ICDSC 2019

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

LTTM

Region Merging Driven by Deep Learning for RGB-D Segmentation and Labeling

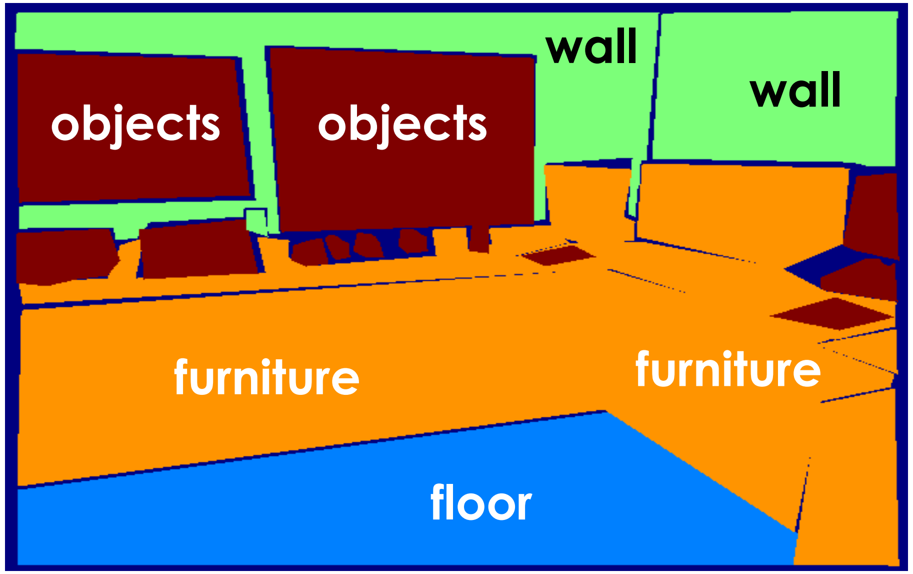
U. Michieli, M. Camporese, A. Agiollo, G. Pagnutti, P. Zanuttigh

September 9th, 2019

Outline

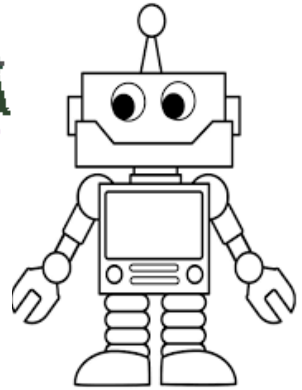
- Semantic Segmentation
- Proposed Framework
  - Pre-processing
  - Over-segmentation and Classification
  - Merging Phase
- Results
- Conclusions and Future Work

Semantic Segmentation

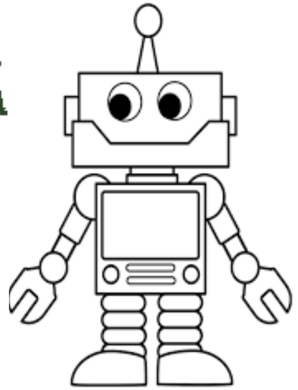


- Segmentation + labeling (pixel-wise classification)
- Deep learning and consumer depth sensors
- Very useful for free navigation systems to explore the surroundings

Semantic Segmentation



Semantic Segmentation



Proposed Framework

Proposed Framework

AIM: propose CNN for region merging and refine boundaries of shapes

Use normalized cuts spectral clustering extended for RGBD
→ but bias toward region of similar sizes

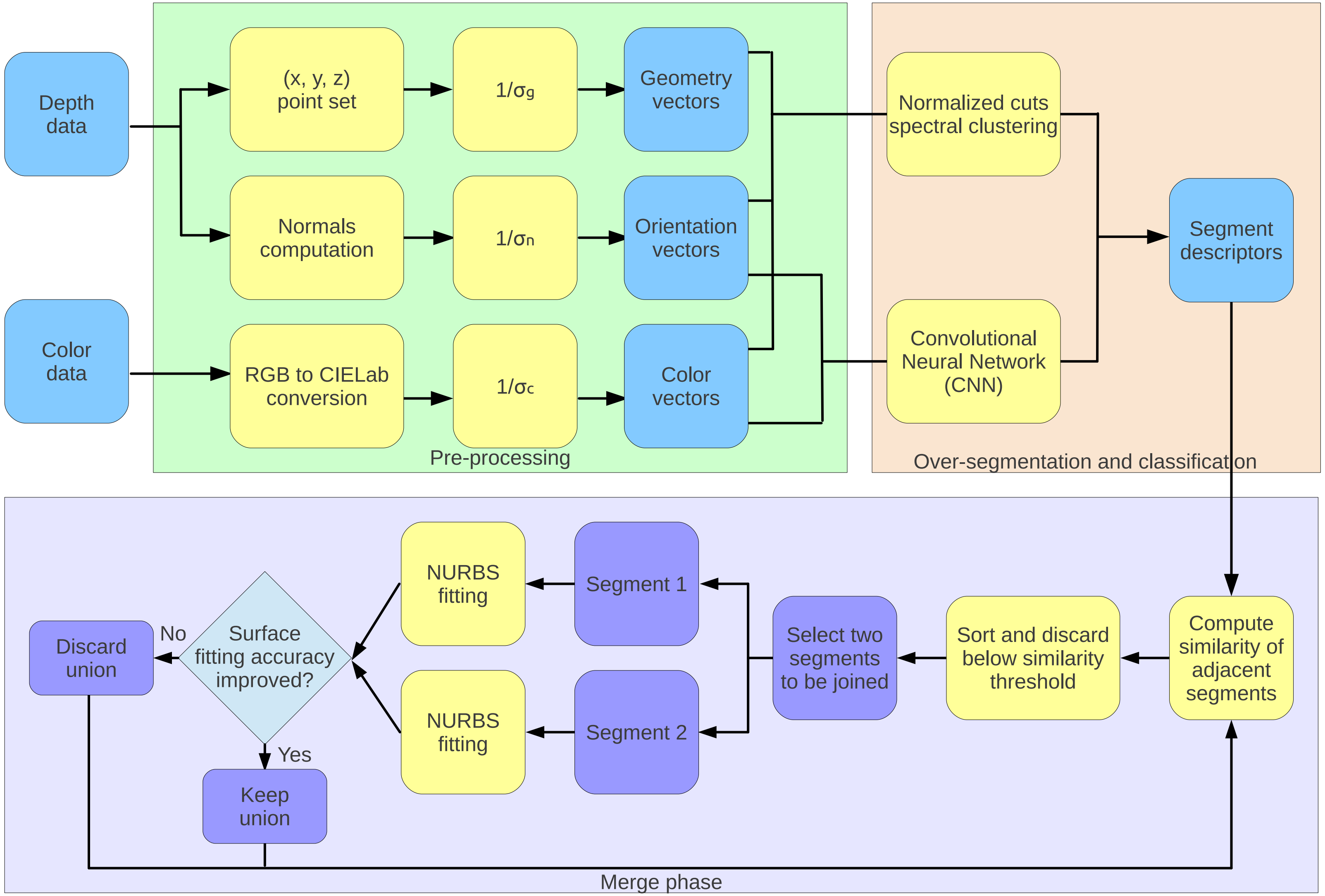
Then 2 steps procedure:

- Initial over-segmentation to properly separate objects
- Region merging procedure to avoid over-segmentation

Framework derived from [1] but much faster and simpler

[1] G.Pagnutti, L. Minto, P. Zanuttigh, "Segmentation and Semantic Labeling of RGBD Data with Convolutional Neural Networks and Surface Fitting ", IET Computer Vision, 2017

Framework of [1]

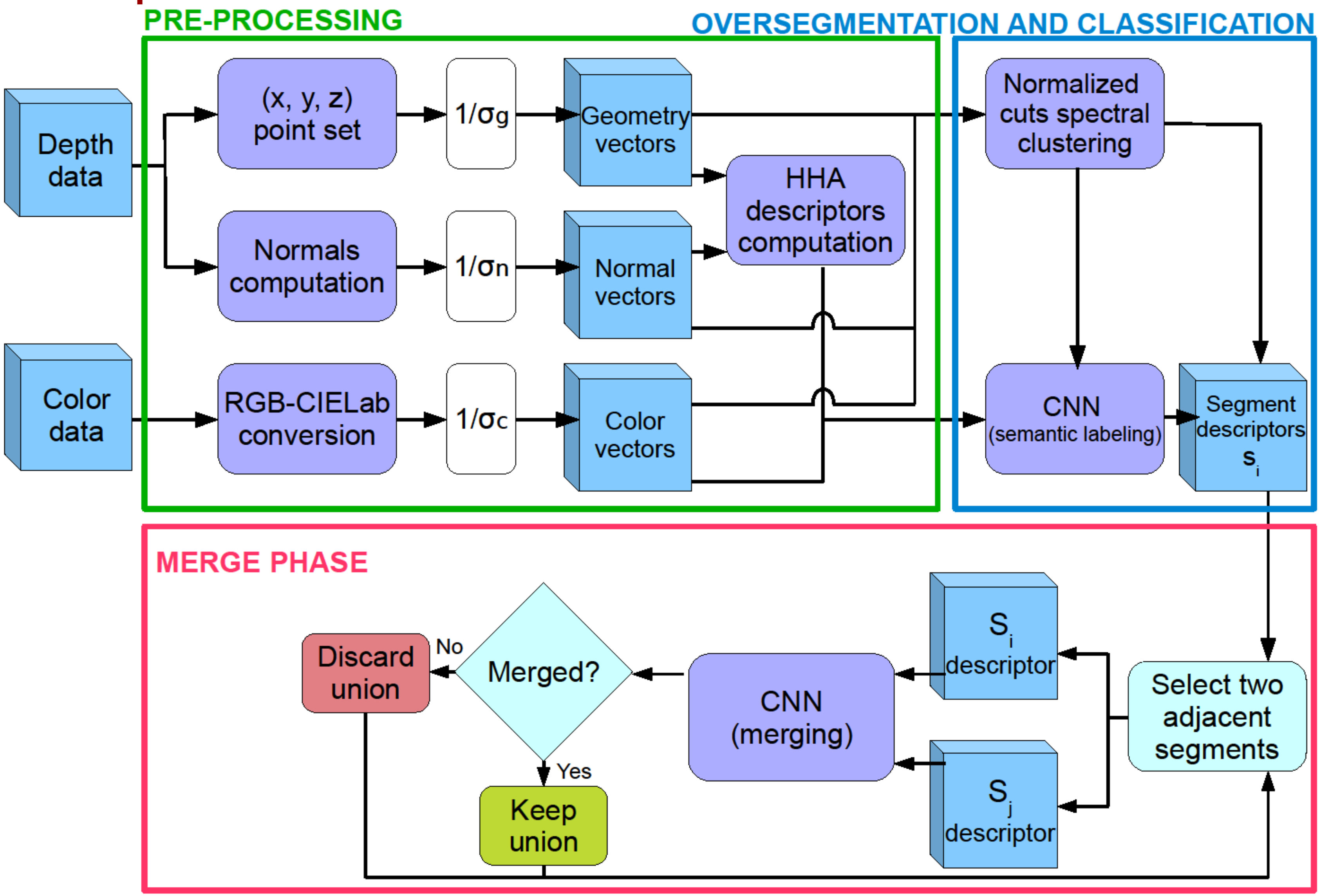


CONs:

- NURBS fitting very slow
- Many hand-tuned thresholds (on depth, color, normals, NURBS fitting)

[1] G.Pagnutti, L. Minto, P. Zanuttigh, "Segmentation and Semantic Labeling of RGBD Data with Convolutional Neural Networks and Surface Fitting ", IET Computer Vision, 2017

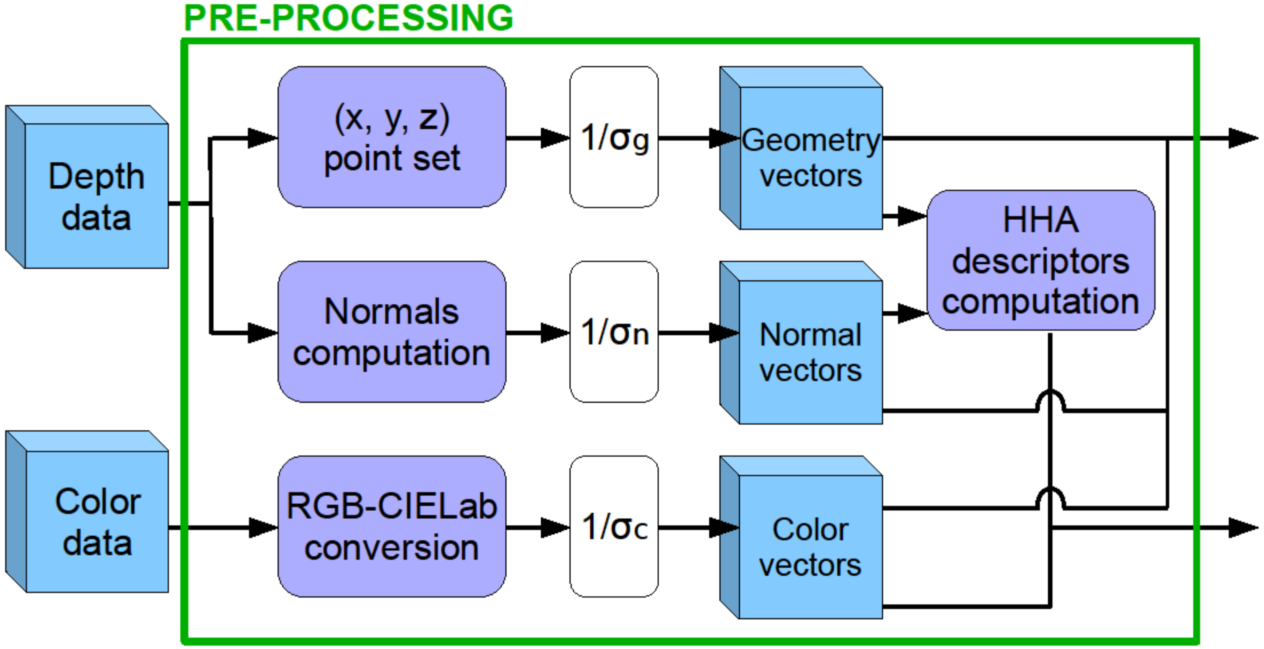
Proposed Framework



PROs:

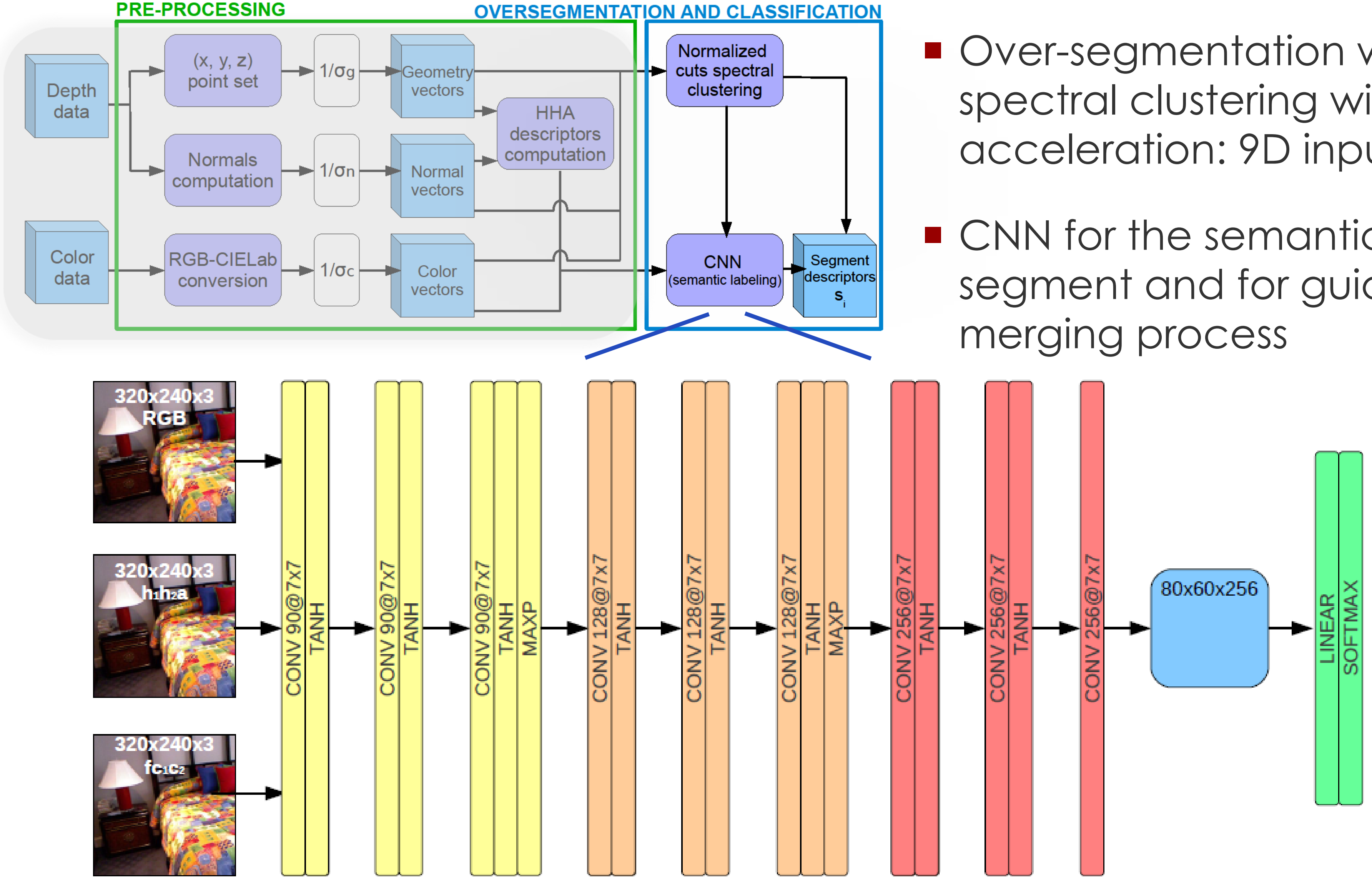
- Much faster
- Fewer thresholds
- Same accuracy

Proposed Framework - Preprocessing



- 3 channels for 3D location
- 3 channels for surface normals
- 3 channels for color representation
  → CIELab for perceptual uniformity
- Normalization to achieve consistent representation across the 3 domains.

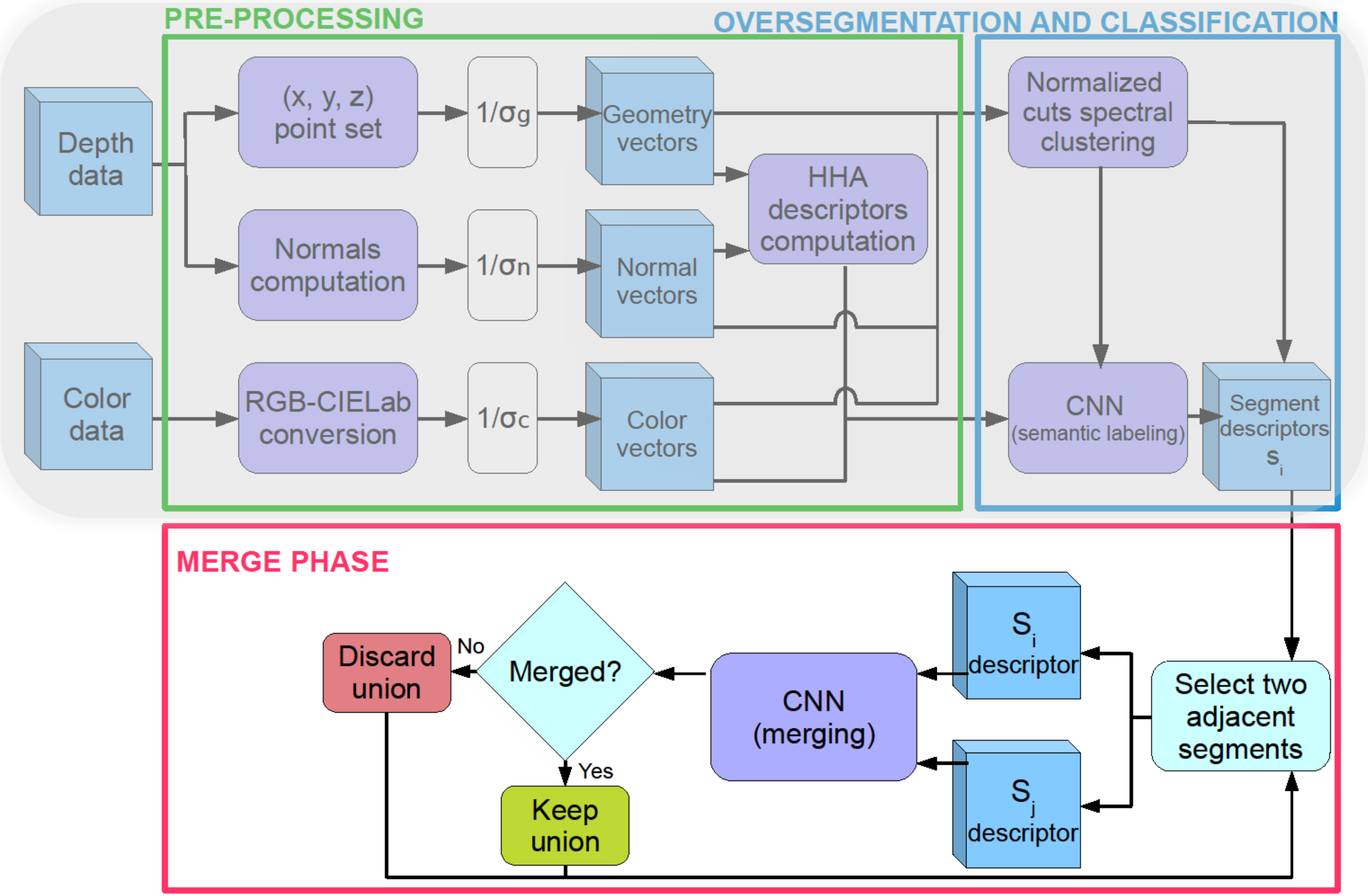
Proposed Framework – Oversegmentation



- Over-segmentation with normalized cuts spectral clustering with Nystrom acceleration: 9D input
- CNN for the semantic labeling of each segment and for guiding the region merging process

- 9 conv layers
- 15 classes
- very simple

Proposed Framework – Region Merging



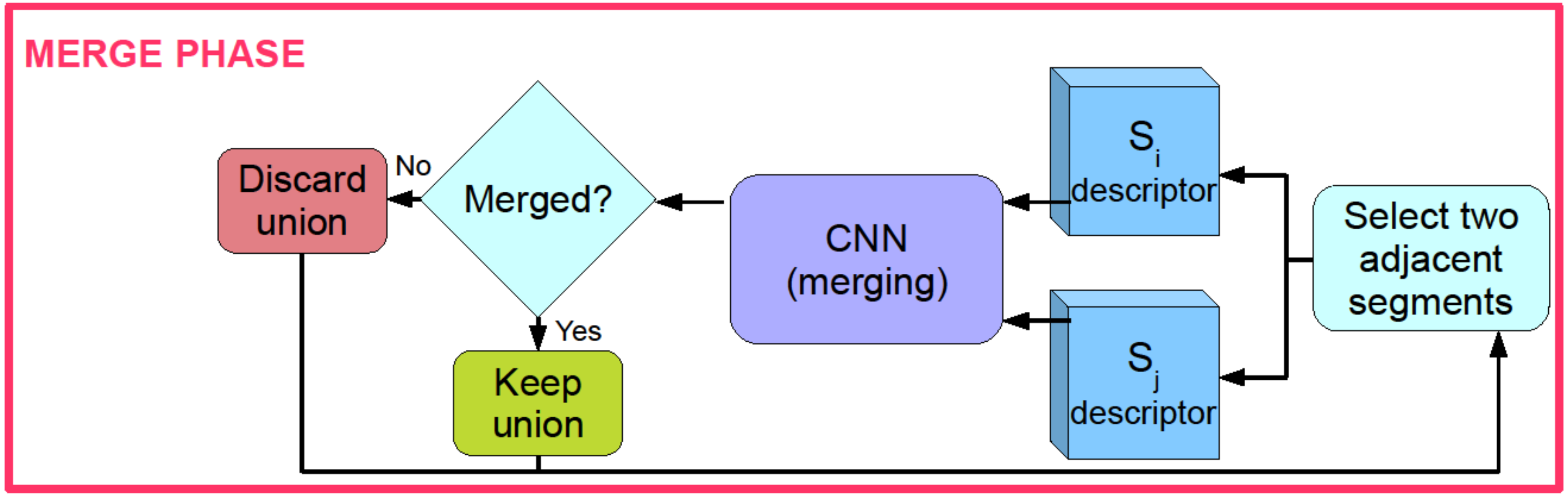
- Compute adjacency map of the segments
- Compute similarity between adjacent segment descriptors with Bhattacharyya coefficient:

$$b_{i,j} = \sum_t \sqrt{s_i^t s_j^t}$$

t: class scores
s_i: descriptors (~PDFs)

- Sort list on the basis of $b_{i,j}$

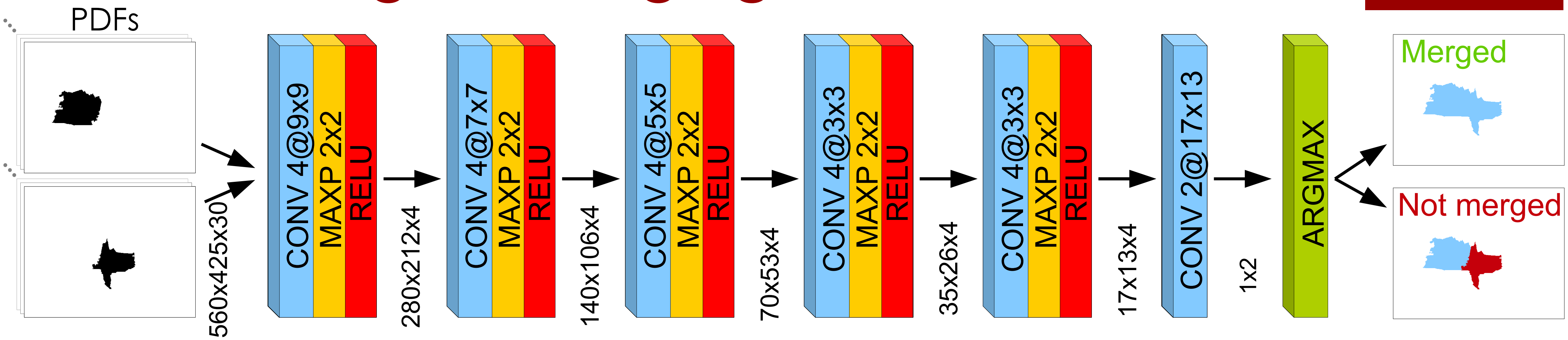
Proposed Framework



Iterative merging procedure

- Select segments with $b_{i,j} > T_{sim}$
- CNN classifier to decide whether the two segments will be joined or not
  - If merged: new segment of the union is created and list updated
  - If not merged: remove segments from the list

CNN for Region Merging - PDFs



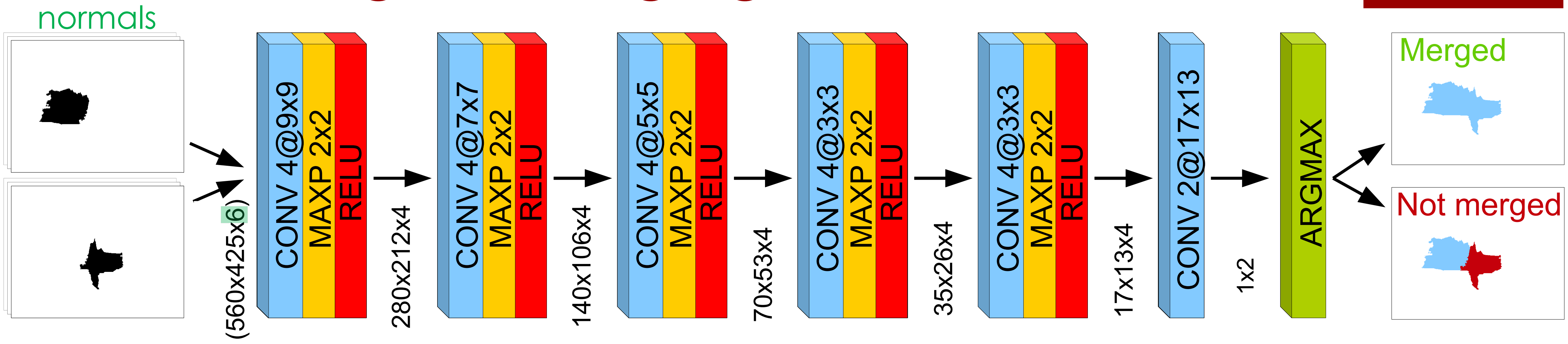
CNN for classification (6 conv. layers, symm. padding, 2x2 maxpool, ReLU)

input: 2 outputs of softmax layer of semantic CNN (15 channels each candidate)

training: 50 epochs, batch size of 32 samples, CE & L2 regularization losses, Adam with $lr = 10^{-4}$, regularization constant $= 10^{-3}$, $T_{sim} = 0.8$

training time: about 11 hours on a NVIDIA Titan X GPU

CNN for Region Merging - Normals



CNN for classification (6 conv. layers, symm. padding, 2x2 maxpool, ReLU)

input: 2 surface normals of the 2 candidate segments (3 channels each)

training: 50 epochs, batch size of 32 samples, CE & L2 regularization losses, Adam with $lr = 10^{-3}$, regularization constant $= 5 \cdot 10^{-5}$, $T_{sim} = 0.75$

training time: about 3 hours on a NVIDIA Titan X GPU

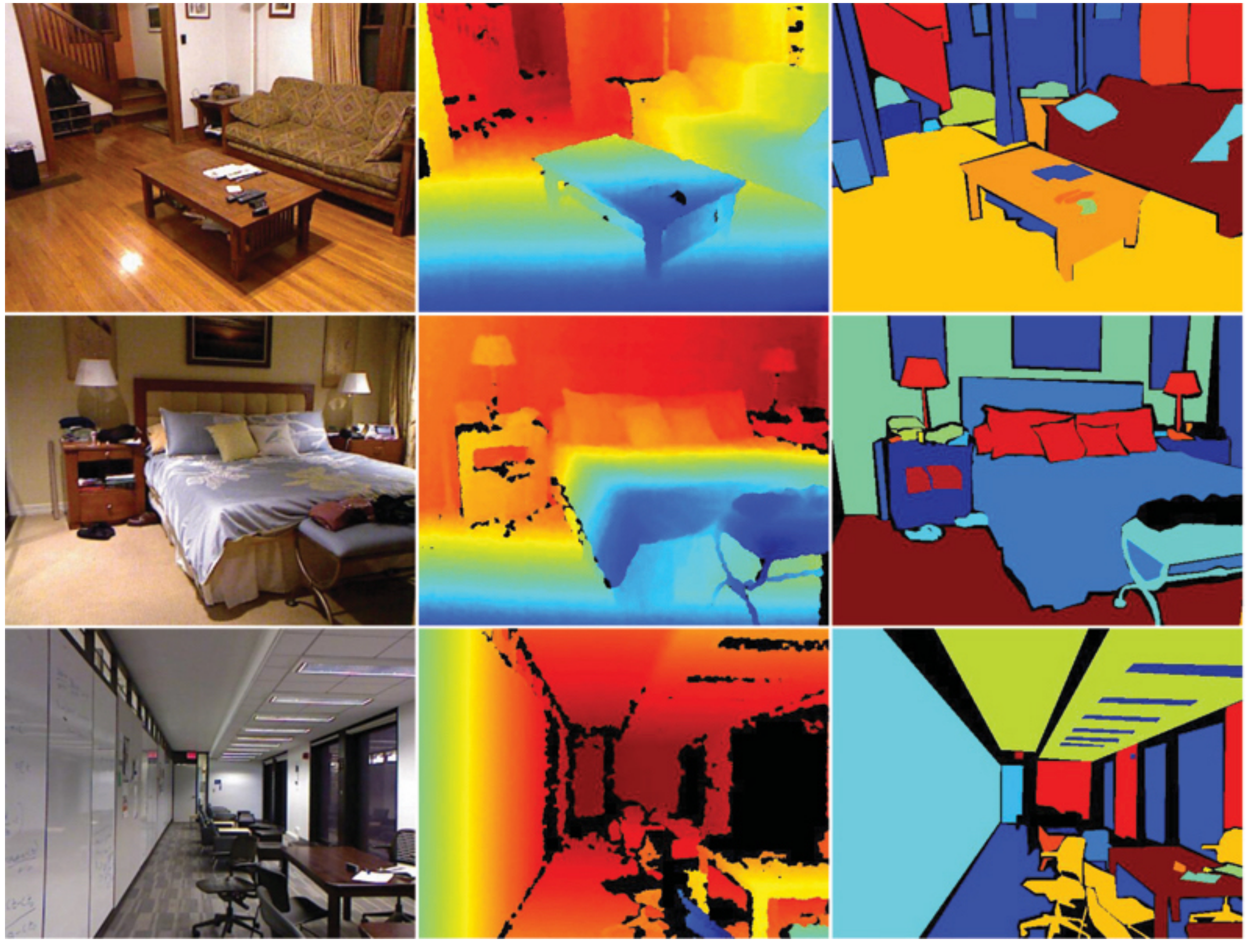
→ PDFs richer descriptions, while normals are faster with limited impact on the final accuracy

Experimental Results

NYUDv2 Dataset [2]

1449 depth maps + color images of indoor scenes with Kinect sensor

RGB raw depth GT



training set: 795 scenes
test set: 654 scenes

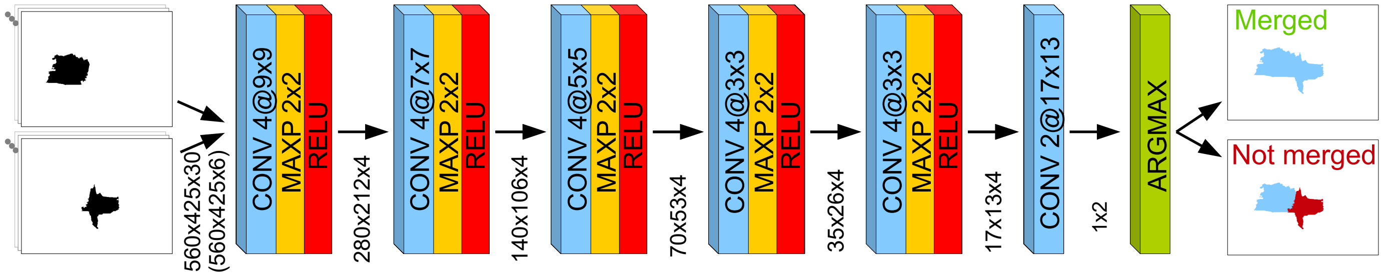
894 classes clustered in 15 classes as [3]

unknown & *unlabeled* classes excluded

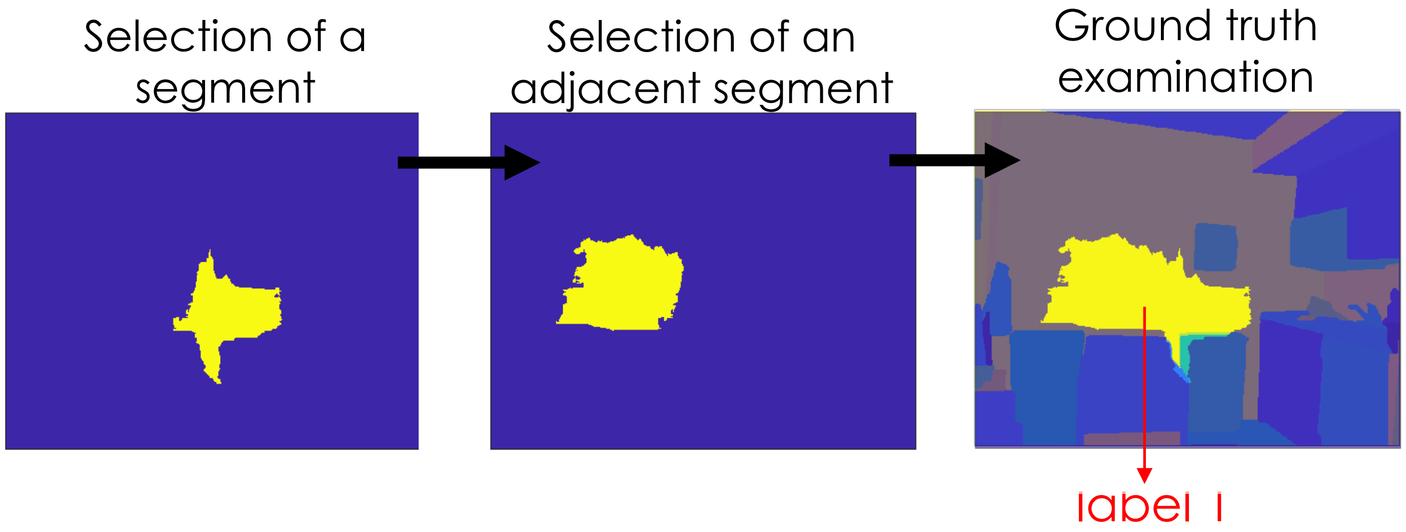
[2] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. 2012. Indoor segmentation and support inference from RGBD images. ECCV. Springer.
[3] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. 2013. Indoor semantic segmentation using depth information. ICLR.

Merging CNN – Ground Truth Generation

Need a dataset to train the merging CNN



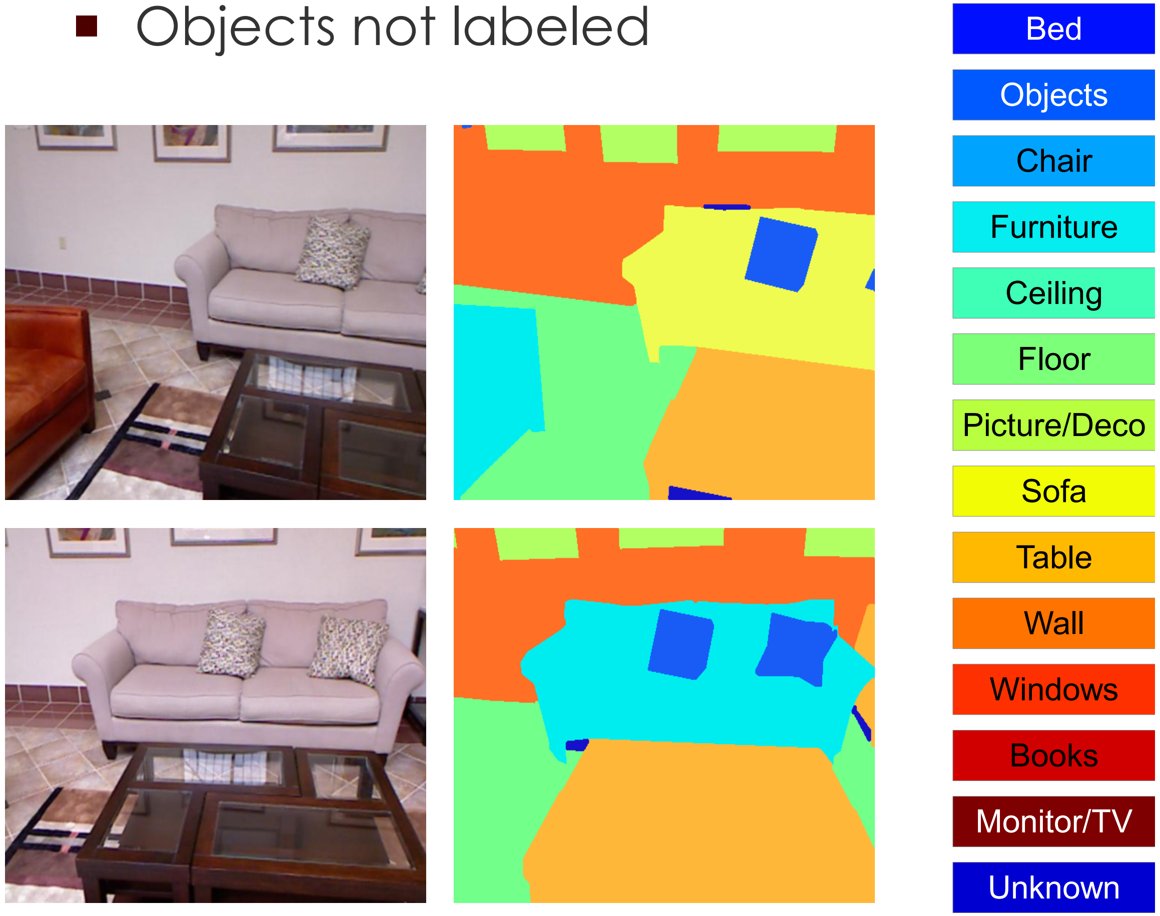
- Randomly select 10 couples of adjacent segments in each image
  - Assign label 1 if more than 85% of the union of the segments belongs to same object in the semantic segmentation ground truth
  - Assign label 0 otherwise

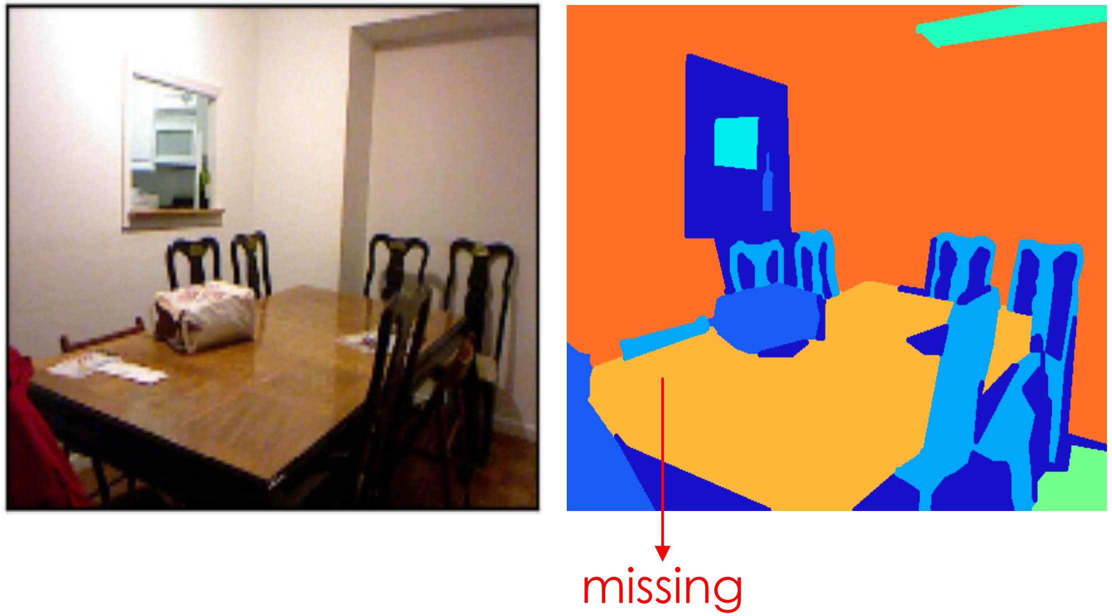


Region appears to be uniform

Merging CNN – GT Ambiguities

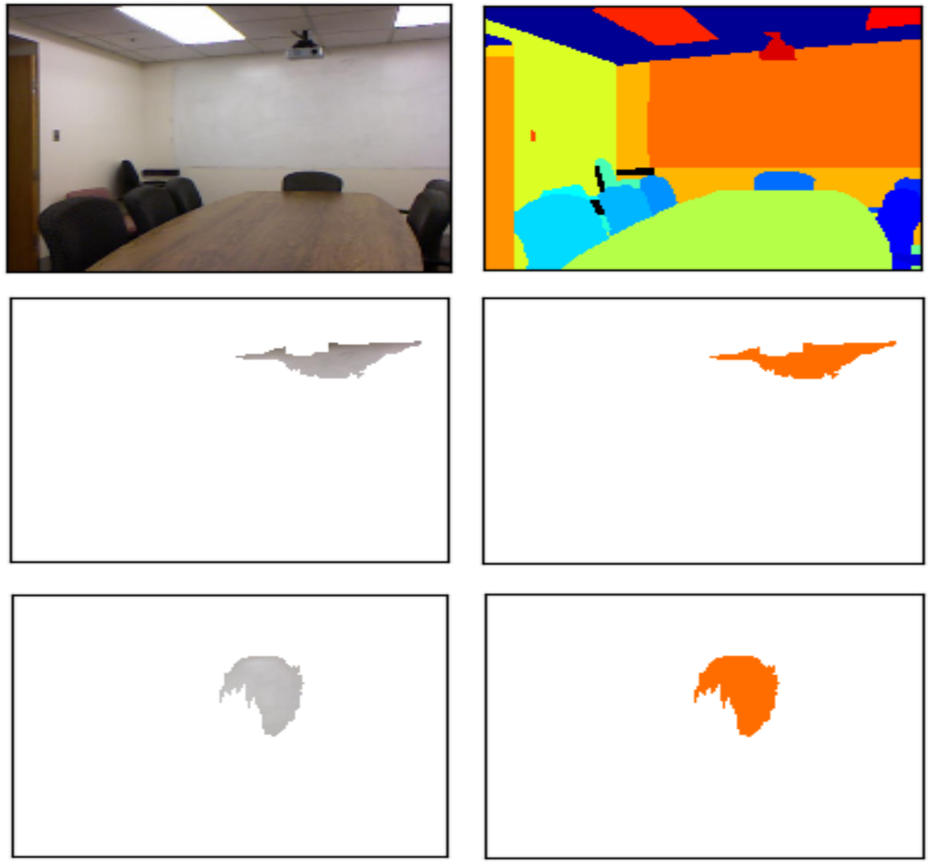
- Examples of ambiguities in ground truth:
  - Inconsistent labeling
  - Objects not labeled



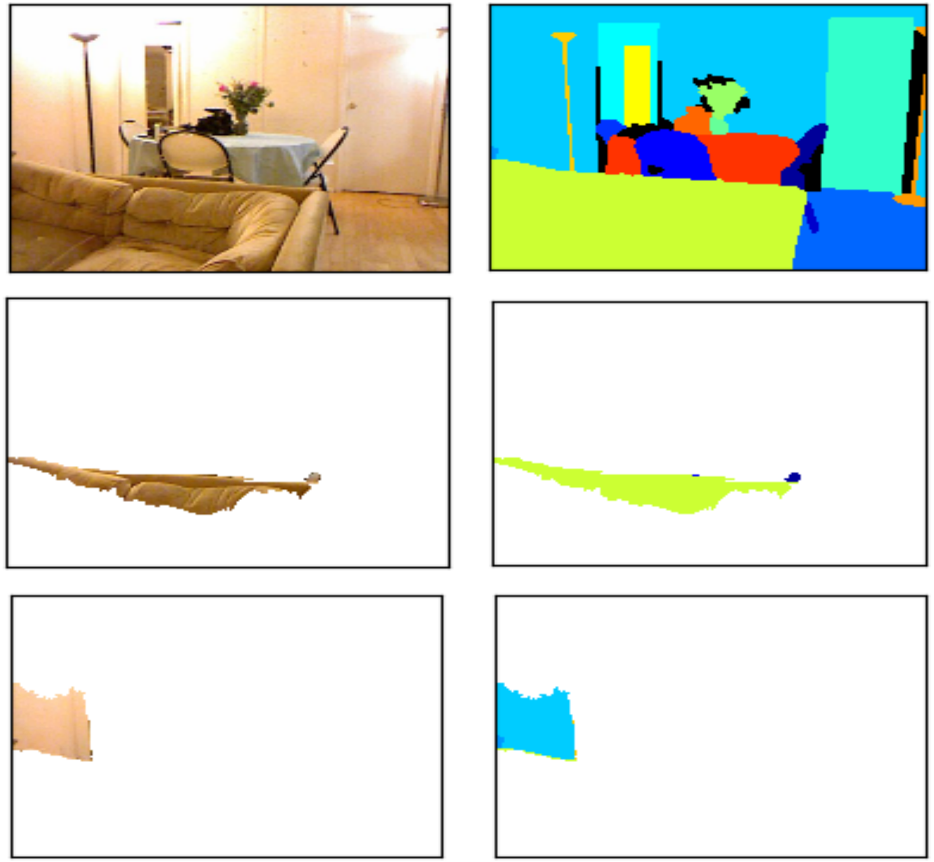


Merging CNN – Results

Predicted: Merge
GT: Merge



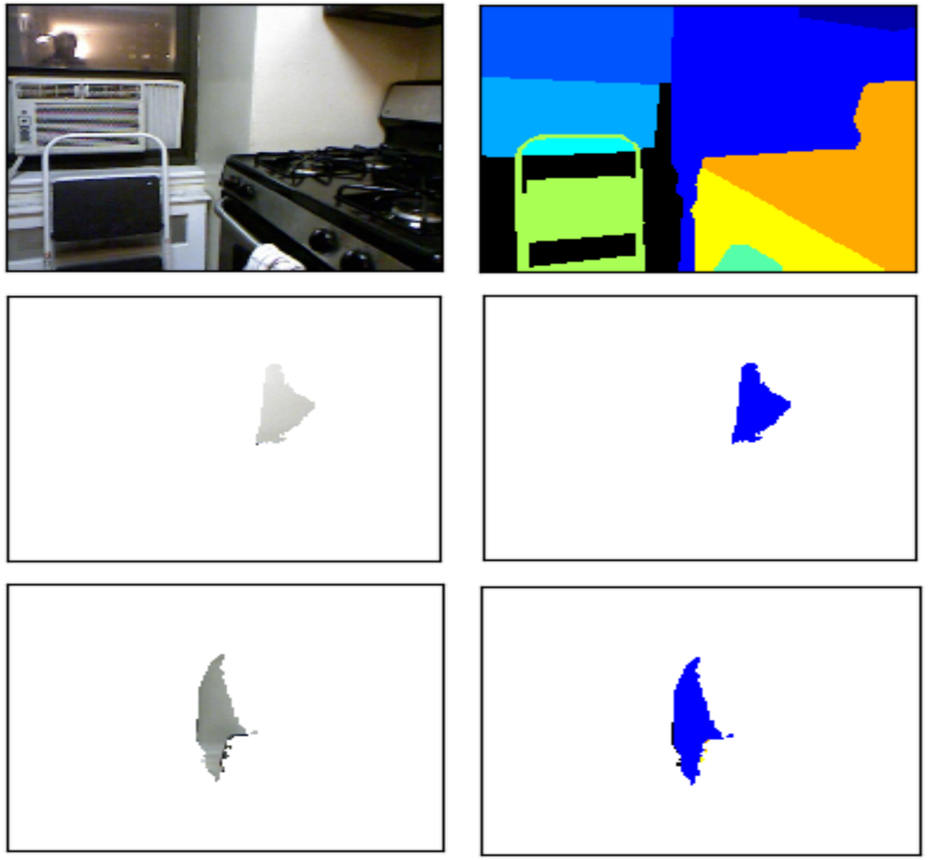
Predicted: Not Merged
GT: Not Merged



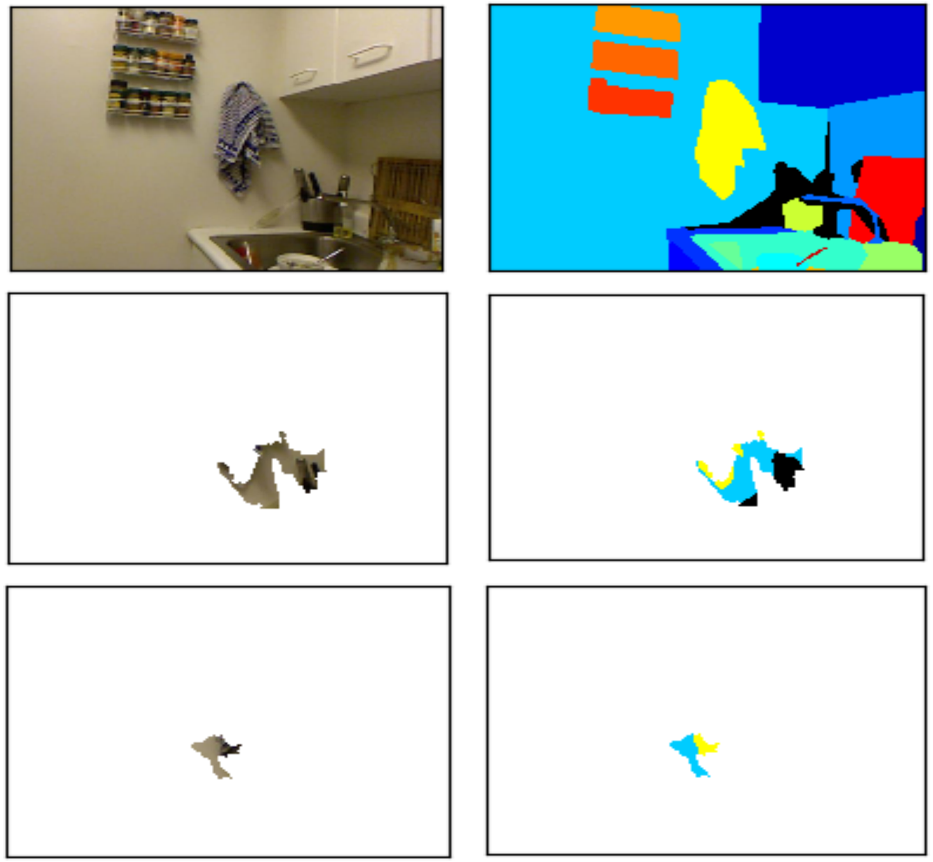
- Good oversegmentation (inter-uniformity)

Merging CNN – Results

Predicted: Not Merged
GT: Merge



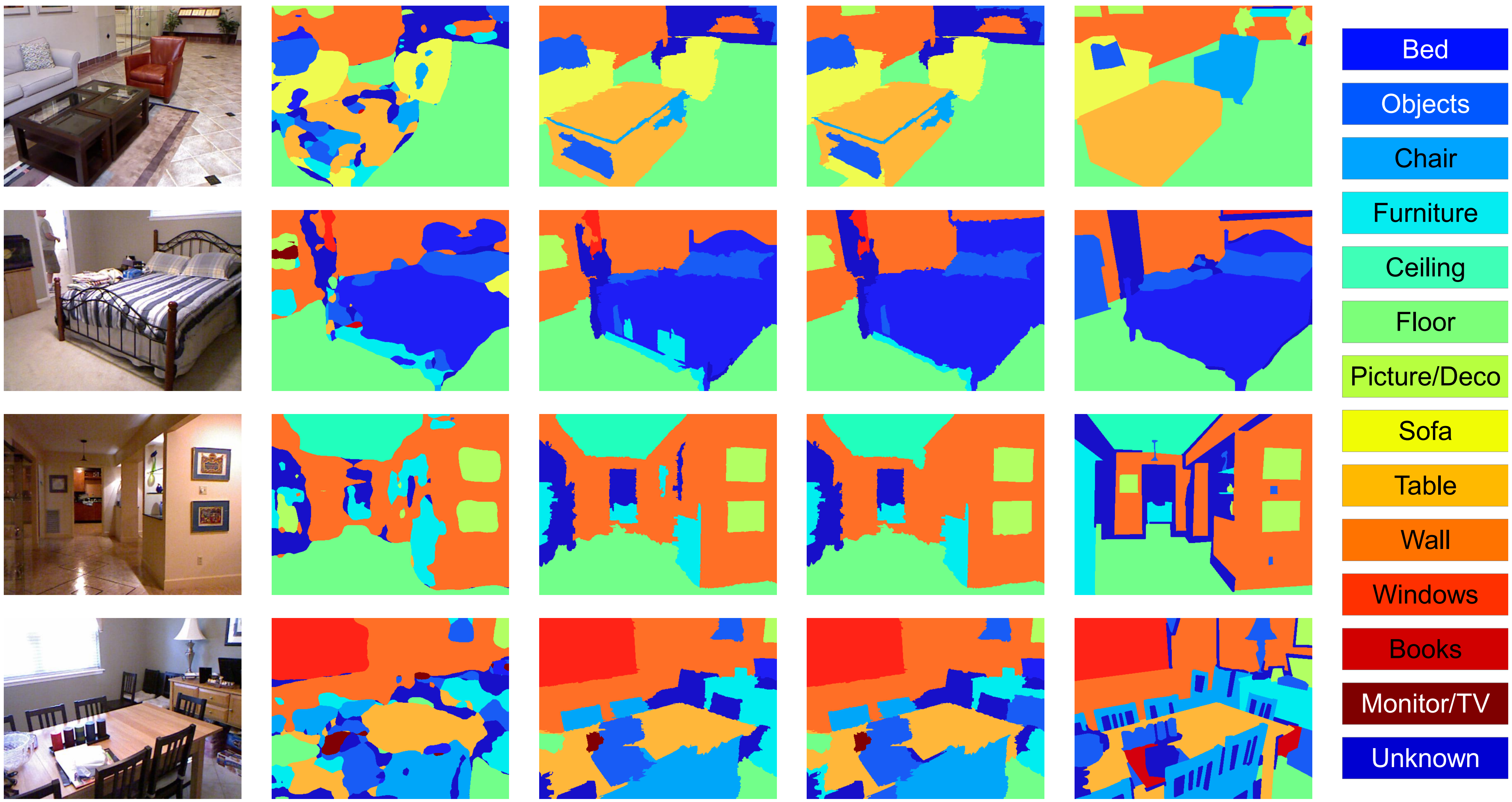
Predicted: Merge
GT: Not Merged



- Bad oversegmentation

Qualitative Results

Color view | Semantic CNN | Pagnutti et al. [1] | Our Approach | Ground Truth



[1] G.Pagnutti, L. Minto, P. Zanuttigh, "Segmentation and Semantic Labeling of RGBD Data with Convolutional Neural Networks and Surface Fitting ", IET Computer Vision, 2017

Quantitative Results

Approach	*Pixel Accuracy*	*Class Accuracy*
Couprie et al. [4]	52.4%	36.2%
Hickson et al. [5]	53.0%	47.6%
A. Wang et al. [6]	46.3%	42.2%
J. Wang et al. [7]	54.8%	52.7%
A. Hermans et al. [8]	54.2%	48.0%
D. Eigen et al. [9]	75.4%	66.9%
Pagnutti et al. [1]	67.2%	54.4%
Semantic CNN	64.4%	51.7%
Our method (normals)	66.6%	53.6%
Our method (PDFs)	67.2%	54.5%

[1] G.Pagnutti, L. Minto, P. Zanuttigh, "Segmentation and Semantic Labeling of RGBD Data with Convolutional Neural Networks and Surface Fitting ", IET Computer Vision, 2017
[4] C. Couprie, C. Farabet, L. Najman, and Y. Lecun. 2014. Convolutional nets and watershed cuts for real-time semantic Labeling of RGBD videos. JMLR 15, 1 (2014), 3489–3511.
[5] S. Hickson, I. Essa, and H. Christensen. 2015. Semantic Instance Labeling Leveraging Hierarchical Segmentation. WCACV. 1068–1075
[6] A. Wang, J. Lu, G. Wang, J. Cai, and T. Cham. 2014. Multi-modal unsupervised feature learning for RGB-D scene labeling. ECCV. 453–467.
[7] J. Wang, Z. Wang, D. Tao, S. See, and G. Wang. 2016. Learning Common and Specific Features for RGB-D Semantic Segmentation with Deconvolutional Networks. ECCV. 664–679.
[8] A. Hermans, G. Floros, and B. Leibe. 2014. Dense 3D semantic mapping of indoor scenes from rgb-d images. ICRA. 2631–2638.
[9] D. Eigen and R. Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. ICCV. 2650–2658.

Quantitative Results

Approach	*Pixel Accuracy*	*Class Accuracy*	*Inference Time**
Pagnutti et al. [1]	**67.2%**	54.4%	58 ms
Our method (normals)	66.6%	53.6%	**2 ms**
Our method (PDFs)	**67.2%**	**54.5%**	10 ms

* on a Intel Core i7-8700K CPU @3.70GHz with NVIDIA GeForce GTX 1070 GPU

- Same over-segmentation
- Similar results
- Much faster
  - no surface fitting
  - In [1] time heavily depends on the area to be fit, here it is constant!
- Fewer hand-tuned thresholds (1 vs. 4)

[1] G.Pagnutti, L. Minto, P. Zanuttigh, "Segmentation and Semantic Labeling of RGBD Data with Convolutional Neural Networks and Surface Fitting ", IET Computer Vision, 2017

Conclusions → Future Work

Conclusions → Future Work

- Agnostic to the over-segmentation method

Conclusions → Future Work

- Agnostic to the over-segmentation method
  - use other methods like superpixels

Conclusions → Future Work

- Agnostic to the over-segmentation method
  - use other methods like superpixels
- Semantic CNN very simple

Conclusions → Future Work

- Agnostic to the over-segmentation method
  - use other methods like superpixels
- Semantic CNN very simple
  - use more complex one (less speed)

Conclusions → Future Work

- Agnostic to the over-segmentation method
  - use other methods like superpixels
- Semantic CNN very simple
  - use more complex one (less speed)
- CNN useful for region merging

Conclusions → Future Work

- Agnostic to the over-segmentation method
  - use other methods like superpixels
- Semantic CNN very simple
  - use more complex one (less speed)
- CNN useful for region merging
  - focus the attention on the edges of the candidates

Conclusions → Future Work

- Agnostic to the over-segmentation method
  - use other methods like superpixels
- Semantic CNN very simple
  - use more complex one (less speed)
- CNN useful for region merging
  - focus the attention on the edges of the candidates
- Smaller computational time

Conclusions → Future Work

- Agnostic to the over-segmentation method
  - use other methods like superpixels
- Semantic CNN very simple
  - use more complex one (less speed)
- CNN useful for region merging
  - focus the attention on the edges of the candidates
- Smaller computational time
  - useful for free-navigation and for other fields

Thank you!

Questions?