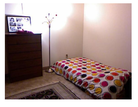
Indoor Scene Parsing with Instance Segmentation, Semantic Labeling and Support Relationship Inference

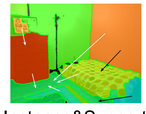
Wei Zhuo [1,3], Mathieu Salzmann [2], Xuming He [1,3], Miaomiao Liu [1,3]
[1] Australian National University, [2] EPFL, [3] Data61-CSIRO

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

Objective:

Analyze a scene by jointly estimating its instances, their semantic labels and support relationships between instances (e.g., the floor supports the desk from below).

Input: 

Output:  Our Semantic  Instance&Support

Motivation:

Strong connections exist among the above mentioned tasks

- good regions respect semantic labels;
- support relationships can only be defined on meaningful regions;
- support relationships strongly depend on semantics.

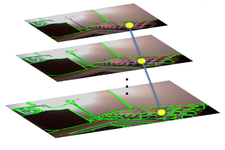
Contribution:

Compared to previous work [2,3], we

- jointly train instance segmentation with support relationships;
- perform prediction from a single RGB image.

Overview:

Given a hierarchical segmentation, we formulate the joint learning problem as selecting the best set of regions. We seek regions that have



- a high probability of being instances;
- homogenous semantic labels;
- a high probability of having valid support relationships.

Reference:

[1] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.
[2] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In ECCV, 2012.
[3] N. Silberman, D. Sontag, and R. Fergus. Instance segmentation of indoor scenes using a coverage loss. In ECCV, 2014.
[4] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In NIPS, 2014.
[5] K.He,X.Zhang,S.Ren,andJ.Sun.Spatial pyramid pooling in deep convolutional networks for visual recognition. In ECCV, 2014.

Models:

We formulate our problem as inference in a CRF, whose energy is

$$E(A,M,S) = \sum_{i=1}^{R} \phi_a(a_i) + \sum_{i=1}^{R} \phi_{ma}(M_i,a_i) + \phi_{tree}(A) + \sum_{i=1}^{R}\sum_{j=0}^{R} \phi_s(S_{ij}) + \sum_{i=1}^{R}\sum_{j=0}^{R} \phi_{sa}(S_{ij},a_i,a_j)$$

with variables

- A: binary variables indicating whether a region is selected;
- M: semantic labels defining the class to which a region belongs, for K classes;
- S_{ij}: variables defining the type of support that region j provides to region i; the support types include {no support, support from below, support from behind};

and potentials

- ϕ_a, ϕ_s: unaries for region selection and support types;
- ϕ_{ma}: probability of predicting a particular semantic label for a region if it is active;
- ϕ_{sa}: dependencies between the support variables and the region selection ones;
- ϕ_{tree}: enforces that only one region is selected in every path from the root of the hierarchy to a leaf.

All potentials rely on deep features [1,4,5] and hand-craft ones [2].

Learning with Structural SVM:

Let $(x^{(n)}, y^{(n)})$ be a set of pairs of images and labels, with $y^{(n)}$ is ground truth labels. Let $\phi(x,y) = [\phi_a, \phi_{ma}, \phi_s, \phi_{sa}]$. We express training as

$$min_{w,\epsilon\leq 0} \frac{1}{2} w^T w + \frac{\lambda}{N}\sum_{n}^{N} \epsilon_n ,$$

$$\text{s.t.} \quad w^T\left[\phi(x^{(n)},y^{(n)}) - \phi(x^{(n)},y)\right] \geq \Delta(y^n,y) - \epsilon_n, \forall y$$

where $\Delta(y^n,y)$ returns the loss of an arbitrary prediction y compared to the best configuration.

$$\Delta(y^n,y) = \frac{w_{sup}^{ls}}{Q}\sum_{i=1}^{R}\sum_{j=0}^{R} 1[S_{ij} \neq S_{ij}^*] + w_r^{ls}\frac{1}{L}\sum_{g\in G} L_{r_g}(min_{i\in A^{(n)}} IoU(r_g - r_i^{(n)})) - w_r^{ls}\frac{1}{L}\sum_{g\in G} L_{r_g}(min_{i\in A} IoU(r_g - r_i))$$

where $r_i^{(n)}$ is the oracle set of regions, which best match ground truth in our hierarchy, S_{ij}^* is the ground truth support label.

Inference:

For both training and test, we do exact inference by Integer Linear Programming. To speed up inference, we trained

- an IoU regressor based on a shallow neural network to reduce the number of regions;
- a binary SVM classifier achieving a high recall on pairs of positive support types to reduce the number of region pairs.

Evaluation:

We evaluate quantitatively on the NYUv2 depth dataset. Correct support predictions are shows as white lines, incorrect ones in black.

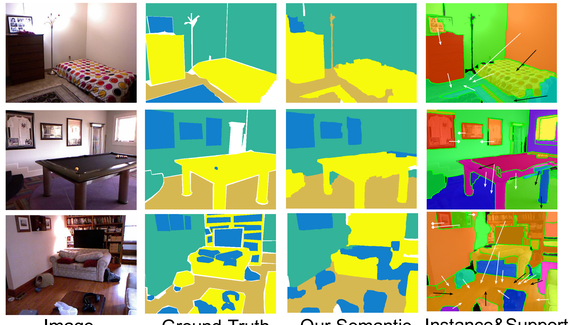


Image Ground-Truth Our Semantic Instance&Support

Ablation study

Model	W.Cov	Sem Avg Acc	Sem Per-Cls Acc	Support Precision	Support Recall
Basic	58.9	-	-	-	-
SC	-	-	-	44.8	39.0
Ours-NS	59.3	73.0	72.0	-	-
Ours-ND	59.3	73.3	72.2	47.0	41.9
Ours	59.4	73.2	72.1	47.6	43.1
Ours(GtSem)	60.1	-	-	48.2	45.0

Comparison to baselines

Model	Orable W.Cov	W.Cov	Sem Avg Acc	Sem Per-Cls Acc	Support Precision	Support Recall
Basic	68.8	61.1	-	-	-	-
SC	-	-	-	-	48.3	37.9
Ours-NS	68.8	62.8	74.8	73.7	-	-
Ours	68.8	62.7	75.3	74.3	49.5	38.6
[3]	70.6	62.5	-	-	-	-
[2]	-	-	-	-	54.5	-

Conclusion:

Our experiments demonstrate that jointly reasoning about the three tasks is in general beneficial, particularly for support relationships.