Learning Shape Completion from Bounding Boxes with CAD Shape Priors

David Stutz

Master Thesis
September 15th 2017

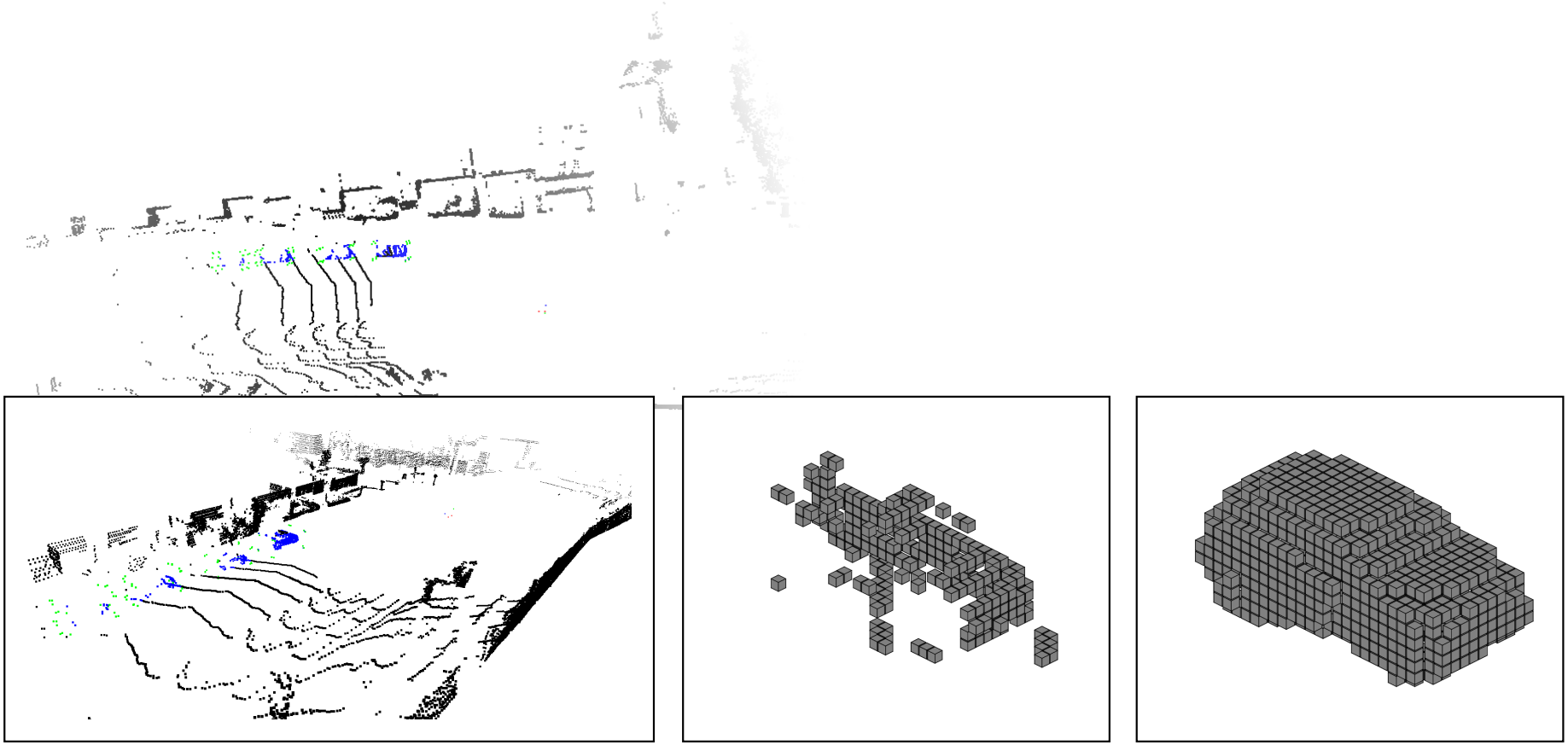
Outline

- Problem
- Related Work
- Proposed Approach
  - Shape Prior
  - Shape Inference
- Experiments
- Conclusion

Problem

Problem Overview

Shape completion from point clouds.



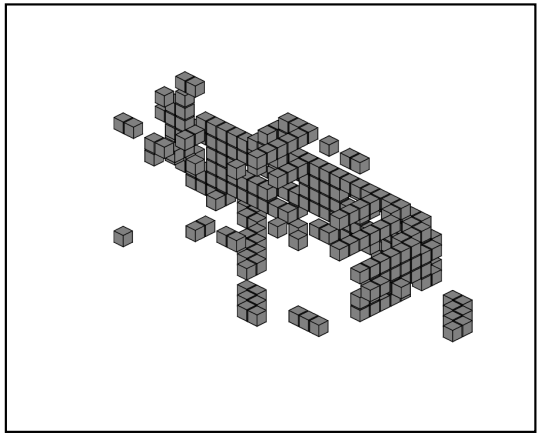
Voxelized Observation

Voxelized Shape

Problem Definition

Problem

Given observations $\mathcal{X} = \{x_1, \ldots, x_N\}$ *and reference shapes* $\mathcal{Y} = \{y_1, \ldots, y_M\}$*, learn a mapping* $x_n \mapsto y(x_n)$ *such that* $y(x_n)$ *fits the unknown target shape* y_n^**.*

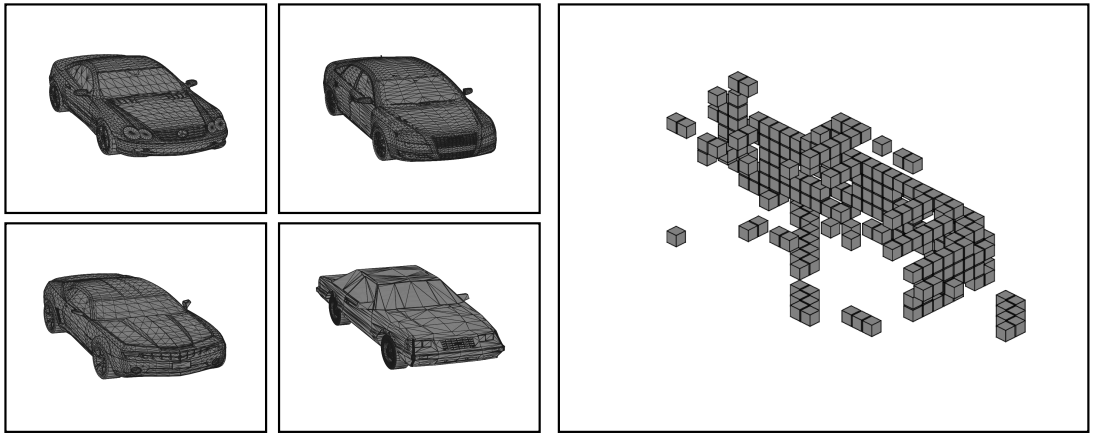


Observation x_n

Problem Definition

Problem

Given observations $\mathcal{X} = \{x_1, \ldots, x_N\}$ *and reference shapes* $\mathcal{Y} = \{y_1, \ldots, y_M\}$*, learn a mapping* $x_n \mapsto y(x_n)$ *such that* $y(x_n)$ *fits the unknown target shape* y_n^**.*

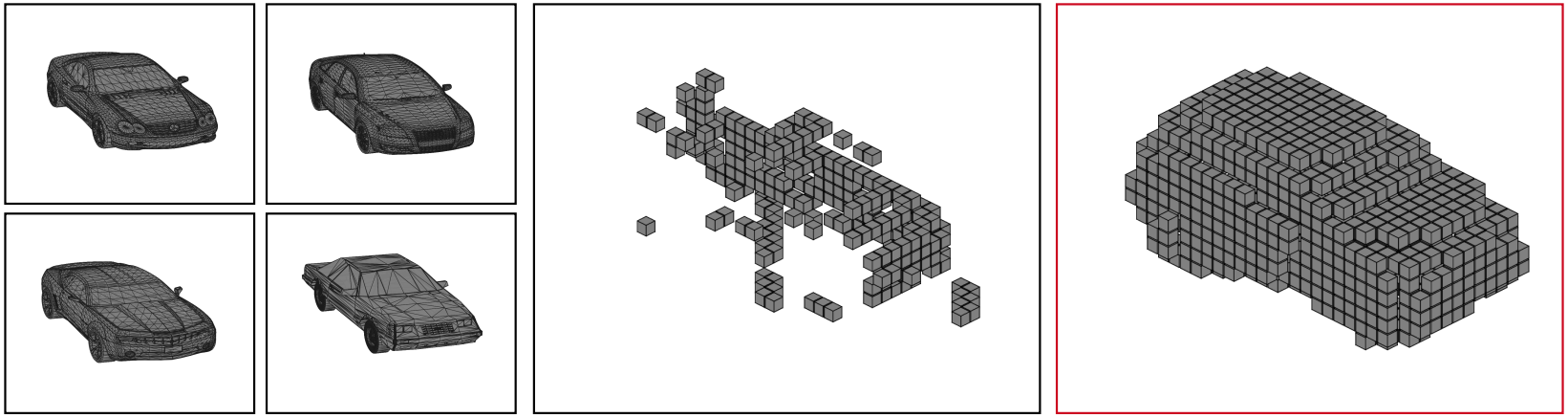


Reference Shapes y_m Observation x_n

Problem Definition

Problem

Given observations $\mathcal{X} = \{x_1, \ldots, x_N\}$ *and reference shapes* $\mathcal{Y} = \{y_1, \ldots, y_M\}$*, learn a mapping* $x_n \mapsto y(x_n)$ *such that* $y(x_n)$ *fits the unknown target shape* y_n^**.*



Reference Shapes y_m Observation x_n Unknown Shape y_n^*

Problem Definition

Problem

Given observations $\mathcal{X} = \{x_1, \ldots, x_N\}$ *and reference shapes* $\mathcal{Y} = \{y_1, \ldots, y_M\}$*, learn a mapping* $x_n \mapsto y(x_n)$ *such that* $y(x_n)$ *fits the unknown target shape* y_n^**.*

Weakly-supervised:

- known object category;
- and bounding boxes required on real data.

Related Work

Related Work

Generative shape modeling, including [Wu et al., 2015, Wu et al., 2016].

Related Work

Generative shape modeling, including
[Wu et al., 2015, Wu et al., 2016].

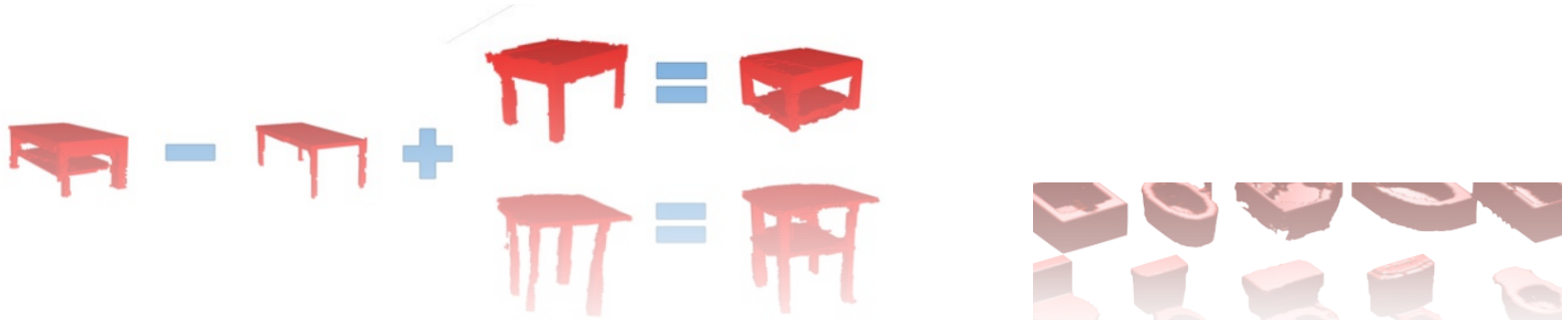
Shape completion, following [Sung et al., 2015]:

- symmetry based approaches;
- data-driven approaches, including
  [Dame et al., 2013, Engelmann et al., 2016];
- and recently learning-based approaches, including
  [Dai et al., 2016].

Selected Related Work

Generative shape modeling, *e.g.* [Wu et al., 2015, Wu et al., 2016]:

- Learn a generative model of shapes, *e.g.* using generative adversarial networks [Wu et al., 2016];
- and use for shape classification, manipulation and generation.

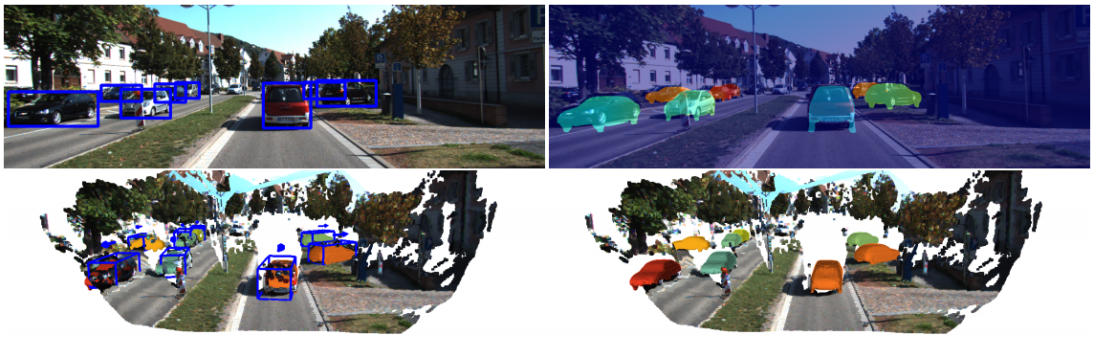


[Wu et al., 2016] [Wu et al., 2015]

Selected Related Work

Data-driven shape completion, *e.g.* [Dame et al., 2013, Engelmann et al., 2016]:

- Learn shape prior, *e.g.* using PCA or GP-LVM;
- and pose shape completion as energy minimization.

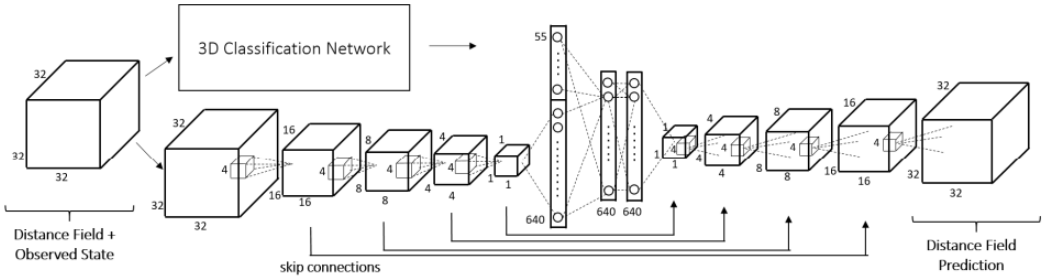


[Engelmann et al., 2016]

Selected Related Work

Learning-based shape completion, *e.g.* [Dai et al., 2016]:

- Learn an encoder-decoder network on synthetic data;
- and post-process if necessary.



[Dai et al., 2016]

Discussion

Two “philosophies”:

- data-driven approaches are applicable to real data, but shape completion involves energy minimization;
- learning-based approaches need supervision, but shape completion is “just a forward pass”.

Discussion

Question

Do strong shape priors allow us to learn shape completion under weak supervision?

Goal:

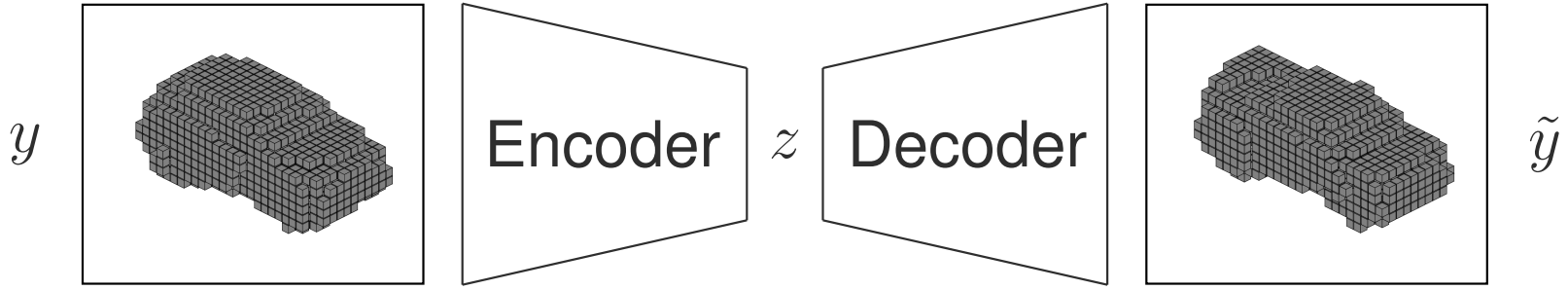
- Efficient shape completion;
- and learning on real data.

Proposed Approach

Proposed Approach

Shape prior.

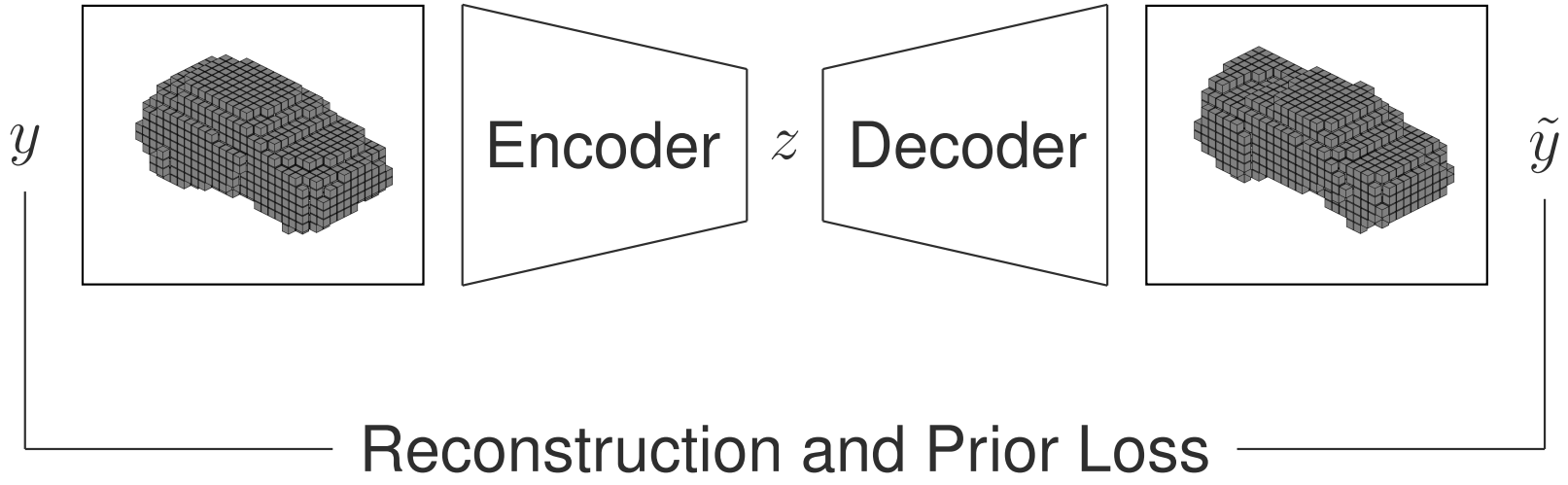
Learn a variational auto-encoder
[Kingma and Welling, 2013]:



Proposed Approach

Shape prior.

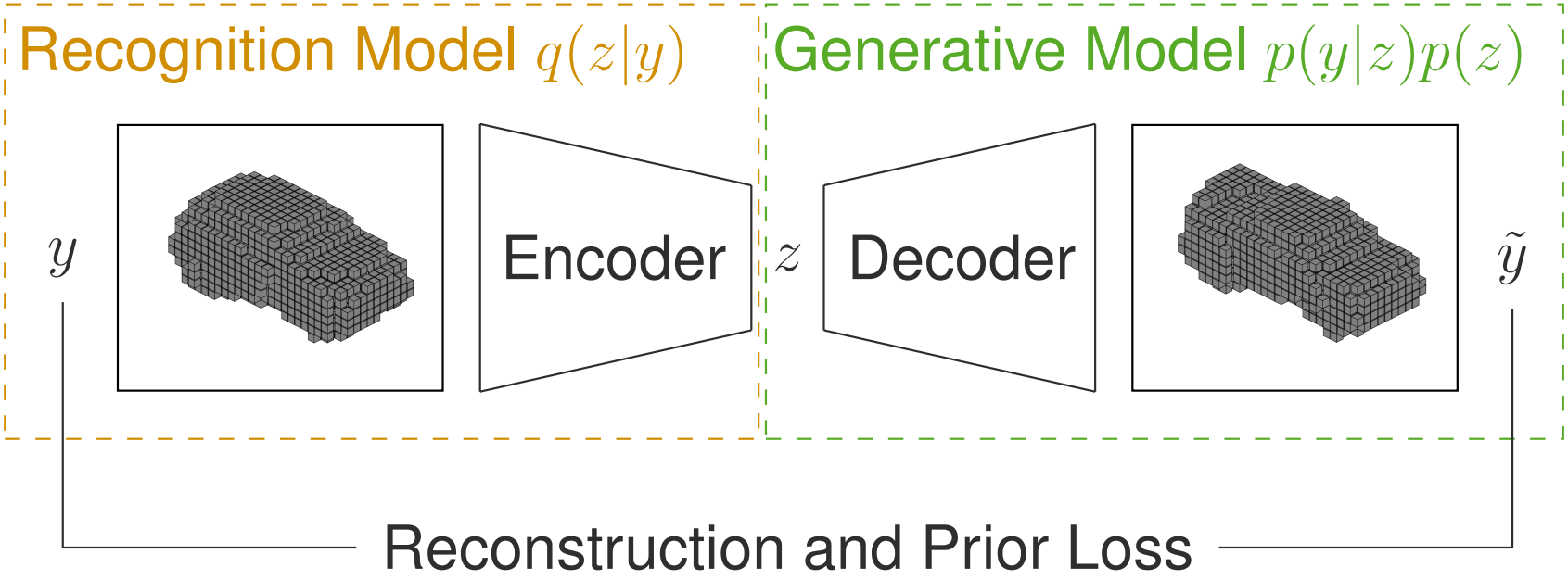
Learn a variational auto-encoder:



Proposed Approach

Shape prior.

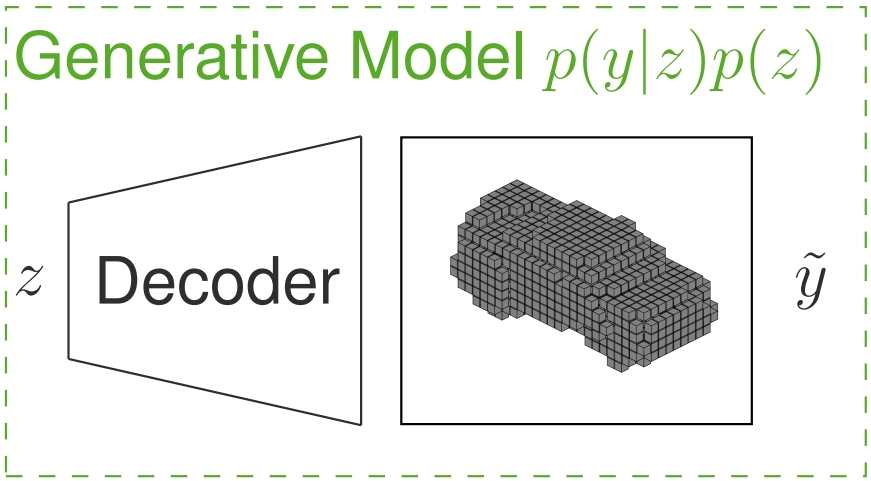
Learn a variational auto-encoder:



Proposed Approach

Shape inference.

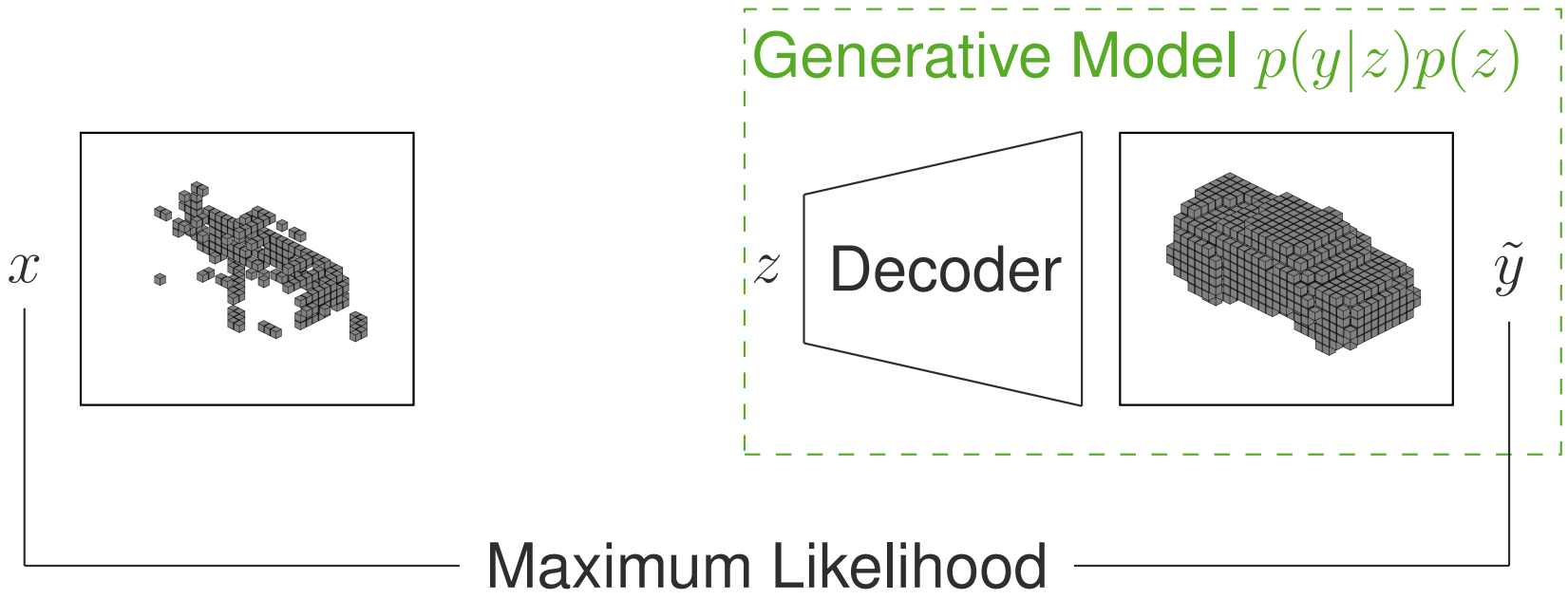
Perform maximum likelihood:



Proposed Approach

Shape inference.

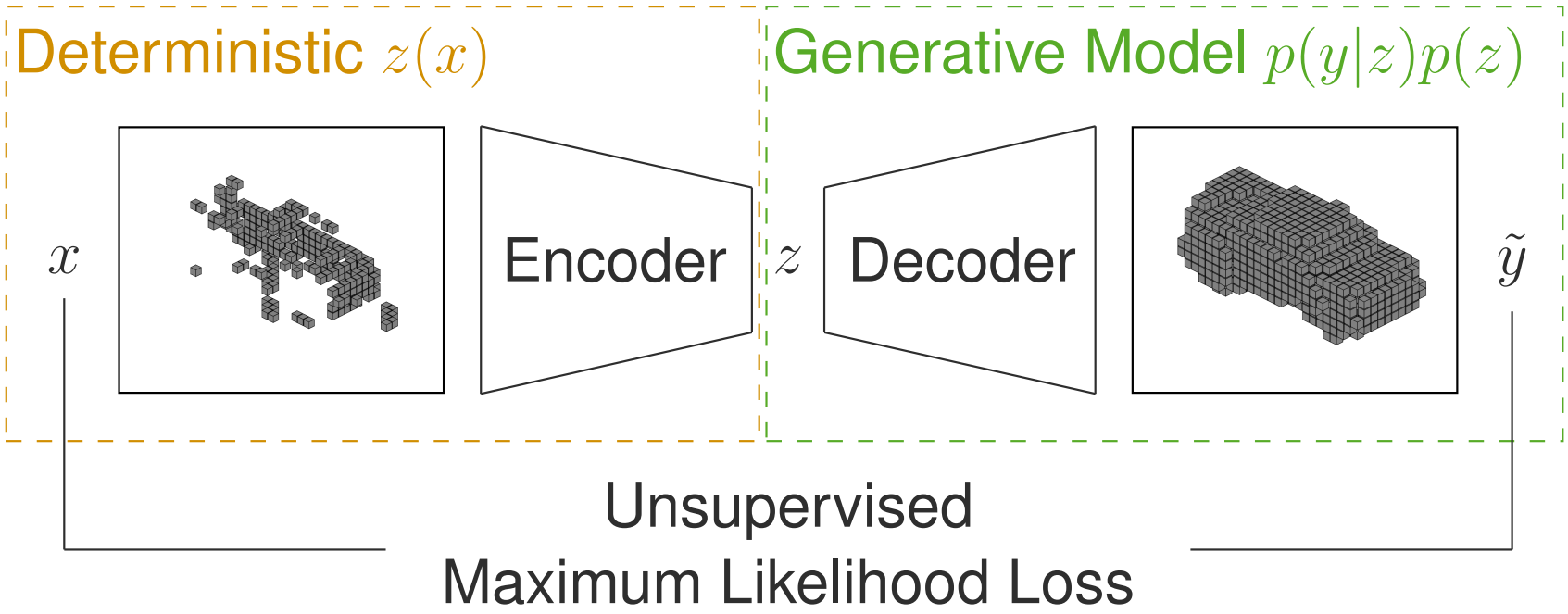
Perform maximum likelihood:



Proposed Approach

Shape inference.

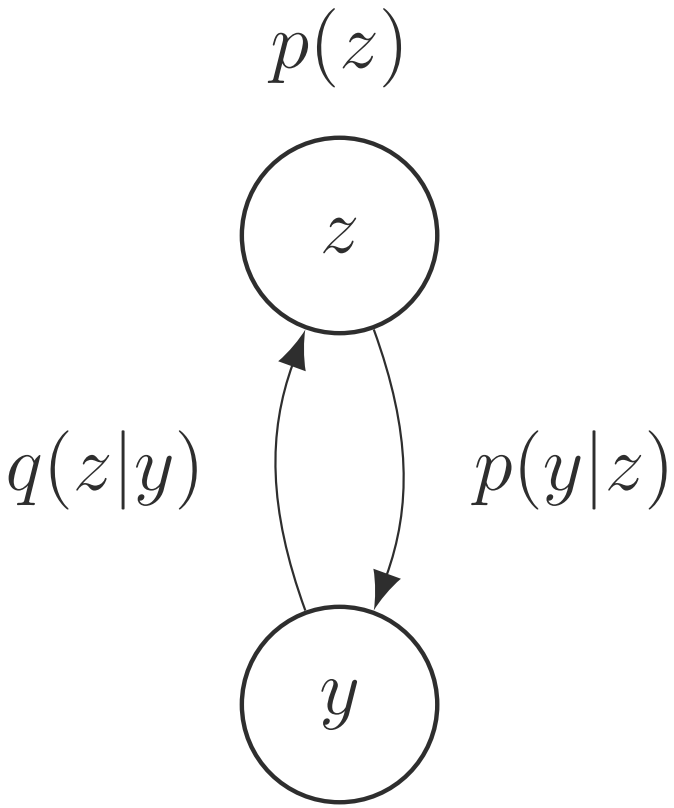
~~Perform~~ Learn – *i.e.* amortize – maximum likelihood:



Shape Prior

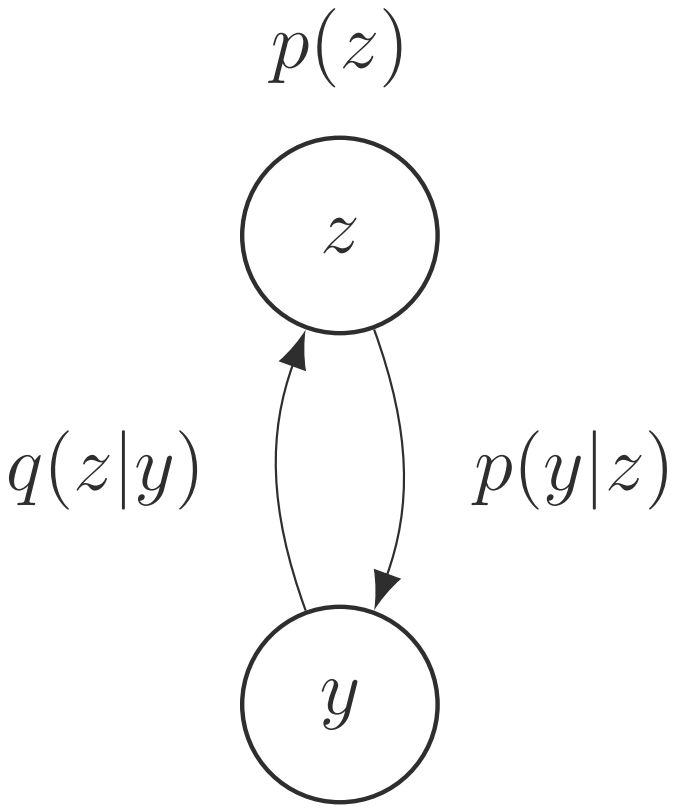
Shape Prior

Learn a variational auto-encoder:



Shape Prior

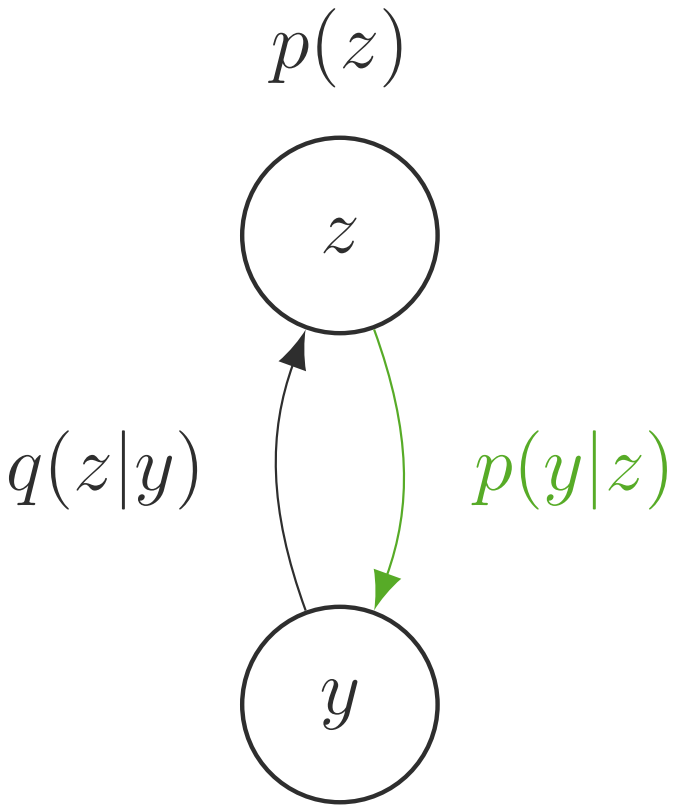
Learn a variational auto-encoder:



Prior $p(z) = \mathcal{N}(z|0, I)$

Shape Prior

Learn a variational auto-encoder:



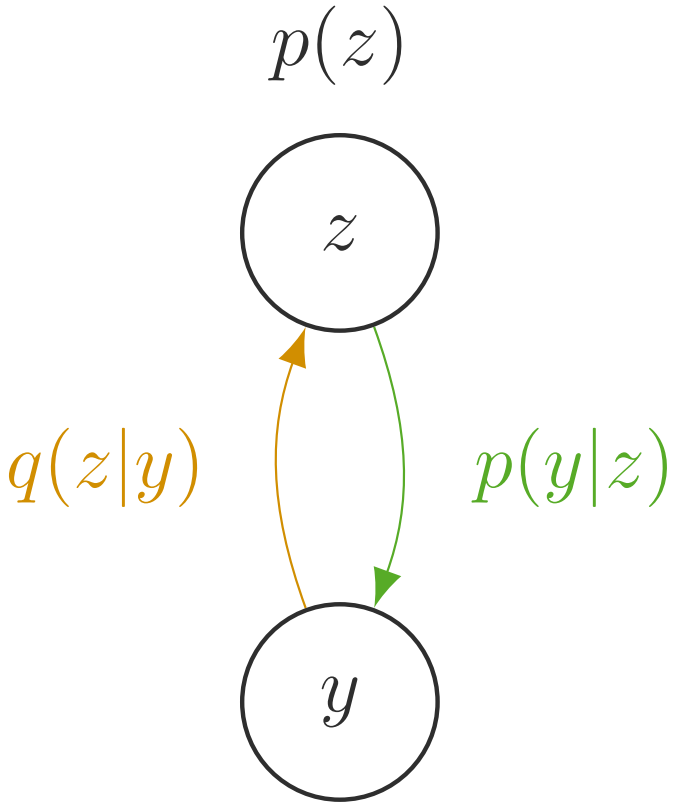
Prior $p(z) = \mathcal{N}(z|0, I)$

Decoder/Generative Model

$$p(y|z) = \prod_i \text{Ber}(y_i|\theta_i(z))$$

Shape Prior

Learn a variational auto-encoder:



Prior $p(z) = \mathcal{N}(z|0, I)$

Decoder/Generative Model

$$p(y|z) = \prod_i \text{Ber}(y_i|\theta_i(z))$$

Encoder/Recognition Model

$$q(z|y) = \mathcal{N}(z|\mu(y), \sigma^2(y))$$

Shape Prior

Maximum likelihood leads to:

$$\mathcal{L}_{\text{ELBO}} = -\mathbb{E}_{q(z|y)}[\log p(y|z)] + \text{KL}(q(z|y)|p(z))$$

- Encoder: $q(z|y) = \mathcal{N}(z|\mu(y), \sigma^2(y))$;
- and decoder: $p(y|z) = \prod_i \text{Ber}(y_i|\theta_i(z))$.

Shape Prior

Maximum likelihood leads to:

$$\mathcal{L}_{\text{ELBO}} = -\mathbb{E}_{q(z|y)}[\log p(y|z)] + \text{KL}(q(z|y)|p(z))$$

prior

- Encoder: $q(z|y) = \mathcal{N}(z|\mu(y), \sigma^2(y))$;
- and decoder: $p(y|z) = \prod_i \text{Ber}(y_i|\theta_i(z))$.

Shape Prior

Maximum likelihood leads to:

$$\mathcal{L}_{\text{ELBO}} = -\mathbb{E}_{q(z|y)}[\log p(y|z)] + \mathcal{P}_{\text{Reg}}(\mu(y), \sigma^2(y))$$

- Encoder: $q(z|y) = \mathcal{N}(z|\mu(y), \sigma^2(y))$;
- and decoder: $p(y|z) = \prod_i \text{Ber}(y_i|\theta_i(z))$.

Shape Prior

Maximum likelihood leads to:

reconstruction

$$\mathcal{L}_{\text{ELBO}} = -\mathbb{E}_{q(z|y)}[\log p(y|z)] + \mathcal{P}_{\text{Reg}}(\mu(y), \sigma^2(y))$$

- Encoder: $q(z|y) = \mathcal{N}(z|\mu(y), \sigma^2(y))$;
- and decoder: $p(y|z) = \prod_i \text{Ber}(y_i|\theta_i(z))$.

Shape Prior

Maximum likelihood leads to:

$$\mathcal{L}_{\text{ELBO}} = \sum_i \mathcal{L}_{\text{BCE}}(\underset{\substack{\shortparallel \\ \theta_i(z)}}{\tilde{y}_i}, y_i) + \mathcal{P}_{\text{Reg}}(\mu(y), \sigma^2(y))$$

- Encoder: $q(z|y) = \mathcal{N}(z|\mu(y), \sigma^2(y))$;
- and decoder: $p(y|z) = \prod_i \text{Ber}(y_i|\theta_i(z))$.

Shape Prior

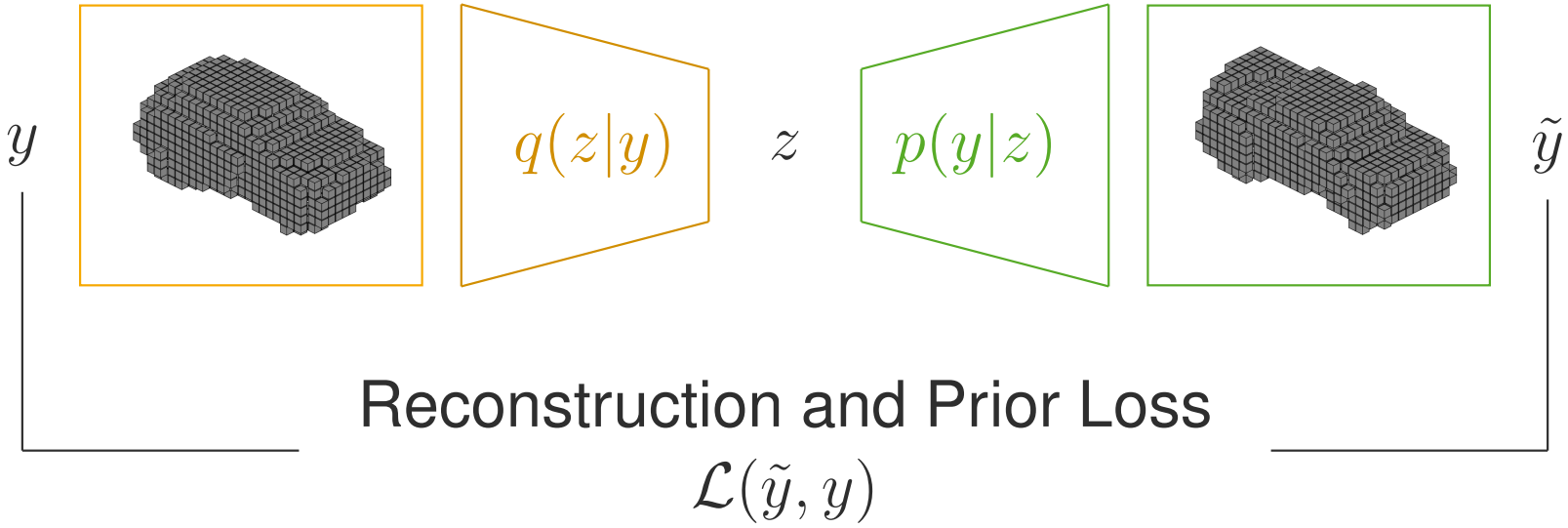
Maximum likelihood leads to:

$$\mathcal{L}(\tilde{y}, y) = \sum_i \mathcal{L}_{\text{BCE}}(\tilde{y}_i, y_i) + \mathcal{P}_{\text{Reg}}(\mu(y), \sigma^2(y))$$

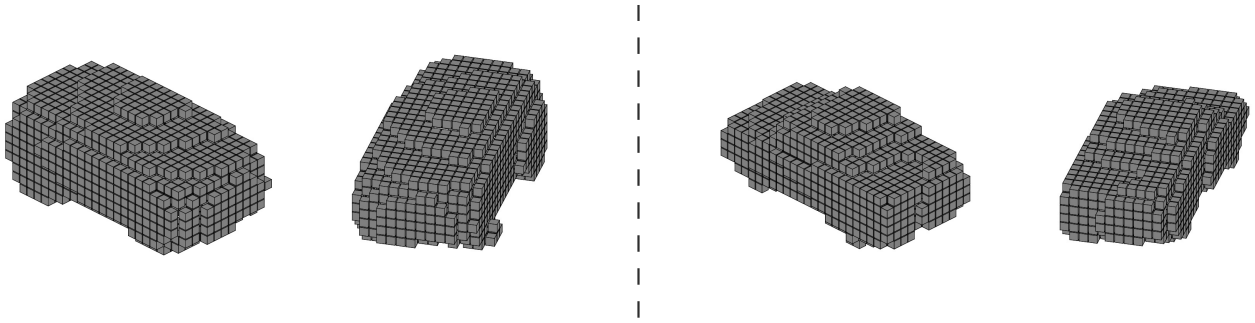
- Encoder: $q(z|y) = \mathcal{N}(z|\mu(y), \sigma^2(y))$;
- and decoder: $p(y|z) = \prod_i \text{Ber}(y_i|\theta_i(z))$.

Shape Prior

Training a variational auto-encoder:



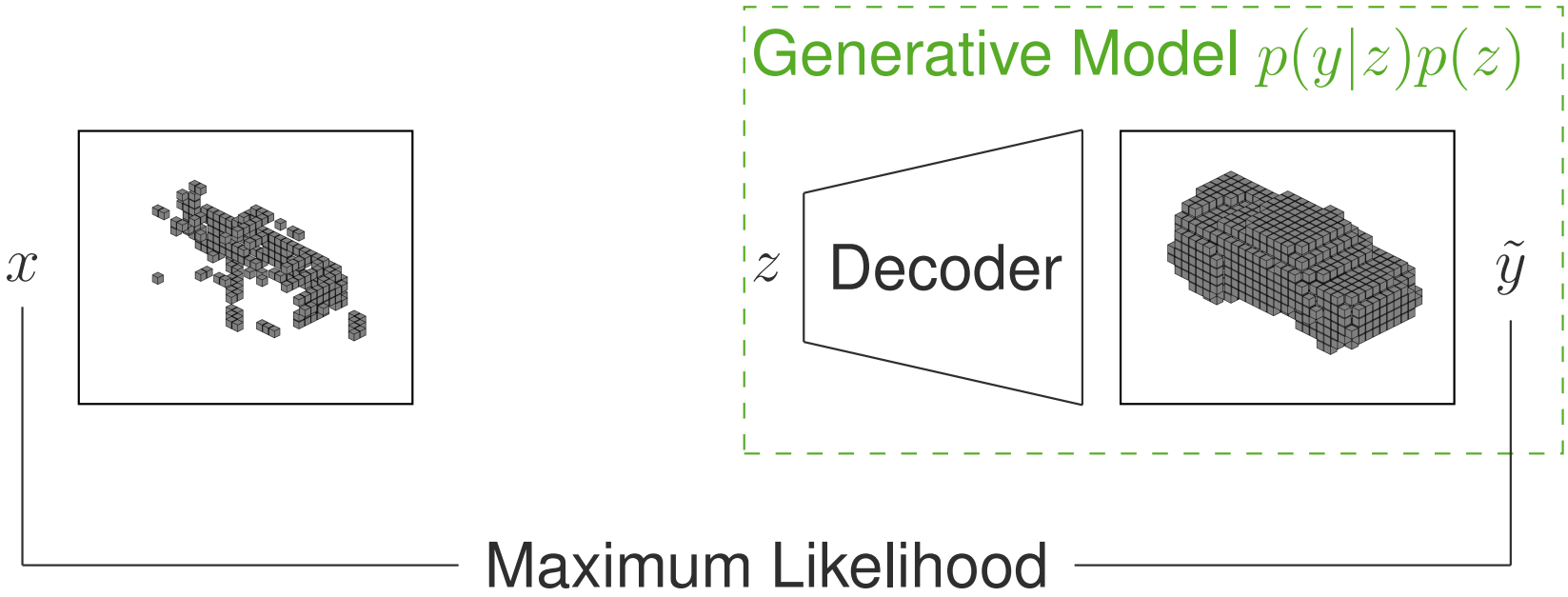
Generating random shapes:



Shape Inference

Shape Inference

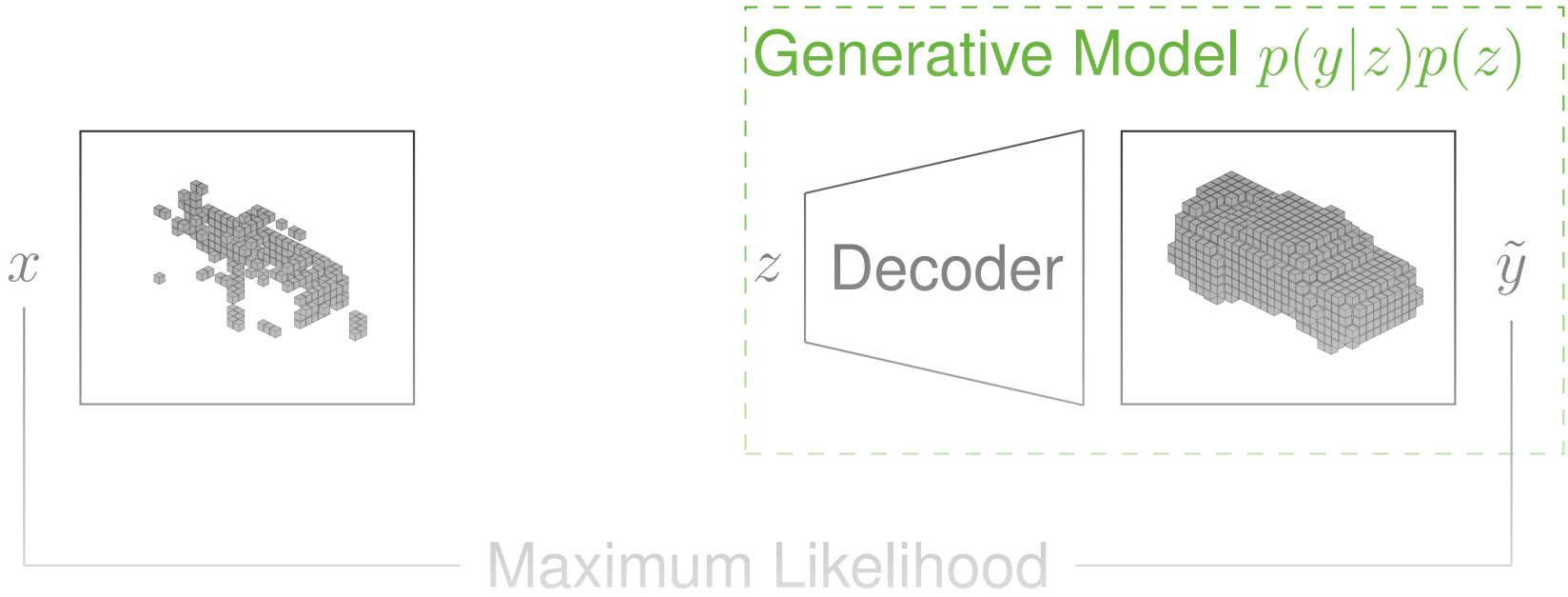
Maximize the likelihood of observation x over the latent space:



Shape Inference

Minimize the negative log-likelihood of observation x over the latent space:

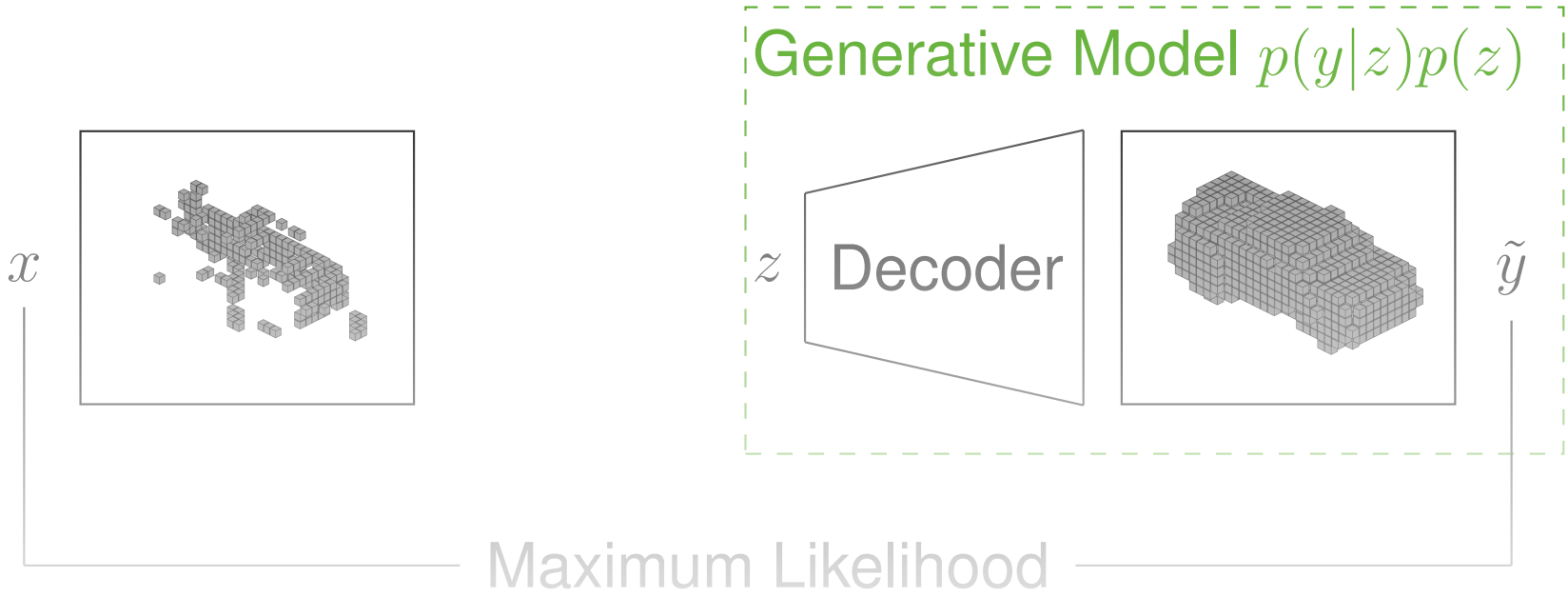
$$\begin{aligned} &\underset{z}{\operatorname{argmin}} - \ln p(y = x, z) \\ &= \underset{z}{\operatorname{argmin}} - \sum_i \ln p(y_i = x_i | z) - \ln p(z) \end{aligned}$$



Shape Inference

Minimize the negative log-likelihood of observation x over the latent space:

$$\begin{aligned} &\underset{z}{\operatorname{argmin}} - \ln p(y = x, z) \\ &= \underset{z}{\operatorname{argmin}} \sum_i \mathcal{L}_{\text{BCE}}(\tilde{y}_i, x_i) - \ln p(z) \end{aligned}$$

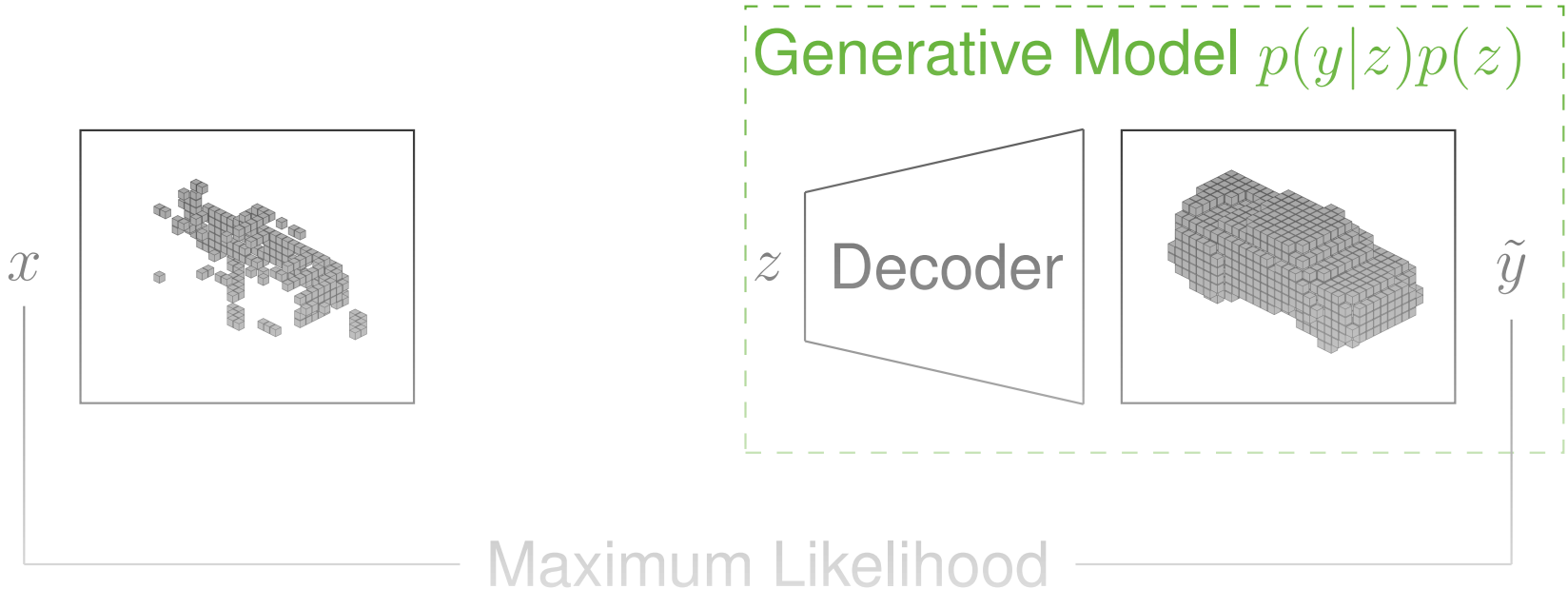


Shape Inference

Minimize the negative log-likelihood of observation x over the latent space:

$$\underset{z}{\operatorname{argmin}} - \ln p(y = x, z)$$

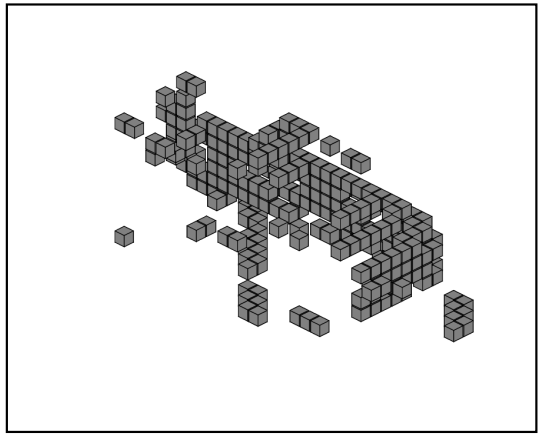
$$= \underset{z}{\operatorname{argmin}} \sum_i \mathcal{L}_{\text{BCE}}(\tilde{y}_i, x_i) + \text{const} + \frac{1}{2}\|z\|_2^2$$



Shape Inference

What observations do we have?

- Occupied voxels $x_i = 1$ (from observed points);

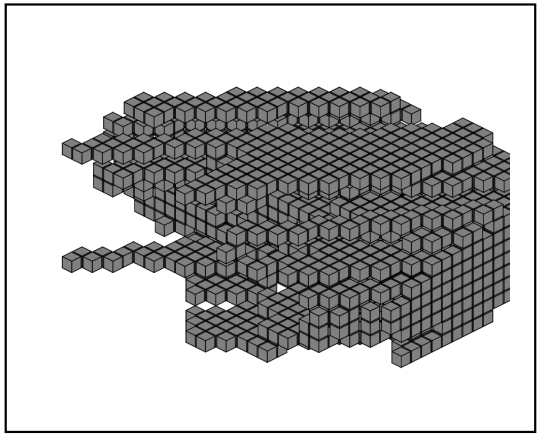


Occupied
Voxels $x_i = 1$

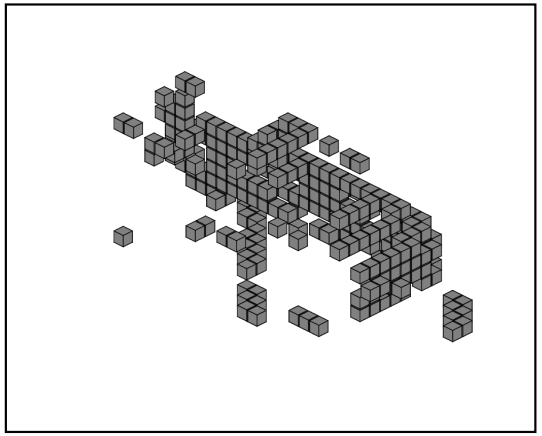
Shape Inference

What observations do we have?

- Occupied voxels $x_i = 1$ (from observed points);
- unoccupied voxels $x_i = 0$ (from free space);



Free Space
Voxels $x_i = 0$

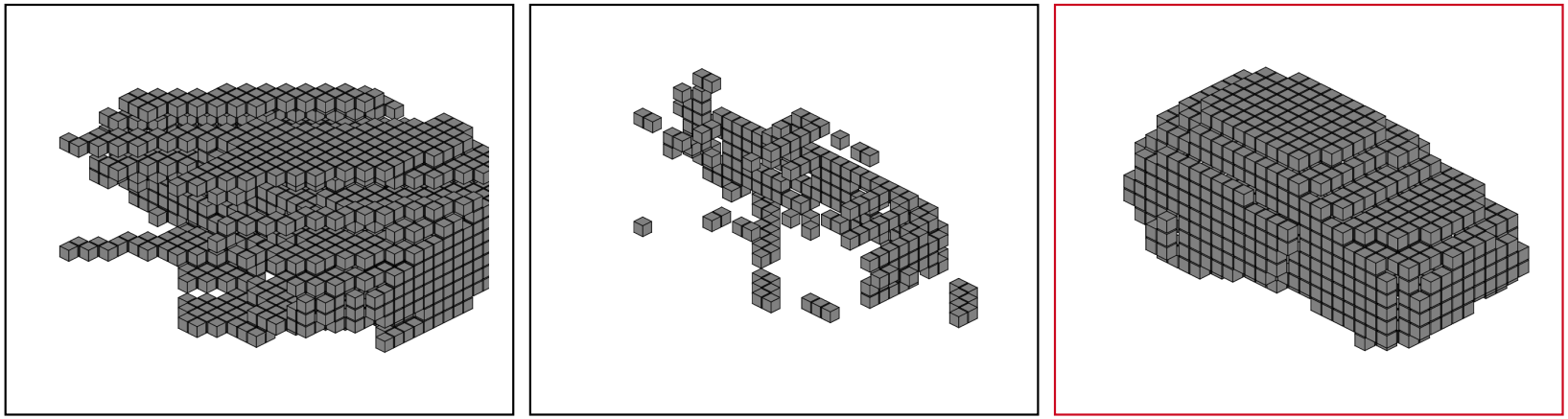


Occupied
Voxels $x_i = 1$

Shape Inference

What observations do we have?

- Occupied voxels $x_i = 1$ (from observed points);
- unoccupied voxels $x_i = 0$ (from free space);
- and unknown voxels "$x_i = \perp$".



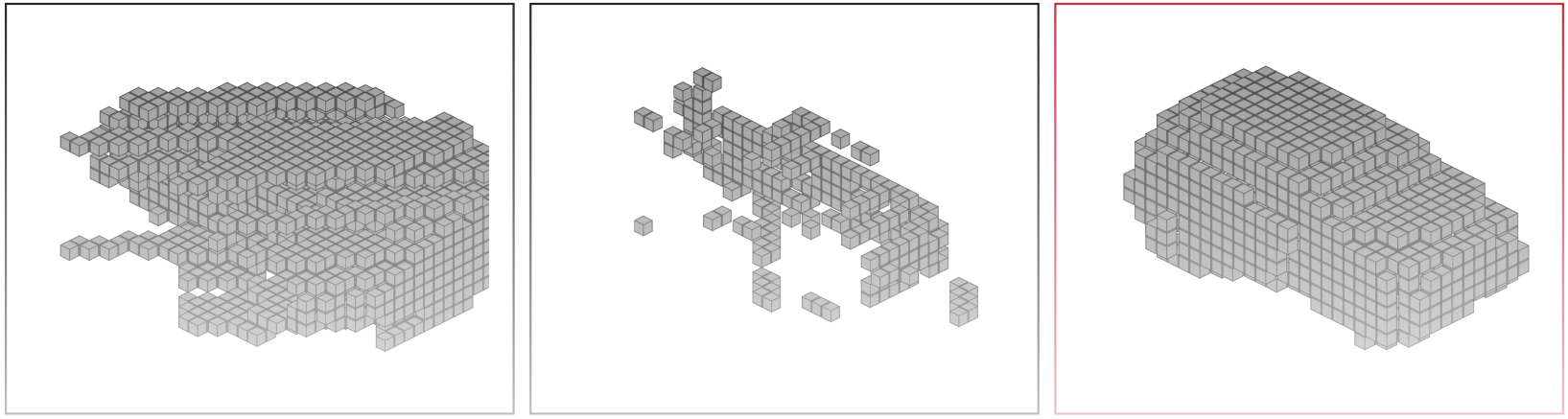
Free Space Voxels $x_i = 0$ | Occupied Voxels $x_i = 1$ | Unknown Shape y^*

Shape Inference

Minimize the negative log-likelihood of observation x over the latent space:

$$\underset{z}{\operatorname{argmin}} - \ln p(y = x, z)$$

$$= \underset{z}{\operatorname{argmin}} \sum_{x_i \neq \perp} \mathcal{L}_{\text{BCE}}(\tilde{y}_i, x_i) + \text{const} + \frac{1}{2}\|z\|_2^2$$



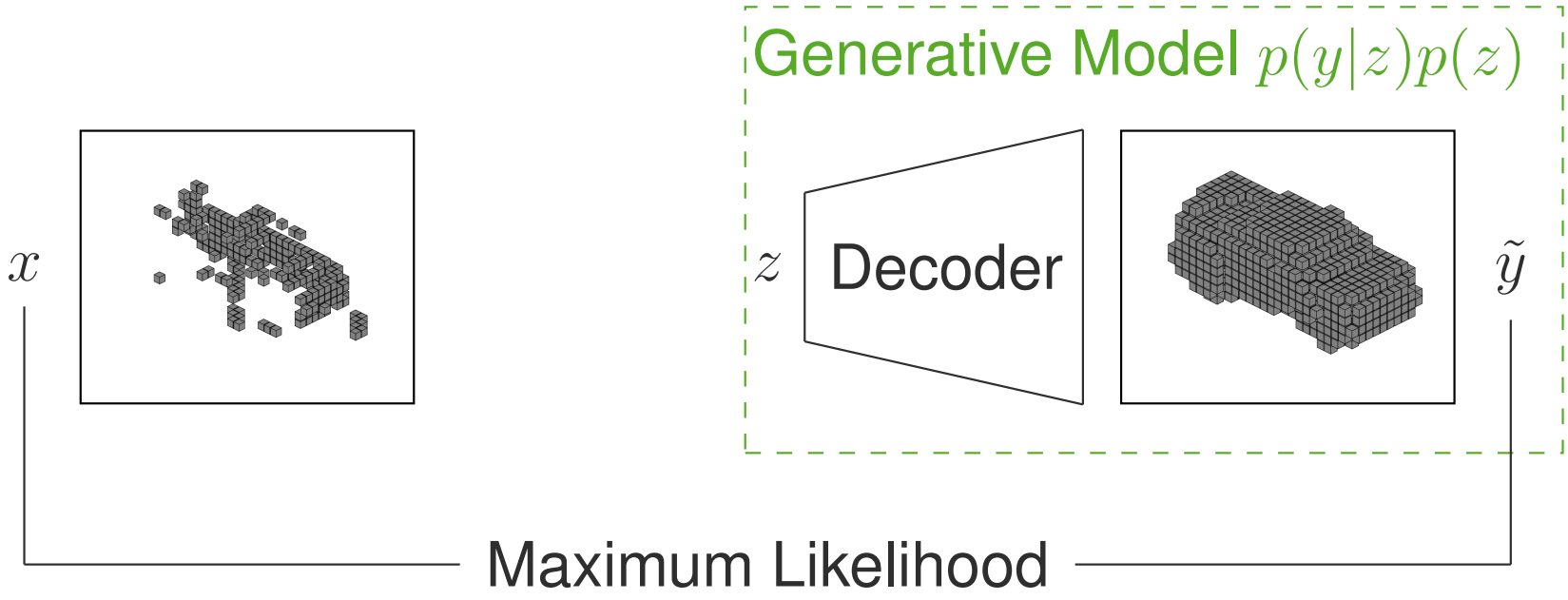
Free Space
Voxels $x_i = 0$

Occupied
Voxels $x_i = 1$

Unknown Shape y^*

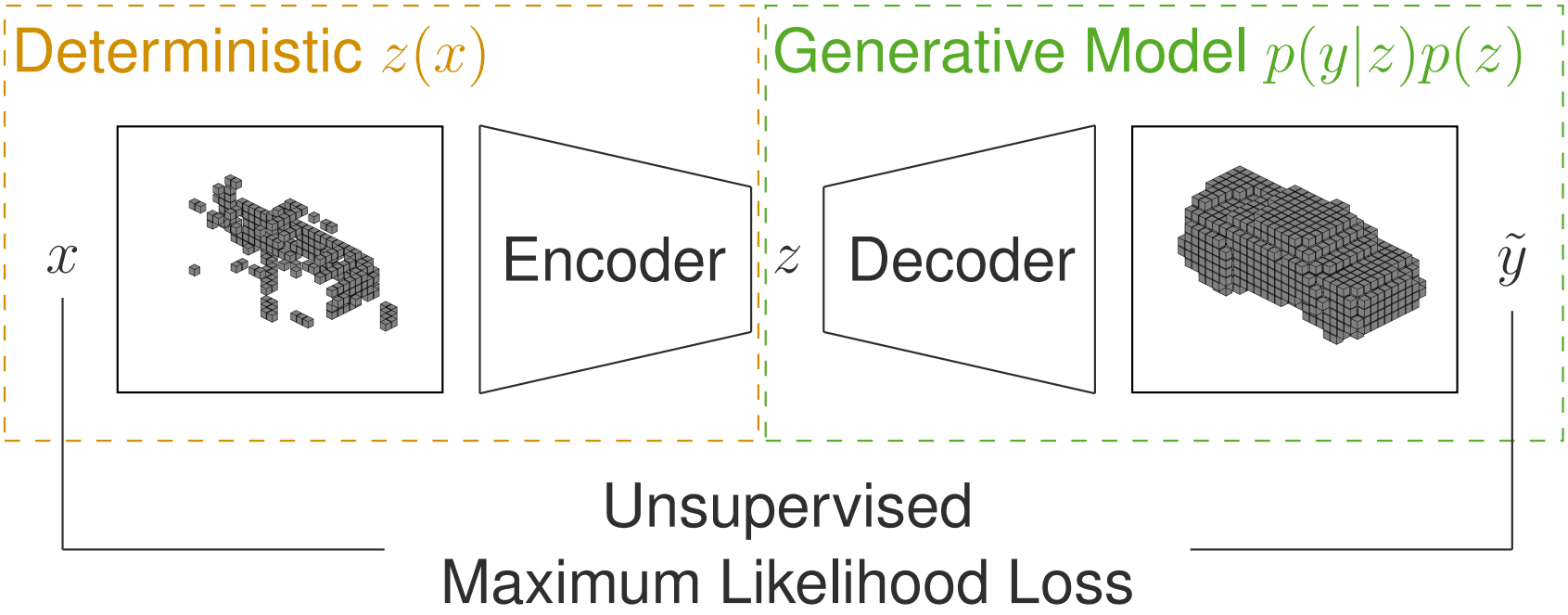
Shape Inference

Minimize the negative log-likelihood of observation x over the latent space:



Shape Inference

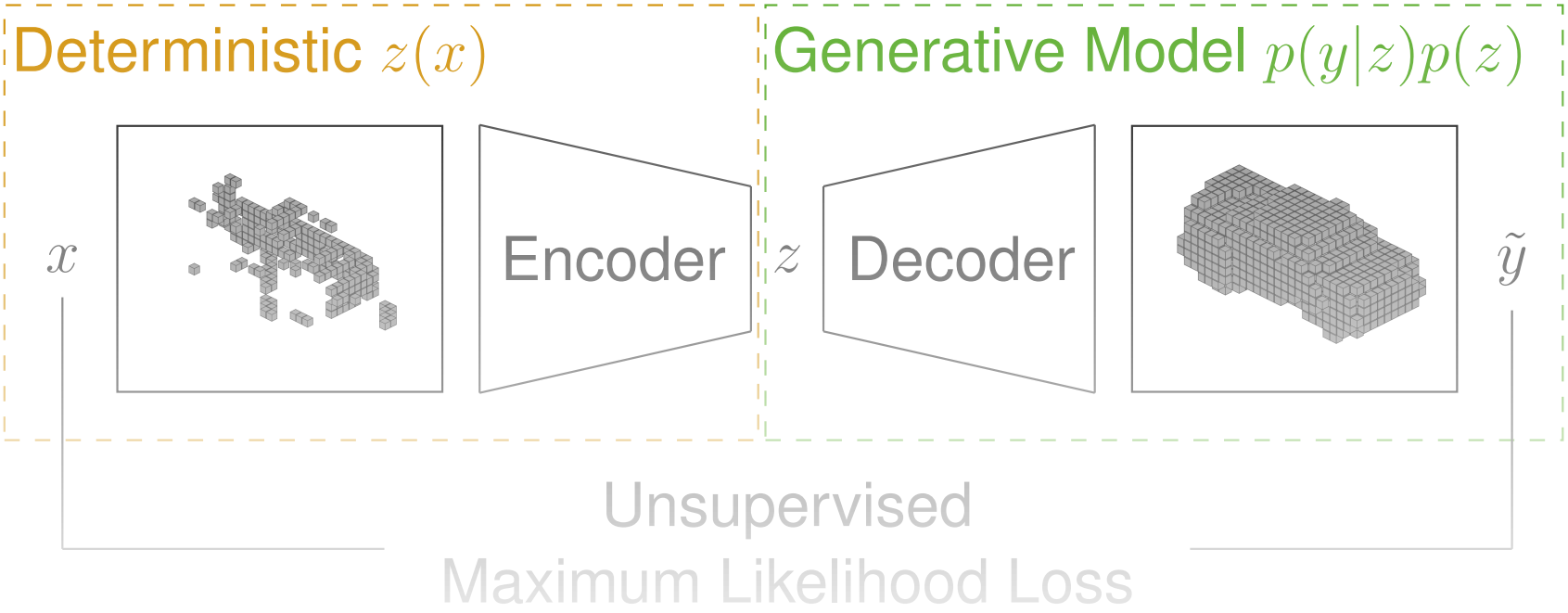
Learn to minimize the negative log-likelihood of observation x over the latent space:



Shape Inference

Learn to minimize the negative log-likelihood of observation x over the latent space:

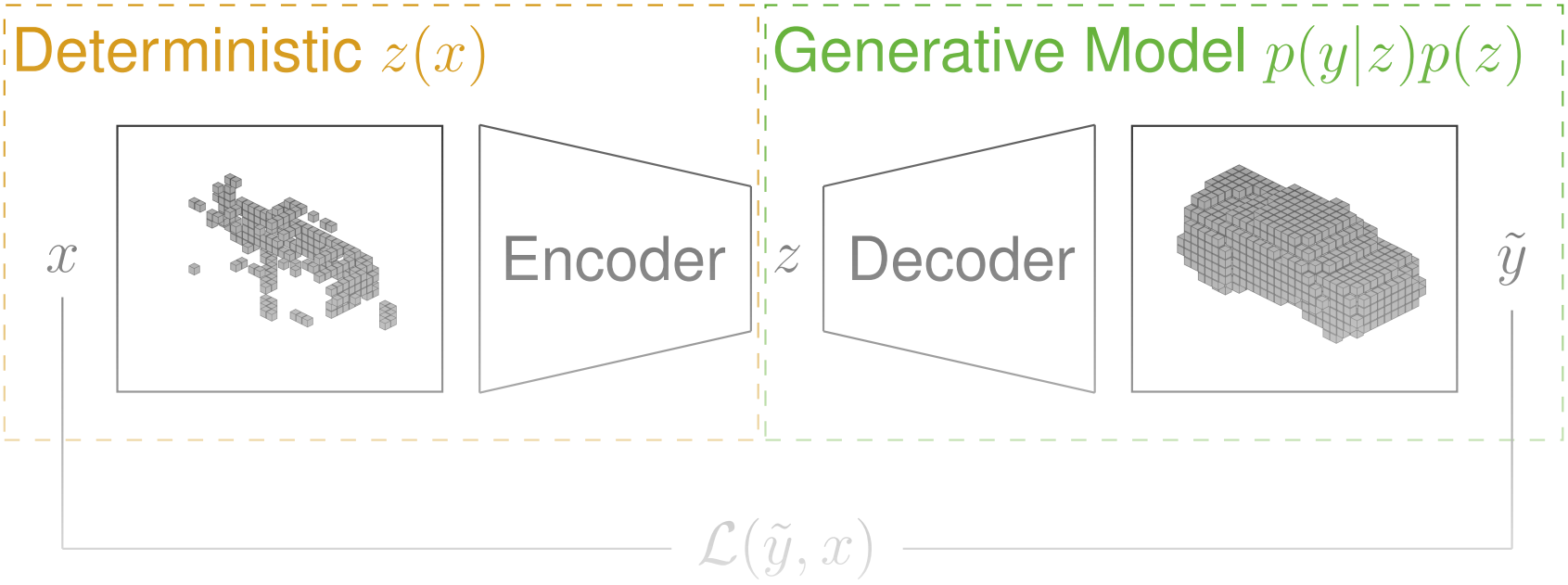
$$x \mapsto \tilde{z}(x) \approx \operatorname*{argmin}_{z} - \sum_{x_i \neq \perp} \ln p(y_i = x_i | z) - \ln p(z)$$



Shape Inference

Learn to minimize the negative log-likelihood of observation x over the latent space:

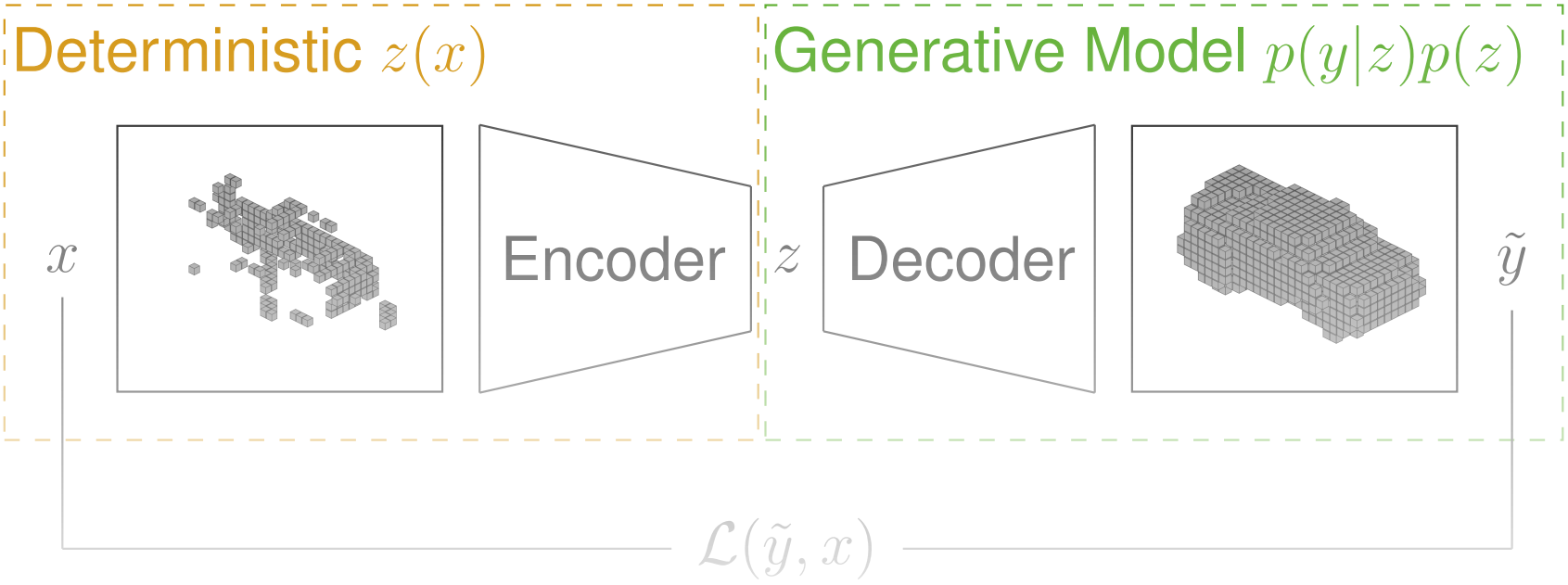
$$\mathcal{L}(\tilde{y}, x) = -\sum_{x_i \neq \perp} \ln p(\tilde{y}_i = x_i | z) - \ln p(z)$$



Shape Inference

Learn to minimize the negative log-likelihood of observation x over the latent space:

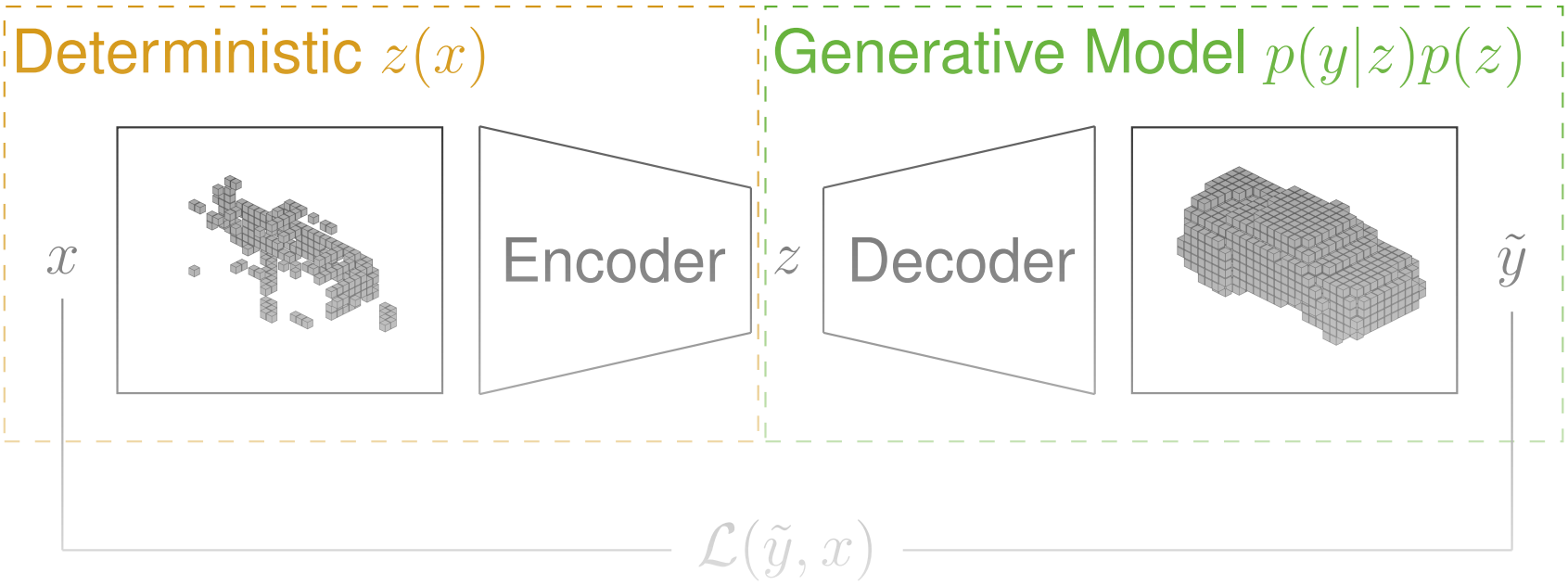
$$\mathcal{L}(\tilde{y}, x) = \sum_{x_i \neq \perp} \mathcal{L}_{\text{BCE}}(\tilde{y}_i, x_i) - \ln p(z)$$



Shape Inference

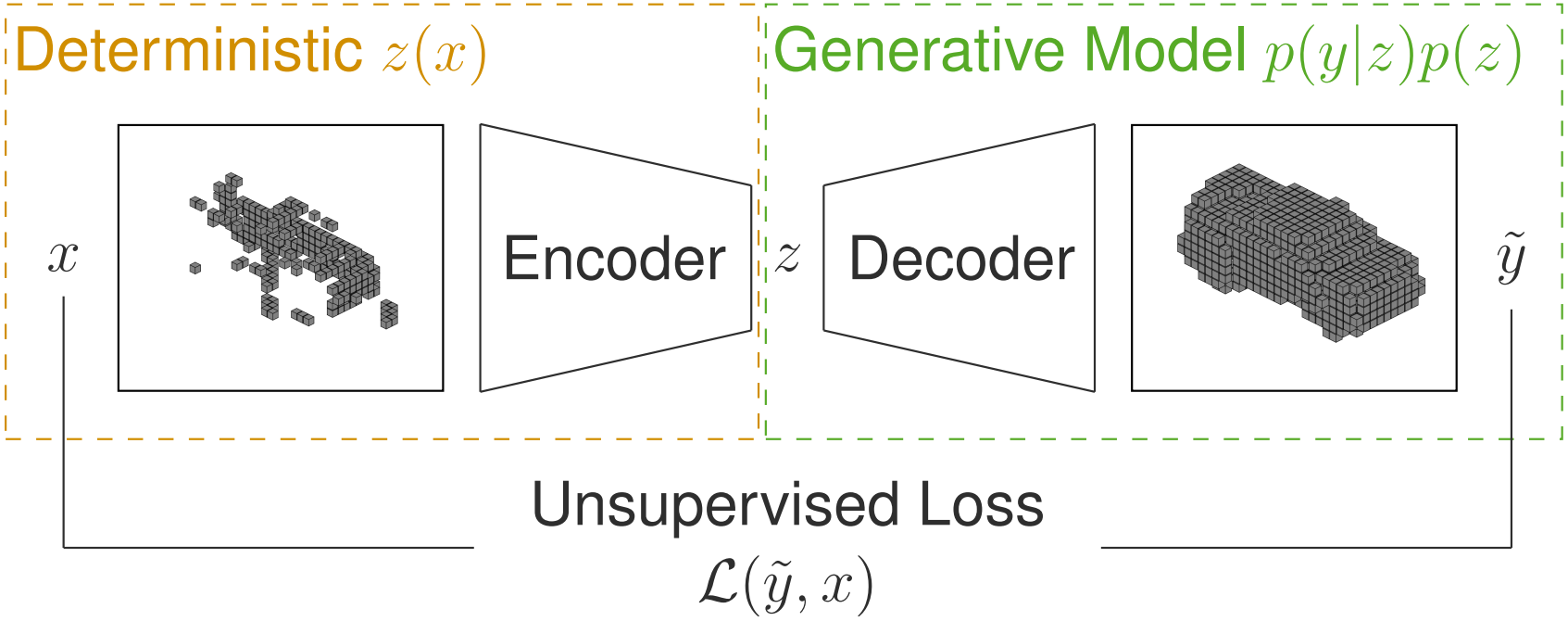
Learn to minimize the negative log-likelihood of observation x over the latent space:

$$\mathcal{L}(\tilde{y}, x) = \sum_{x_i \neq \perp} \mathcal{L}_{\text{BCE}}(\tilde{y}_i, x_i) + \text{const} + \frac{1}{2}\|z(x)\|_2^2$$



Shape Inference

Amortized maximum likelihood:

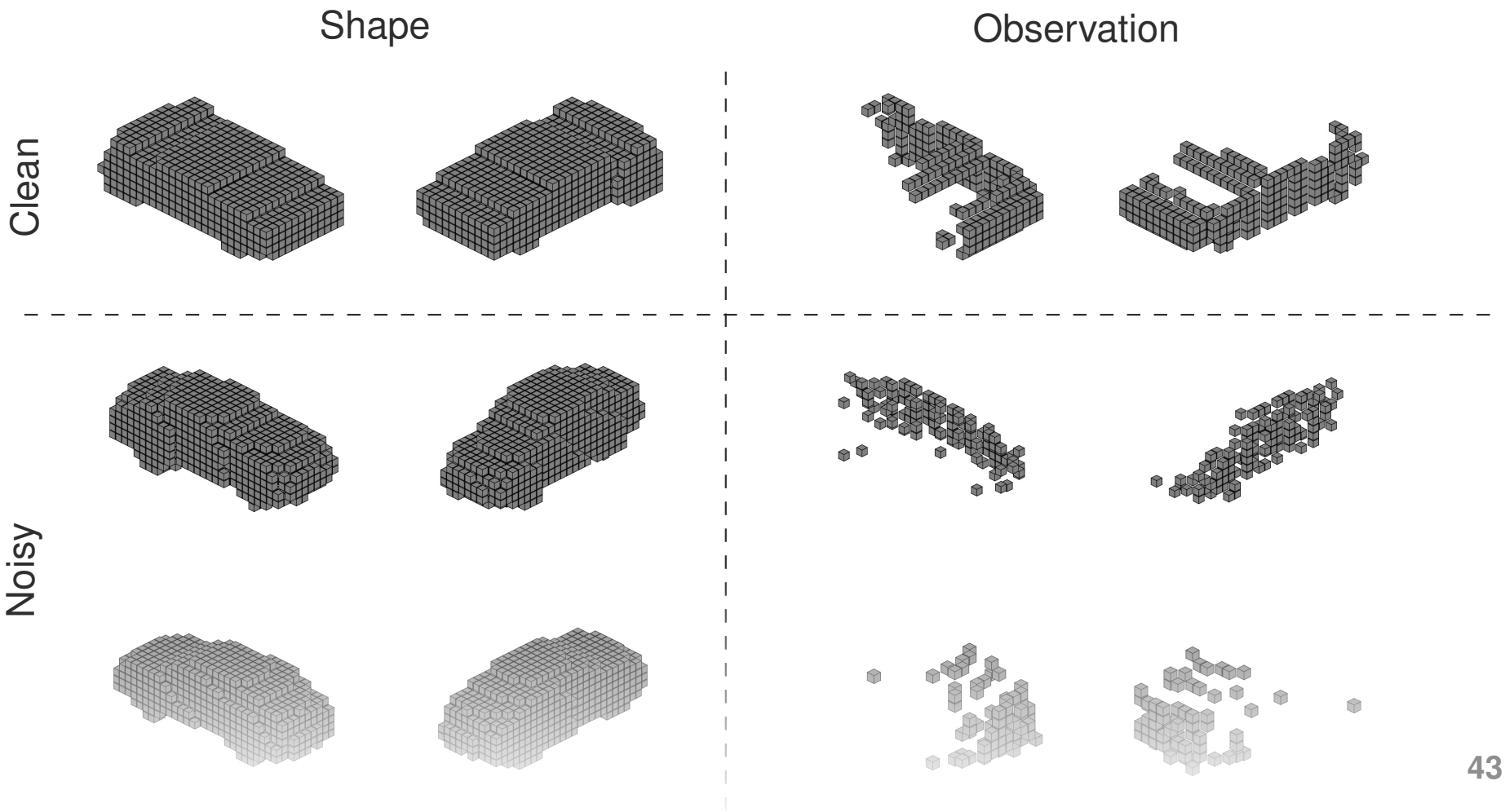


- Can be extended to signed distance functions;
- learning on real data and efficient inference.

Experiments

ShapeNet

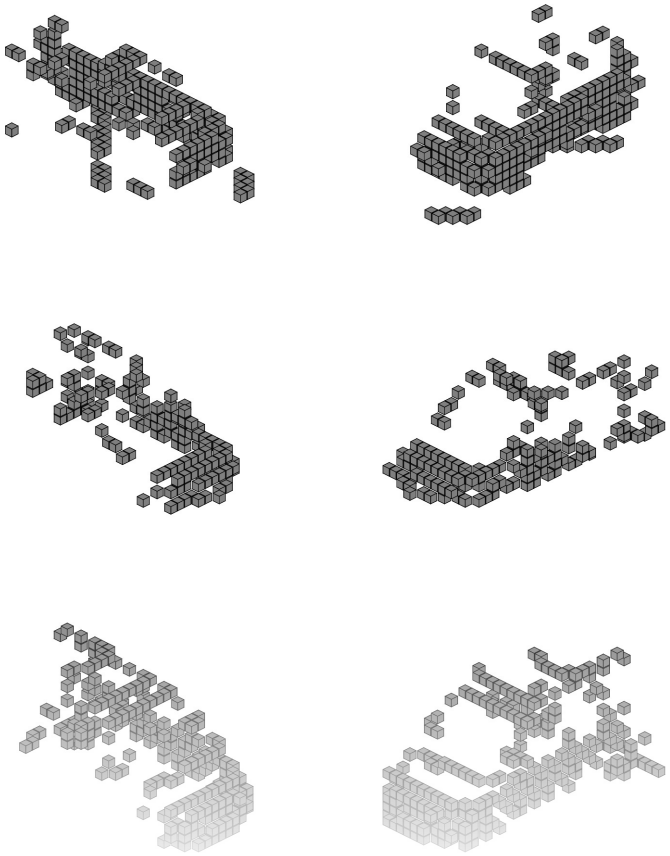
[Chang et al., 2015]; $\sim 3k$ car models, voxelized to 32^3, and synthetically generated observations.



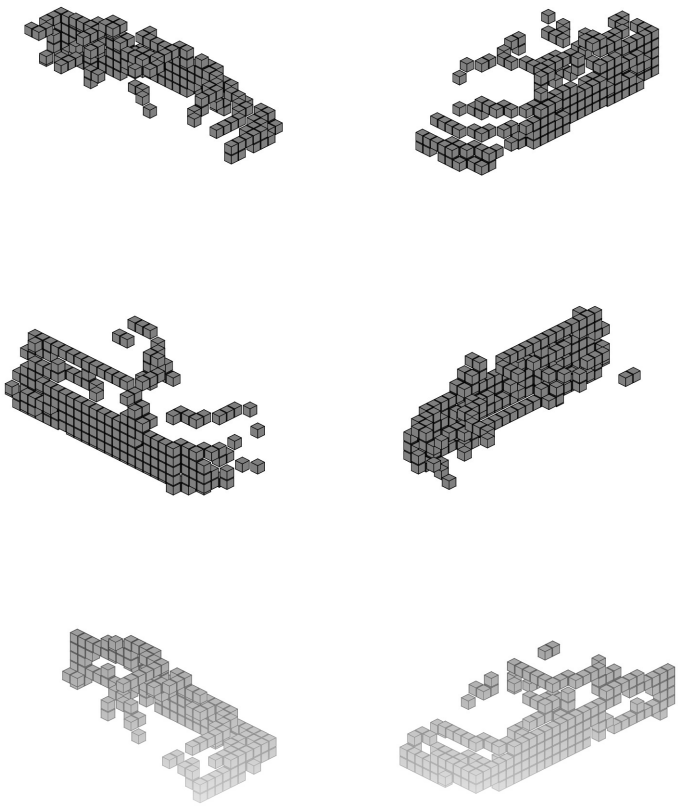
KITTI

[Geiger et al., 2012]; ground truth 3D bounding boxes, voxelized to 32^3, without target shapes.

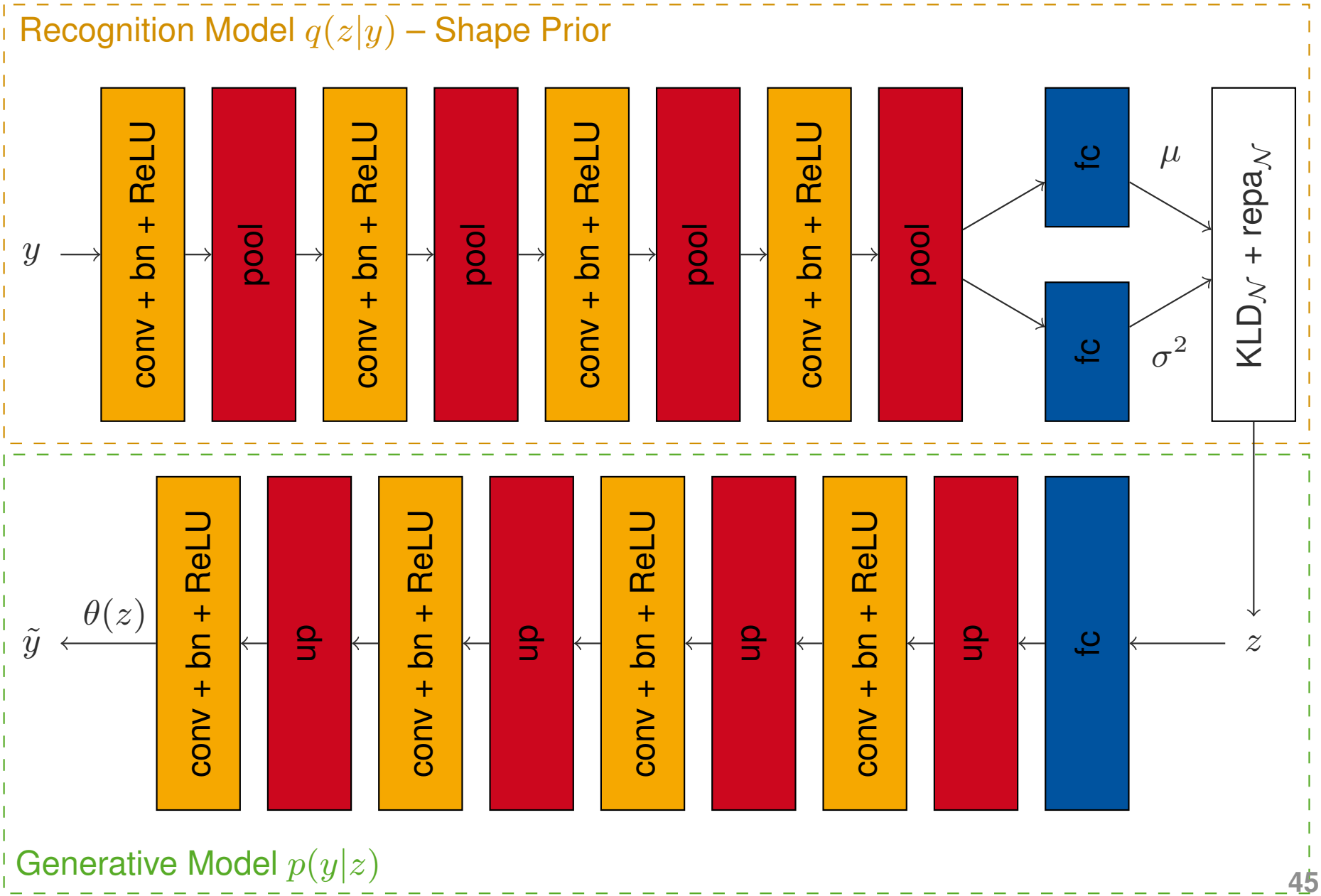
Observation



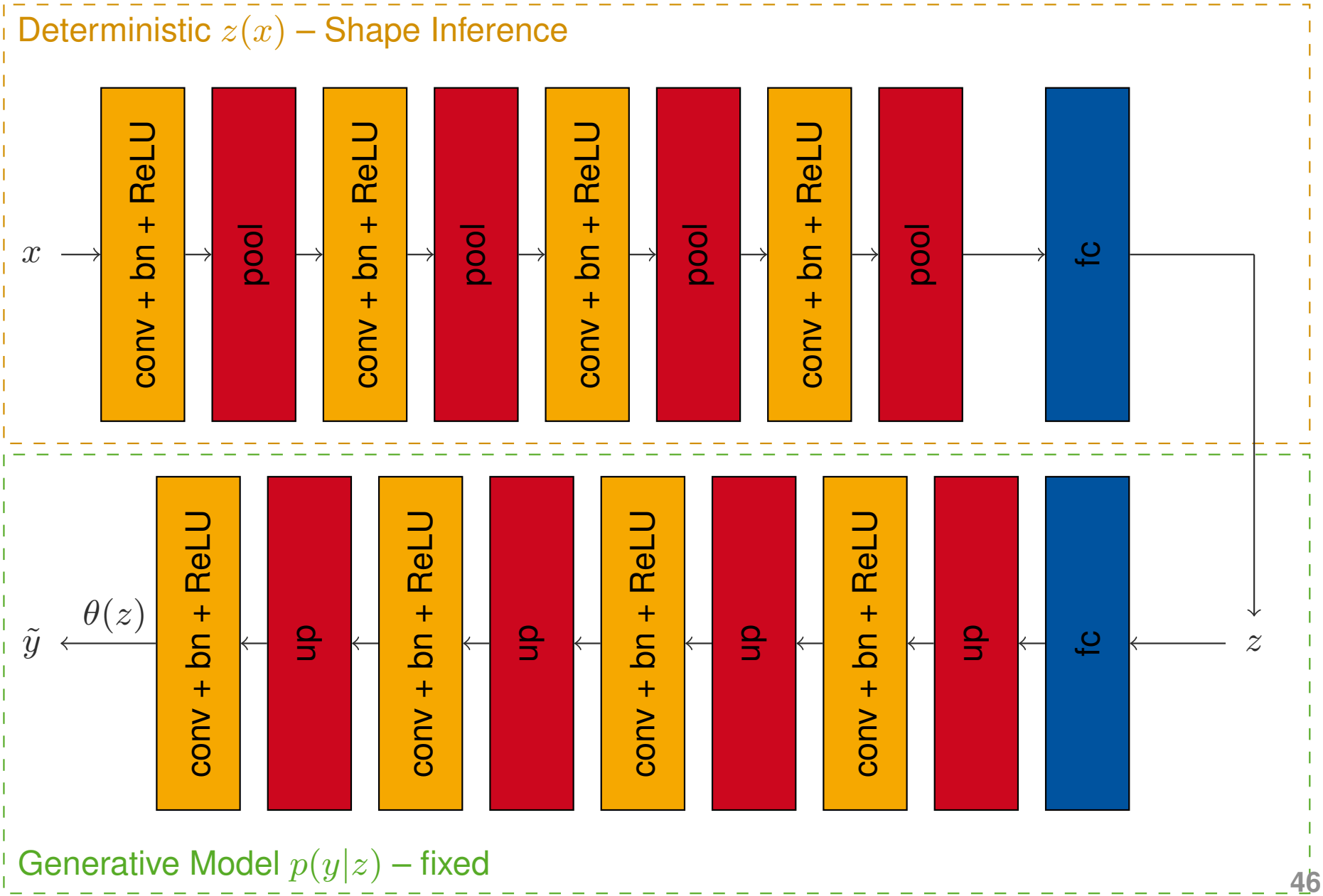
Observation



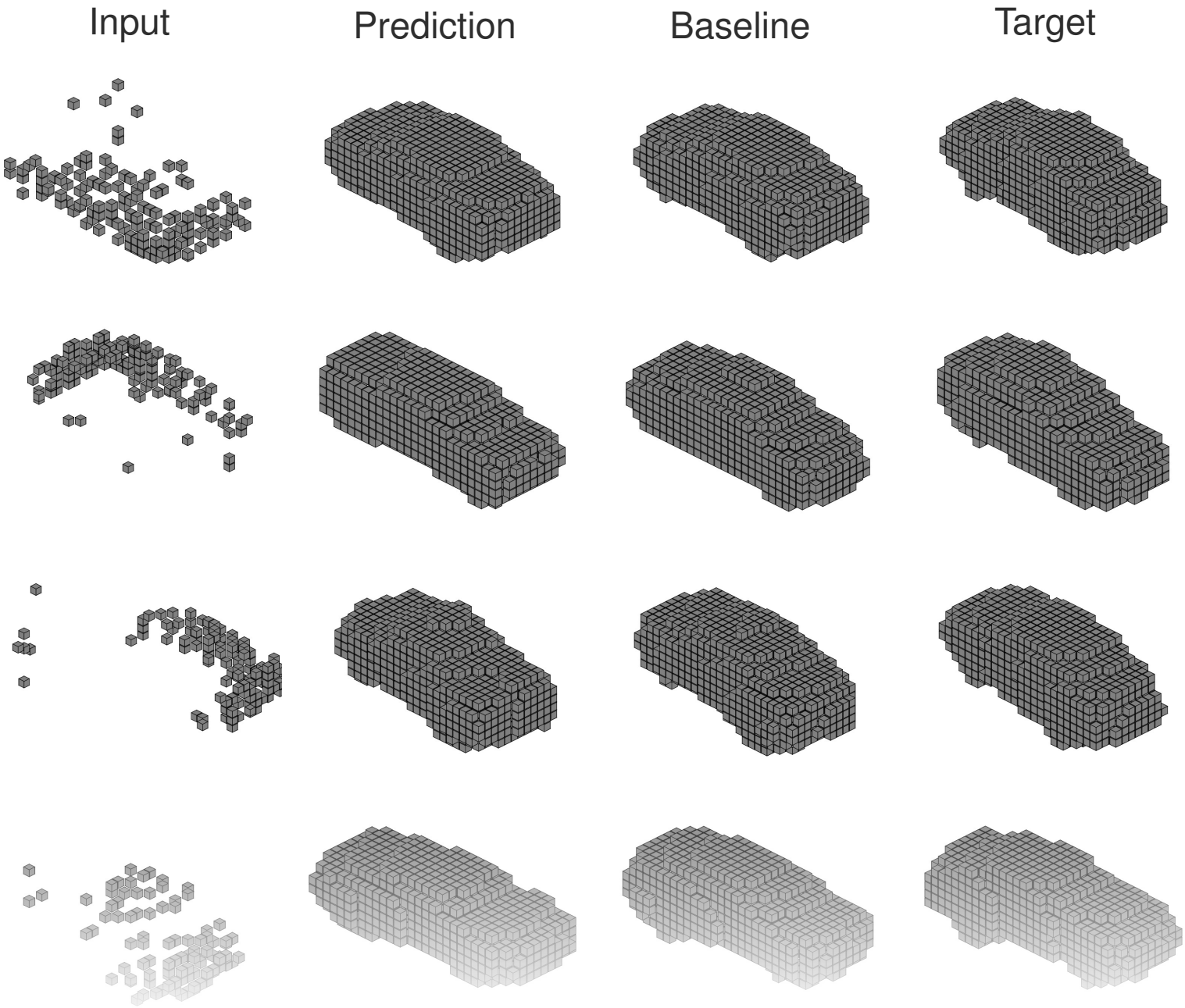
Architecture



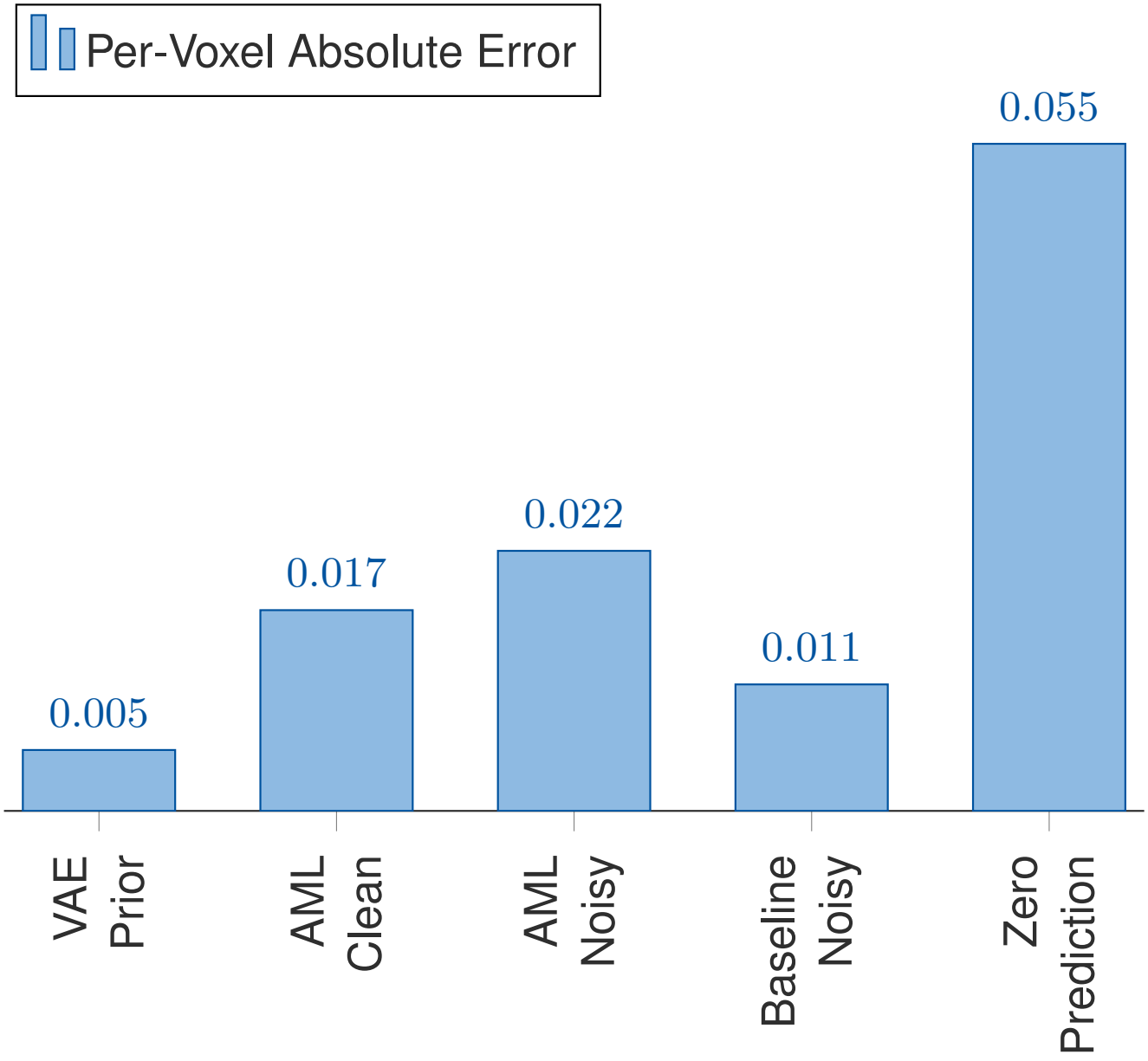
Architecture



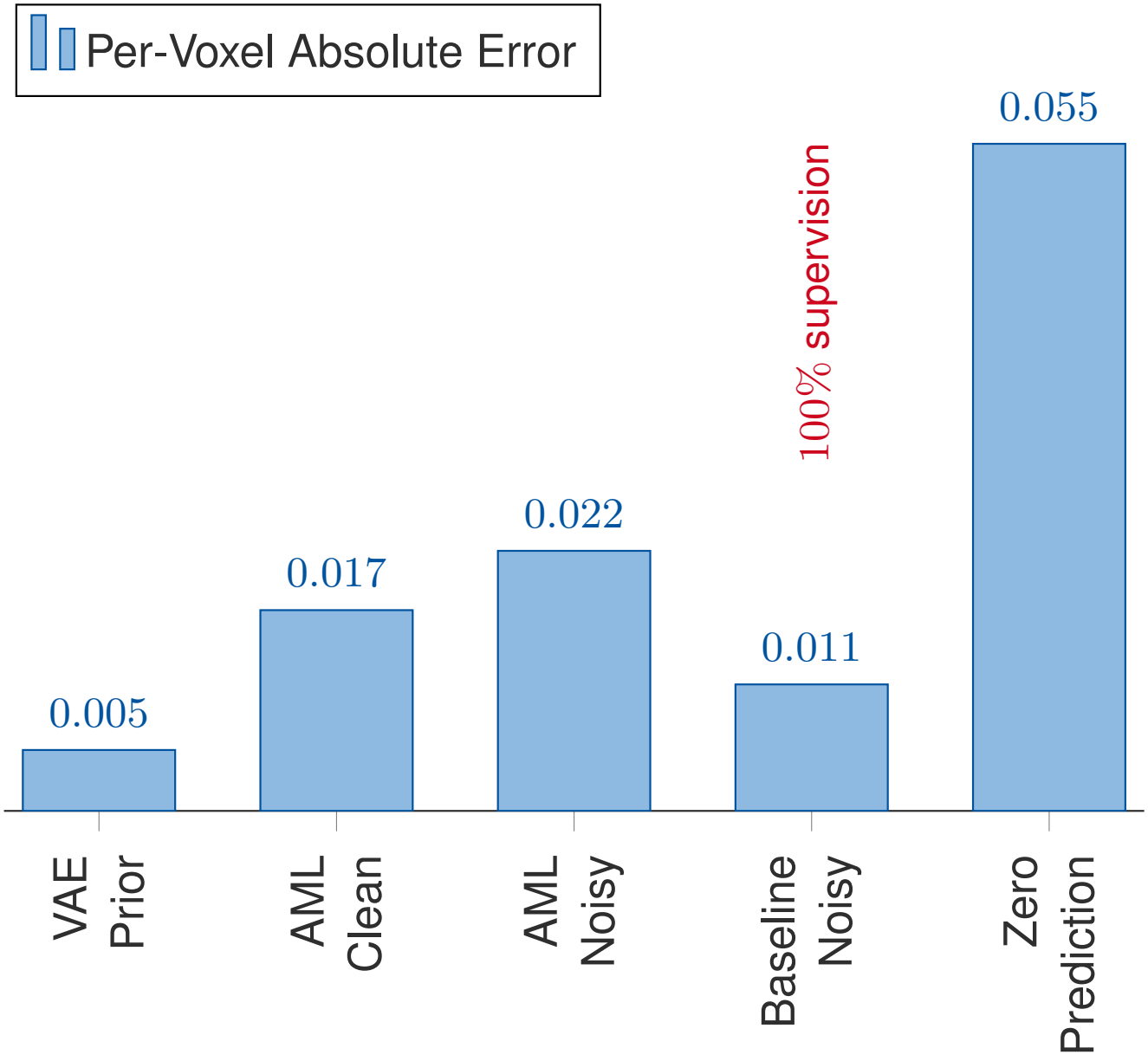
Results on ShapeNet



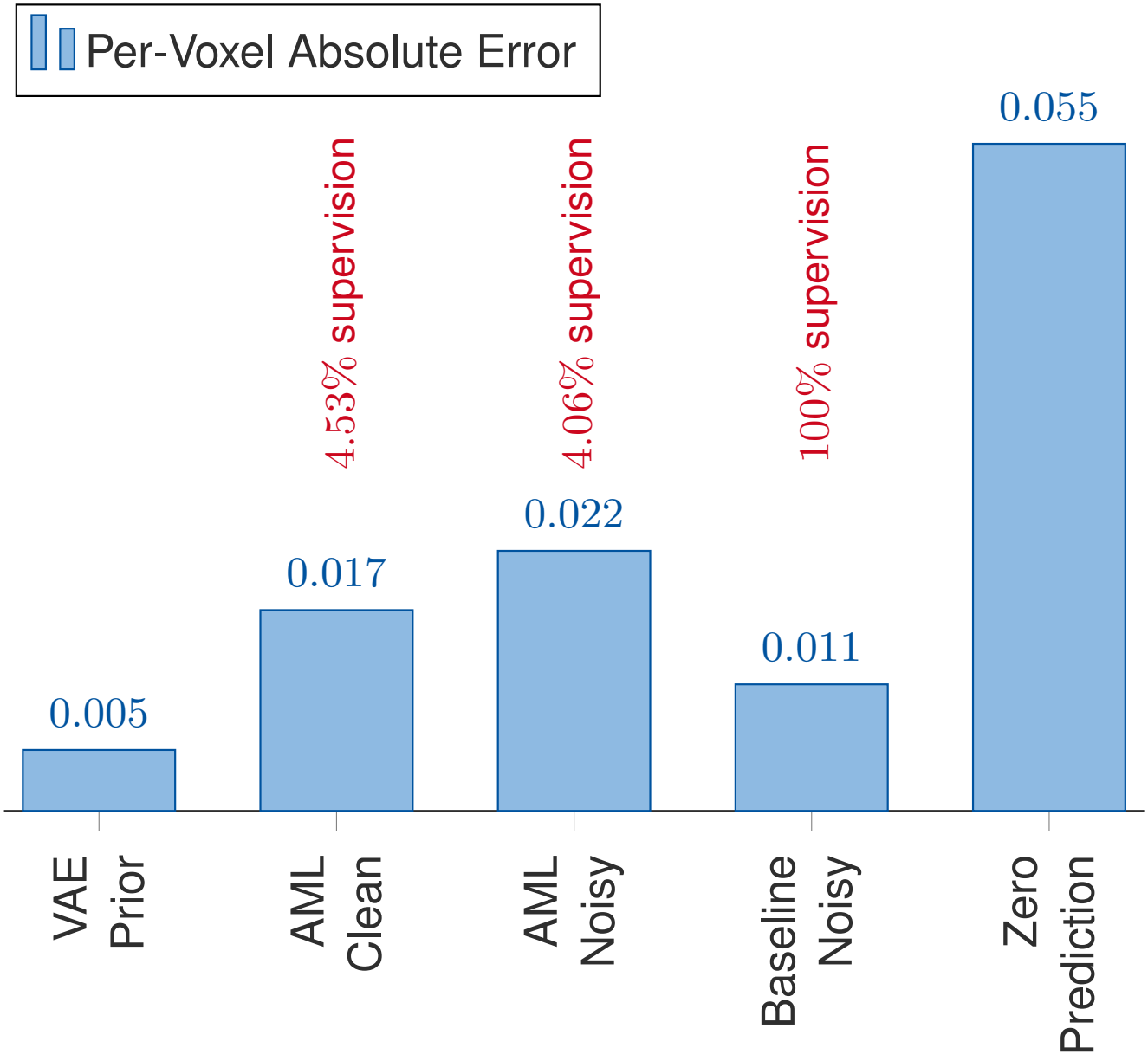
Results on ShapeNet



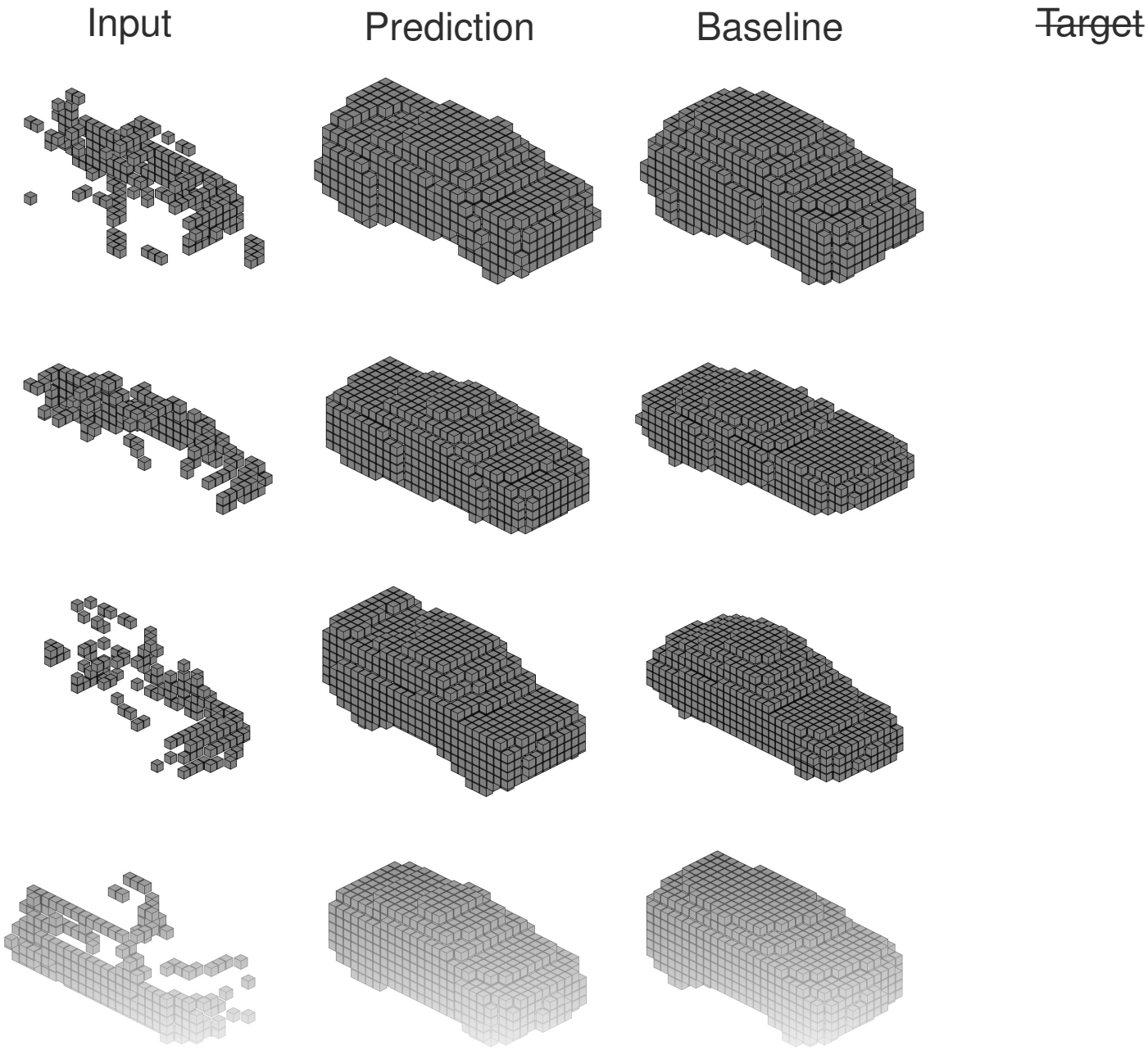
Results on ShapeNet



Results on ShapeNet



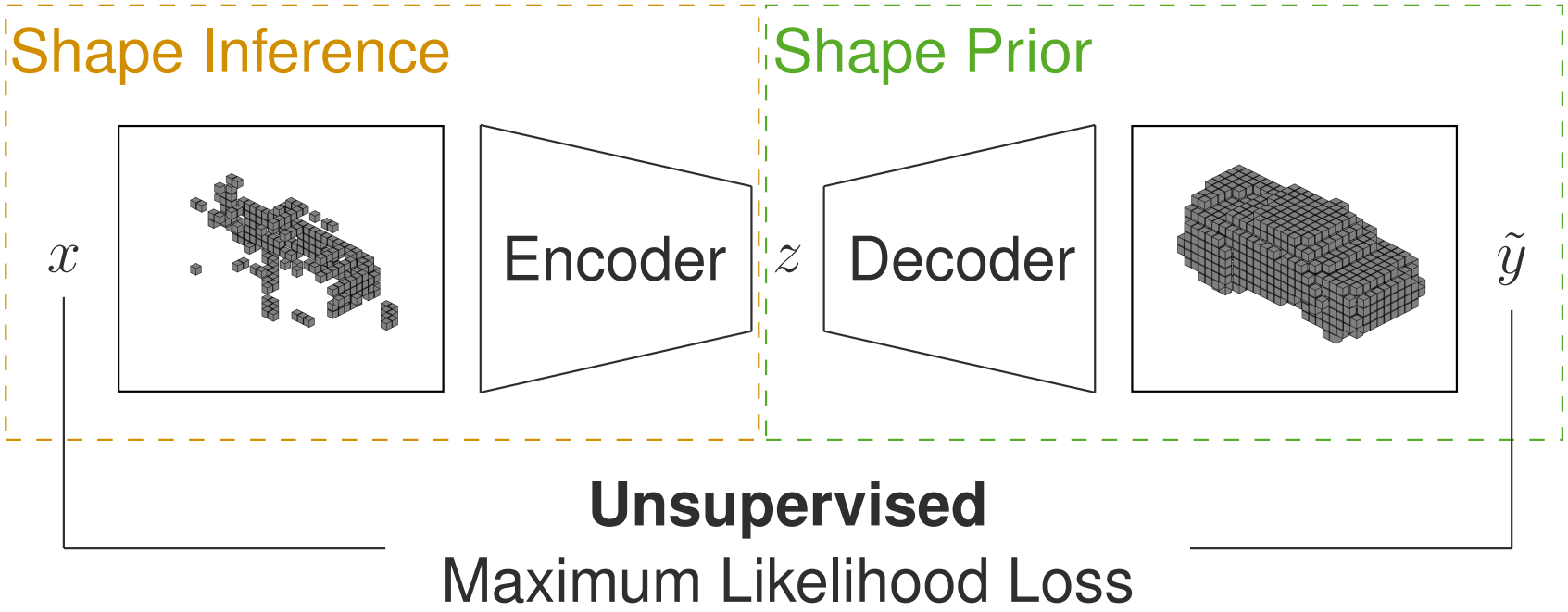
Results on KITTI



Conclusion

Summary

We proposed an amortized maximum likelihood framework for **learning** shape completion in a **weakly-supervised** setting.



Conclusion

Our experiments suggest that:

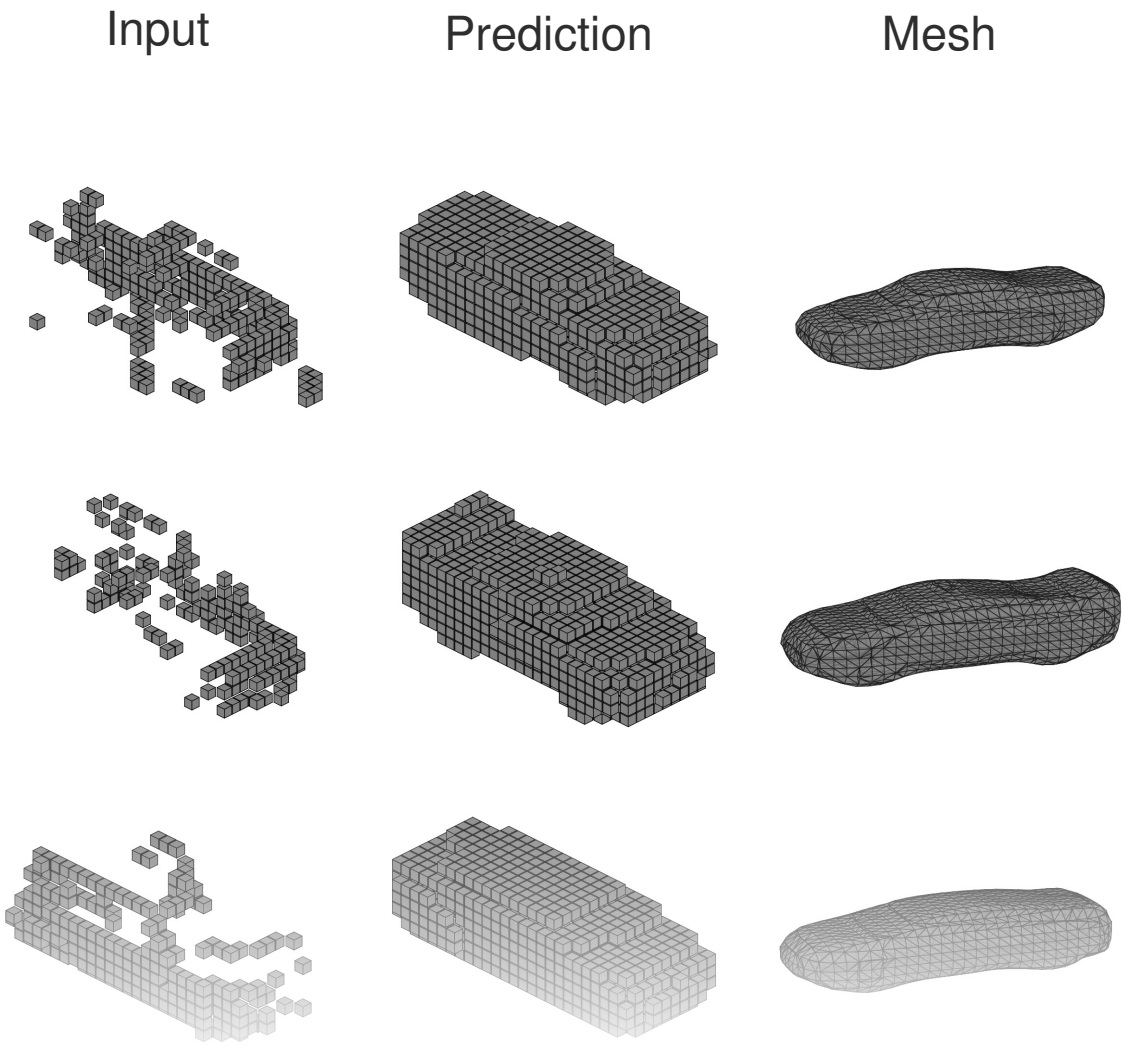
Hypothesis

Strong shape priors allow to learn shape completion under weak supervision.

- Additionally, inference involves “a simple forward pass”.

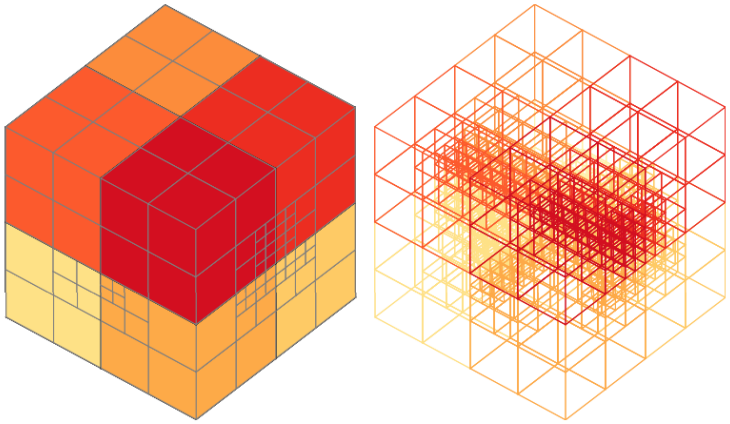
Future Work

Improve proposed amortized maximum likelihood framework on signed distance functions.

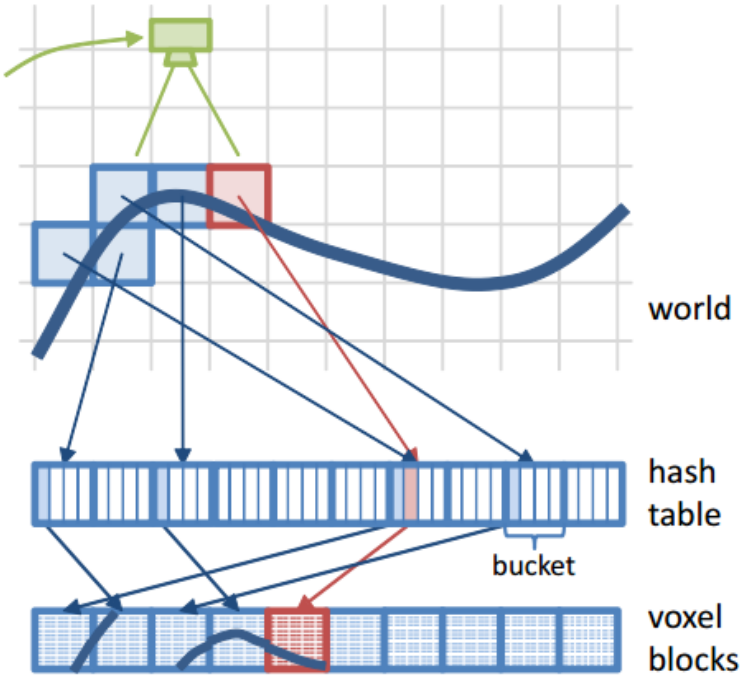


Future Work

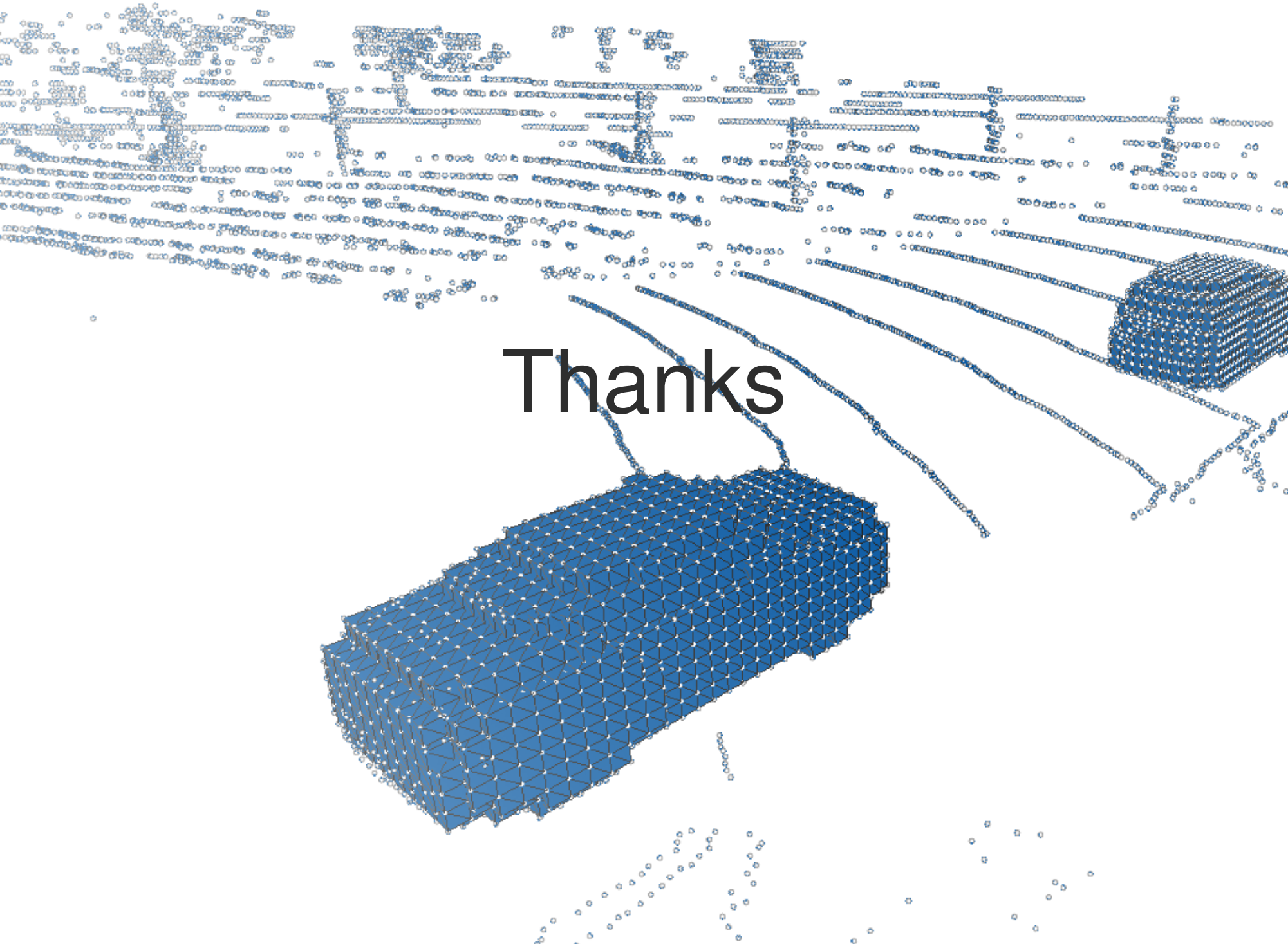
Increase resolution to learn details.



Input Reconstructed 3D point cloud



[Riegler et al., 2016] [Nießner et al., 2013]
[Fan et al., 2016]

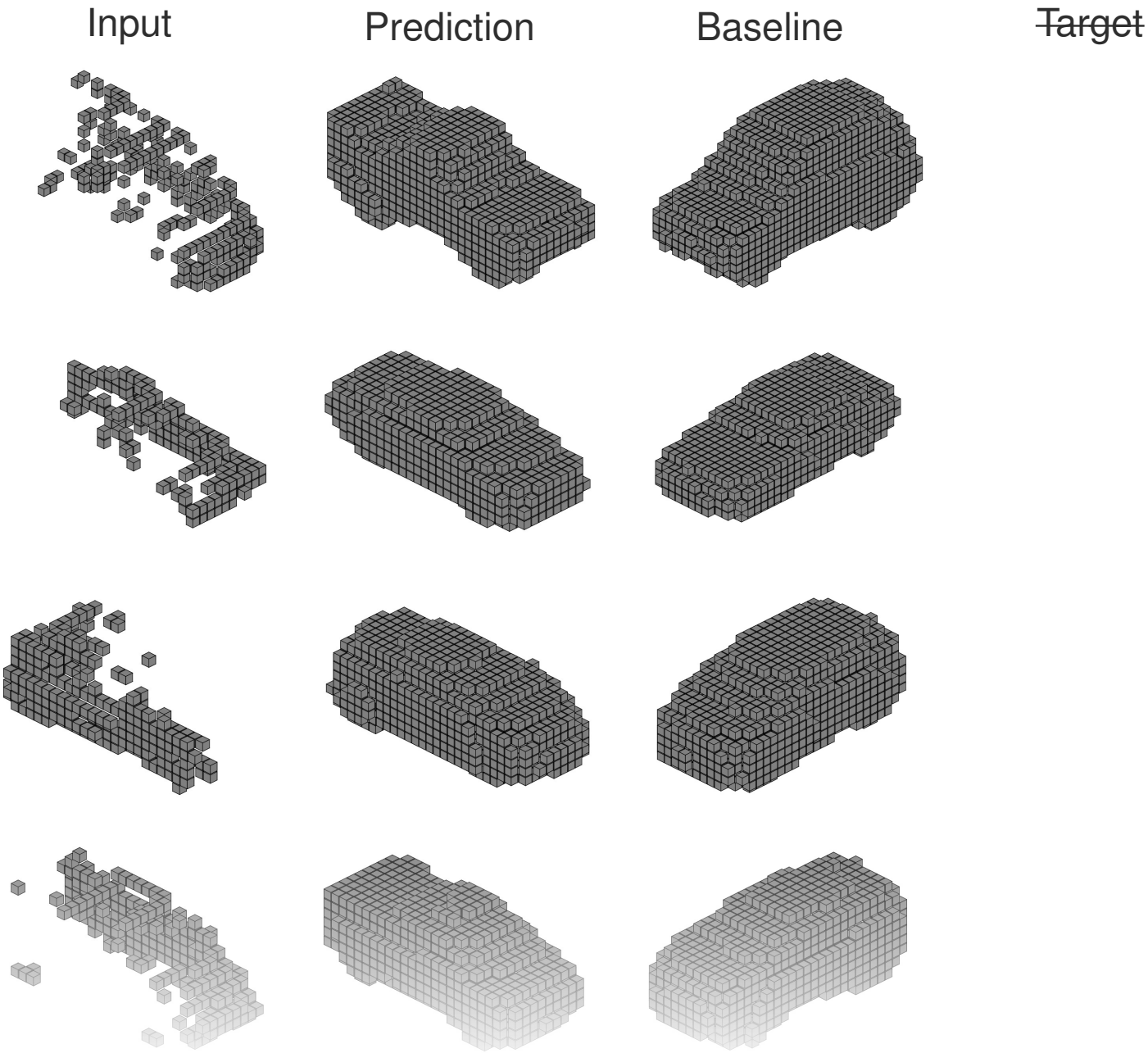

Thanks

Appendix – Related Work

Rough categorization following [Sung et al., 2015]:

- symmetry based approaches, *e.g.* [Thrun and Wegbreit, 2005, Pauly et al., 2008, Zheng et al., 2010, Kroemer et al., 2012];
- data-driven approaches, *e.g.* [Pauly et al., 2005, Li et al., 2015, Nan et al., 2012, Gupta et al., 2015, Dame et al., 2013, Engelmann et al., 2016];
- and recently learning-based approaches, *e.g.* [Firman et al., 2016, Smith and Meger, 2017, Dai et al., 2016, Sharma et al., 2016, Rezende et al., 2016, Fan et al., 2016].

Appendix – Results on KITTI



References I

Chang, A. X., Funkhouser, T. A., Guibas, L. J., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F. (2015).

Shapenet: An information-rich 3d model repository.

CoRR, abs/1512.03012.

Dai, A., Qi, C. R., and Nießner, M. (2016).

Shape completion using 3d-encoder-predictor cnns and shape synthesis.

CoRR, abs/1612.00101.

Dame, A., Prisacariu, V. A., Ren, C. Y., and Reid, I. D. (2013).

Dense reconstruction using 3d object shape priors.

In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 1288–1295.

References II

Engelmann, F., Stückler, J., and Leibe, B. (2016).

Joint object pose estimation and shape reconstruction in urban street scenes using 3d shape priors.

In *Pattern Recognition - 38th German Conference, GCPR 2016, Hannover, Germany, September 12-15, 2016, Proceedings*, pages 219–230.

Fan, H., Su, H., and Guibas, L. J. (2016).

A point set generation network for 3d object reconstruction from a single image.

CoRR, abs/1612.00603.

Firman, M., Mac Aodha, O., Julier, S. J., and Brostow, G. J. (2016).

Structured prediction of unobserved voxels from a single depth image.

In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5431–5440.

References III

Geiger, A., Lenz, P., and Urtasun, R. (2012).

Are we ready for autonomous driving? the KITTI vision benchmark suite.

In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3354–3361.

Gupta, S., Arbeláez, P. A., Girshick, R. B., and Malik, J. (2015).

Aligning 3d models to RGB-D images of cluttered scenes.

In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4731–4740.

Kingma, D. P. and Welling, M. (2013).

Auto-encoding variational bayes.

CoRR, abs/1312.6114.

References IV

Kroemer, O., Amor, H. B., Ewerton, M., and Peters, J. (2012).

Point cloud completion using extrusions.

In *12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012), Osaka, Japan, November 29 - Dec. 1, 2012*, pages 680–685.

Li, Y., Dai, A., Guibas, L. J., and Nießner, M. (2015).

Database-assisted object retrieval for real-time 3d reconstruction.

Comput. Graph. Forum, 34(2):435–446.

Nan, L., Xie, K., and Sharf, A. (2012).

A *search-classify* approach for cluttered indoor scene understanding.

ACM Trans. Graph., 31(6):137:1–137:10.

References V

Nießner, M., Zollhöfer, M., Izadi, S., and Stamminger, M. (2013).

Real-time 3d reconstruction at scale using voxel hashing.

ACM Trans. Graph., 32(6):169:1–169:11.

Pauly, M., Mitra, N. J., Giesen, J., Gross, M. H., and Guibas, L. J. (2005).

Example-based 3d scan completion.

In *Third Eurographics Symposium on Geometry Processing, Vienna, Austria, July 4-6, 2005*, pages 23–32.

Pauly, M., Mitra, N. J., Wallner, J., Pottmann, H., and Guibas, L. J. (2008).

Discovering structural regularity in 3d geometry.

ACM Trans. Graph., 27(3):43:1–43:11.

References VI

Rezende, D. J., Eslami, S. M. A., Mohamed, S., Battaglia, P., Jaderberg, M., and Heess, N. (2016).

Unsupervised learning of 3d structure from images.

In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4997–5005.

Riegler, G., Ulusoy, A. O., and Geiger, A. (2016).

Octnet: Learning deep 3d representations at high resolutions.

CoRR, abs/1611.05009.

Sharma, A., Grau, O., and Fritz, M. (2016).

Vconv-dae: Deep volumetric shape learning without object labels.

In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, pages 236–250.

References VII

Smith, E. and Meger, D. (2017).

Improved adversarial systems for 3d object generation and reconstruction.

CoRR, abs/1707.09557.

Sung, M., Kim, V. G., Angst, R., and Guibas, L. J. (2015).

Data-driven structural priors for shape completion.

ACM Trans. Graph., 34(6):175:1–175:11.

Thrun, S. and Wegbreit, B. (2005).

Shape from symmetry.

In *10th IEEE International Conference on Computer Vision (ICCV 2005), 17-20 October 2005, Beijing, China*, pages 1824–1831.

References VIII

Wu, J., Zhang, C., Xue, T., Freeman, B., and Tenenbaum, J. (2016).

Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling.

In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 82–90.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. (2015).

3d shapenets: A deep representation for volumetric shapes.

In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1912–1920.

References IX

Zheng, Q., Sharf, A., Wan, G., Li, Y., Mitra, N. J., Cohen-Or, D., and Chen, B. (2010).

Non-local scan consolidation for 3d urban scenes.

ACM Trans. Graph., 29(4):94:1–94:9.