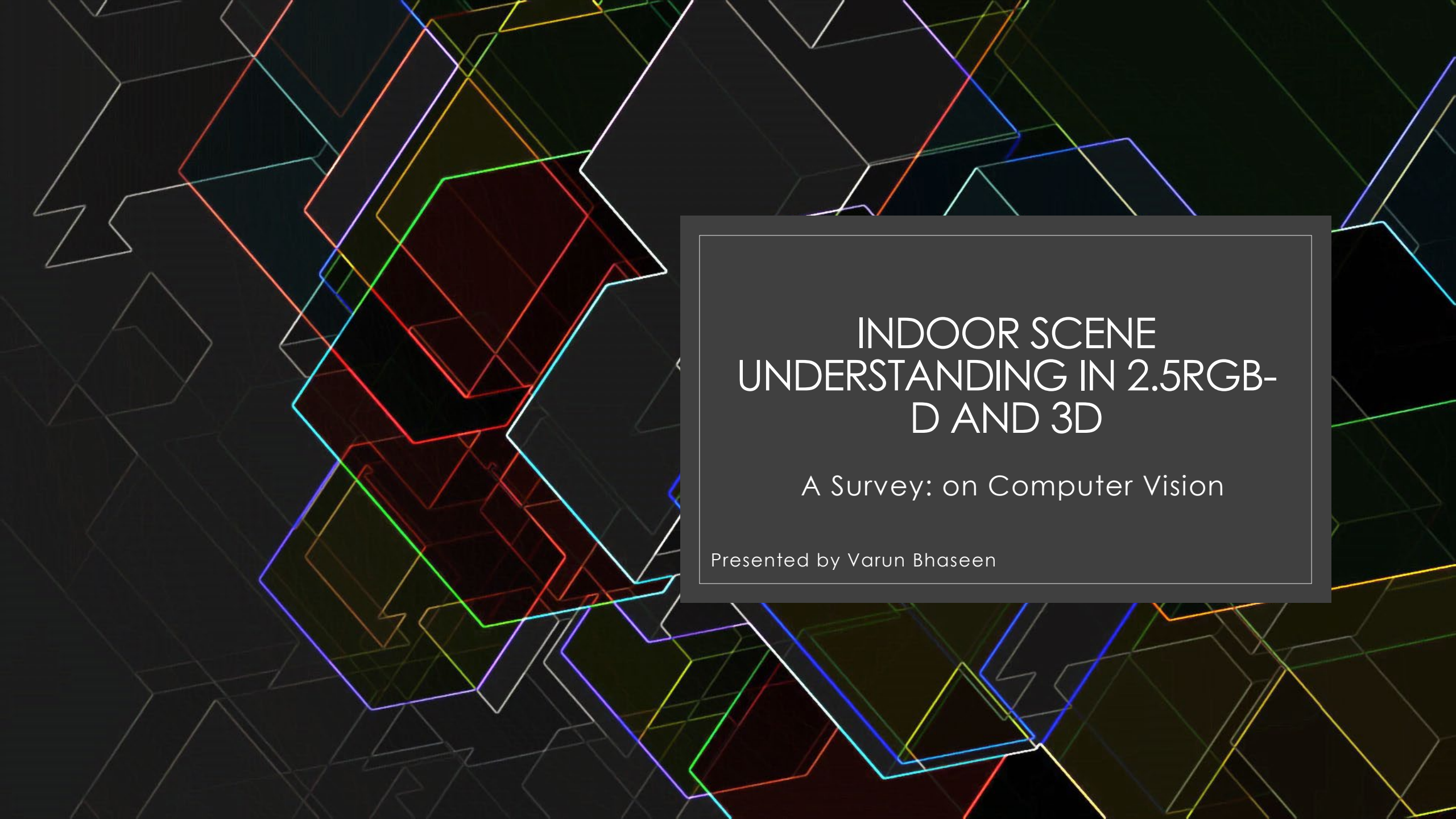


INDOOR SCENE UNDERSTANDING IN 2.5RGB-D AND 3D

A Survey: on Computer Vision

Presented by Varun Bhaseen

INDOOR SCENE UNDERSTANDING IN 2.5RGB-D AND 3D

AUTHORED BY:

HONGWEI YONG, JIANQIANG HUANG, XIANSHENG HUA, LEI ZHANG

DATE PUBLISHED:

10 JANUARY 2019

ARTICLE LINK:

HTTPS://ARXIV.ORG/ABS/1803.03352

Introduction

An image to a machine can be represented as nothing but a grid of numbers.

In order to develop a model which has a comprehensive understanding of visual content, it is necessary to uncover the underlying geometric and semantic clues between various scene elements present in its field of vision.

It is also required to comprehend both the apparent and hidden relationships present between scene elements.

Approach for Computer Vision Modelling

Image Recognition

Object Detection

Semantic Segmentation

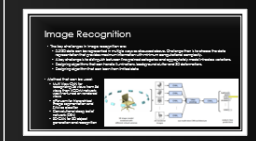
Physics based reasoning

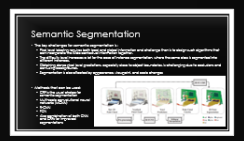
Object pose estimation

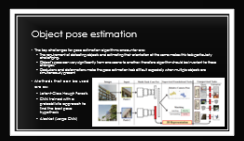
3d reconstruction from RGB-D

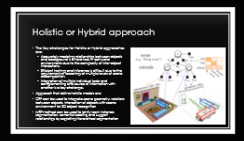
Saliency Prediction

Holistic or hybrid approach









Data Representation

Point Cloud

Voxel Representation

3d Mesh

Depth Channel and Encodings

Octree representation

Stixels

Truncated Signed Distance function

Constructive Solid Geometry

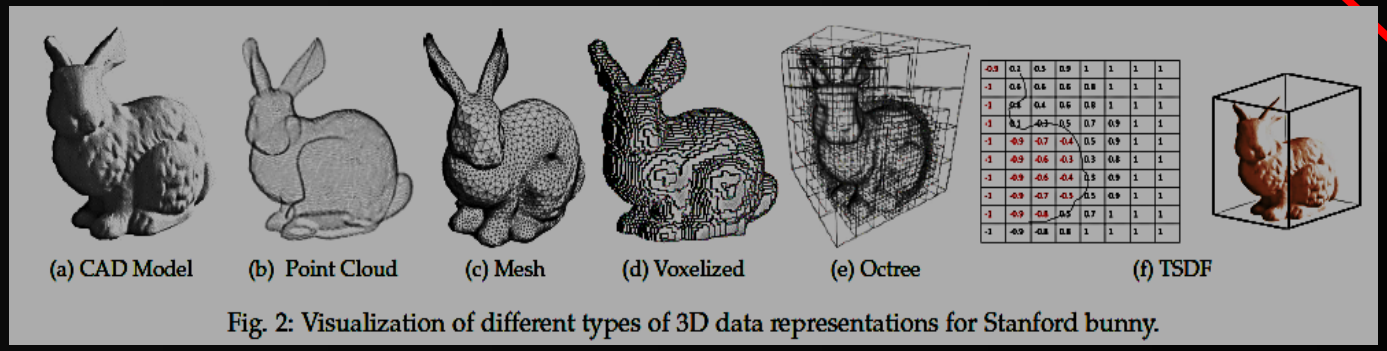


Fig. 2: Visualization of different types of 3D data representations for Stanford bunny.

Datasets: List of popular datasets

Dataset	NYUv2 [10]	SUN3D [29]	SUN RGB-D [30]	Building Parser [31]	Matterport 3D [32]	ScanNet [33]	SUNCG [34]	RGBD Object [35]	SceneNN [36]	SceneNet RGB-D [37]	PiGraph [38]	TUM [39]	Pascal 3D+ [40]
Year	2012	2013	2015	2017	2017	2017	2016	2011	2016	2016	2016	2012	2014
Type	Real	Real	Real	Real	Real	Real	Synthetic	Real	Real	Synthetic	Synthetic	Real	Real
Total Scans	464	415	-	270	-	1513	45,622	900	100	57	63	39	-
Labels	1449 images	8 scans	10k images	70k images	194k images	1513 scans	130k images	900 scans	100 scans	5M images	21 scans*	39 scans	24k images
Objects/Scenes	Scene	Scene	Scene	Scene	Scene	Scene	Scene	Object	Scene	Scene	Scene	Scene	Object
Scene Classes	26	254	47	11	61	707	24	-	-	5	30	✗	-
Object Classes	894	-	800	13	40	50 - at least	84	51	50 - at least	255	5 subjects	✗	12
In/Outdoor	Indoor	Indoor	Indoor	Indoor	Indoor	Indoor	Indoor	Indoor	Indoor	Indoor	Indoor	Indoor	In+Out
Available Data Types													
RGB	✓	-	✓	✓	✓	-	-	✓	-	✓	✗	-	✓
Depth	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✗	-	✗
Video	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓	✗
Point cloud	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	-	✗	✗
Mesh/CAD	✗	✗	✗	✓	✓	✓	✓	✗	✓	✗	-	✗	✓
Available Annotation Types													
Scene Classes	✓	✗	✓	✓	✓	✗	✗	✗	✓	✓	✗	✗	✗
Semantic Label	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗
Object BB	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✗	✗	✓
Camera Poses	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
Object Poses	✓	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✓
Trajectory	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗
Action	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗

-: means information not available, ∗: Average reported; 4.9 actions annotated per scan and there are 298 actions with 8.4s length available.

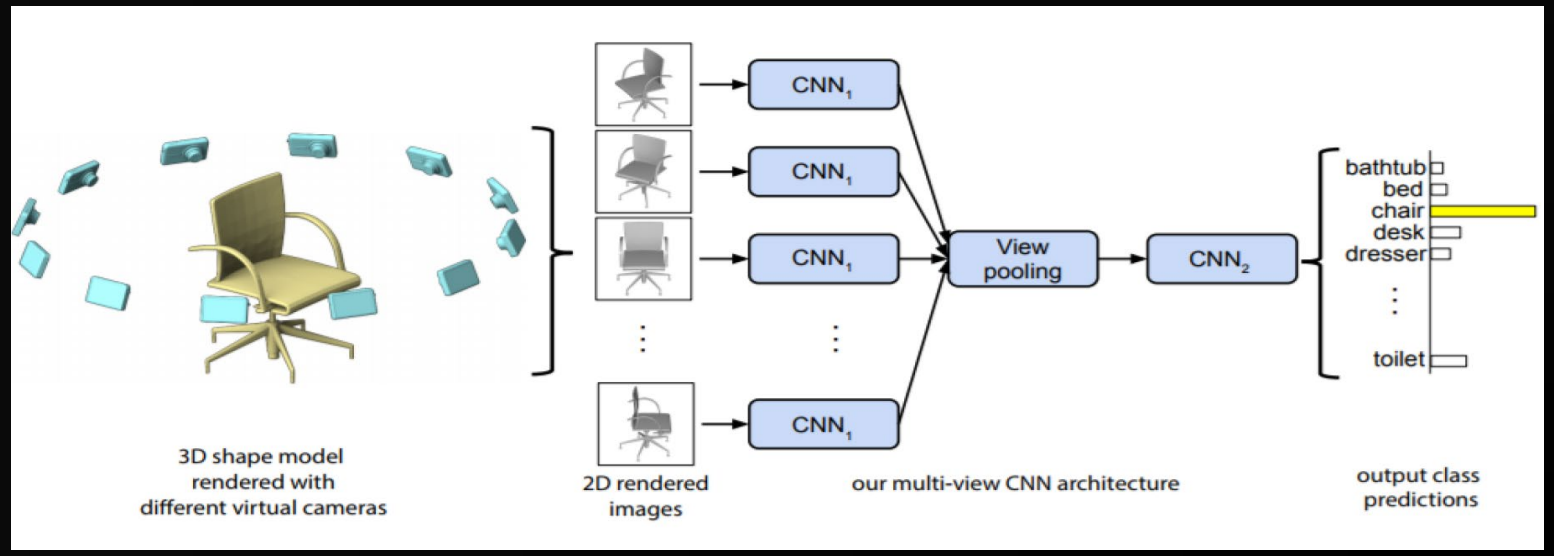
THANK YOU

Article Link: https://medium.com/@varun.bhaseen/an-approach-for-computer-vision-indoor-scene-understanding-in-2-5rgb-d-and-3d-b77a133574a0

Image Recognition

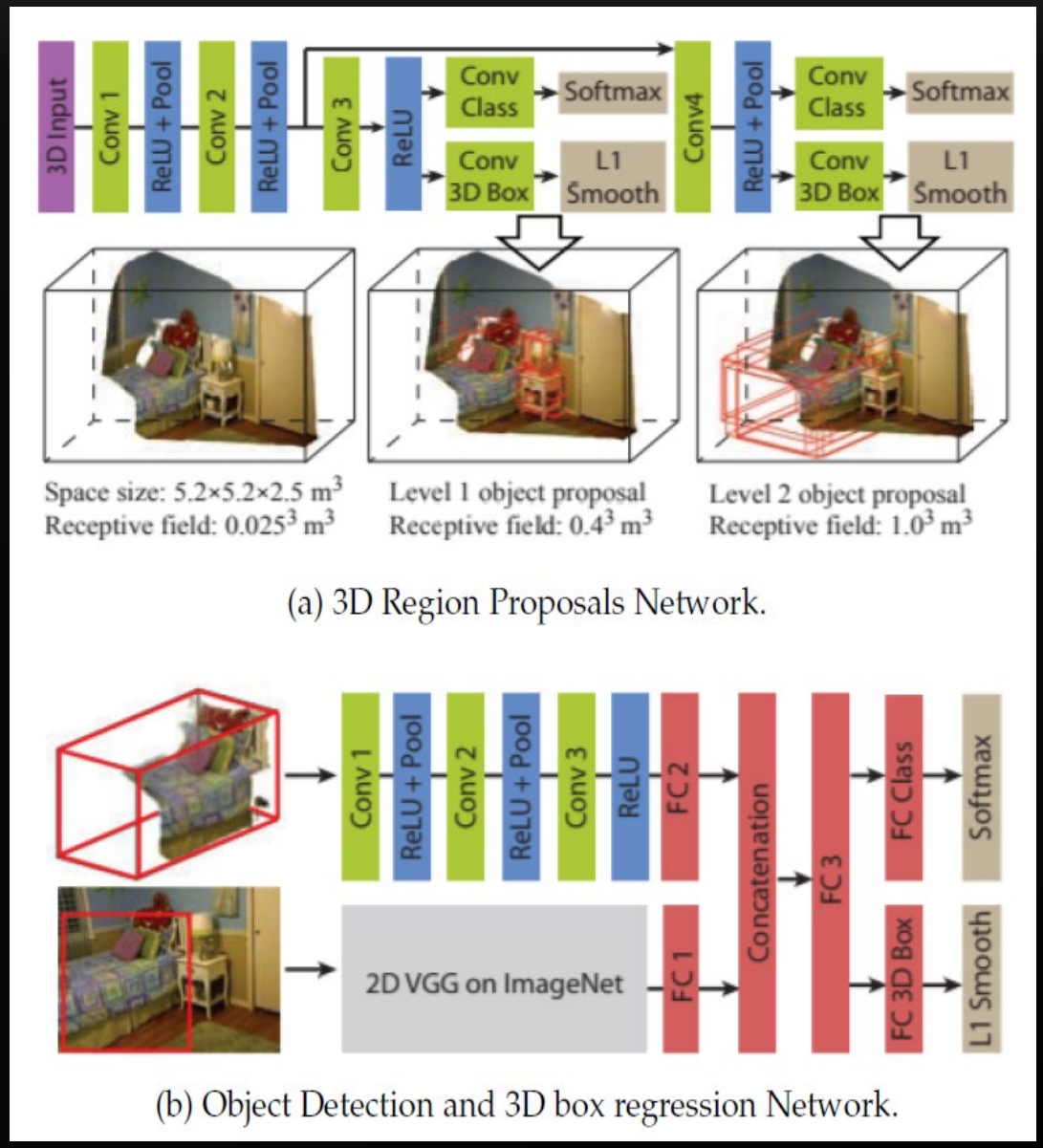
- The key challenges in image recognition are:
  - 2.5/3D data can be represented in multiple ways as discussed above. Challenge then is to choose the data representation that provides maximum information with minimum computational complexity.
  - A key challenge is to distinguish between fine grained categories and appropriately model intraclass variations.
  - Designing algorithms that can handle illuminations, background clutter and 3D deformations.
  - Designing algorithm that can learn from limited data

- Method that can be used:
  - Multi View CNN for recognizing 2d views from 3d views then VGG-M network was fine-tuned on rendered views
  - gPb-ucm for hierarchical image segmentation and SVM as classifier
  - Convolutional deep belief network (DBN)
  - 3D-GAN for 3D object generation and recognition



Object Detection

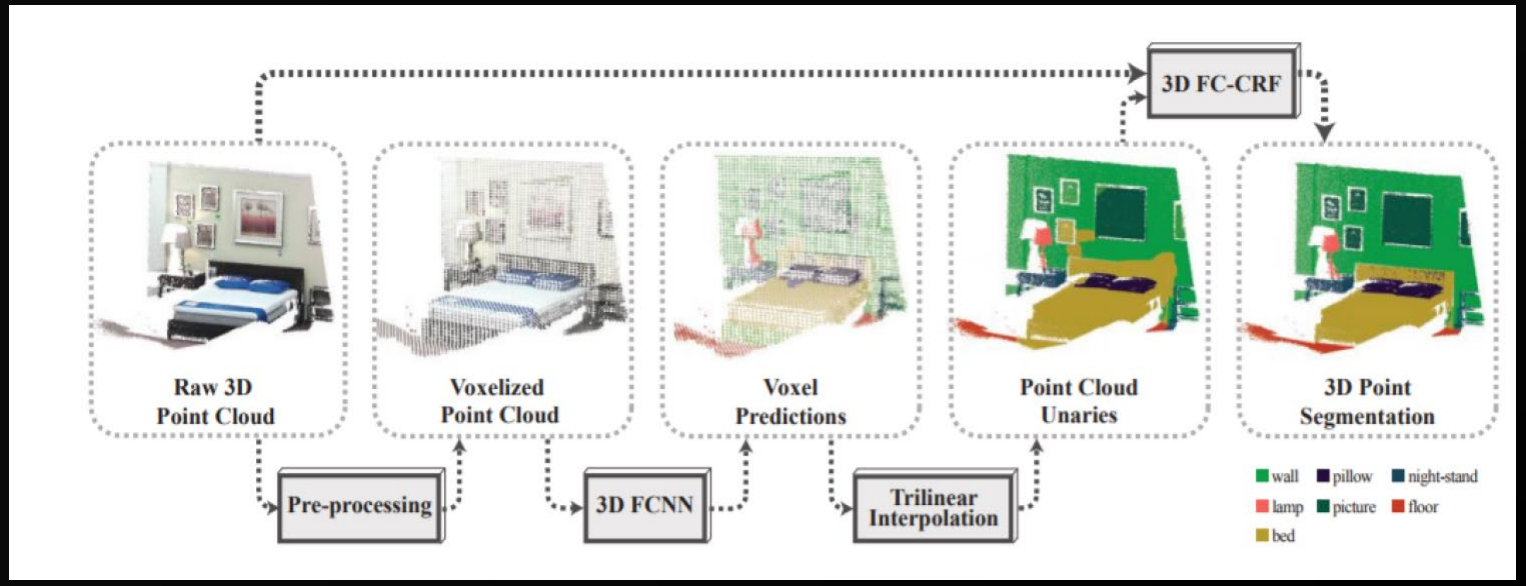
- Key challenges for object detection are:
  - Real world environments can be highly cluttered and object identification in such environments is very
  - challenging.
  - Detection algorithm should also be able to handle viewpoint and illuminations variations and deformations.
  - In many scenarios, it is necessary to understand the scene context to successfully detect objects.
  - Objects categories have a long-tail (imbalanced) distribution, which makes it challenging to model the infrequent classes
- Methods that can be used are:
  - A CRF model can be used to model relationship between objects and regions
  - RPN can be used to produce 3d regional proposal followed by 3d & 2d CNN
  - Fast R-CNN can also be used.



Semantic Segmentation

- The key challenges for semantic segmentation is :
  - Pixel level labeling requires both local and global information and challenge then is to design such algorithms that can incorporate the wide contextual information together.
  - The difficulty level increases a lot for the case of instance segmentation, where the same class is segmented into different instances.
  - Obtaining dense pixel level predictions, especially close to object boundaries, is challenging due to occlusions and confusing backgrounds.
  - Segmentation is also affected by appearance, viewpoint, and scale changes

- Methods that can be used:
  - CRF is the usual choice for semantic segmentation
  - Multi-scale convolutional neural networks (MCNN)
  - R-CNN
  - FCN
  - Also combination of both CNN and CRFs for improved segmentations

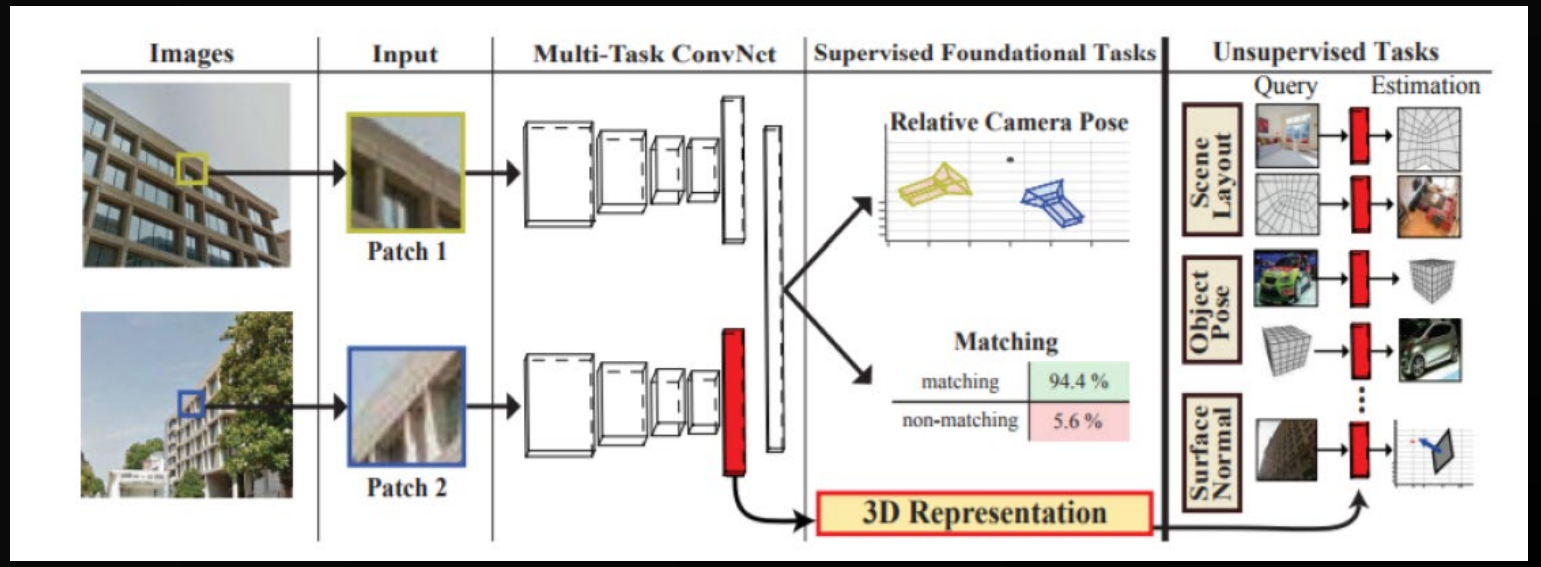


Physics based reasoning

- The key challenges that has been faced for physics-based reasoning is:
  - This task requires starting with extremely limited information (e.g., a still image) and performing extrapolation to predict rich information about scene dynamics.
  - A desirable characteristic is to adequately model prior information about the physical world.
  - Physics based reasoning requires algorithms to reason about the contextual information
- Can be subdivided into multiple areas like:
  - Dynamics prediction
  - Support relationships
  - Stability Analysis
  - Hazard detection
  - Occlusion reasoning

Object pose estimation

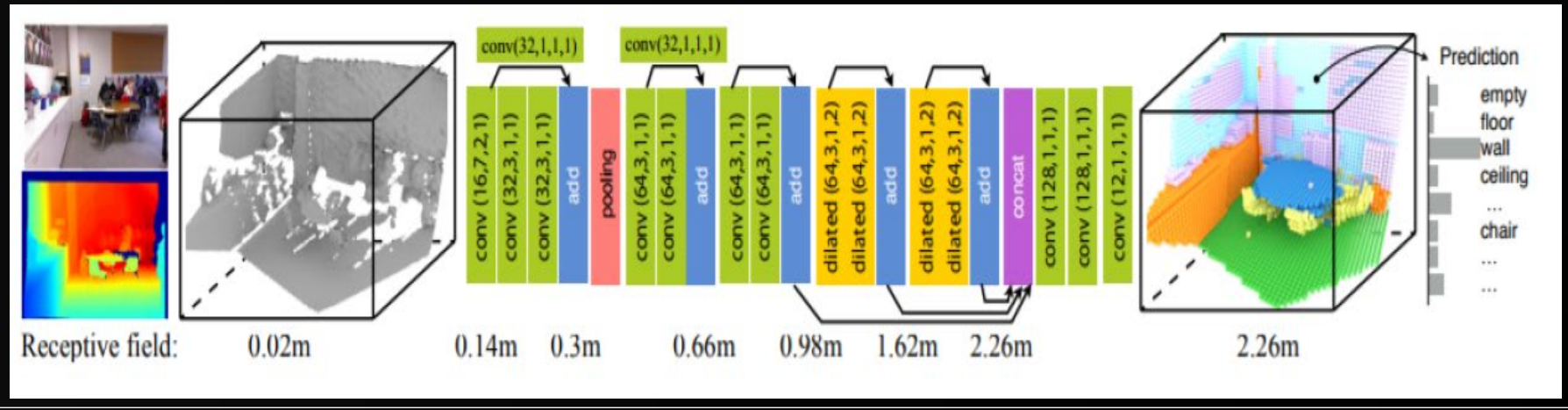
- The key challenges for pose estimation algorithms encounter are:
  - The requirement of detecting objects and estimating their orientation at the same makes this task particularly challenging.
  - Object's pose can vary significantly from one scene to another; therefore algorithm should be invariant to these changes.
  - Occlusions and deformations make the pose estimation task difficult especially when multiple objects are simultaneously present
- Methods that can be used are as:
  - Latent-Class Hough Forests
  - CNN trained with a probabilistic approach to find the best pose hypothesis
  - AlexNet (Large CNN)



3D Reconstruction from RGB-D

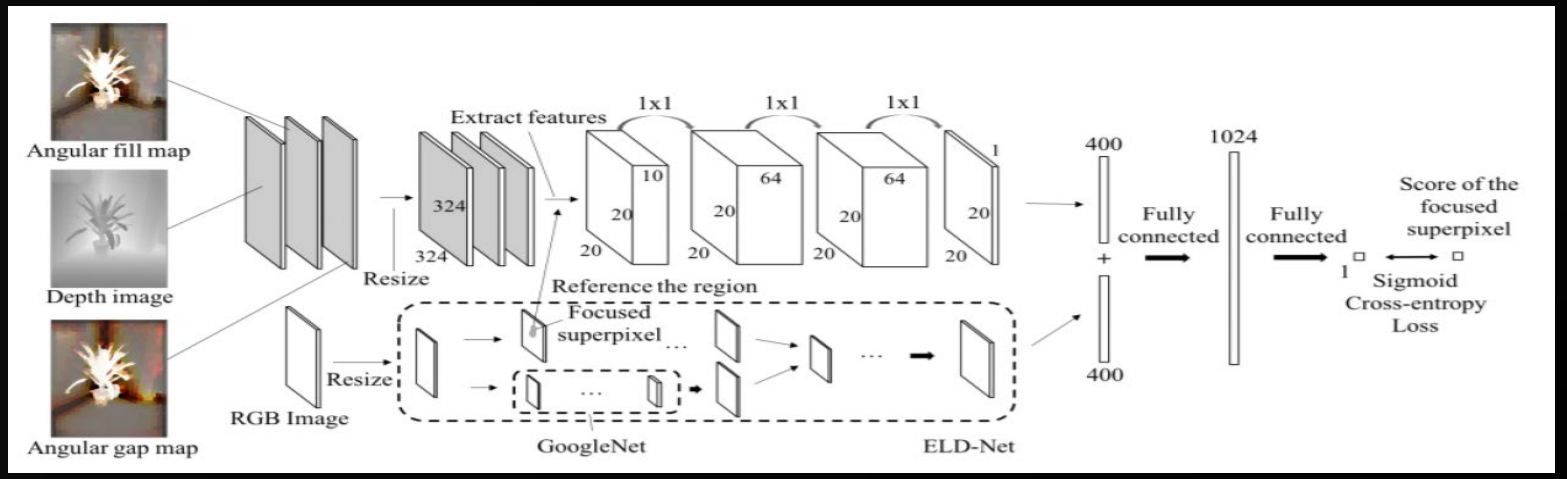
- The key challenges for 3D reconstruction are:
  - Complete 3D reconstruction from incomplete information is an ill-posed problem with no unique solution.
  - This problem poses a significant challenge due to sensor noise, low depth resolution, missing data and quantization errors.
  - It requires appropriately incorporating external information about the scene or object geometry for a successful reconstruction

- Methods that can be used are:
  - A 3D CNN can be used to jointly perform semantic voxel labeling and scene completion from a single RGB-D image
  - Convolutional LSTM based recurrent network can be used as well
  - TSDF based network
  - OctNet and SSCnet



Saliency Prediction

- The key challenges for saliency prediction task are:
  - Saliency is a complex function of different factors including appearance, texture, background properties, location, depth etc. It is a challenge to model these intricate relationships.
  - It requires both top-down and bottom-up cues to accurately model objects saliency.
  - A key requisite is to adequately encode the local and global context
- Methods that can be used for Saliency prediction are:
- Multistage models need to be used for saliency prediction where local, global and background contrast-based cues are used to predict a rough estimate of saliency



Holistic or Hybrid approach

- The Key challenges for Holistic or Hybrid approaches are:
  - Accurately modeling relationships between objects and background is a hard task in real-world environments due to the complexity of inter-object interactions.
  - Efficient training and inference is difficult due to the requirement of reasoning at multiple levels of scene decomposition.
  - Integration of multiple individual tasks and complementing one source of information with another is a key challenge.
- Approach that define holistic models are:
- CRF can be used to integrate scene geometry, relations between objects, interaction of objects with scene environment for 3D object recognition
- MRF method can be used to jointly learn instance segmentation, semantic labeling and support relationships by exploiting hierarchical segmentation