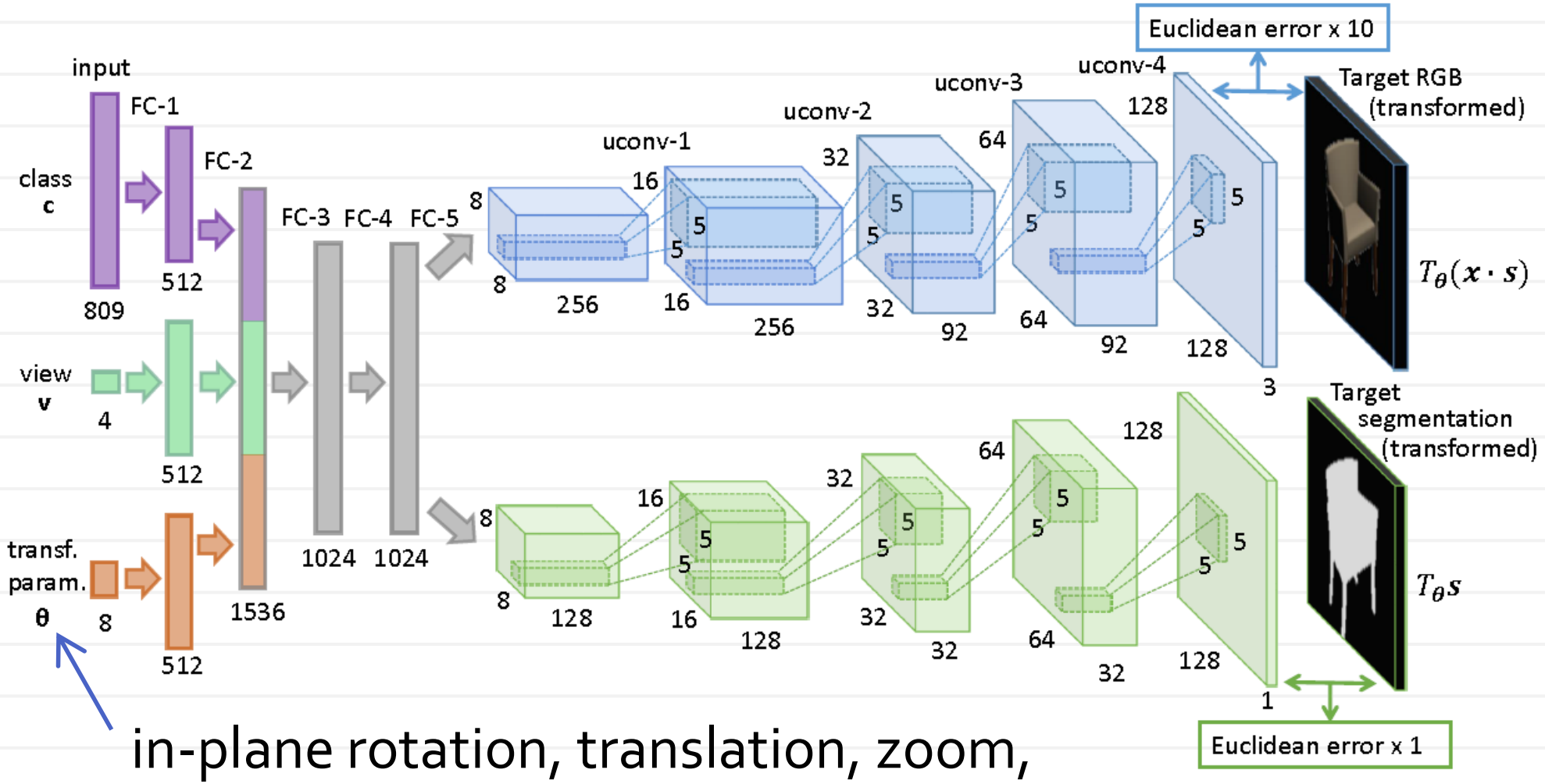
Lecture 7: Latent Models and Autoencoders

Feed-forward conditional generation



in-plane rotation, translation, zoom, stretching horizontally or vertically, changing hue, changing saturation, changing brightness

[Dosovitskiy et al. CVPR 2015]

High-res generation with perceptual losses

[Chen and Koltun ICCV17]

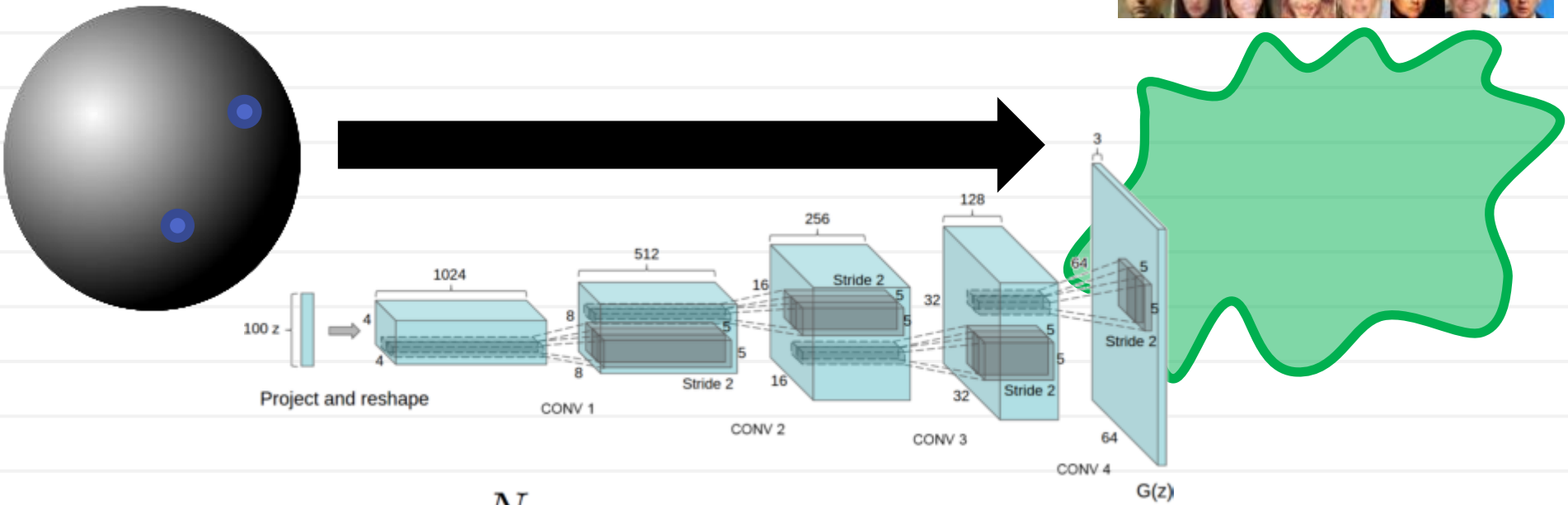
Latent models of images

- Previous results (chairs, Cityscapes "fakes") are very impressive yet "intensely" supervised
- What about learning from raw images?
- Long-long history of research, e.g. Eigenface model ([Sirovich&Kirby 1987], [Turk&Pentland 1991])

Latent models of images

[Bojanowski et al. 2017]: the simplest deep latent model for images:

$$\mathcal{Z} = \mathcal{B}(r, d, p) = \left\{ z \in \mathbb{R}^d : \|z\|_p \leq r \right\}$$



$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \left[\min_{z_i \in \mathcal{Z}} \ell\left(g_\theta(z_i), x_i\right) \right]$$

Latent models of images: reconstructions

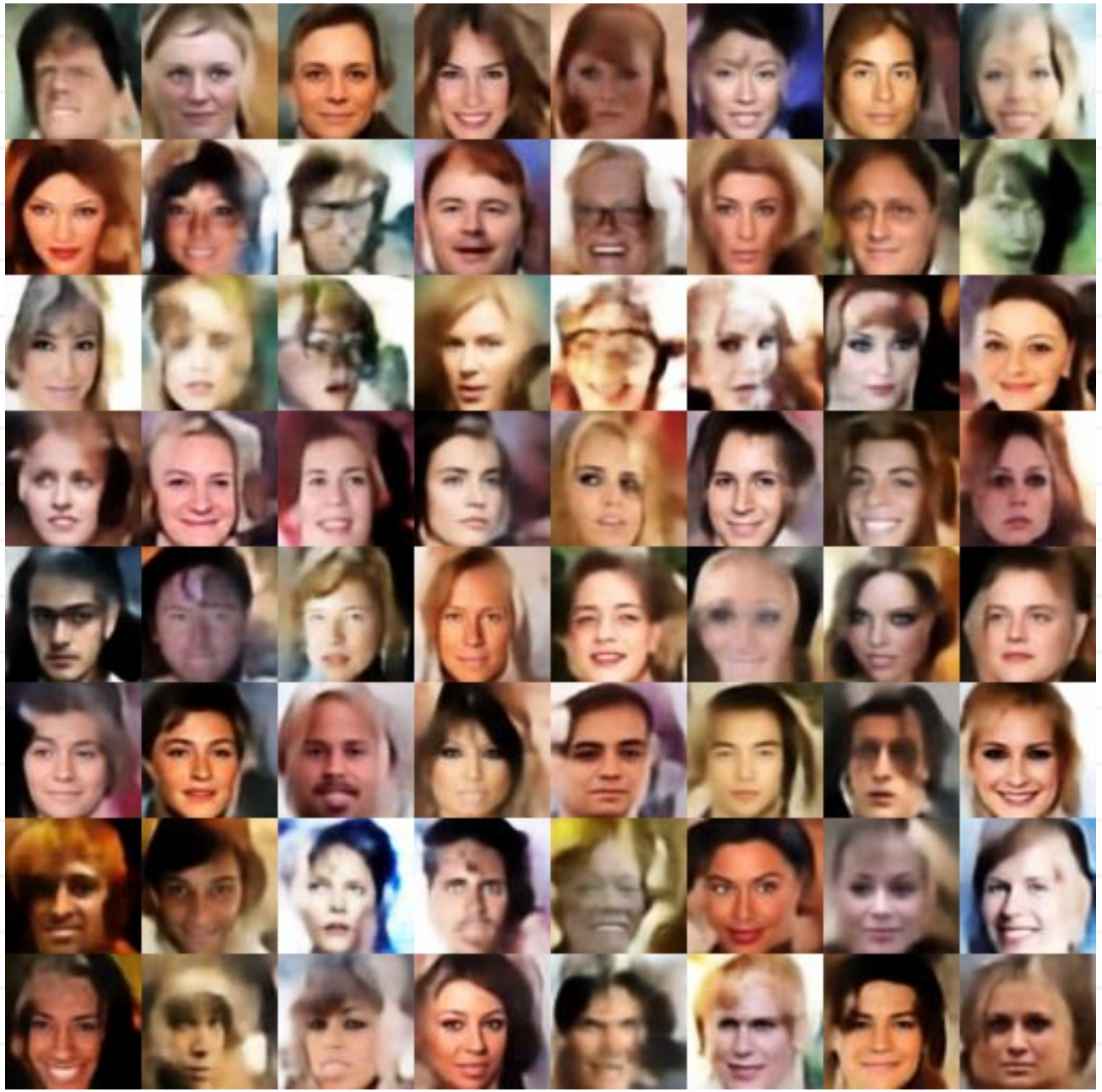
d =512

[Bojanowski et al. 2017]

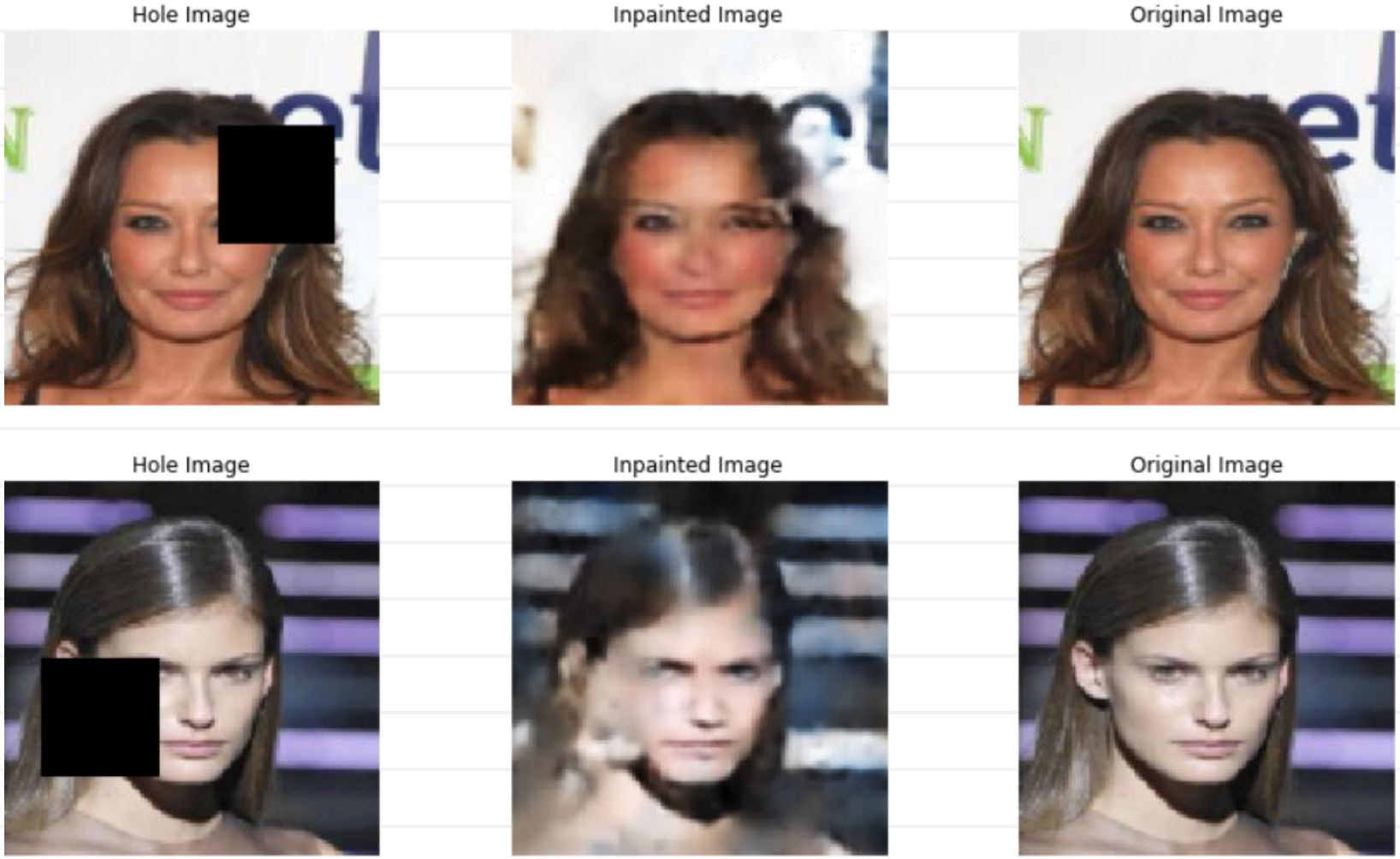
Latent models of images: samples

1. Fit Gaussian in the latent space
2. Sample and generate

[Bojanowski et al. 2017]



Latent models of images: restoration



$$\min_{z \in \mathcal{Z}} \|(g_\theta(z) - x) \odot m\|$$

[thanks to ShahRukh Athar]

Problems with direct optimization

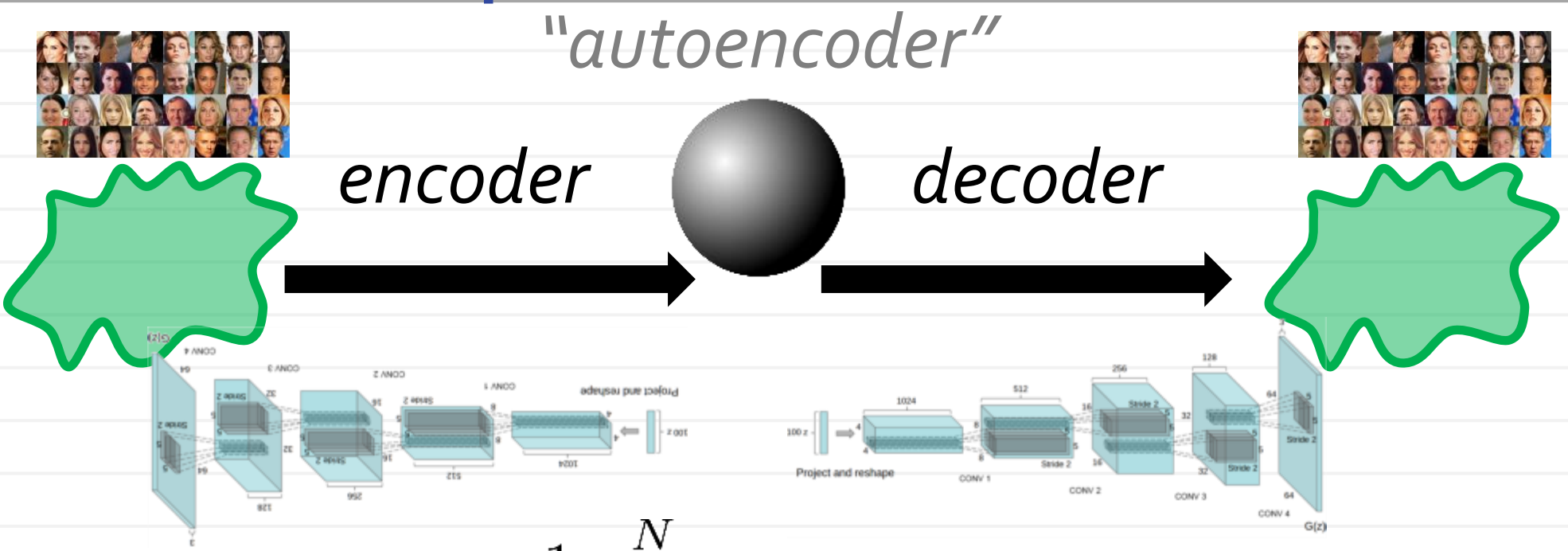
- Previous model requires optimization to fit new images:

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \left[\min_{z_i \in \mathcal{Z}} \ell\left(g_\theta(z_i), x_i\right) \right]$$

- Previous model requires storing latent vectors during learning (not scalable)

- **Idea**: predict latent vectors with a new network (*encoder*) from the image

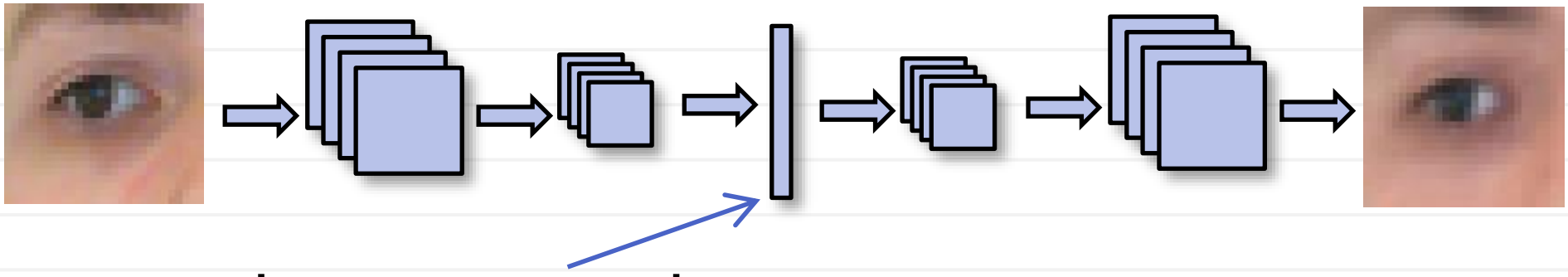
From direct optimization to Autoencoders



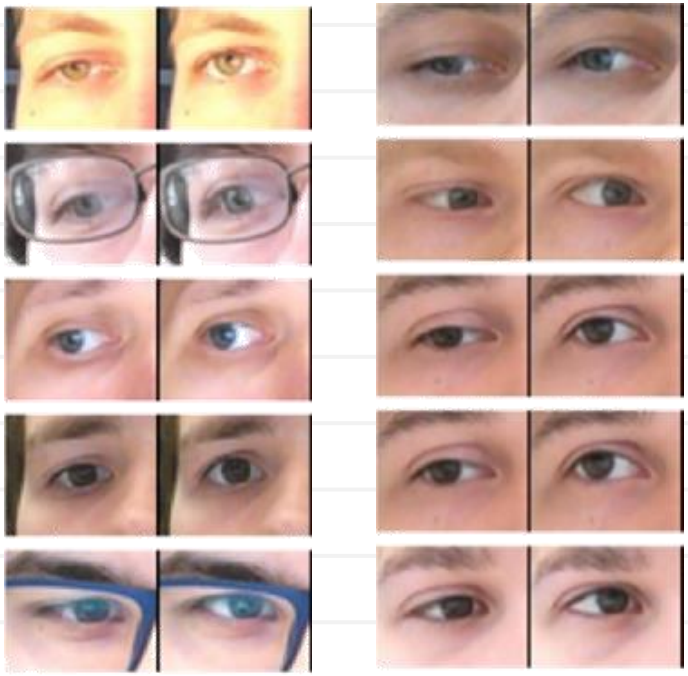
$$\min_{\phi,\theta} \frac{1}{N} \sum_{i=1}^{N} l\left(g_\theta(e_\phi(x_i)), x_i\right)$$

- Learning still unsupervised
- No scalability issues
- A lot depends on the loss

Smart image editing

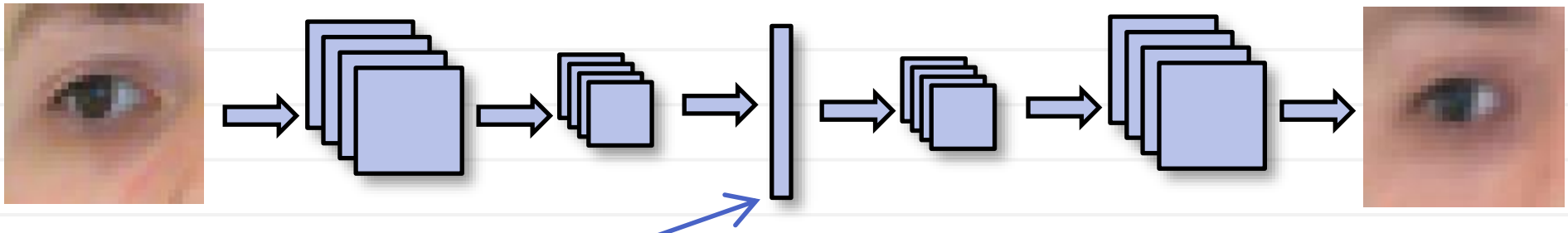


Low-dim space, where we can estimate semantically meaningful directions

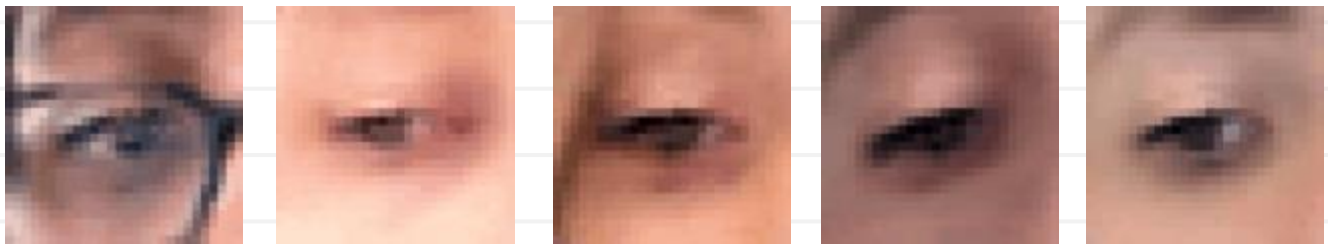


Given a bunch of pairs we can estimate a vector for gaze redirection

Smart image editing



Low-dim space, where we can estimate semantically meaningful directions

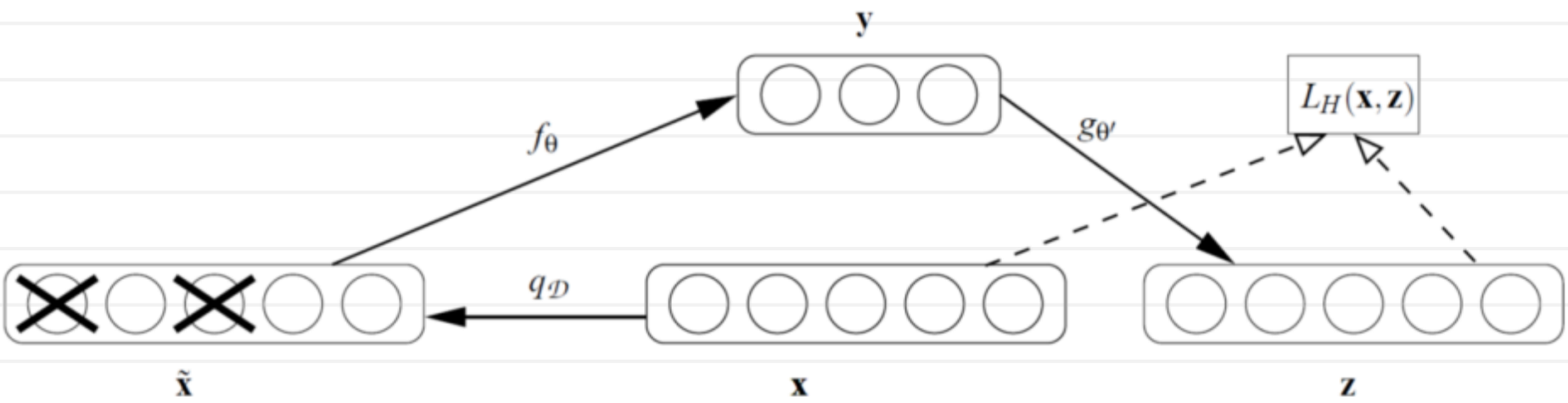


Thanks to L.Yekimov
(+F.Chervinsky,D.Kononenko,D.Sungatullina,Y.Ganin)

Preventing overfitting in autoencoders

- Autoencoders underfit (especially with L2-loss)
- Autoencoders overfit (data compression performance)
- Standard tricks have all been tried (weight decay, activation penalization, **noise**)
- Denoising autoencoders are particularly popular

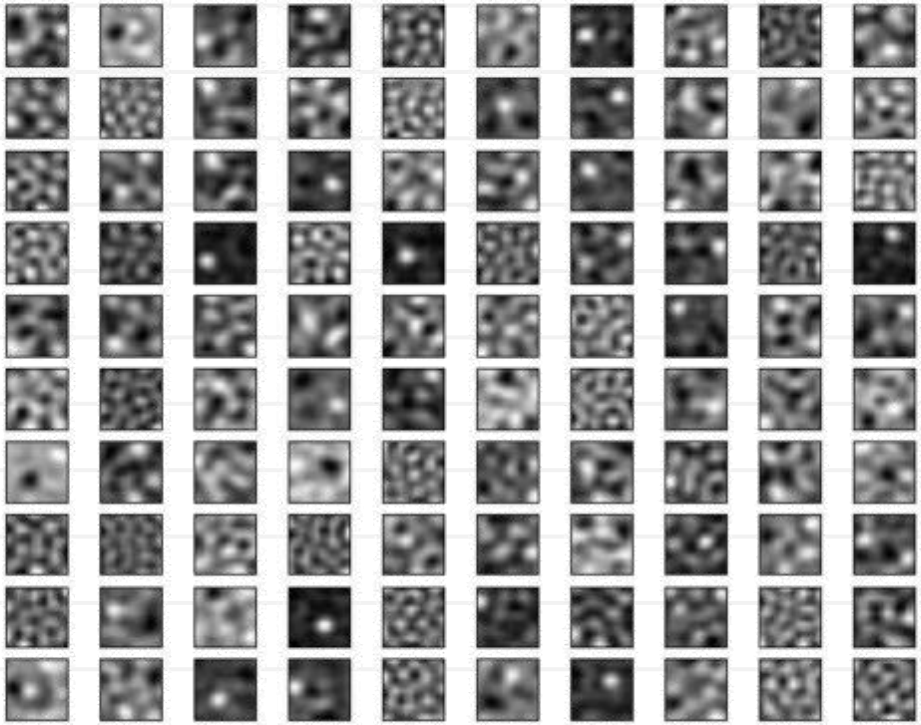
Stacked denoising autoencoders



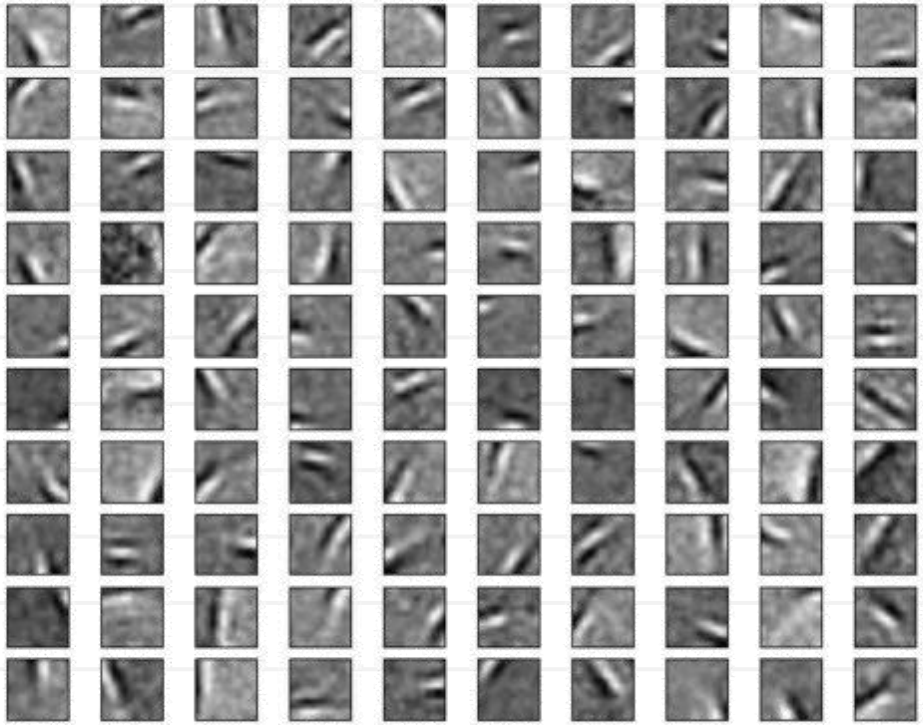
- Autoencoder receives corrupted input and must denoise
- Layer-wise/stacked training was once popular (no longer)

[Vincent et al. 2010]

Stacked denoising autoencoders



Autoencoder on image patches learned with weight decay

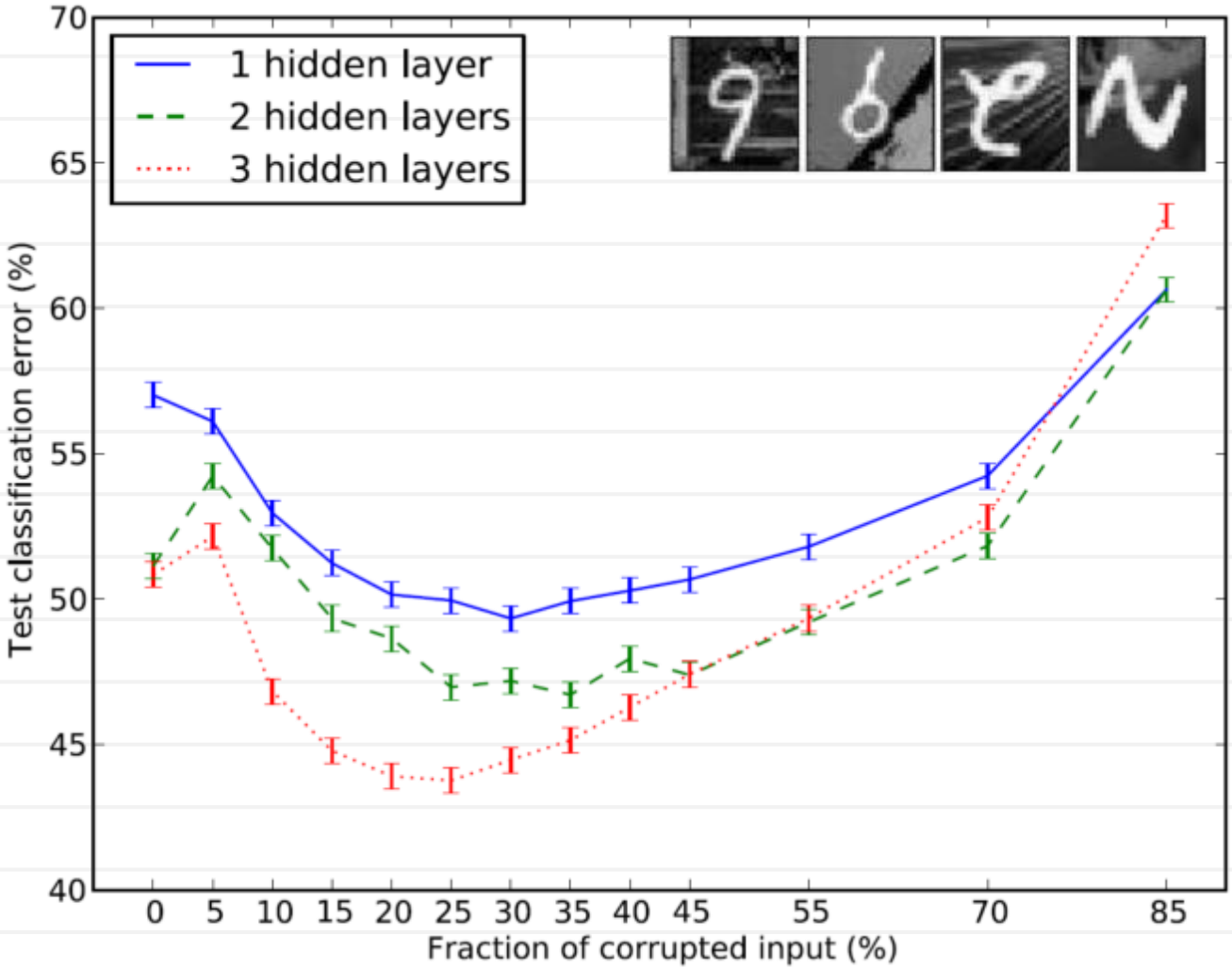


Autoencoder on image patches learned with denoising

[Vincent et al. 2010]

Stacked denoising autoencoders

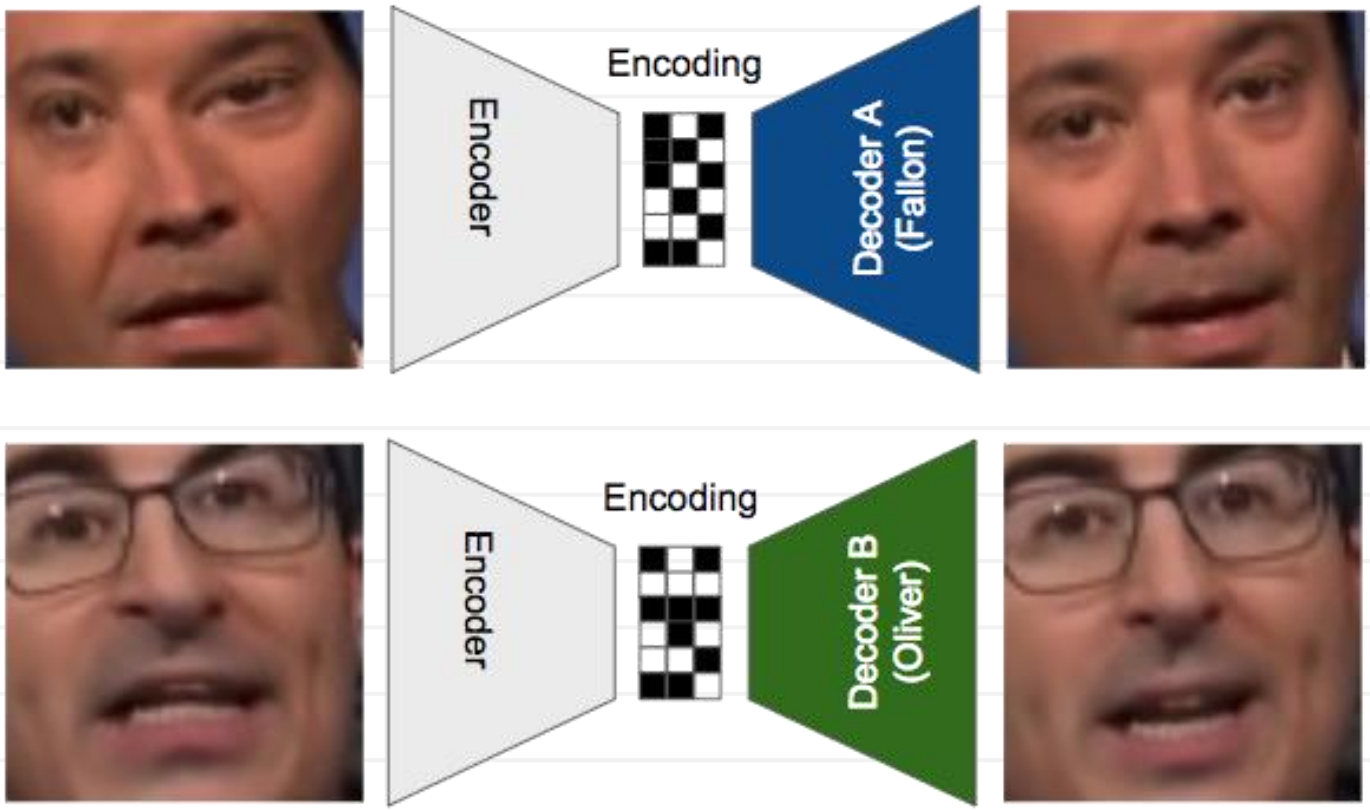
Supervised classification accuracy after training on features learned by the autoencoder:



[Vincent et al. 2010]

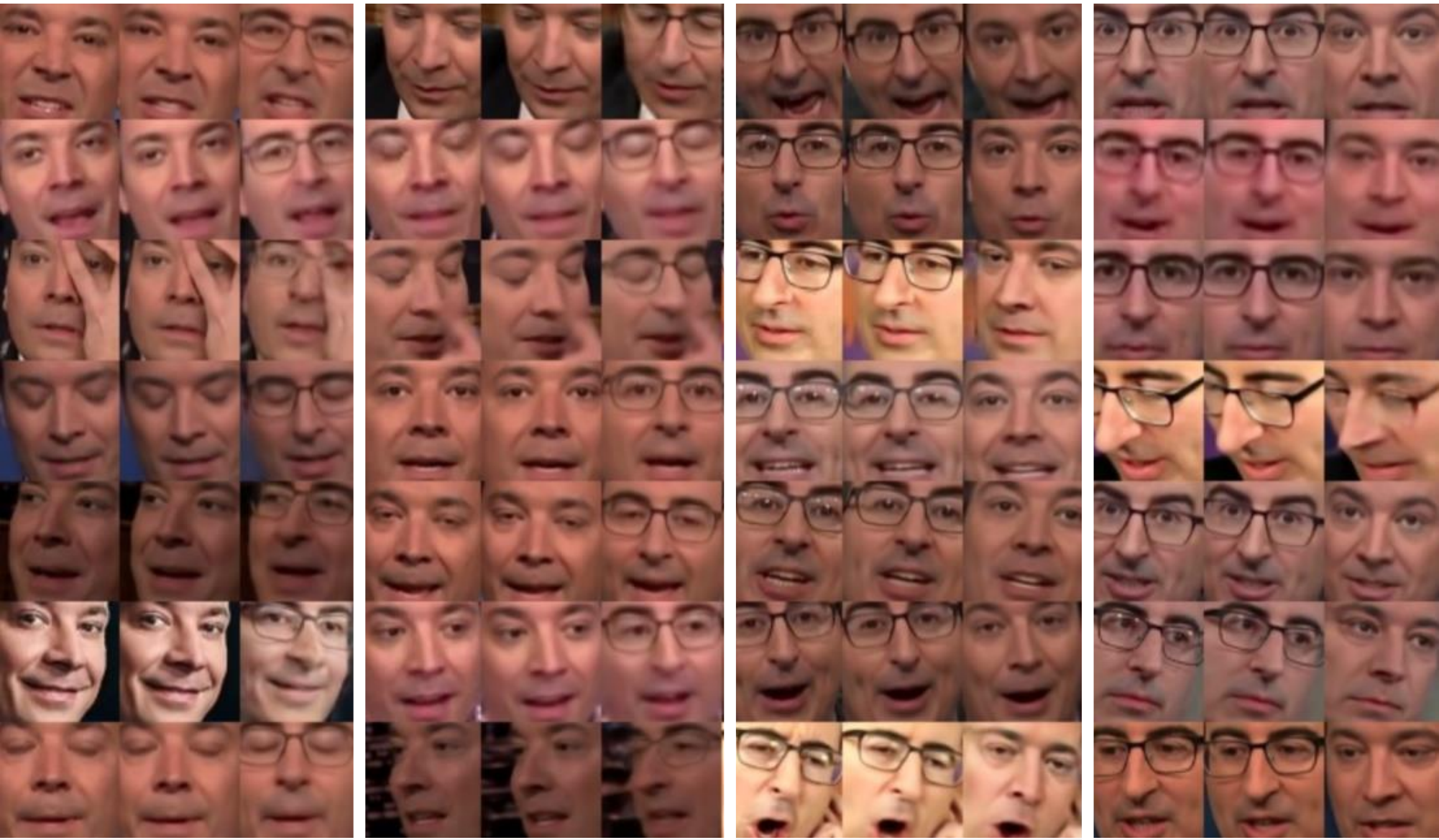
DeepFake methods

Paired "dewarping" autoencoders:



[images source: Gaurav Oberoi blog]

DeepFake system



[images source: Gaurav Oberoi blog]

DeepFake system

[video source: Gaurav Oberoi blog]

Probabilistic latent model: VAE

Ideally, we want to do ML learning in such model: $\frac{1}{N}\sum_{i=1}^{N}\log\left(\int_z p_\theta(x_i|z)dz\right) \to \max_\theta$

$p_\theta(z|x) = ?$

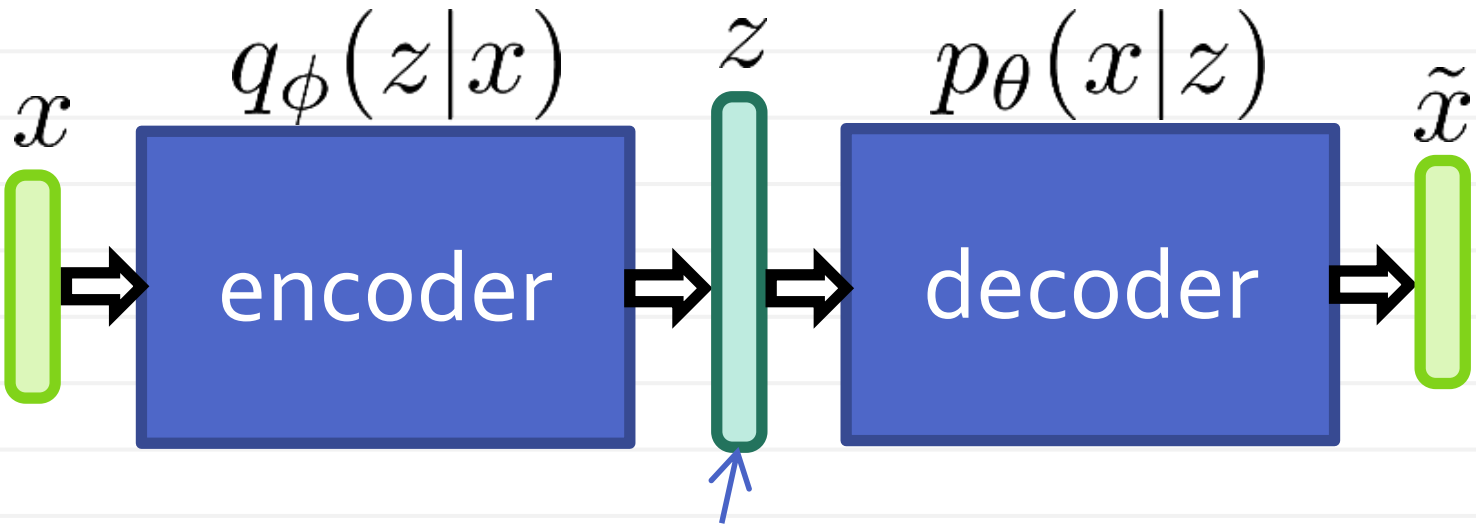
$p_\theta(x|z)$

Idea 1: use $q_\phi(z|x)$ instead of $p_\theta(z|x)$

Idea 2: Learn the parameters of both direct and inverse mappings simultaneously

[Kingma & Welling 14]

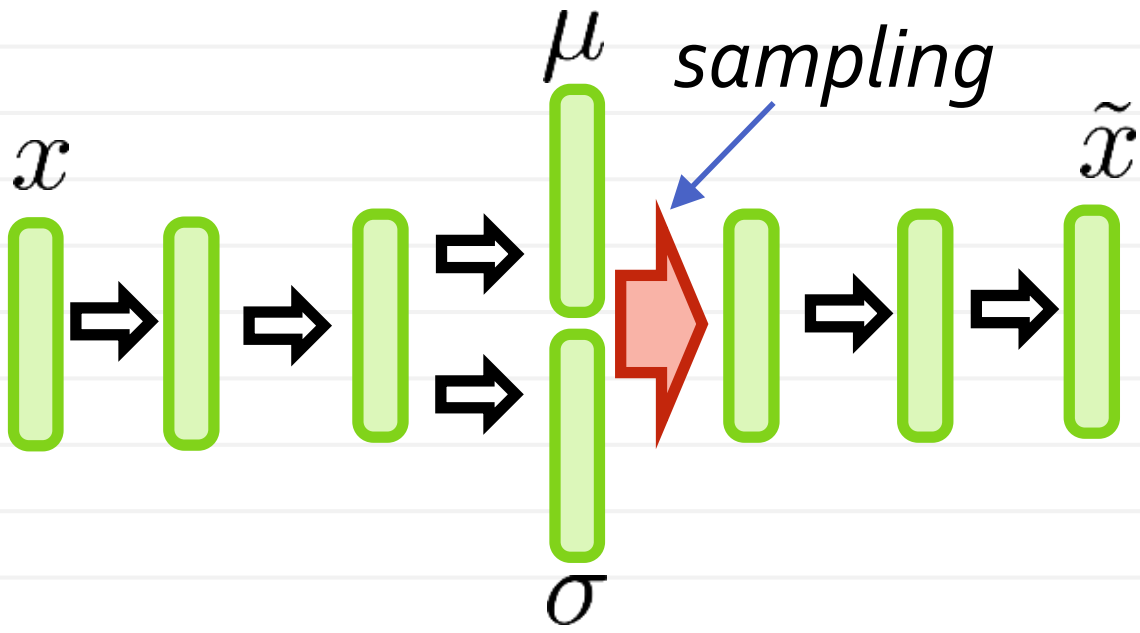
Variational autoencoder (VAE)



[Kingma & Welling 14]

Variational autoencoder (VAE)

E.g.:



$$[\mu, \sigma] = e_\phi(x) \quad q_\phi(z|x) = \mathcal{N}\left(\mu, \text{diag}(\sigma^2)\right)$$

$$\tilde{z} \sim \mathcal{N}\left(\mu, \text{diag}(\sigma^2)\right) \quad \tilde{x} = d_\theta(\tilde{z})$$

$$p_\theta(x|z) = \mathcal{N}\left(\tilde{x}, \alpha\mathbf{I}\right)$$

[Kingma & Welling 14]

Variational lower bound

$$\log p(x) = \log \int_z p(x,z) dz$$

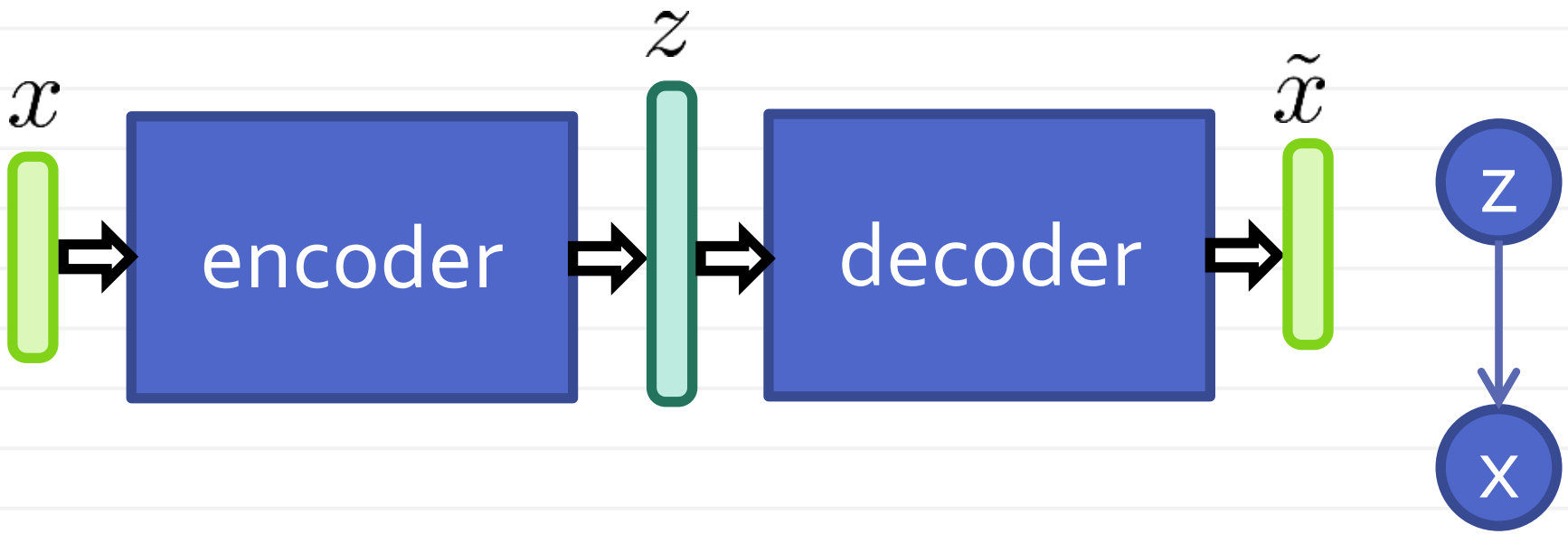
$$= \log \int_z p(x,z) \frac{q(z|x)}{q(z|x)} dz = \log \mathbb{E}_{q(z|x)} \frac{p(x,z)}{q(z|x)}$$

$$= \log \mathbb{E}_{q(z|x)} \frac{p(x|z)\, p(z)}{q(z|x)}$$

$$\geq \mathbb{E}_{q(z|x)} \log p(x|z) + \mathbb{E}_{q(z|x)} \log \frac{p(z)}{q(z|x)}$$

$$= \mathbb{E}_{q(z|x)} \log p(x|z) - \mathrm{KL}\big(q(z|x) \,\|\, p(z)\big)$$

Variational lower-bound



$$\log p(x) \geq \underbrace{-\mathrm{KL}\big(q(z|x) \,\|\, p(z)\big)}_{\text{regularization}} + \underbrace{\mathbb{E}_{q(z|x)} \log p(x|z)}_{\sim\text{denoising auto-encoder}}$$

[Kingma & Welling 14]

Minimizing regularization term

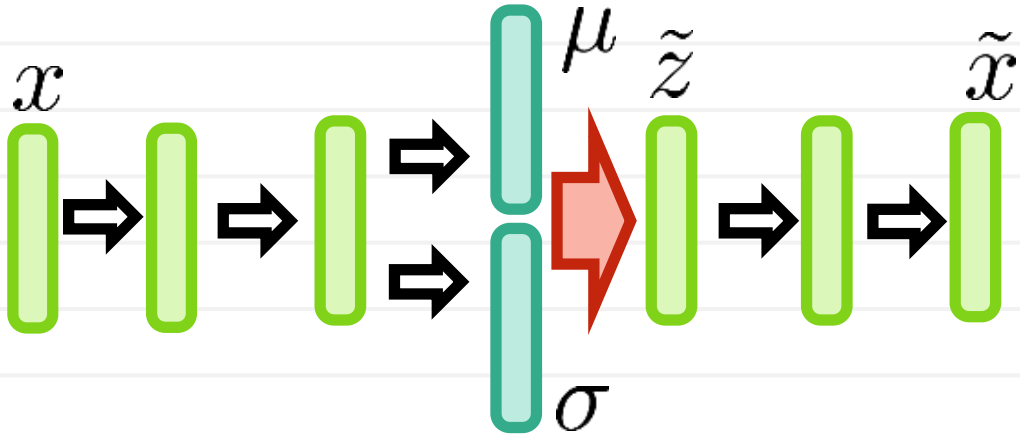
$$\log p(x) \geq \underbrace{-\mathrm{KL}\big(q(z|x) \,\|\, p(z)\big)}_{\text{regularization}} + \mathbb{E}_{q(z|x)} \log p(x|z)$$

$$p(z) = \prod_i \mathcal{N}(z_i|0,1)$$

$$\mathrm{KL}\big(q_\phi(z|x) \,\|\, p(z)\big) = \frac{1}{2}\sum_i (\mu_i^2 + \sigma_j^2 - 1 - \log \sigma_j^2)$$

[Kingma & Welling 14]

Optimizing the reconstruction term



$$\mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) \to \max_\theta$$

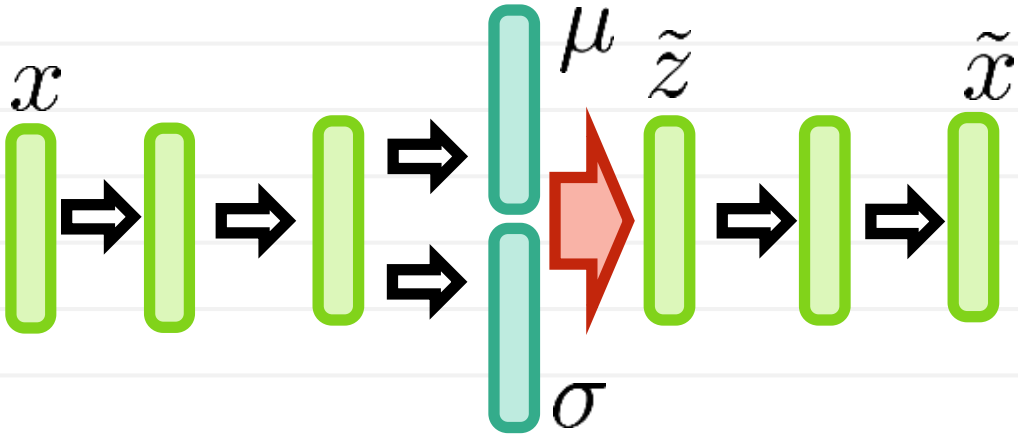
$$[\mu, \sigma] = e_\phi(x) \quad q_\phi(z|x) = \mathcal{N}\big(z|\mu, \text{diag}(\sigma^2)\big)$$

$$\tilde{z} \sim \mathcal{N}\big(\mu, \text{diag}(\sigma^2)\big) \quad \tilde{x} = d_\theta(\tilde{z})$$

$$p_\theta(x|z) \approx \mathcal{N}\big(x|\tilde{x}, \alpha\mathbf{I}\big)$$

$$\mathbb{E}_{q_\phi(z|x)} \|d_\theta(z) - x\|^2 \to \min_\theta$$

Reparameterization trick



$$\mathbb{E}_{q_\phi(z|x)}\|d_\theta(z) - x\|^2 \to \min_\theta$$

$$\tilde{z} \sim \mathcal{N}\left(\mu, \text{diag}(\sigma^2)\right) \quad \tilde{x} = d_\theta(\tilde{z})$$

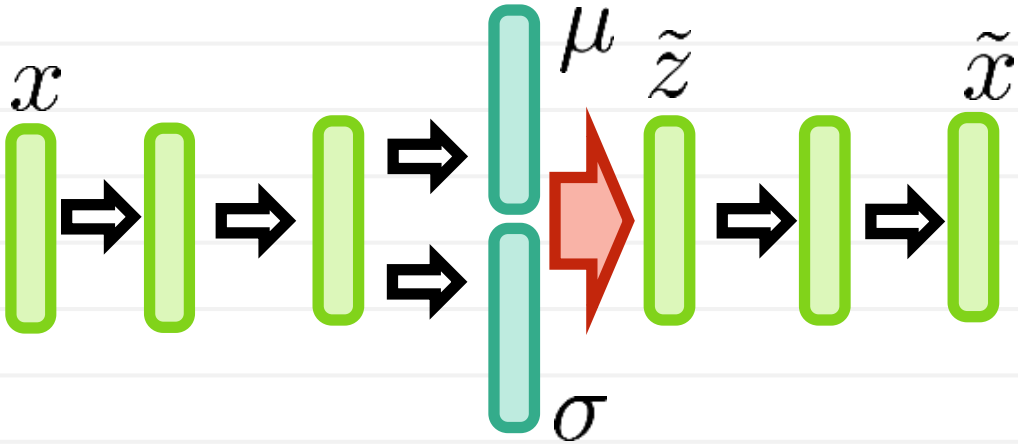
$$\epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad \tilde{z} = \mu + \sigma \odot \epsilon$$

Sampling at every iteration

$$L(\phi, \theta) = \sum_i \left\| d_\theta\left(\mu_\phi(x_i) + \sigma_\phi(x_i) \odot \epsilon_i\right) - x_i \right\|^2$$

We can optimize it in a standard way!

Backprop through sampling is easy



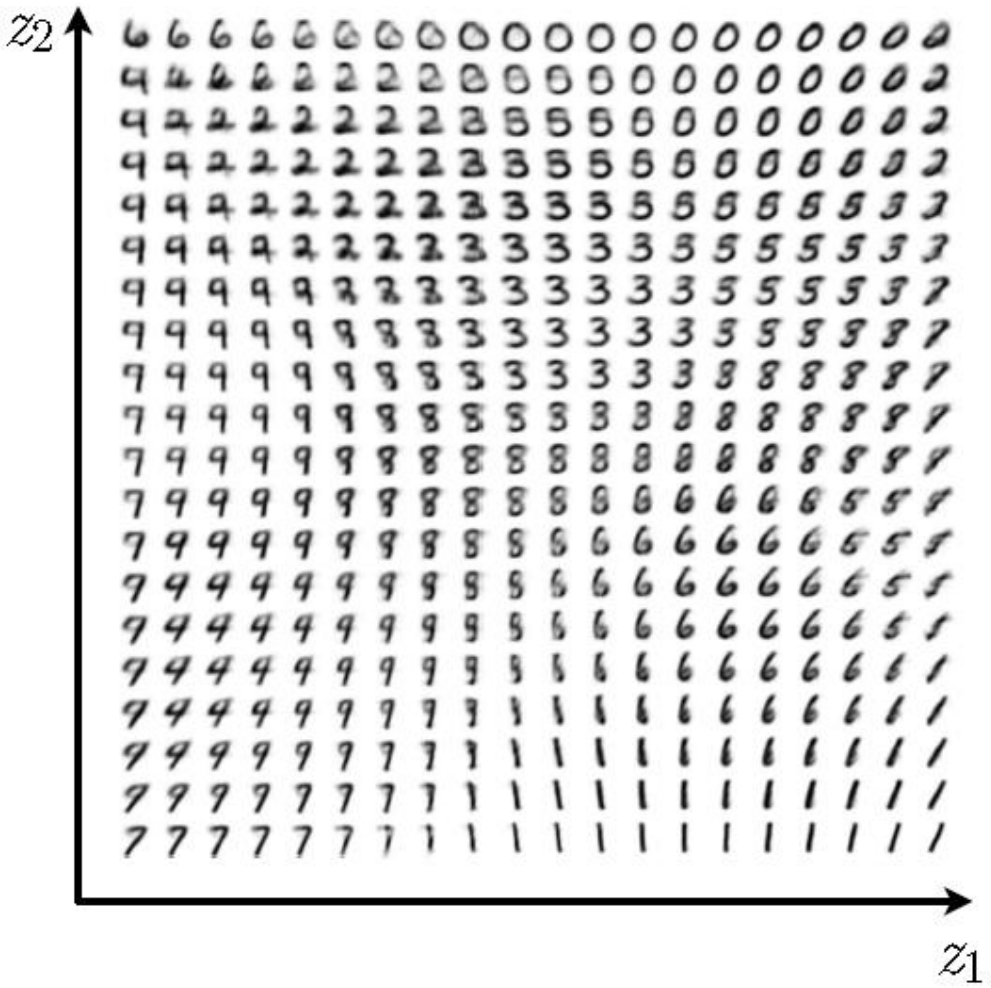
$$\epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad \tilde{z} = \mu + \sigma \odot \epsilon$$

$$L(\phi, \theta) = \sum_i \left\| d_\theta\left(\mu_\phi(x_i) + \sigma_\phi(x_i) \odot \epsilon_i\right) - x_i \right\|^2$$

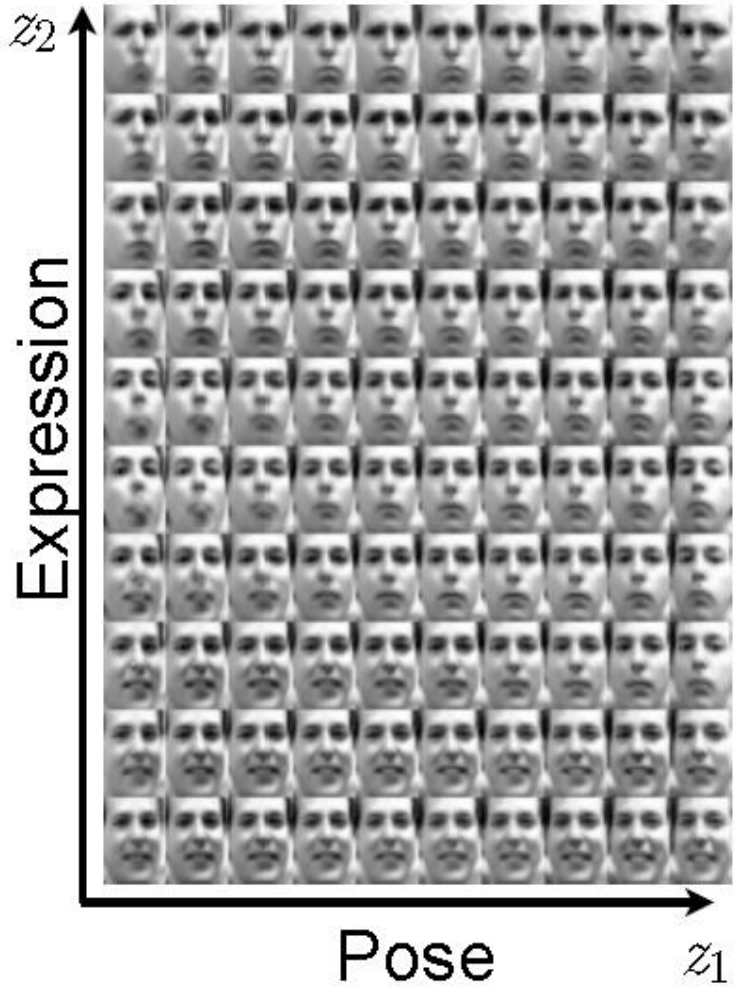
$$\frac{dL}{d\mu} = \frac{dL}{d\tilde{z}} \qquad \frac{dL}{d\sigma} = \frac{dL}{d\tilde{z}} \odot \epsilon$$

VAE-learned manifolds

MNIST:

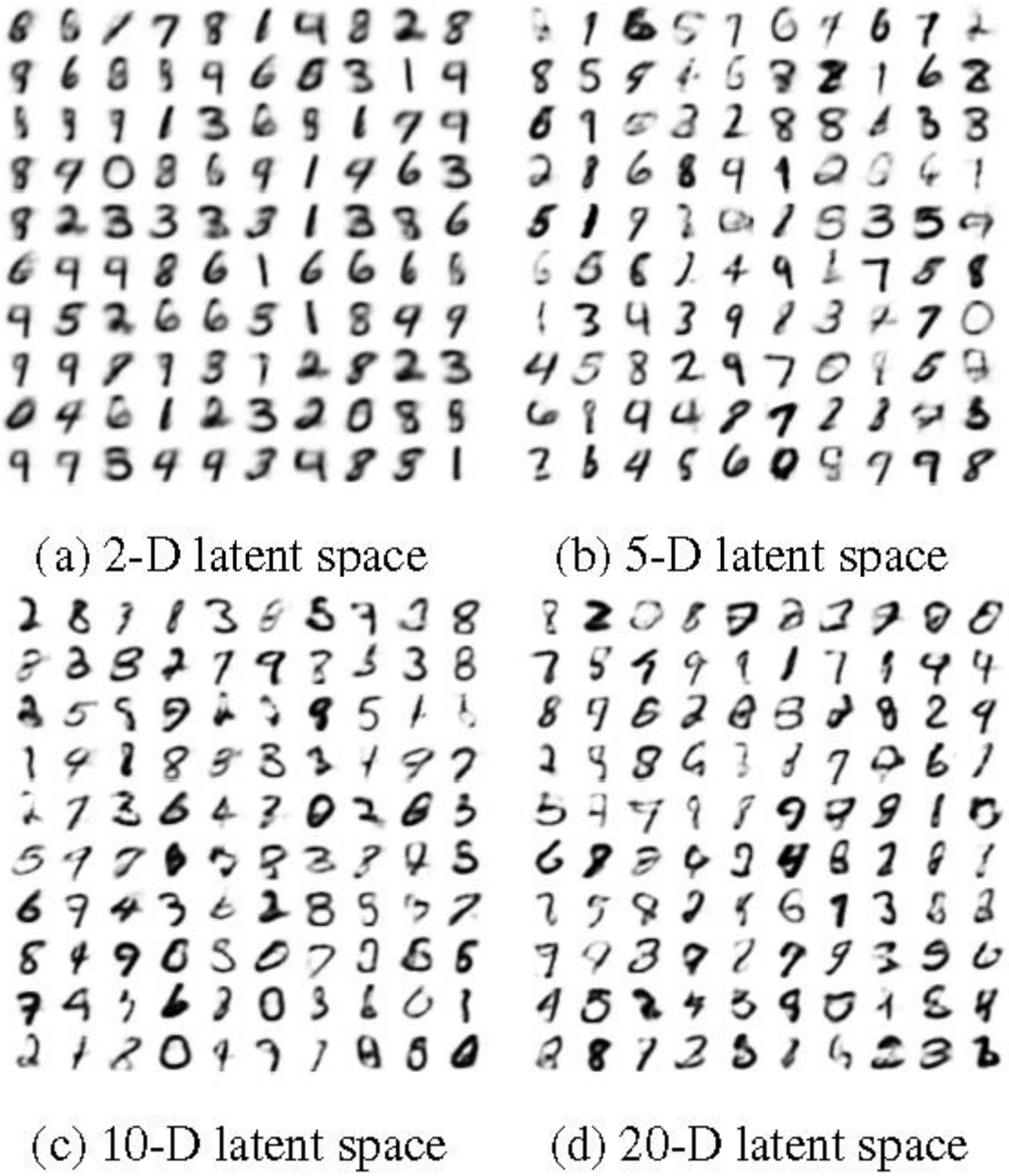


Frey Face dataset:



[Kingma & Welling 14]

VAE-learned manifolds



(a) 2-D latent space

(b) 5-D latent space

(c) 10-D latent space

(d) 20-D latent space

[Kingma & Welling 14]

Collapsing components

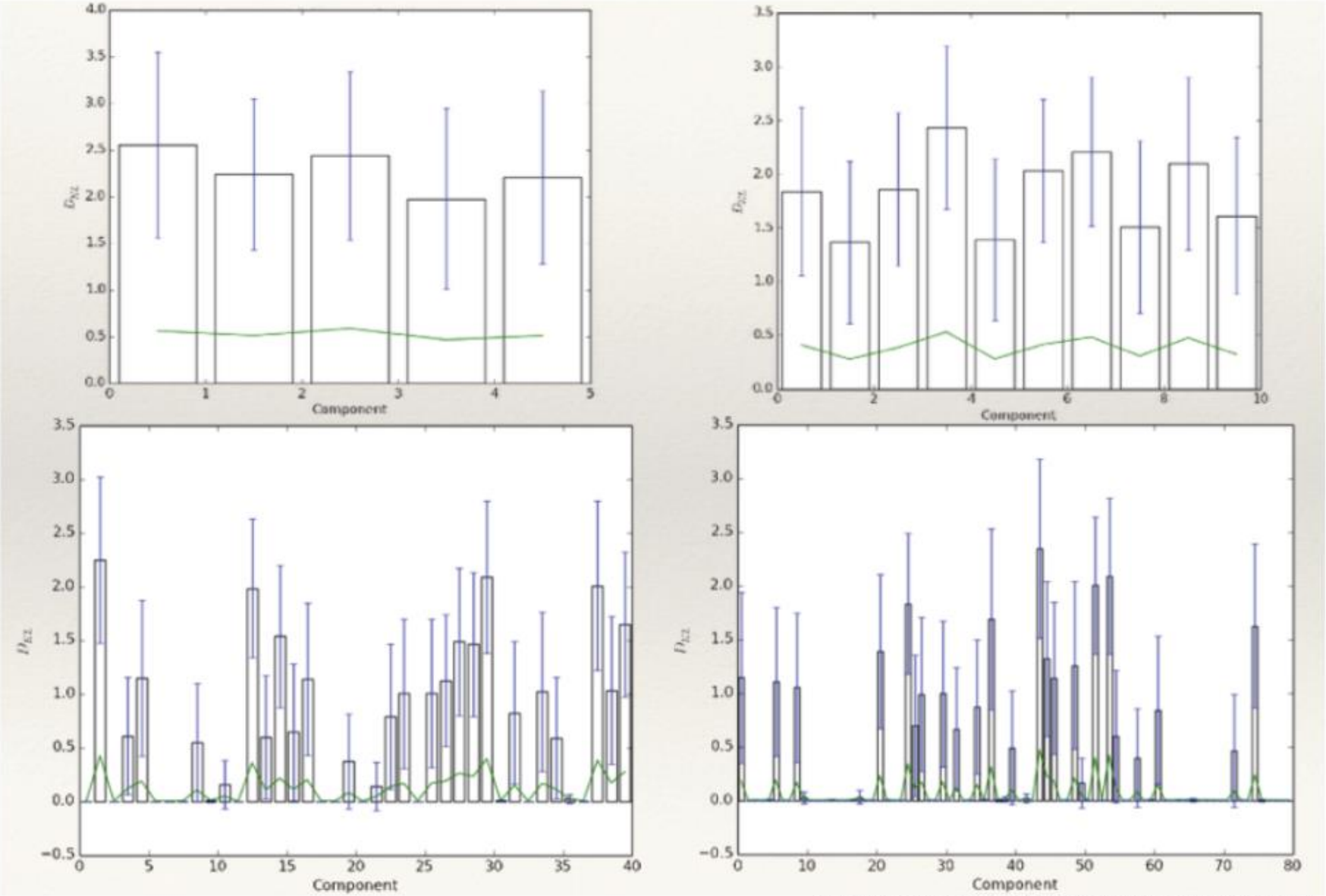


Figure from Laurent Dinh & Vincent Dumoulin

Collapsing components

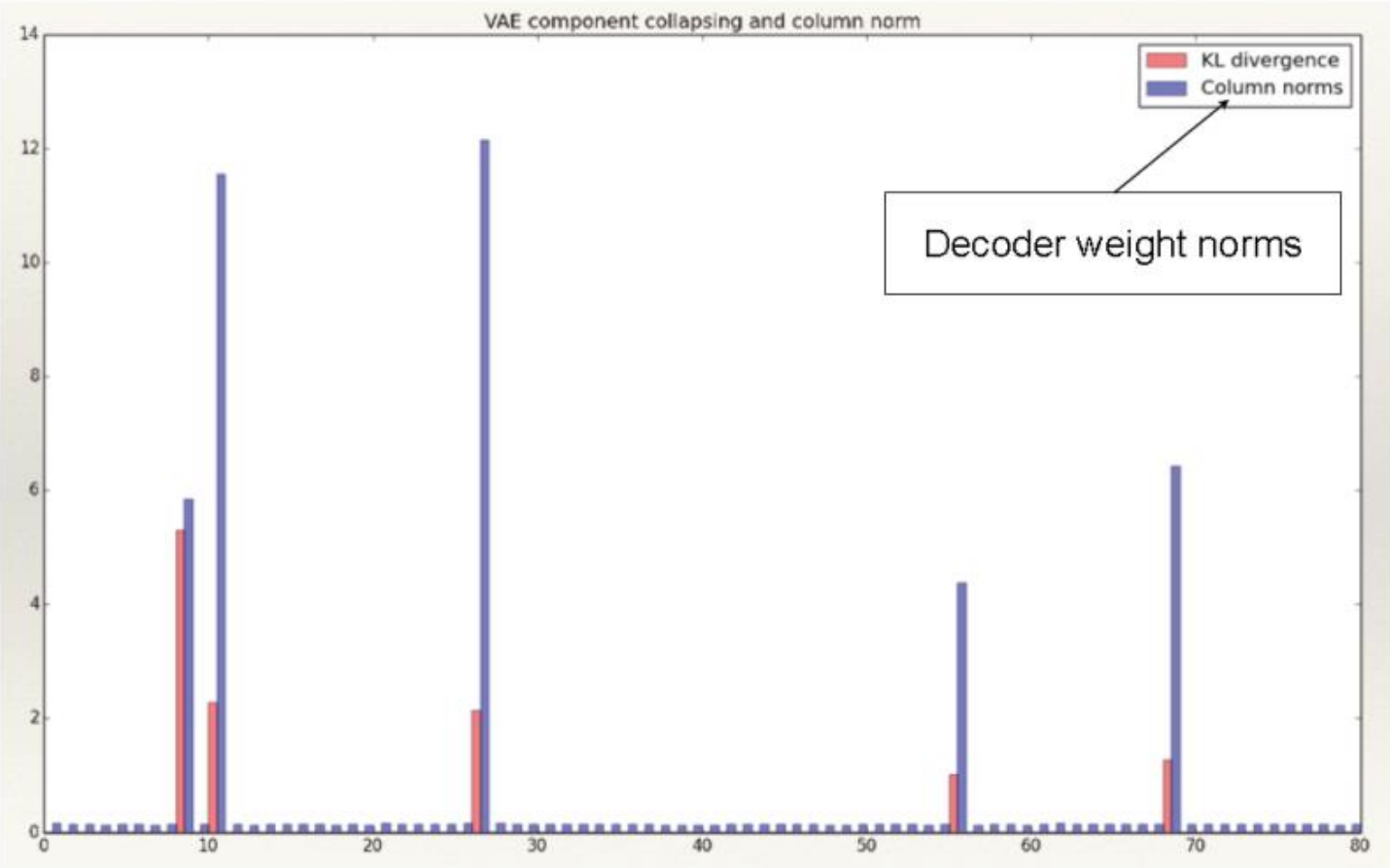


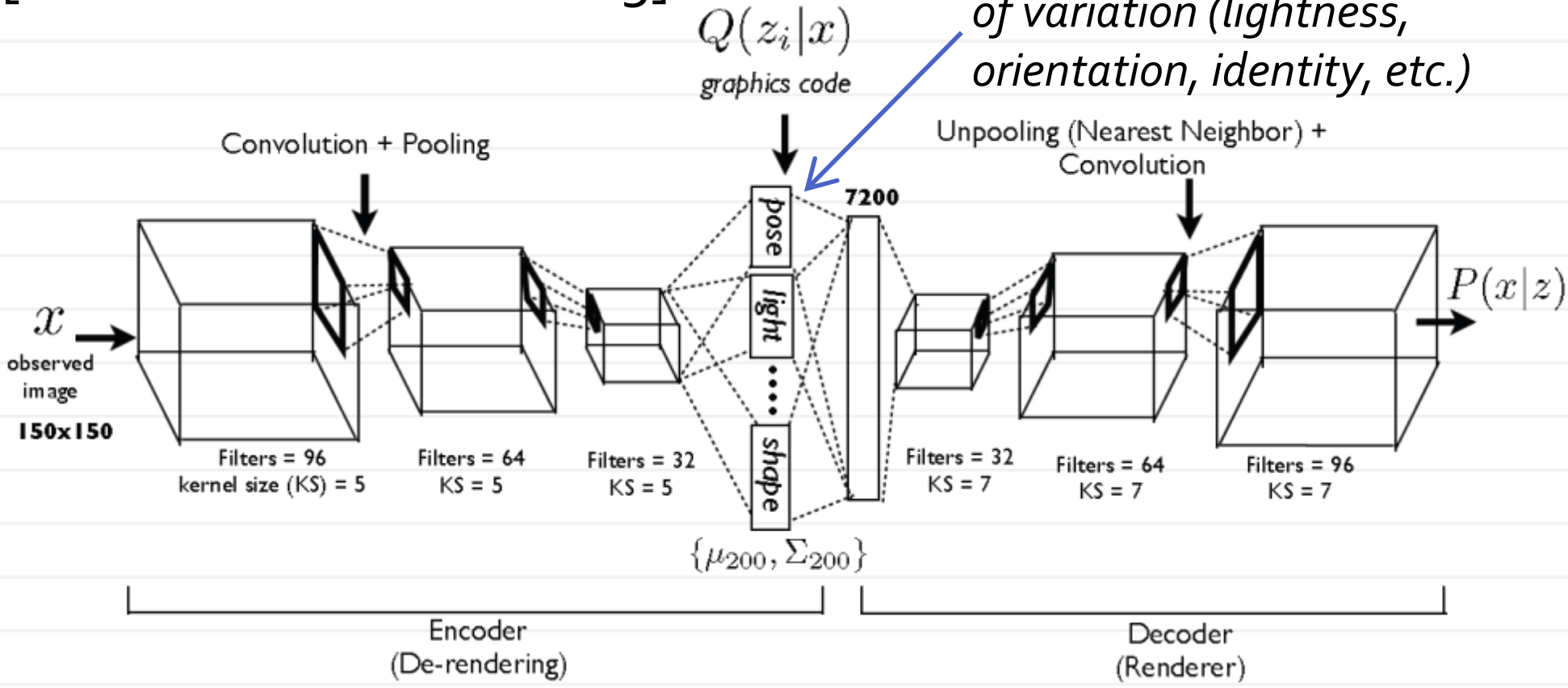
Figure from Laurent Dinh & Vincent Dumoulin

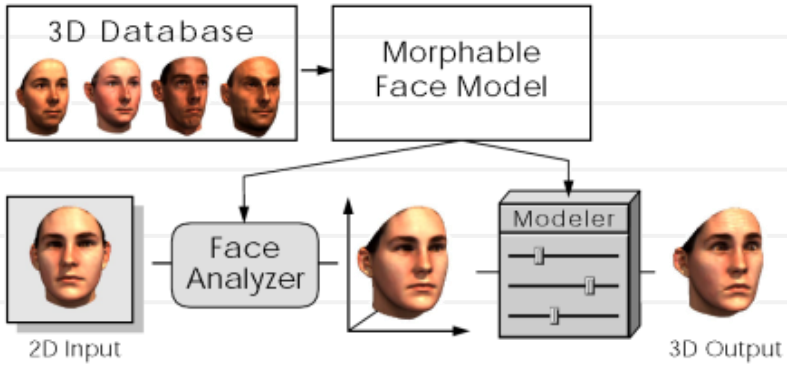
- Components that are shrunk to *N(0,1)* are not used by decoder

Deep Inverse Graphics

[Kulkarni et al. NIPS 2015]

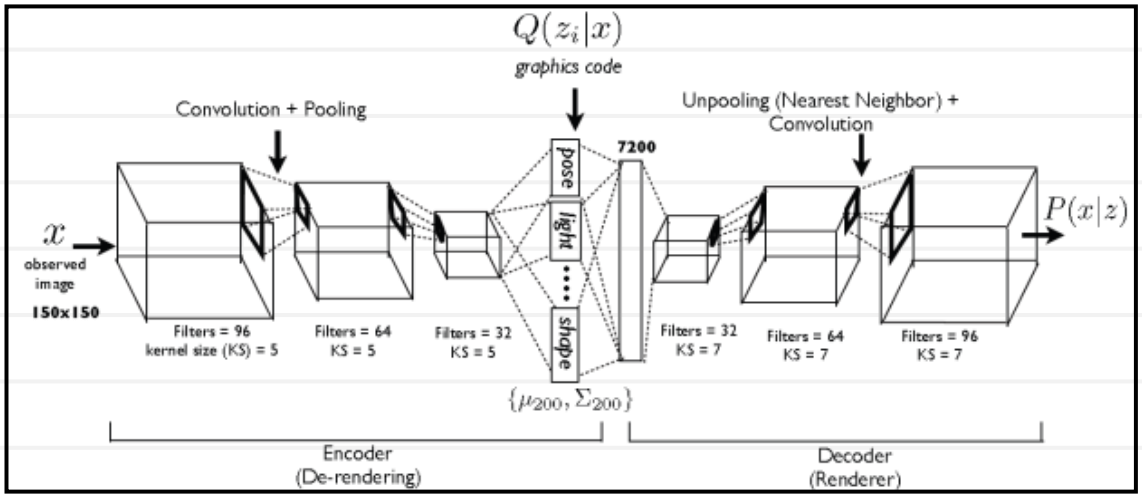
Different variables z are assigned to different factors of variation (lightness, orientation, identity, etc.)





[Blanz and Vetter 1999] morphable model

Ensuring semantic meaning of dimensions

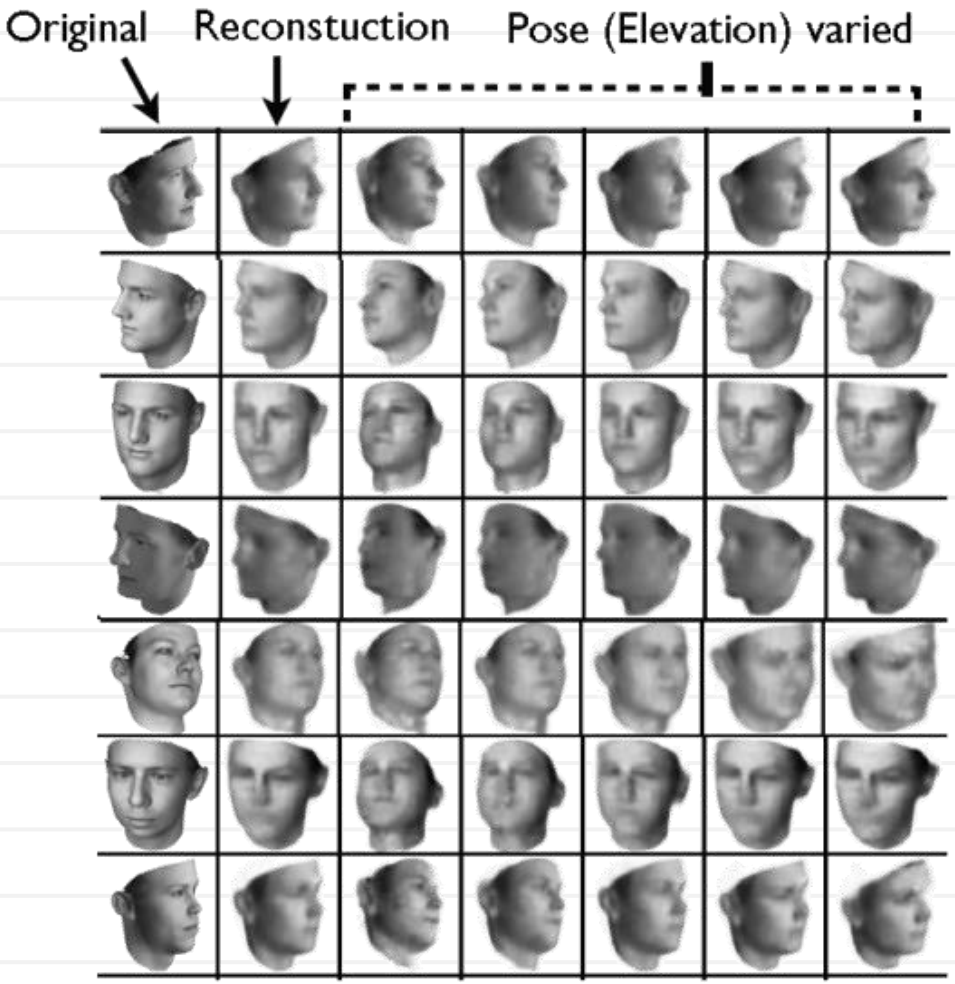


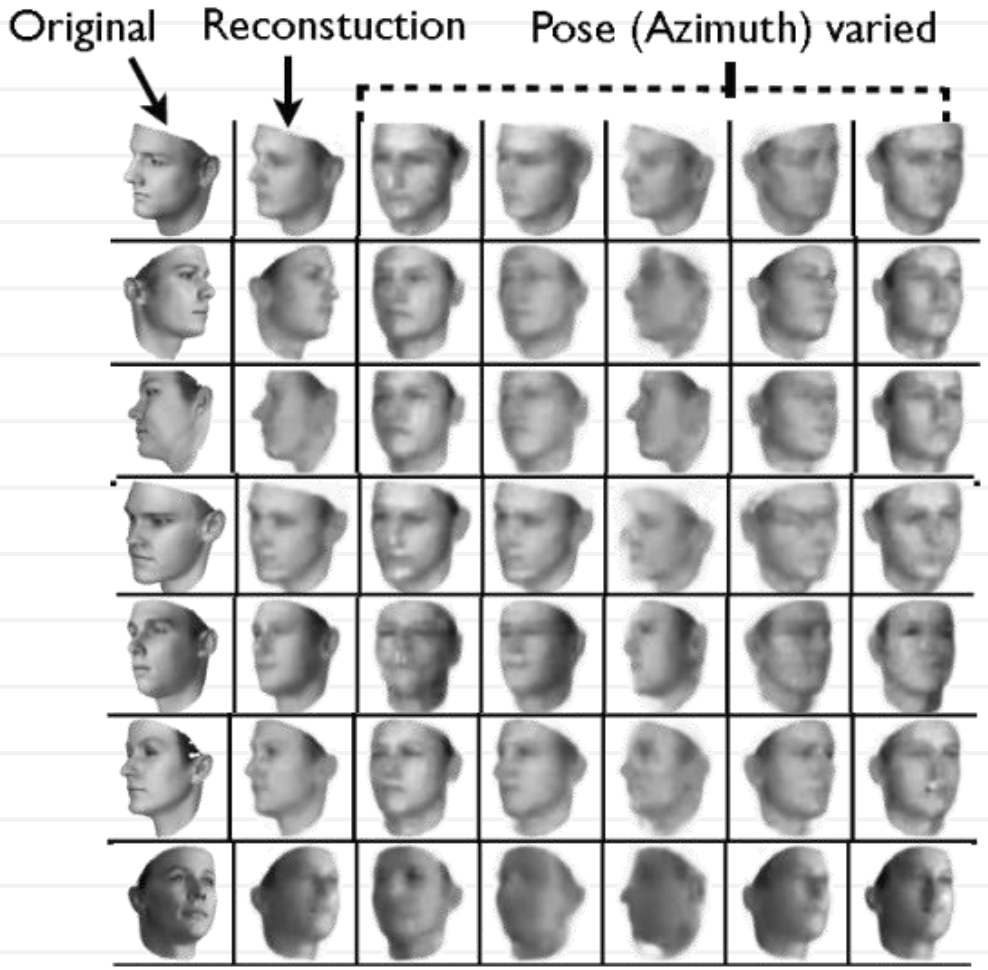
Weak supervision

Training with clamping:

- Draw minibatches with some factors fixed
- During forwardprop replace corresponding z with averages $z_j^i = \bar{z}_j$
- Add clamping error to the training loss

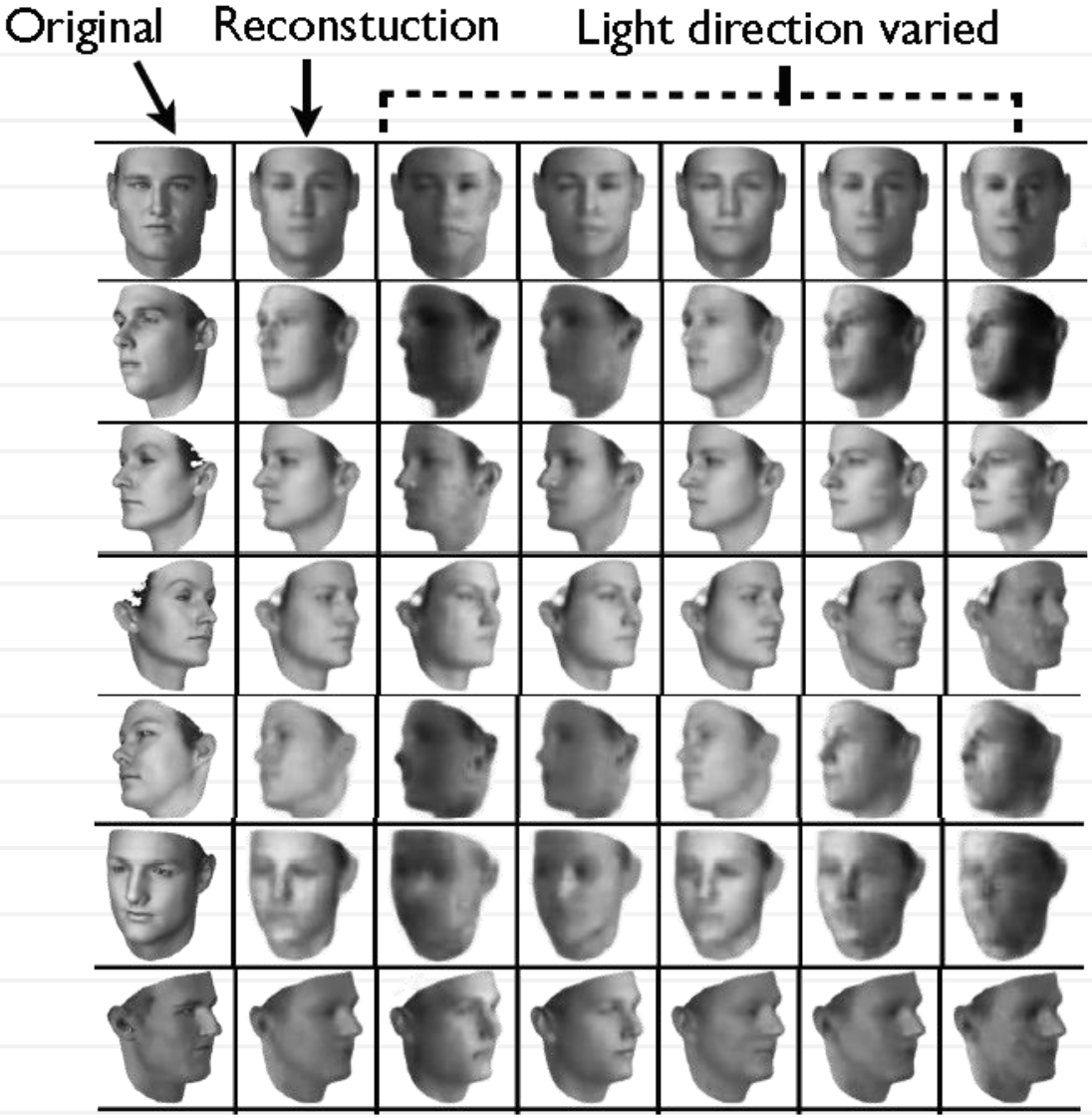
Deep Inverse Graphics





[Kulkarni et al. NIPS 2015]

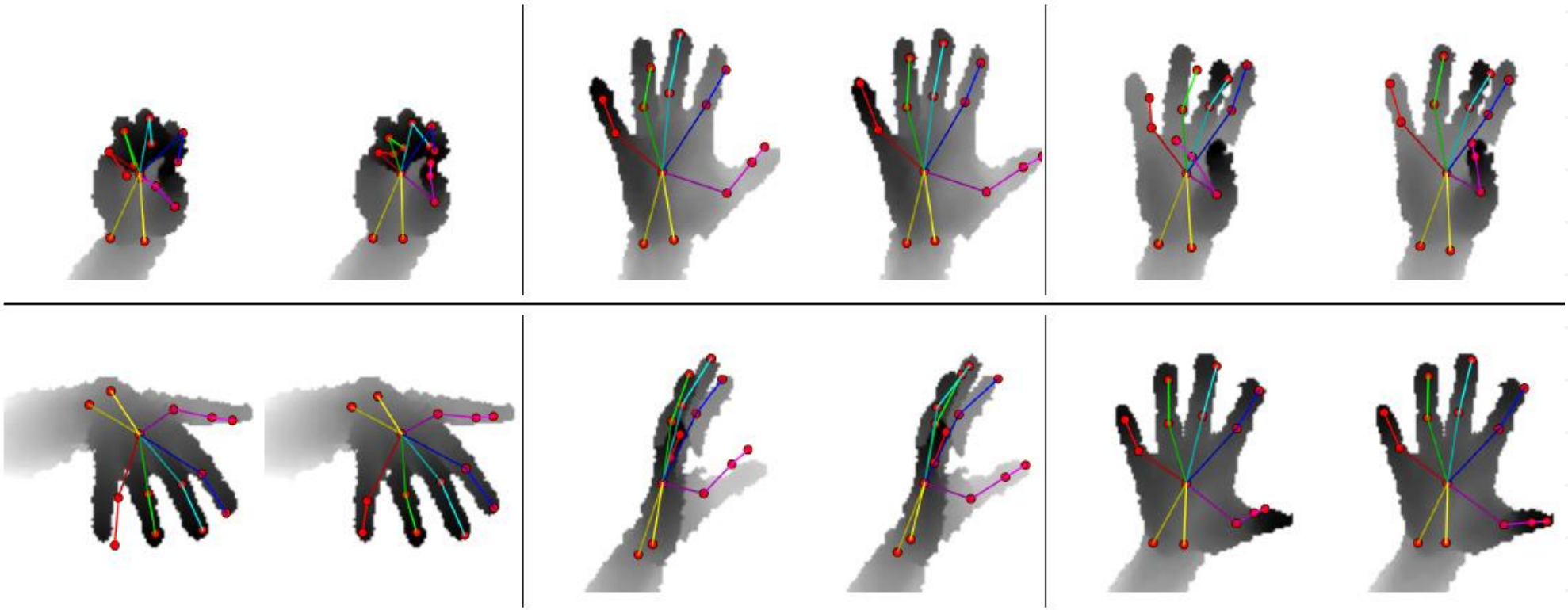
Deep Inverse Graphics



[Kulkarni et al. NIPS 2015]

Vision with Feedback loop: background

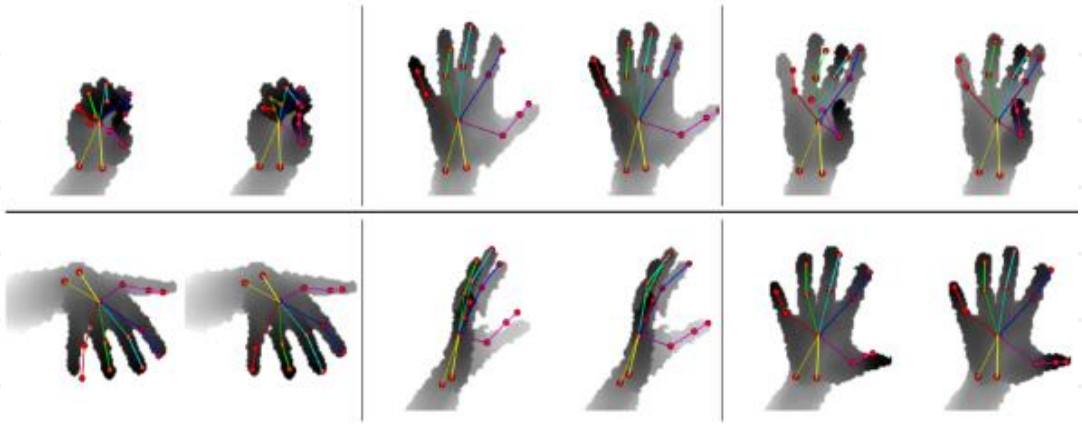
Problem setting:



- Difficult, even humans far from perfect
- Synthetic data is reasonable

[Oberweger, Wohlhart, Lepetit ICCV15]

Vision with Feedback loop : background

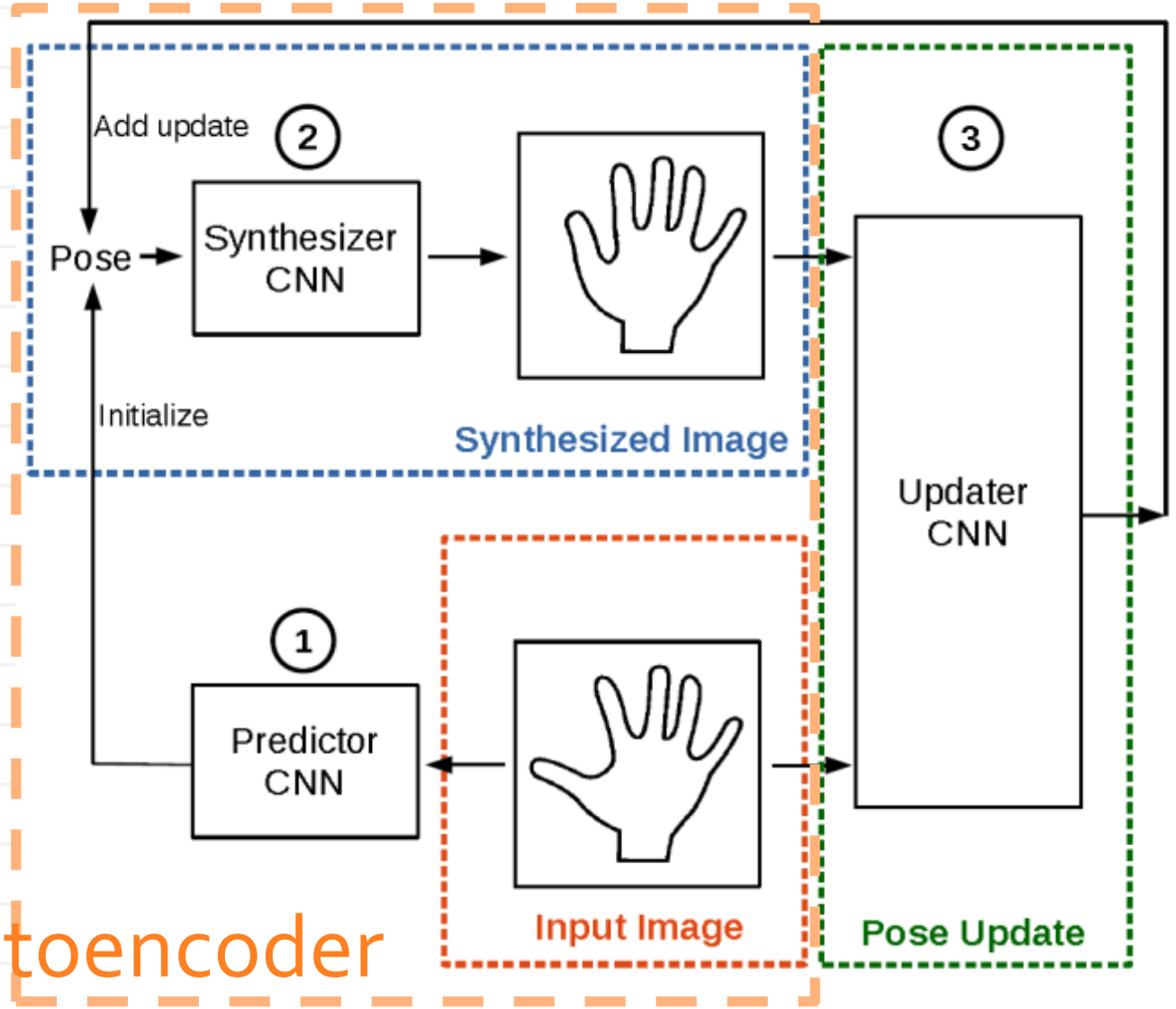


SoTA: $\|I - \text{Rendering}(p)\| \to \min_p$

"gold standard algorithm"

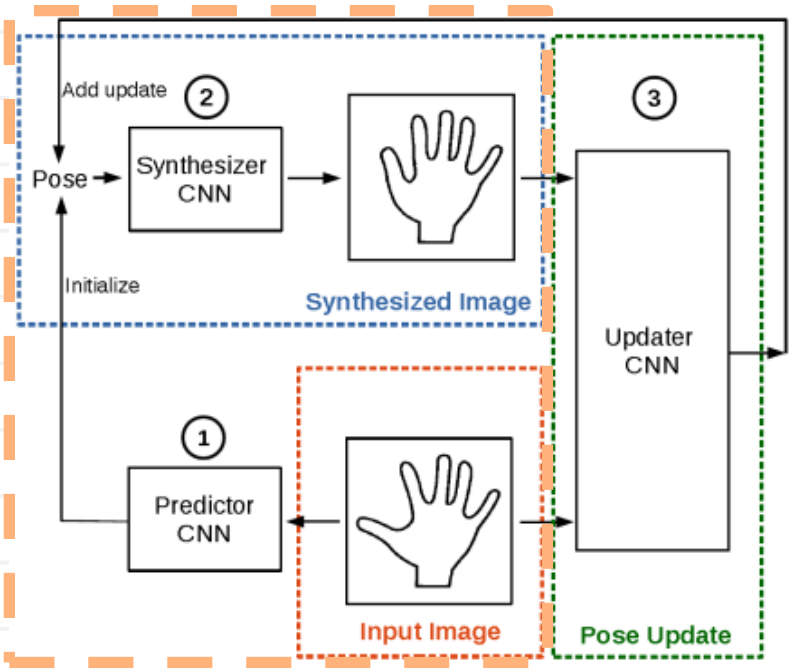
- Slow
- Get stuck in poor minima
- Needs good model
- When properly engineered, very hard to beat

Vision with Feedback loop



[Oberweger, Wohlhart, Lepetit ICCV15]

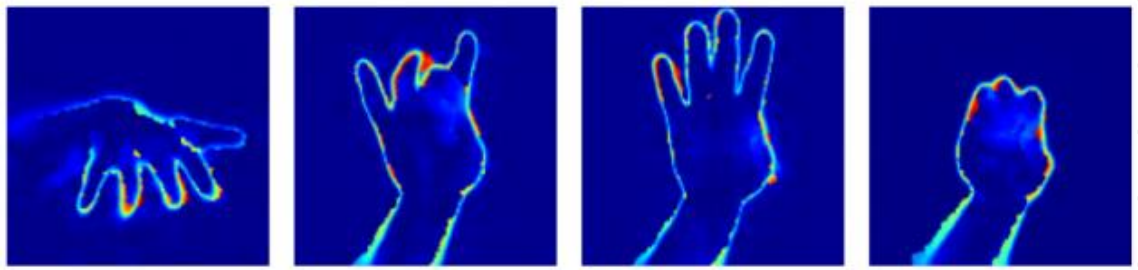
Vision with Feedback loop



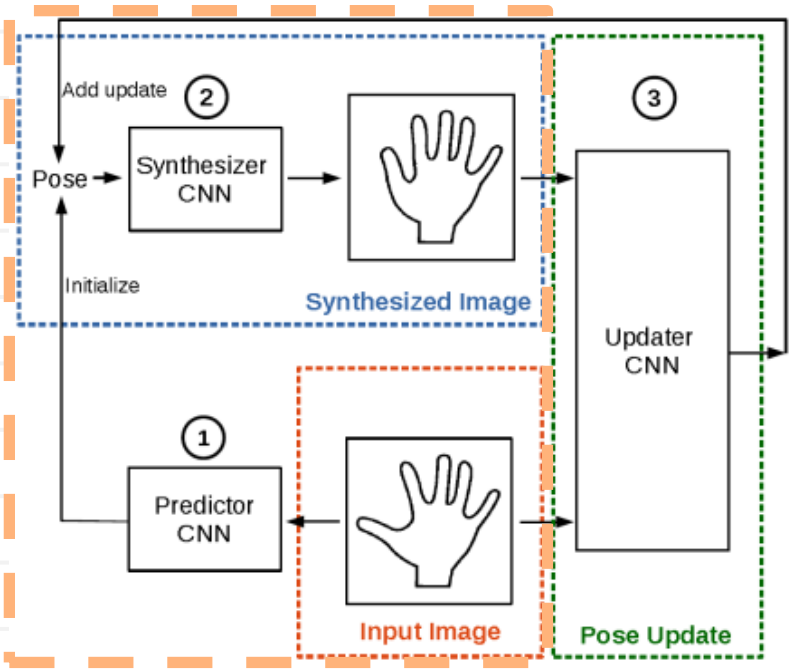
[Oberweger et al. ICCV15]

Autoencoder reconstruction:

Difference:

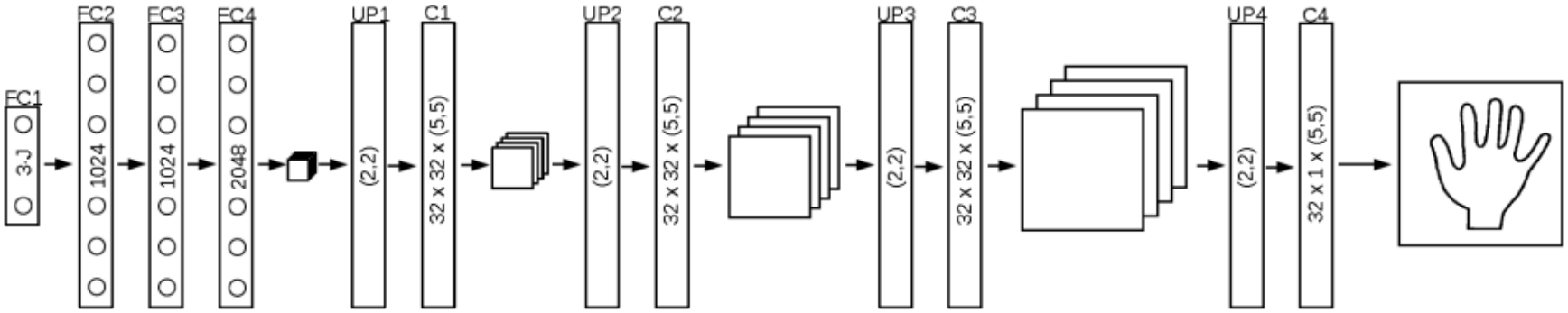


Vision with Feedback loop

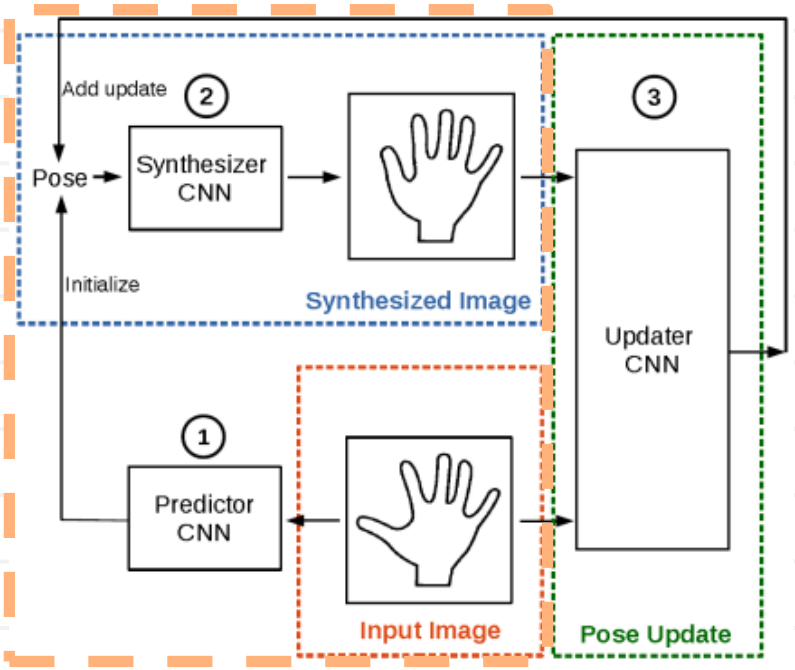


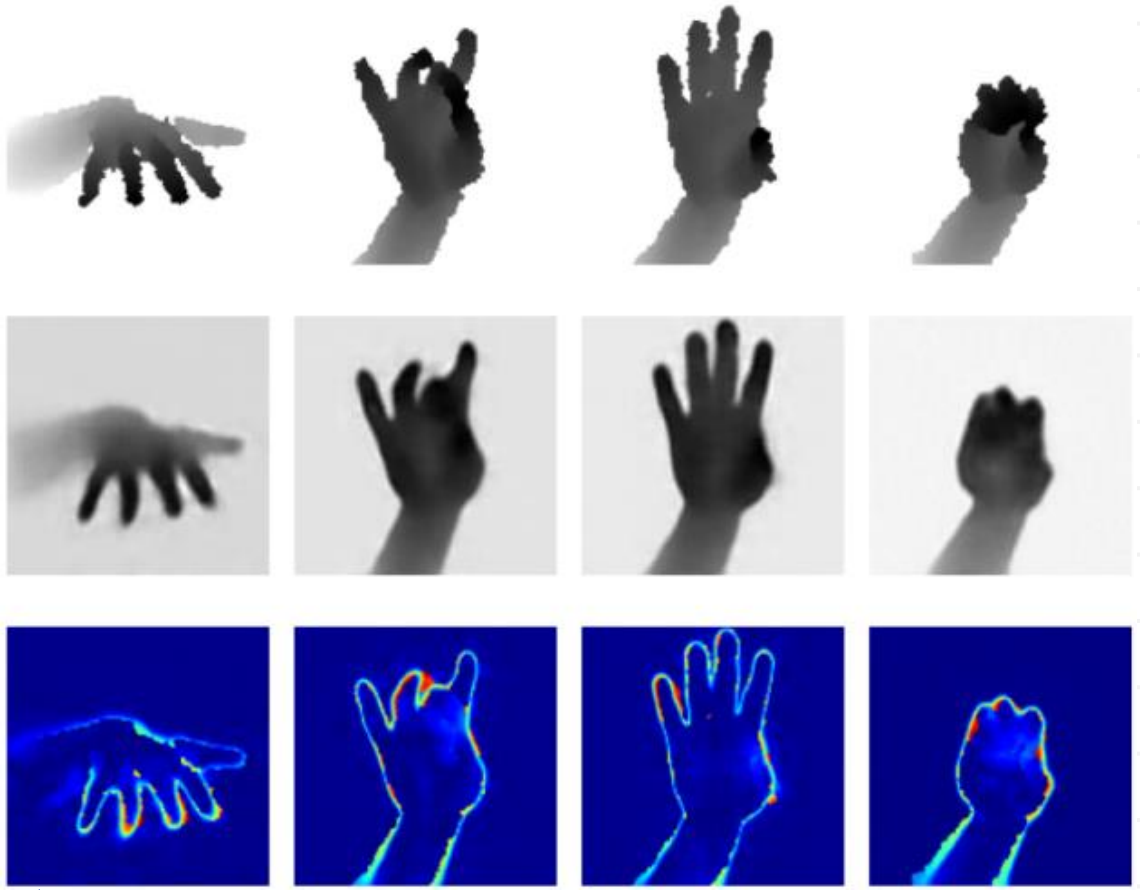
[Oberweger et al. ICCV15]

Synthesizer (decoder):



Vision with Feedback loop

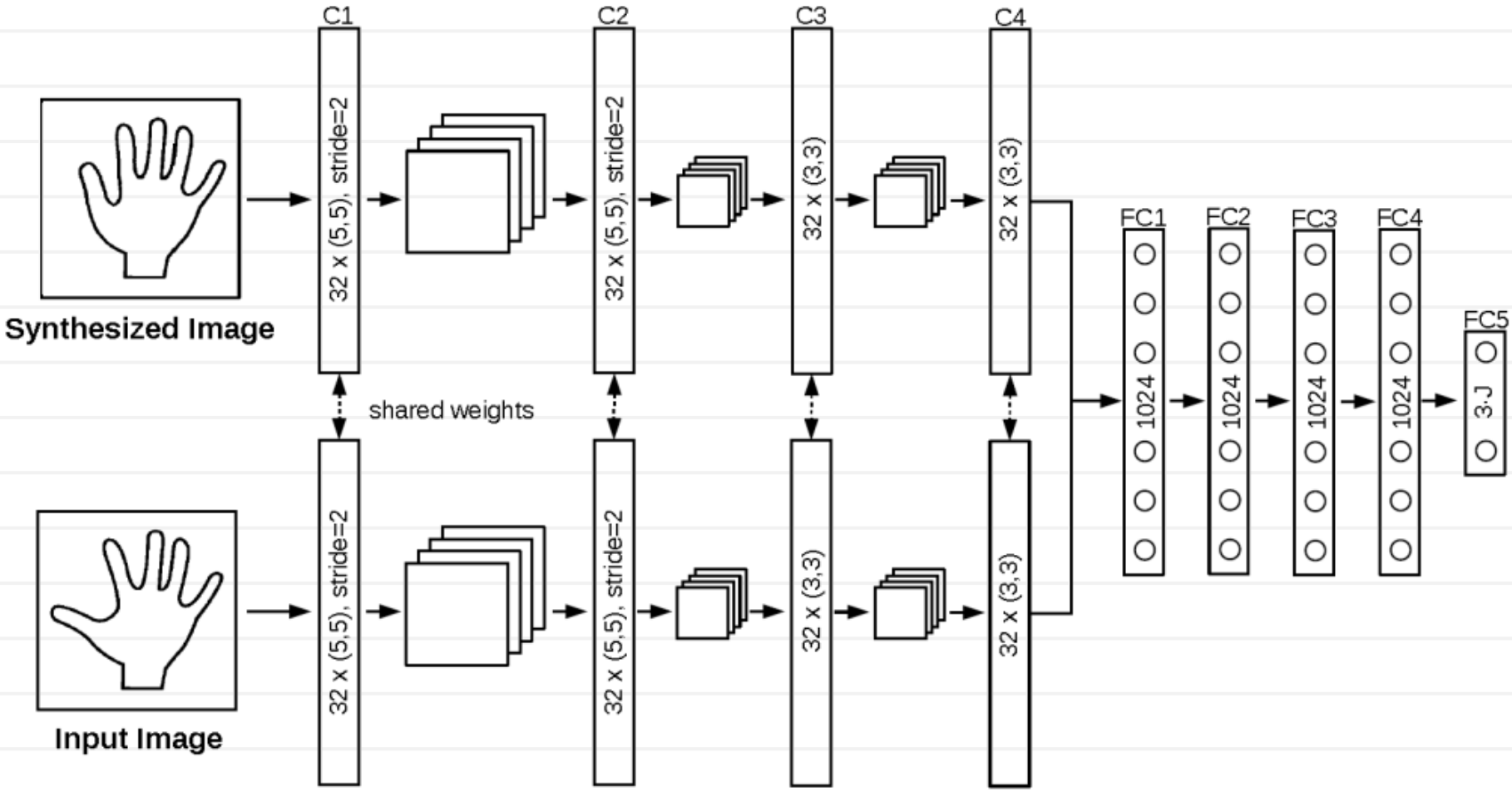




Where to improve, how to improve

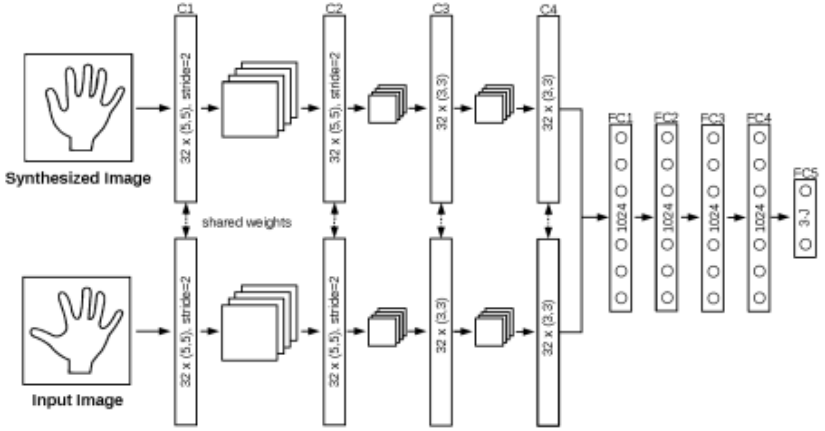
[Oberweger, Wohlhart, Lepetit ICCV15]

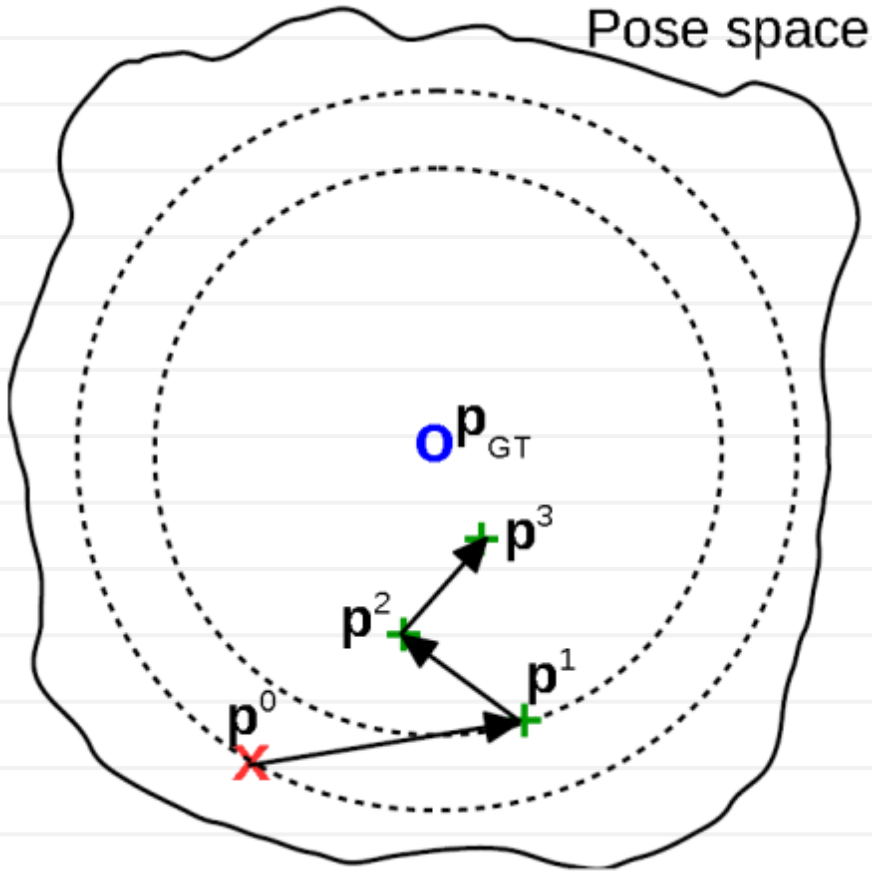
Updater network



[Oberweger, Wohlhart, Lepetit ICCV15]

Updater network



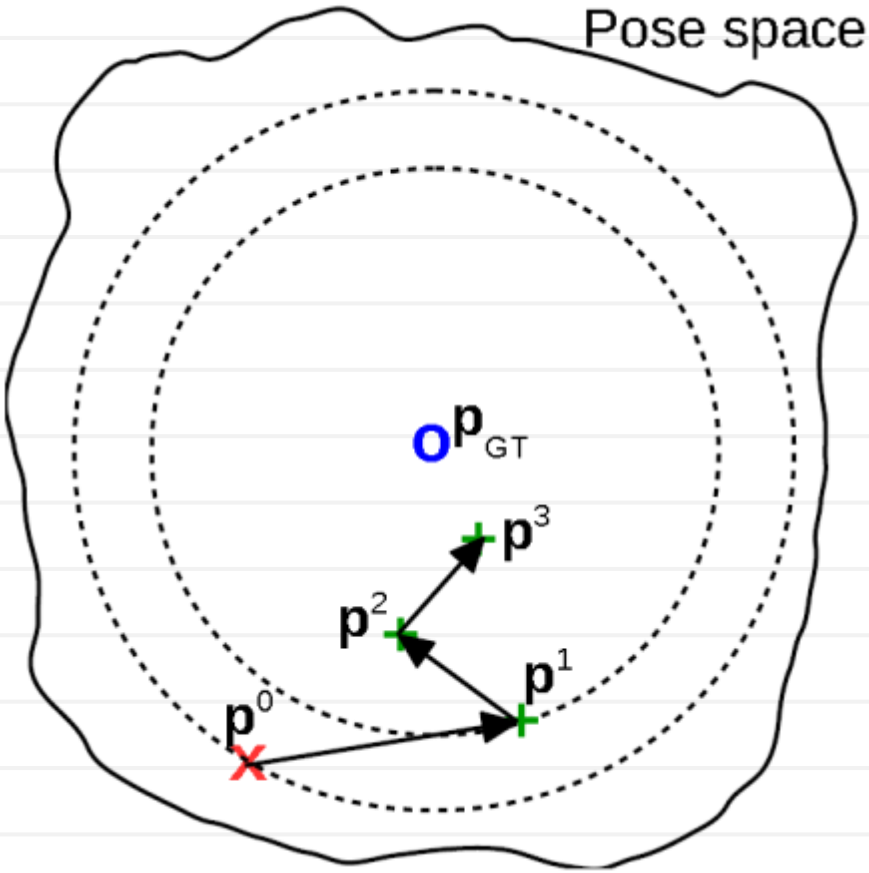


- Training with L2 loss does not work (problem too hard)
- Instead:

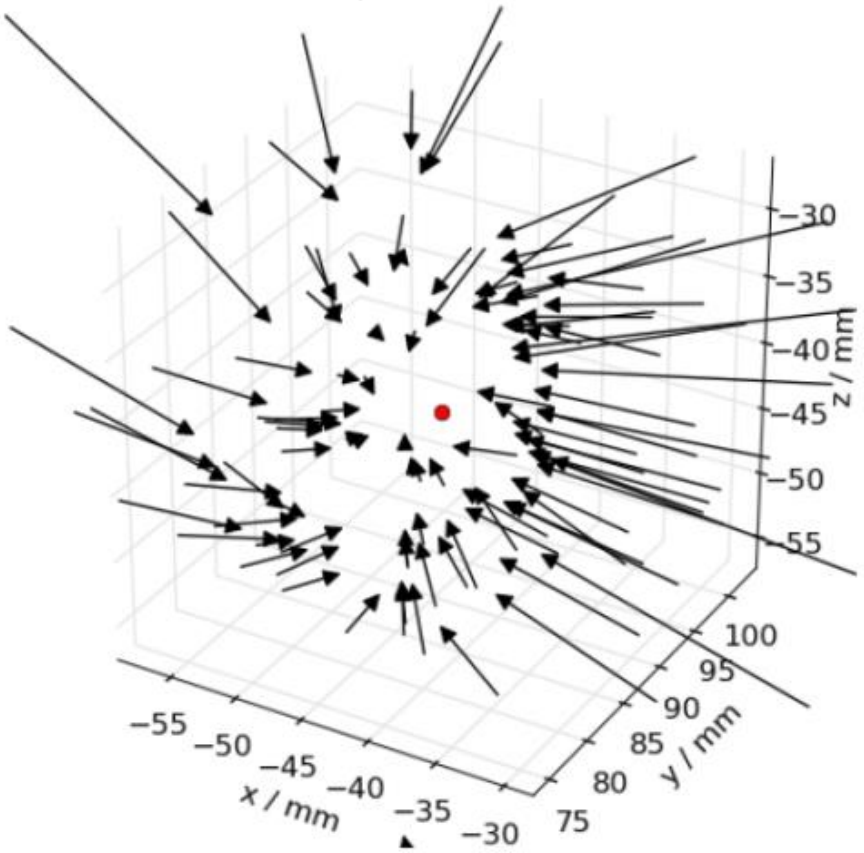
$$\mathcal{L} = \sum_{(\mathcal{D},\mathbf{p})\in\mathcal{T}} \sum_{\mathbf{p}'\in\mathcal{T}_{\mathcal{D}}} \max(0, \|\mathbf{p}''-\mathbf{p}\|_2 - \lambda\|\mathbf{p}'-\mathbf{p}\|_2)$$

[Oberweger et al.ICCV15]

Resulting updates



One of joints:

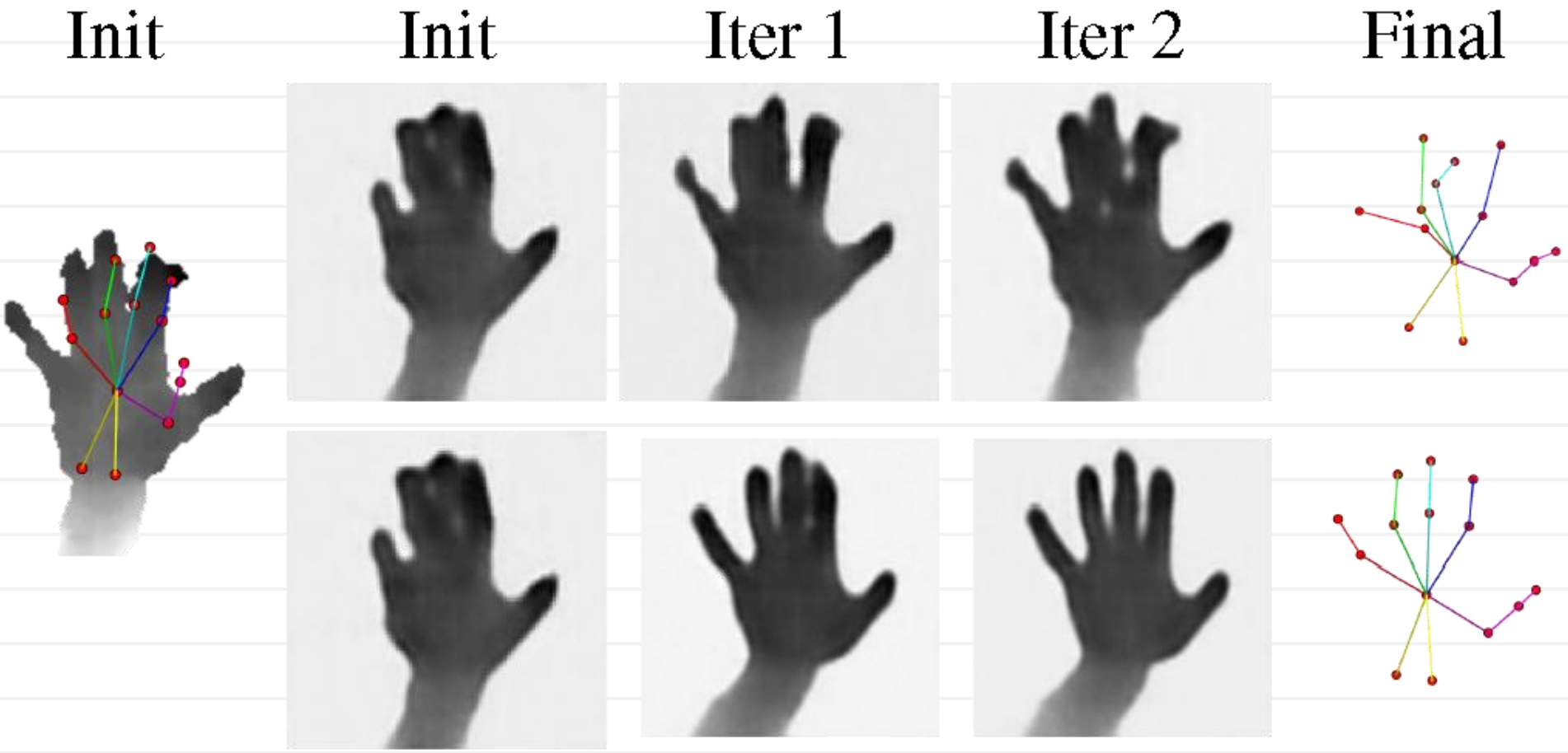


$$\mathcal{L} = \sum_{(\mathcal{D},\mathbf{p})\in\mathcal{T}} \sum_{\mathbf{p}'\in\mathcal{T}_{\mathcal{D}}} \max(0, \|\mathbf{p}''-\mathbf{p}\|_2 - \lambda\|\mathbf{p}'-\mathbf{p}\|_2)$$

[Oberweger et al.ICCV15]

Qualitative comparison

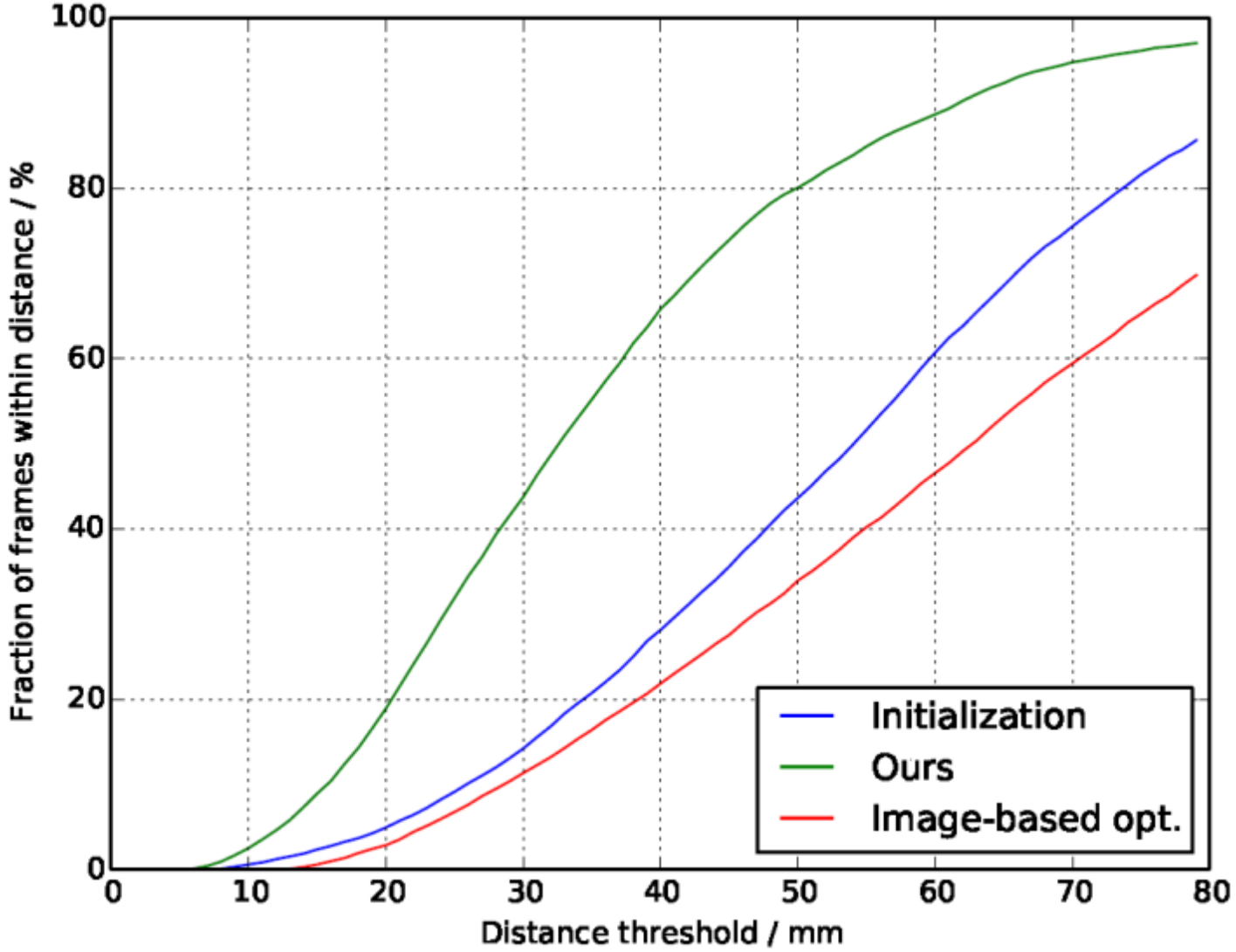
"Gold standard"



Theirs

[Oberweger, Wohlhart, Lepetit ICCV15]

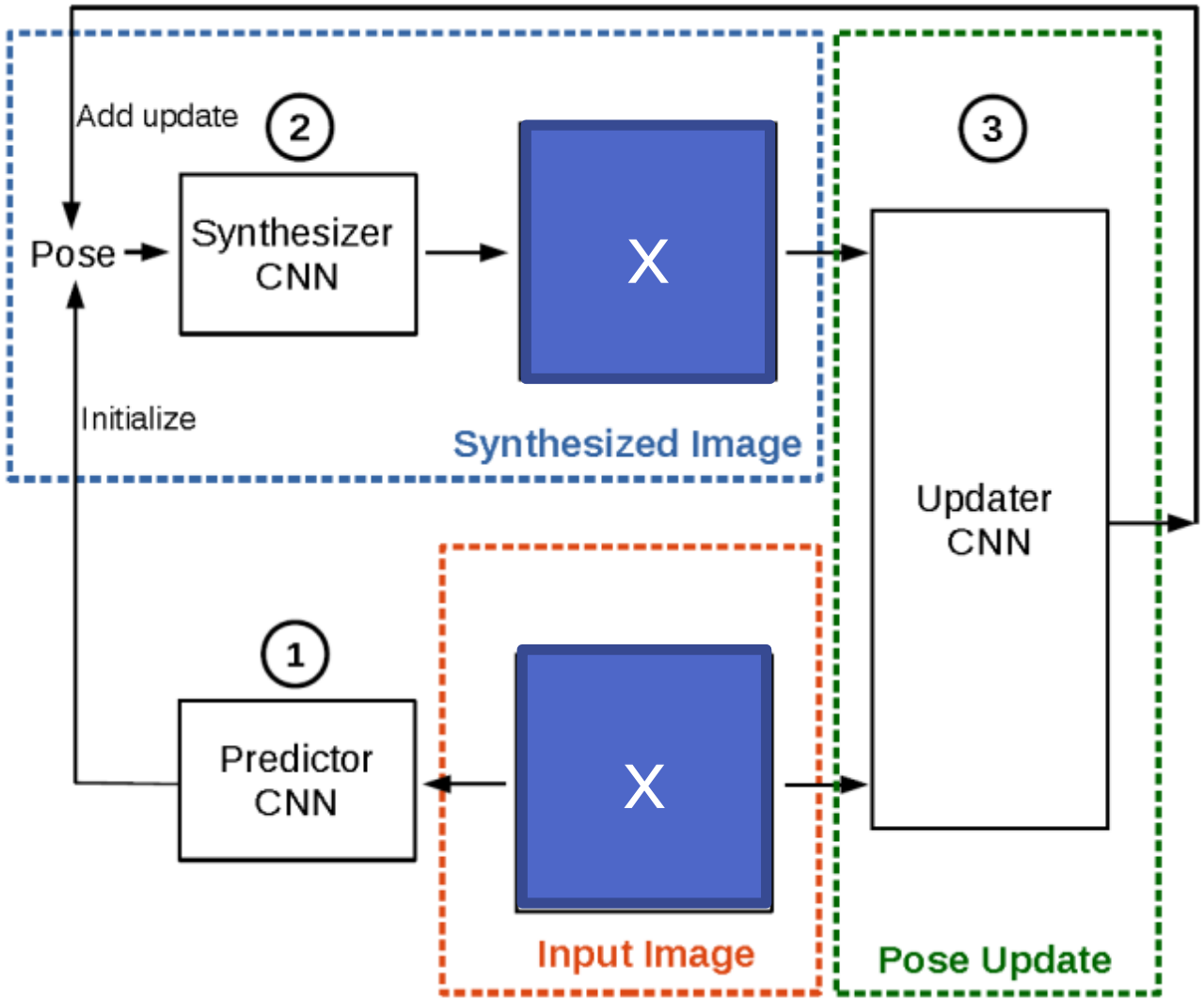
Quantitative comparison



[Oberweger, Wohlhart, Lepetit ICCV15]

Vision with Feedback loop

Very general/promising:



[Oberweger, Wohlhart, Lepetit ICCV15]

Recap

- Latent models of images are active research area (to be continued)
- Autoencoders are natural ways to tackle unsupervised learning with DL
- Natural ways to regularize: weight decay, denoising, *variational*
- Can be part of the bigger systems (c.f. last example)

Bibliography

Piotr Bojanowski, Armand Joulin, David Lopez-Paz, Arthur Szlam: Optimizing the Latent Space of Generative Networks. CoRR abs/1707.05776 (2017)

Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle: Greedy Layer-Wise Training of Deep Networks. NIPS 2006: 153-160

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol: Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. Journal of Machine Learning Research 11: 3371-3408 (2010)

Diederik P. Kingma, Max Welling:
Auto-Encoding Variational Bayes. ICLR 2014

Aaron Courville, Variational Autoencoders and Extensions, videolectures.net

Bibliography

Tejas D. Kulkarni, Will Whitney, Pushmeet Kohli, Joshua B. Tenenbaum: Deep Convolutional Inverse Graphics Network. NIPS 2015

Volker Blanz, Thomas Vetter:
A Morphable Model for the Synthesis of 3D Faces. SIGGRAPH 1999: 187-194

Markus Oberweger, Paul Wohlhart, Vincent Lepetit:
Training a Feedback Loop for Hand Pose Estimation. ICCV 2015: 3316-3324