INDIANA UNIVERSITY
School of
Informatics, Computing, and Engineering

Automatic Annotation for Semantic Segmentation in Indoor Scenes

Md Alimoor Reza[1], Akshay U. Naik[1], Kai Chen[2], David Crandall[1]

[1]School of Informatics, Computing, and Engineering, Indiana University, [2]School of Computer Science, Fudan University

1. Introduction

- ❑ **Motivation:** State of art semantic segmentation models are mostly based on Deep Neural Networks which require large quantities of training labels to learn millions of parameters. However, human semantic annotation is very **expensive and time-consuming,** especially for videos.
- ❑ **Goal:** Generate densely annotation for images of an indoor video **without any human annotation** which can be used to better train a data-hungry Deep Neural Network for semantic segmentation (e.g. FCN[1]) on SUN3D dataset

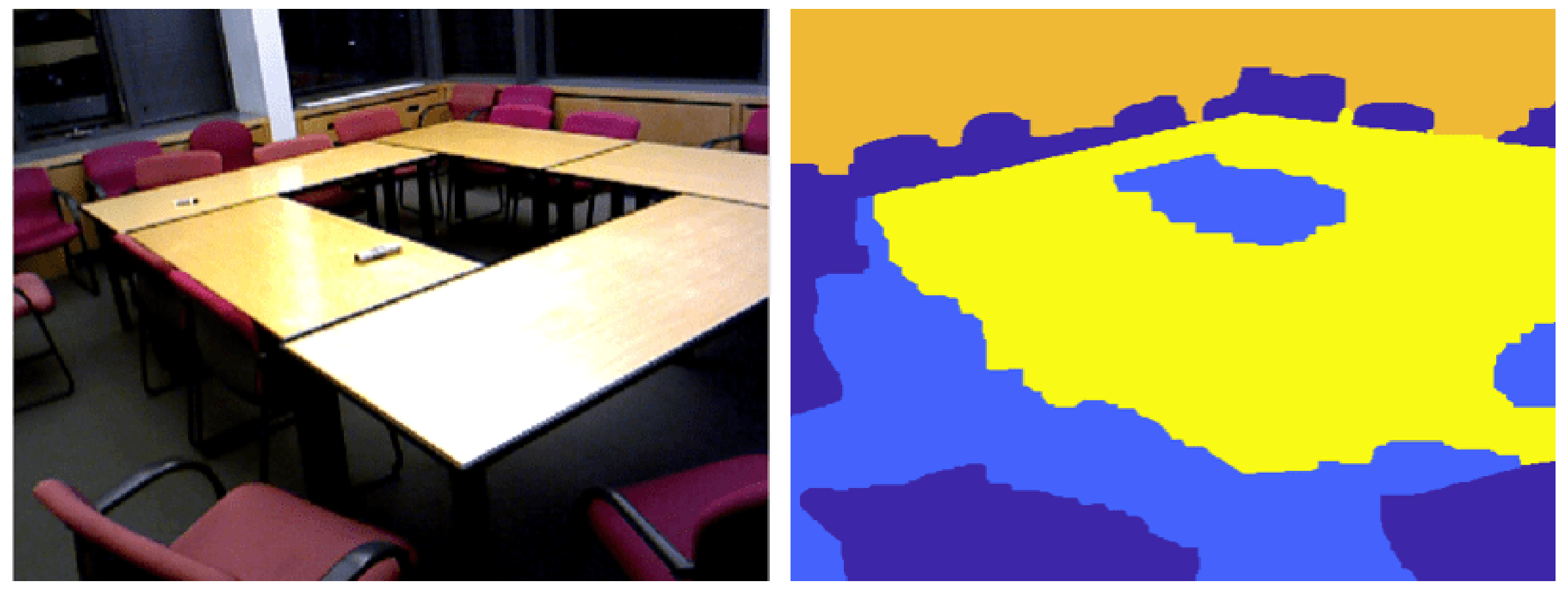

Fig. 1. Images (left) are automatically annotated (right) based on off-the-shelf object detectors and a 3D room layout estimator.

2. Solution

- ❑ A typical approach for scene understanding is to separate an indoor scene into two parts: **foreground** (e.g. chairs) and **background** (e.g. walls). We can capture them separately and then **combine** them to generate a final semantic annotation.
- ❑ We use a state of the art **object detector** (Mask RCNN [2]) to obtain instance masks for foreground objects and **3D scene layout segmentation** model ([3]) to get background information.

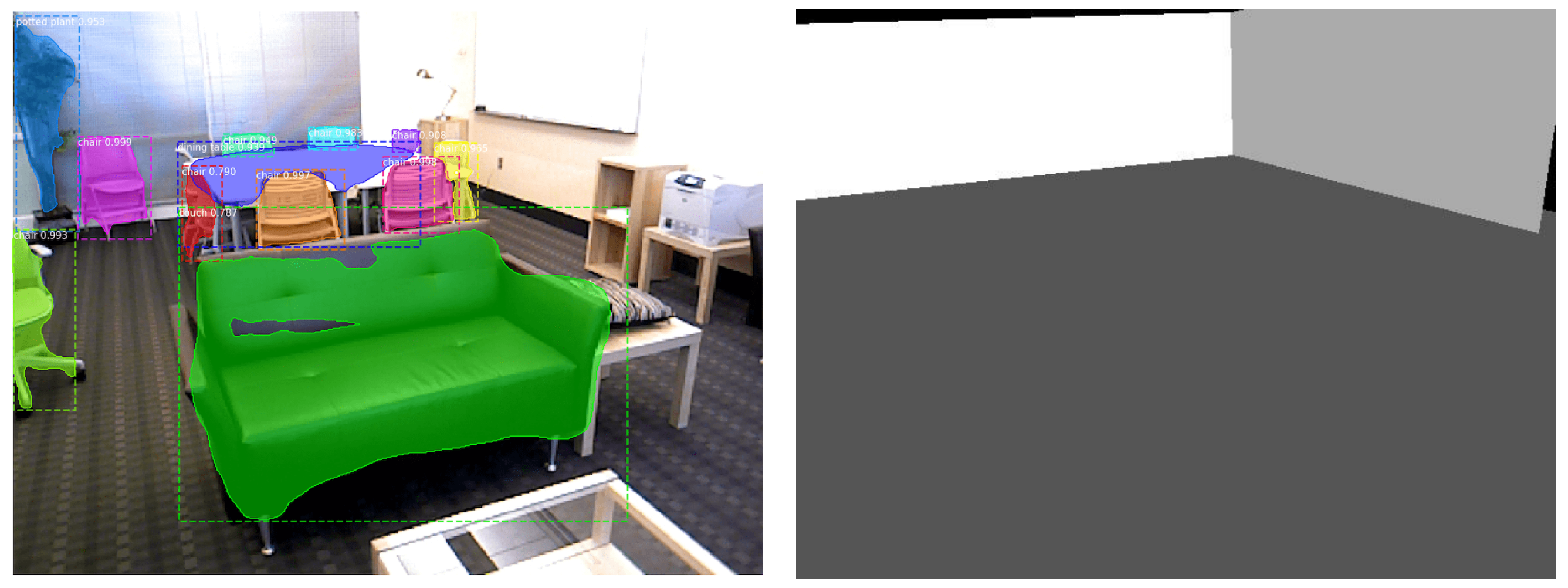

Fig. 2. Sample detection (left) and 3D room layout results (right) from SUN3D dataset.

- ❑ We formulate the pixel-level annotation in a **Conditional Random Field (CRF) energy minimization framework** to combine these two source of information together and produce consistent annotation over the entire video.

3. Model Architecture

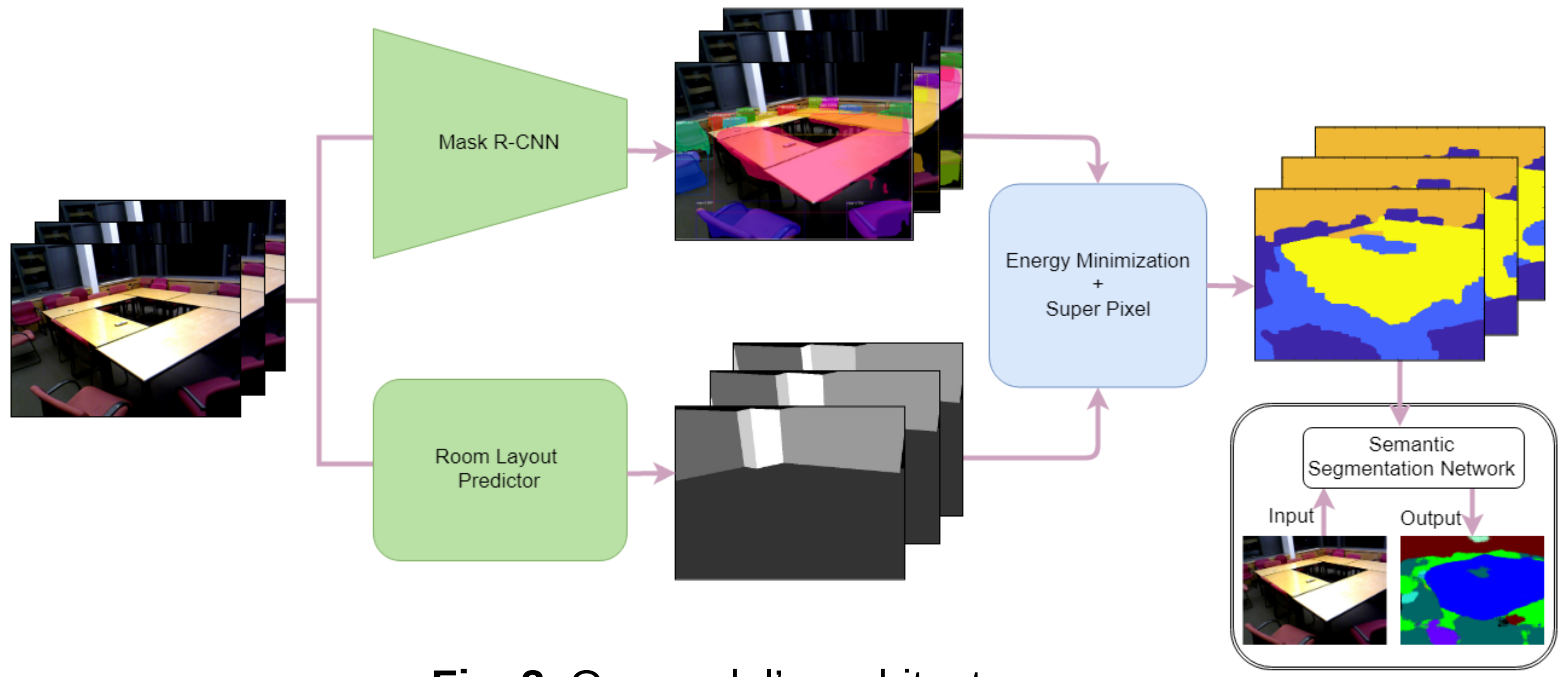

Fig. 3. Our model's architecture.

- ❑ Generate **foreground and background** semantic segmentation maps using Mask RCNN and room layout predictor
- ❑ Generate **superpixels** to accelerate processing and prepare for energy minimization.
- ❑ Pixel annotation: Energy minimization function (definition below)

$$\begin{aligned} E(X_k|I_k, I_{k-1}, I_{k-2}, I_{k-3}) = &\sum_{i\in V} \theta_i(x_i; I_k) \\ &+ \sum_{i\in V} \phi_i(x_i; I_{k-1}, I_{k-2}, I_{k-3}) \\ &+ \sum_{(i,j)\in\zeta} \psi_{ij}(x_i, x_j; I_k), \end{aligned}$$

Definition: Here $\{I_1, I_2 \dots \dots I_N\}$ is the given video sequence. For each frame I_k a CRF graph G=(V, ζ) is defined over all pixels and 4-connected neighbors.

- **First unary term**: For each superpixel we calculate the percentage of pixels belonging to a given class as a score. The first unary term equals the **negative** value of maximum score.
- **Second unary term**: Transfer three immediately preceding frames to the current frame using **optical flow** and take average of the **total** unary energy term at each pixel location.
- **Pairwise term**: Pairwise term: Encourage smoothness by penalizing heterogeneous.

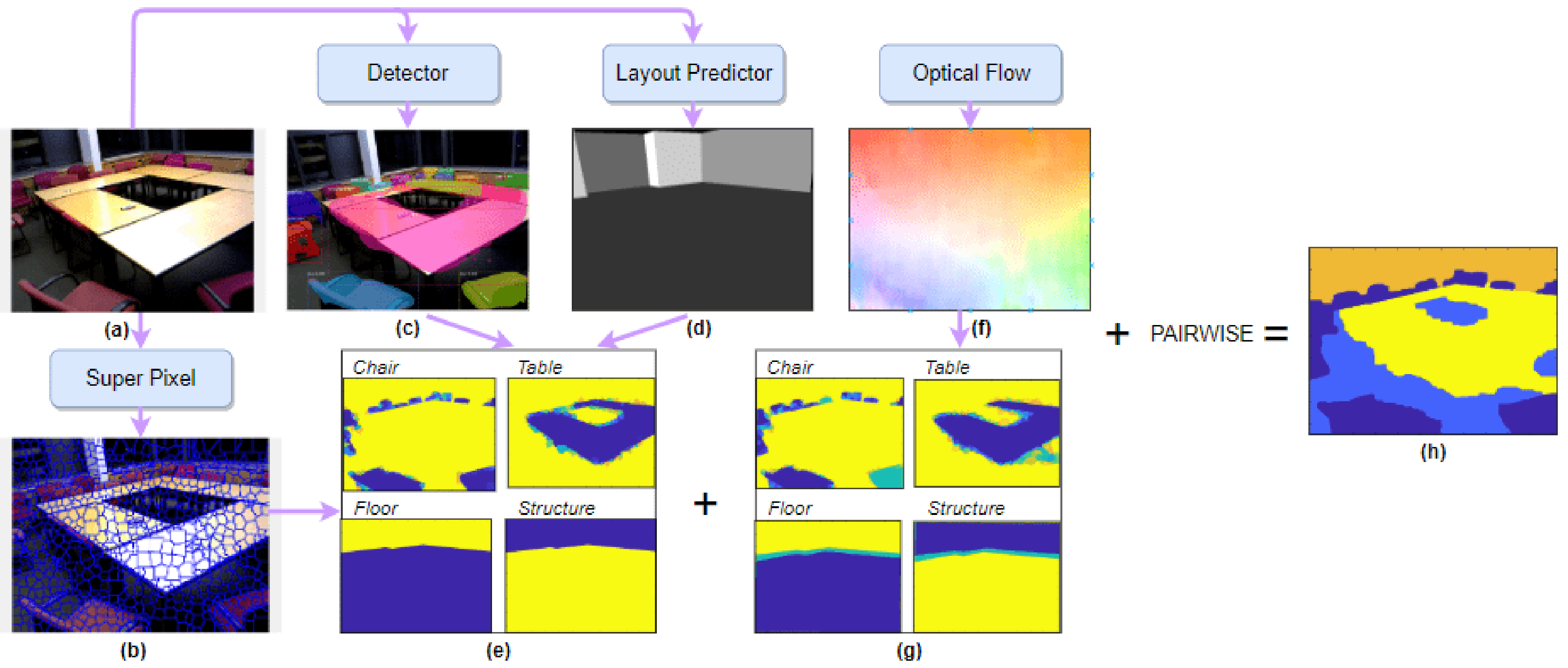

Fig. 4. Visualization of our energy minimization formulation.

4. Experimental Results

- ❑ **SUN3D dataset** consists of 8 different indoor scene videos and all of them have no more than 5% frames with human annotated annotation. We use **4 videos to train** and **the other 4 videos for testing**.
- ❑ We generate annotation for each video frame automatically and use them together with human hand-labeled annotation to train a *Fully Convolution Network* and report performance according to two metrics: *per-class accuracy (left) and per-class IOU (right).*

Video	Bed	Ceiling	Chair	Floor	Furniture	Props	Structure	Table	TV	Mean across category
hotel-umd	81.9 / 60.0	60.6 / 33.6	51.3 / 39.0	56.7 / 37.3	12.9 / 05.6	21.9 / 10.6	66.6 / 59.1	—	54.4 / 52.7	50.8 / 37.2
hv-c5	—	0 / 0	77.6 / 66.9	83.8 / 49.9	0 / 0	64.0 / 05.7	81.7 / 76.4	84.8 / 80.2	—	56.0 / 39.9
studyroom	—	0 / 0	74.8 / 64.8	74.2 / 59.6	36.0 / 31.0	23.9 / 07.5	87.5 / 70.9	48.0 / 45.4	—	49.2 / 39.9
mit-32	—	—	72.4 / 66.0	91.5 / 73.5	—	35.6 / 09.6	69.9 / 62.1	59.6 / 55.4	—	65.9 / 53.3
hv-c6	—	—	77.5 / 68.0	70.2 / 39.7	0 / 0	23.9 / 04.3	87.5/ 84.2	84.4 / 76.8	—	57.2 / 45.5
hv-c8	—	18.5 / 10.9	79.5 / 10.9	95.7 / 68.6	0 / 0	70.5 / 07.1	74.9 / 73.4	77.0 / 74.4	—	59.4 / 44.3
dorm	85.7 / 84.0	49.9 / 42.6	96.8 / 73.4	89.7 / 06.9	14.9 / 08.6	47.1 / 35.5	69.3 / 58.5	47.7 / 44.5	—	62.6 / 44.3
mit-lab	—	0 / 0	99.2 / 75.3	78.2 / 68.9	99.8 / 54.4	18.1 / 15.9	80.5 / 77.0	88.9 / 38.5	—	66.4 / 47.1
Mean across video sequence	**83.8 / 72**	21.5 / 14.5	**78.6 / 58.0**	**80 / 50.6**	23.4 / 14.2	38.1 / 12.0	**77.2 / 70.2**	**70.1 / 59.3**	54.4 / 52.7	—

Table 1: Quantitative evaluation result comparing our automatic annotation on test images and test ground truth.

Train Set	Bed	Ceiling	Chair	Floor	Furniture	Props	Structure	Table	Average value
GT	39.3 / 33.6	68.5 / 54.7	73.3 / 40.2	66.3 / 39.4	16.0 / 10.5	15.4 / 12.0	84.2 / 71.8	88.3 / 74.2	56.4/42.0
Auto	1.2 / 1.0	11.3 / 9.5	53.1 / 21.8	72.0 / 15.7	0.9 / 0.7	4.0 / 2.9	70.8 / 53.9	67.6 / 55.3	35.1/20.1
GT + Auto-sample	15.1 / 13.4	33.8 / 30.7	**81.0** / 38.1	**73.3** / 22.9	10.5 / 7.0	6.4 / 5.4	78.3 / 61.9	**89.6** / 60.2	48.5/30.0

Table 2: Semantic segmentation results using different training label sources.

5. Conclusion

- ❑ In this work, our method relies on two complementary sources: **object detectors and scene layout estimators** to generate dense pixel-level annotation, which is effective for training a deep neural network for semantic segmentation task.
- ❑ Results show that we observe improvement in some classes but there is still space for improvement.
- ❑ In the future, we plan to augment the method to generate annotation for a large number of fine-grained indoor object categories.

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in CVPR, 2015.
[2] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in ICCV, 2017.
[3] C. Taylor and A. Cowley, "Parsing indoor scenes using RGB-D imagery," in RSS, 2012.

***Acknowledgements**: This work was supported in part by the National Science Foundation (CAREER IIS-1253549). Kai Chen participated in this project while visiting Indiana University through the IU Global Talent Attraction Program (GTAP) program.*