SEMANTIC SCENE PARSING BY INFORMATION PURSUIT

Donald Geman
Johns Hopkins University

SIAM Conference on Imaging Science
Albuquerque, New Mexico
May 25, 2016

COLLABORATORS

- Ehsan Jahangiri (former PhD student, JHU)
- Erdem Yoruk (former PhD student, JHU)
- Laurent Younes
- Rene Vidal
- Many of the basic ideas go back to work with Bruno Jedynak on “active testing” for tracking and with Yali Amit on decision trees and coarse-to-fine indexing for object detection.

MACHINES VS. HUMANS

- Interpreting scenes is effortless and instantaneous for people, even generating rich semantic annotations (“telling a story”).
- Machines lag very far behind in understanding images, and building a *description machine* remains a fundamental A.I. challenge.
- This remains true even for the restricted task of detecting and localizing all instances from a set of object categories.

STREET SCENES

Table Scenes




Outline

- General Query Model
- Information Pursuit
- Table Settings

Scene Variables

- $I \in \mathcal{I}$: observed image of an underlying visual scene.
- $Z \in \mathcal{Z}$: latent description or interpretation of the scene.
- $U \in \mathcal{U}$: other typically unobserved components, e.g., camera properties and view angles.
- Assume that Z, U and I are random variables on some probability space.
- Goal: Reconstruct (as much as possible about Z) from the observation of I.

Motivation for a Model

- Despite a mass migration to deep learning, due in part to genuine advances and ubiqitous success stories, there is no evident path from object detection to deep semantic annotation, e.g., story telling.
- Interesting attributes of natural vision and decision-making which are missing in most machine vision systems:
  - Exploiting *context* to remove amibiguities.
  - Analyzing scenes at different levels of resolution which is often coarse-to-fine.
  - Excelling at games like “Twenty Questions” by asking the right questions in the right order.

QUERIES

- We acquire evidence from different levels of semantic and geometric resolution and integrate the evidence by updating likelihoods.
- Evidence is collected from the answers to a series of queries q about I from a specified set $\mathcal{Q}$.
- $Y_q = f_q(Z, U)$: a bit of information or "annobit" (not necessarily binary).
- The dependency on U allows q to depend on locations relative to the observed image.
- Strategy: Progressively estimate annobits by running matching, unit cost classifiers, $X_q(I), q \in \mathcal{Q}$.
- We think of Y_q as the true answer and X_q as an imperfect answer in a "Twenty Questions" game.

EXAMPLES OF ANNOBITS

- *Scene context:* global labels such as “indoor” and ”street scene.” etc. Not used here.
- *“Part-of” descriptors:* Indicate that a region *R* belongs to a larger structure (e.g.,road, building, table).
- *Existence descriptors:* Indicate the presence of objects with certain attributes (e.g., category and pose).
- *Derived annobits* are also useful, e.g., a list of object categories with instances visible within a region.

Query Model

- Assume $Y_{\mathcal{Q}} = (Y_q, q \in \mathcal{Q})$ is a sufficient statistic for $X_{\mathcal{Q}}$:

$$P(X_{\mathcal{Q}}|Z) = P(X_{\mathcal{Q}}|Y_{\mathcal{Q}}).$$

- Assume a prior scene distribution for Z and a prior camera/viewpoint distribution for U.
- The prior model $p(z)$ encodes knowledge about likely and unlikely configurations (spatial context).
- Combining the prior with the data model $P(X_{\mathcal{Q}}|Y_{\mathcal{Q}})$ would in principle allow us to sample from the posterior $P(Z|X_{\mathcal{Q}})$, which modulates or *contextualizes* raw classifier output.

Outline

- General Query Model
- Information Pursuit
- Table Settings

Query Ordering

- Set $q_1 = \arg\max_{q \in \mathcal{Q}} \mathcal{I}(X_q, Y_{\mathcal{Q}})$.
- Thereafter, for $k > 1$ and given an image I,

$$q_k(I) = \arg\max_{q \in \mathcal{Q}} \mathcal{I}(X_q, Y_{\mathcal{Q}} | \mathbf{e}_{k-1}(I))$$

  where $\mathbf{e}_{k-1}(I)$ is the "evidence" or "history" after $k-1$ queries: $\mathbf{e}_{k-1}(I) = \{X_{q_\ell} = x_\ell,\ \ell = 1, \ldots, k-1\}, x_\ell = X_{q_\ell}(I)$.
- We sometimes denote the history by $(q_1, \ldots, q_{k-1}, x_1, \ldots, x_{k-1})$.
- Note that the conditional distribution for the mutual information is the posterior distribution of $(X_q, Y_{\mathcal{Q}})$ given $X_{q_1}, \ldots, X_{q_{k-1}}$.

First Characterization

- Since $\mathcal{I}(X_q, Y_{\mathcal{Q}}|\mathbf{e}_{k-1}(I)) =$

$$H(Y_{\mathcal{Q}}|\mathbf{e}_{k-1}(I)) - H(Y_{\mathcal{Q}}|X_q, \mathbf{e}_{k-1}(I)),$$

  the next question $q_k(I)$ for image I is the one whose addition will minimize $H(Y_{\mathcal{Q}}|X_q, \mathbf{e}_{k-1}(I))$.
- Here, again, $Y_{\mathcal{Q}}$ and X_q are random variables, and the conditional entropy is computed for the conditional probability $P(\cdot|\mathbf{e}_{k-1}(I))$.

Second Characterization

- We also have $\mathcal{I}(X_q, Y_{\mathcal{Q}} | \mathbf{e}_{k-1}(I)) = H(X_q | \mathbf{e}_{k-1}(I)) - H(X_q | Y_{\mathcal{Q}}, \mathbf{e}_{k-1}(I))$.
- This implies that the next question for image I is selected such that
  - $H(X_q | \mathbf{e}_{k-1}(I))$ is large, so that its answer is as unpredictable as possible given the current evidence and
  - $H(X_q | Y_{\mathcal{Q}}, \mathbf{e}_{k-1}(I))$ is small, i.e., X_q is a "good" classifier.
- The two criteria are however balanced, so that one could accept a (currently) relatively poor classifier if it is (currently) highly unpredictable.

An Approximation

- Depending on $P(X_{\mathcal{Q}}, Y_{\mathcal{Q}})$, these conditional entropies may not be easy to compute.
- For table settings, we neglect the error made by X_q at the selection stage, replacing X_q by Y_q.
- Consequently,

$$q_k = \underset{q \in \mathcal{Q} \setminus \{q_1, \ldots, q_{k-1}\}}{\arg\max} H(Y_q | \mathbf{e}_{k-1}(I)).$$

- However, $X_{\mathcal{Q}}$ and $Y_{\mathcal{Q}}$ are not assumed to coincide in the conditioning event $\{\mathbf{e}_{k-1}(I)\}$ (which depends on the X variables) so that the accuracy of the classifiers is still accounted for when updating the posterior.

Conditional Independence

- Selection is simplified if one assumes that the classifier outputs are conditionally independent given $Y_{\mathcal{Q}}$.
- In that case, using the fact that $\mathbf{e}_{k-1}$ only depends on the realizations of X, one has

$$H(X_q|Y_{\mathcal{Q}}, \mathbf{e}_{k-1}(I)) = \begin{cases} H(X_q|Y_{\mathcal{Q}}) \text{ if } q \notin \{q_1, \ldots, q_{k-1}\} \\ 0 \text{ otherwise} \end{cases}$$

- Therefore the conditional entropy is directly computable from the data model.
- The other term $H(X_q|\mathbf{e}_{k-1}(I))$, can be computed using Monte-Carlo simulations with a complexity similar to the oracle approximation.

Z VERSUS Y

- We have used the annobits $Y_{\mathcal{Q}}$ to represent the unknown scene Z in selecting questions.
- $H(Z|\mathbf{e}_{k-1}(I)) - H(Y_{\mathcal{Q}}|\mathbf{e}_{k-1}(I)) = H(Z|Y_{\mathcal{Q}}, \mathbf{e}_{k-1}(I))$.
- The difference is small if the residual uncertainty of Z given $Y_{\mathcal{Q}}$ is small, which is the approximation made here.

OUTLINE

- General Query Model
- Information Pursuit
- Table Settings

Prior Model

- Here $U = (\mathcal{W}, T)$, where
  - $\mathcal{W}$ is the set of intrinsic (calibration matrix) and extrinsic (pose in 3D) camera parameters;
  - T specifies the table dimensions and we assume the table is centered at the origin and lies in the xy-plane of the 3D coordinate system.
- The homography H is a function of $\mathcal{W}$.
- The r.v.s (Z, T) and $\mathcal{W}$ are assumed independent and the prior model is then $P(Z, T, \mathcal{W}) = P(\mathcal{W})P(T)P(Z|T)$.
- Will skip $P(\mathcal{W})$ and $P(T)$.
- Recall $Y_q = f_q(Z, U)$; hence the distribution of $(Y_q, q \in \mathcal{Q})$ is determined by the prior.

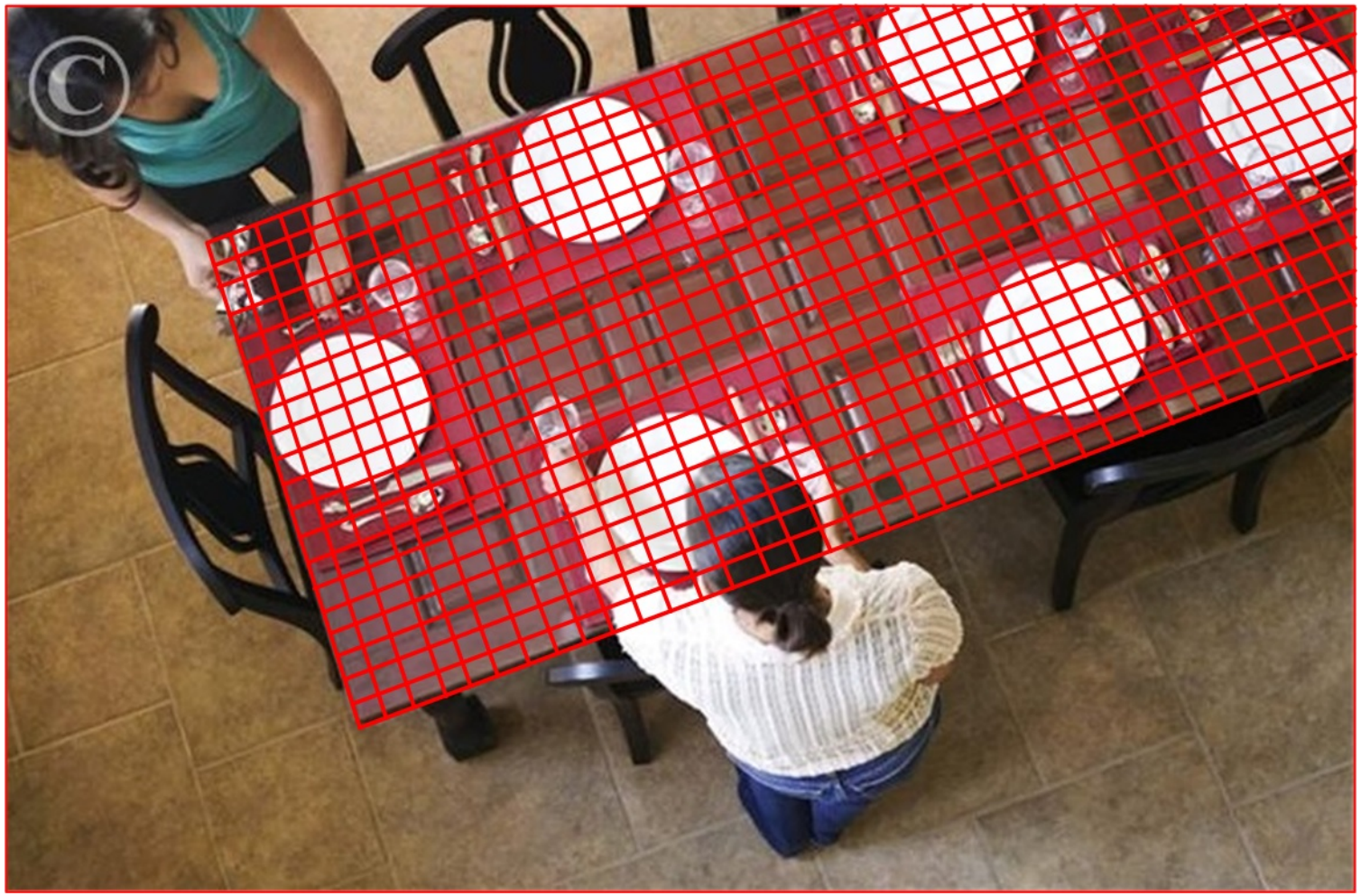
Scene Distribution

- $\mathcal{M}_0(T)$: A partition of the table into 5cm $\times$ 5cm "cells".
- $Z = \{Z_{c,m}, c \in \mathcal{C}, m \in \mathcal{M}_0(T)$: binary variables indicating a the presence of at least one instance of category c centered in m.
- Scale is determined by the table dimensions (for our categories).
- $P(Z|T)$: A Gibbs distribution

$$P_\lambda(Z|T) \propto \sum_i \lambda_i g_i(Z)$$

where g_i is distinguished set of "features".

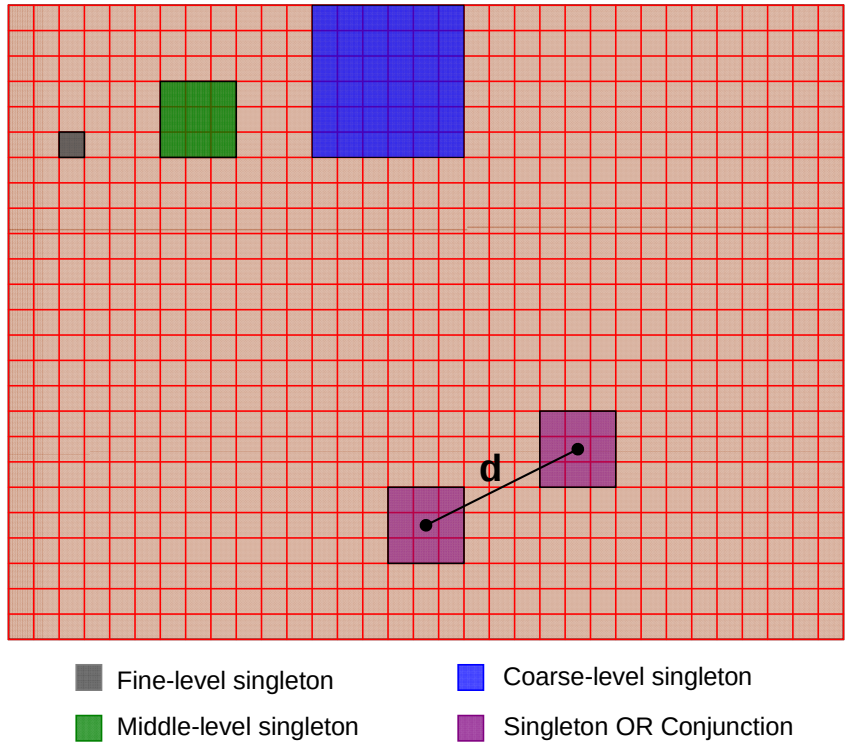
Table Mesh



FEATURES

- The set of features includes each $Z_{c,m}$.
- It also includes coarser locators $Z_{c,m}$ which are indexed by two coarser partitions $m \in M_1(T) \cup M_2(T)$ using $15cm \times 15cm$ cells and $45cm \times 45cm$ cells, repectively.
- Notice that for $m \in M_1(T)$ we have $Z_{c,m} = \max_{m' \in M_0(T), m' \subset m} Z_{c,m'}$ and similarly for $Z_{c,m}$, $m \in M_2(T)$.
- In addition, there are conjunction features $g(z) = Z_{c,m} Z_{c',m'}$, $m, m' \in M_1(T)$, for "nearby" m, m'.
- Such features allow for inhibiting nearby co-occurrences of some categories (e.g., two plates), or of two different categories, and promoting co-occurences other categories

FEATURES (CONT)



- The singleton features accommodate the overall empirical statistics for localized object instances.
- The conjunction feature functions incorporate contextual relations between different object categories.

JHU Table-setting Dataset

LEARNING

- We exploit symmetry in table-settings to reduce the number of parameters, for instance, given T, group features whose weights λ are expected to be the same.
- Still, estimating the (surviving) λ is *difficult* due to the large number of parameters and relatively small number of annotated table settings.
- To overcome this, we learn the parameters from *synthetic data* - samples from a high-resolution, generative attribued graph model in the world coordinate system.
- The model has many fewer, and far more interpretable, parameters and can be efficiently learned from limited number of manually annotated images.

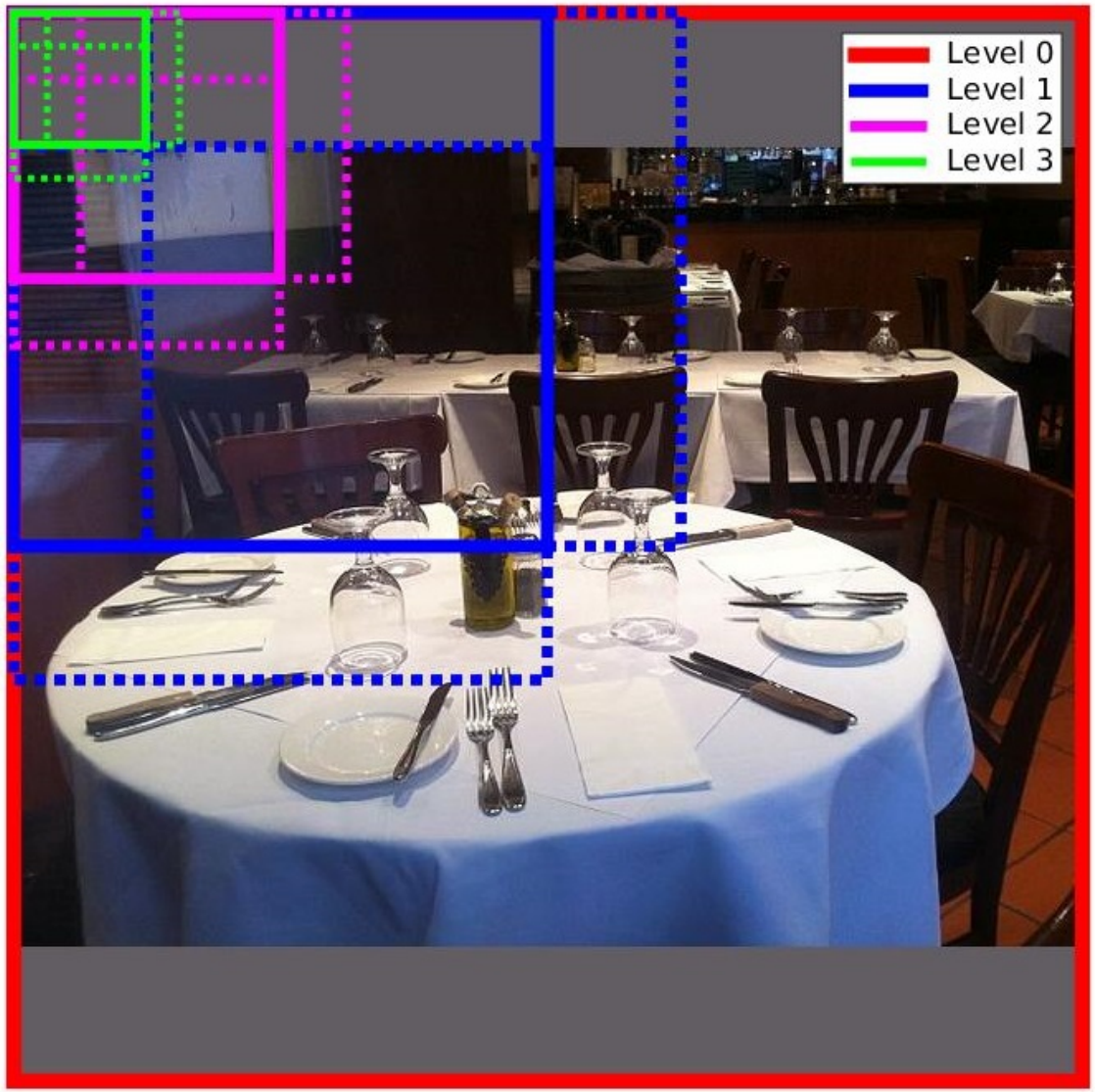
Learning (cont)

- Essentially a multi-type branching process with vertices attributed by category and 3D pose.
- Direct (unconditional) sampling is trivial, and we can generate an arbitrarily large synthetic training set for estimating $p_\lambda(z)$.
- We learned 10 models $P(z|T)$ for 10 different table sizes using stochastic gradient descent, iteratively minimizing the KL divergence between the Gibbs and empirical distribution.

Existence Annobits

- Objects are attributed a category and an apparent pose in the image coordinate system, taken here as the location of the center and the size (e.g., diameter).
- The pose space is then $\mathcal{D} \times (0, +\infty)$ where $\mathcal{D}$ is the image domain normalized to $D = [0, 1]^2$.
- *Object instance:* (c, x, s) with $c \in \mathcal{C}$, $x \in D$ and $s > 0$.
- $\mathcal{A}$: finite set of "windows" or "annocells" $W \subset D$ arranged in a 4-level hierarchy of varying sizes; $|\mathcal{A}| = 1036$.
- $\mathcal{J}$: finite set of size intervals J.
- $Y_{c,W,J} = 1$ if an instance from c with size in J is visible in W. (Write $Y_{c,W}$ if $J = (0, +\infty)$.)

ANNOCELL HIERARCHY



Derived Annobits

- Can be defined to match the type of classifiers available.
- We use the collection

$$Y_W = (Y_{c,W}, c \in \mathcal{C}).$$

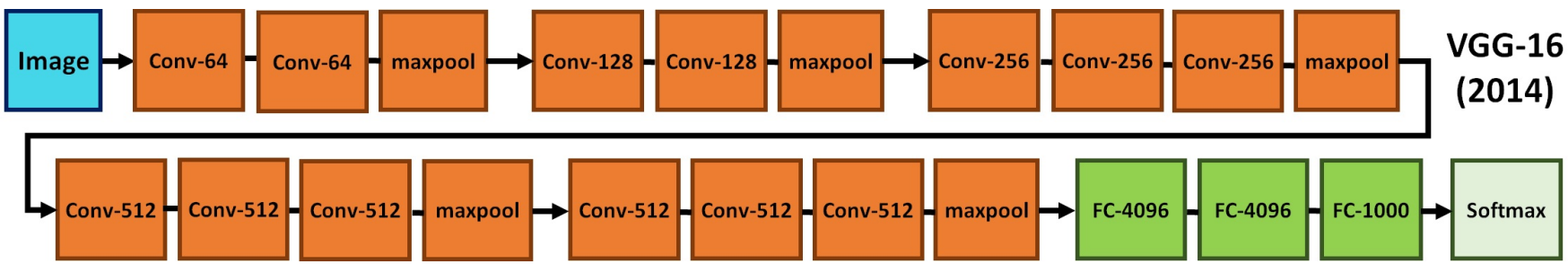
- We also use $Y^s_{W,J}$, a binary variable indicating whether the average size of the objects present in W belongs to J, and define $Y^t_W = 1$ if W is part of the table and $Y^t_W = 0$ otherwise.

CLASSIFIERS

- Variables X_W, $W \in \mathcal{A}$, predict Y_W by providing "weights" on categories $c \in \mathcal{C}$ and "background."
- Variables X^s_W, $W \in \mathcal{A}$, provide "weights" on $\mathcal{J}$.
- Variables X^t_W, $W \in \mathcal{A}'$ (where $\mathcal{A}'$ is a subset of $\mathcal{A}$) to predict Y^t_W.

CNN CLASSIFIERS

- We trained (the last layers of) three deep CNNs, all based on the VGG-16 network (up to layer 15):
  - CatNet: for category classification,
  - ScaleNet: to estimate the scale of detected object instances,
  - TableNet: to detect the table surface area in a given image.
- The CatNet is a CNN with a 5-way softmax output layer used to predict Y^W.



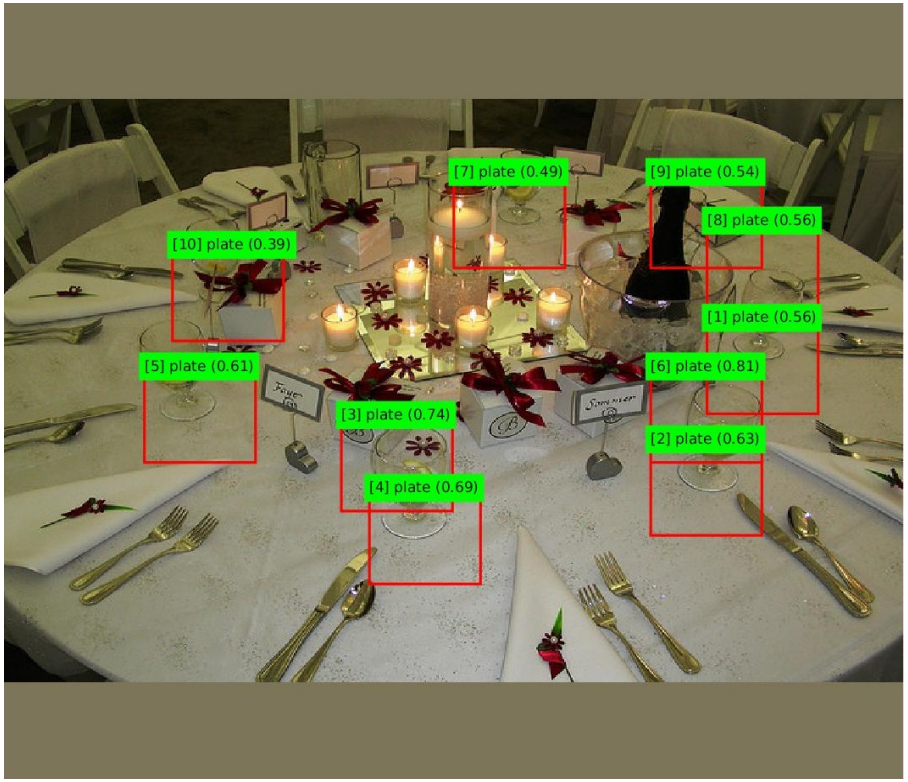
CatNet Training

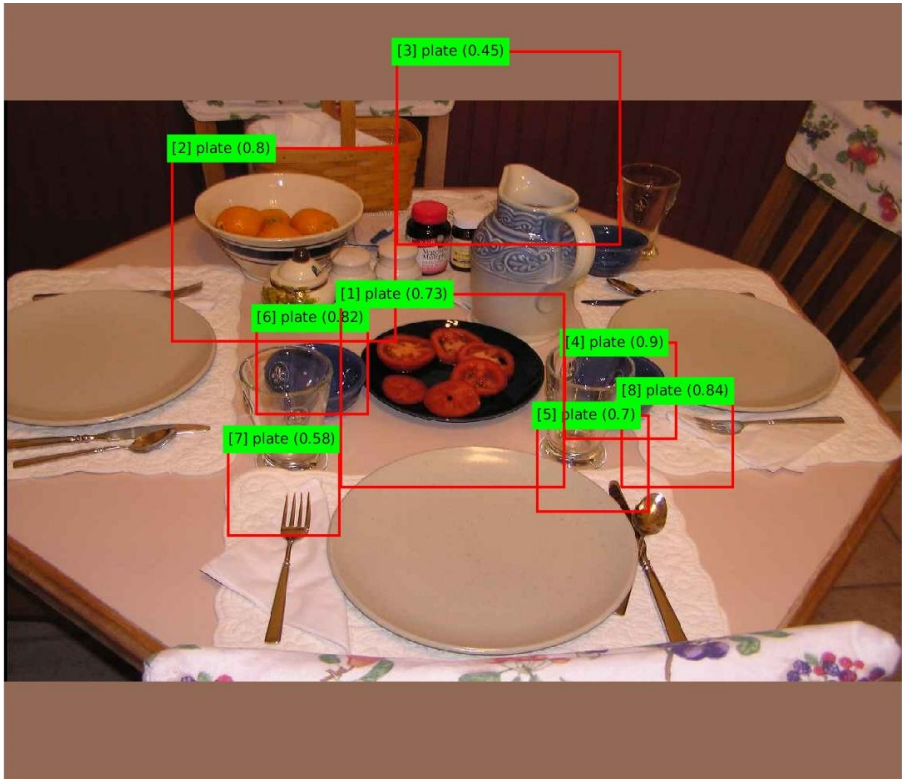
- A patch including multiple object instances appears multiple times in the training set, each time with the category label of one of the existing instances.
- The CatNet was trained by minimizing the cross-entropy loss function using stochastic gradient descent.
- Training took about 24 hours when the first 15 weight layers were initializing by the first 15 weight layers from the VGG-16 network.

ScaleNet and TableNet

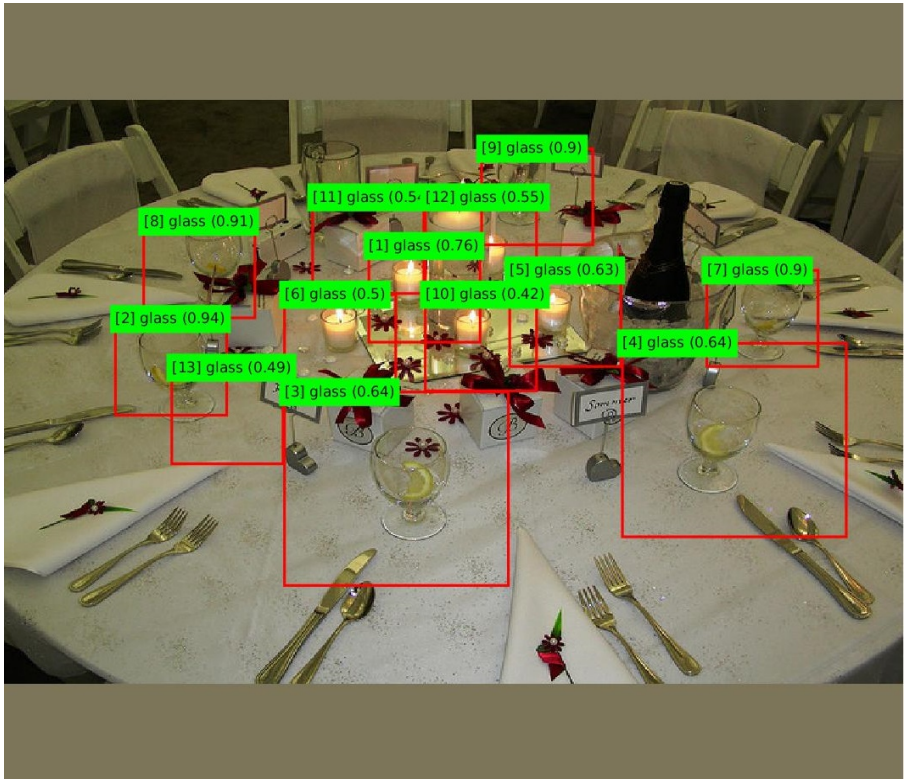
- ScaleNet estimates the ratio of average object scale (in pixels) to the size of the input patch, which stays unchanged after resizing the original input to 224×224.
- TableNet is a CNN trained to label a patch A as "part of a table setting" or "not part of a table setting." Ouputs used to estimate the boundary of the table.

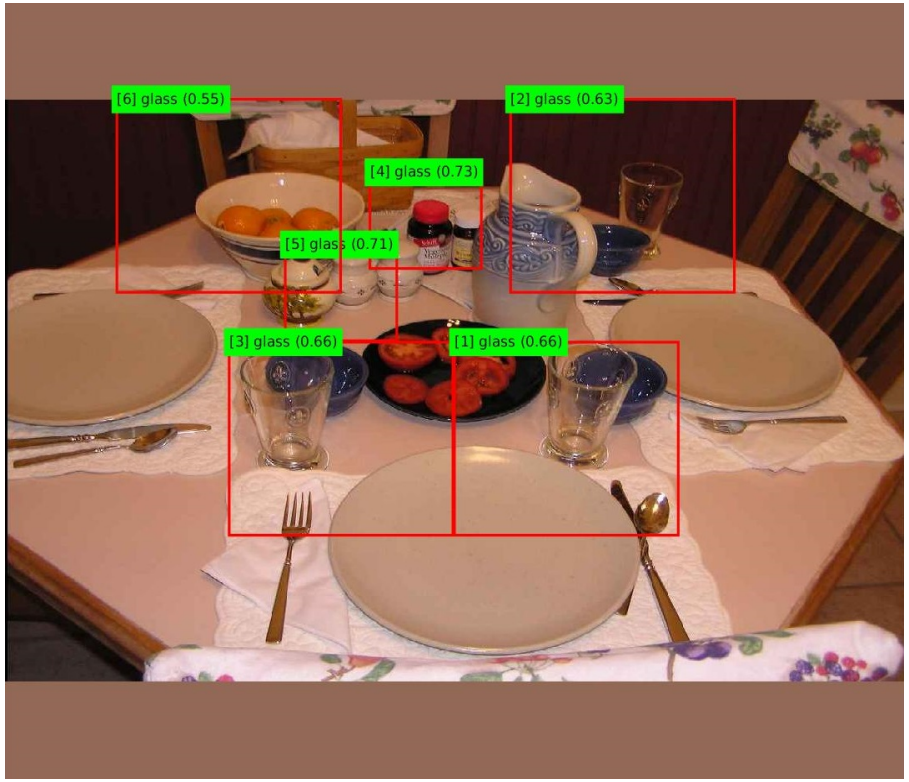
POOR "PLATE" DETECTIONS BY CNNS



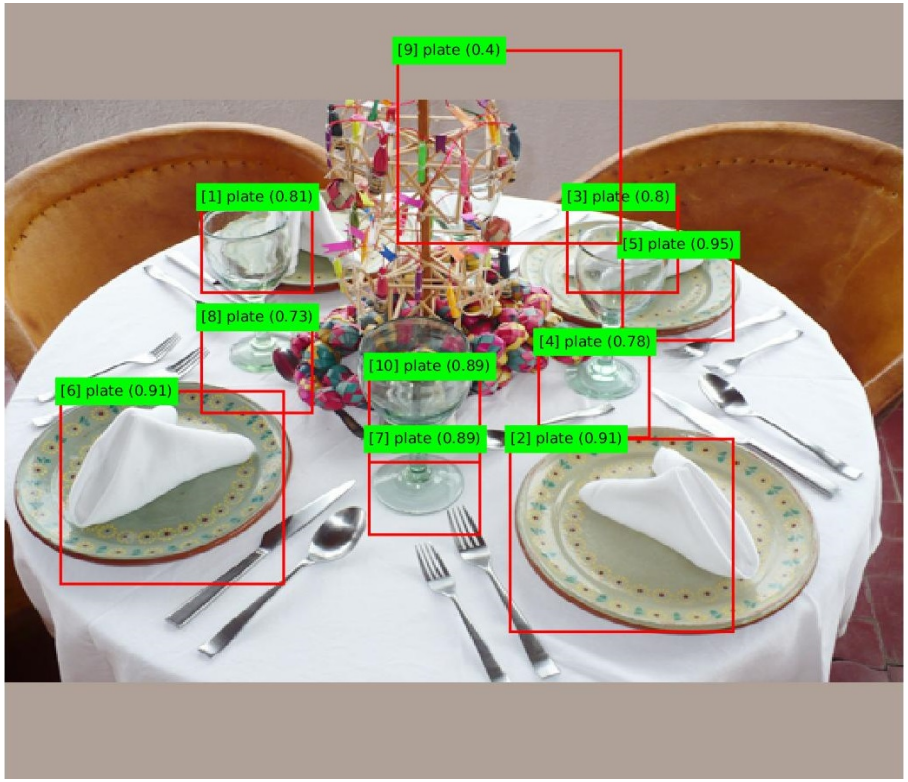


“GLASS” DETECTIONS BY CNNS





Contextually Inconsistent Detections



CNN DETECTION EXAMPLES

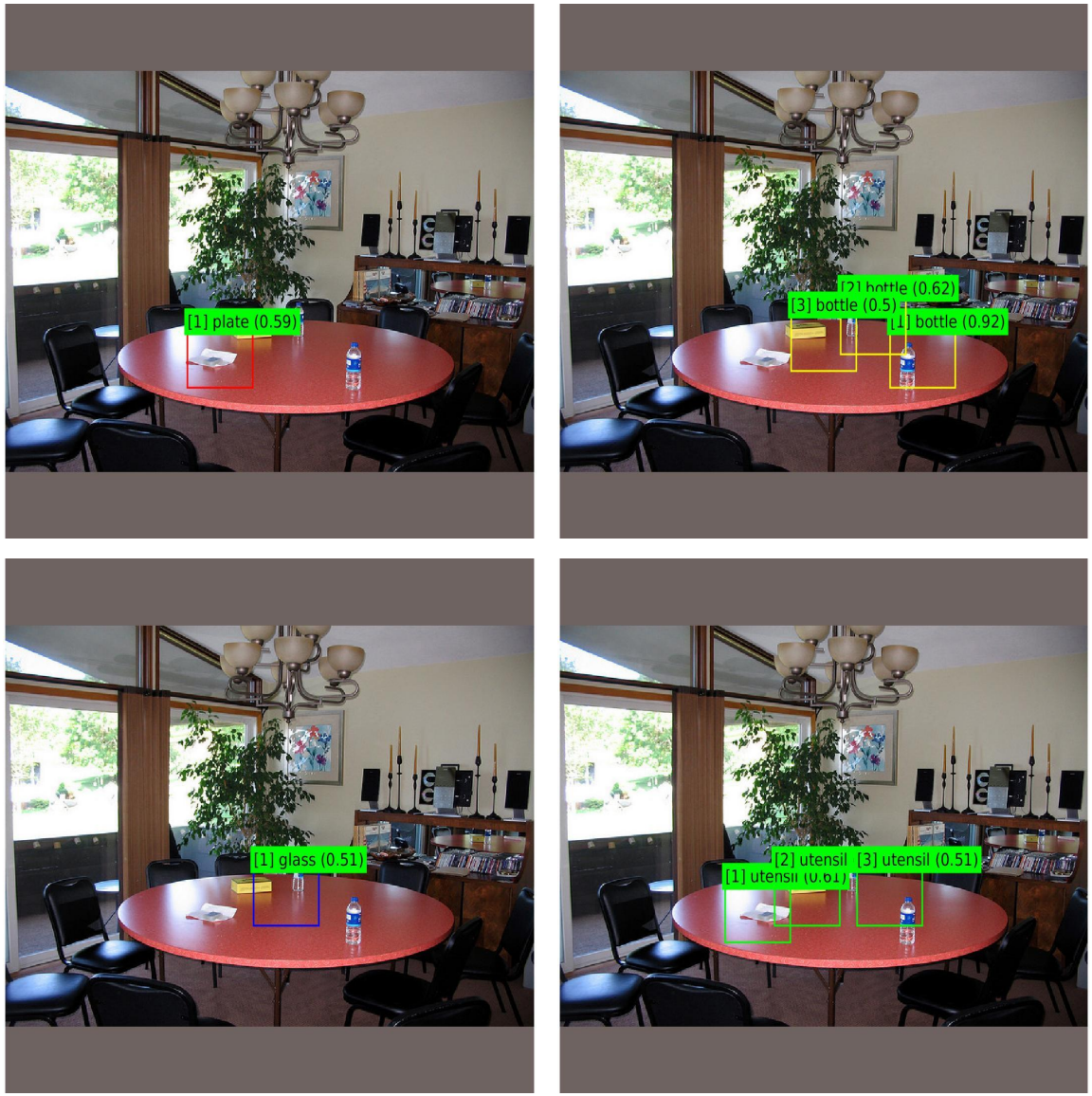
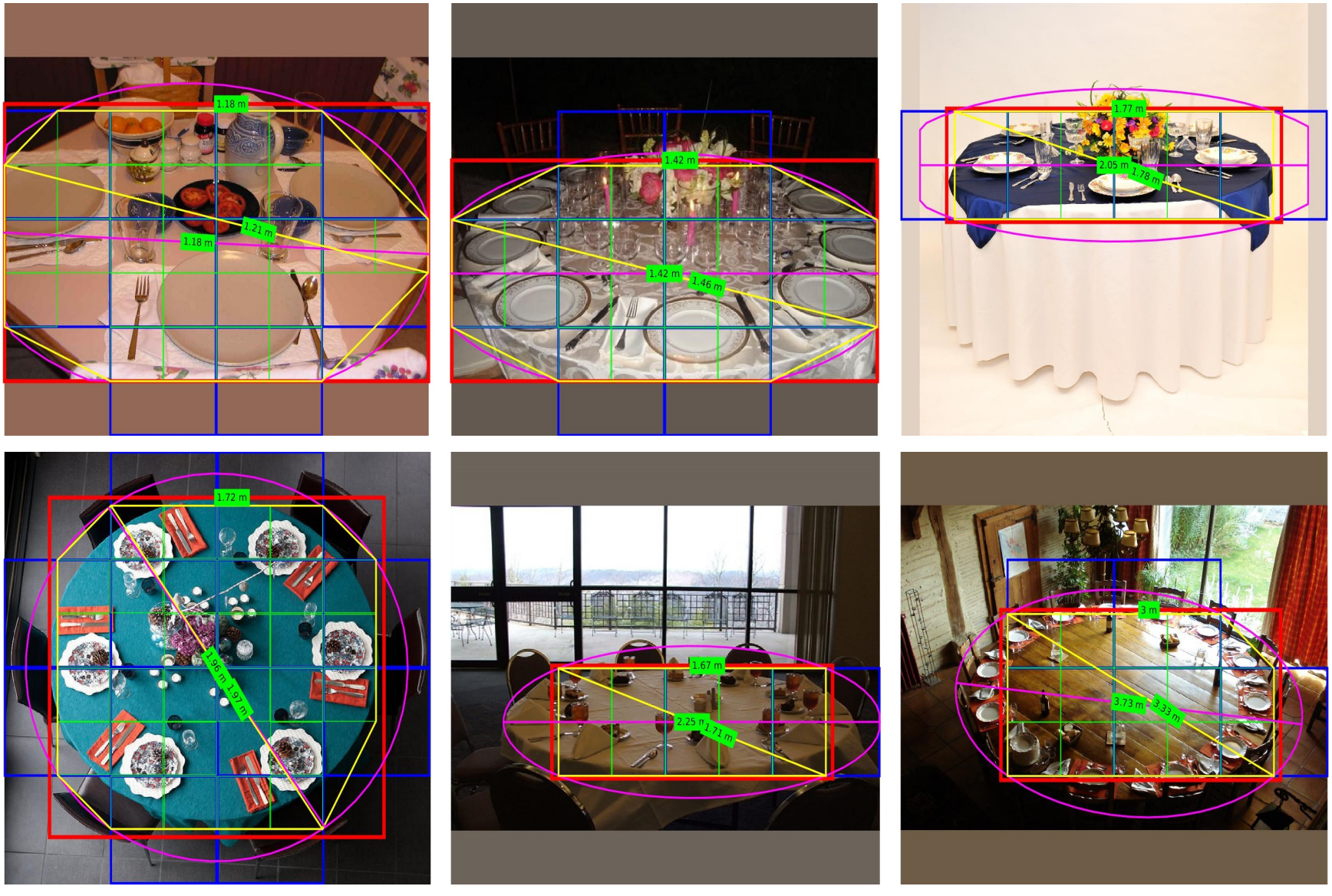


TABLE DETECTION BY TABLENET



DIRICHLET DATA MODEL FOR CATNET

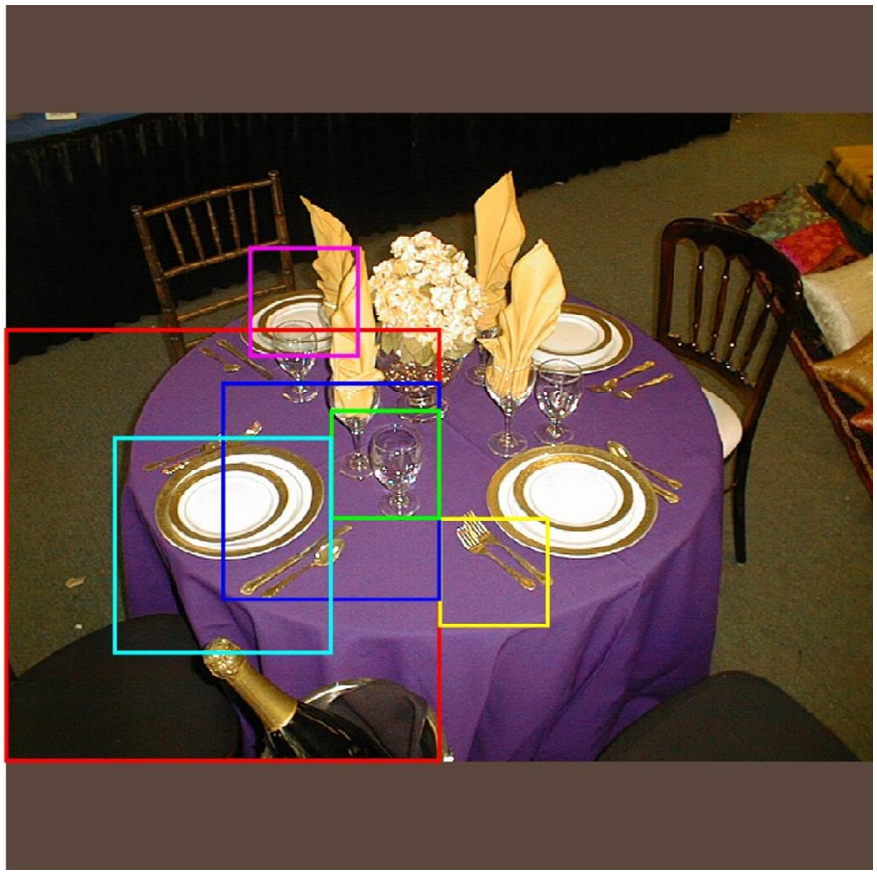
- The CNN classifier X_W predicts which object categories have instances inside W, returning a probability vector of dimension $K = 1 + |\mathcal{C}|$, the extra dimension for "none".
- We model the conditional distribution of X_W given the annobits $Y_W = (Y_{c,W}, c \in \mathcal{C})$ is as a K-dimensional Dirichlet distribution.
- We learned 16 conditional CatNet data models (MLE) for the 16 possible subsets of four object categories.
- The training data are obtained by running the CNNs on patches with matching configuration.
- Similarly for ScaleNet.

Maximum Likelihood Estimates

Category	Ground Truth	MLE
Plate	✓	✓
Bottle	×	×
Glass	✓	✓
Utensil	✓	✓

Category	Ground Truth	MLE
Plate	×	×
Bottle	×	×
Glass	✓	✓
Utensil	✓	✓

Category	Ground Truth	MLE
Plate	✓	✓
Bottle	×	×
Glass	×	×
Utensil	×	✓



Category	Ground Truth	MLE
Plate	✓	✓
Bottle	×	×
Glass	×	×
Utensil	×	×

Category	Ground Truth	MLE
Plate	×	×
Bottle	×	×
Glass	✓	✓
Utensil	×	×

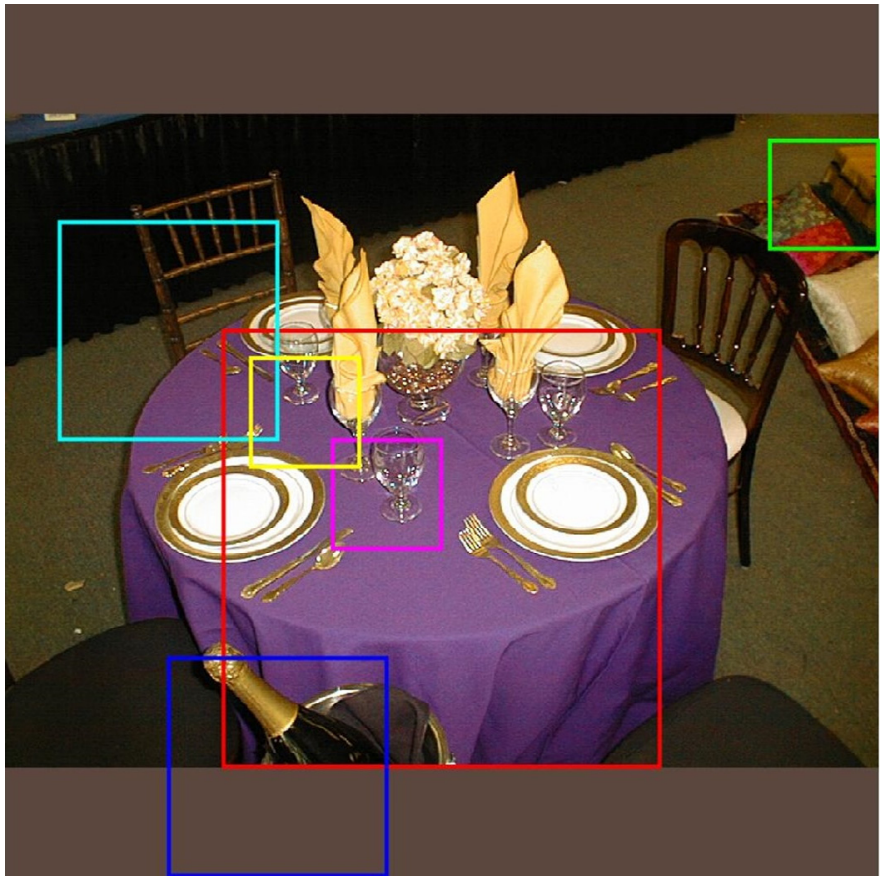
Category	Ground Truth	MLE
Plate	×	×
Bottle	×	×
Glass	×	×
Utensil	✓	✓

Maximum Likelihood Estimates (cont)

Category	Ground Truth	MLE
Plate	✓	×
Bottle	×	×
Glass	✓	✓
Utensil	✓	✓

Category	Ground Truth	MLE
Plate	×	×
Bottle	×	×
Glass	×	×
Utensil	×	✓

Category	Ground Truth	MLE
Plate	×	×
Bottle	×	×
Glass	×	×
Utensil	✓	✓



Category	Ground Truth	MLE
Plate	×	✓
Bottle	×	×
Glass	×	✓
Utensil	×	×

Category	Ground Truth	MLE
Plate	×	×
Bottle	×	×
Glass	×	×
Utensil	×	×

Category	Ground Truth	MLE
Plate	×	×
Bottle	×	×
Glass	×	×
Utensil	×	✓

POSTERIOR SAMPLING

- Posterior sampling was carried out in three nested loops corresponding to factoring the posterior at step k:

$$P(Z, T, H|\mathbf{e}_k) = P(T|\mathbf{e}_k)P(H|T, \mathbf{e}_k)P(Z|T, H, \mathbf{e}_k).$$

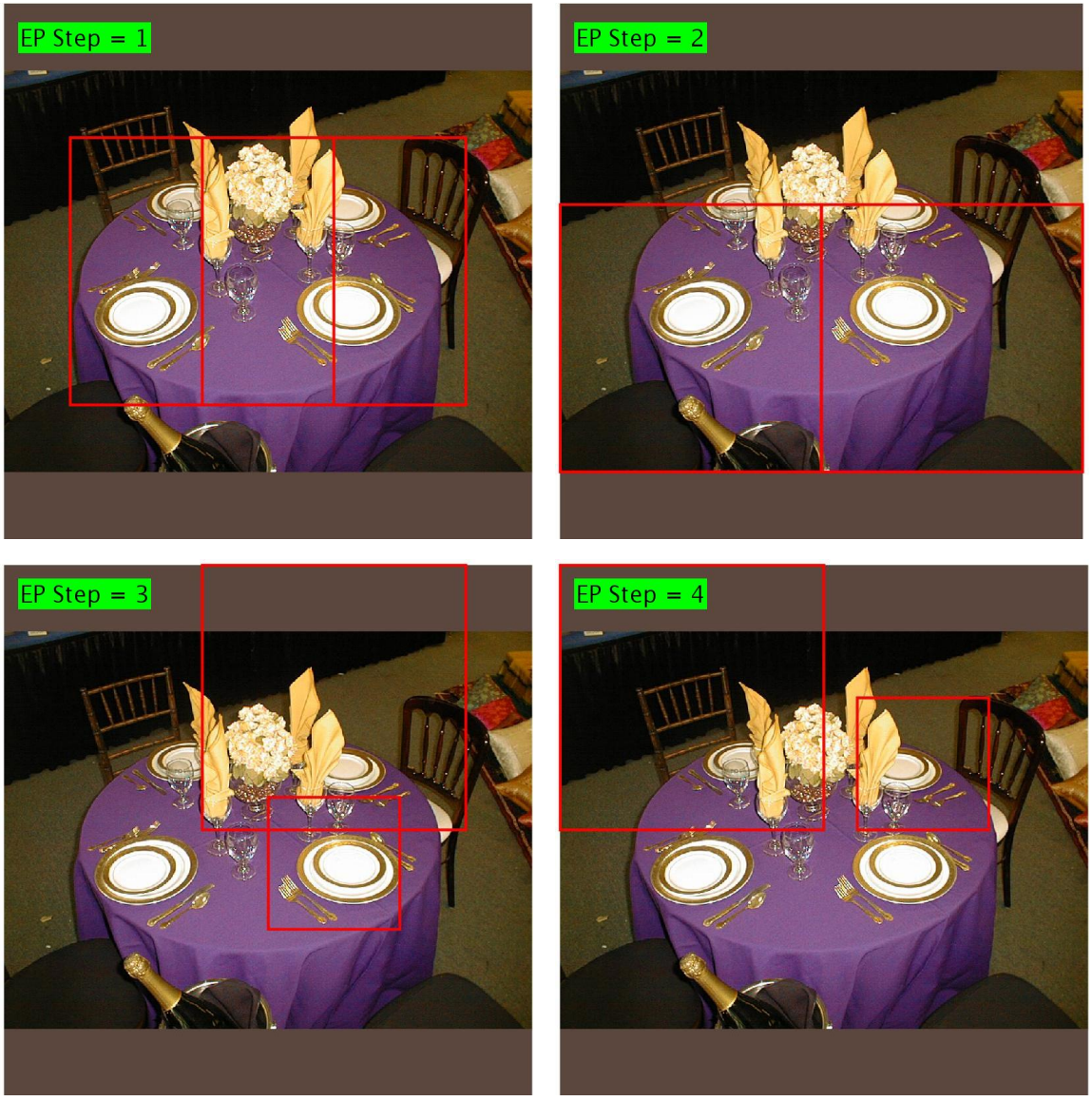
  - Outer Loop: sampling table size (Metropolis-Hastings)
  - Middle Loop: sampling homography (Metropolis-Hastings)
  - Inner Loop: sampling MRF model (Gibbs sampling)
- Given posterior samples of (Z, H), directly obtain posterior samples of $Y_{\mathcal{Q}}$, and hence can estimate $H(Y_q|\mathbf{e}_k)$ for all new q.

FULL POSTERIOR DETECTIONS

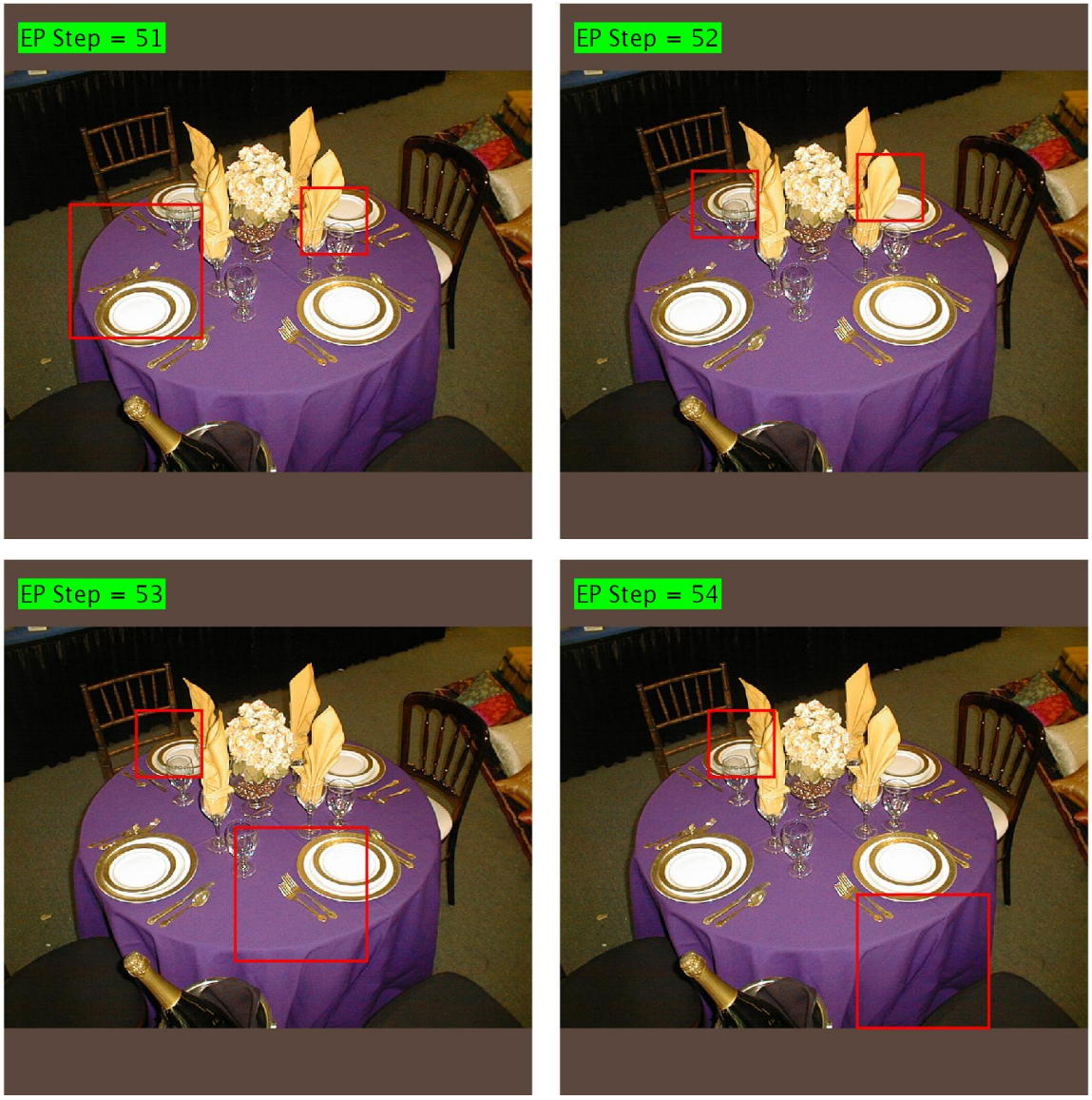
EP DETECTIONS (STEP 40)

CNN DETECTIONS

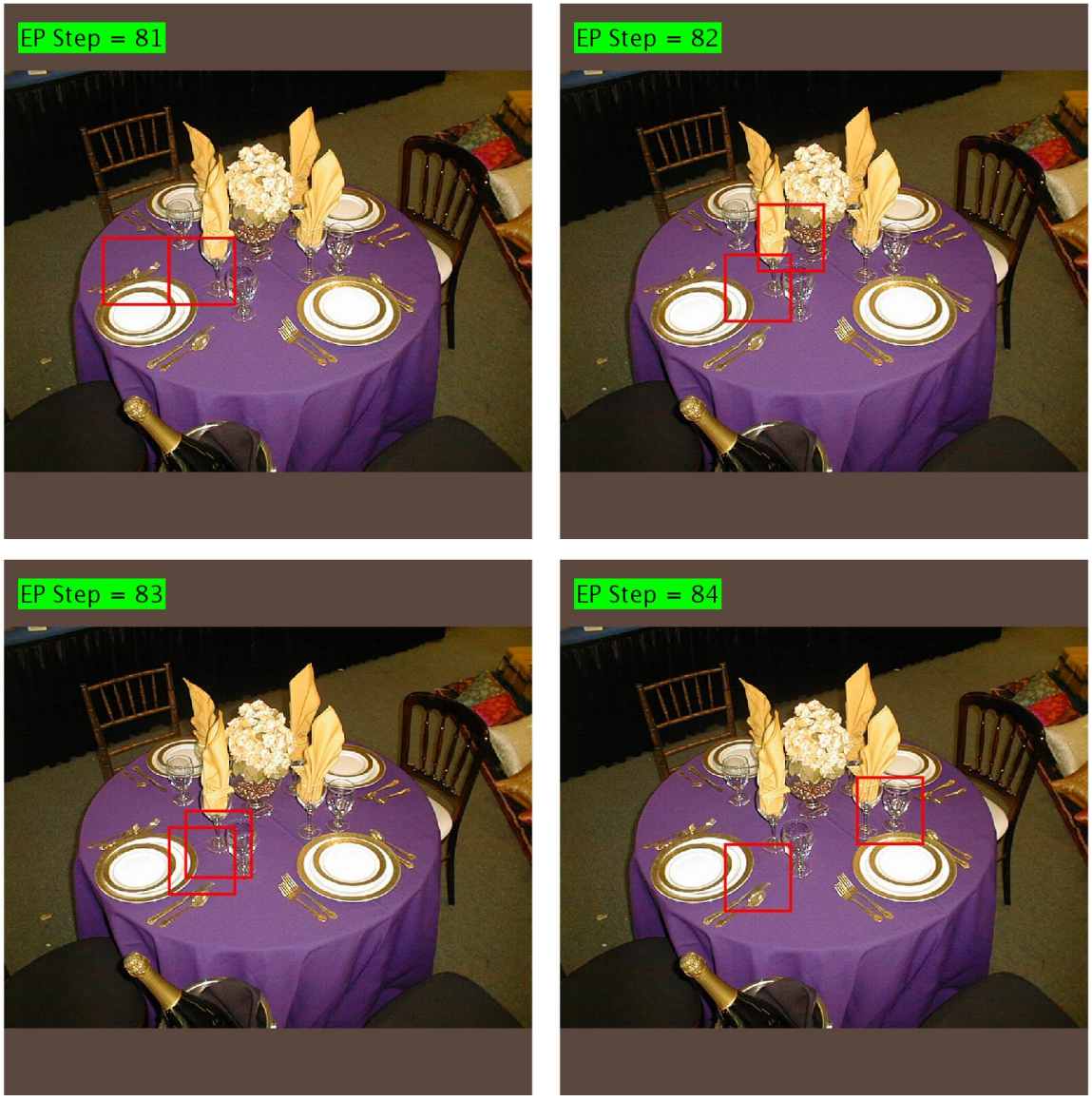
EP QUESTIONS (STEPS 1-4)



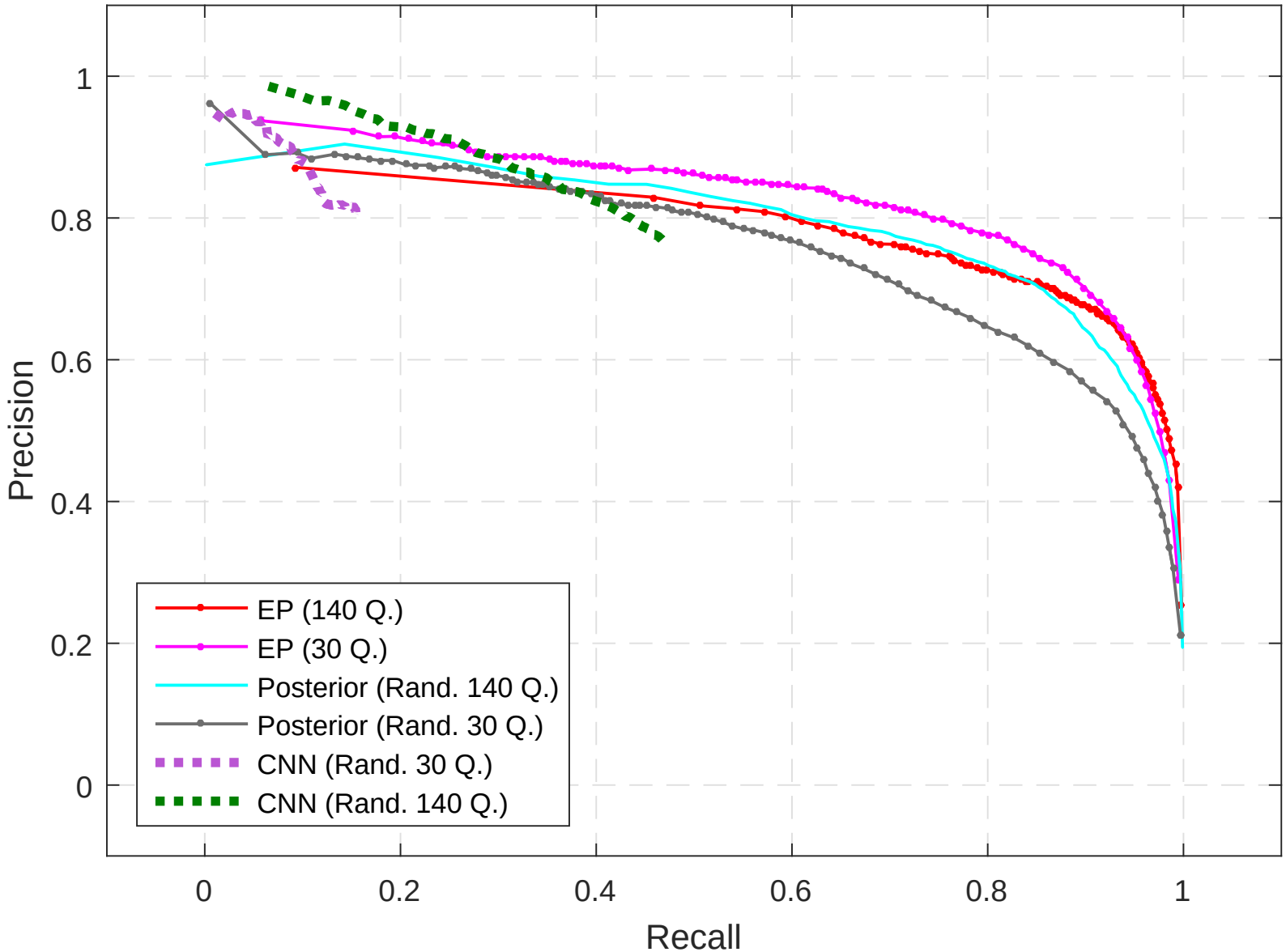
EP QUESTIONS (STEPS 51-54)



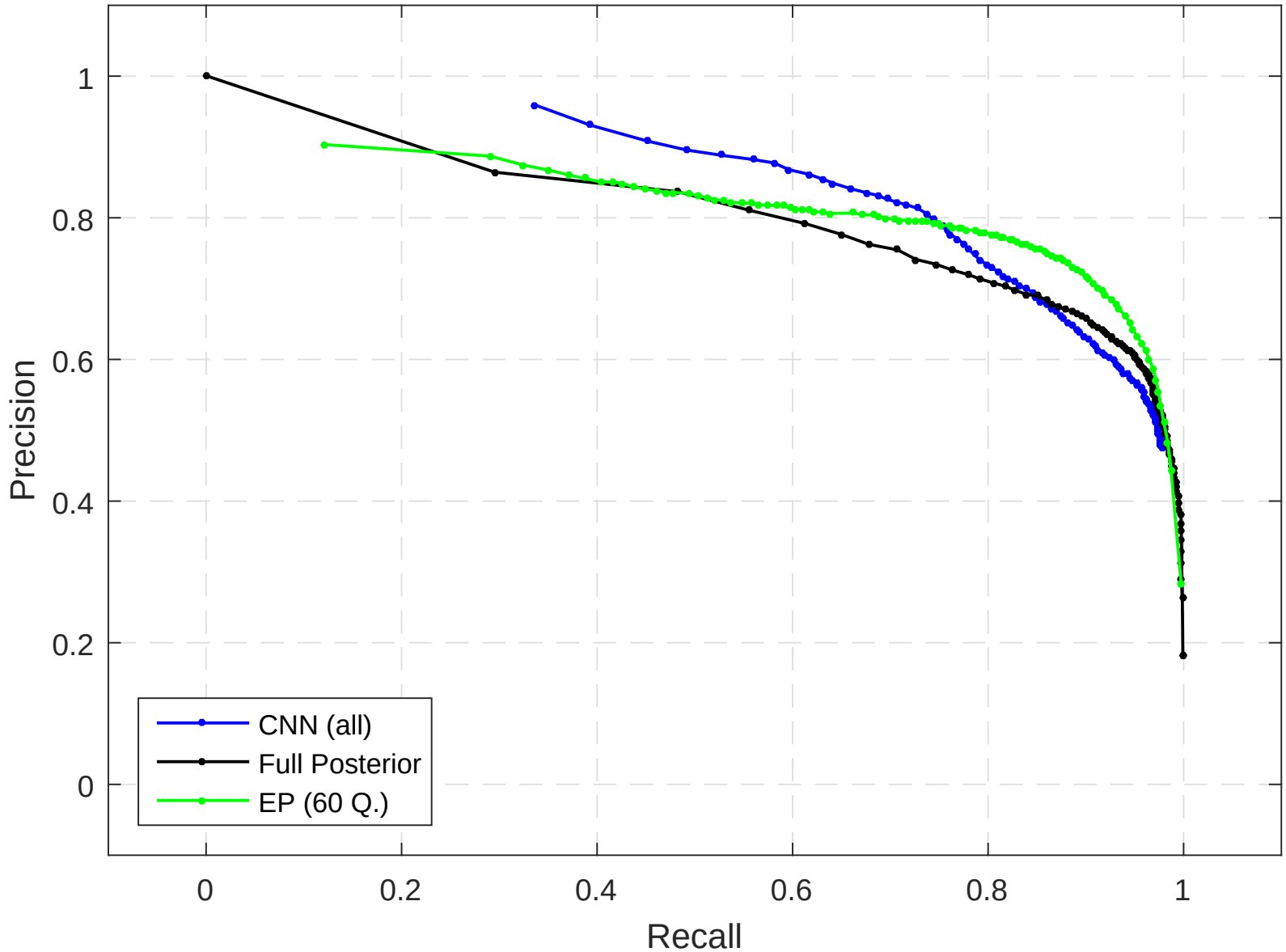
EP QUESTIONS (STEPS 81-84)



Precision-Recall Curves



Precision-Recall Curves



Concluding Remarks

- Assets:
  - *Coarse-to-fine search emerges naturally.*
  - *Ambiguities due to conflicting evidence sometimes resolved.*
  - *A fraction of the classifiers may be sufficient.*
- Liabilities:
  - *The treatment of context is limited. Compositional models do it the right way.*
  - *Many moving parts.*
  - *Replacing X_q by Y_q in query selection is unncessary and may degrade performance.*