Introduction to Generative Adversarial Networks

Ian Goodfellow, OpenAI Research Scientist

NIPS 2016 Workshop on Adversarial Training
Barcelona, 2016-12-9

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)

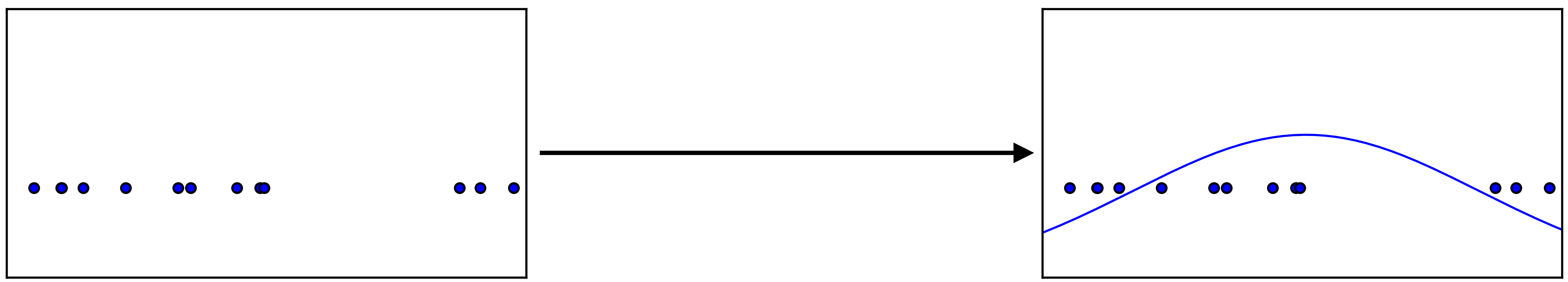
Generative Adversarial Networks

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

(Goodfellow 2016)

Generative Modeling

- Density estimation

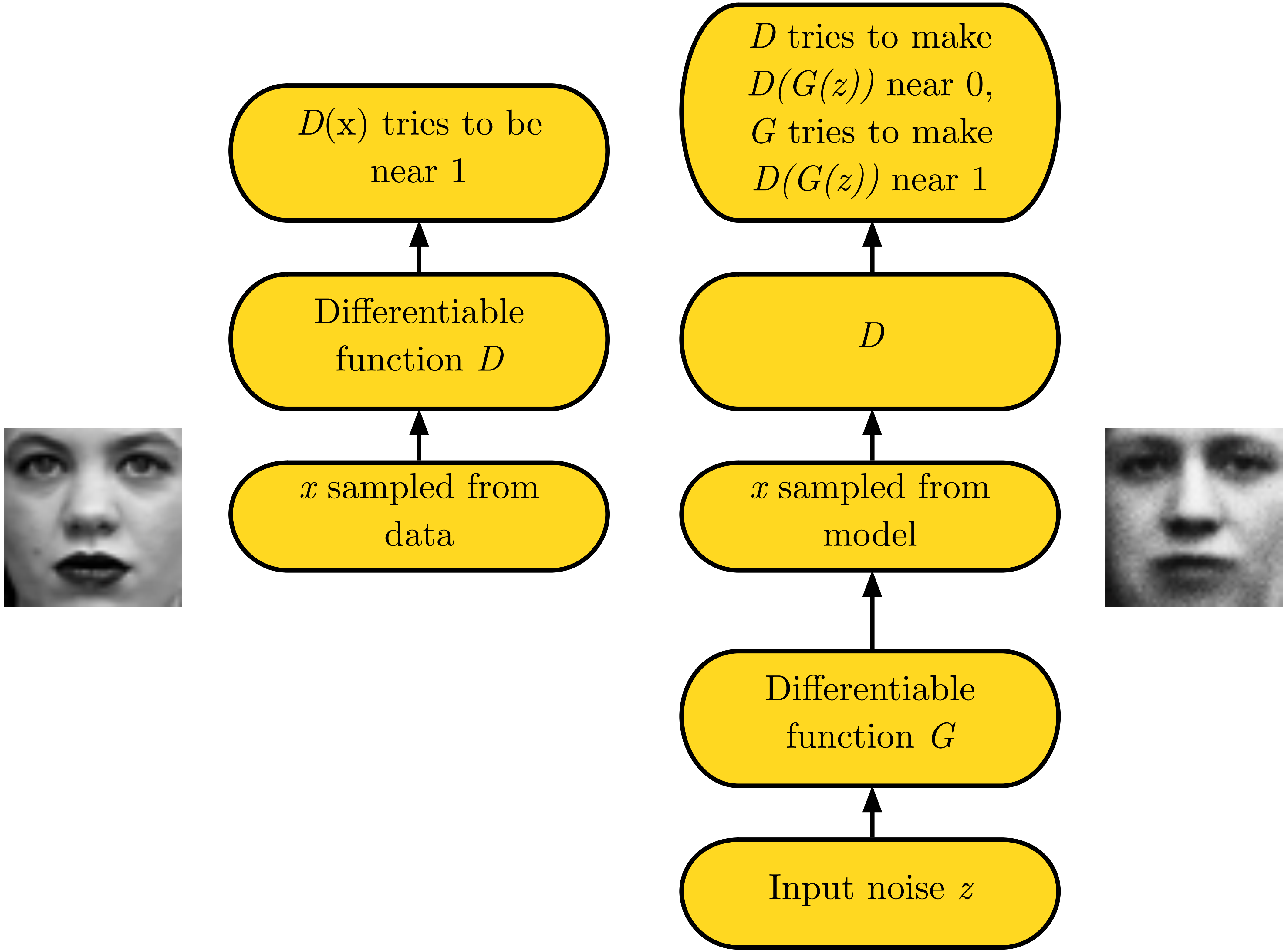


- Sample generation

Training examples Model samples

(Goodfellow 2016)

Adversarial Nets Framework



(Goodfellow 2016)

Minimax Game

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

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

-Equilibrium is a saddle point of the discriminator loss

-Resembles Jensen-Shannon divergence

-Generator minimizes the log-probability of the discriminator being correct

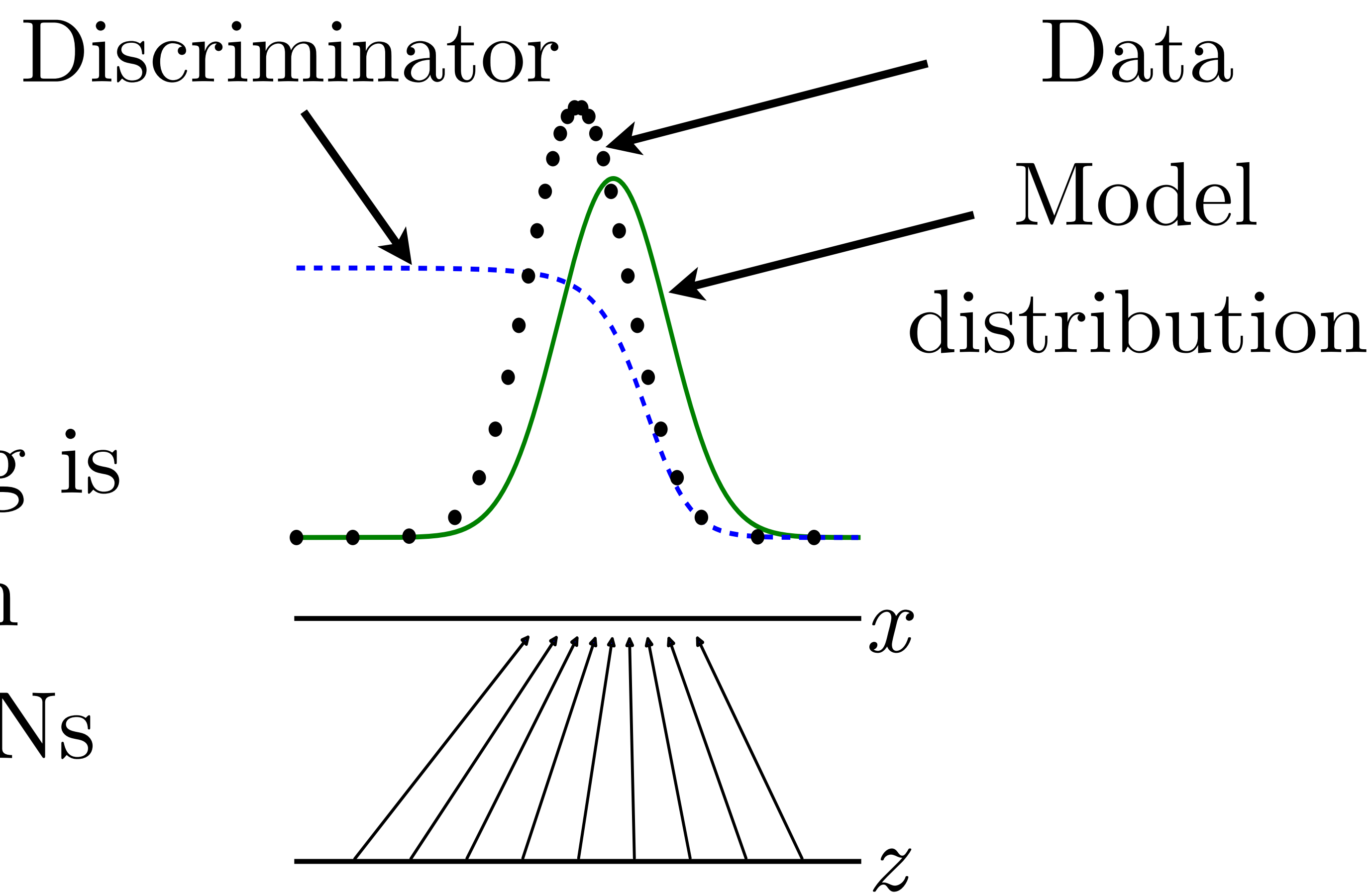
(Goodfellow 2016)

Discriminator Strategy

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



(Goodfellow 2016)

Non-Saturating Game

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

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

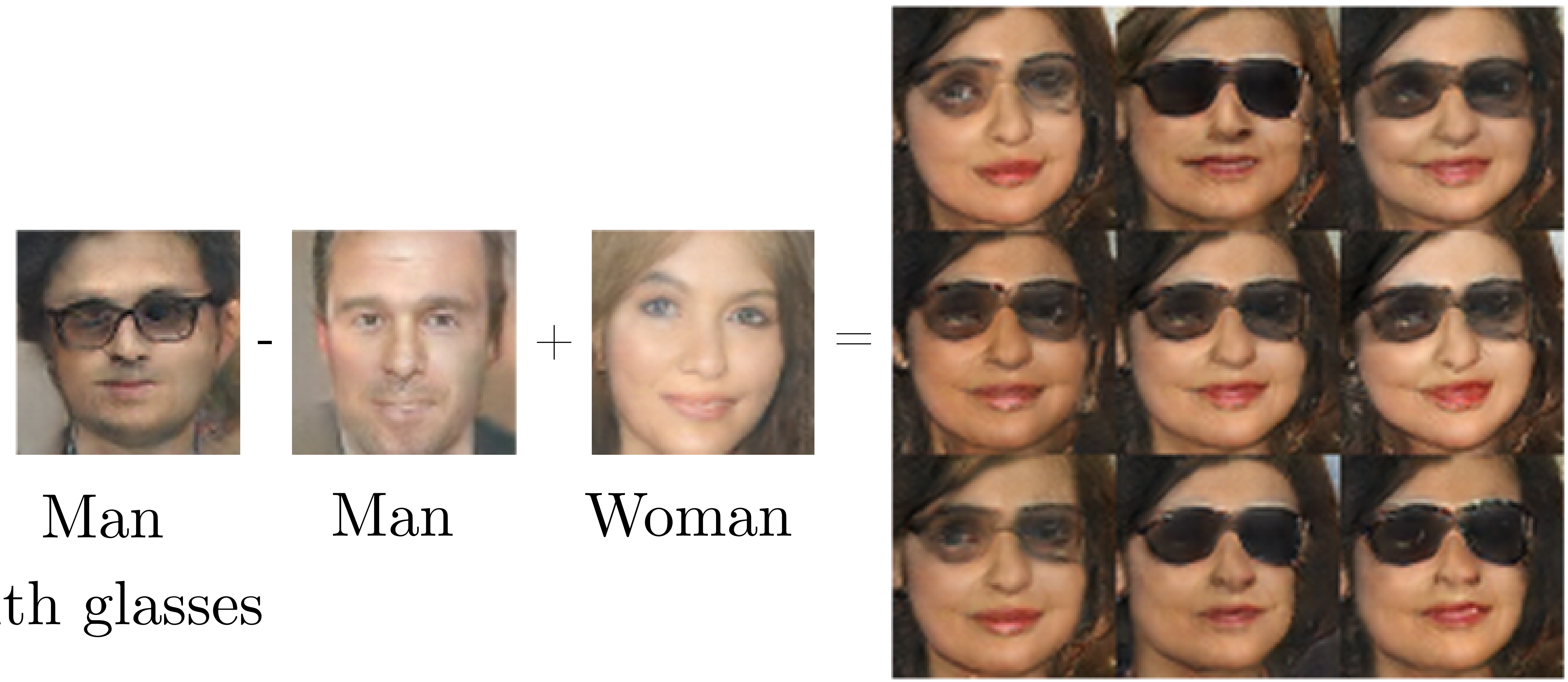
-Equilibrium no longer describable with a single loss

-Generator maximizes the log-probability of the discriminator being mistaken

-Heuristically motivated; generator can still learn even when discriminator successfully rejects all generator samples

(Goodfellow 2016)

Vector Space Arithmetic



Man with glasses — Man + Woman = Woman with Glasses

(Radford et al, 2015)

Non-convergence

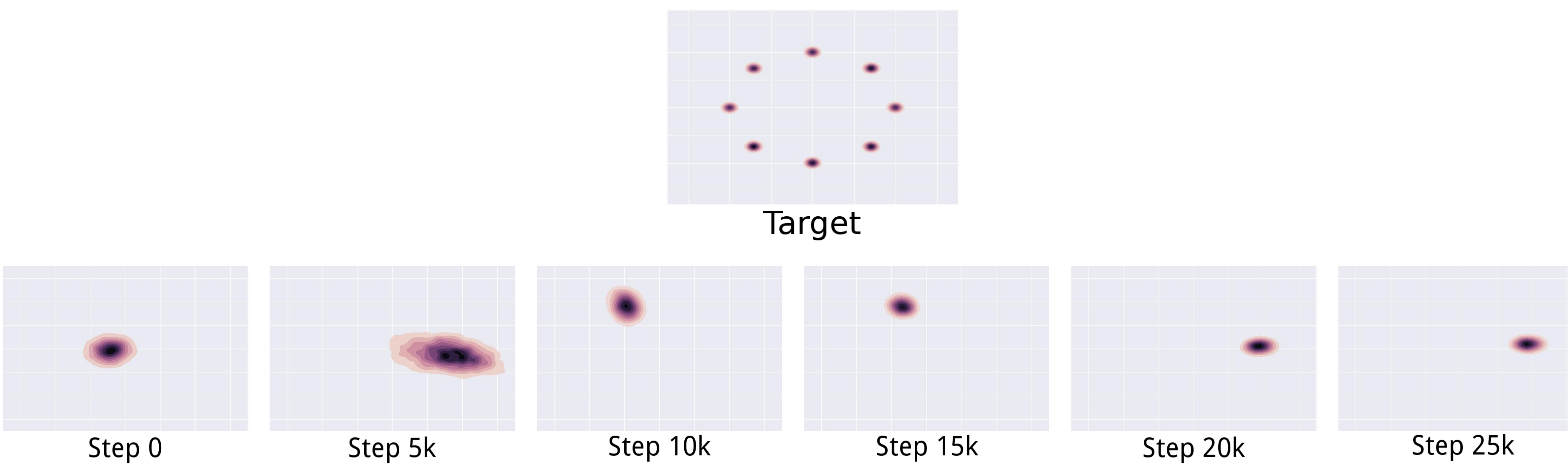
- Optimization algorithms often approach a saddle point or local minimum rather than a global minimum

- 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



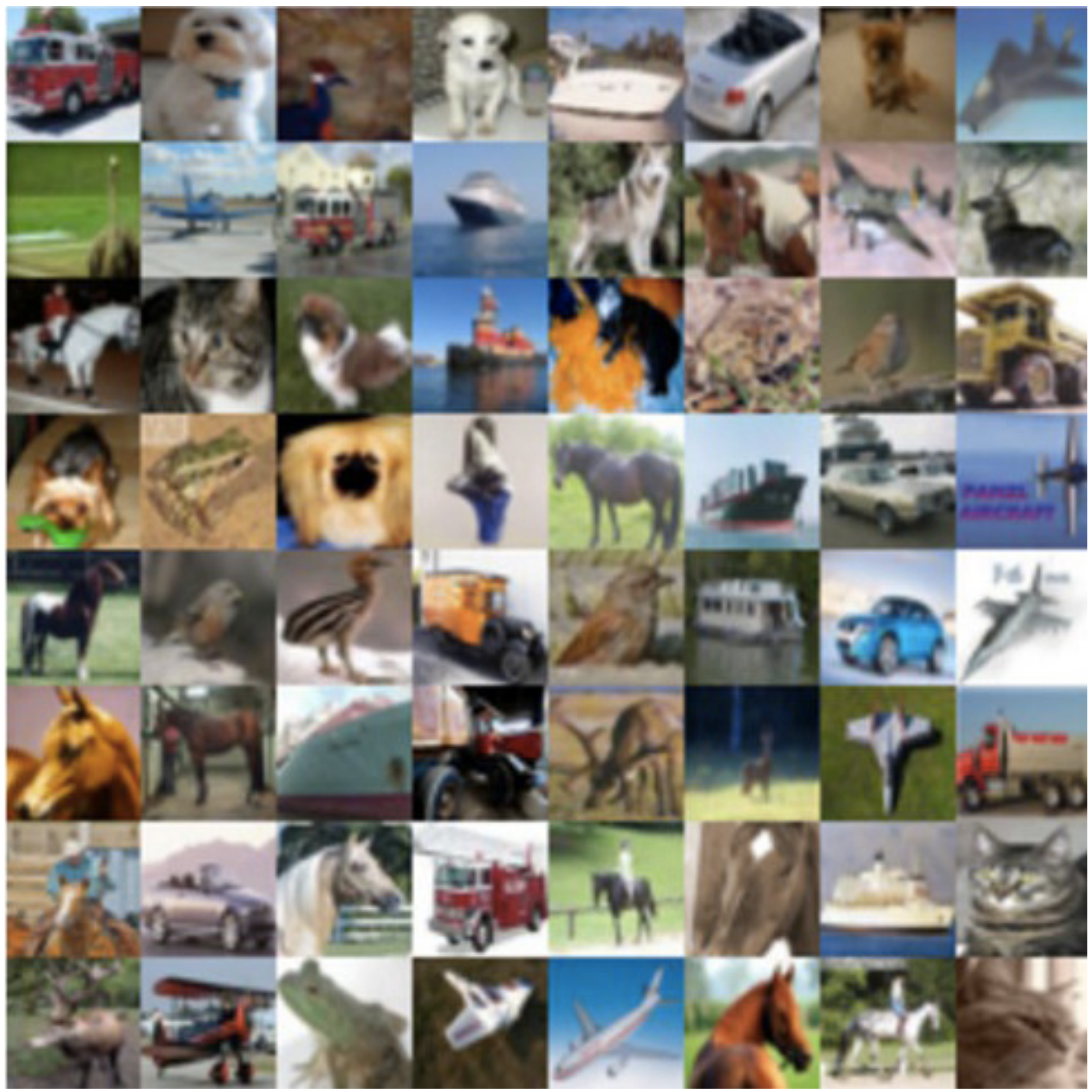
(Metz et al 2016)

Minibatch Features

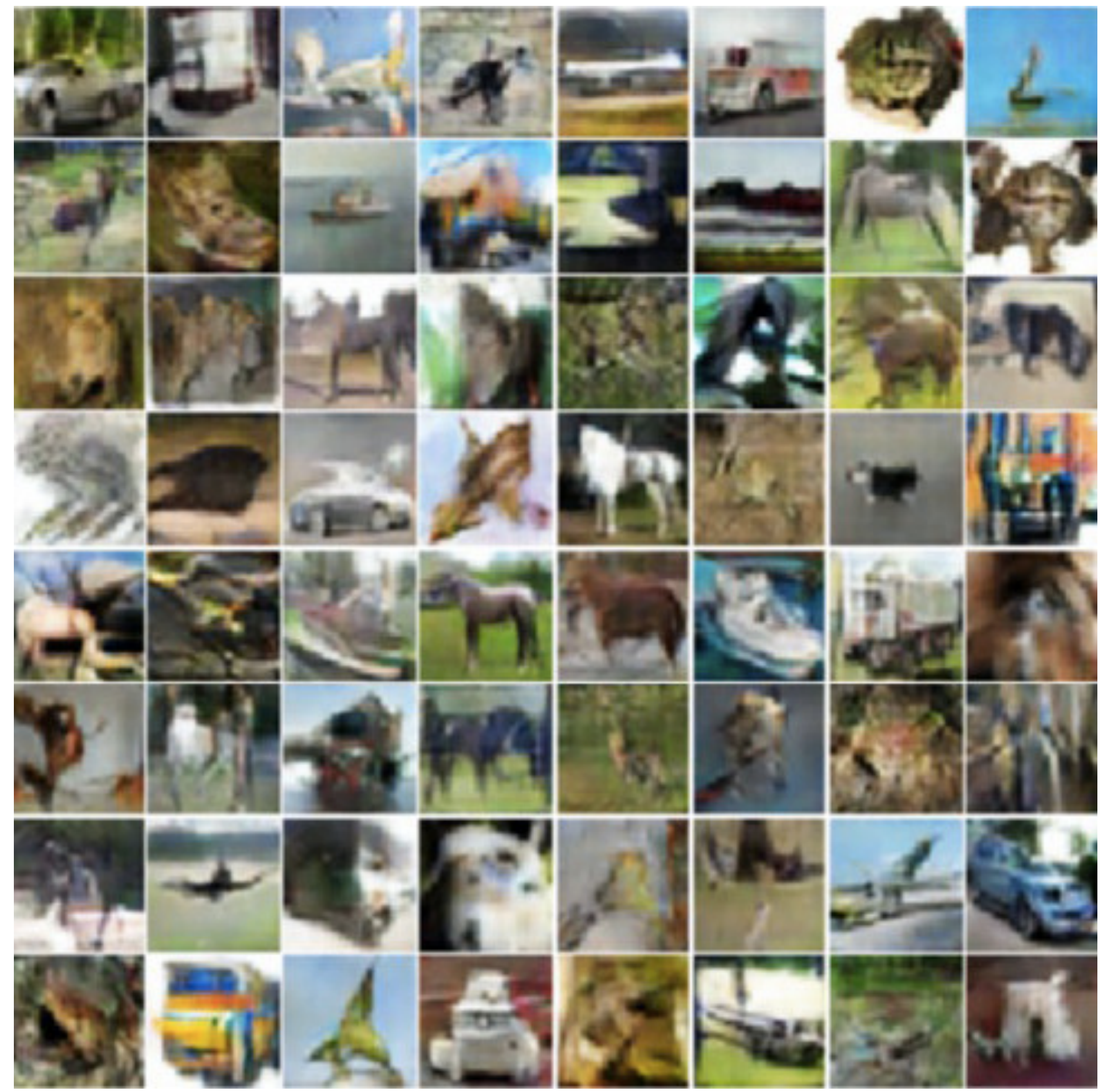
- Add minibatch features that classify each example by comparing it to other members of the minibatch (Salimans et al 2016)

- Nearest-neighbor style features detect if a minibatch contains samples that are too similar to each other

Minibatch GAN on CIFAR



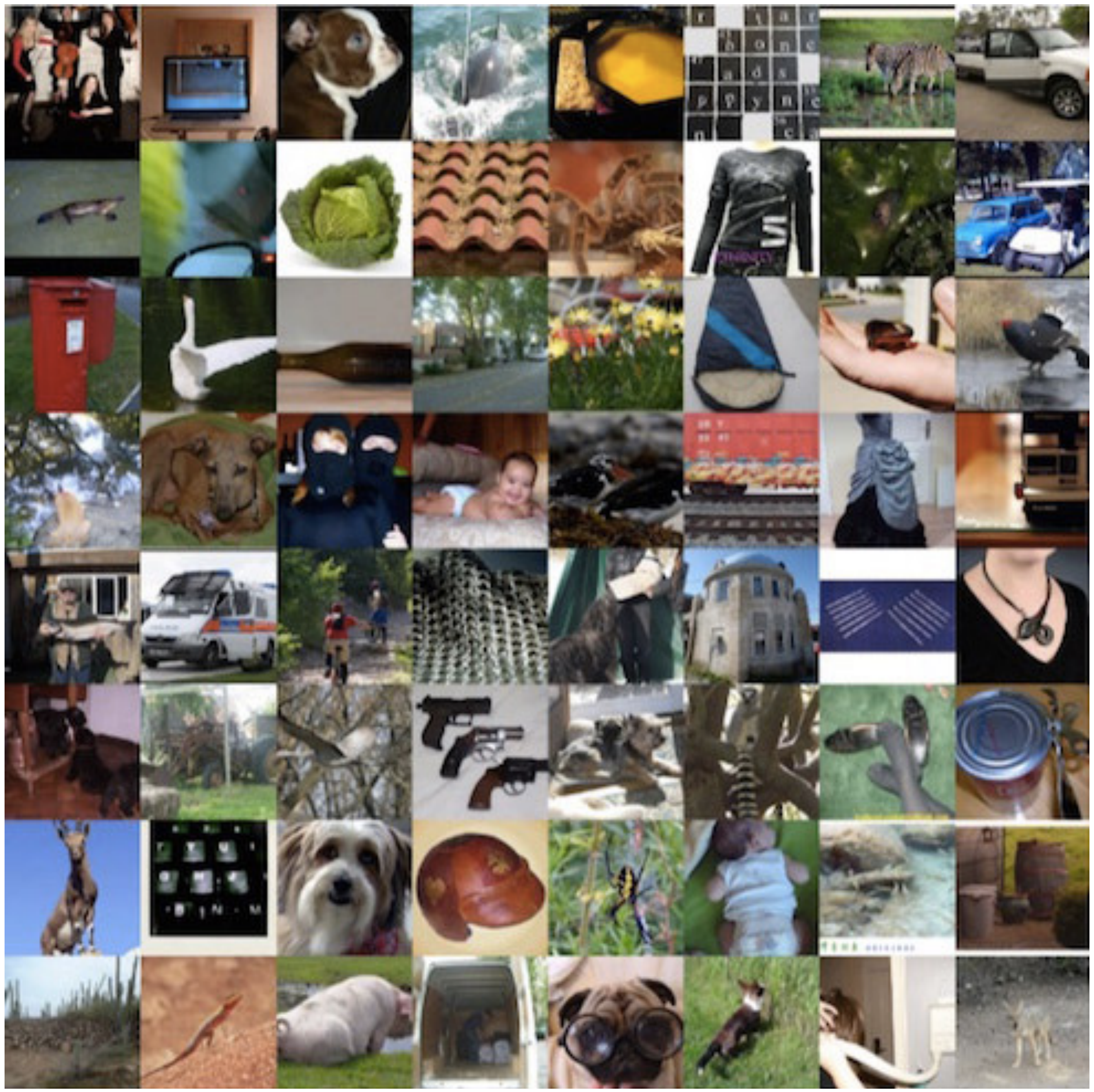
Training Data



Samples

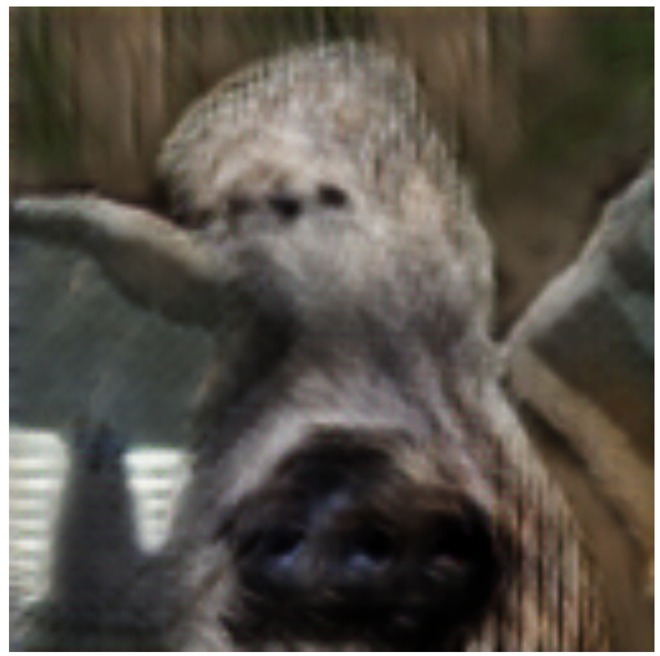
(Salimans et al 2016)

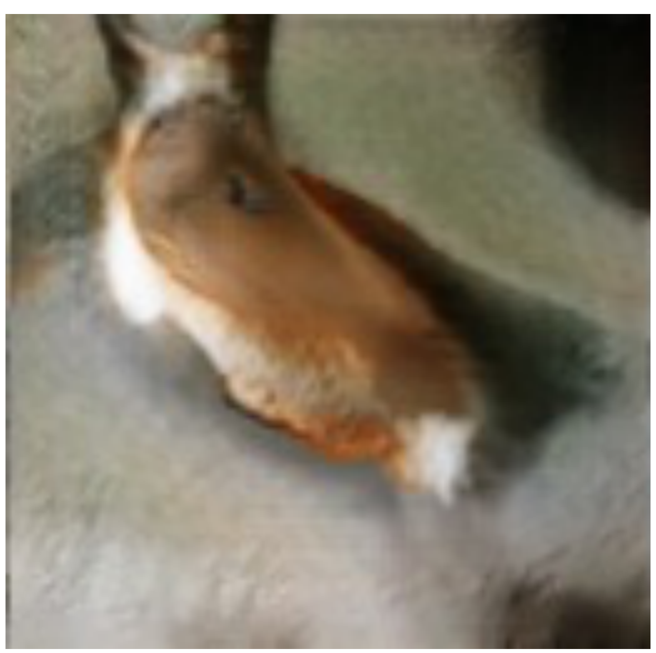
Minibatch GAN on ImageNet



(Salimans et al 2016)

Cherry-Picked Results

(Goodfellow 2016)

Problems with Counting





(Goodfellow 2016)

Problems with Perspective

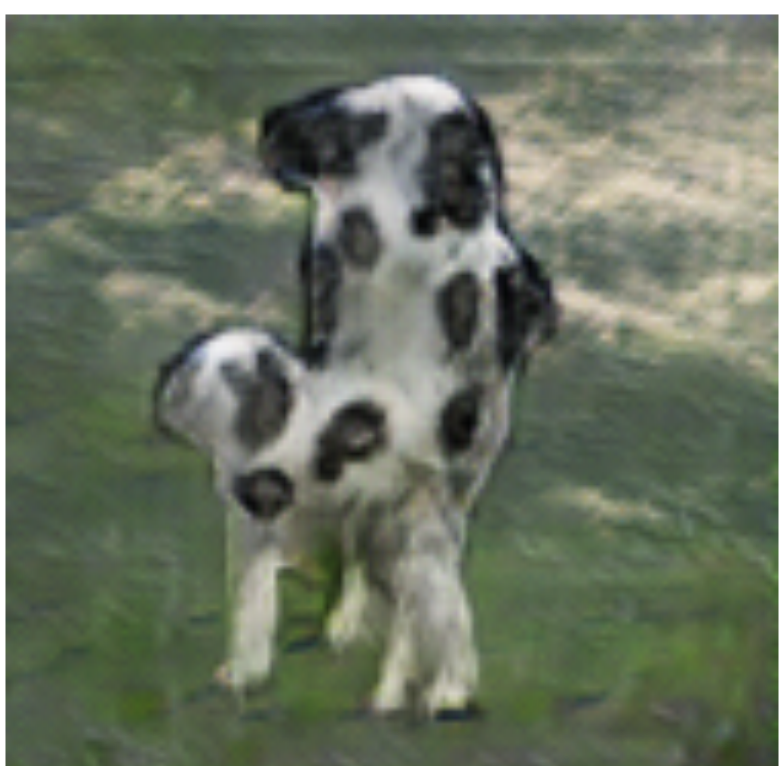


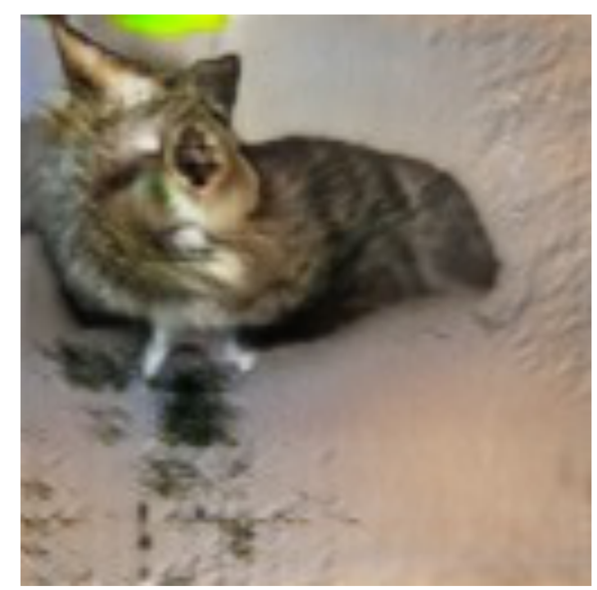



(Goodfellow 2016)

Problems with Global Structure





(Goodfellow 2016)

Image to Image Translation

(Isola et al 2016)

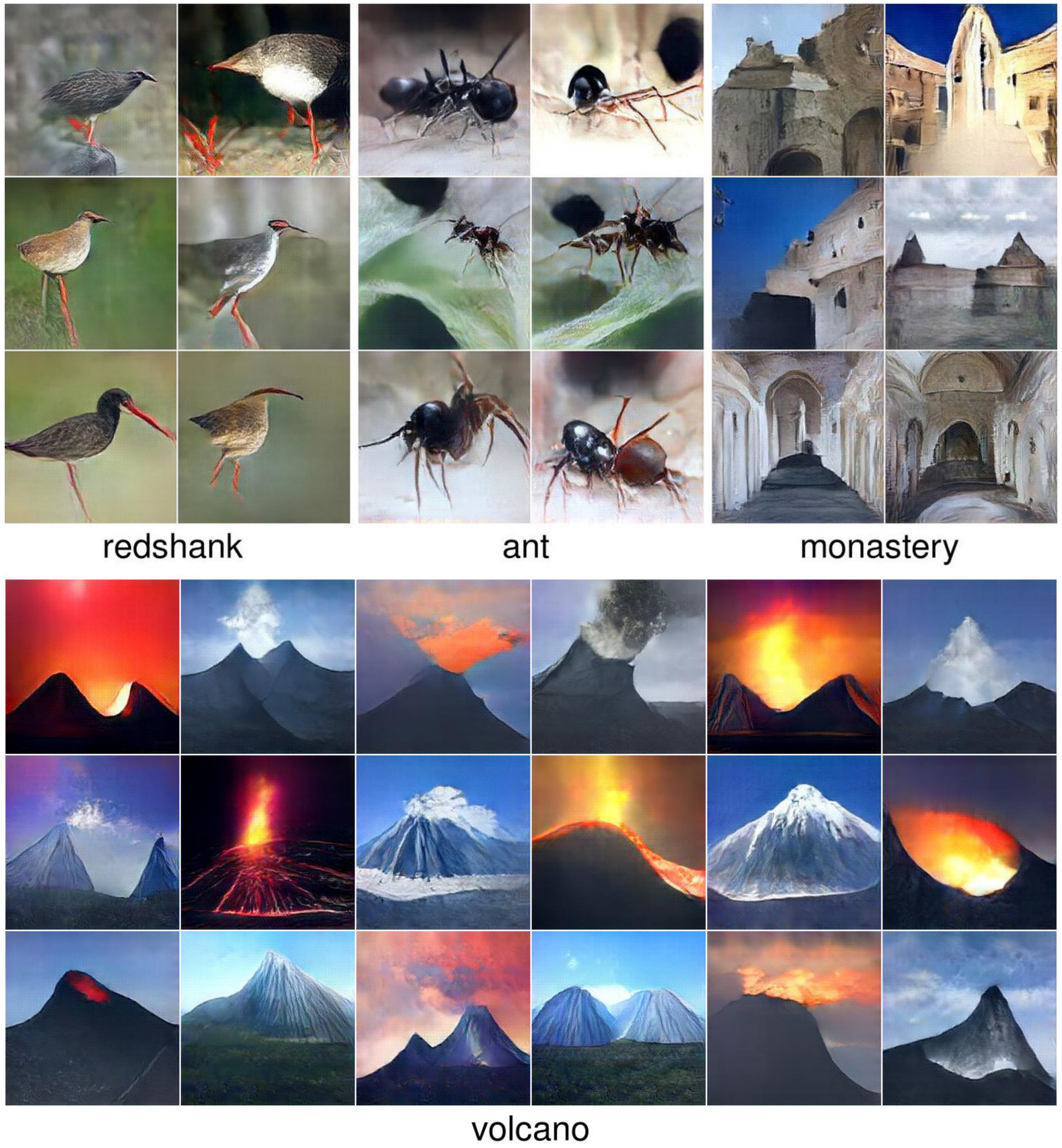
Plug and Play Generative Models

- 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

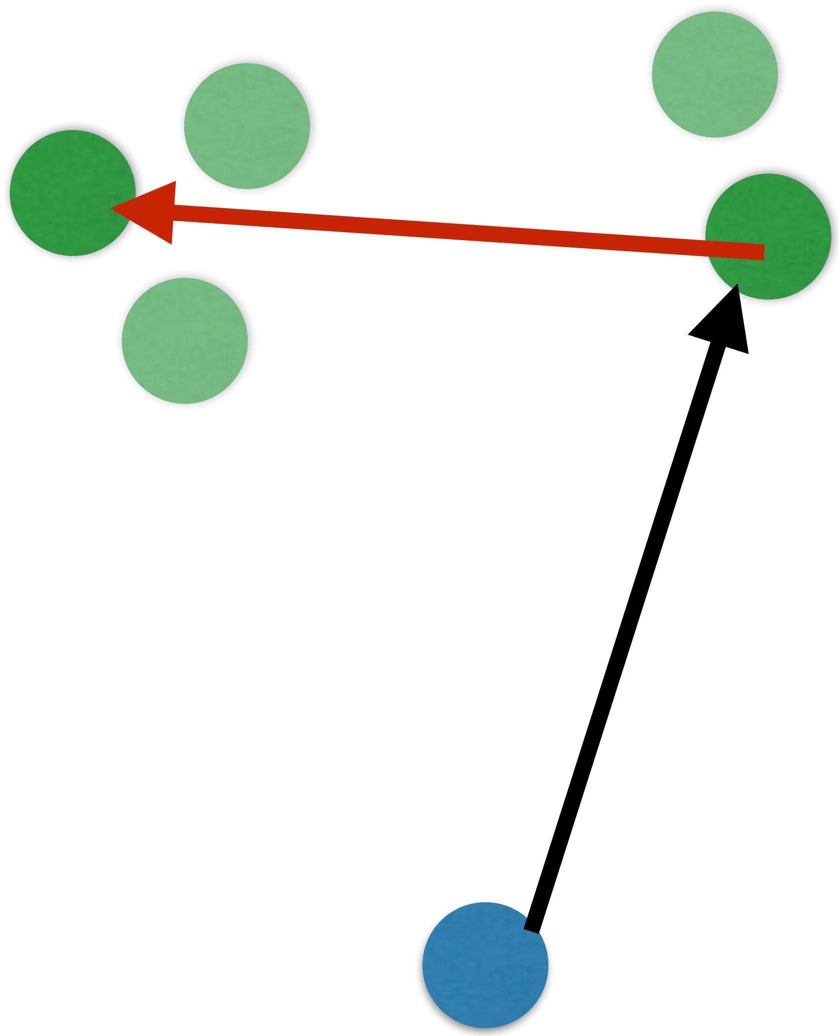
PPGN Samples



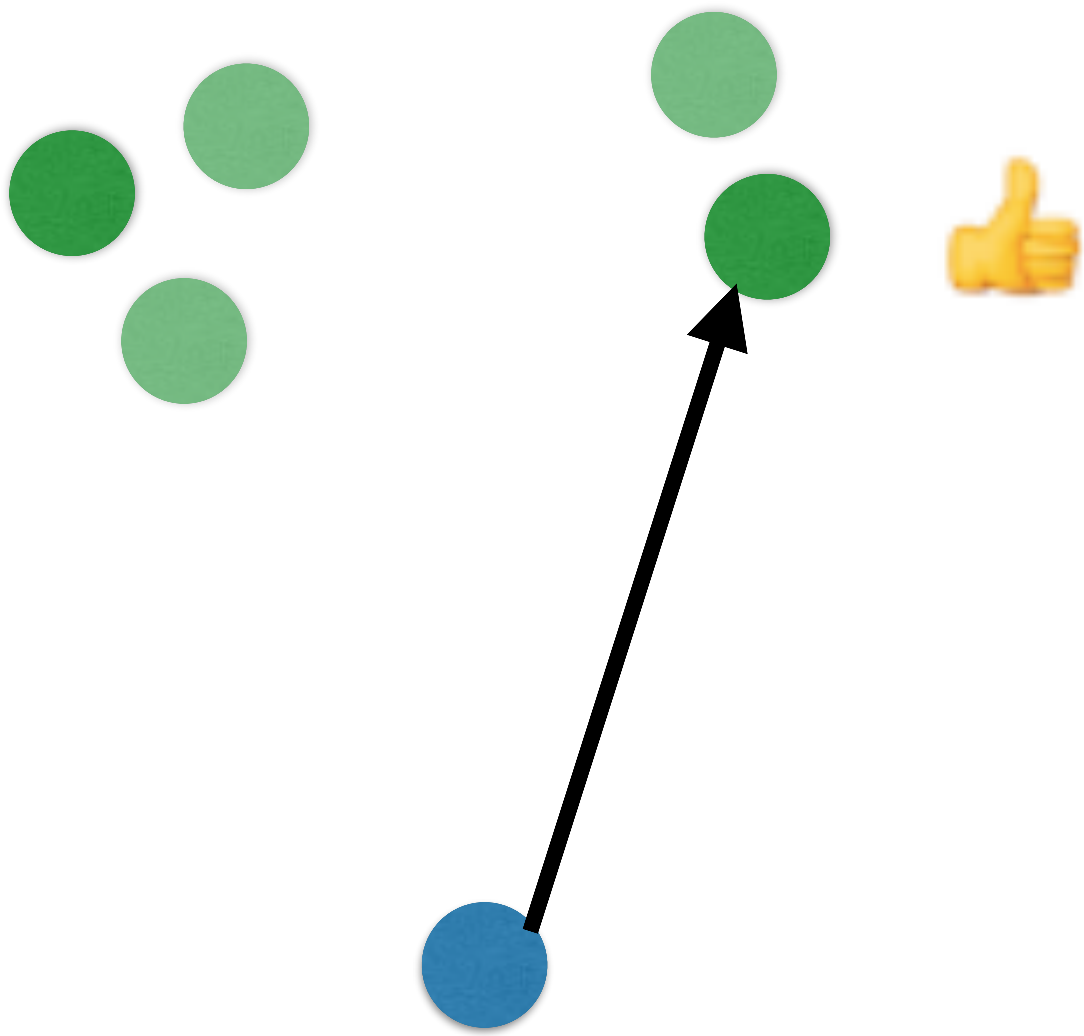
(Nguyen et al 2016)

GANs allow many answers

Mean Squared Error

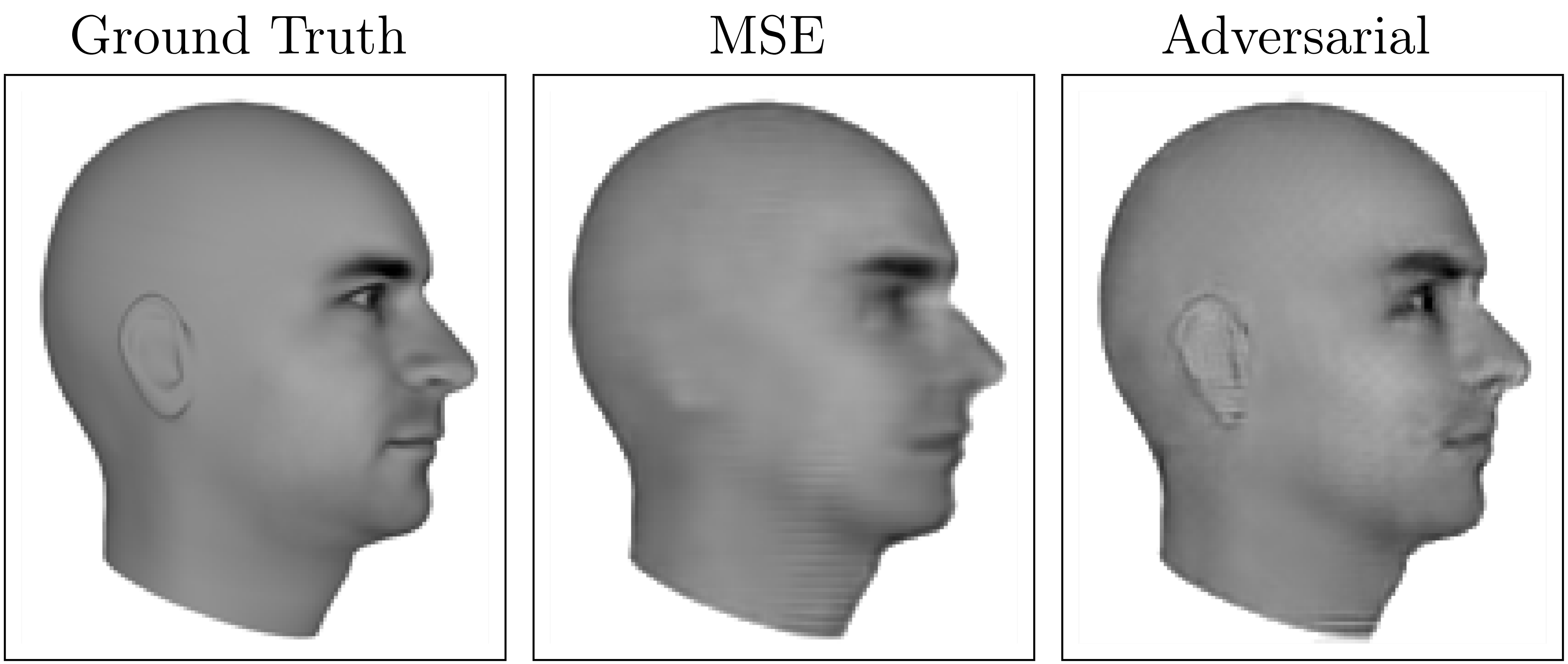


GANs



(Goodfellow 2016)

Next Video Frame Prediction



(Lotter et al 2016)

Adversarial training for people

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

(Goodfellow 2016)

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