IEEE Transaction on Pattern Analysis and Machine Intelligence

Hierarchical Surface Prediction

Christian Häne, Shubham Tulsiani, Jiterndra Malik

University of California, Berkeley(UCB)

Published 30 January 2019

Slides compiled by Mengzhou Li

Contribution

- Proposed the hierarchical surface prediction (HSP) method which is able to generate 3D geometry prediction with high resolution voxel grids from one color/depth image.

Background

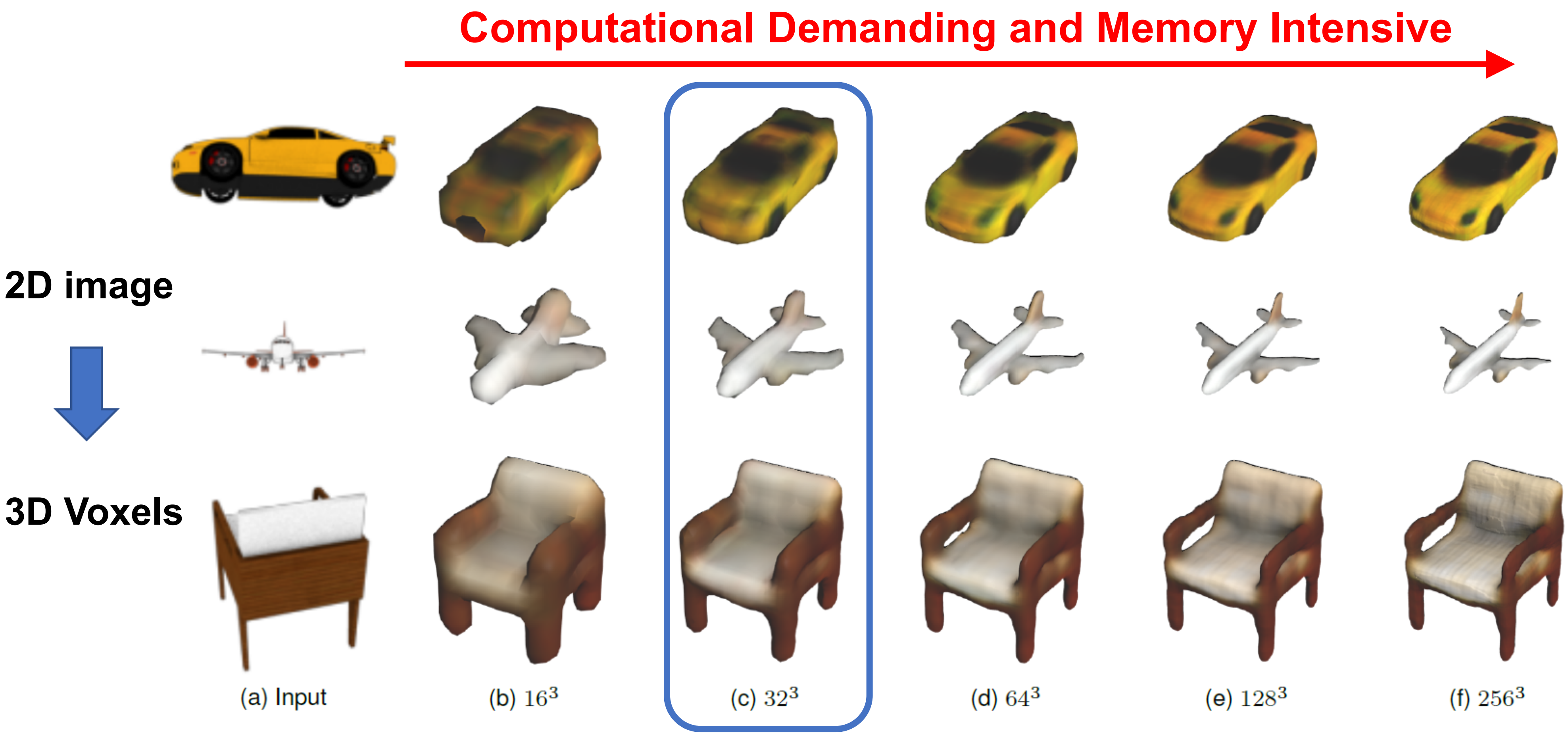
- ➢Task: 3D voxels prediction from 2D image inputs

- ➢Traditional method (from a large collection of multi-view images):
  - Dense matching
  - Minimization of reprojection errors

- ➢Recent method (from single images)
  - CNNs which directly map an input image to the geometry voxel grids

Limitations of current CNNs

➢ **Cubic growth of the volume with increasing resolution**



Common resolution

32^3

Computational Infeasible for Higher Resolution

Principle behind HSP

➢Cubic Growth => Quadratic Growth

3D object = 3D volume or 2D mesh surface

Surfaces are only two dimensional.

$X^3 \Rightarrow X^2$

2D image	2D image
↓	↓
3D Voxels	2D Surfaces

➢Observation basis

Only **a few of the voxels** are in **the vicinity of** the object's **surface**.
Most voxels are "boring" either completely inside or outside the object.

➢Principle

To only predict voxels around the surface

Computationally efficient
$32^3 \Rightarrow 256^3$

HSP ideas

2D image

3D Voxels

Labels for voxels:

- Free space (outside)
- Occupied space (inside)

One to End

Benefits:

- Computationally efficient for high resolution
- Better for surface properties, i.e., color
  - Colors are defined around the surface, while voxel far away from the surface won't get assigned
  - But assignment is unclear for traditional method

2D image

2D Surfaces

5-Level Coarse to Fine Resolution

Labels for voxels:

- Free
- Boundary
- Occupied

subdivide

1 => 8

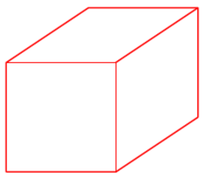
One voxel to 8 child nodes

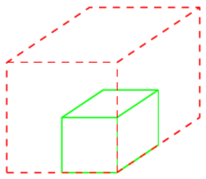
Free, Boundary, Occupied

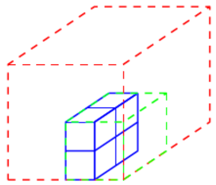
Free, Boundary, Occupied

Free, Boundary, Occupied

Free, Boundary, Occupied







Networks

- ➢ **Basic CNNs**

Encoder --- Decoder

Input / color/depth image

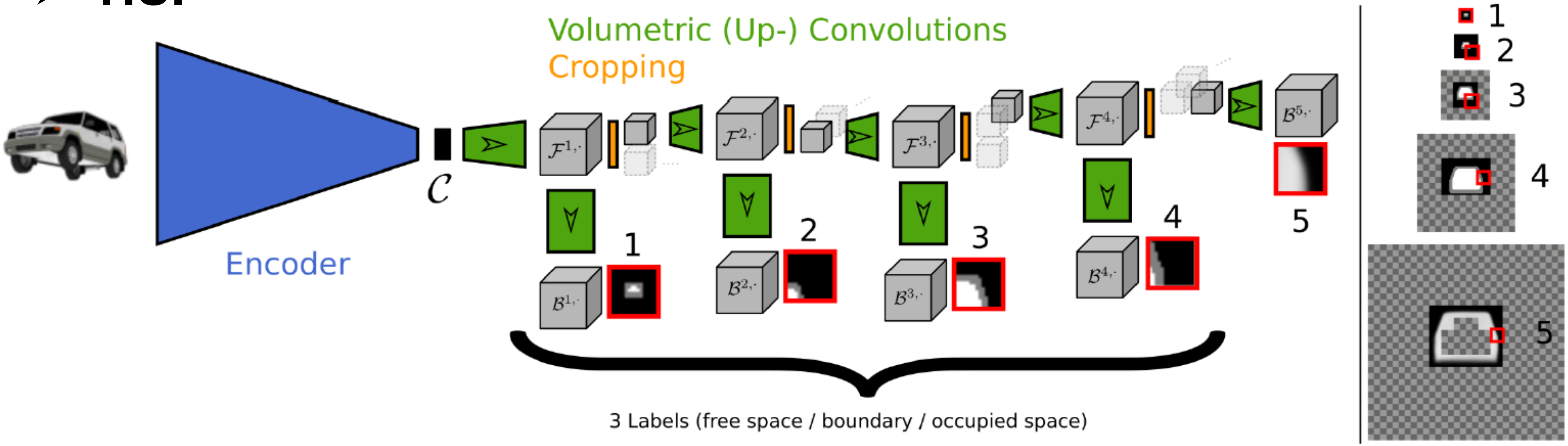
Shape code *C* a feature vector

Voxel grid *V* 3D geometry

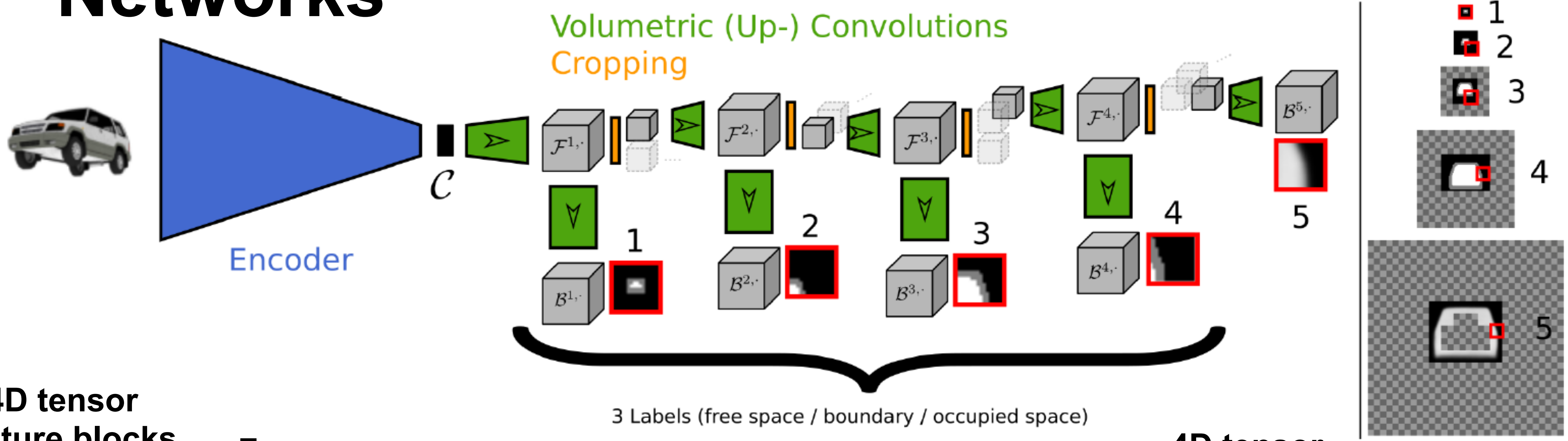
Thresholding

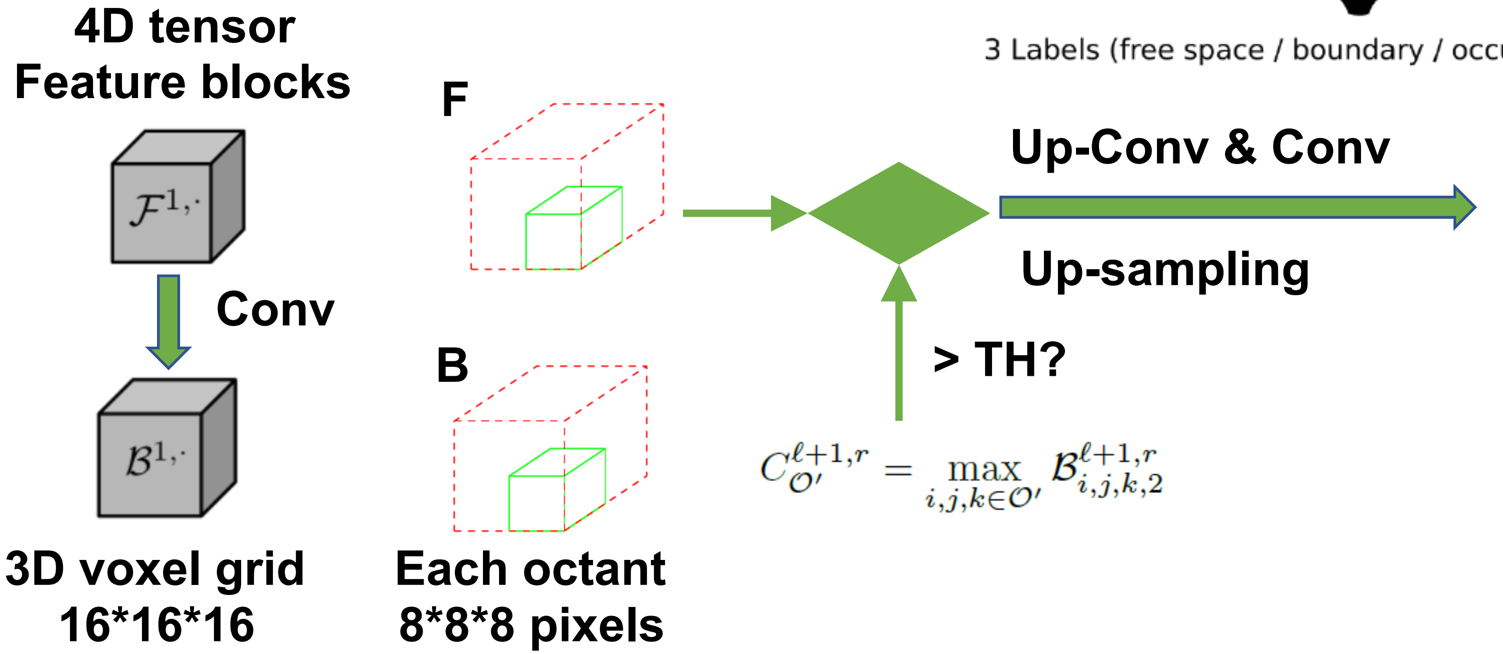
- **Free**
- **Occupied**

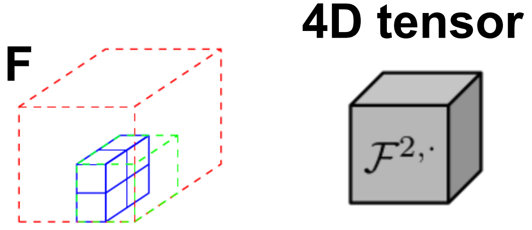
- ➢ **HSP**



Networks







- Each level of the tree has a ground truth;
- Each level has their individual filters.

Type	kW	kH	sW	sH	oC	oW	oH
Input	-	-	-	-	1/3	128	128
Conv	5	5	1	1	16	128	128
MP + IN + R	2	2	2	2	16	64	64
Conv	3	3	1	1	32	64	64
MP + IN + R	2	2	2	2	32	32	32
Conv	3	3	1	1	64	32	32
MP + IN + R	2	2	2	2	64	16	16
Conv	3	3	1	1	128	16	16
MP + IN + R	2	2	2	2	128	8	8
Conv	3	3	1	1	256	8	8
MP + IN + R	2	2	2	2	256	4	4
Conv	3	3	1	1	512	4	4
MP + IN + R	2	2	2	2	512	2	2
Conv + IN	3	3	1	1	1024	2	2
RS + R	-	-	-	-	4096	1	-

TABLE 1: Color/Depth Encoder

Type	kW	kH	kD	sW	sH	sD	oC	oW	oH	oD
FC	-	-	-	-	-	-	4096	1	-	-
RS + IN + R	-	-	-	-	-	-	512	2	2	2
UpConv + IN + R	4	4	4	2	2	2	256	4	4	4
UpConv + IN + R	4	4	4	2	2	2	128	8	8	8
UpConv + R	4	4	4	2	2	2	128	16	16	16
Conv + R	3	3	3	1	1	1	64	16	16	16
UpConv + R	4	4	4	2	2	2	64	32	32	32
Conv	3	3	3	1	1	1	1	32	32	32

TABLE 2: Baseline Decoder

Conv	Convolution
UpConv	Up-convolution
FC	Fully connected
MP	Max Pooling
R	ReLU
RS	Reshape
IN	Instance normalization

kW, kH, kD	Kernel sizes in the three dimensions
sW, sH, sD	Strides in the three dimensions
oC	Number of output feature channels
oW, oH, oD	Output sizes in the three dimensions

Type	kW	kH	kD	sW	sH	sD	oC	oW	oH	oD
FC	-	-	-	-	-	-	13824	-	-	-
RS + IN + R	-	-	-	-	-	-	512	3	3	3
UpConv + IN + R	4	4	4	2	2	2	256	6	6	6
UpConv + IN + R	4	4	4	2	2	2	128	12	12	12
UpConv + R	4	4	4	2	2	2	128	22	22	22
Conv + R	3	3	3	1	1	1	64	20	20	20

TABLE 3: Decoder module, bottleneck to feature block $\mathcal{F}^{1,1}$

Type	kW	kH	kD	sW	sH	sD	oC	oW	oH	oD
UpConv + R	4	4	4	2	2	2	64/32	22	22	22
Conv + R	3	3	3	1	1	1	64/32	20	20	20

TABLE 4: Upsampling module

Type	kW	kH	kD	sW	sH	sD	oC	oW	oH	oD
Conv + R	3	3	3	2	2	2	32/16	18	18	18
Conv	3	3	3	1	1	1	3/6	16	16	16

TABLE 5: Intermediate output module

Type	kW	kH	kD	sW	sH	sD	oC	oW	oH	oD
UpConv + R	4	4	4	2	2	2	16	18	18	18
Conv	3	3	3	1	1	1	1/4	16	16	16

TABLE 6: Full output module

Networks Training

Loss functions:

- Occupancy Loss
  - Cross-Entropy for the occupancy prediction
- Color Loss * 10
  - Mean absolute difference for the color prediction
  - For voxel not on the boundary assign 0 loss

Loss balance for levels:

- Occupancy loss
  - Divided by 8^{l-1}
- Color loss
  - Divided by 4^{l-1}

Network Training

- Subsampling of the child nodes
  - Trees get traversed in a depth first manner
  - The child node is traversed with a certain probability
- Gradient step
  - different sample => different tree
  - Traverse the tree for each example individually
  - Sum up all the gradients and only do a gradient step when a forward and backward traversal of all trees of the whole mini-batch have been done
- Dataset

ShapeNetCar	7497 3D models from the category Car
ShapeNet3	18320 3D models from the categories: Car, Chair, Aeroplane
ShapeNet13	43784 3D models from the categories: Car, Chair, Aeroplane, Table, Couch, Rifle, Lamp, Vessel, Bench, Speaker, Cabinet, Display, Telephone

Experiments

Baselines: traditional CNNs with two different ground truth labels.

LR H:

downsample the HR ground truth to 32^3 (voxel value 0 or 1, contain boundary or not),

then trilinearly upsample to 256^3

LR S:

downsample the HR ground truth to 32^3 (voxel value represent the boundary to space ratio),

then trilinearly upsample to 256^3

Experiments

Computation Efficiency

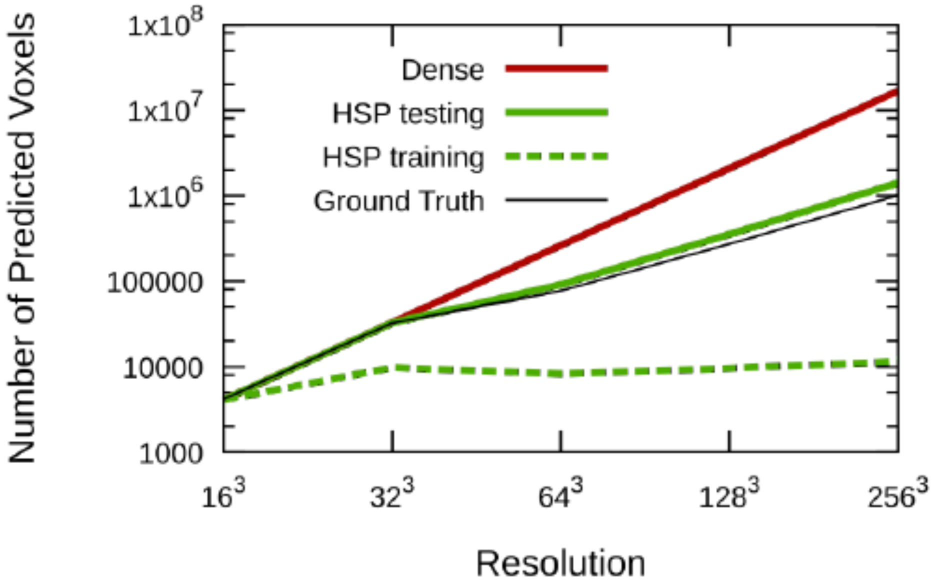


Fig. 6: Number of predicted voxels at different resolutions for a dense baseline and our hierarchical prediction. As additional reference we also plot the number of voxels the ground truth voxel block octrees contain. The numbers were computed on the dataset ShapeNet13 with RGB images as input data.

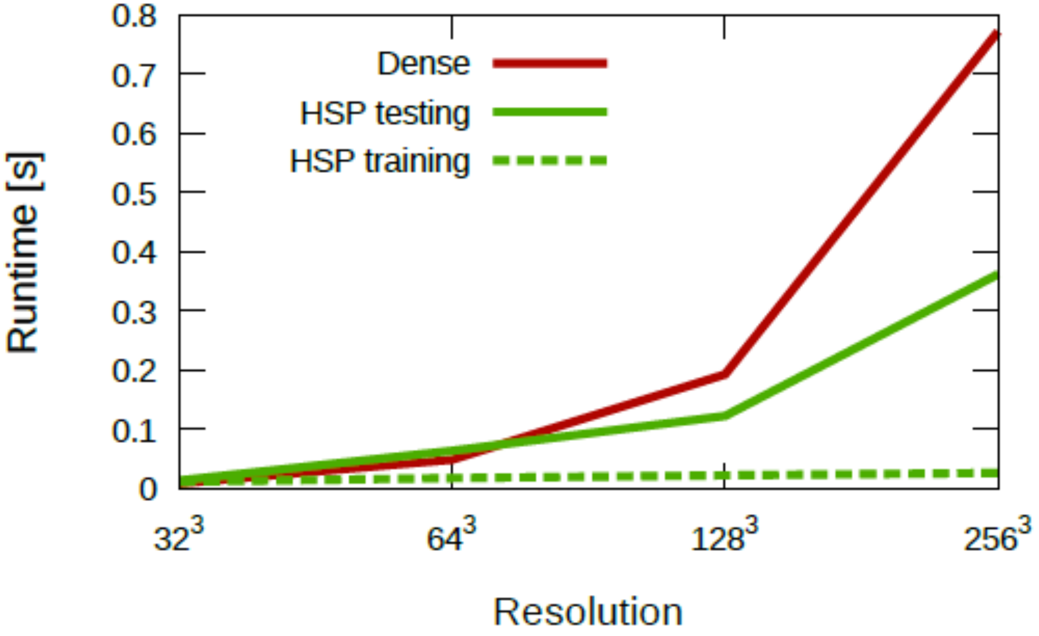


Fig. 7: Runtime of a forward pass for different resultions using an NVIDIA Quadro M6000 GPU. The numbers were computed on the ShapeNet13 dataset with RGB images as input data.

Experiments

Prediction performance

IoU The Intersection over Union is defined as

$$\mathrm{IoU}(pred, gt) = \frac{|\,\mathrm{occ}(pred) \cap \mathrm{occ}(gt)|}{|\,\mathrm{occ}(pred) \cup \mathrm{occ}(gt)|}, \quad (2)$$

where $\mathrm{occ}(\cdot)$ returns the set of occupied voxels and $|\cdot|$ is the set cardinality.

CD We first define the asymmetric Chamfer Distance between volumes vol_1 and vol_2

$$\mathrm{CD}_{\mathrm{as}}(v_1, v_2) = \frac{1}{|\,\partial(v_1)|} \sum_{p \in \partial(v_1)} \min_{q \in \partial(v_2)} \|p - q\|_2. \quad (3)$$

$$\mathrm{CD}(pred, gt) = \frac{\mathrm{CD}_{\mathrm{as}}(pred, gt) + \mathrm{CD}_{\mathrm{as}}(gt, pred)}{2}. \quad (4)$$

Metric	Method	Car	Chair	Aero	Mean
IoU	LR H	0.642	0.372	0.443	0.486
	LR S	0.678	0.385	0.505	0.523
	HSP	**0.709**	**0.414**	**0.557**	**0.560**
	HSP Color	0.691	0.379	0.519	0.530
CD	LR H	0.0161	0.0229	0.0171	0.0187
	LR S	0.0189	0.0269	0.0202	0.0220
	HSP	**0.0116**	**0.0201**	**0.0131**	**0.0149**
	HSP Color	0.0121	0.0241	0.0165	0.0176

TABLE 7: Results for RGB input on the ShapeNet3 dataset.

Experiments

Metric	Method	Car	Chair	Aero	Table	Couch	Rifle	Lamp	Vessel	Bench	Speaker	Cabinet	Display	Phone	Mean
IoU	LR H	0.624	0.389	0.411	0.349	0.556	0.383	0.232	0.437	0.277	0.511	0.547	0.377	0.604	0.438
	LR S	0.675	0.374	0.487	0.351	0.589	0.354	0.241	0.436	0.166	0.530	0.583	0.383	0.585	0.443
	HSP	**0.696**	**0.408**	**0.531**	**0.412**	**0.600**	**0.423**	**0.280**	**0.457**	**0.312**	**0.542**	**0.605**	**0.406**	**0.616**	**0.484**
CD	LR H	0.0205	0.0223	0.0199	0.0226	0.0267	0.0208	0.0417	0.0264	0.0222	0.0294	0.0220	0.0273	**0.0183**	0.0246
	LR S	0.0198	0.0288	0.0228	0.0267	0.0288	0.0213	0.0495	0.0296	0.0263	0.0340	0.0249	0.0326	0.0276	0.0287
	HSP	**0.0121**	**0.0223**	**0.0150**	**0.0195**	**0.0235**	**0.0155**	**0.0337**	**0.0227**	**0.0197**	**0.0271**	**0.0176**	**0.0270**	0.0185	**0.0211**

TABLE 8: Results for RGB input on the ShapeNet13 dataset.

Metric	Method	Car	Chair	Aero	Table	Couch	Rifle	Lamp	Vessel	Bench	Speaker	Cabinet	Display	Phone	Mean
IoU	LR H	0.589	0.370	0.386	0.320	0.542	0.355	0.226	0.416	0.223	0.495	0.537	0.365	0.556	0.414
	LR S	0.636	0.358	0.430	0.321	0.550	0.334	0.234	0.417	0.155	0.516	0.554	0.378	0.541	0.417
	HSP	**0.717**	**0.455**	**0.555**	**0.454**	**0.661**	**0.441**	**0.318**	**0.511**	**0.340**	**0.581**	**0.637**	**0.463**	**0.708**	**0.526**
CD	LR H	0.0273	0.0272	0.0249	0.0271	0.0298	0.0236	0.0436	0.0291	0.0297	0.0329	0.0276	0.0324	0.0256	0.0293
	LR S	0.0215	0.0329	0.0290	0.0294	0.0299	0.0298	0.0632	0.0323	0.0494	0.0349	0.0274	0.0329	0.0282	0.0339
	HSP	**0.0111**	**0.0192**	**0.0129**	**0.0161**	**0.0179**	**0.0149**	**0.0395**	**0.0192**	**0.0172**	**0.0235**	**0.0141**	**0.0214**	**0.0119**	**0.0184**

TABLE 9: Results for depth input on the ShapeNet13 dataset.

Experiments

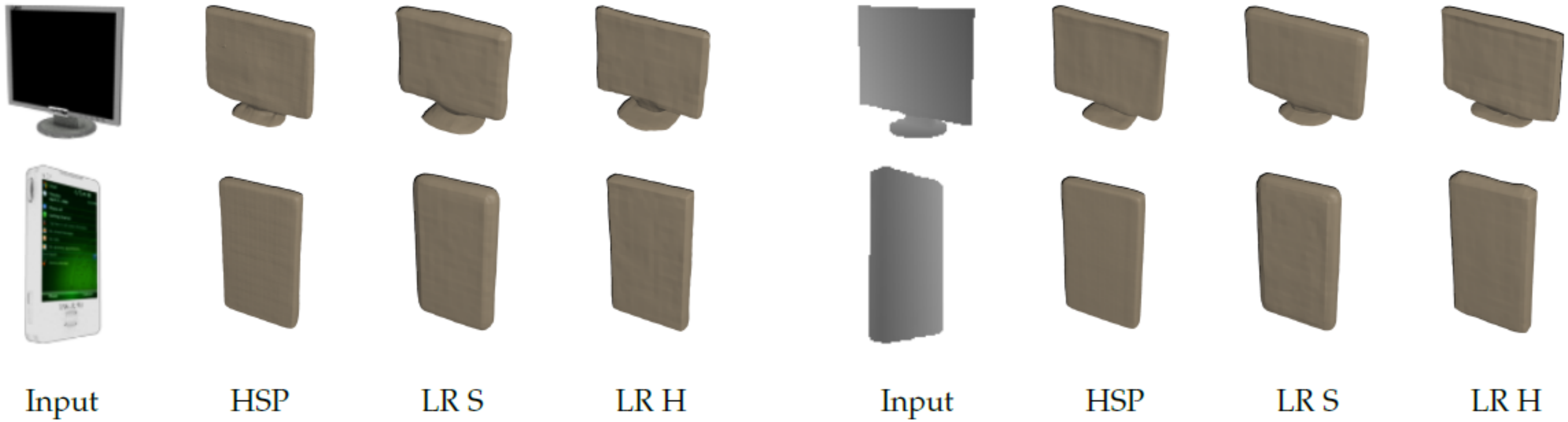


Fig. 11: Selected examples on the task of occupancy prediction form RGB and Depth input on the ShapeNet13 dataset, continued.

Thanks for your attention