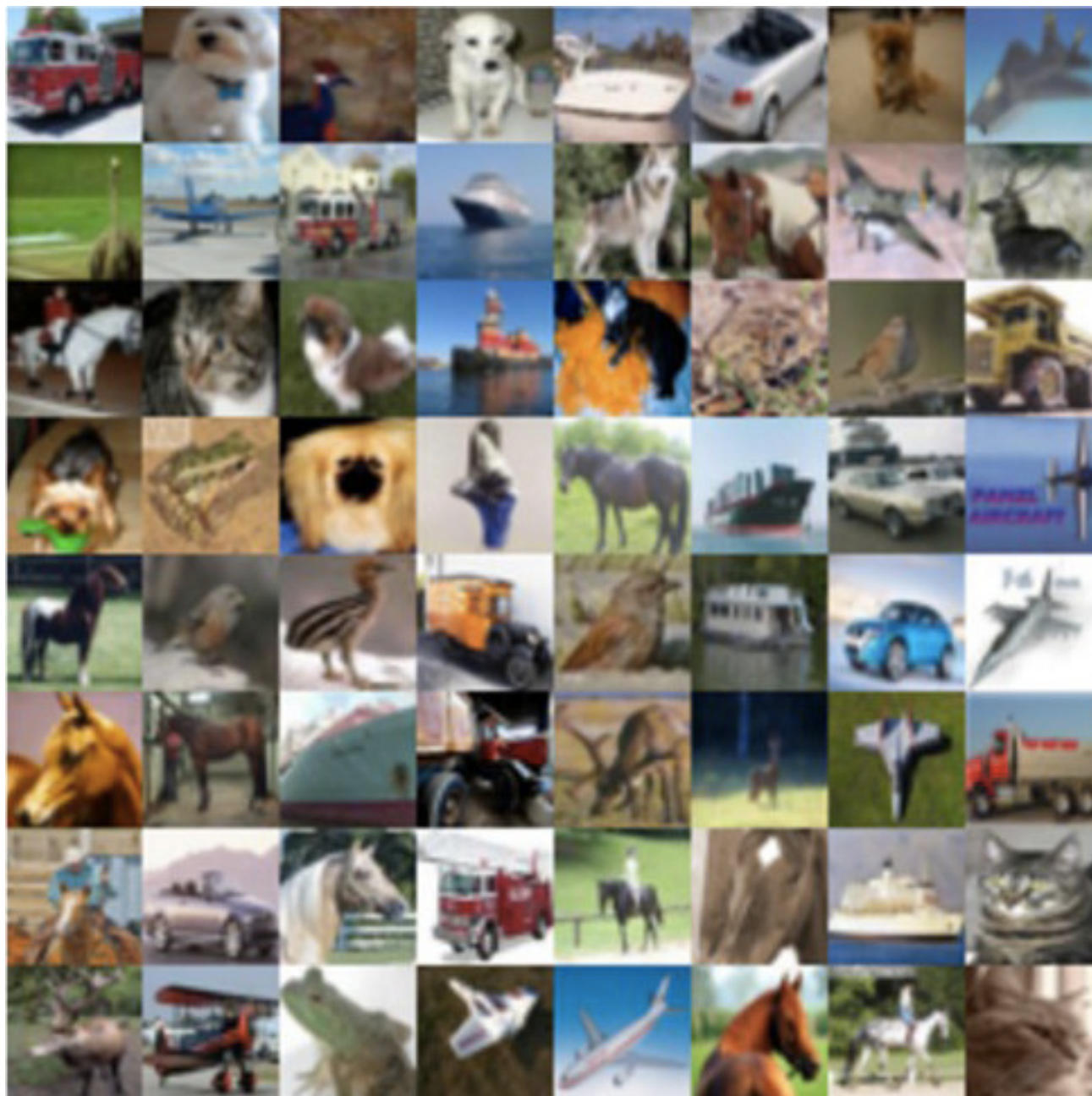


(Metz et al 2016)

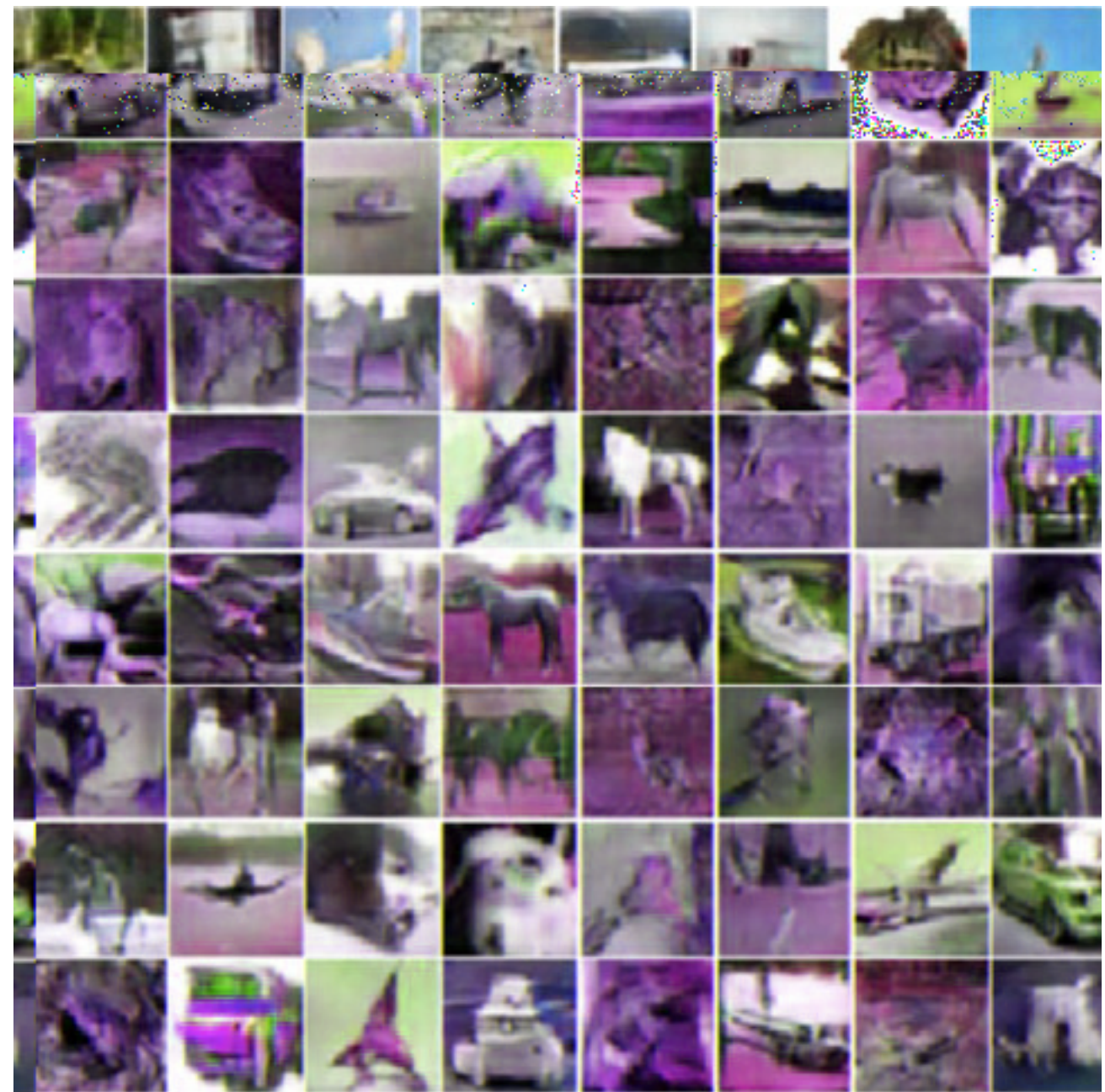
Minibatch Features

- Add minibatch features that classify each example by comparing it to other members of the minibatch (Salimans et al 2016)
- Nearest-neighbor style features detect if a minibatch contains samples that are too similar to each other

Minibatch GAN on CIFAR



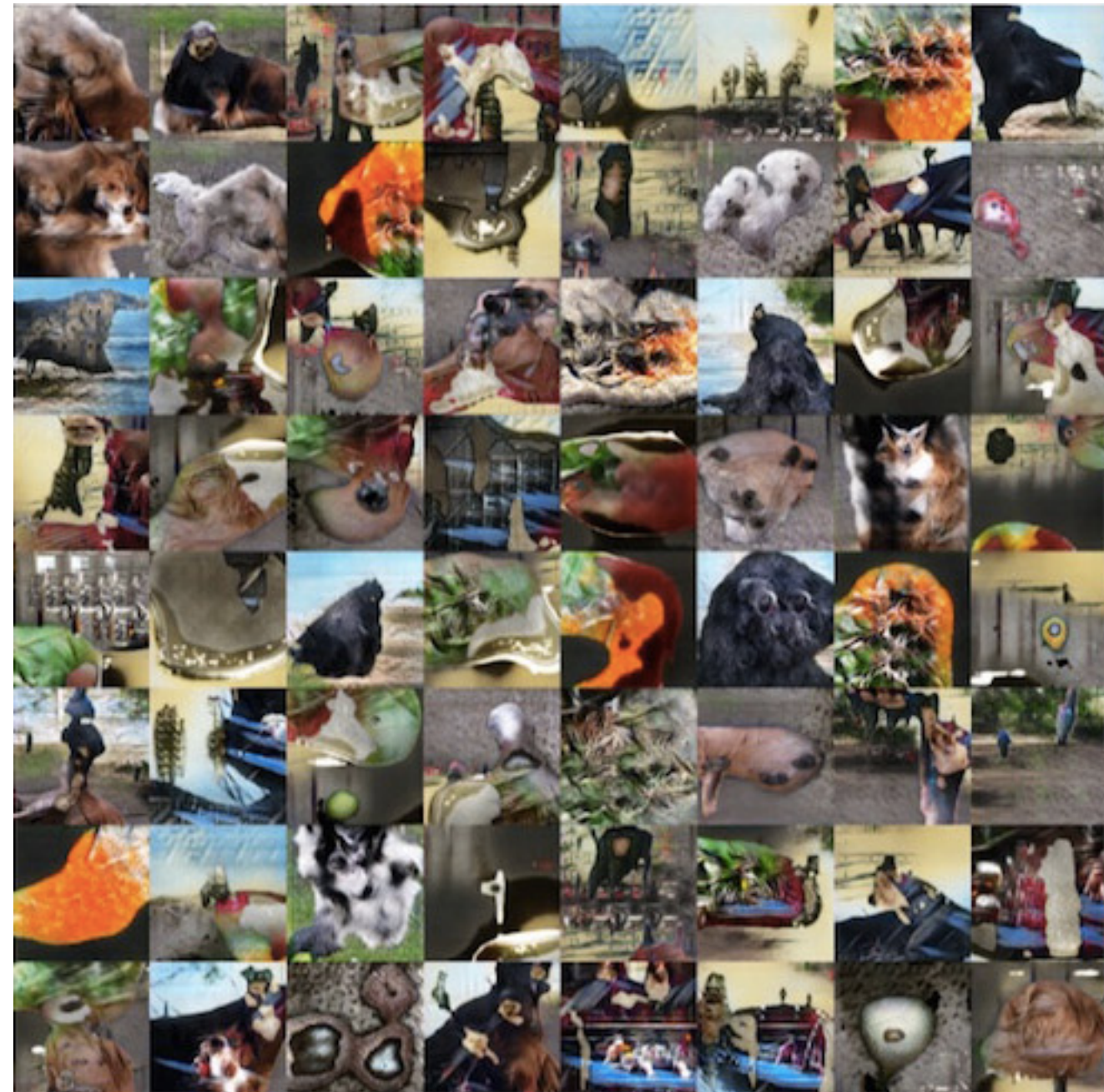
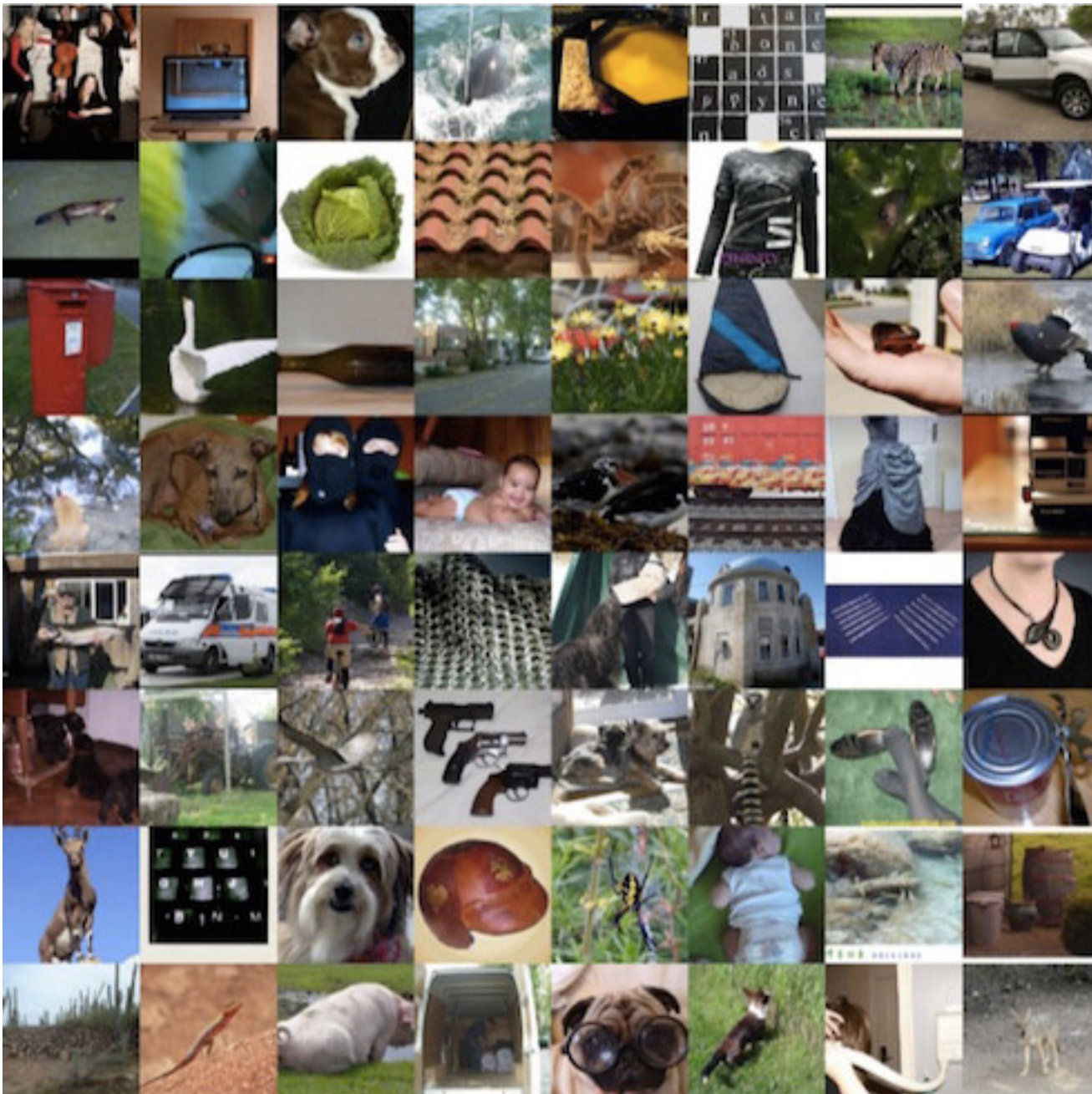
Training Data



Samples

(Salimans et al 2016)

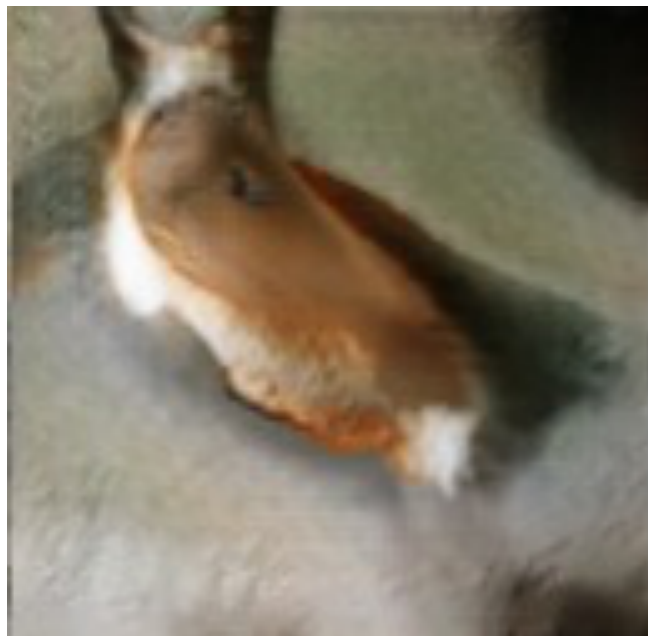
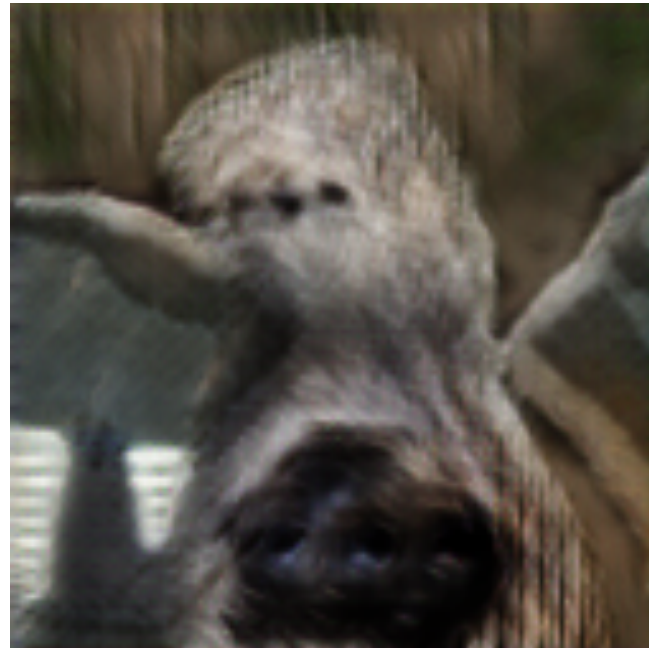
Minibatch GAN on ImageNet



(Salimans et al 2016)

(Goodfellow 2016)

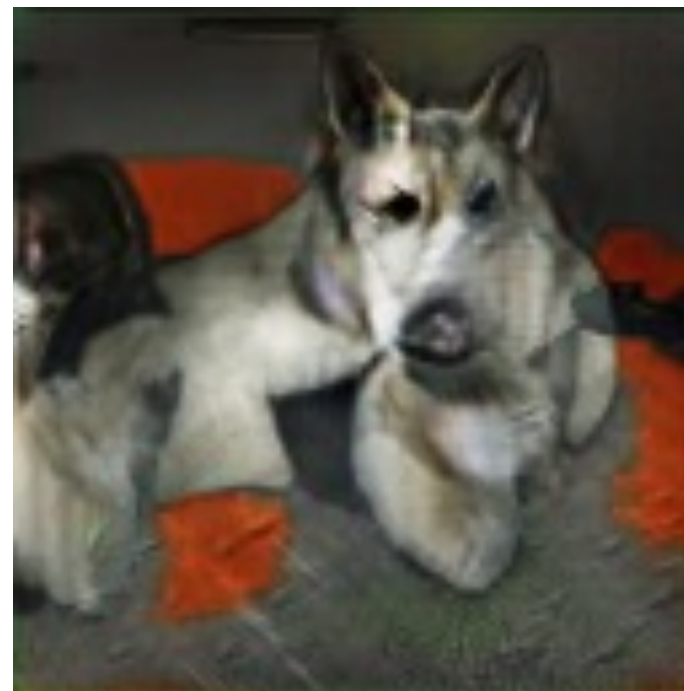
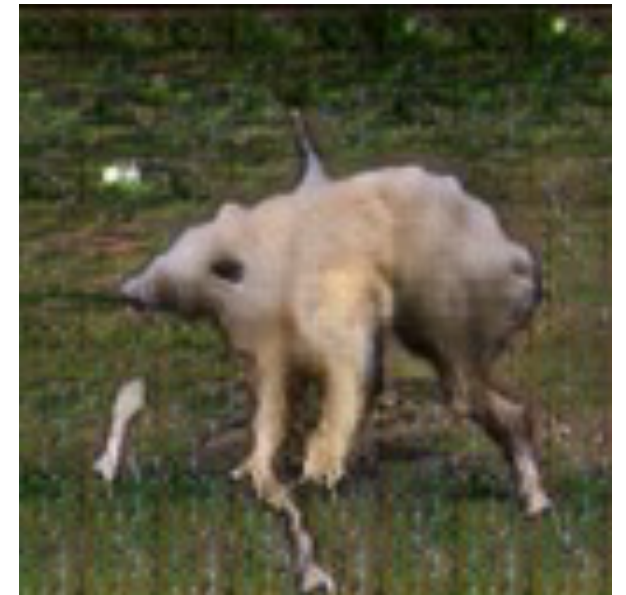
Cherry-Picked Results



Problems with Counting



Problems with Perspective



Problems with Global Structure

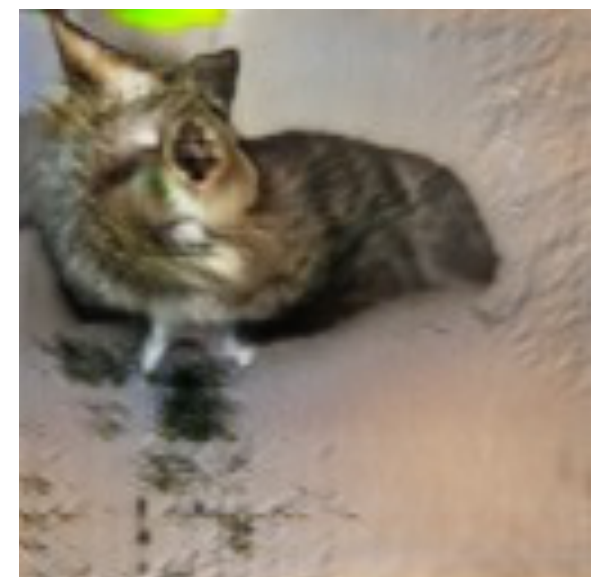
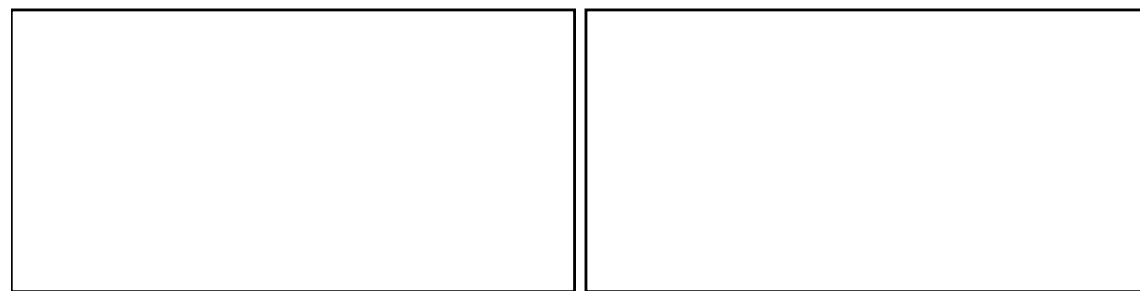


Image to Image Translation

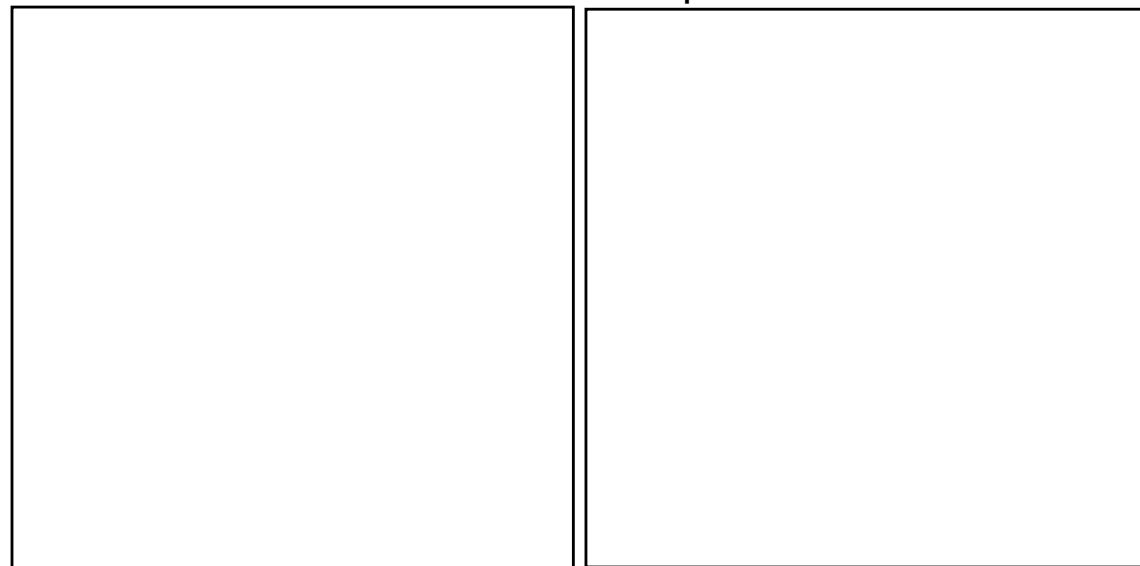
Labels to Street Scene



input

output

Aerial to Map



input

output

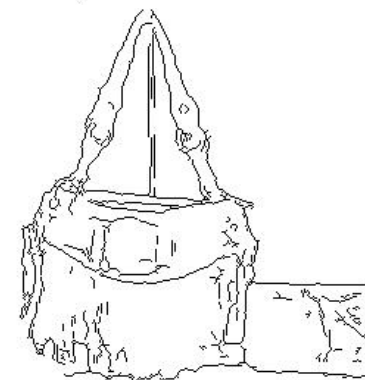
Input



Ground truth



Output



(Isola et al 2016)

Plug and Play Generative Models

- New state of the art generative model (Nguyen et al 2016) released days before NIPS
- Generates 227x227 realistic images from all ImageNet classes
- Combines adversarial training, moment matching, denoising autoencoders, and Langevin sampling

PPGN Samples



redshank

ant

monastery

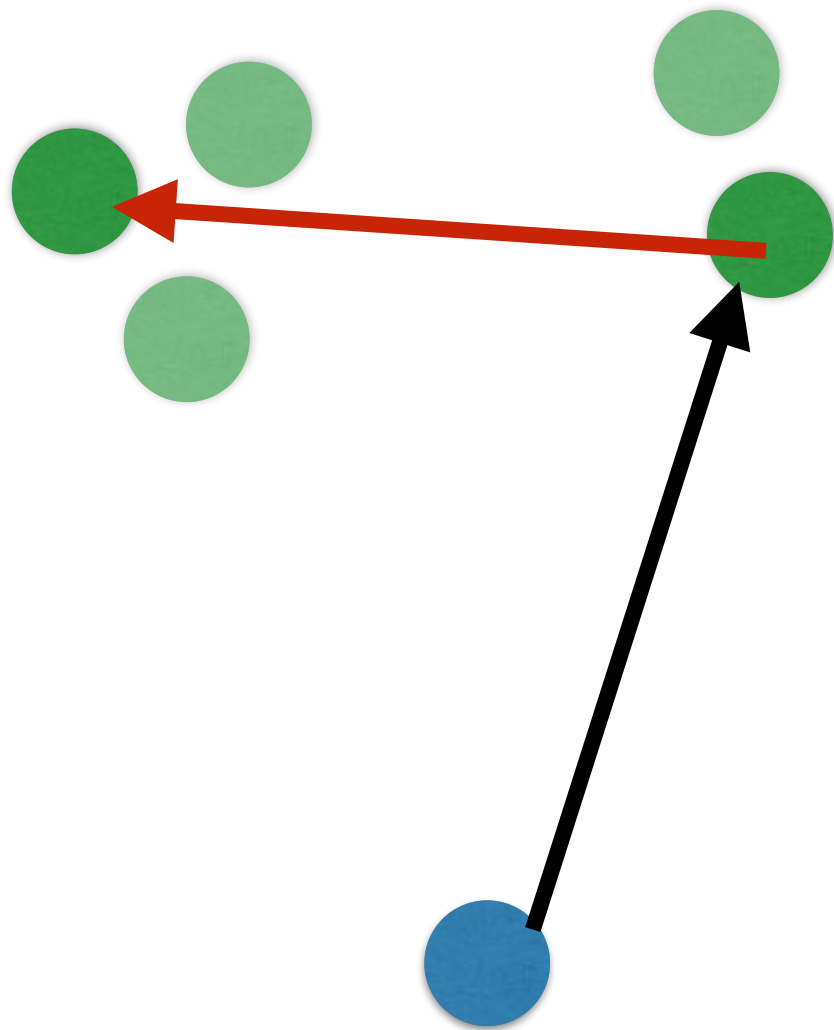


volcano

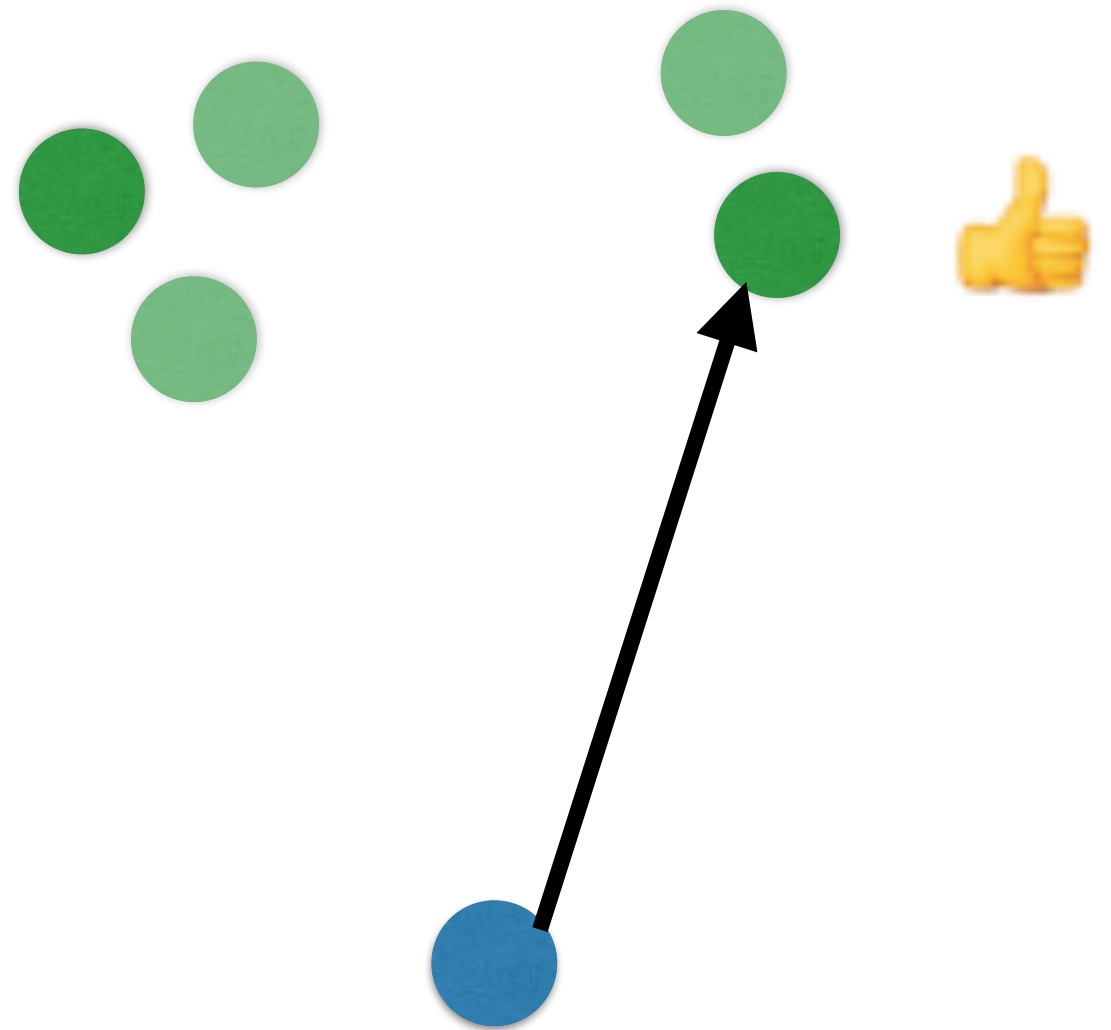
(Nguyen et al 2016)

GANs allow many answers

Mean Squared Error

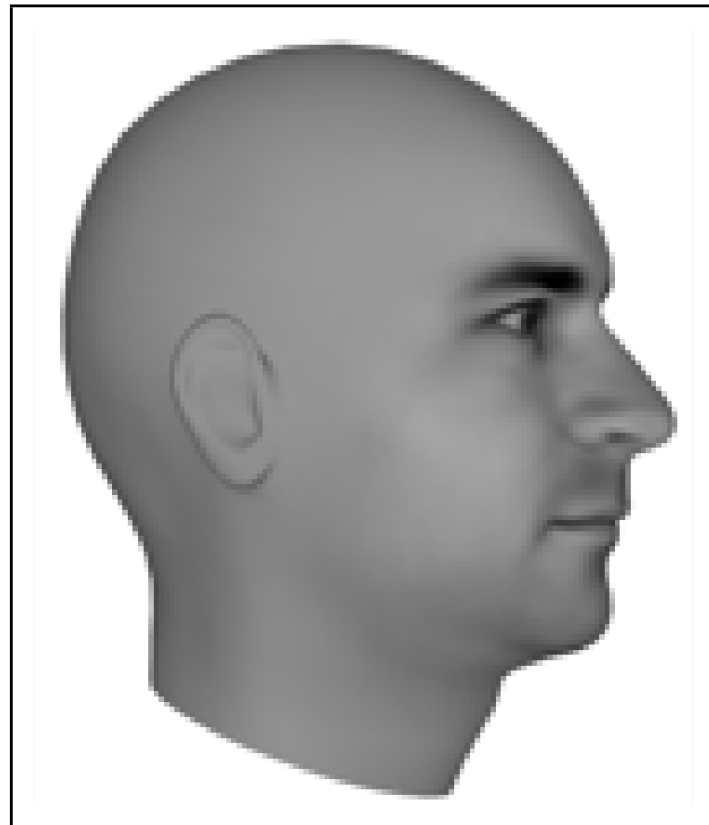


GANs

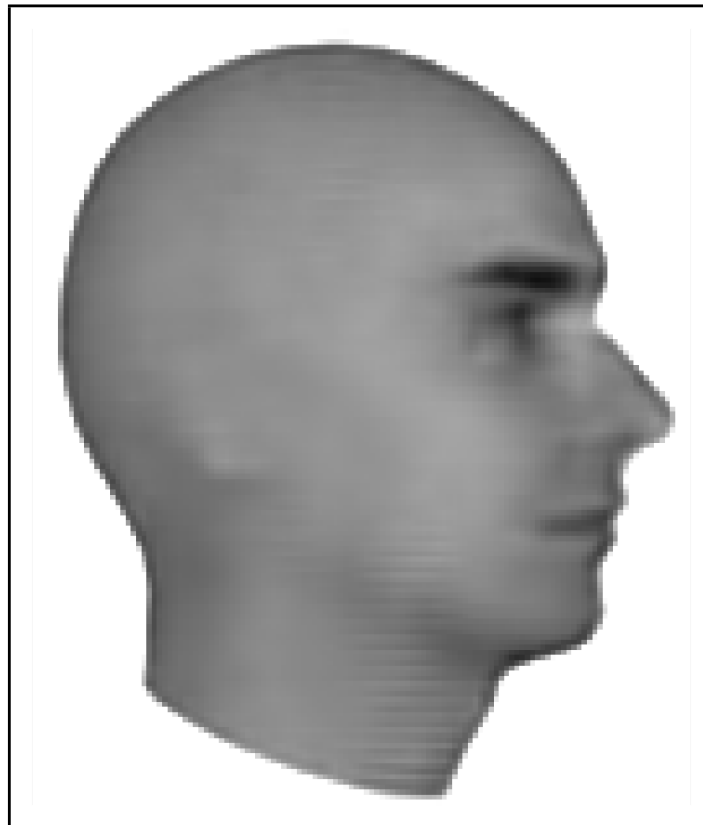


Next Video Frame Prediction

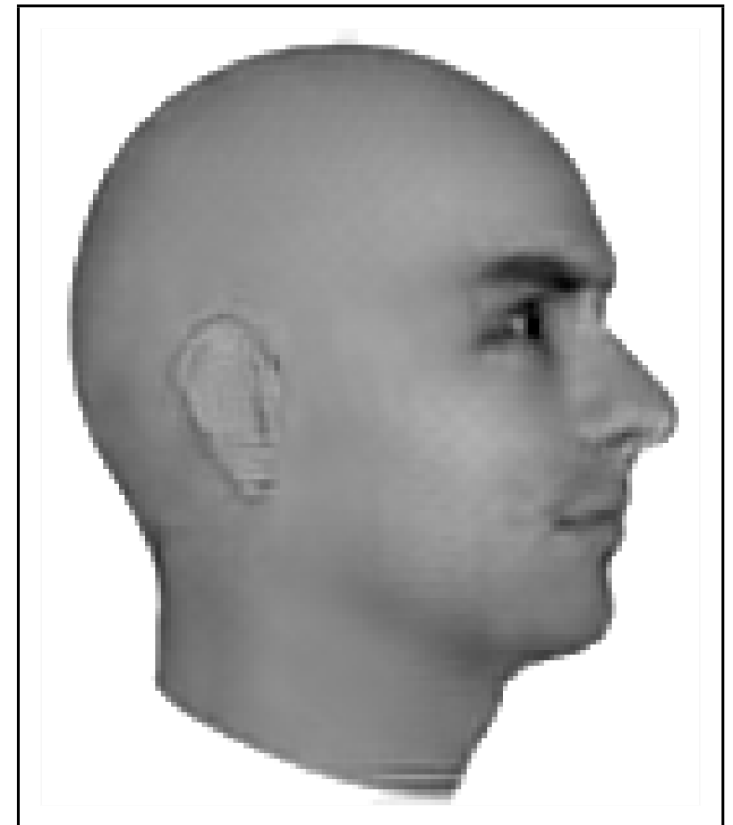
Ground Truth



MSE



Adversarial



(Lotter et al 2016)

Adversarial training for people

- Markets
 - Cycles due to non-convergence?
- Auctions, taxation, etc.
- Deliberate practice (Ericsson et al 1993)

Conclusion

- Adversarial training is a term encompassing old and new work
- GANs are a generative models that use supervised learning to estimate a density ratio
- GANs allow a model to learn that there are many correct answers
- Adversarial training can be useful for people as well as machine learning models