Generative Sparse Detection Networks for 3D Single-shot Object Detection

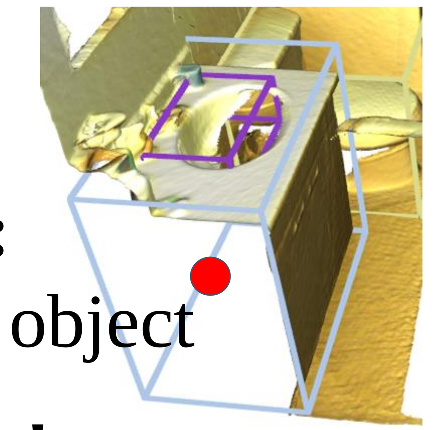
JunYoung Gwak†, Christopher Choy‡, Silvio Savarese†

†Stanford University ‡NVIDIA

1. Motivation

Disjoint input and output space:

- Input 3D scan: surface of the object
- Output anchor space: center of the bounding box



Sparse convolution / PointNet: Learn only on the surface of the object

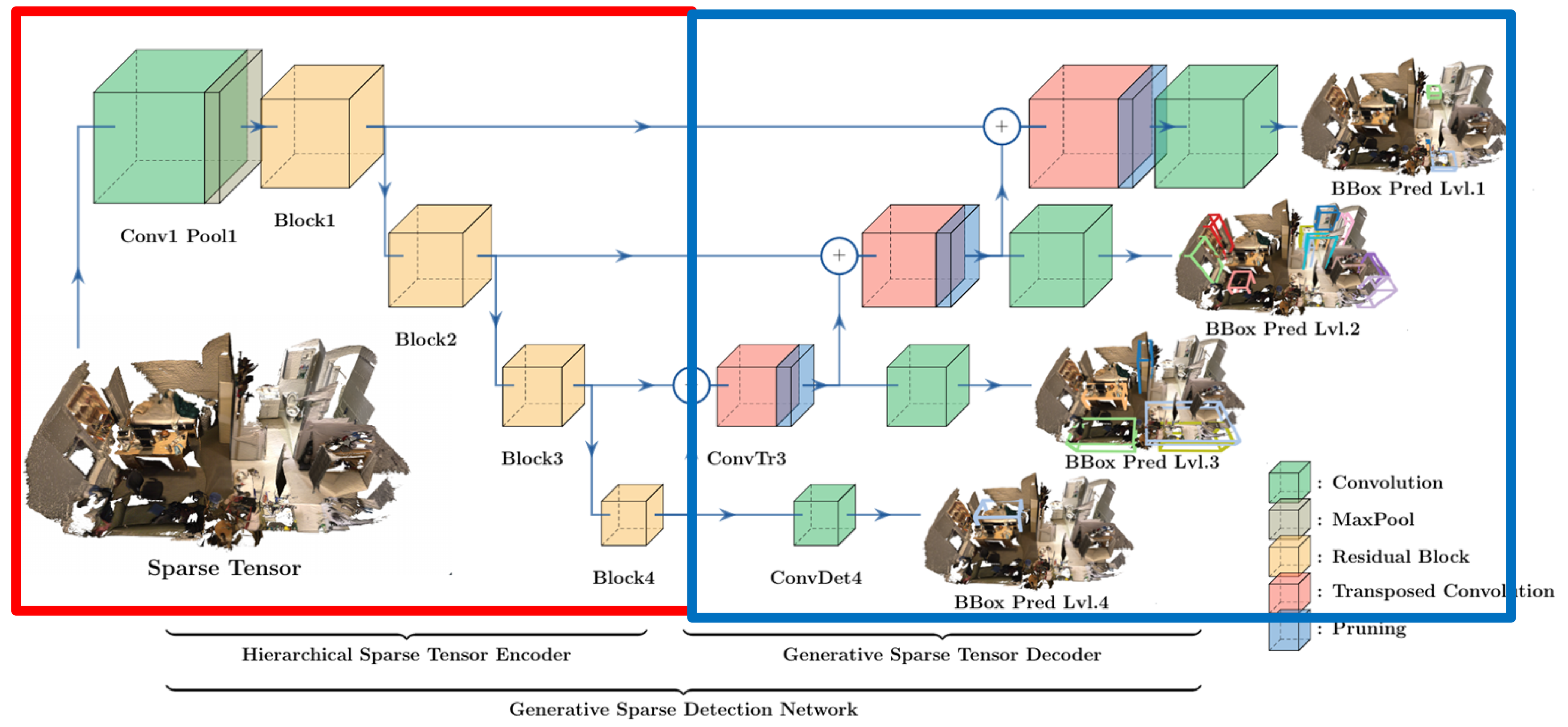
⇒ Output space is unreachable!

2. Previous Works

1. Make predictions at the surface of the object (Point R-CNN, AVOD, …) ⇒ Nontrivial to decide which part of the surface is responsible for the prediction
2. Convert sparse tensor to dense tensor (3DMV) ⇒ Give up efficiency in sparsity
3. For every point, predict relative center of the instance (VoteNet) ⇒ Requires center aggregation

Ours: Object centers are close to the object surface. Can we **generate** object centers **efficiently**?

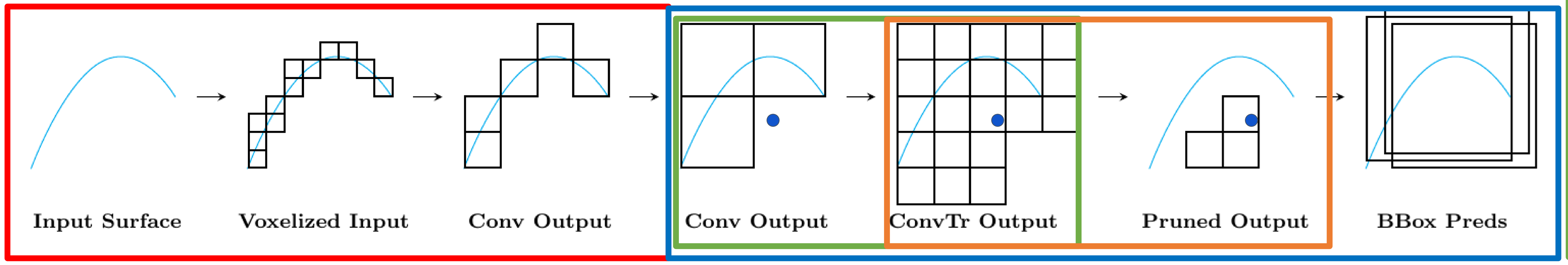
3. Method

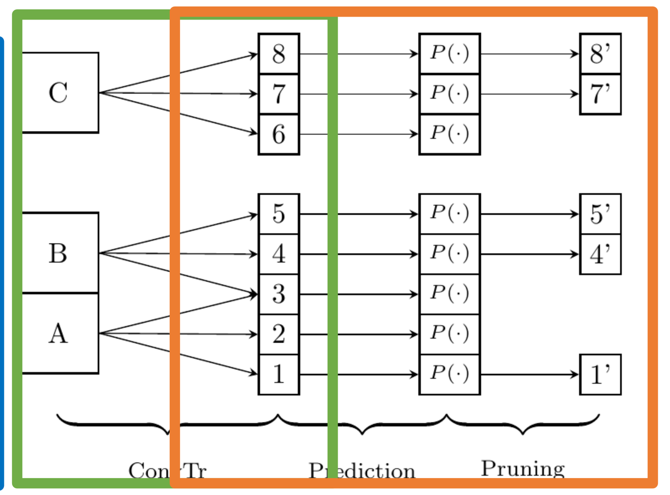


Hierarchical Sparse Tensor Encoder: Efficiently encodes 3D scene using *Sparse Convolution*.

Generative Sparse Tensor Decoder: *Generates* and *prunes* new coordinates to support anchor box centers

- **Sparse Transposed Convolution**
  - Outer-product of the convolution kernel shape on the input coordinates
  - Generates surrounding coordinates of the input coordinates (expands support)
- **Sparsity Pruning**
  - For each generated point, predict whether to prune the coordinate
  - Prune coordinates that are not bounding box centers

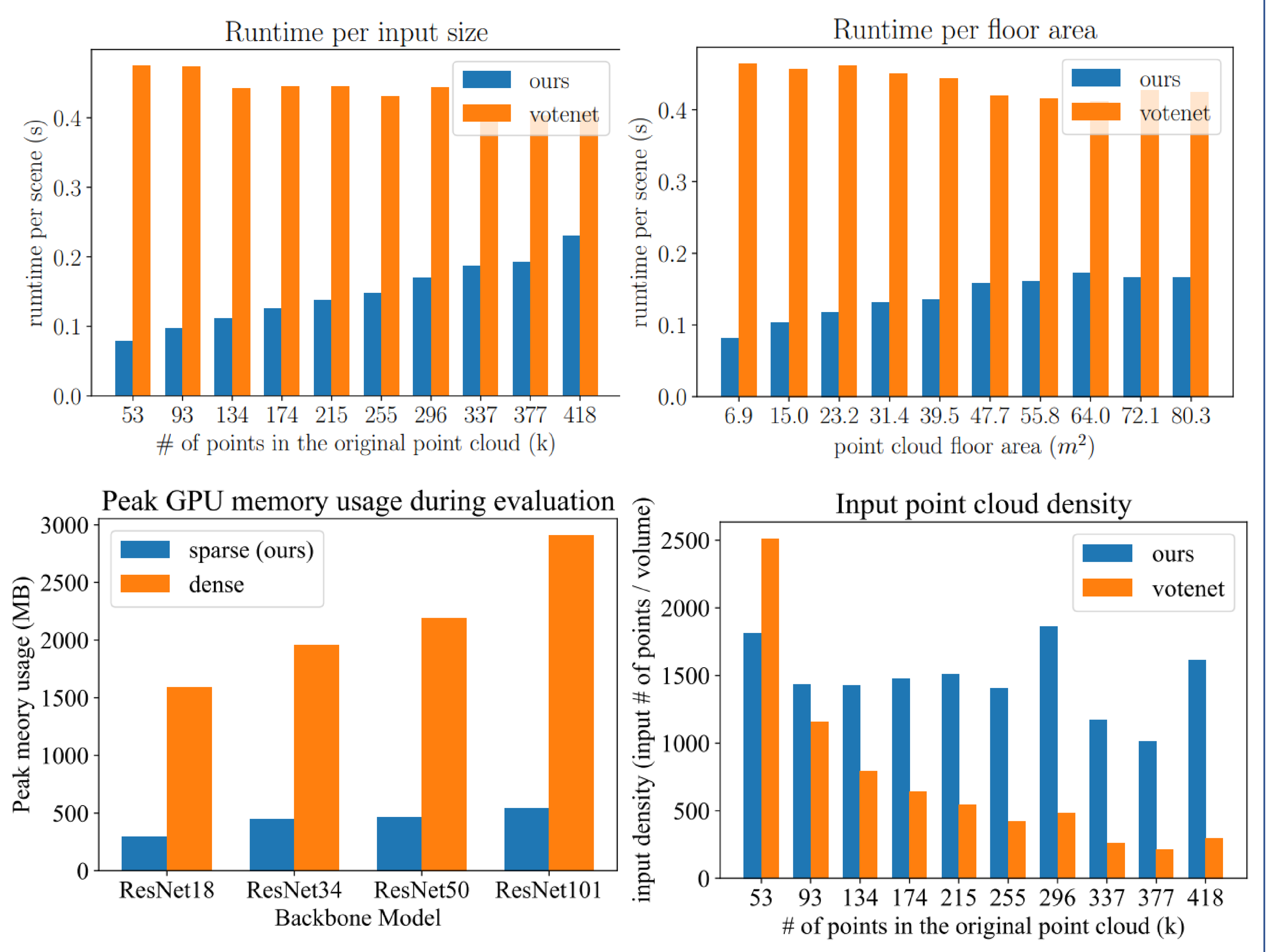




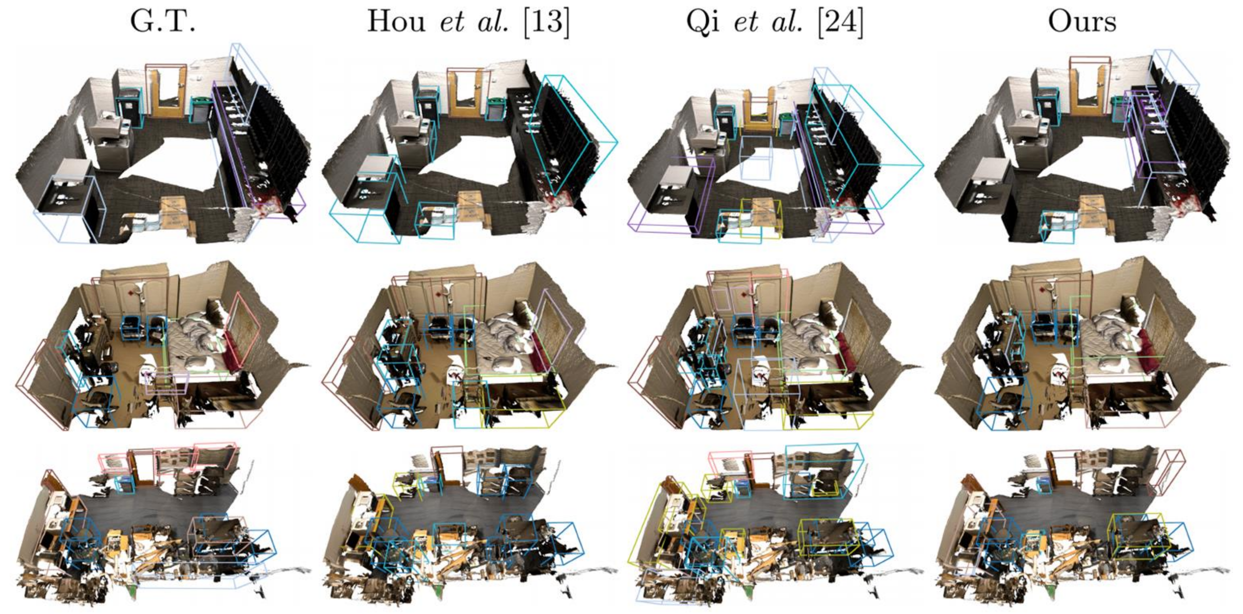
4. Experiments

ScanNet

- Outperforms previous state-of-the-art by **4.2 mAP@0.25**
- While being **x3.7 faster**
- **x6 memory efficient** to dense counterpart
- Maintains constant input density

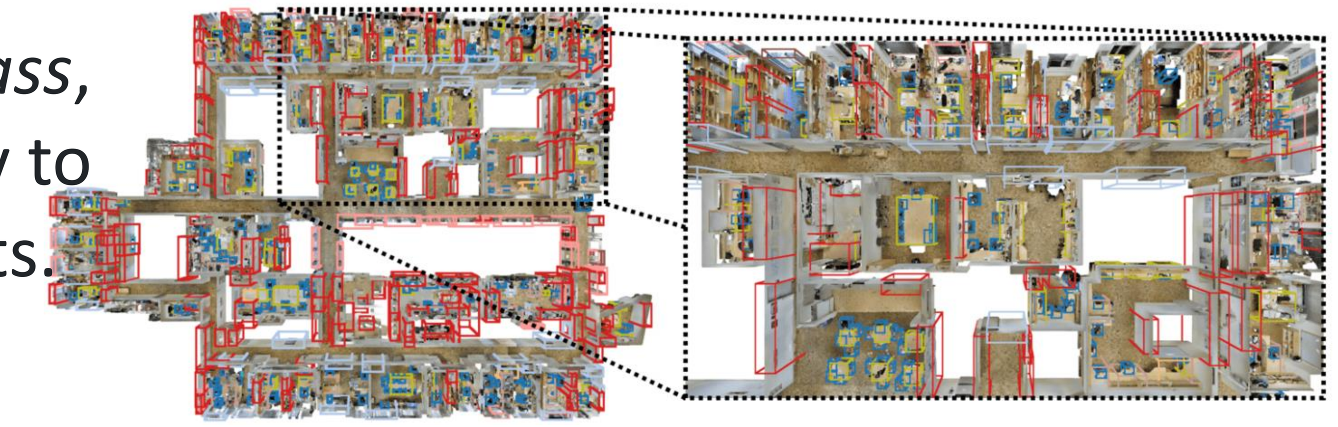


Method	Single Shot	mAP@0.25	mAP@0.5
DSS [28,13]	✗	15.2	6.8
MRCNN 2D-3D [11,13]	✗	17.3	10.5
F-PointNet [25]	✗	19.8	10.8
GSPN [37,24]	✗	30.6	17.7
3D-SIS [13]	✓	25.4	14.6
3D-SIS [13] + 5 views	✓	40.2	22.5
VoteNet [24]	✗	58.6	33.5
GSDN (Ours)	✓	**62.8**	**34.8**



S3DIS

- Outperforms baseline method
- Can process the entire building 5 (13984m^3, 53 room) of S3DIS dataset in a *single fully convolutional feed-forward pass*, only using 5G of GPU memory to detect 573 instances of objects.



Gibson

- Our model trained on single room of ScanNet dataset generanlizes to multi-story buildings without any ad-hoc pre-processing or post-processing.