

Utonia: One Encoder for All 3D Point Clouds

Kabui, Charles

2026-03-13

[Read at ToKnow.ai](#)

The infographic features a dark blue background with a grid pattern. At the top right, there is a faint 3D point cloud visualization. The main title 'Utonia: One Encoder for All Point Clouds' is prominently displayed in white and light blue. Below the title, a subtitle reads 'Five 3D domains unified in a single self-supervised model'. Three key statistics are presented in separate boxes with vertical bars: '5 Domains' (Satellite, LiDAR, indoor, objects, and video-lifted point clouds), '81.1%' (ScanNet indoor segmentation SOTA with one unified model), and '82.1%' (Robotic grasping success rate, up from 74.7% with prior models). The date 'March 13, 2026' is at the bottom left, and the 'ToKnow.ai' logo is at the bottom right.

Utonia: One Encoder for All Point Clouds

Five 3D domains unified in a single self-supervised model

- 5 Domains**
Satellite, LiDAR, indoor, objects, and video-lifted point clouds
- 81.1%**
ScanNet indoor segmentation SOTA with one unified model
- 82.1%**
Robotic grasping success rate
Up from 74.7% with prior models

March 13, 2026

ToKnow.ai

Researchers at The University of Hong Kong and Xiaomi released [Utonia](#), the first self-supervised point cloud encoder trained across five 3D domains at once: satellite scans, outdoor street LiDAR, indoor room scans, standalone object models, and point clouds built from regular video. Built on a 137-million parameter [Point Transformer V3](#), it was trained on 250,000 cross-domain scenes plus 1 million 3D object assets. Three techniques make joint training work: randomly hiding color and surface data during training so the model stays robust when sensors differ, rescaling all point clouds to a shared spatial unit, and a position encoding

that avoids locking geometry to fixed grid coordinates. One model matches or beats separate domain-specific encoders across the board: **81.1%** on ScanNet indoor segmentation, **82.2%** on nuScenes outdoor segmentation, and **95.2%** on ScanObjectNN object classification. Robotic grasping success jumps to **82.1%** when robot policies use Utonia’s 3D features, up from 74.7% with older encoders.

3D understanding has been fragmented: one model for indoor rooms, another for self-driving, another for small objects. Teams building robots, AR headsets, or autonomous vehicles each trained their own encoder from scratch. Utonia replaces all of them with one pretrained model, and it reveals emergent behaviors that only appear during joint training. The model automatically learns that a toy car in a CAD dataset and a real car on a street share the same structure, something no single-domain model could discover. [Weights are on HuggingFace](#) under a CC-BY-NC 4.0 license.

This follows the same trajectory as language and vision: specialized models giving way to unified foundations that benefit all tasks. In 3D, fragmentation persisted longer because point cloud formats vary drastically across sensors. Utonia suggests the unification era for sparse 3D data has started.

Read More: Utonia’s cross-domain spatial understanding connects to the broader challenge of spatial consistency discussed in [The Trinity of Consistency](#).

Sources:

- [Utonia Paper \(arXiv:2603.03283\)](#)
- [Utonia GitHub Repository \(469 stars\)](#)
- [Utonia Model Weights on HuggingFace](#)
- [HuggingFace Paper Discussion \(164 upvotes, #1 paper of the day\)](#)

Disclaimer: For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. Read more: [/terms-of-service](#)