Fast Surface Reconstruction from Voxels by Learning Hierarchical Latent Code Sets

신지윤, 이정우

서울대학교

jyezs5150@snu.ac.kr, junglee@snu.ac.kr

Fast Surface Reconstruction from Voxels by Learning Hierarchical Latent Code Sets

Jiyoon Shin, Jungwoo Lee

Seoul National University

요 약

Neural implicit functions have proved successful in representing and reconstructing 3D shapes at arbitrary resolutions and with high fidelity. The functions are mainly learned by sparse explicit representations such as voxel grids, point clouds, or 2D views taken from several viewpoints. Unfortunately, between those, learning neural fields and reconstructing from discrete voxels remain limited due to the computational complexity it carries. Specifically, existing methods suffer from slow convergence; model training time is measured in days to weeks.
To overcome this issue, an easy-to-hard learning paradigm is introduced, leveraging the proposed curriculum transformer decoder to accumulate a set of hierarchical latent codes. The sparse voxelized shape is first encoded into different levels of feature grids, and the given query point is grid-sampled on each feature, creating a hierarchical latent code set. The latent codes are then separately included in the curriculum transformer decoder, beginning with latents of coarse-level features and gradually adding latents of more low-level features. Representations from all levels of features are included in the optimization phase for another curriculum strategy. Faster convergence and improved generalization are achieved by the idea of curriculum learning embedded; about $10 \times$ speedup is observed with on-par high-fidelity results. Experiments also verify that our method shows robustness against different shape categories, gains potential for being useful in the wild, and gains representation power by outperforming various baselines in point cloud completion tasks.

Ⅰ. 서 론

Representing and reconstructing shapes are two inseparable tasks that have long been fundamental problems in 3D computer vision. In recent years, neural fields for representing shapes in an implicit way have gained popularity by achieving faithful reconstruction performances. Specifically, discrete and sparse shapes represented as voxel grids, point clouds, or 2D multi-views are encoded into latent codes. The latent codes are then decoded with a given query location to learn the particular implicit function. At inference time, continuous 3D surfaces can be reconstructed by the mapping of coordinates to the implicit function value with arbitrary precision. Unfortunately, learning implicit representations and reconstructing from voxel grids remain inactive and limited to a few papers ([1], [2], [3], [4], [5]). While voxels are the early building blocks of shapes and can be easily processed by adapting learning-based image processing techniques, they suffer from the computation cost and efficiency that grows cubically with the resolution (i.e.,, model training time measured in days to weeks, see Figure 1).

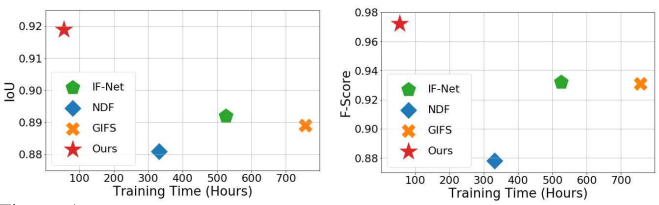

Figure 1.

Ⅱ. 본론

To overcome these challenges, we devise a novel easy-to-hard learning paradigm leveraging the proposed curriculum transformer decoder to accumulate a set of hierarchical latent codes. The framework gains power through three steps:

(1) The sparse voxel input is first encoded into a series of multi-scale feature grids. Grids preserve features with different scopes of receptive fields, thereby including coarse to fine-level features in each grid. Then given a query point, a hierarchical latent code set is acquired by grid sampling the continuous coordinates to each feature grid.

(2) The proposed curriculum transformer decoder, which embeds the idea of curriculum learning, includes the latent codes to the basic block in order of global to local. Thus, the inputs for the curriculum transformer are enabled to be updated to a denser feature.Integrating relevant information between the features by cross-attention layers, representations can start at a coarse geometry and gradually add finer-featured details that are relatively hard to capture.

(3) By exposing output predictions from all stages of latent codes, the neural field representing the continuous surface is learned.

A mild training regime is applied to representations derived from global stages, whereas strict training is employed to those derived from local stages. As part of the curriculum strategy, this optimization process also allows the model to begin small and progressively add more specific components.

The core of our method's curriculum-fashion neural field learning are: multi-stage inputs, which are hierarchical latent codes to be fed into the curriculum transformer decoder block in stages; and multi-stage losses, which are losses from the latent codes with varying contributions to the learning process. Both introduce an easy way for representations to be learned and furthermore, allow the geometry to be recognized quickly. This can be explained by the nature of curriculum strategies that encourage faster convergence and improved generalization. The major strength of our method is demonstrated in Figure 1, compared with previous voxel reconstruction methods: achieving approximately an order of magnitude speedup in training time while also achieving higher performance.

Ⅲ. 결론

In this work, we introduced a novel easy-to-hard paradigm to achieve fast and powerful learning of neural fields and reconstructions from discrete voxel grids. By embedding the idea from curriculum learning to the transformer architecture, a set of hierarchical latent codes including features of global and local is enabled to be learned in order. Specifically, different latent codes are included in the cross-attention layer of the decoder block, accumulating feature information at every step. Involving all decoder outputs in the optimization process also contributes to the curriculum strategy, encouraging higher generalization and fast convergence.

Experiments demonstrate that our method not only achieves a magnitude of speedup compared to baseline methods - but also shows faithful reconstructions from sparse voxels and point clouds and gains the potential for being useful in the wild.

ACKNOWLEDGMENT

Put sponsor acknowledgments.

참 고 문 헌

[1] L. M. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger.Occupancy networks: Learning 3d reconstruction in function space.In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pages 4460-4470, 2019.

[2] S. Peng, M. Niemeyer, L. M. Mescheder, M. Pollefeys, and A. Geiger. Convolutional occupancy networks. In Proc. European Conf. on Computer Vision (ECCV), pages 523-540, 2020.

[3] J. Chibane, T. Alldieck, and G. Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pages 6968-6979, 2020.

[4] J. Chibane, A. Mir, and G. Pons-Moll. Neural unsigned distance fields for implicit function learning. In Proc. Advances in Neural Information Processing Systems (NIPS), 2020.

[5] J. Ye, Y. Chen, N. Wang, and X. Wang. Gifs: Neural implicit function for general shape representation.In Proc. Conf. on Computer Vision and Pattern Recognition (CVPR), pages 12829-12839, 2022.