CSC2457 3D & Geometric Deep Learning

Virtual Multi-view Fusion for 3D Semantic Segmentation

Date: February. 2nd, 2021

Presenter: Xiang Cao

Instructor: Animesh Garg

UNIVERSITY OF TORONTO

The Paper

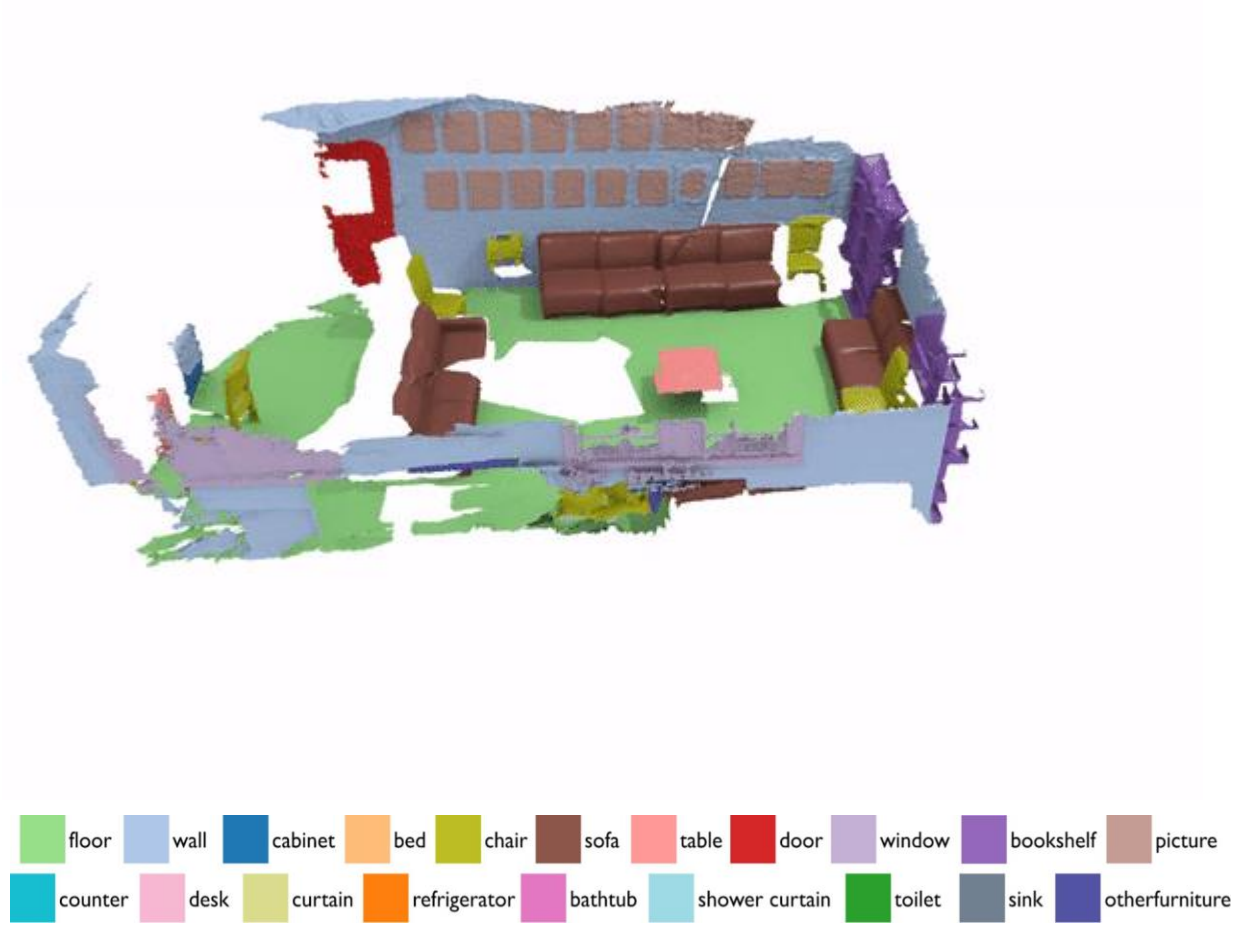
Virtual Multi-view Fusion for 3D Semantic Segmentation

Abhijit Kundu, Xiaoqi Yin, Alireza Fathi, David Ross,
Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru

Google Research

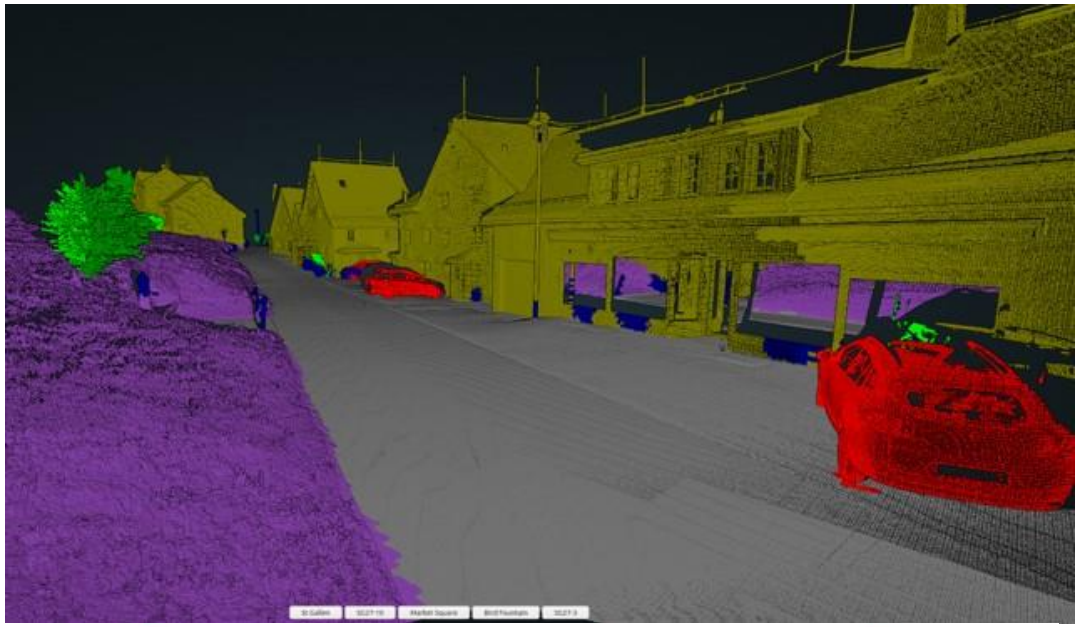
Abstract. Semantic segmentation of 3D meshes is an important problem for 3D scene understanding. In this paper we revisit the classic multiview representation of 3D meshes and study several techniques that make them effective for 3D semantic segmentation of meshes. Given a 3D mesh reconstructed from RGBD sensors, our method effectively chooses different virtual views of the 3D mesh and renders multiple 2D channels for training an effective 2D semantic segmentation model. Features from multiple per view predictions are finally fused on 3D mesh vertices to predict mesh semantic segmentation labels. Using the large scale indoor 3D semantic segmentation benchmark of ScanNet, we show that our virtual views enable more effective training of 2D semantic segmentation networks than previous multiview approaches. When the 2D per pixel predictions are aggregated on 3D surfaces, our virtual multiview fusion method is able to achieve significantly better 3D semantic segmentation results compared to all prior multiview approaches and competitive with recent 3D convolution approaches.

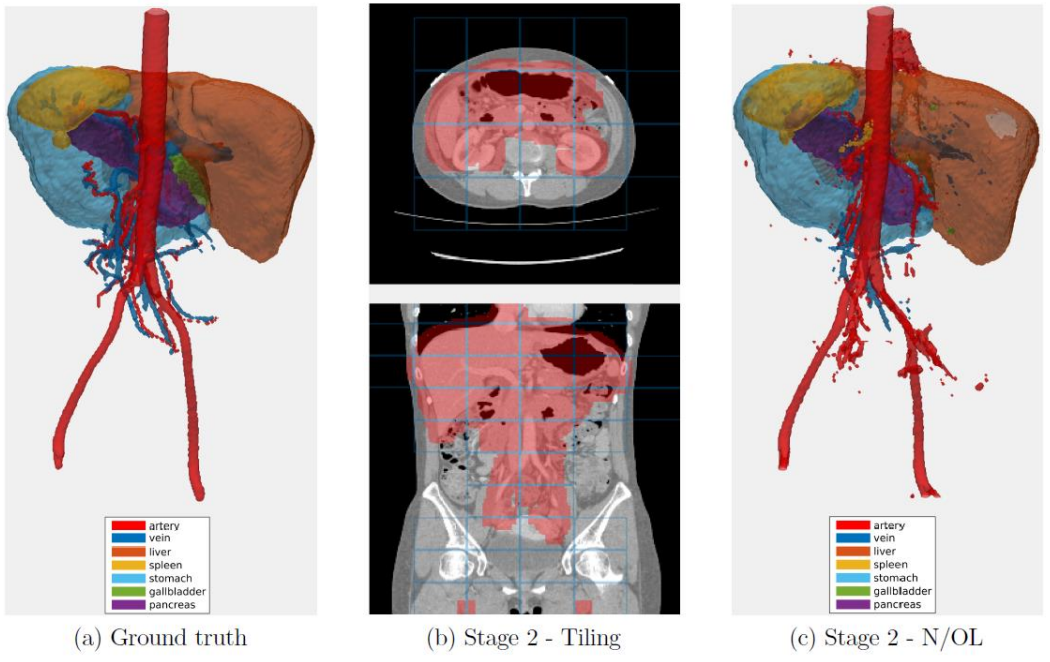
Keywords: 3D semantic segmentation, Scene Understanding



Motivation

- 3D Semantic Segmentation Applications:
- Autonomous Driving
- Robotics
- Biomedical imaging





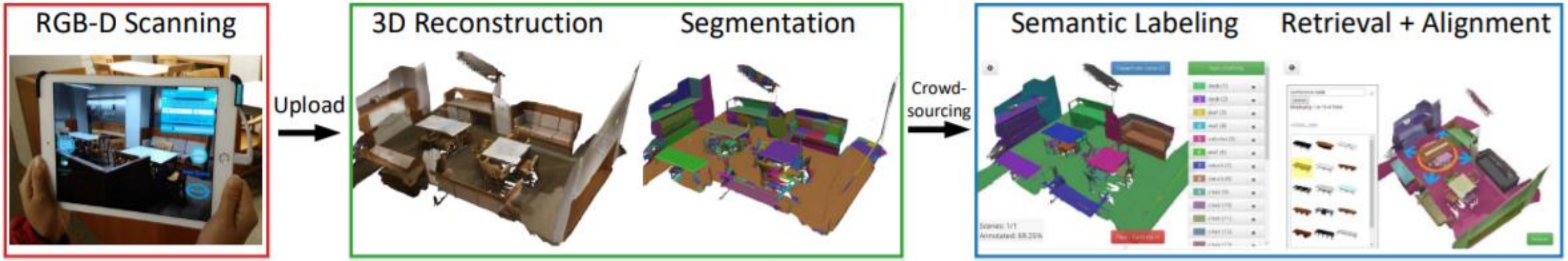
(a) Ground truth (b) Stage 2 - Tiling (c) Stage 2 - N/OL

Main Problem: 3D Semantic Segmentation

- Extract 3D geometry of a region of interest through segmentation
- Understand objects' shape, category and position
- Difficulties in surface determination in the 3D
- Previous works:
    1. Multi view labelling
    2. Native 3D Convolution
    3. Synthetic Data

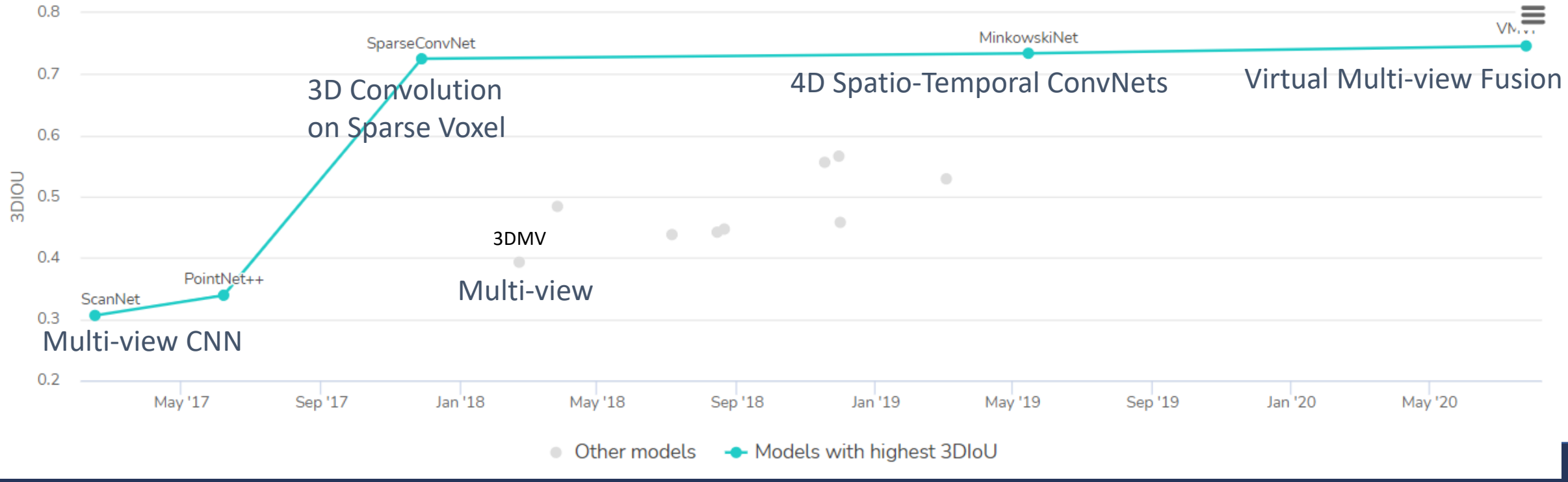
General Background : ScanNet dataset

- Richly-annotated 3D Reconstructions of Indoor Scenes
- 1513 indoor scenes with 2.5 Million Views
- Contains 3d camera poses and extrinsics
- *Volumetric and multi-view CNNs for object classification on 3D data.*



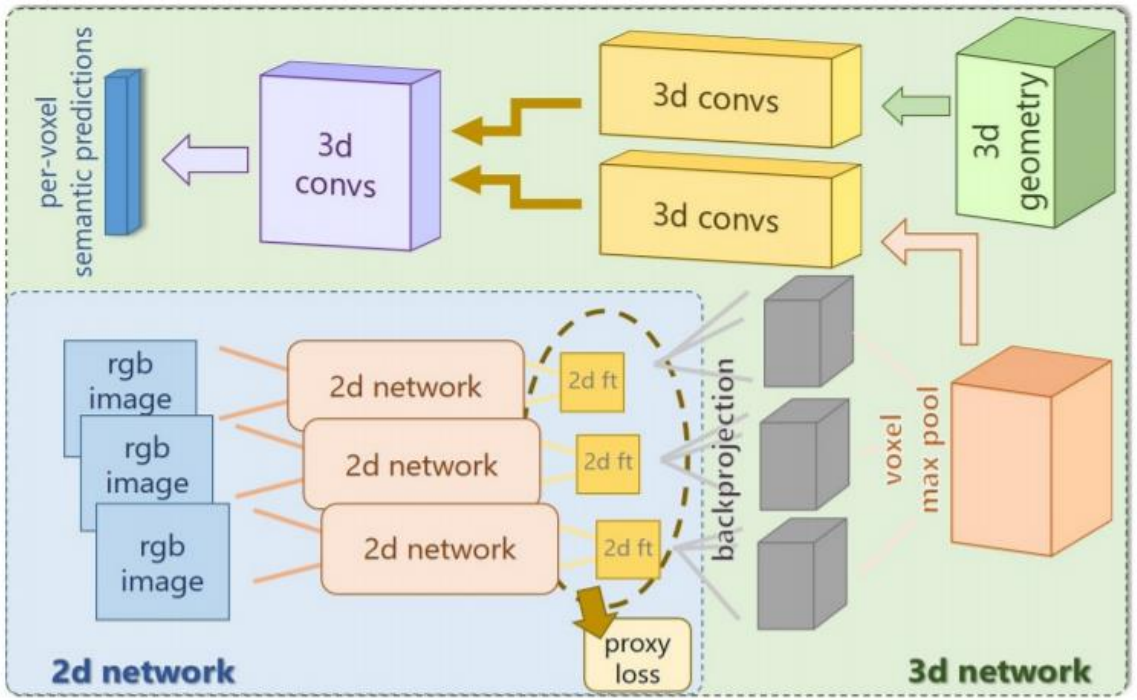
Prior Works and Their Limits

- Multi-view labeling: Used 2D-3D projection. restricted by limited fields of view, not scale invariant, lighting of certain views
- Native 3D Convolution : Point cloud or Sparse Voxel. Limited resolution comparing to 2D
- Synthetic data: limited domain adaption.



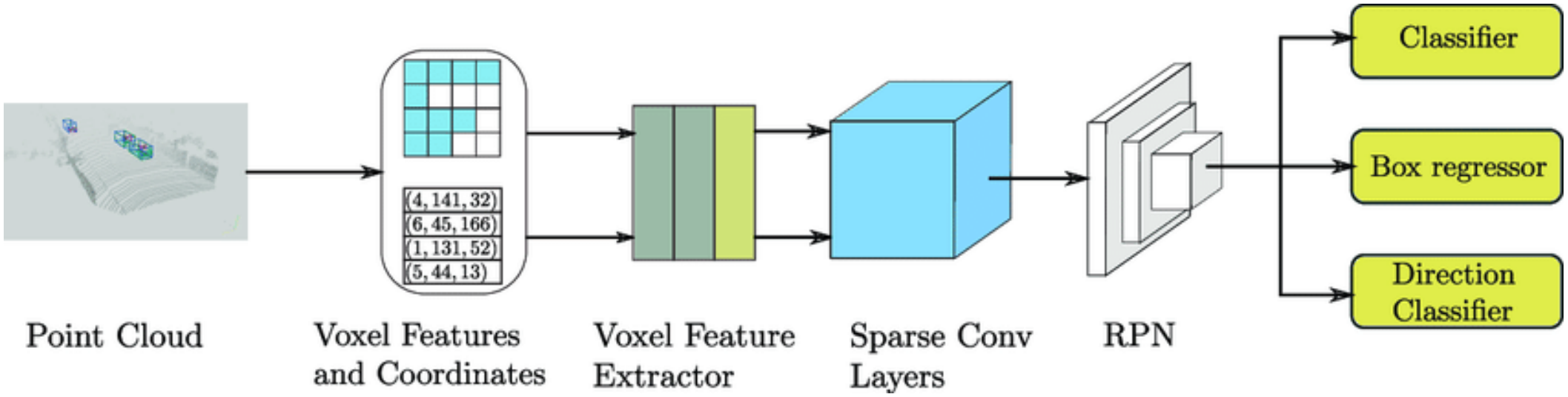
3DMV: Joint 3D-Multi-View Prediction for 3D Semantic Scene Segmentation

- Same author of ScanNet Dataset
- Use Multi-view method



3D Semantic Segmentation with Submanifold Sparse Convolutional Networks

- Native 3D Convolutional Networks
- Submanifold sparse convolutional networks SSCN
- Efficiency on high-dimensional sparse input data



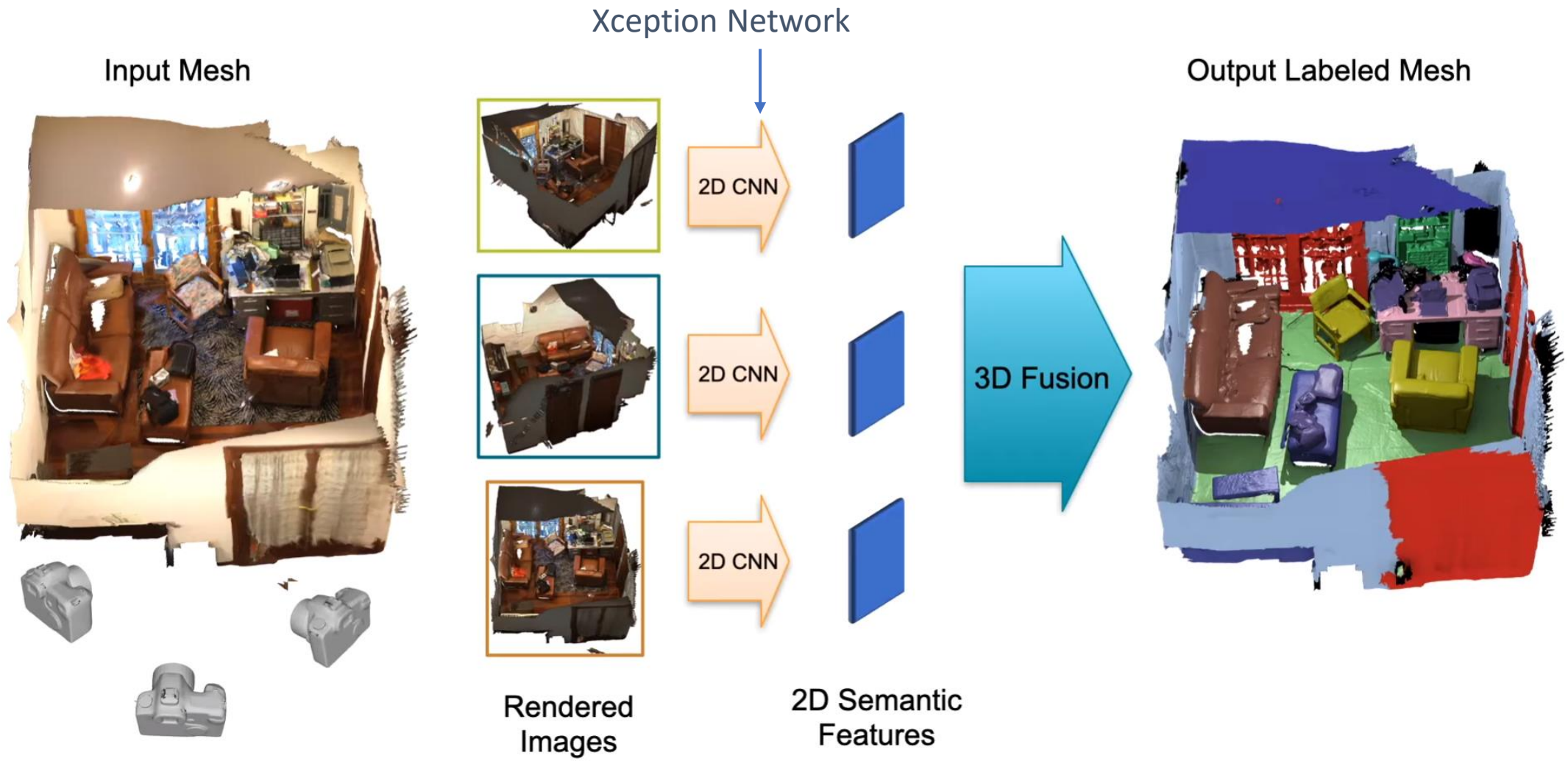
Contributions

- Problem: 3D Semantic Segmentation
- STOA performance achieved with simpler classic models
- Key innovation: 4 steps in data augmentation including adding virtual multi-view images
- Reopen the direction of Multi-view for future research

Problem Setting

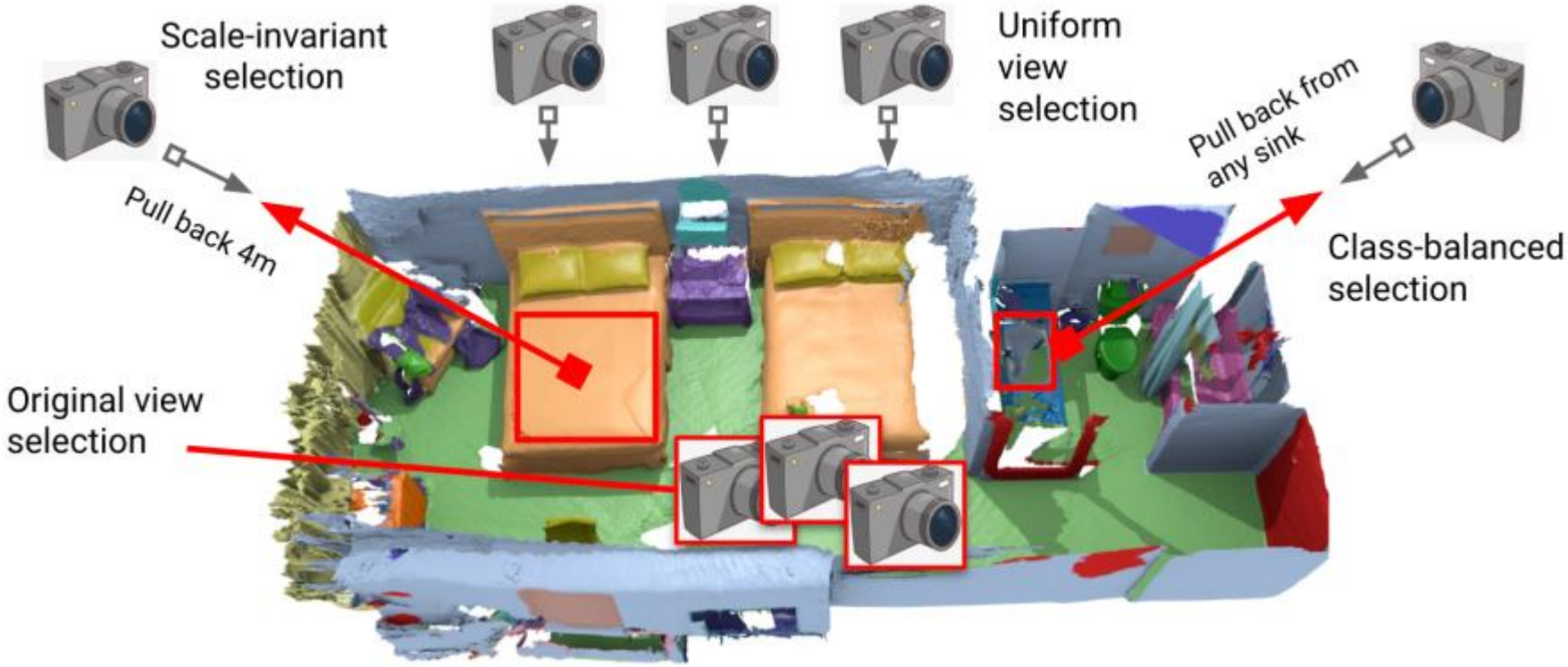
- A new multi view-based approach for 3D Semantic segmentation
- Input data: 3D mesh from RGB-D camera
- Use synthetic images rendered from virtual views
- Complete further data augmentation
- Performance Metrics: 3D IOU (Intersection over Union)

Approach :Overview



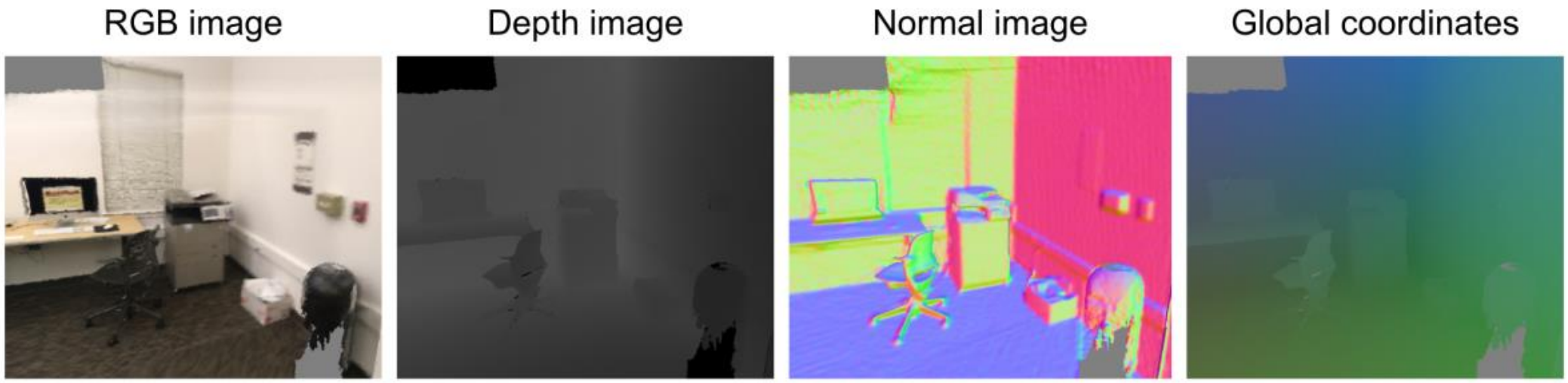
Approach : Step 1

- Virtual Multi-view synthesis



Approach : Step 2

- Adding additional channels

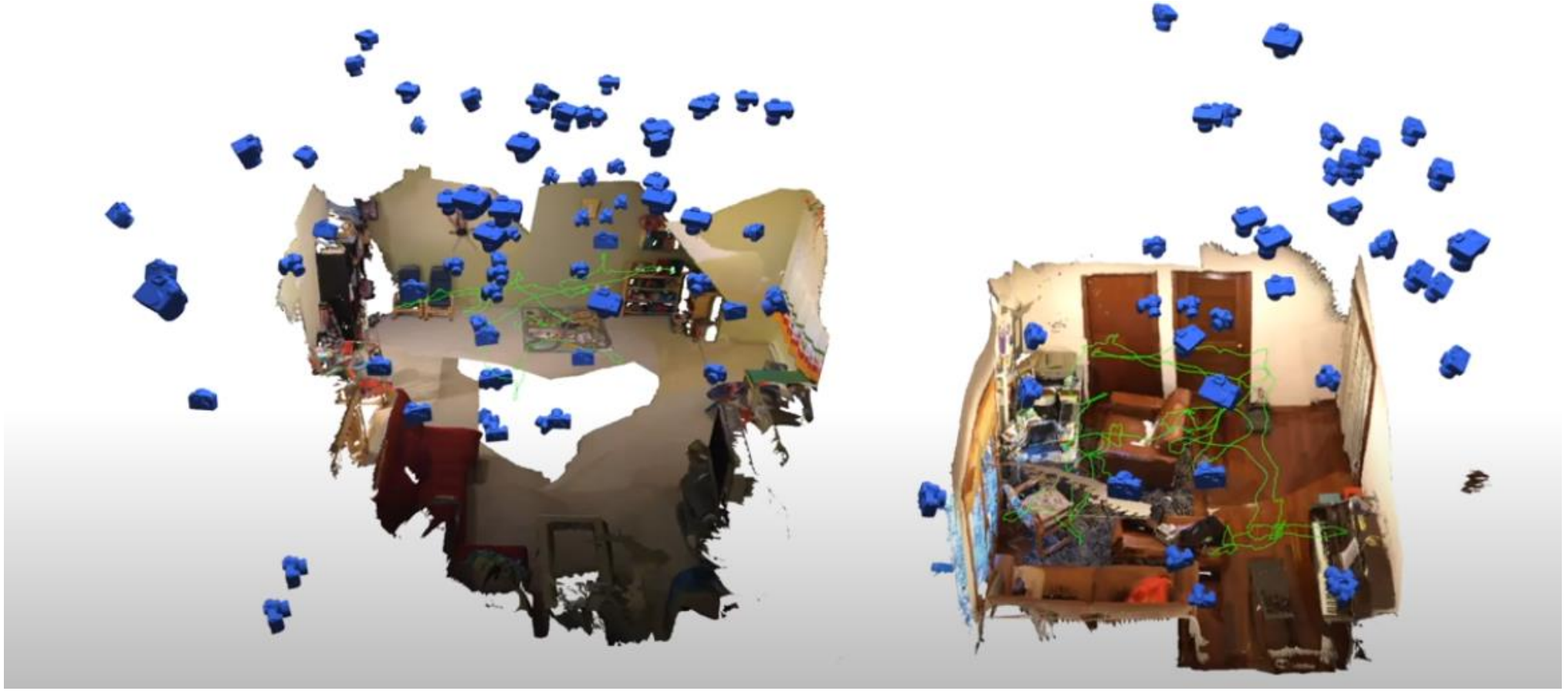


Approach : Step 3

- Expanding higher field of view
- Achieve larger spatial context

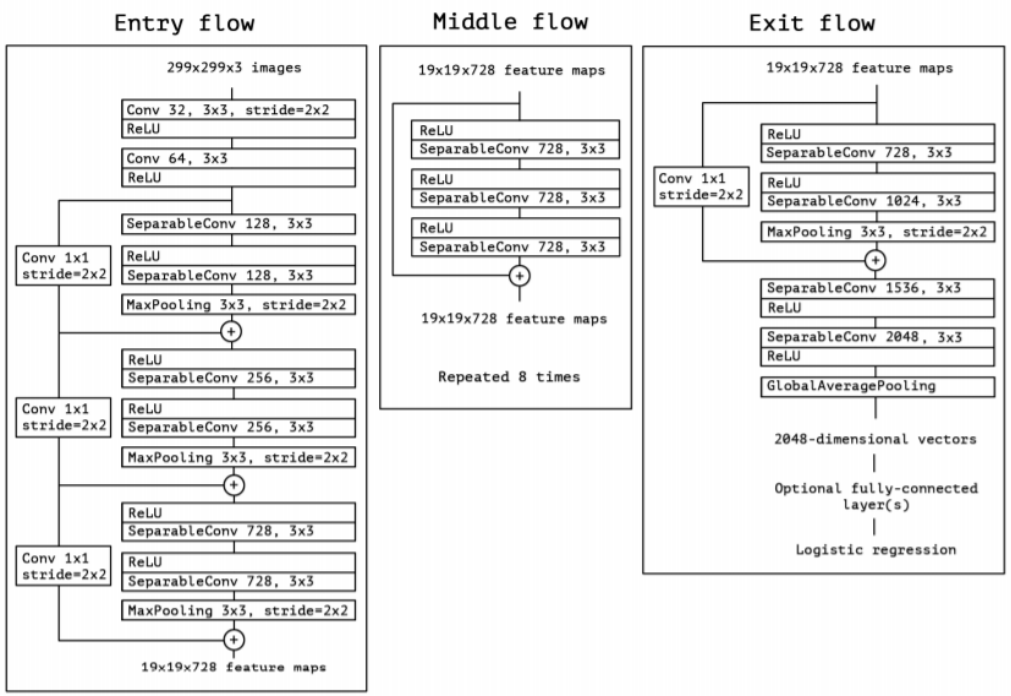
Approach : Step 4

- View Sampling



Approach: Xception Network

- Use transfer learning from 2D CNN
- Leverage larger 2D dataset: ImageNet, COCO
- Depth wise Separable Convolutions



Approach: 3D Fusion of 2D Features

- Project from 2D back to 3D
- Render depth channel on virtual views
- Project back to each virtual views according to depth.

$$\mathbf{x}_{k,i} = \mathbf{K}_i(\mathbf{R}_i\mathbf{X}_k + \mathbf{t}_i) \qquad c_{k,i} = \left\|\mathbf{X}_k - \mathbf{R}_i^{-1}\mathbf{t}_i\right\|_2$$

$$\mathcal{F}_k = \{\mathbf{f}_i(\mathbf{x}_{k,i}) \mid \mathbf{x}_{k,i} \in \mathcal{A}_i, |d_i(\mathbf{x}_{k,i}) - c_{k,i}| < \delta, \forall i \in \mathcal{I}\}$$

Experimental Results

- 1st place on ScanNet Val set
- 2nd place on test set
- Qualitative result:

Method	3D mIoU (val split)	3D mIoU (test split)
PointNet [30]	53.5	55.7
3DMV [10]	-	48.4
SparseConvNet [11]	69.3	72.5
PanopticFusion [28]	-	52.9
PointConv [43]	61.0	66.6
JointPointBased [5]	69.2	63.4
SSMA [39]	-	-
KPConv [38]	69.2	68.4
MinkowskiNet [7]	72.2	73.6.
PointASNL [44]	63.5	66.6
OccuSeg [13]	-	**76.4**
JSENet [16]	-	69.9
Ours	**76.4**	74.6

Discussion of Results : Ablation Study

SensorType	Channels	Intrinsics	Extrinsics	3D mesh IoU
Real	RGB	Original	Original	60.1
Virtual	RGB	Original	Original	63.2
Virtual	RGB + Normal+ Coordinates	Original	Original	66.1
Virtual	RGB + Normal+ Coordinates	High FOV	Original	67.9
Virtual	**RGB + Normal + Coordinates**	**High FOV**	**View sampling**	**70.1 (+ 10.1)**

Limitations

- Require highly articulated feature engineering
- Scalability
- Computation cost on simulation instead of training
- Noise/distortion exist on virtual images

Contributions (Recap)

- Problem: 3D Semantic Segmentation
- STOA performance achieved with simpler classic models
- Key innovation: 4 steps in data augmentation including adding virtual multi-view images
- Reopen the direction of Multi-view for future research