Unsupervised Learning of 3D Structure from Images

https://goo.gl/8pGKOG

October 21, 2016

Objective

Find a statistical representation of data (images or volume data) that is generalizable to new unseen data (new views of an object).

We want to infer a 3D representation (polyhedral mesh or dense volumes of voxels) from 2D images that we can render new instances of and reason about.

Other approaches rely heavily on visual feature engineering, the paper’s approach is to learn to infer 3D representation directly from 2D images.

Applications

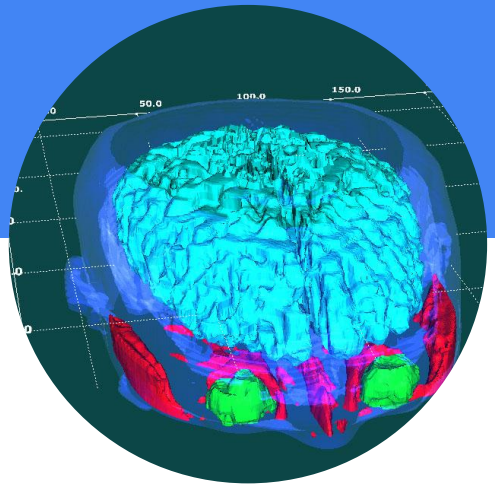
3D representations allows easier downstream properties of objects to use for:

- Physical reasoning of objects including interaction, and navigation,
- Scene completion,
- Denoising,
- Compression, and
- Generative virtual reality

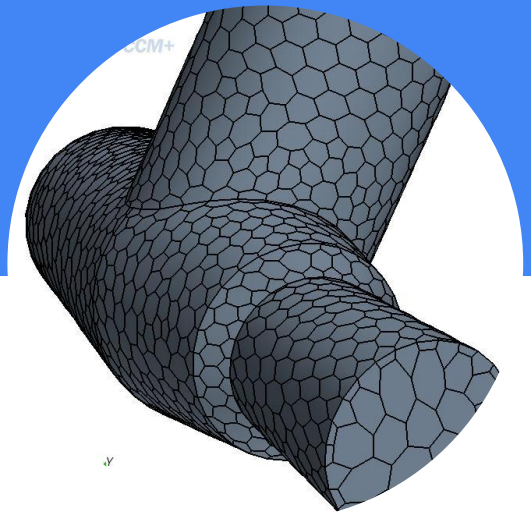
Challenges

1. Inherently ill-posed: an infinite number of possible 3D structures to give rise to a particular 2D observation.
    - Learn statistical models to find most likely representation
2. Inference is intractable: mapping image pixels to 3D representations, handling multi-modality of representations
3. Unclear how to best represent 3D structures
    - Polyhedral mesh and dense volume of voxels are used in the paper

3D Representations



Dense Volume of Voxels



Polyhedral Mesh

Model

Conditional Generative Model

Given observed volume or image **x** and a context **c**, infer 3D representation **h**, and render 2D image from either a Neural Net or OpenGL engine.

Context c is either: nothing, object class label, or one or more views of a scene.

The generative models with latent variables describe probability densities p(**x**) though marginalization of the set of latent variables **z**.

$$p\left(\mathbf{x}\right) = \int p_\theta\left(\mathbf{x} \mid \mathbf{z}\right) p\left(\mathbf{z}\right) d\mathbf{z}$$

Bound Marginal likelihood p(x) by

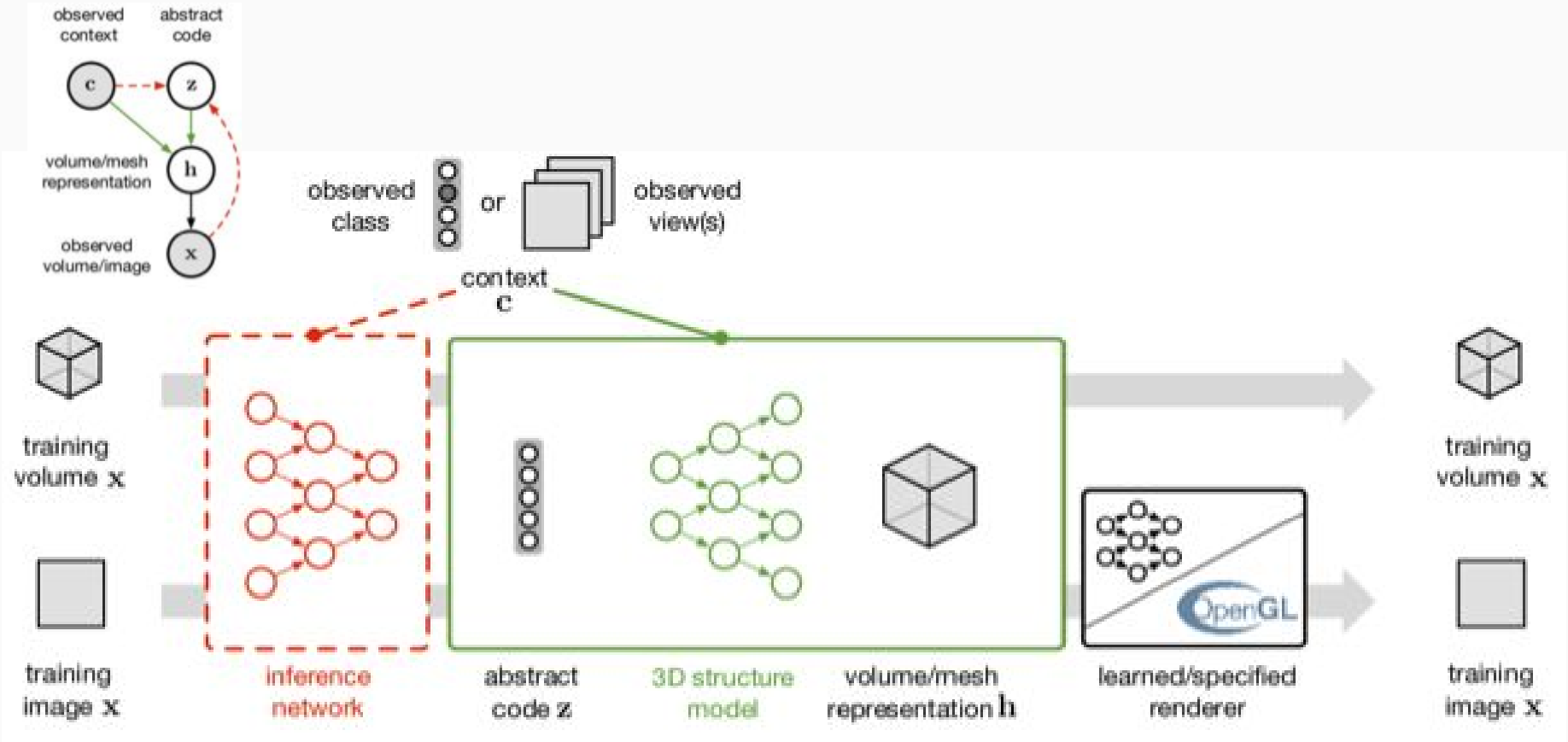
$$\mathcal{F} = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log p_\theta\left(\mathbf{x} \mid \mathbf{z}\right)\right] - KL\left[q_\phi\left(\mathbf{z} \mid \mathbf{x}\right) \parallel p\left(\mathbf{z}\right)\right]$$

Architectures

Applied recent work on sequential generative models (similar to RNN) to capture complex distributions of 3D structures.

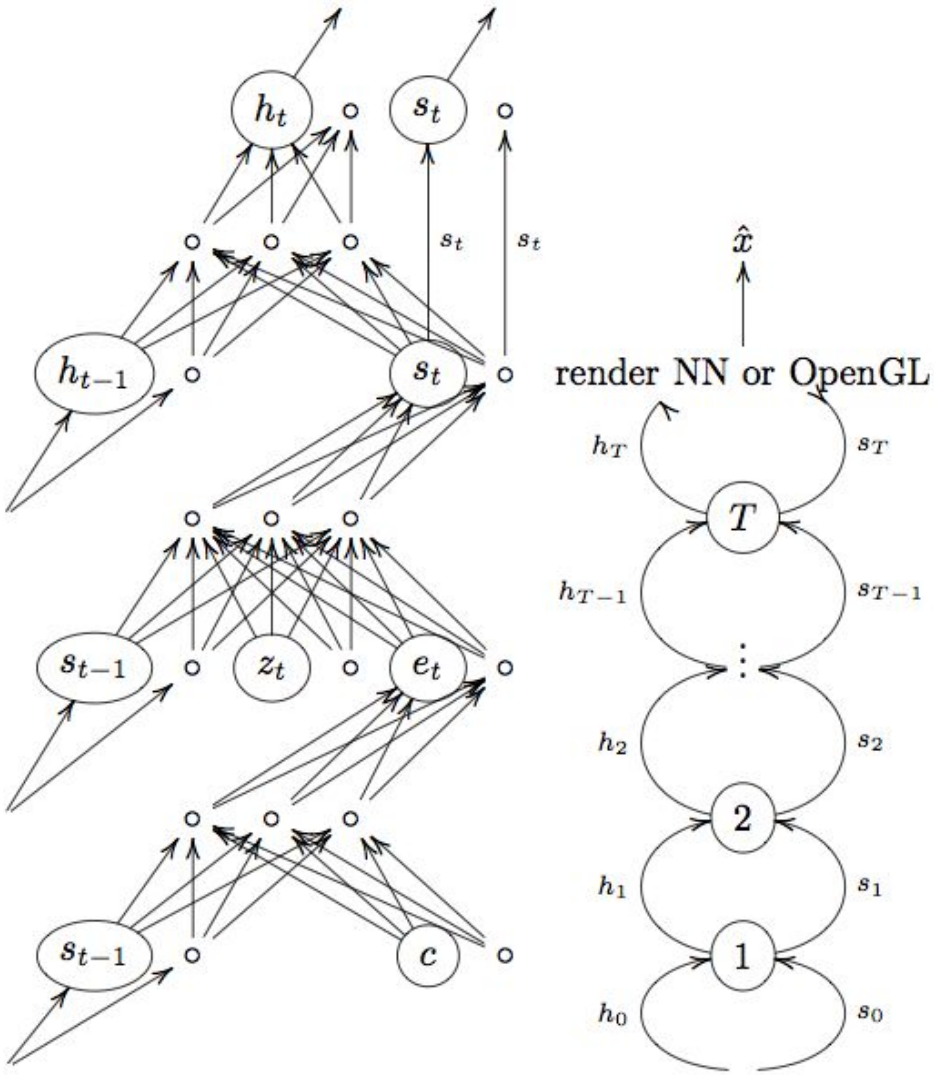
Sequentially transform independent Gaussian latent variables into refinements of h (the “canvas”).

Architectures

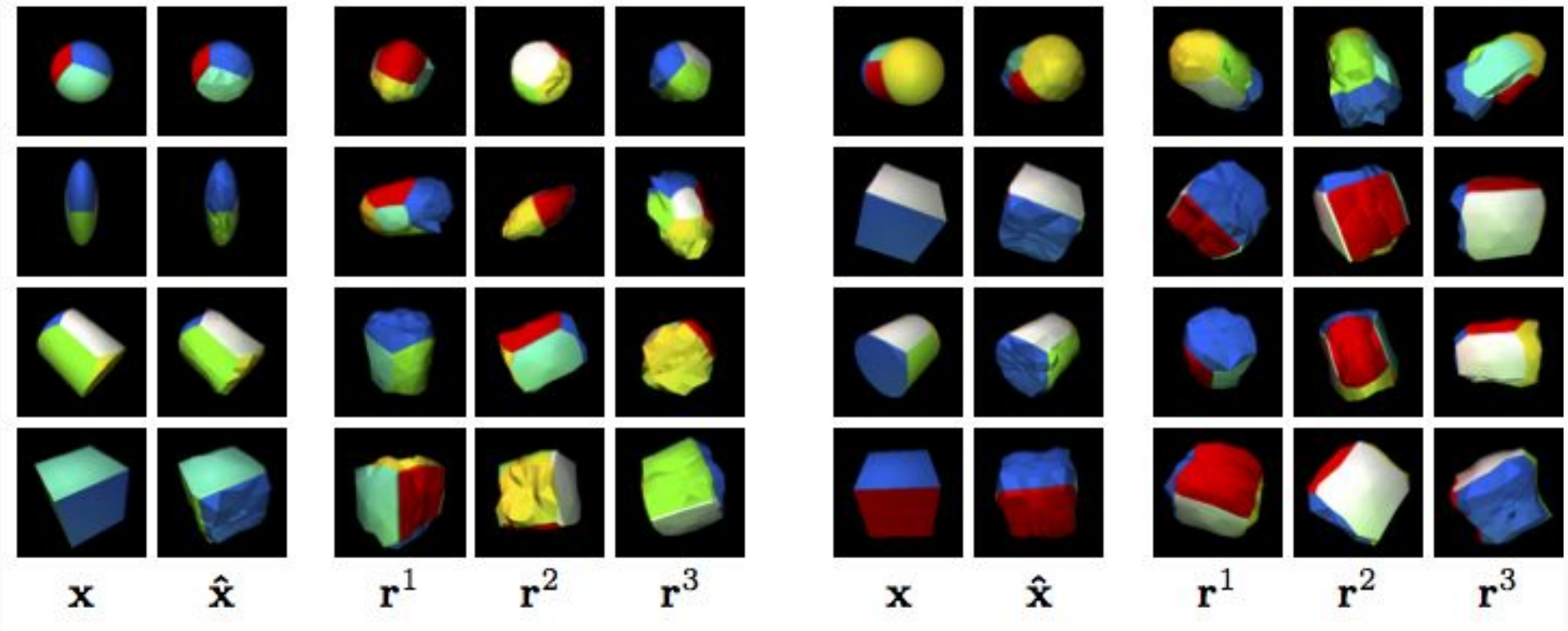


Neural Framework

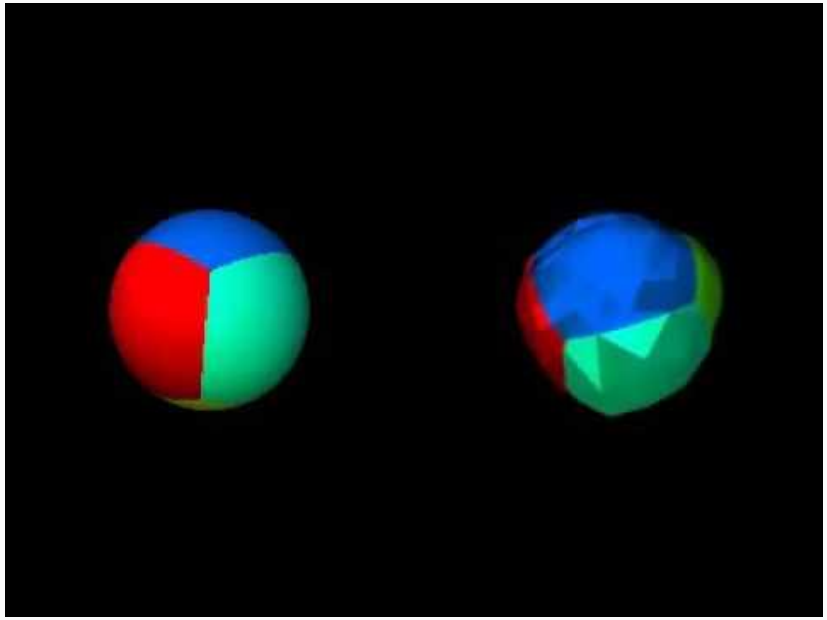
class or image	c
Latents	$z_t \sim N(0,1)$
Encoding	$\mathbf{e_t} = f_{\text{read}}(\mathbf{c}, \mathbf{s_{t-1}}; \theta_r)$
Hidden State	$\mathbf{s_t} = f_{\text{state}}(\mathbf{s_{t-1}}, \mathbf{z_t}, \mathbf{e_t}; \theta_s)$
3D representation	$\mathbf{h_t} = f_{\text{write}}(\mathbf{s_t}, \mathbf{h_{t-1}}; \theta_w)$
2D projection	$\hat{\mathbf{x}} = \text{Proj}(\mathbf{h_t}, \mathbf{s_t}; \theta_p)$
Observation	$\mathbf{x} \sim p(\mathbf{x} \mid \hat{\mathbf{x}})$
	$\theta = \{\theta_r, \theta_s, \theta_w, \theta_p\}$

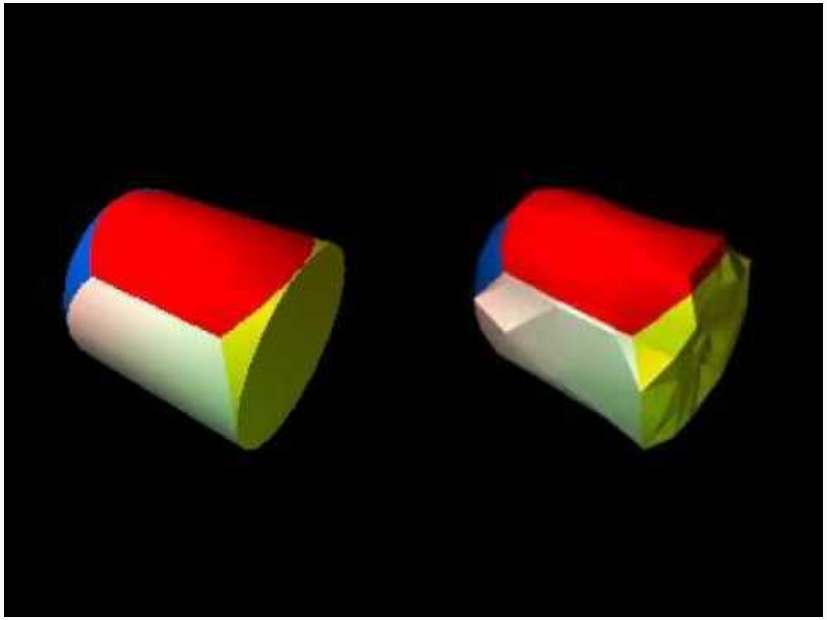


Results - Mesh

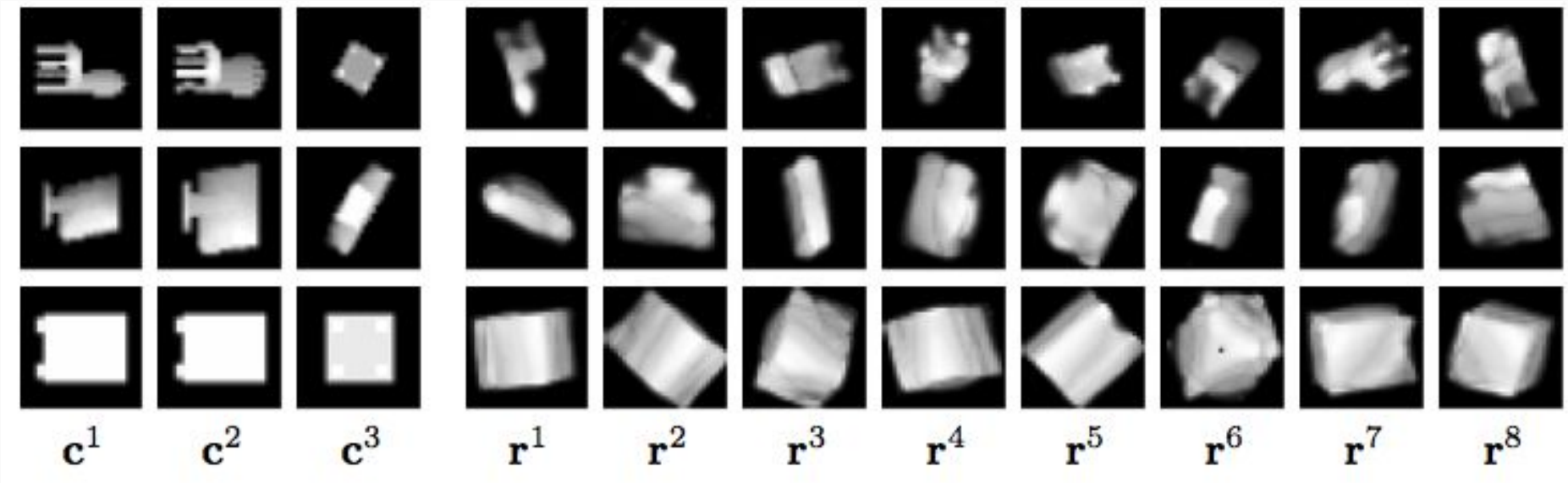


Results - Mesh

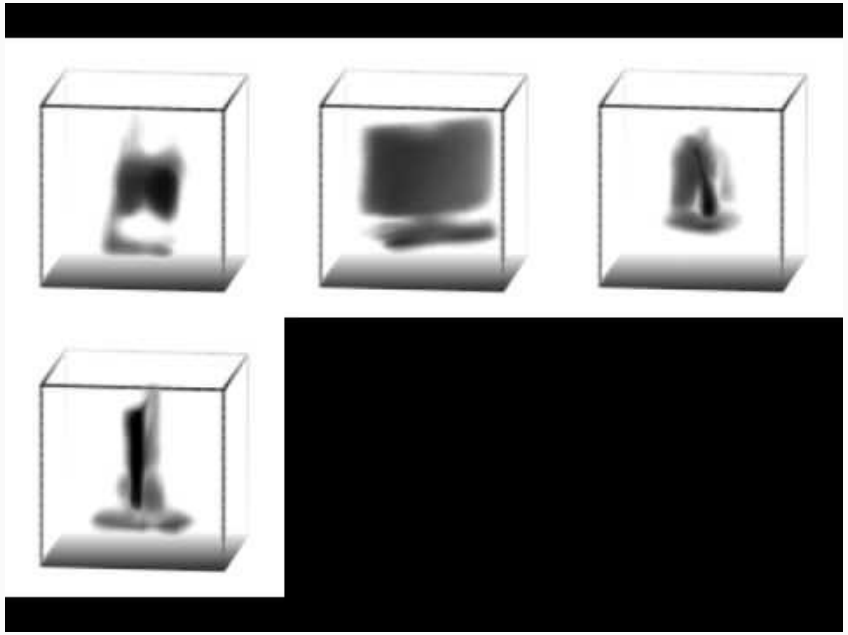


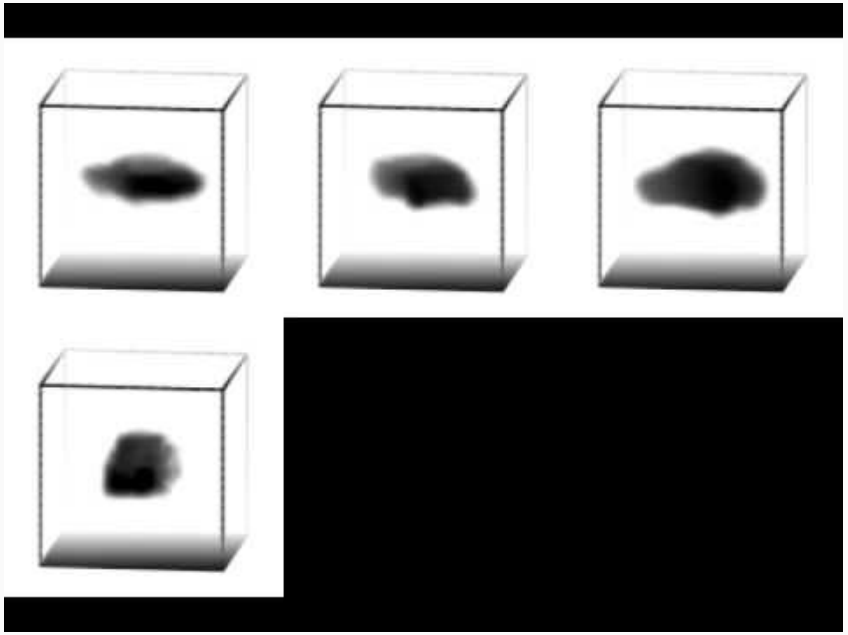


Results - Volume



Results - Volume





Conclusion

The advantage of forcing inference into a specific format is that the NN can be chained together to perform various tasks.

Questions?