Learning 3D Shape Completion under Weak Supervision

David Stutz

Abstract

Obtaining complete 3D representations of relevant objects in the environment is essential for safe autonomous driving. To this end, we propose an efficient, learning-based but weakly-supervised approach for learning 3D shape completion of cars from sparse 3D point clouds. On ShapeNet [1] and KITTI [2], we demonstrate that our approach outperforms related approaches [3, 4] while requiring less supervision or runtime.

1. Introduction

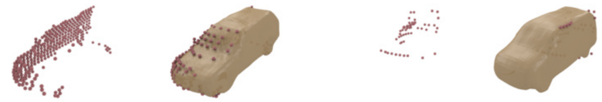
Autonomous vehicles, as in the case of KITTI [2], are commonly equipped with LiDAR scanners providing a 360 degree point cloud of the environment in real-time. This point cloud, however, is inherently incomplete: back and bottom of objects are typically occluded and the observations are sparse and noisy (Fig. 1, right). However, in order to make informed decisions (e.g., for path planning and navigation), it is of utmost importance to efficiently establish a representation of the environment which is as complete as possible. Thus, efficient 3D shape completion of point clouds for important objects such as cars is essential for safe autonomous driving.

2. Related Work

Recent approaches to 3D shape completion can be categorized into data-driven and learning-based methods. Data-driven approaches, e.g., [3], rely on learned shape priors and formulate shape completion as an optimization problem over the corresponding latent space. These approaches have demonstrated good performance on real data, e.g., on KITTI [2], but are often slow in practice. Learning-based approaches, e.g., [4], learn shape completion end-to-end on synthetic data, e.g., on ShapeNet [1]. These approaches are efficient, however, require full supervision during training. Unfortunately, even multiple, aggregated views will not be complete due to occlusion and sparse sampling of views.

3. Contributions

We propose an amortized maximum likelihood approach for 3D shape completion avoiding slow optimization and full supervision. We use a variational auto-encoder [5] to learn a low-dimensional, latent shape space (Fig. 2, left). Then, 3D shape completion can be formulated as maximum likelihood fitting over the learned latent shape space – without requiring full supervision. To avoid expensive optimization, we amortize, i.e. *learn*, the maximum likelihood problem by training a new encoder (Fig. 2, right). On ShapeNet [1] and KITTI [2], we compare our approach to state-of-the-art data-driven and learning-based approaches [3, 4].



(a) ShapeNet [1] (Synthetic) (b) KITTI [2] (Real)

Figure 1: 3D shape completion results of cars (point cloud observations in red; completed meshes in beige). Learning shape completion on real data is challenging due to sparse and noisy observations and missing ground truth.

This paper summarizes the master thesis [6][1]; the results have also been presented at CVPR'18 [7] and submitted to an IJCV special issue on robotic vision [8].

4. Method

Problem: We tackle a weakly-supervised formulation of 3D shape completion: Given (incomplete) observations $\mathcal{X} = \{x_n\}_{n=1}^N$ (Fig. 2, top right) and reference shapes $\mathcal{Y} = \{y_m\}_{m=1}^M$ (Fig. 2, top left) both of the same, known object category, learn a mapping $x_n \mapsto \tilde{y}(x_n)$ such that the predicted shape $\tilde{y}(x_n)$ matches the observation x_n while being plausible considering the set of reference shapes $\mathcal{Y}$. Here, $y_m \in \mathbb{R}^{H \times W \times D}$ are occupancy grids, i.e., voxel $y_{m,i} = 1$ iff the voxel lies on or inside the shape's surface, or signed distance functions (SDFs), i.e., voxel $y_{m,i}$ holds the distance to the surface and its sign indicates inside/outside allowing to derive sub-voxel accurate meshes. For the observations, we write $x_n \in \{0, 1, \bot\}^{H \times W \times D}$ to make missing information explicit: $x_{n,i} = \bot$ corresponds to unobserved voxels, while $x_{n,i} = 1$ and $x_{n,i} = 0$ correspond to occupied and unoccupied voxels (i.e., observed points and free space), respectively.

Shape Prior: We use the reference shapes $\mathcal{Y}$ to learn a prior of 3D shapes over a low-dimensional latent space $\mathcal{Z} =$

[1] Submitted September 13, 2017 at RWTH Aachen University; supervised by Prof. Andreas Geiger, Max Planck Institute for Intelligent Systems and University of Tübingen, and Prof. Bastian Leibe, RWTH Aachen University. Code and data available at davidstutz.de/mint2018.

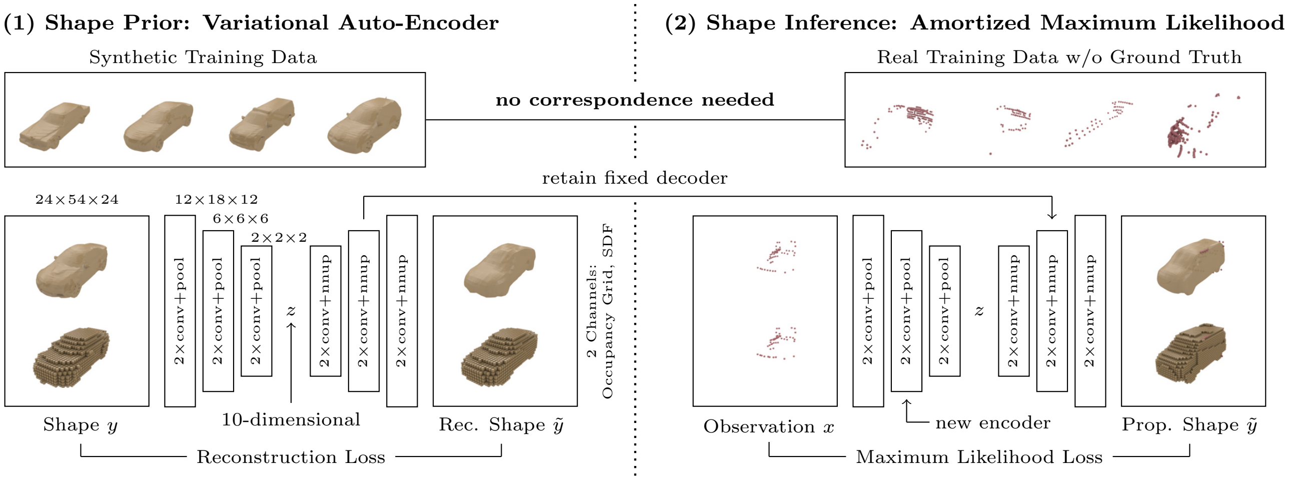


Figure 2: Step 1: a variational auto-encoder (VAE) [5] is trained on cars from ShapeNet [1]. Step 2: the VAE's decoder is fixed and retained in order to train a new deterministic encoder on KITTI [2]. The fixed decoder constrains the predictions to valid cars while the maximum likelihood loss aligns the predictions with the observations.

$\mathbb{R}^Q$. The prior is learned using a variational auto-encoder (VAE) [5] where the joint distribution $p(y, z)$ decomposes into $p(y|z)p(z)$ with $p(z) = \mathcal{N}(z; 0, I_Q)$ being a standard Gaussian. For training the generative model $p(y|z)$, we also need to approximate the posterior $q(z|y) \approx p(z|y)$. Then, recognition model $q(z|y)$ and generative model $p(y|z) = \prod_i p(y_i|z)$ take the following form:

$$q(z|y) = \mathcal{N}(z; \mu(y), \text{diag}(\sigma^2(y))) \tag{1}$$

$$p(y_i|z) = \begin{cases} \text{Ber}(y_i; \theta_i(z)) & \text{for occupancy grids} \\ \mathcal{N}(y_i; \mu_i(z), \sigma^2) & \text{for SDFs} \end{cases} \tag{2}$$

where $\mu(y), \sigma^2(y) \in \mathbb{R}^Q$ and $\mu_i(z)$ or $\theta_i(z)$ are predicted using the encoder and decoder, respectively – both implemented as 3D convolutional neural networks (cf. Fig. 2).

The VAE is trained by minimizing the following loss:

$$\mathcal{L}_{\text{VAE}}(w) = -\mathbb{E}_{q(z|y)}[\ln p(y|z)] + \text{KL}(q(z|y)|p(z)). \tag{3}$$

where w are the weights of the encoder and decoder (hidden in $q(z|y)$ and $p(y|z)$). The Kullback-Leibler divergence KL is computed analytically and the negative log-likelihood $-\ln p(y|z)$ corresponds to a cross-entropy error for occupancy grids or a sum-of-squared error for SDFs. In practice, the expectation is computed using one sample $z \sim q(z|y)$ per iteration – we refer to [5] for details.

Shape Inference: After learning the shape prior $p(y, z) = p(y|z)p(z)$, shape completion can be formulated as a maximum likelihood (ML) problem over the low-dimensional latent space $\mathcal{Z}$. The corresponding negative log-likelihood $-\ln p(y, z)$ to be minimized is

$$\mathcal{L}_{\text{ML}}(z) = -\sum_{x_i \neq \perp} \ln p(y_i = x_i|z) - \ln p(z). \tag{4}$$

As the prior $p(z)$ is Gaussian, the negative log-probability $-\ln p(z)$ is a regularizer proportional to $\|z\|_2^2$ and ensures that high-probability shapes are favored. As before, the generative model $p(y|z)$ decomposes over voxels; here, we can only consider actually observed voxels $x_i \neq \perp$. Instead of solving Eq. (4) for each observation independently, we train a new encoder $z(x; w)$ to *learn*, i.e., amortize, ML – resulting in an amortized ML (AML) approach. To this end, we keep the generative model $p(y|z)$ fixed and train the weights w of the encoder $z(x; w)$ using the ML objective:

$$\begin{aligned} \mathcal{L}_{\text{AML}}(w) = -&\sum_{x_i \neq \perp} \ln p(y_i = x_i|z(x; w)) \\ &- \lambda \ln p(z(x; w)) \end{aligned} \tag{5}$$

where λ controls the importance of the shape prior. For both shape representations, the loss results in a cross-entropy error (for SDFs, a reparameterization is used, cf. [8]).

5. Experiments

Architecture: Encoder and decoder (Fig. 2) consist of three stages, each comprising two convolutional layers (including batch normalization and ReLU activations) and max pooling/nearest neighbor upsampling. Our latent space is $Q = 10$-dimensional; we use $\log \sigma^2 = -2$ in Eq. (2). Occupancy grids and SDFs are provided in two separate channels (and occupancy grids are predicted using Sigmoid activations). For data augmentation, we apply slight random rotations, scalings and translations.

Data: On ShapeNet, we generated 500 car shapes for training the shape prior, 5000 observations with noise for training the inference model (from separate 500 shapes) and 1000 observations with ground truth shapes for testing. On KITTI, we extracted car observations from the

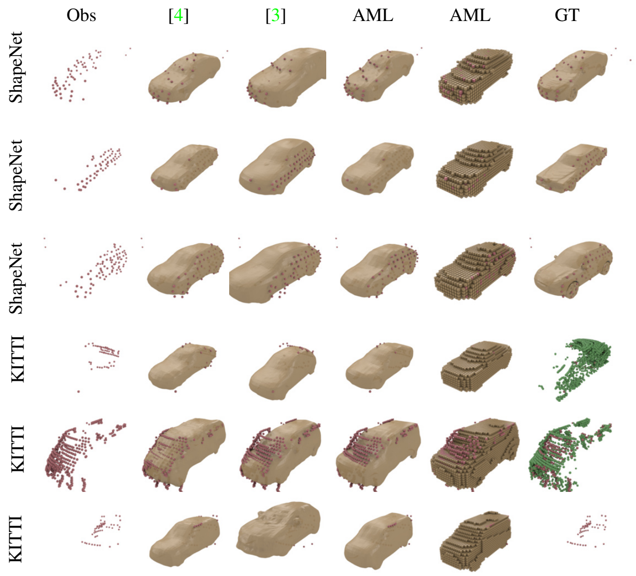


Figure 3: Qualitative results on ShapeNet [1] and KITTI [2] comparing AML to [3] and [4] (observations in red and completed shapes in beige). AML visually outperforms both [3] and [4] while being more efficient and requiring less supervision.

Velodyne point clouds using the ground truth 3D bounding boxes. Additionally, we constructed partial ground truth (green in Fig. 3) from multiple views. Overall, we obtained 8442 / 9194 observations for training / testing. All observations and shapes are voxelized at a resolution of $H \times W \times D = 24 \times 54 \times 24$.

Baselines: We consider maximum likelihood (ML) (solving Eq. (4) iteratively), the data-driven approach by Engelmann et al. [3] and the learning-based approach by Dai et al. [4] as baselines.

Evaluation: On ShapeNet, we compute the distance from the reconstructed mesh to the ground truth mesh (accuracy) and vice-versa (completeness) in voxels [vx]. On KITTI, we compute the average distance of the partial ground truth to the reconstructed mesh (completeness) in meters [m]. In all cases, *lower is better*.

Results: On ShapeNet (Fig. 4, left), our approach clearly outperforms [3] as well as our ML baseline, illustrating that *learning* 3D shape completion is beneficial. Requiring only 2ms per observation, our approach is roughly 84 times faster than [3]. We also demonstrate comparable performance to [4] while using only 3.86% supervision. On KITTI (Fig. 4, right), quantitative evaluation is difficult due to the incomplete ground truth; still, our approach slightly outperforms [3] and [4]. Here, compared to [4], our approach requires only 6.79% supervision. Qualitatively (Fig. 3), our approach outperforms both [4] and [3], especially on KITTI where [3] tends to overfit to noise and [4] (trained on ShapeNet) is unable to generalize.

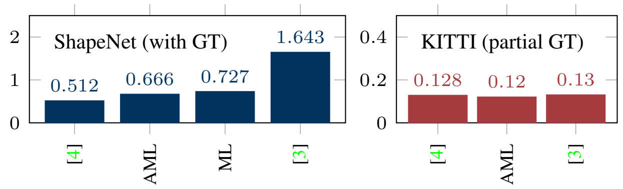


Figure 4: ▮▮ Average of accuracy and completeness in [vx] on ShapeNet and ▮▮ completeness in [m] on KITTI. On ShapeNet, AML outperforms ML and [3] and is able to compete with [4]. On KITTI, AML slightly outperforms both [3] and [4], while being faster and requiring less supervision.

We refer to [8] for more details and experiments.

6. Conclusion

In this work, we presented a novel, learning-based but weakly-supervised approach to 3D shape completion from sparse and noisy point clouds. We learn a variational auto-encoder [5] shape prior on synthetic data and formulate 3D shape completion as maximum likelihood problem which can be amortized, i.e., *learned*. On ShapeNet [1] and KITTI [2], we demonstrated that our approach outperforms related data-driven approaches [3], while being 84 times faster, and can compete with learning-based approaches [4], while requiring 93% less supervision.

References

[1] A. X. Chang et al. “ShapeNet: An Information-Rich 3D Model Repository”. In: *arXiv.org* 1512.03012 (2015).

[2] A. Geiger et al. “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite”. In: *CVPR*. 2012.

[3] F. Engelmann et al. “Joint Object Pose Estimation and Shape Reconstruction in Urban Street Scenes Using 3D Shape Priors”. In: *GCPR*. 2016.

[4] A. Dai et al. “Shape Completion using 3D-Encoder-Predictor CNNs and Shape Synthesis”. In: *CVPR*. 2017.

[5] D. P. Kingma and M. Welling. “Auto-Encoding Variational Bayes”. In: *ICLR* (2014).

[6] D. Stutz. “Learning Shape Completion from Bounding Boxes with CAD Shape Priors”. MA thesis. Aachen, Germany: RWTH Aachen University, 2017.

[7] D. Stutz and A. Geiger. “Learning 3D Shape Completion from Laser Scan Data with Weak Supervision”. In: *CVPR*. 2018.

[8] D. Stutz and A. Geiger. “Learning 3D Shape Completion under Weak Supervision”. In: *arXiv.org* abs/1805.07290 (2018).