AttentionSDF: Locally-conditioned Shape Representation with Cross-Attention

Kristen Zhang, Annie Feng

dxzhang1@mit.edu,azf@mit.edu

Abstract

We propose AttentionSDF, a model to represent 3D shapes and predict signed distance fields from point cloud data. Our method modifies DeepSDF with three techniques: locally-conditioned latent codes, positional encoding, and cross-attention layers. From the ablation study on the positional encoding hyperparameter and the ablation study on cross-attention inputs, we chose the optimal settings for the final model, which shows ***three-fold*** *improvement compared to our baseline, using Chamfer distance as a metric. The code for this project is located at: https://github.com/Kristen-Z/Attention_sdf.*

1. Introduction

3D data are recently becoming more feasible to work with due to some advances in shape representation, computing power, and deep learning. For example challenges, handling input shape data (meshes, point-based, voxels) for 3D reconstruction are typically expensive, and deep convolutional networks grow quickly in computational complexity when generalized to 3D. Since Bengio's paper on word embeddings [1], learning representations with semantic meaning has been a popular goal for many domains, such as images and text; models like CLIP and word2vec are two examples. For 3D shapes, DeepSDF demonstrates a method to create a compact latent vector that effectively captures shape information of common objects without explicitly denoting the geometry [15]. DeepSDF uses a probablistic auto-decoder in a method to learn embeddings of 3D shapes, which can be used in applications such as latent shape space interpolation, shape completion, and 3D reconstruction.

Based on the concept of signed distance function, we proposed AttentionSDF, a memory efficient and expressive method to represent shapes and reconstruct meshes from point clouds. We make several improvements to the DeepSDF model and evaluate on a 3D reconstruction task.

2. Related Works

We will review 3 main areas related to our work: different ways to represent shapes (Sec. 2.1), techniques for shape representation learning (Sec. 2.2) and current locally conditioning mechanism for 3D reconstruction task (Sec. 2.3).

2.1. 3D representations

Point-based method is a sparse 3D representation that represents objects using a set of individual points that are connected by edges or other relationships. These representations can be generated from various sources, such as 3D scanning devices and depth sensors. PointNet [17] uses max-pool strategies to extract global shape features from point clouds and then is widely used as an encoder for point learning networks [16]. However, it requires non-trivial steps to generate watertight surfaces due to a large number of points and the lack of topology in point clouds.

Mesh-based representation is a method for representing 3D objects as a collection of connected polygons or triangles that form a surface. In this representation, the surface of the 3D object is discretized into a set of vertices, edges, and faces, forming a mesh structure. Mesh is a preferred representation for many uses, such as visualizations and simulations. But they are harder to directly produce from a neural network (vertex regression and face construction) [14].

Voxel-based method represents 3D shapes as a grid of 3D pixels, called voxels. Each voxel represents a small volume element of the 3D space whose value can be binary (occupied or not occupied) or continuous (e.g., representing color or material properties). Voxels have been the most natural way to represent 3D objects [7, 16]. However, it can be computationally expensive due to the high resolution and large number of voxels required to represent detailed shapes.

Implicit representations don't rely on any explicit geometry. Instead, it represents the surface of the object as the zero level-set of a continuous function. There has been growing interest in exploring 3D representations in implicit

fields as they provide a continuous and topology-agnostic means of representing a 3D surface/volume and can be estimated by a neural network. These functions to represent 3D objects implicitly describe the occupancy of a given 3D point [2, 12, 16] or its distance to the surface signed [15] or unsigned. We used the signed distance function(SDF) that maps any point in R^3 to its distance to the represented surface, signed based on whether the point is inside of it or outside.

2.2. Implicit representations learning techniques

In 2019, three works [5, 12, 15] simultaneously proposed to parametrize implicit fields using deep neural networks. Since then, there has been a surge of interest in exploring how to do it more accurately and efficiently fit implicit functions. Various methods have been proposed to achieve the generalization of neural fields, such as latent vector conditioned MLPs [15], hyper-networks as introduced by Sitzmann et al. [21], and set-latent transformer representations [20]. Recently Rebain et. al [18] run many experiments modeling signals with neural field employing these 3 conditioning strategies and find that attention-based conditioning outperforms other approaches in a variety of settings while being memory efficient as the latent code dimension increases. However, Rebain et al.'s work primarily focuses on high-dimensional conditioning variables in the context of neural fields. In our research, we extend this conclusion to the domain of implicit fields and mesh reconstruction tasks. By incorporating this approach, we aim to enhance the representation and reconstruction capabilities of our model.

2.3. Locally conditioning mechanism for scene representation

Traditionally, a single latent code is commonly used to represent an object. However, when dealing with complex scenes, the exponential number of possible combinations of object parts makes it insufficient to rely on a single latent code [16]. Consequently, researchers have begun exploring the use of locally-conditioned latent codes, where each point in the scene is assigned a unique code based on its location. This approach allows for more fine-grained representation and encoding of the intricate details and variations within the scene. However, typically locally-conditioned codes [3, 6, 10, 16] were stored in a grid which is memory inefficient. While the state-of-the-art model for 3D reconstruction tasks in ShapeNet [4]—POCO [2]—introduced an attention-based approach combined with element-wise locally-conditioned latent codes, these codes often lack semantic meaning and are not editable. However, the effectiveness of combining attention mechanisms and local-conditioning mechanisms in 3D reconstruction tasks serves as the foundation of our project. We propose an innovative method to more effectively achieve local conditioning with rich semantic meaning. By leveraging this approach, we aim to enhance the interpretability and manipulability of the reconstructed 3D shapes while maintaining high accuracy.

3. Methodology

Our hypothesis is that the addition of three methods to DeepSDF's model will improve upon the baseline on the 3D reconstruction task: cross-attention, positional-encoding, and locally-conditioned latent codes.

3.1. Dataset and Preprocessing

The model will be evaluated on the chairs, lamps, and planes classes in ShapeNetv2 [4]. We use DeepSDF's preprocessing module and selected subset of ShapeNetv2 for evaluation of our model. The preprocessing module samples from the ShapeNet meshes to get xyz-coordinates of the point cloud, and its SDF values. For our task, we disregard the SDF values and only take the xyz-coordinates as input. For an overview of our method, see Figure 1.

We train on 1780 planes, 3281 chairs, and 897 lamps. The testing set has 456 planes, 832 chairs, and 213 lamps.

3.2. Model Architecture

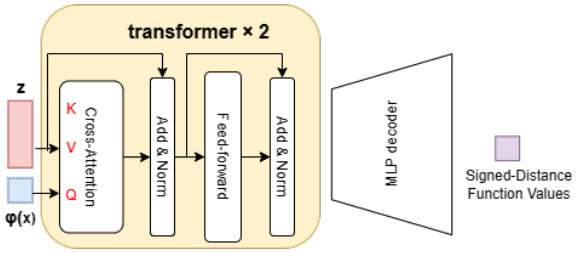

Figure 2. Overview of our model. Our model consists of a cross attention transformer and a decoder. x is the coordinate of a point, and the function of coordinates, denoted as $\phi(x)$, serves as the positional encoding function. The latent code for an object is denoted as z. Given the coordinate of a point and initialized latent code, the transformer takes the concatenation of z and x as inputs. The outputs of transformer will then be decoded by the MLP to attain the corresponding SDF value.

In the context of attention mechanisms, there are three key inputs: query, key, and value, as described in the Transformer model [22]. In self-attention, the same embeddings are used as inputs for query, key, and value. On the other hand, in cross-attention, multi-modality features are employed as inputs. However, the specific implementation of cross-attention can vary significantly in different contexts, as demonstrated in various works [9, 11, 19]. Therefore, in our research, we conducted practical experiments to determine the most suitable cross-attention mechanism for the

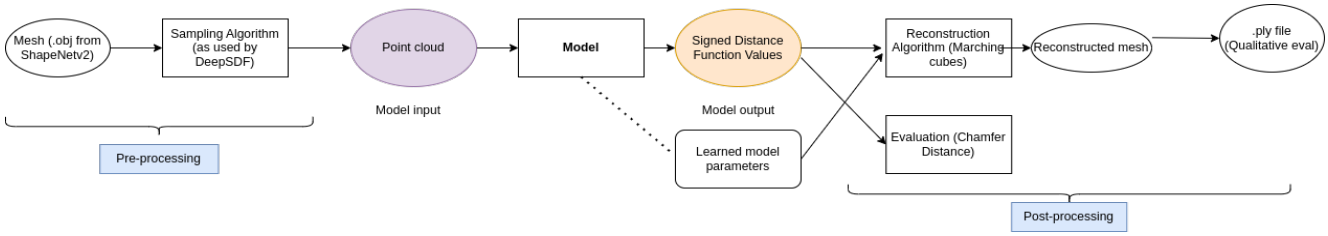

Figure 1. Overview of pipeline. This is where the model, and model inputs and outputs fit into the pre-processing and post-processing steps. Note that in addition to the point cloud coordinates input, the model initializes latent vectors for each set of points and updates them in back-propagation.

reconstruction task. As shown in Table 1, various variables can be employed as the query input. In our experiments, we observed that the concatenation of latent codes and coordinates outperforms the other two structures. This approach not only achieves superior performance but also demonstrates the ability to capture complex topology features. We suppose that using a single modality signal as the input may present challenges for the MLP in learning a comprehensive representation.

3.3. Positional Encoding

In transformers, positional encodings give the model a notion of order by associating input tokens with their indices; however, the positional encoding we use is a mapping from continuous coordinates to a higher dimensional space so that the network can approximate a higher-frequency function [13]. This is described in detail by the NERF paper, which uses the encoding function:

$$\gamma(p) = \sin(2^0\pi p), \cos(2^0\pi p), \cdots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p) \tag{1}$$

In both NERF and our implementation, we apply the encoding function element-wise to the normalized xyz-coordinates.

In addition, we did ablation study about the effect of varying hyperparameter L in the encoding function. Since L controls the frequency (increases with L), intuitively, a higher frequency will let the reconstruction have finer details, but may be too complex and introduce noise. On the other hand, a lower frequency may not capture the important features in the data. As explained in Figure 3, the best experimental value of L was L=3.

3.4. Locally-conditioned representation

Unlike POCO [2], our approach does not rely on a pre-trained backbone to generate point-level latent codes, which may lack semantic meaning. Instead, we employ a separate MLP that is specific to each shape to compute the latent code at each input point. This strategy enables the network to learn local features and capture the semantics of the shapes more effectively. Furthermore, the weight-encoded MLP itself serves as a memory-efficient and effective representation for the shape. By encoding the shape information in the weights of the MLP, we can capture important features while reducing the overall memory requirements. This approach allows for efficient storage and retrieval of shape information during the reconstruction process.

3.5. Evaluation Metric

To evaluate the reconstruction, we use the Chamfer distance [8] to calculate the difference between two sets of points:

$$\begin{aligned} d_{CD}(S_1, S_2) = &\frac{1}{S_1}\sum_{x\in S_1} min_{y\in S_2}||x-y||_2^2 \\ &+\frac{1}{S_2}\sum_{y\in S_2} min_{x\in S_1}||y-x||_2^2 \end{aligned} \tag{2}$$

Qualitatively, we use DeepSDF's code to recover the mesh from SDF and visualize the mesh to judge if the objects were adequately reconstructed.

4. Experiments

To evaluate the effectiveness of the three proposed changes (local-condition, attention, positional encoding), we ran experiments with different combinations of them.

In addition to baseline and other reference models, we trained models with combinations of these 3 changes for 2000 epochs on the planes dataset, and evaluated on the planes test set. The model types and results are in Table 2. The model using locally-conditioned latent codes and cross-attention performed the best overall. This is likely because the locally-conditioned model allows the network to learn more information about the shapes, by learning a latent code per object point compared to the other models that learn a latent code per object.

Based on the experimental results, we can conclude that the three proposed changes we made have significantly enhanced the model's performance in 3D reconstruction, as evaluated by the Chamfer Distance metric. Among these

	self attention	cross attention1	cross attention2
Query	cat[latent codes, coordinates]	coordinates	latent codes
Key,Value	cat[latent codes, coordinates]	cat[latent codes, coordinates]	cat[latent codes, coordinates]

Table 1. Experiments for cross attention structure. Different combinations of variables for input to the attention mechanism. When query, key, and value are the same input, called "self-attention"; generally called "cross-attention".

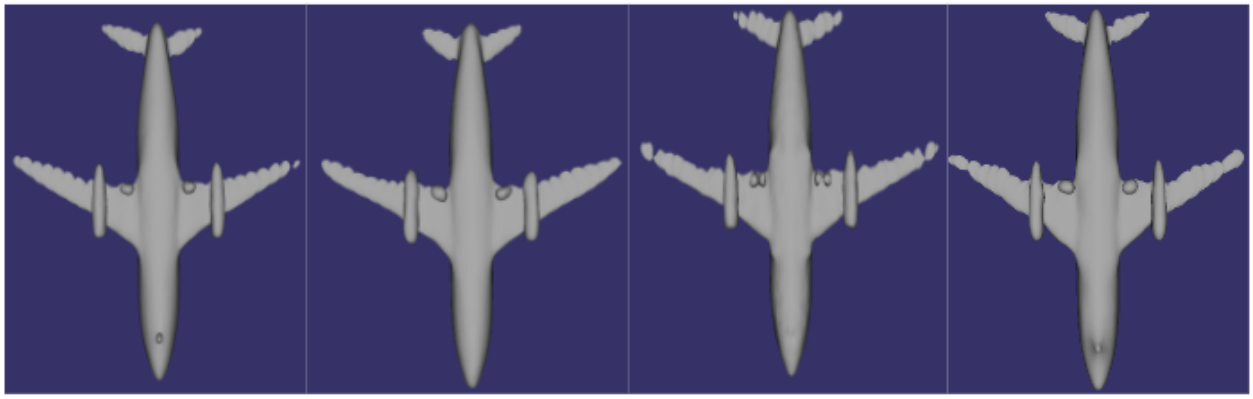

Figure 3. The mesh reconstruction and visualization (plane bottom view) for different values of L in equation (1). As L increases, more detailed characteristics appear on the plane: the tail gets sharper, and the turbines and wheels look more realistic. However, starting in L=7, the wings start to break apart. From the qualitative reconstructions, best value of L is determined to be 3.

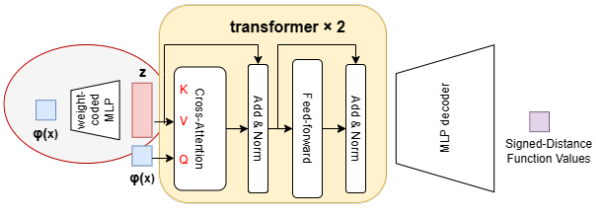

Figure 4. Locally-conditioned AttentionSDF model architecture: The weight-encoded MLP generates shape-specific latent codes at each point, while the transformer and Decoder compute the SDF values using these codes.

Model	CD (Mean)	CD (Median)
DeepSDF (Baseline) [15]	0.287	0.71
LIG [10]	0.150	-
ConvONet [16]	0.091	-
AttentionSDF (ours)	0.173	**0.025**
AttentionSDF with PE (ours)	0.143	0.028
Local AttentionSDF (ours)	**0.080**	0.026

Table 2. Comparison of Chamfer Distance metric of different models' performance on the planes test split. ConvONet and LIG are classical locally-conditioned models. Our best 3 models use self-attention (cross-attention where Q,K,V are all the same vector: a concatenation of latent codes and xyz coordinates). Our best overall model used local-conditioned method and self-attention in the transformer layers.

changes, the locally-conditioned AttentionSDF approach stands out with remarkable performance, achieving a three-fold improvement compared to the baseline method. This finding highlights the effectiveness and superiority of the locally-conditioned AttentionSDF approach in generating high-quality 3D reconstructions.

5. Conclusion and Discussion

The AttentionSDF demonstrates significant improvements compared to our baseline model DeepSDF and other classical 3D reconstruction methods. We have incorporated three major enhancements into our model, each of which has positively impacted its performance.

The attention mechanism and positional encoding both play important roles in improving the performance of models, but the attention mechanism tends to have a larger impact compared to positional encoding. While positional encoding helps capture the high frequency information, the attention mechanism enables the model to dynamically weigh and attend to different elements, enhancing its ability to capture important patterns and dependencies in the data. Particularly noteworthy is the introduction of the locally

Figure 5. Visual comparison of a reconstructed plane. From left: reconstruction of ground truth, DeepSDF baseline, AttentionSDF without local-conditioning, and AttentionSDF with local-conditioning.

Figure 6. Visual comparison of reconstructed chair. Left and right are two chair instances reconstructed by our AttentionSDF model.

conditioning mechanism, which has yielded remarkable results. Additionally, we have proposed a novel approach of employing a weight-encoded network to compute latent codes. This innovation will be particularly valuable in large scene reconstructions with complicated topology.

Due to time constraints, our model was only applied to a subset of the ShapeNet dataset, specifically focusing on planes, chairs, and lamps. The evaluation was performed using a single metric. In order to establish that our model is generalizable, it is necessary to extend the evaluation to a wider range of classes in the ShapeNet dataset and assess its performance using additional metrics such as IoU and F-Score. This comprehensive evaluation would provide a more thorough understanding of the model's capabilities across different object categories and enable a more robust assessment of its performance.

Our weight encoded MLP provides a powerful and expressive representation of shapes. Further research can focus on developing advanced techniques for shape generation and manipulation. This can include tasks such as interactive shape editing, shape completion, and deformation, allowing users to intuitively modify shapes while maintaining their semantic meaning. For large scene reconstructions, and real-time robotics applications, there is a search to find computationally efficient and sound methods to interpret and manipulate shape data in complex environments. We hope that this work is one step towards this goal.

6. Contributions

6.1. Kristen

1. Ran preprocessing of data
2. Implemented attention, locally-conditioned method, and positional encodings
3. Ran training experiments for the models, and analyzed logs and train plots. Get all the final results.
4. Implemented the model architecture
5. Helped write presentation and report

6.2. Annie

1. Download, upload, compress/uncompress data for the project
2. Search for computing resource (GCP) for the project, and prepare environment for training experiment and version control
3. Researched literature for context and motivation of project
4. Helped write the presentation and report, and create visuals and figures

References

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155, mar 2003. 1

[2] Alexandre Boulch and Renaud Marlet. POCO: point convolution for surface reconstruction. *CoRR*, abs/2201.01831, 2022. 2, 3

[3] Rohan Chabra, Jan Eric Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard A. Newcombe. Deep local shapes: Learning local SDF priors for detailed 3d reconstruction. *CoRR*, abs/2003.10983, 2020. 2

[4] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 2

[5] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *CoRR*, abs/1812.02822, 2018. 2

[6] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. *CoRR*, abs/2003.01456, 2020. 2

[7] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. *CoRR*, abs/1604.00449, 2016. 1

[8] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. *CoRR*, abs/1612.00603, 2016. 3

[9] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2

[10] Chiyu "Max" Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas A. Funkhouser. Local implicit grid representations for 3d scenes. *CoRR*, abs/2003.08981, 2020. 2, 4

[11] Hezheng Lin, Xing Cheng, Xiangyu Wu, and Dong Shen. Cat: Cross attention in vision transformer. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 2

[12] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. *CoRR*, abs/1812.03828, 2018. 2

[13] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *CoRR*, abs/2003.08934, 2020. 3

[14] Charlie Nash, Yaroslav Ganin, S. M. Ali Eslami, and Peter W. Battaglia. Polygen: An autoregressive generative model of 3d meshes. *CoRR*, abs/2002.10880, 2020. 1

[15] Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *CoRR*, abs/1901.05103, 2019. 1, 2, 4

[16] Songyou Peng, Michael Niemeyer, Lars M. Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. *CoRR*, abs/2003.04618, 2020. 1, 2, 4

[17] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016. 1

[18] Daniel Rebain, Mark J. Matthews, Kwang Moo Yi, Gopal Sharma, Dmitry Lagun, and Andrea Tagliasacchi. Attention beats concatenation for conditioning neural fields. *Transactions on Machine Learning Research*, 2023. 2

[19] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6229–6238, 2022. 2

[20] Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lucic, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas A. Funkhouser, and Andrea Tagliasacchi. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. *CoRR*, abs/2111.13152, 2021. 2

[21] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *CoRR*, abs/1906.01618, 2019. 2

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2