ECCV'20 ONLINE

23-28 AUGUST 2020

16TH EUROPEAN CONFERENCE ON

COMPUTER VISION

WWW.ECCV2020.EU

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

JunYoung Gwak, Christopher Choy, Silvio Savarese

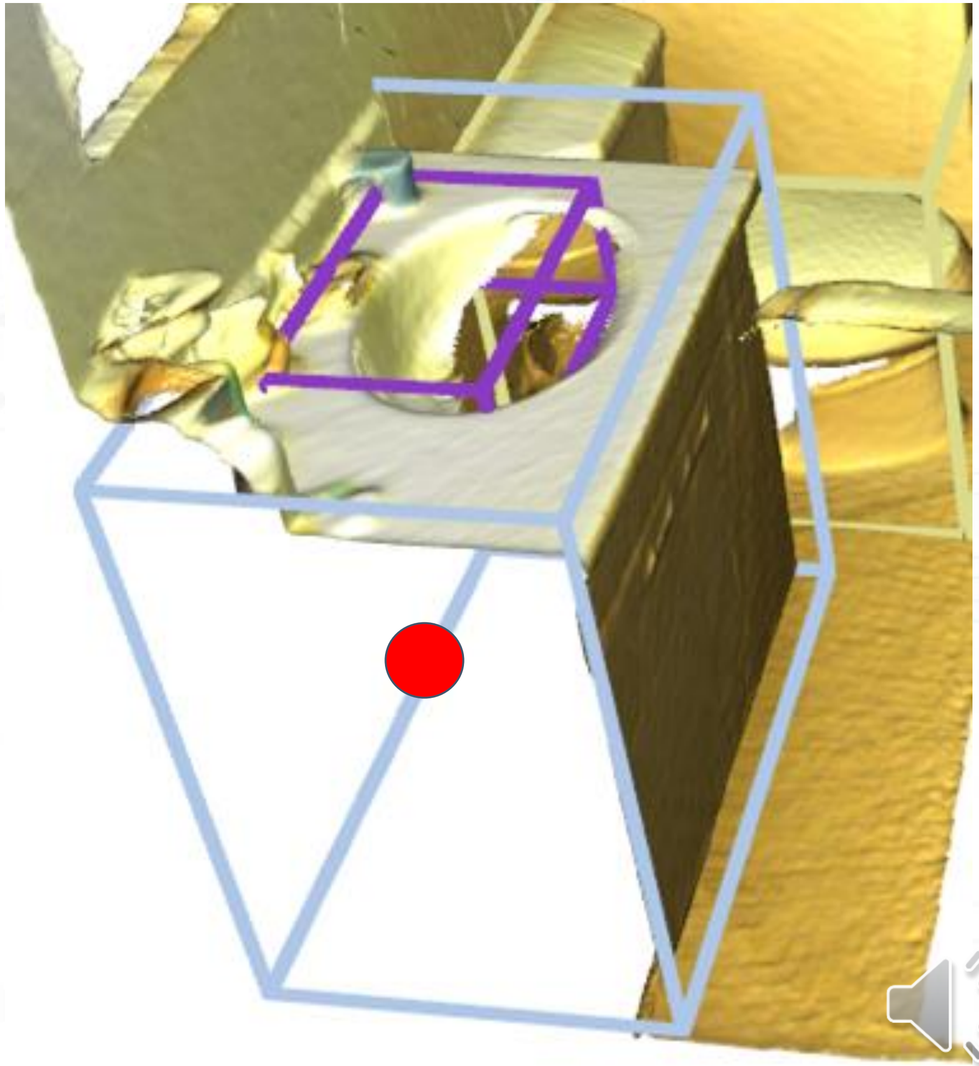
Key Challenge of 3D Object Detection

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

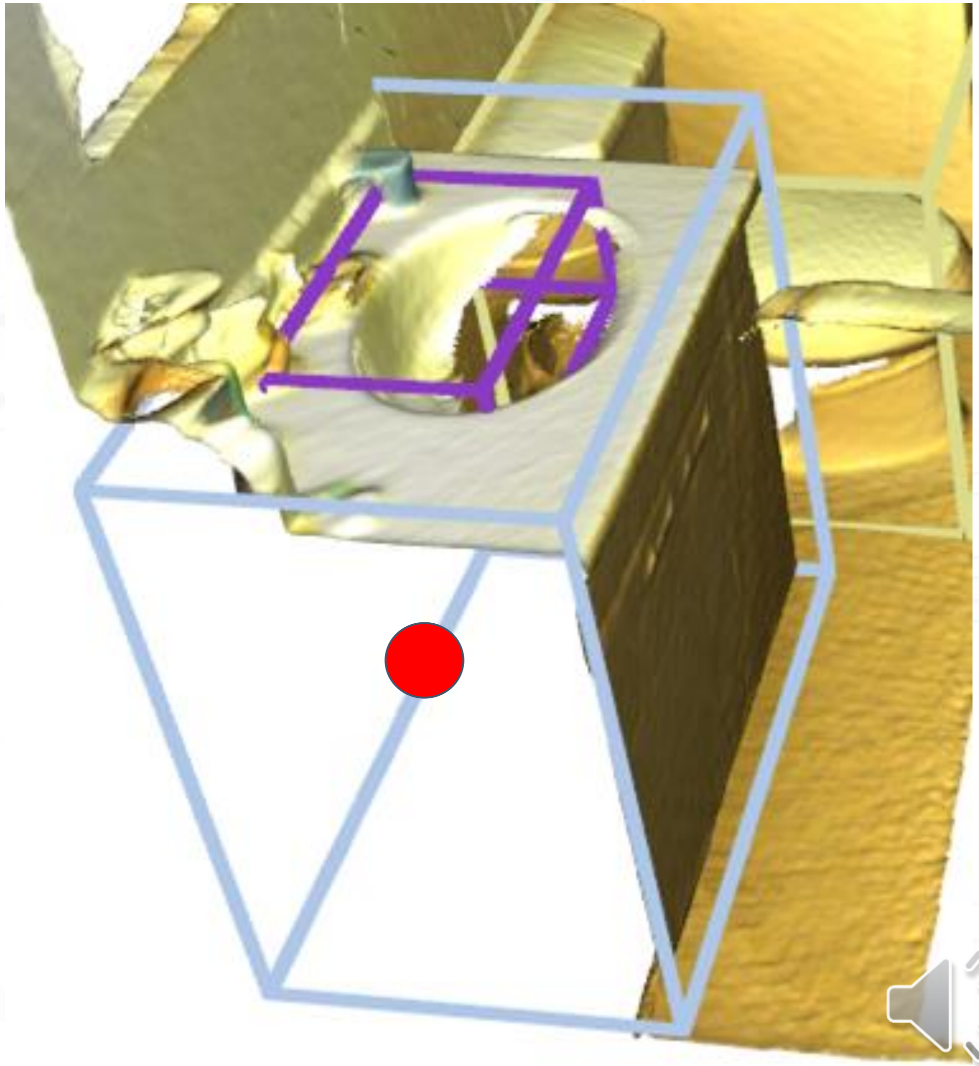
⇒ Output space is unreachable!



Key Challenge of 3D Object Detection

Possible solutions? (previous works)

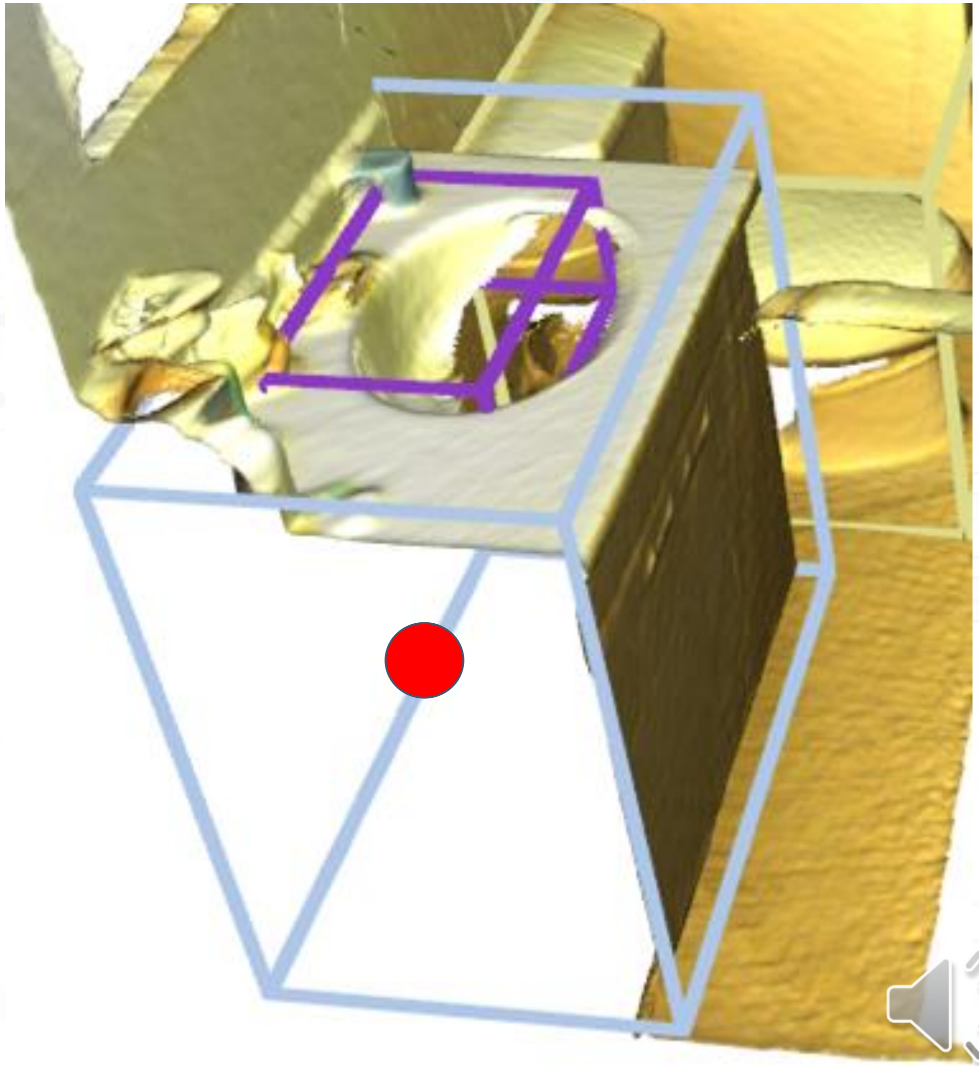
- Ignore this problem and make predictions at the surface of the object
  - Nontrivial to decide which part of the surface is responsible for the prediction



Key Challenge of 3D Object Detection

Possible solutions? (previous works)

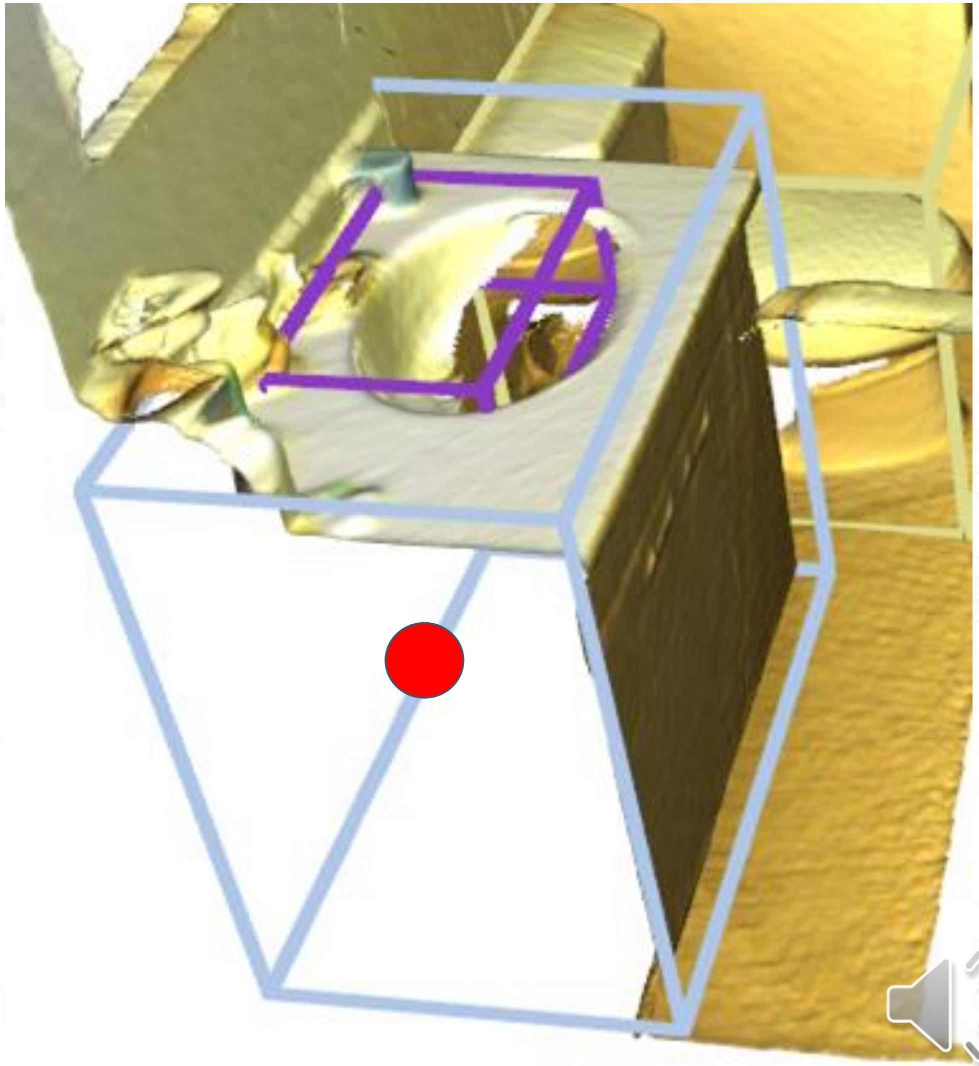
- Ignore this problem and make predictions at the surface of the object
  - Nontrivial to decide which part of the surface is responsible for the prediction
- Convert sparse tensor to dense tensor
  - Give up efficiency in sparsity



Key Challenge of 3D Object Detection

Possible solutions? (previous works)

- Ignore this problem and make predictions at the surface of the object
  - Nontrivial to decide which part of the surface is responsible for the prediction
- Convert sparse tensor to dense tensor
  - Give up efficiency in sparsity
- For every point, predict relative center of the instance
  - Requires center aggregation (clustering), inefficient

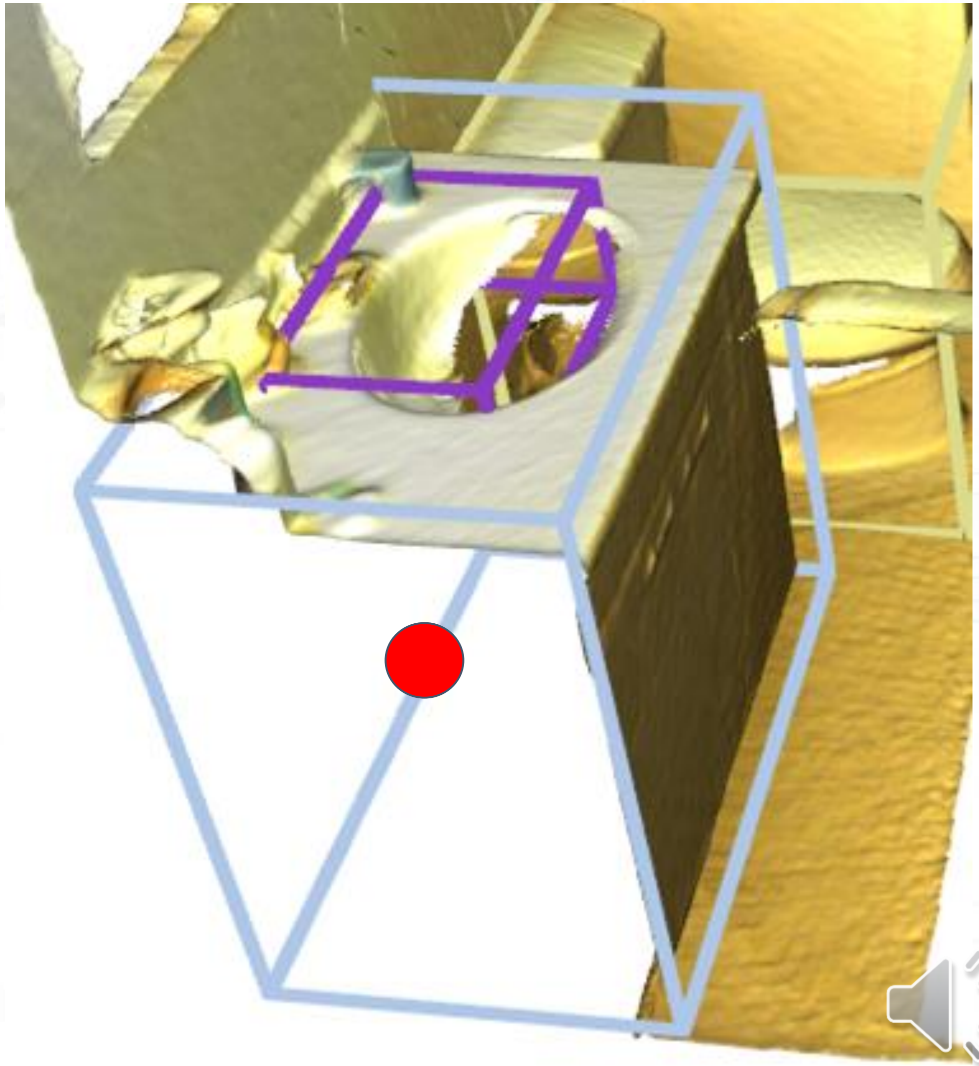


Key Challenge of 3D Object Detection

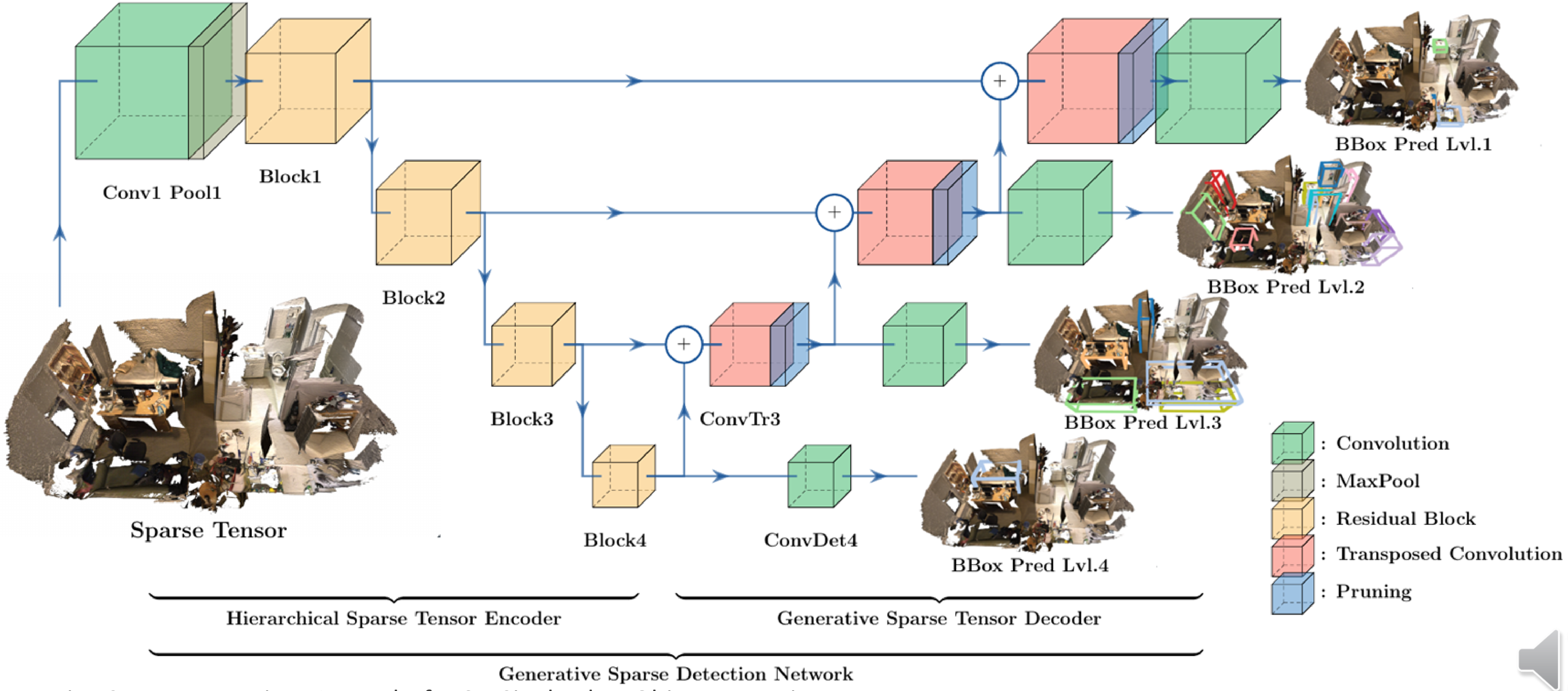
Key observation:

Object centers are close to the object surface

Can we **generate** object centers **efficiently**?

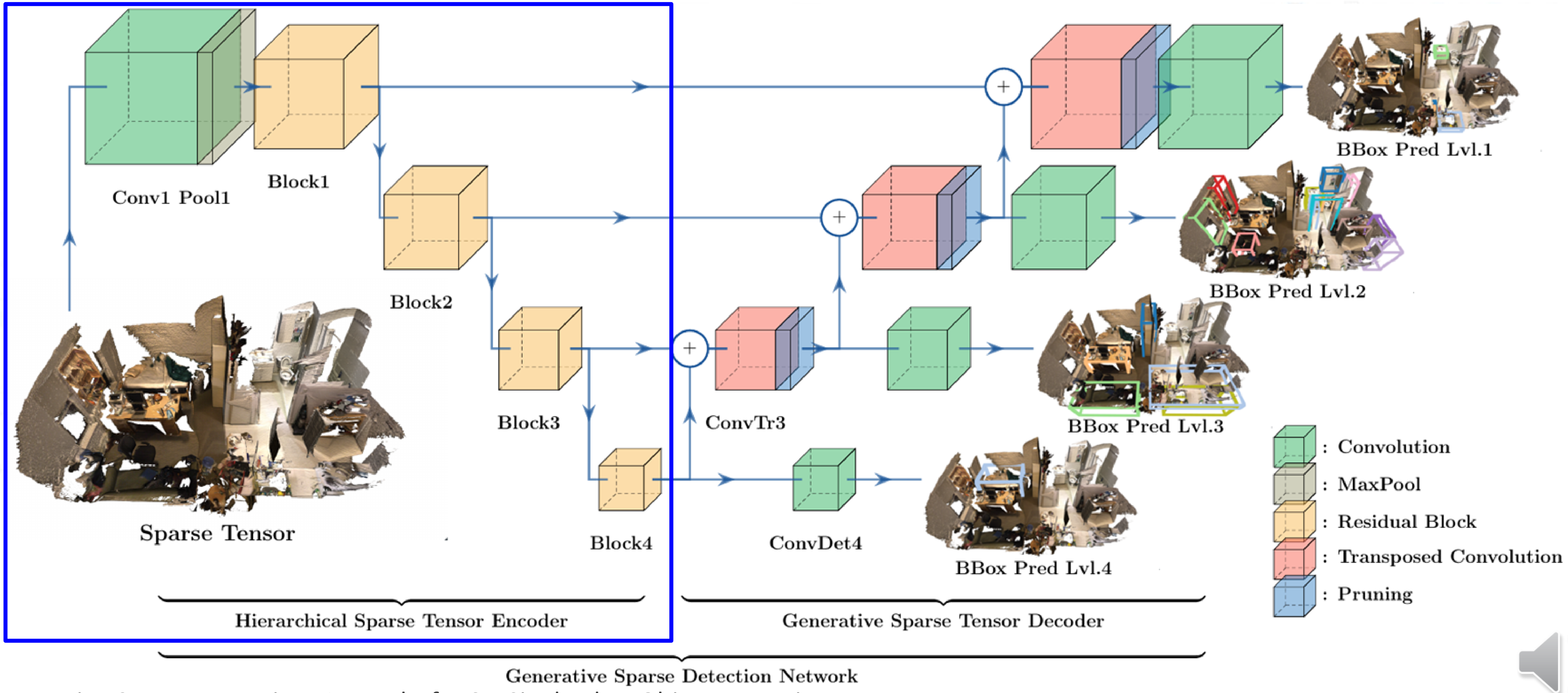


Method Overview



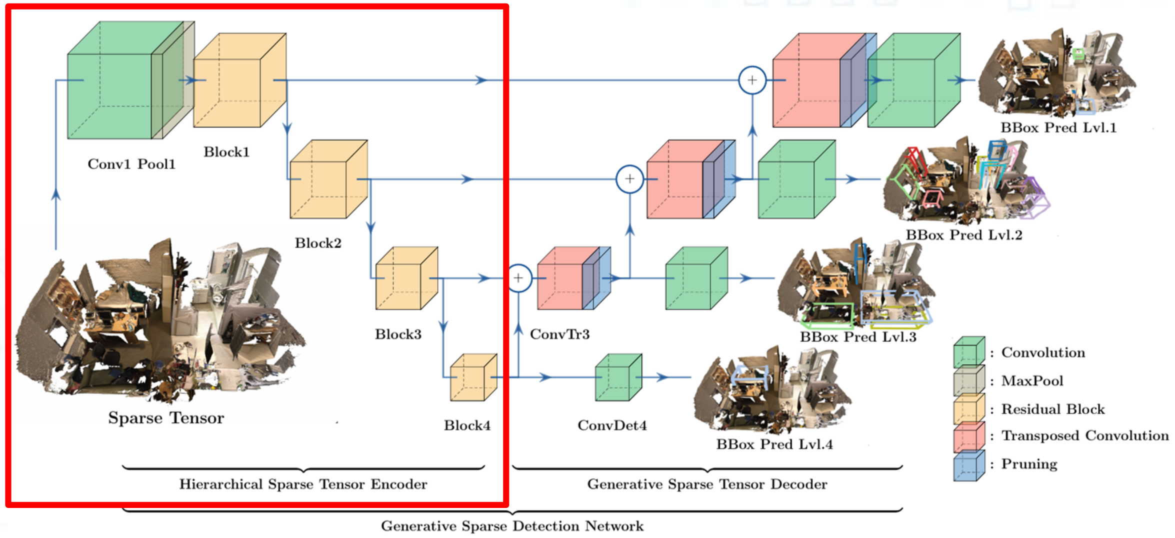
Generative Sparse Detection Networks for 3D Single-shot Object Detection

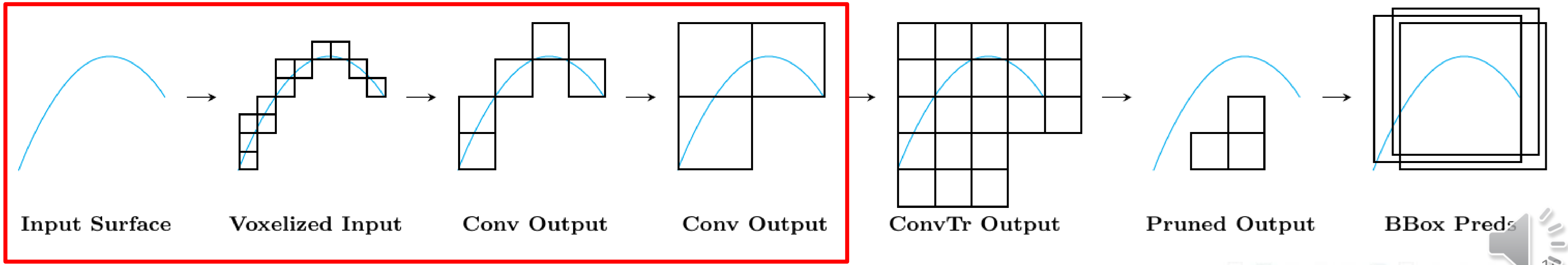
Hierarchical Sparse Tensor Encoder



Hierarchical Sparse Tensor Encoder

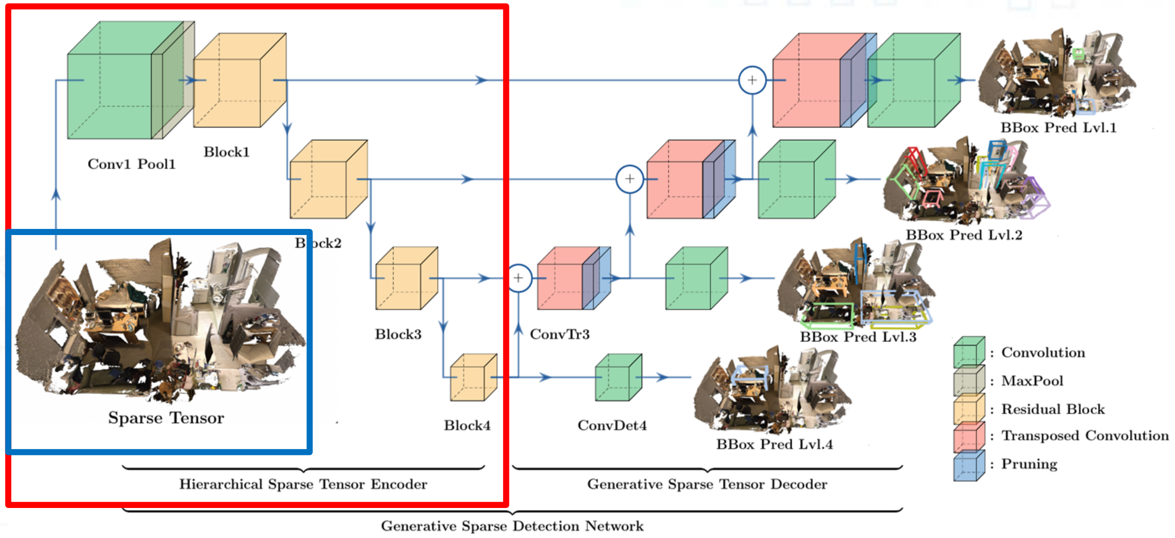
- Generates hierarchical sparse tensor features with sparse 3D ResNet
- Analogous to ResNet encoders commonly used in of 2D detectors

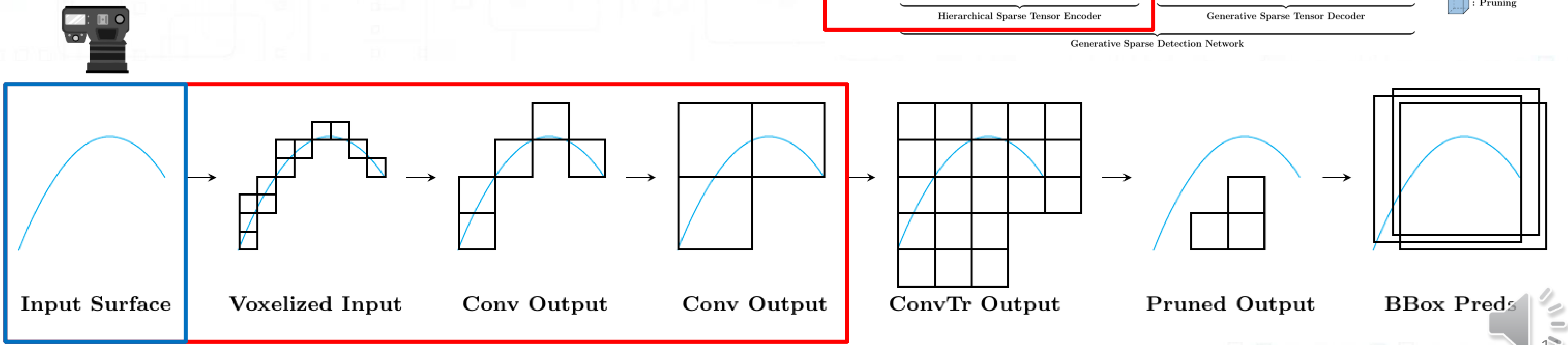




Hierarchical Sparse Tensor Encoder

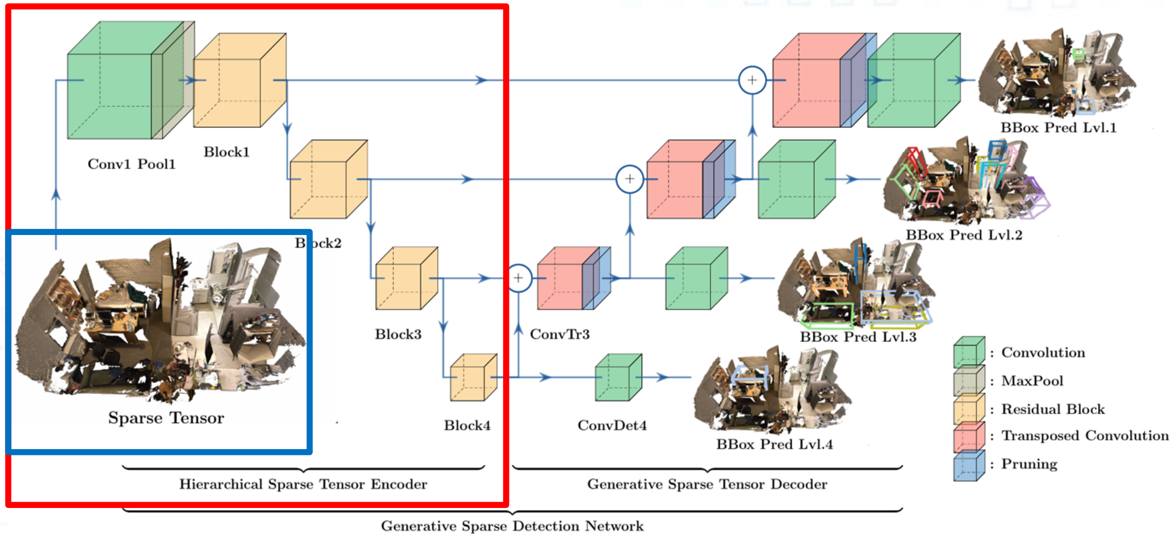
- Generates hierarchical sparse tensor features with sparse 3D ResNet
- Analogous to ResNet encoders commonly used in of 2D detectors

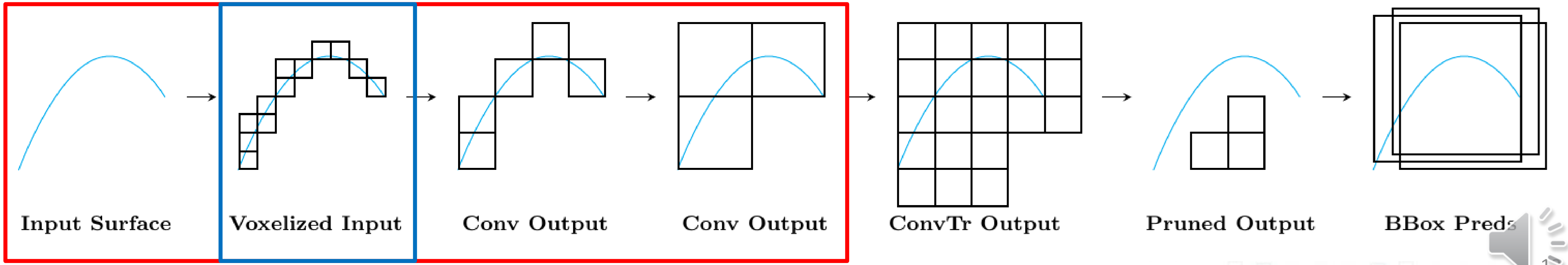




Hierarchical Sparse Tensor Encoder

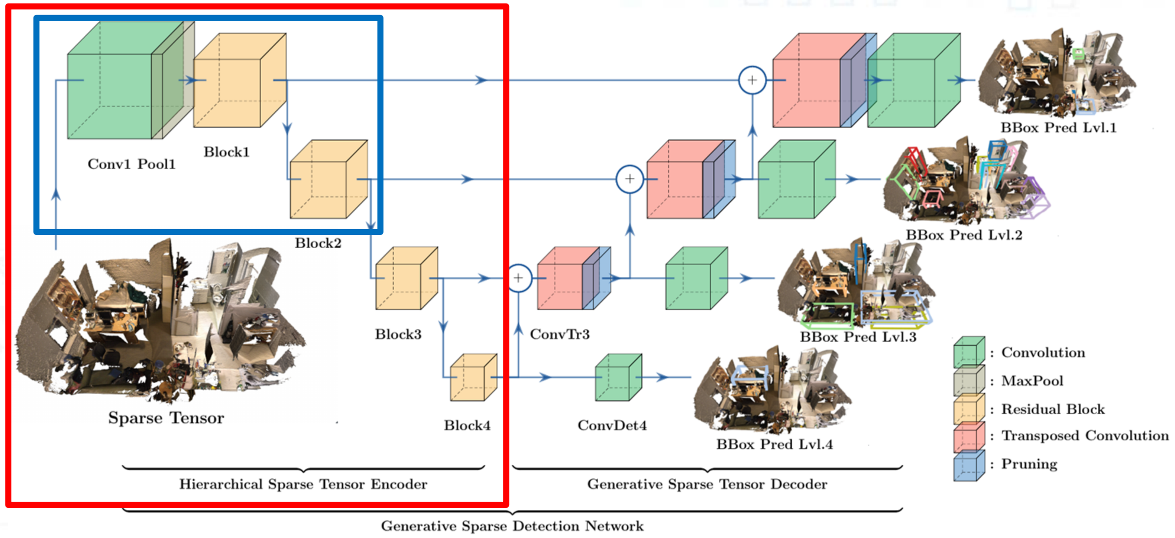
- Generates hierarchical sparse tensor features with sparse 3D ResNet
- Analogous to ResNet encoders commonly used in of 2D detectors

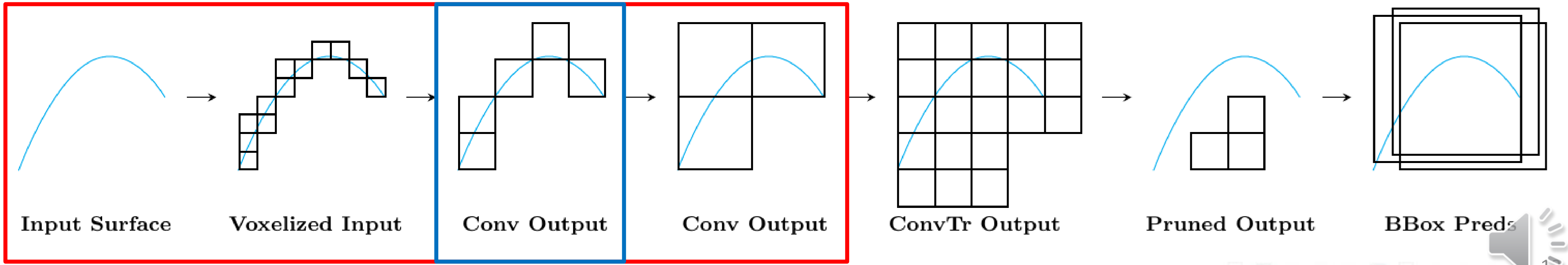




Hierarchical Sparse Tensor Encoder

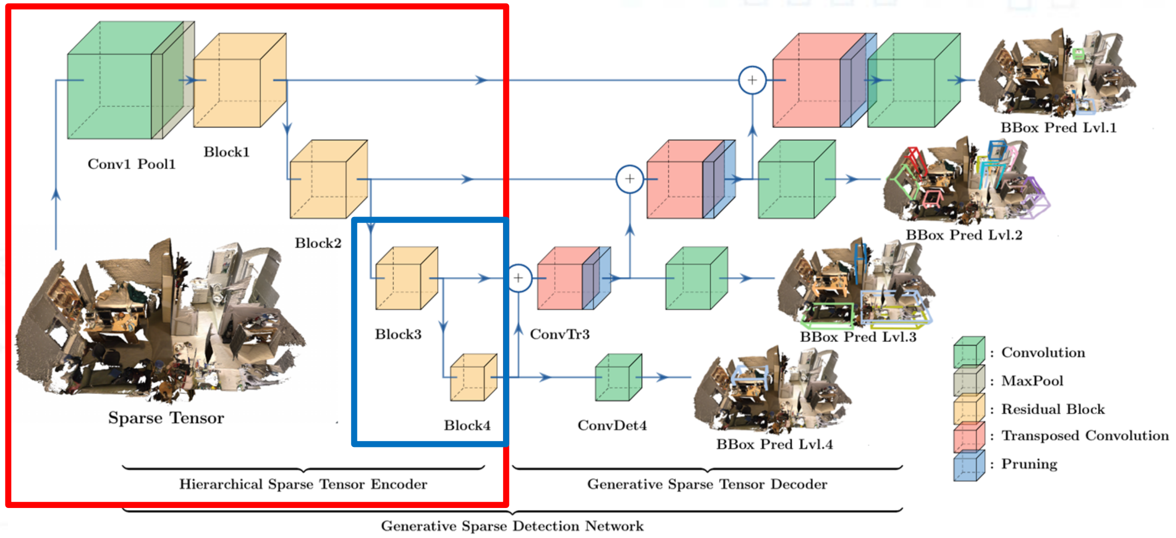
- Generates hierarchical sparse tensor features with sparse 3D ResNet
- Analogous to ResNet encoders commonly used in of 2D detectors

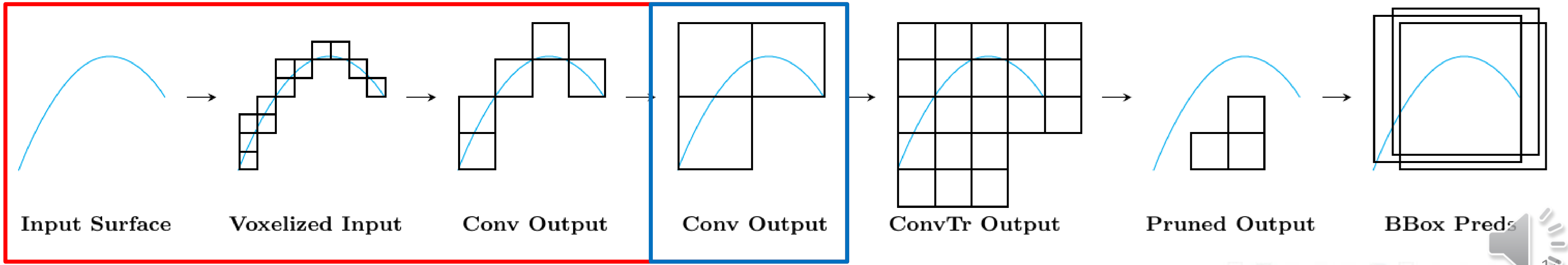




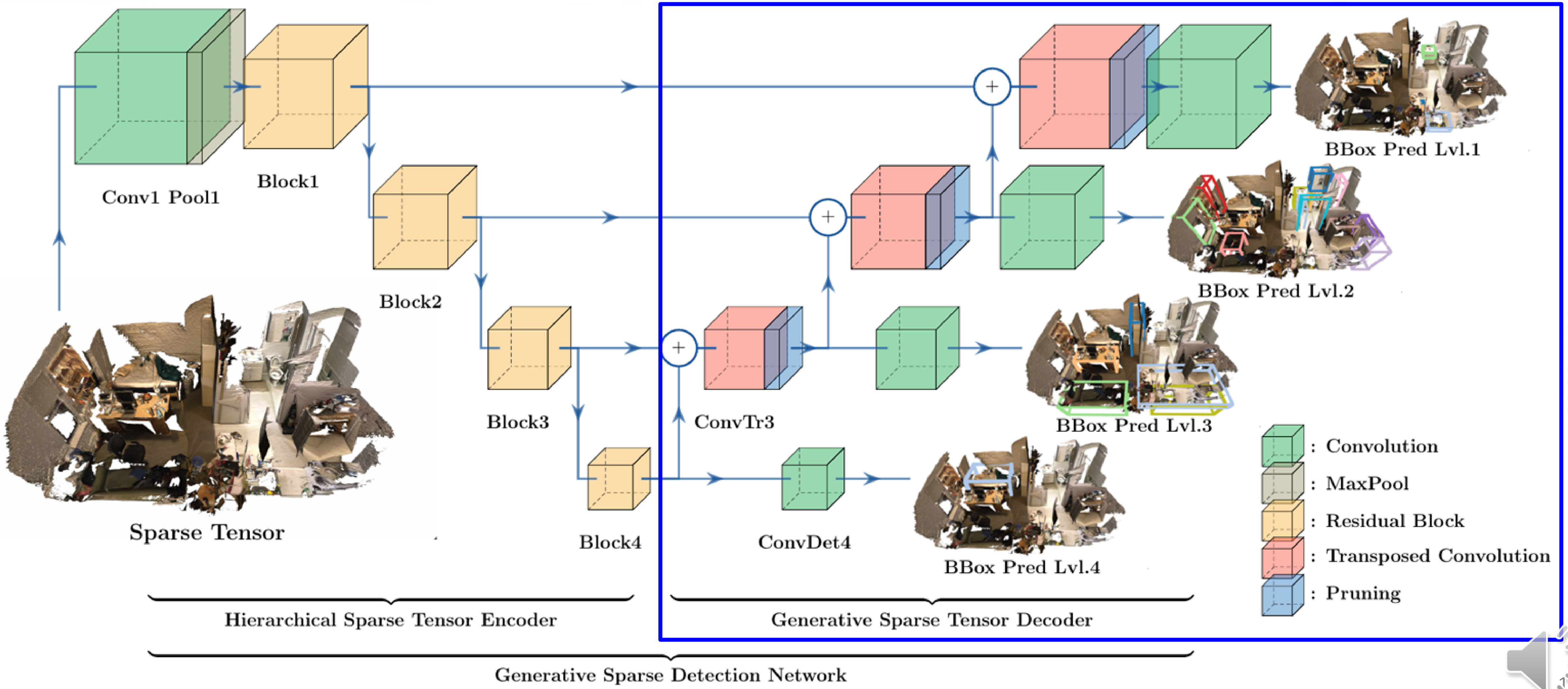
Hierarchical Sparse Tensor Encoder

- Generates hierarchical sparse tensor features with sparse 3D ResNet
- Analogous to ResNet encoders commonly used in of 2D detectors

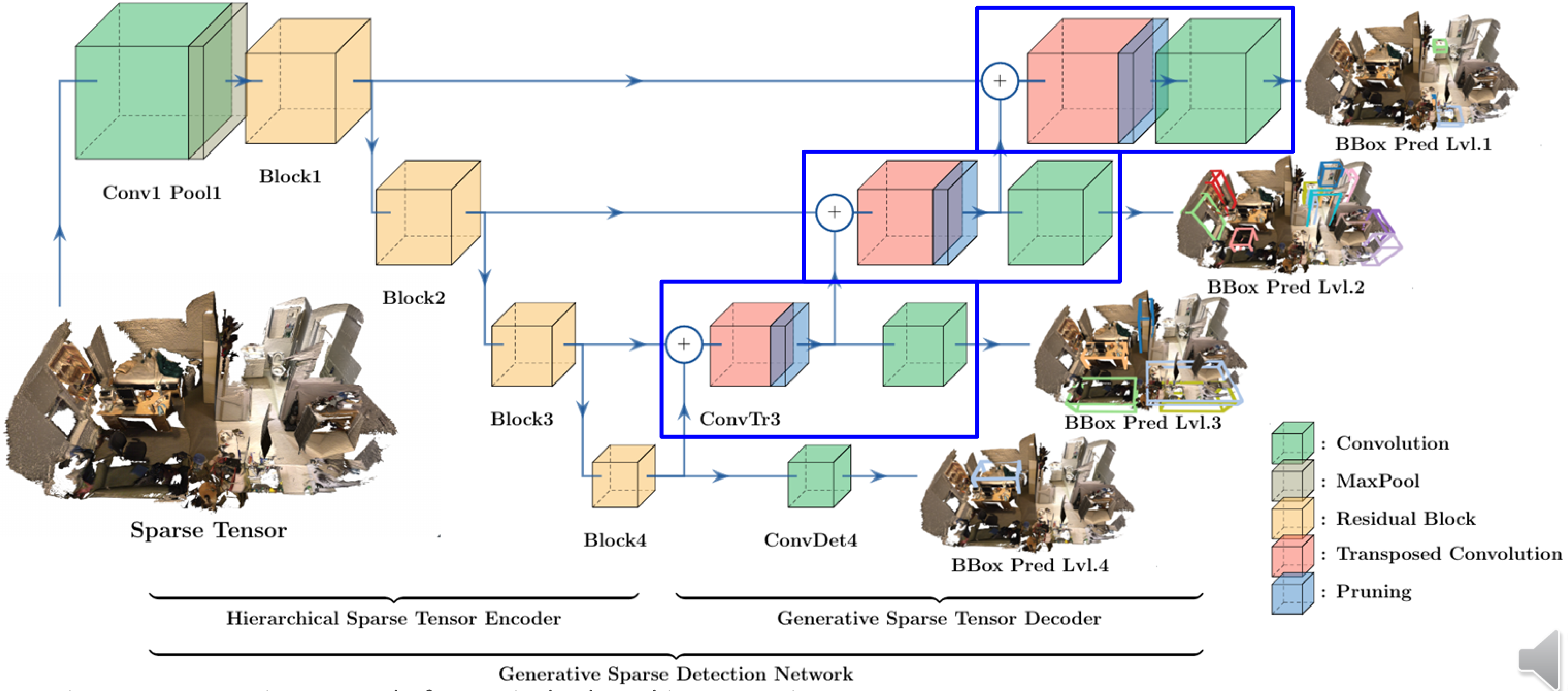




Generative Sparse Tensor Decoder

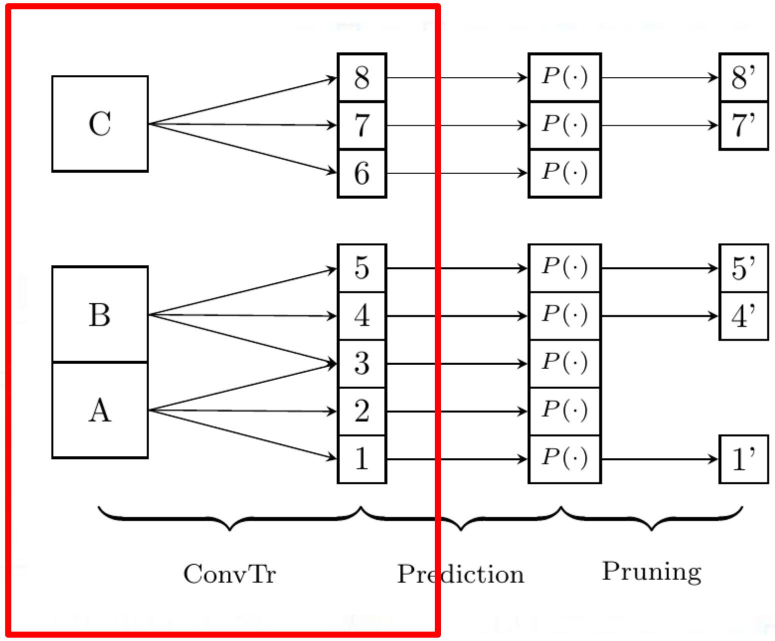


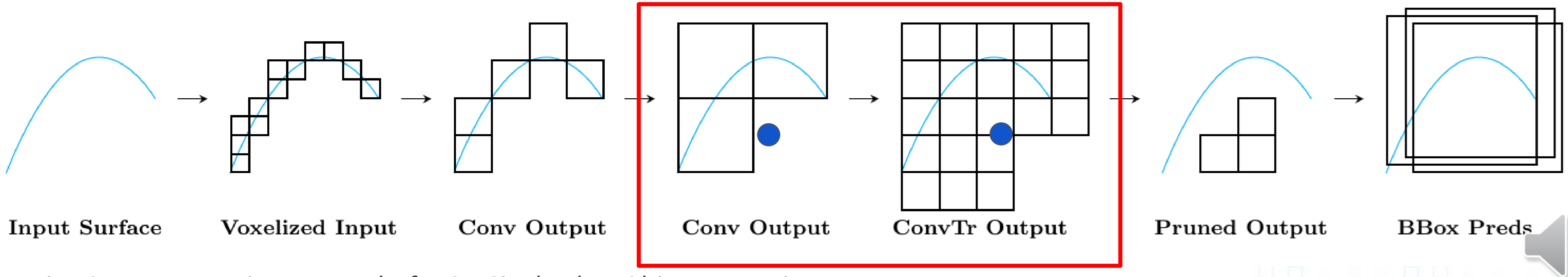
Transposed Convolution + Sparsity Pruning



Transposed Convolution + Sparsity Pruning

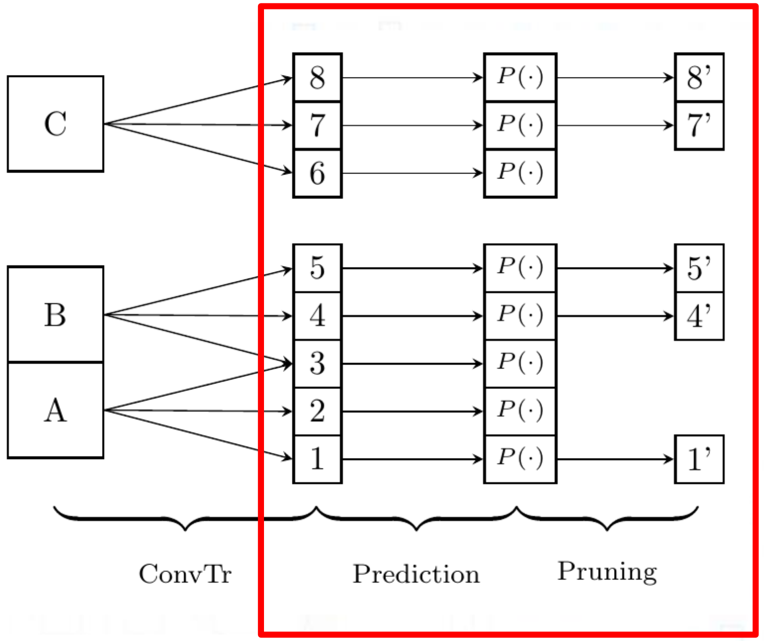
- **Sparse Transposed Convolution**
  - Outer-product of the convolution kernel shape on the input coordinates
  - **Generates surrounding coordinates of the input coordinates (expands support)**
- Sparsity Pruning

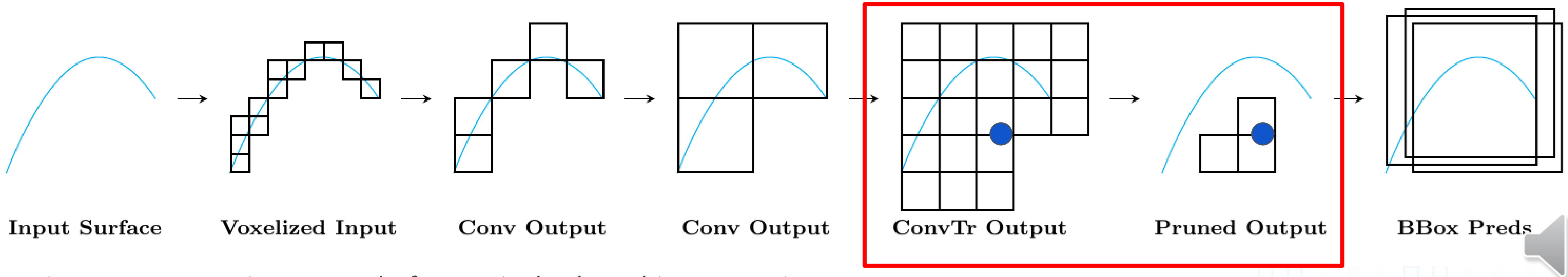




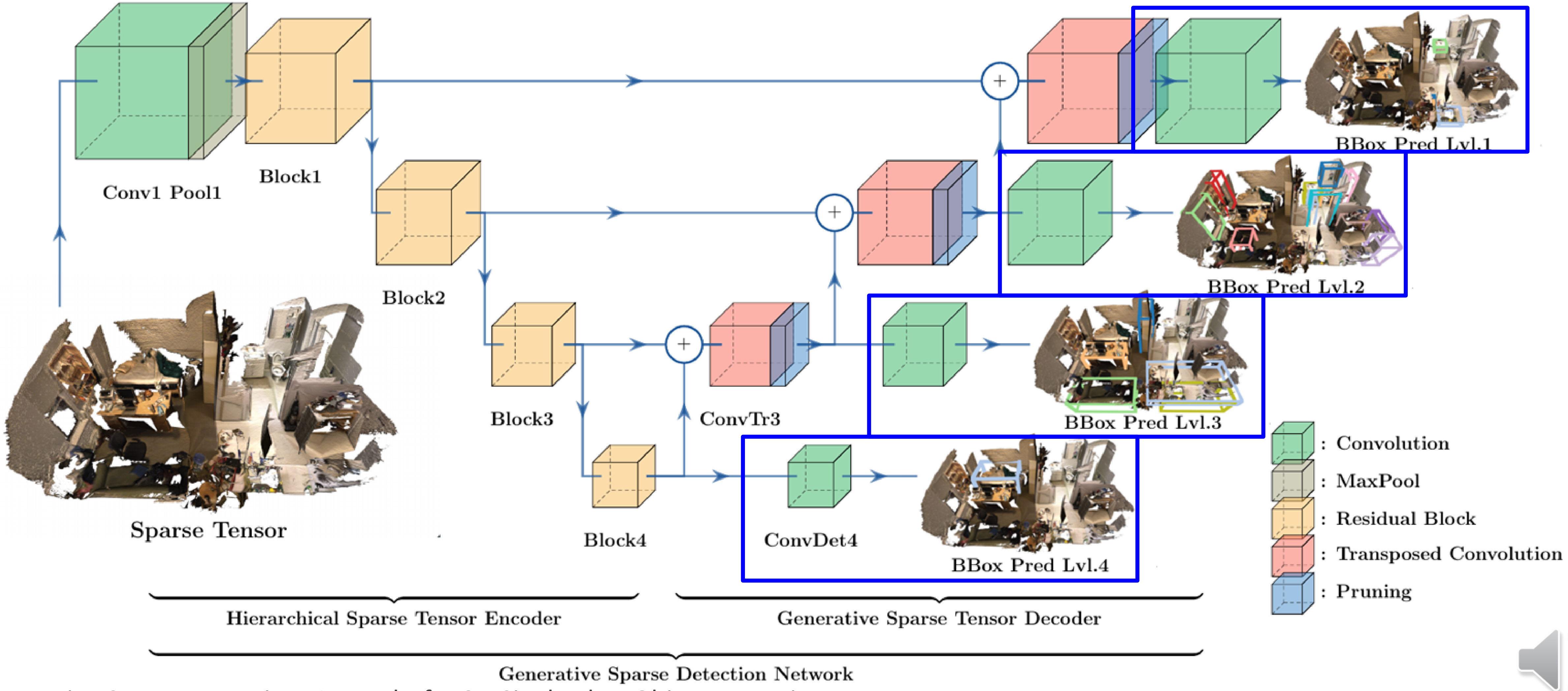
Transposed Convolution + Sparsity Pruning

- Sparse Transposed Convolution
- **Sparsity Pruning**
  - For each generated point, predict whether to prune the coordinate
  - **Prune** coordinates that are not bounding box centers





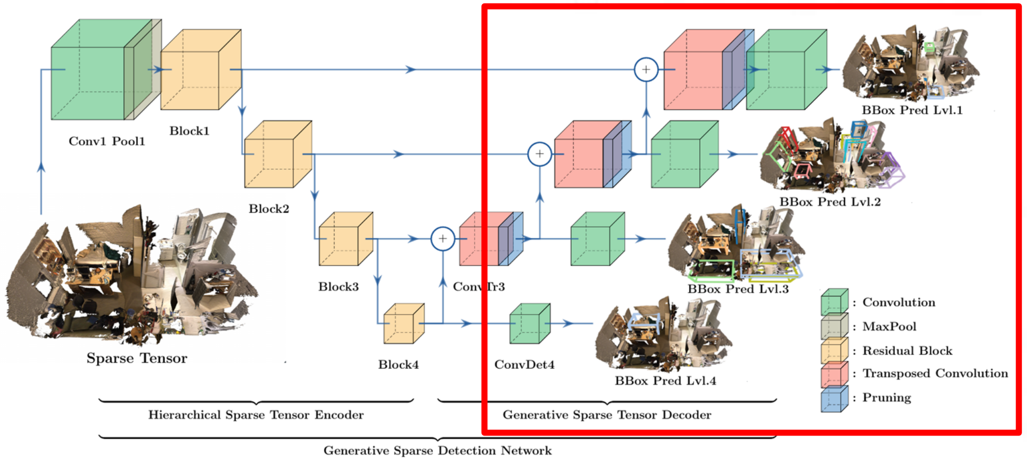
Bounding box prediction

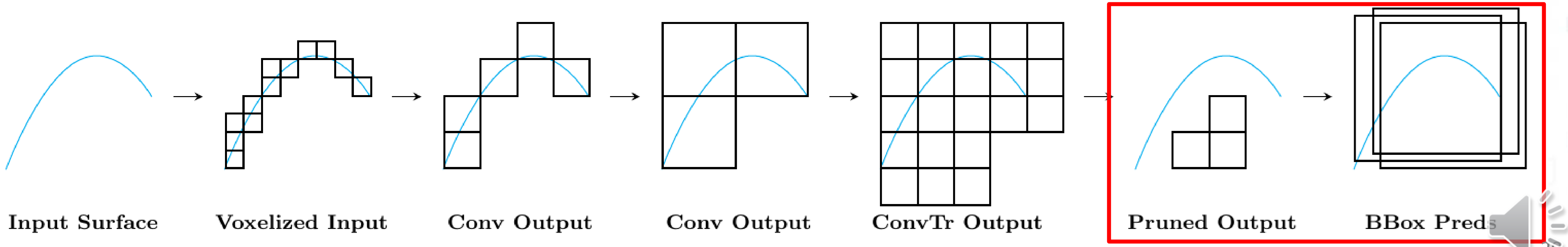


Generative Sparse Detection Networks for 3D Single-shot Object Detection

Bounding box prediction

- For every point that are not pruned, predict
  - Anchor classification
  - Bounding box regression
  - Semantic classification
- Hierarchical multi-scale prediction on pyramid network





Advantages of Our Method

Full 3D search space

- Search for object center up to ±1.6m of any observable surface

Fully sparse: Minimal runtime and memory footprint

- Sparse Convolution Encoder
- Conv Transpose and Pruning to only generate anchor centers

Fully-convolutional

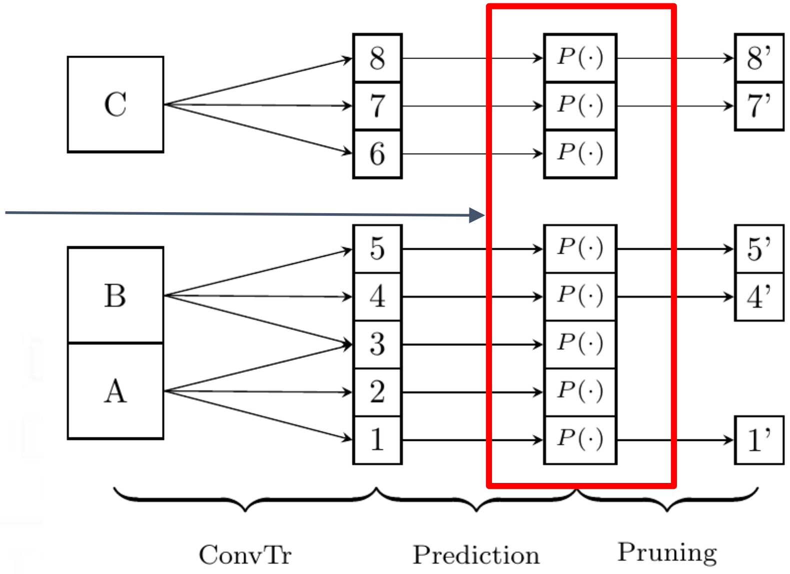
- Simple architecture
- No clustering, no crop and merge, just convolutions

Losses

- Sparsity Prediction: Balanced Cross Entropy
- Anchor Prediction: Balanced Cross Entropy
- Semantic Prediction: Cross Entropy
- Bounding Box Regression: Huber Loss

Losses

- Sparsity Prediction: Balanced Cross Entropy
- Anchor Prediction: Balanced Cross Entropy
- Semantic Prediction: Cross Entropy
- Bounding Box Regression: Huber Loss



Balanced Cross Entropy

Overcome heavy label bias by equally penalizing positive and negative samples

$$L_{\mathrm{b}}(\hat{\mathbf{y}}, \mathbf{y}) = -\frac{1}{2|\mathcal{P}|}\sum_{i\in\mathcal{P}} \log(P(\hat{\mathbf{y}}_i)) - \frac{1}{2|\mathcal{N}|}\sum_{i\in\mathcal{N}} \log(1 - P(\hat{\mathbf{y}}_i))$$

Losses

- Sparsity Prediction: Balanced Cross Entropy
- Anchor Prediction: Balanced Cross Entropy
- Semantic Prediction: Cross Entropy
- Bounding Box Regression: Huber Loss

Bounding box parameters

Balanced Cross Entropy

Overcome heavy label bias by equally penalizing positive and negative samples

$$L_{\mathrm{b}}(\hat{\mathbf{y}}, \mathbf{y}) = -\frac{1}{2|\mathcal{P}|}\sum_{i\in\mathcal{P}}\log(P(\hat{\mathbf{y}}_i)) - \frac{1}{2|\mathcal{N}|}\sum_{i\in\mathcal{N}}\log(1 - P(\hat{\mathbf{y}}_i))$$

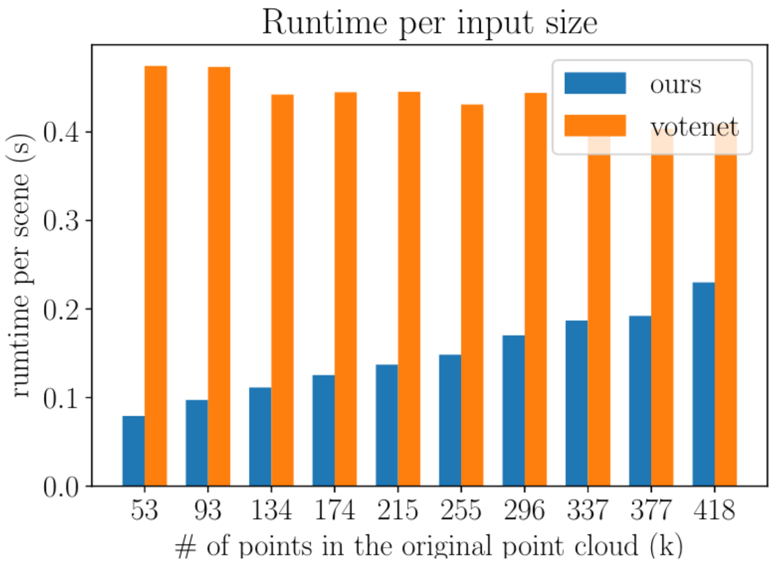
Comparison with previous SOTA - ScanNet

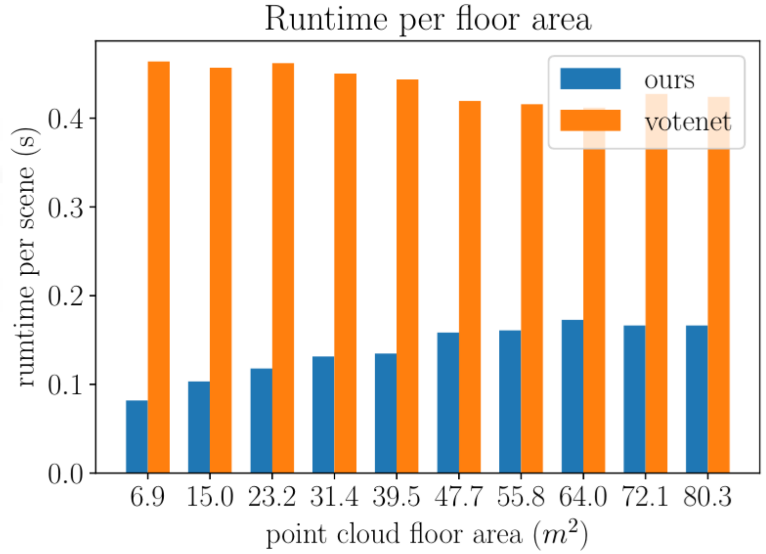
- Outperforms previous state-of-the-art by **4.2 mAP@0.25**
  - While being a single-shot detection

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**

Comparison with previous SOTA - ScanNet

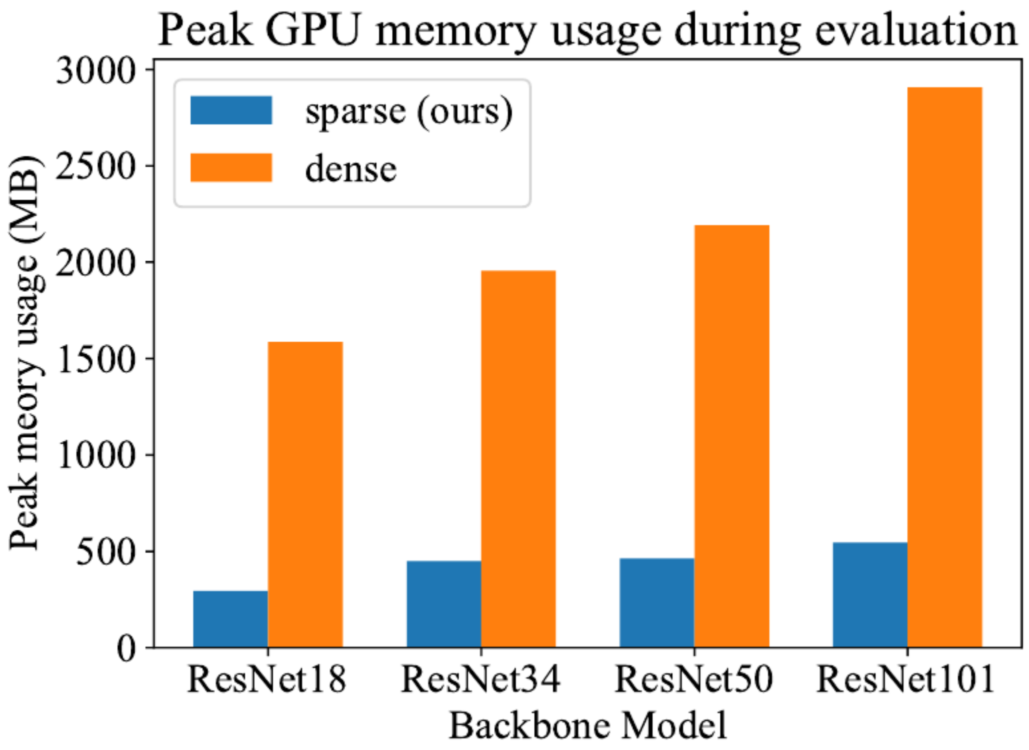
- Outperforms previous state-of-the-art by **4.2 mAP@0.25**
  - While being a single-shot detection
- While being **x3.7 faster**
  - runtime linear to # of points
  - runtime sublinear to floor area
  - ⇒ **free from curse of dimensionality!!**





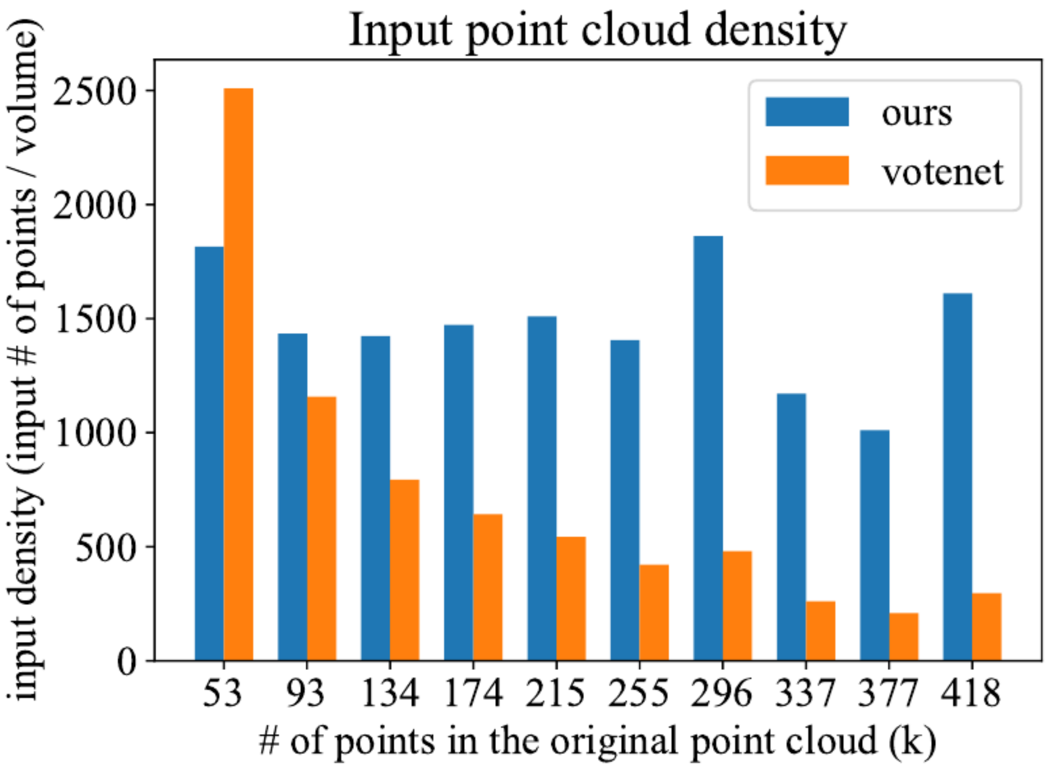
Comparison with previous SOTA - ScanNet

- Outperforms previous state-of-the-art by **4.2 mAP@0.25**
  - While being a single-shot detection
- While being **x3.7 faster**
  - runtime linear to # of points
  - runtime sublinear to floor area
  - ⇒ **free from curse of dimensionality!!**
- Minimal memory footprint
  - **x6** efficient to dense counterpart

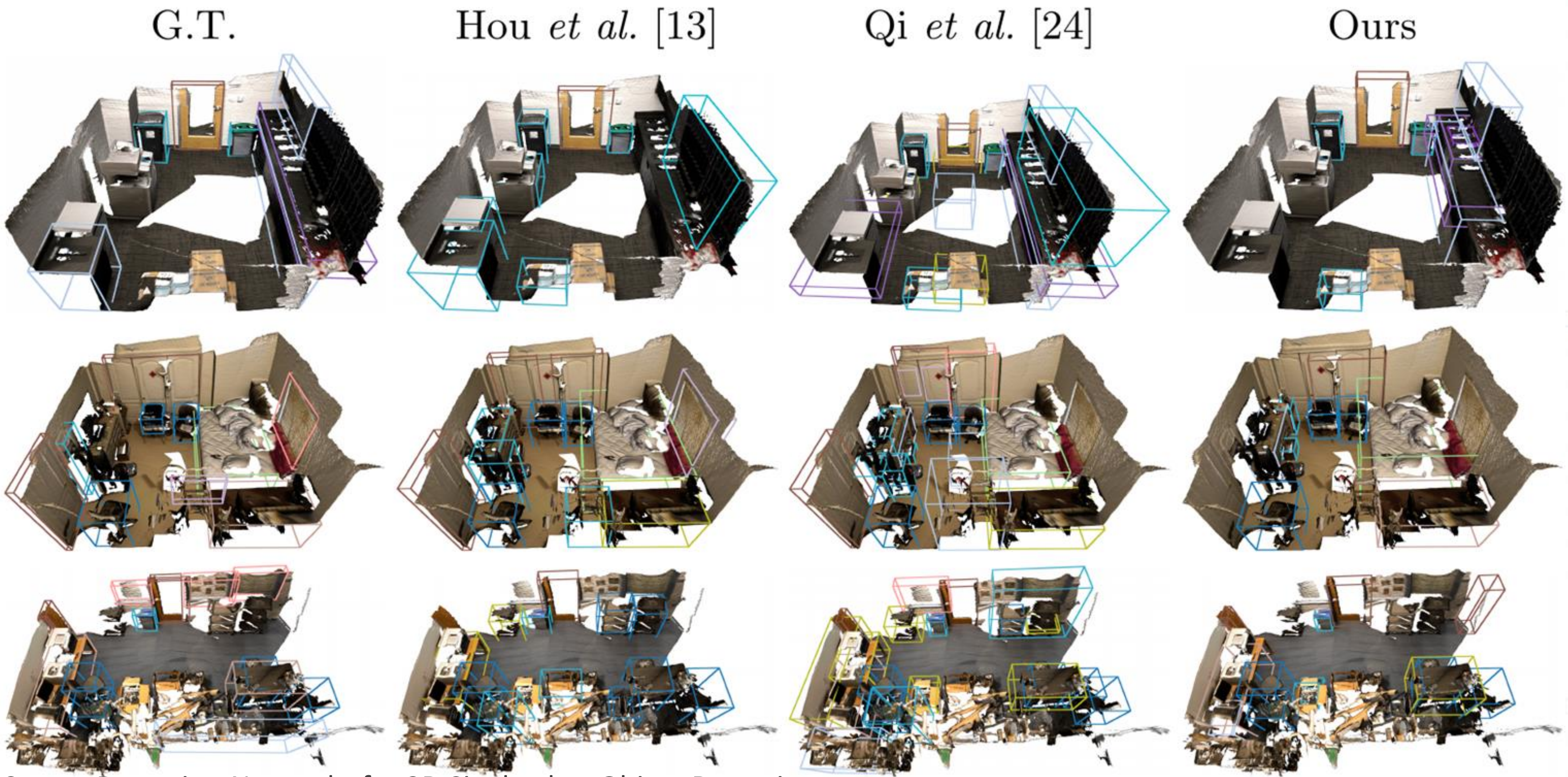


Comparison with previous SOTA - ScanNet

- Outperforms previous state-of-the-art by **4.2 mAP@0.25**
  - While being a single-shot detection
- While being **x3.7 faster**
  - runtime linear to # of points
  - runtime sublinear to floor area
  - ⇒ **free from curse of dimensionality!!**
- Minimal memory footprint
  - **x6** efficient to dense counterpart
- Maintains constant input density
  - Consistent information for scalability



Comparison with previous SOTA - ScanNet

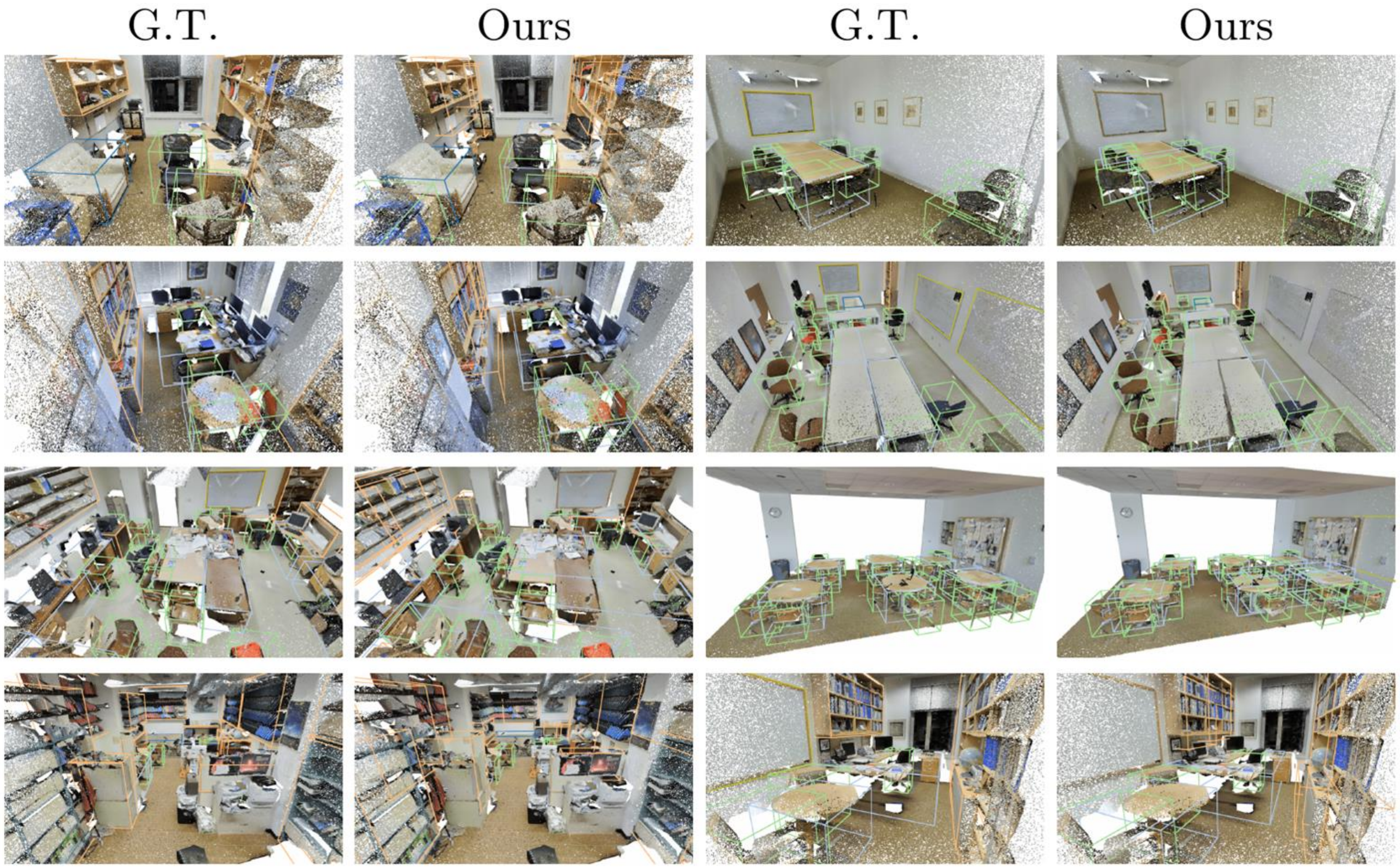


Comparison with previous SOTA - S3DIS

- Achieves state-of-the-art result
- Our method doesn't require crop-and-stitch post-processing unlike Yang et al.

IoU Thres.	Metric	Method	table	chair	sofa	bookcase	board	avg
0.25	AP	Yang *et al.* [36]*	27.33	53.41	9.09	14.76	29.17	26.75
		GSDN (ours)	73.69	98.11	20.78	33.38	12.91	47.77
	Recall	Yang *et al.* [36]*	40.91	68.22	9.09	29.03	50.00	39.45
		GSDN (ours)	85.71	98.84	36.36	61.57	26.19	61.74
0.5	AP	Yang *et al.* [36]*	4.02	17.36	0.0	2.60	13.57	7.51
		GSDN (ours)	36.57	75.29	6.06	6.46	1.19	25.11
	Recall	Yang *et al.* [36]*	16.23	38.37	0.0	12.44	33.33	20.08
		GSDN (ours)	50.00	82.56	18.18	18.52	2.38	34.33

Comparison with previous SOTA - S3DIS



Ablation study

Train without sparsity pruning

➔ Fails to train due to out of memory error

Train without Generative Sparse Tensor Decoder

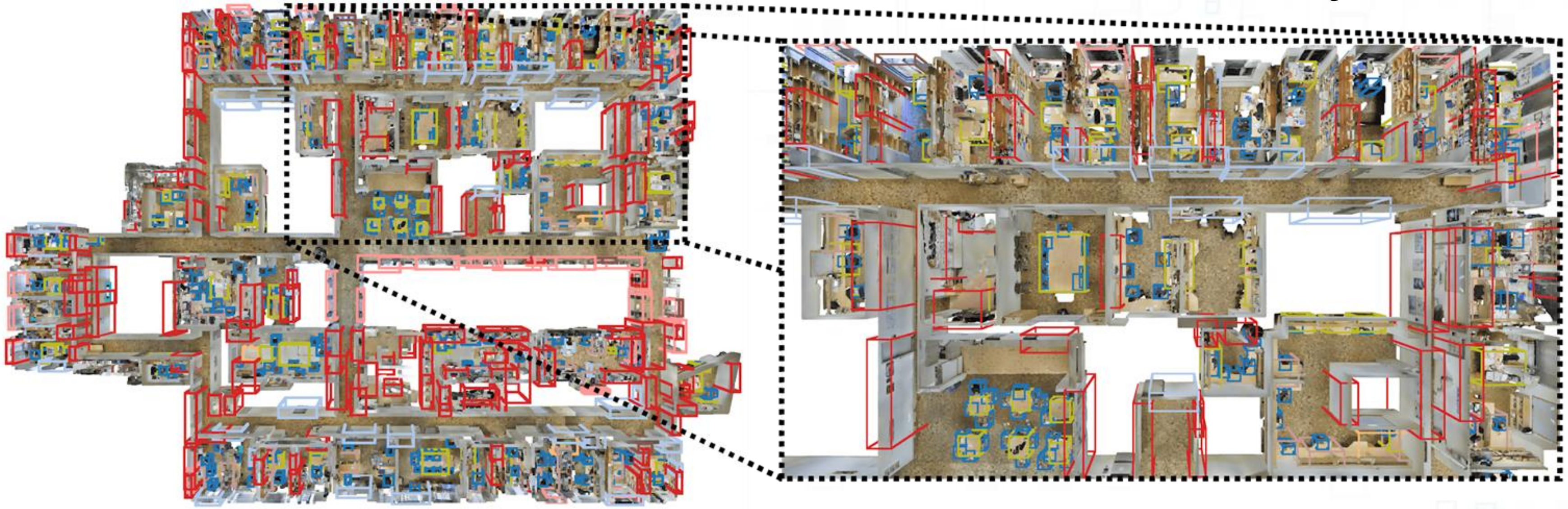
➔

backbone model	mAP@0.25	mAP@0.5
No decoder	52.1	24.6
Ours	**57.2**	**29.7**

Scalability and generalization - S3DIS

Train on small rooms, test on the the entire building 5 of S3DIS

- 78M points, 13984m^3 volume, and 53 rooms
- *Single* fully-convolutional network feed-forward
- Takes 20 seconds *including* data pre-processing and post-processing
- Use 5G GPU memory to detect 573 instances of 3D objects



Scalability and generalization - S3DIS

How does our method achieve high scalability and generalization capacity?

Consistent information regardless of the size of input:

- Fully-convolutional: translation invariant
- Consistent density of input: voxels. no fixed-sized random subsampling

Minimal runtime and memory footprint

- Fully sparse
  - Sparse encoder: sparse convolution
  - Sparse decoder: pruning to prevent cubic growth of generated coordinates

Conclusion

We propose **Generative Sparse Detection Networks**

- Efficiently processes large-scale 3D scene using **Sparse Convolution**
- **Generates** and **prunes** new coordinates to support anchor box centers

Which achieves

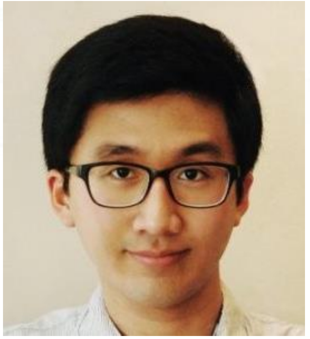
- Outperforms previous state-of-the-art by **4.2 mAP@0.25**
- While being **x3.7 faster** (and runtime grows **sublinear** to the volume)
- With **minimal memory footprint** (**x6** efficient than dense counterpart)
- Processes **unprecedently large scene** in a **single network feed-forward**

ECCV'20 ONLINE 23-28 AUGUST 2020

Thank you!

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

Collaborators



JunYoung Gwak
Stanford University

Chris Choy
NVIDIA

Silvio Savarese
Stanford University