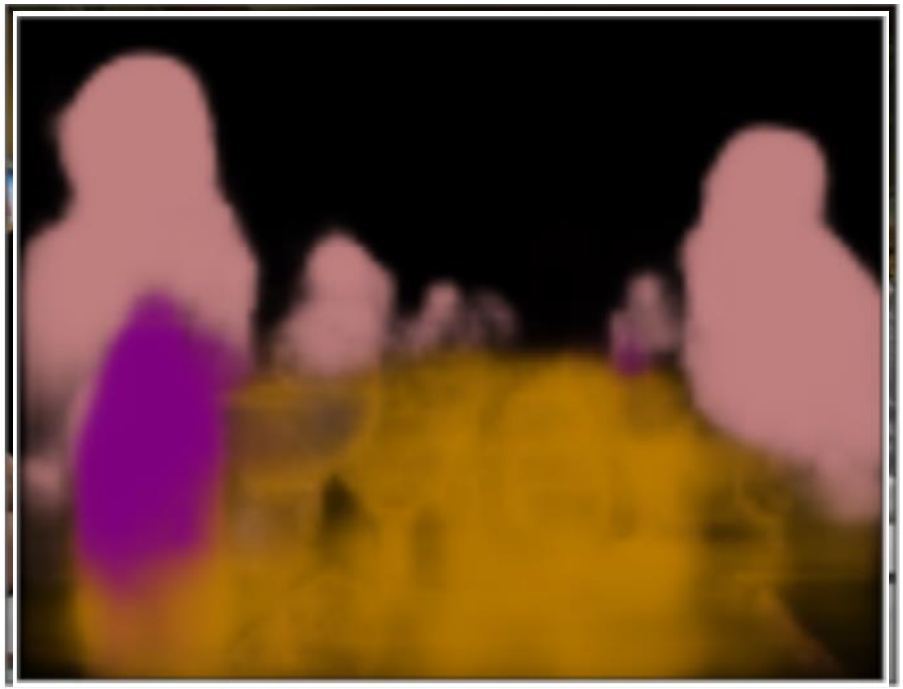
Multi Visual Task Fusion with Deep CNN and Conditional Random Field

Peng Wang, UCLA

Why it is important to fuse multi-tasks in vision

Human are performing multi-tasks simultaneously and register them well.

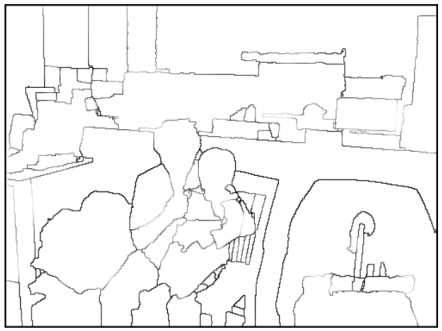


Only by understanding fully and densely to the given scene, we can have confidence to do visual question and answering.

Example results from Kokinnos Arxiv 1609.02132

Why it is important to fuse multi-tasks in vision

Single task could be biased due to a single loss from the system is almost always limited, which can be regularized by other tasks.



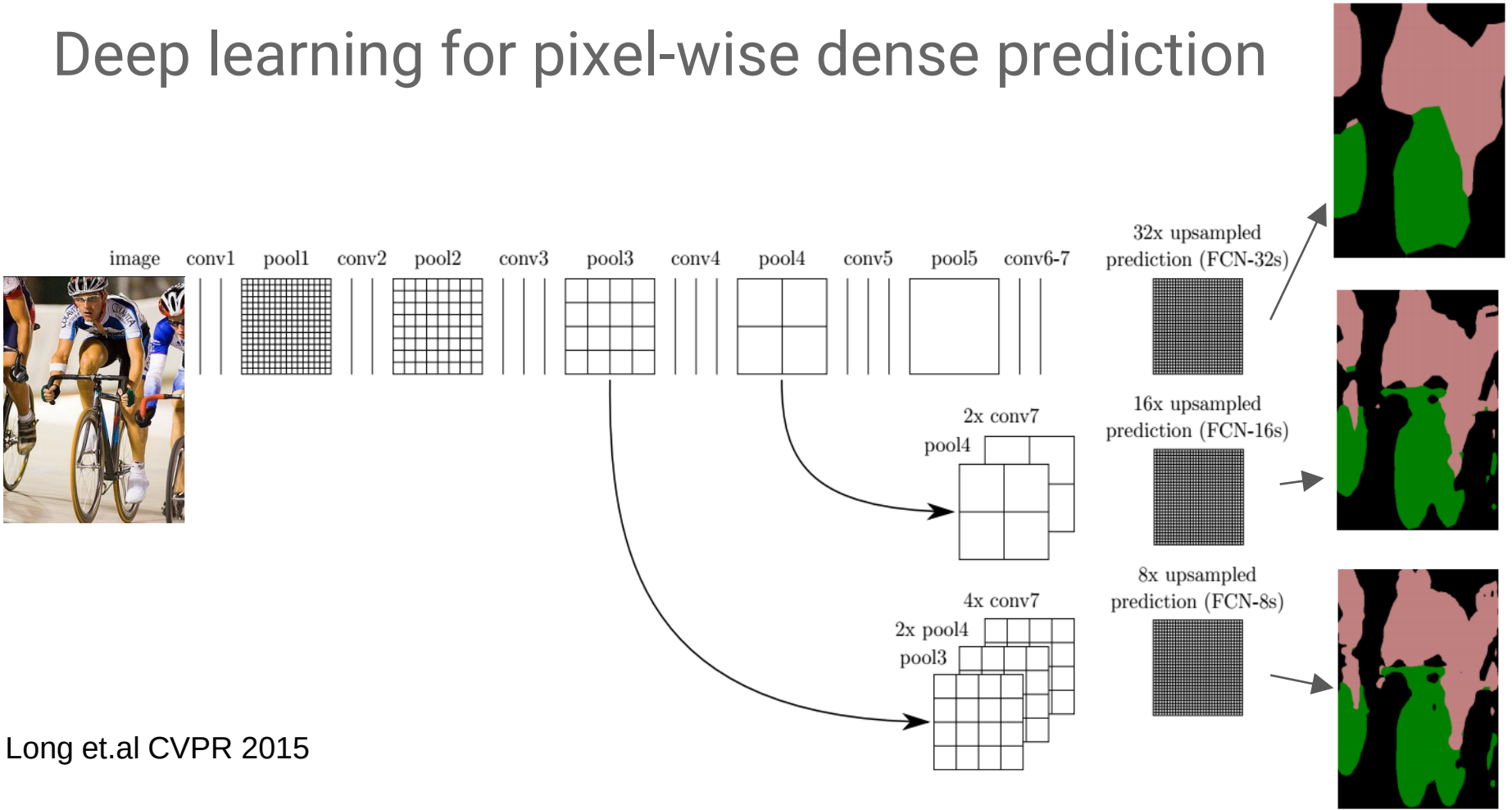
FCN

Bertasius et.al CVPR 2016

Another example of optical flow

Sevilla-Lara et.al CVPR 2016

Deep learning for pixel-wise dense prediction



Long et.al CVPR 2015

Extension afterwards

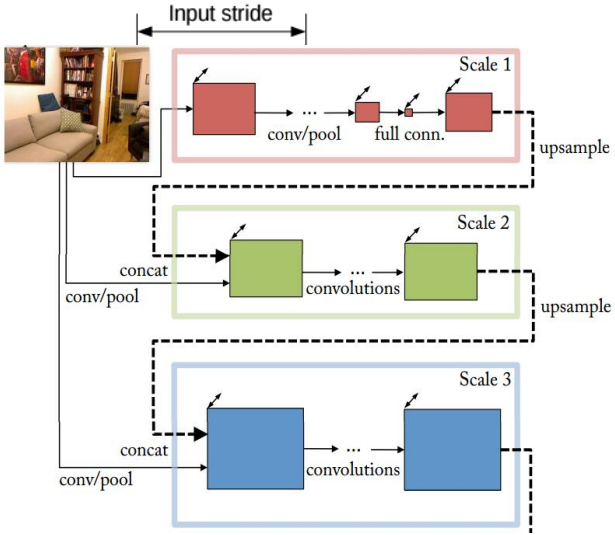
Image

FCN Network

Atrous FCN

Multi-scale FCN

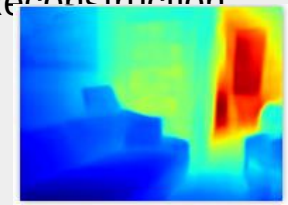
Chen et.al ICLR 2015
Eigen&Fergus ICCV 15

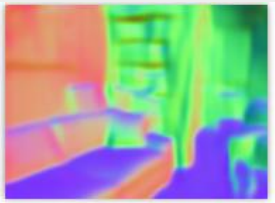


Edge prediction

Kokinnos Arxiv 1609.02132

Reconstruction





Eigen&Fergus ICCV 15

Pose estimation

Insafutdinov et.al ECCV 201

Detection, low level processing, style transfer ...

Extension afterwards

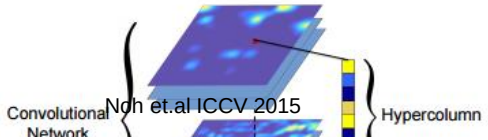
Image

FCN Network

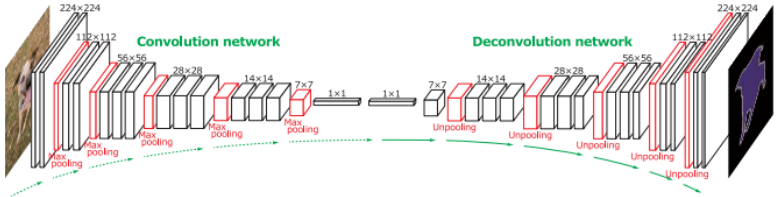
Hypercolumn FCN

Hariharan CVPR 2015

Encoder-Decoder



Noh et.al ICCV 2015

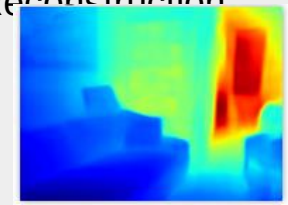


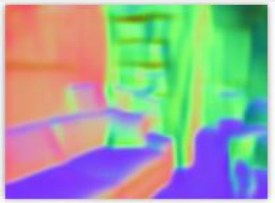
VGG, Inception, Resnet, Inception Resnet etc...

Edge prediction

Kokinnos Arxiv 1609.02132

Reconstruction





Eigen&Fergus ICCV 15

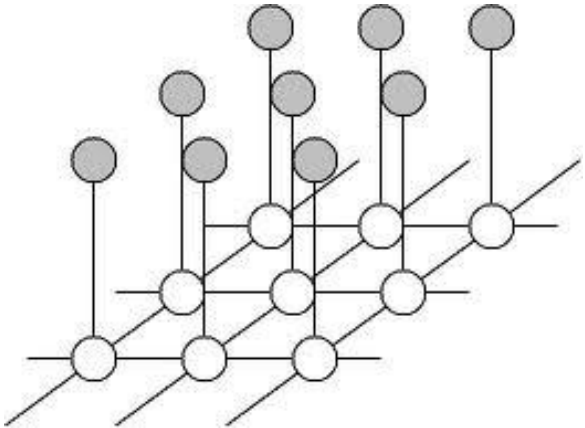
Pose estimation

Insafutdinov et.al ECCV 201

Detection, low level processing, style transfer ...

Conditional Random Field (CRF)

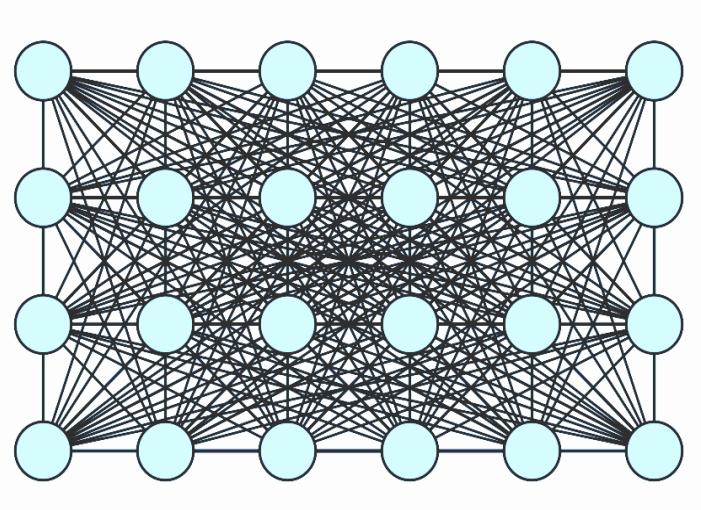
Useful for structure learning and reference, which could be modeled to look at neighbor context and smooth the predictions



$$\Pr(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathrm{Z}(\mathbf{x})} \exp(-E(\mathbf{y}, \mathbf{x})),$$

$$E(\mathbf{y}, \mathbf{x}) = \sum_{p \in \mathcal{N}} U(y_p, \mathbf{x}) + \sum_{(p,q) \in \mathcal{S}} V(y_p, y_q, \mathbf{x}).$$

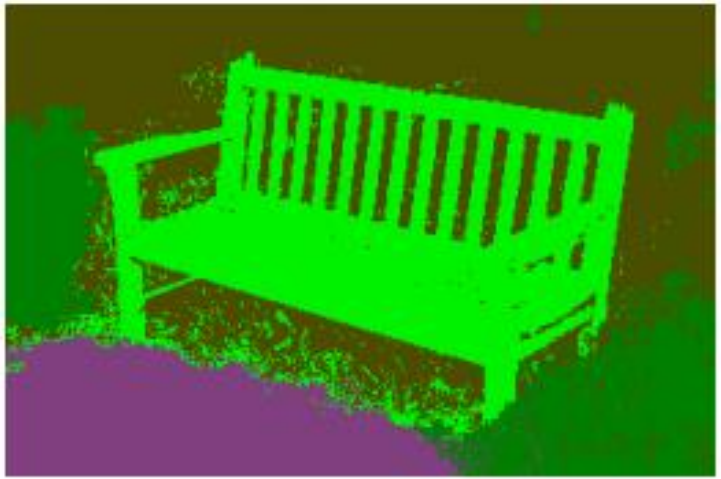
Fully connected CRF



$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j),$$

Connect every pair

Difference



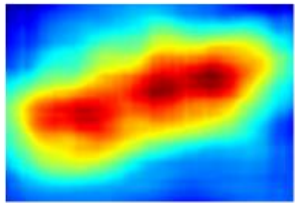
Krahenbuhl & Koltun NIPS 2012

Access long range context in bilateral space

Recent applications

Discrete labels

Image



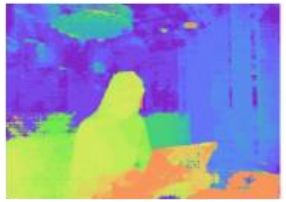
Network output

Refine output

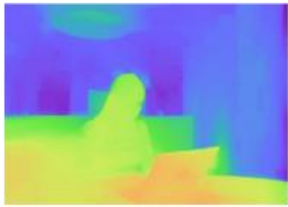
[1] Krähenbühl et.al ICML 2013
[2] Chen et.al ICLR 2014
[3] Zheng et.al ICCV 2015

Continuous labels (Bilateral smoothing)

Image



Stereo output



Refine output

[1] Barron et.al CVPR 2015
[2] Barron et.al arXiv 2016

CRF has long been commonly used in single or multi tasks

Pre-CNN period

SIFT (HOG) + SVM (Structured SVM) for unary energy over pixel or super-pixel, e.g.

C 

s afterwards

CNN period (Just replace the unary ? What else we have from CNN?)

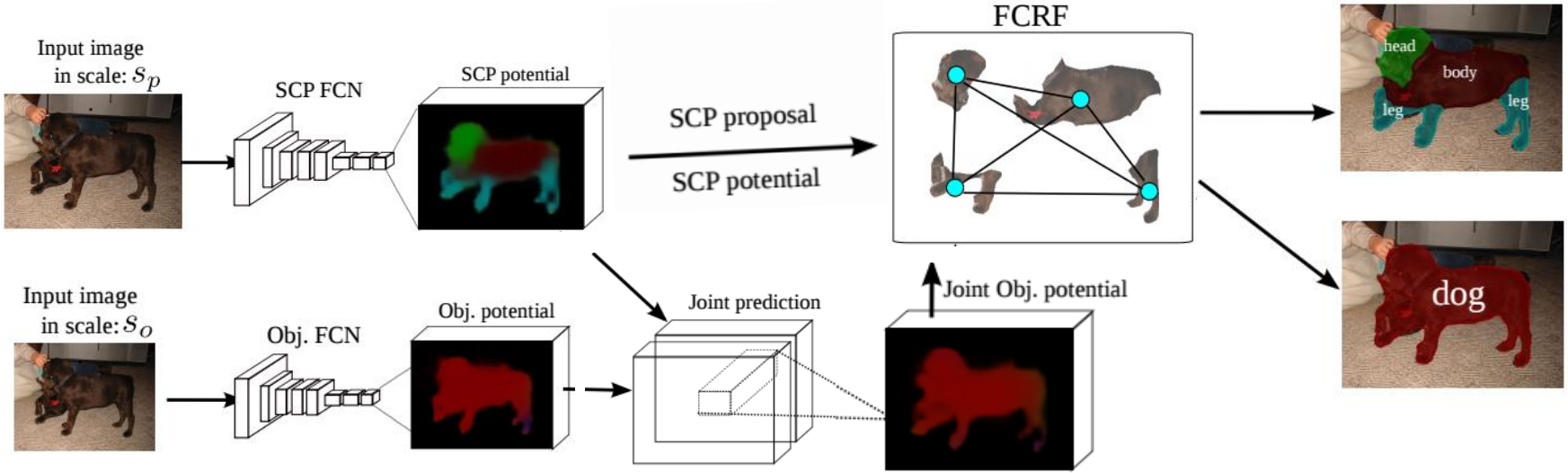
More efficient, unified and robust features from deep learning, which allows us to model multi-tasks more effectively

Two applications from the intuition

[1] **Peng Wang**, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, Alan Yuille, *Joint Object and Part Segmentation using Deep Learned Potentials*, **ICCV** 2015

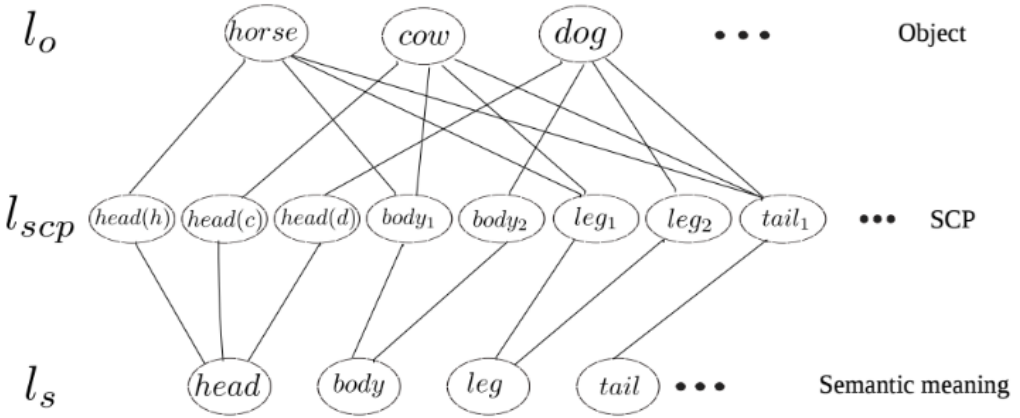
[2] **Peng Wang**, Xiaohui Shen, Bryan Russel, Scott Cohen, Brian Price, Alan Yuille, *SURGE: Surface Regularized Geometric Estimation from a Single Image*, **NIPS** 2016

Joint Object and Part Segmentation

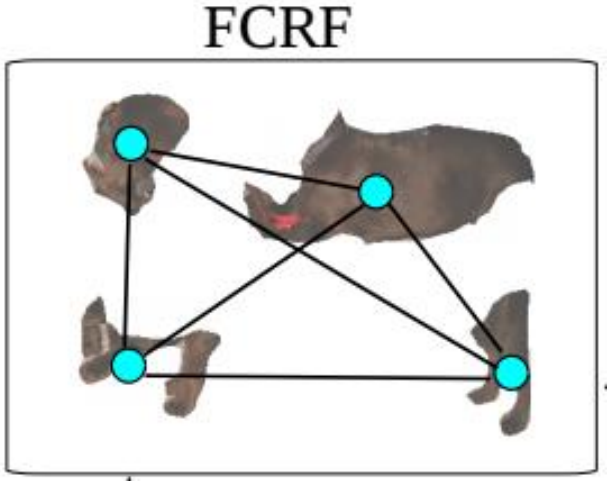


Part sharing

Handle the growth of joint label space



Joint FCRF formulation



$$\min_{\mathcal{L}} \sum_{i \in \mathcal{V}} \psi_i(l^i_{op}) + \lambda_e \sum_{i,j \in \mathcal{V}, i \neq j} \psi_{i,j}(l^i_{op}, l^j_{op})$$

$$\psi_i(l^i_{op}) = \eta(l^i_o, l^i_p)(\psi^o_i(l^i_o) + \lambda_p \psi^p_i(l^i_p))$$

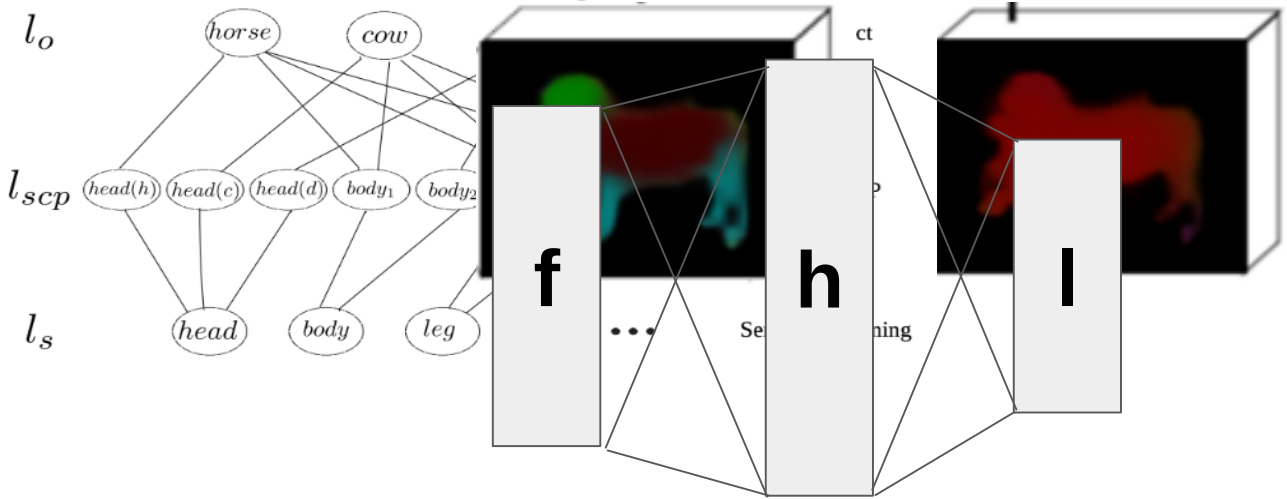
$$\psi_{i,j}(l^i_{op}, l^j_{op}) = \eta(l^i_o, l^i_p)\eta(l^j_o, l^j_p)\psi^{op}_{i,j}(l^i_o, l^j_o, l^i_p, l^j_p)$$

Unary

Pairwise

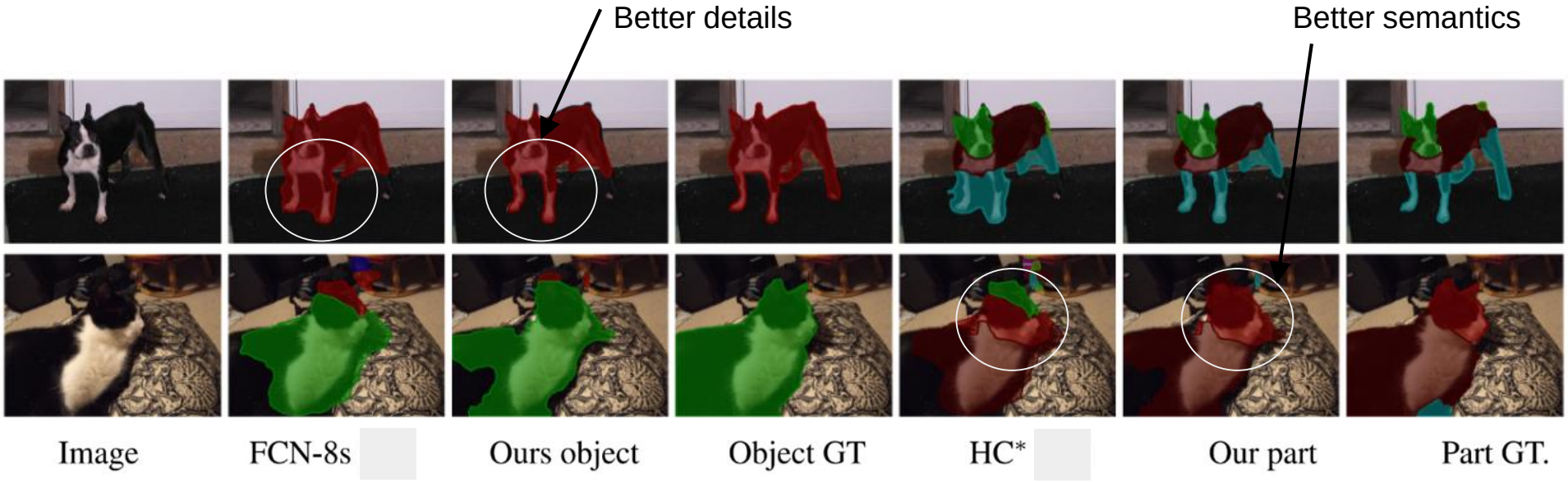
$$\eta(l_o^i, l_p^i)(\psi_i^o(l_o^i) + \lambda_p \psi_i^p(l_p^i)) \qquad \eta(l_o^i, l_p^i)\eta(l_o^j, l_p^j)\psi_{i,j}^{op}(l_o^i, l_o^j, l_p^i, l_p^j)$$

$$\mathbf{f}_{ij} = [\mathbf{f}_i^T, \mathbf{f}_j^T, \kappa_{ij}, \theta_{j|i}]^T, \text{ where } \mathbf{f}_i = [\mathbf{f}_{oi}^T, \mathbf{f}_{pi}^T, \mathbf{f}_{ai}]^T$$



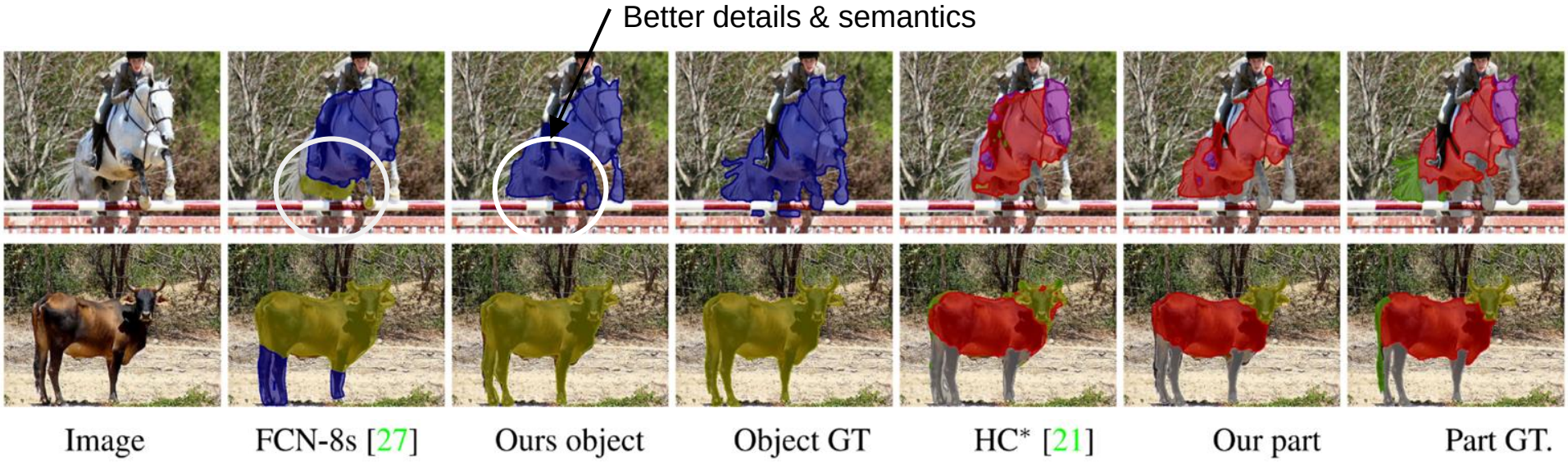
Results

Less confusion and more details due to larger context and joint task performed.

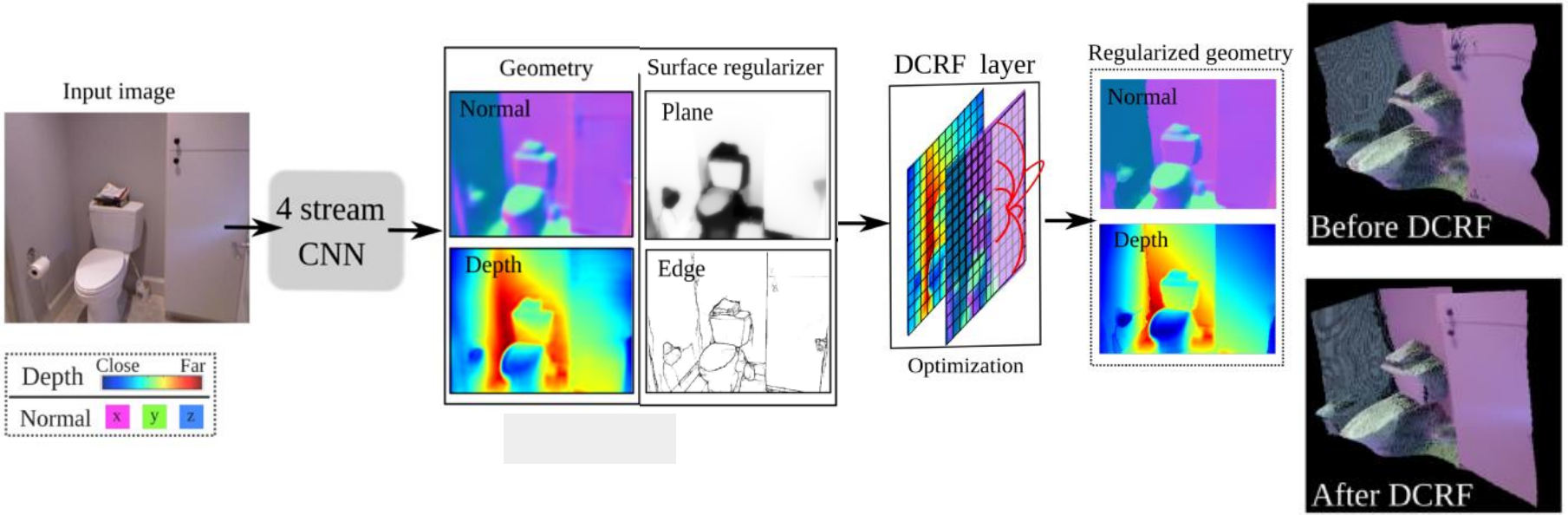


Additional results

Less confusion and more details due to larger context and joint task performed.



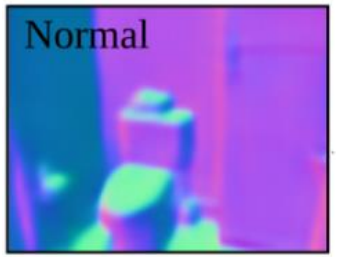
3D geometry reconstruction (Depth & Normal)

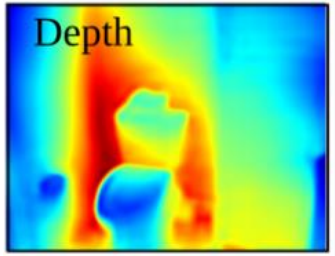


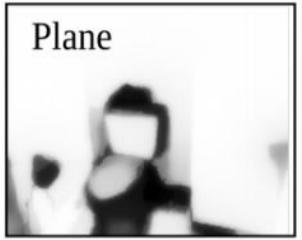
Formulation of the DCRF

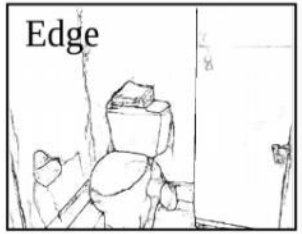
$$\min_{\mathbf{N},\mathbf{D}}\{\sum\nolimits_i \psi_u(\mathbf{n}_i, d_i|\mathbf{N}_o, \mathbf{D}_o) + \lambda \sum\nolimits_{i,j} \psi_r(\mathbf{n}_i, \mathbf{n}_j, d_i, d_j|\mathbf{P}_o, \mathbf{E}_o)\}, \text{ with, } \|\mathbf{n}_i\|_2 = 1$$

$$\sum_i \psi_{\mathbf{n}}(\mathbf{n}_i) + \sum_i \psi_d(d_i) \qquad \sum_{i,j} [\mu_{\mathbf{n}}(\mathbf{n}_i, \mathbf{n}_j) + \mu_d(d_i, d_j|\mathbf{N})]\mathbf{A}_G(i, j|\mathbf{P}, \mathbf{E})\}$$

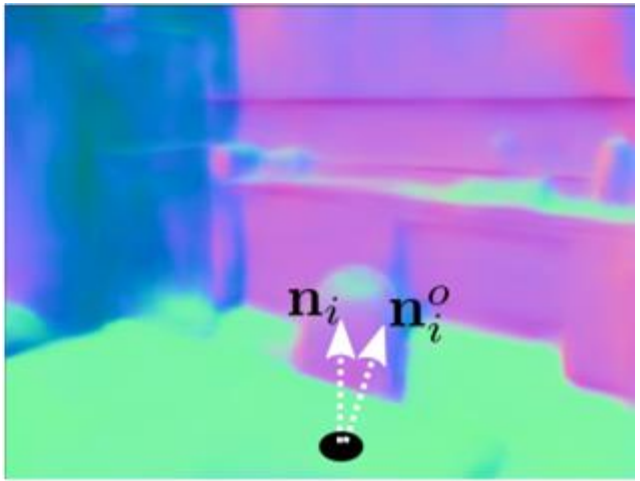






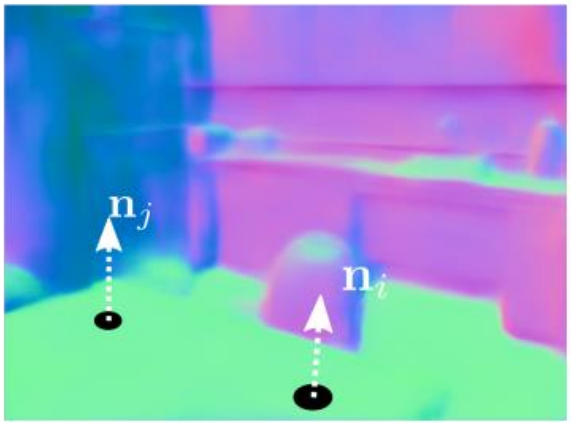


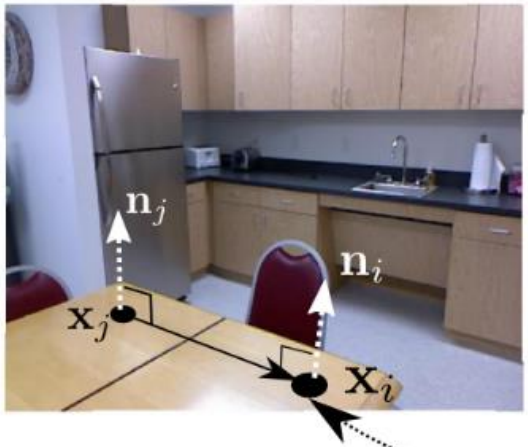
$$\min_{\mathbf{N},\mathbf{D}}\{\sum_i \psi_{\mathbf{n}}(\mathbf{n}_i) + \sum_i \psi_d(d_i) \; + \lambda \sum_{i,j}[\mu_{\mathbf{n}}(\mathbf{n}_i,\mathbf{n}_j) + \mu_d(d_i,d_j|\mathbf{N})]\mathbf{A}_G(i,j|\mathbf{P},\mathbf{E})\}$$

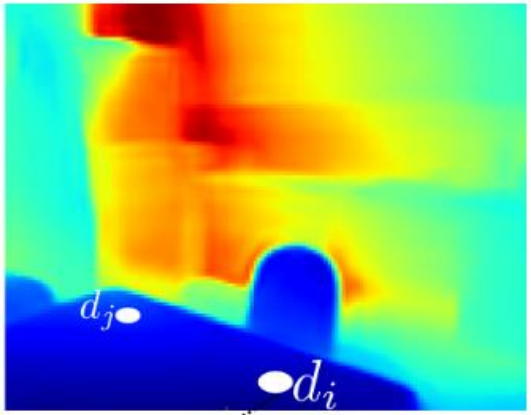


$$\min_{\mathbf{N},\mathbf{D}}\{\sum_i \psi_{\mathbf{n}}(\mathbf{n}_i) + \sum_i \psi_d(d_i) \; + \lambda \sum_{i,j}[\mu_{\mathbf{n}}(\mathbf{n}_i,\mathbf{n}_j) + \mu_d(d_i,d_j|\mathbf{N})]\mathbf{A}_G(i,j|\mathbf{P},\mathbf{E})\}$$

orthogonal compatibility





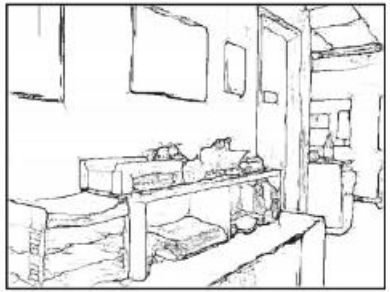


$$\min_{\mathbf{N},\mathbf{D}}\{\sum_i \psi_{\mathbf{n}}(\mathbf{n}_i) + \sum_i \psi_d(d_i) \; + \lambda \sum_{i,j} [\mu_{\mathbf{n}}(\mathbf{n}_i, \mathbf{n}_j) + \mu_d(d_i, d_j|\mathbf{N})]\mathbf{A}_G(i, j|\mathbf{P}, \mathbf{E})\}$$

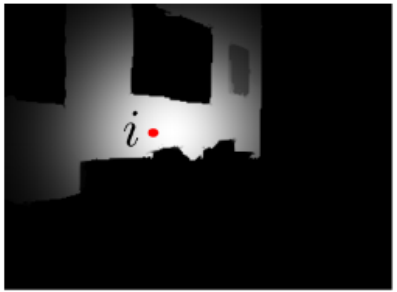
Planar Affinity

Image

Plane



Edge



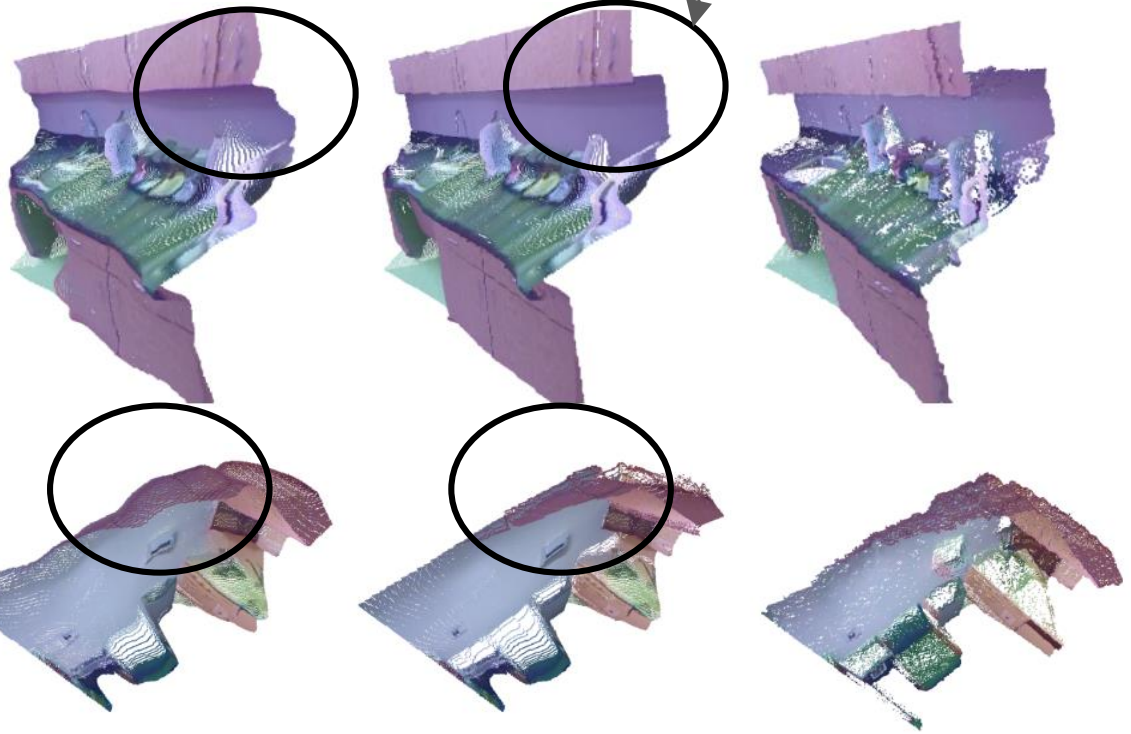
Pairwise planar affinity

Finally, we make the DCRF layer trainable for both normal and depth.

Results

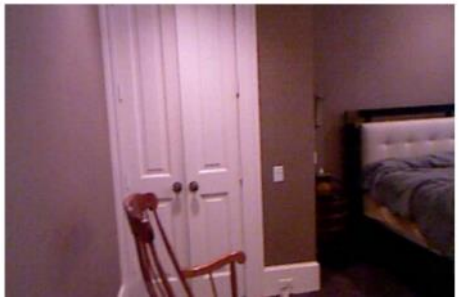
Better 3D planar

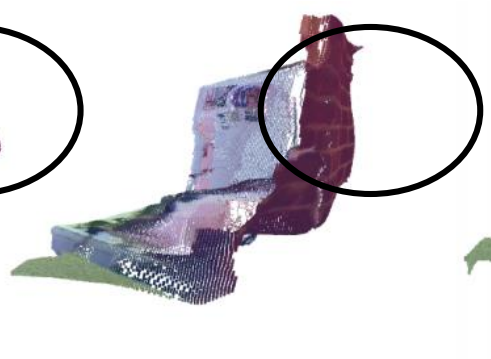
Image | Network output | Regularization | Ground truth

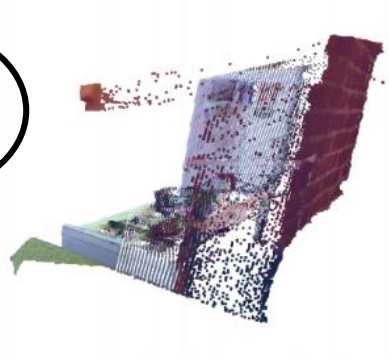


Results

Image | Network output | Regularization | Ground truth







Take home message

1. Performing multi-tasks and register them well could help visual tasks.

1. CNN and CRF could be served as an easy starting approach to model relationships.

1. Discover the complementary property could be either learned if you have large data or discovered from observations.

1. Still long way to go, and a lot of opportunities to combine and register tasks.