Composite Statistical Learning and Inference for Semantic Segmentation

Fuxin Li
Georgia Institute of Technology
fli@cc.gatech.edu

Joao Carreira
University of Coimbra
joaoluis@isr.uc.pt

Guy Lebanon
Georgia Institute of Technology
lebanon@cc.gatech.edu

Cristian Sminchisescu
Lund University
cs@math.lth.se

June 23, 2014

Abstract

In this paper we present a learning and inference framework, Composite Statistical Learning and Inference (CSLI), for random fields with extremely high order interactions. Instead of conventional probabilistic approaches that build models on clique potentials, we propose to focus on subset statistics from overlapping random variable subsets and employ composite likelihood approaches for learning and inference. Focusing on subset statistics avoids the need to consider normalization constants in both learning and inference. Learning becomes a conventional (weighted) regression problem, and inference is also greatly simplified since the subset statistics are of much lower dimensionality than the initial random variables. We present an inference algorithm on the semantic segmentation problem in computer vision, where the statistic of choice is maximal class-specific overlap. Continuous parameters are defined on superpixels obtained by multiple intersections of segments, then the optimal segments are outputted from the inferred superpixel statistics. The algorithm is capable of recombine and refine initial mid-level segment proposals, as well as handle multiple interacting objects, even from the same class, using an EM algorithm maximizing the composite likelihood. In the PASCAL VOC segmentation challenge, the proposed approach obtains high accuracy and successfully handles images of complex object interactions.

1 Introduction

In machine learning and computer vision, many problems can be represented as random fields. A random field model is described by an undirected graph, where random variables are represented as nodes and dependencies among variables are represented by graph edges. The joint probability model is factorized over graph cliques, and potentials (under a Boltzmann distribution) are often defined on such cliques for parameter

learning and inference. Such models have achieved great success in the past twenty years in many different problems in computer vision, natural language processing and other domains.

In practice, most of the clique potentials considered in these models are unary (order 1) or pairwise (order 2). Higher order cliques are often ignored mainly because of the computational difficulties for using them in learning and inference. This raises a question mark for certain complex computer vision problems, such as scene understanding, where it is extremely important to consider higher order potentials because local dependencies are not sufficient to model the complex dependencies over large image regions.

A concrete example is the semantic segmentation problem [1, 14, 15, 17, 22, 23, 25, 34, 43], an important aspect of scene understanding. The goal of semantic segmentation is to detect objects from different categories and identify their spatial layout simultaneously. Each pixel in the image must be classified as a foreground object of a certain category, or be assigned as background. Suppose the labels of each pixel contain two categorical random variables, one indicating its category and another indicating the object it belongs to, the semantic segmentation problem can be defined on a random field where the nodes are the labels of each pixel, and links are specified by dependencies among image regions (pixel subsets). It can then be argued that higher order dependencies are necessary: the label of one pixel is not only dependent on its spatial neighbors, but also the labels of many pixels that are spatially far away from that pixel.

An empirical approach we have pursued with some success to solve this problem can be called 'sliding segments' [27, 5], starting from an unsupervised generation of many possibly conflicting holistic figure-ground segment proposals with object-sized spatial support [7] (Fig. 10). The segment proposals are then passed to classifiers or regressors that determine their category. Full image interpretations are in turn assembled sequentially from individual segments. Feature extraction on segments can better capture global dependencies such as object shape, object-level color and texture distributions, leading to more accurate classifications than earlier local classifiers focusing on patches around pixels [36]. This type of approach has been shown to deliver top performances in difficult benchmarks [27, 5] and is a backbone of most state-of-the-art systems on this problem [44, 28, 42].

The use of predictions on mutually overlapping segments requires a new learning and inference framework for graphical models, that is inherently based on large correlated variable subsets instead of local unary/pairwise connections. In order to do that, we need to translate the insights observed from the empirical problem into statistical language. Note that *segments* can be regarded as *random variable subsets* by noting that each image segment is a subset of image pixels, thus defines a subset of random variables in the aforementioned random field. The category label of the segment, or the spatial overlap of the segment with a ground truth object, can be considered as *subset statistics* that are higher order moments on the subset of random variables defined by the segment. Therefore, the learning and inference based on such a set of mutually overlapping segments is related to a classic approach, composite likelihood [31], that has not been widely used in computer vision and machine learning. Composite likelihood performs inference by modeling probabilities on many random variable subsets,

and is asymptotically consistent [10]. The difference between the aforementioned approach and composite likelihood is that instead of probability models, statistical estimates are employed. A (set of) statistics are usually of much lower dimensionality than the probability distribution defined on many variables, thus one can expect that the ensuing probability models would be less complicated, which would in turn make learning and inference both easier.

The main focus of this paper is to formalize and develop such a statistical framework. We name our framework *Composite Statistical Learning and Inference* (CSLI), which consists of a learning stage (Composite Statistical Learning, CSL) and an inference stage (Composite Statistical Inference, CSI). During CSL, statistics are defined on variable subsets sampled from the σ-algebra of a probability measure, and predictors of subset statistics are learned from observations, using standard techniques such as classification or regression. During CSI, variable subsets are sampled, learned predictors are applied on these subsets to predict statistics, and then inference is performed to uncover the MAP/MLE solution of each random variable.

CSLI handles higher order learning and inference in a clear and tractable manner. A very desirable feature of it is the simple learning stage. Most of the time, learning is performed just as simple classification/regression problems which have been extremely well-studied. Convex solutions often exist and generalization bounds from the training set to testing data have been well-established. During the inference stage, the dimensionality reduction effect by moving from probability to statistical estimates make the problem tractable. In semantic segmentation, the low dimensionality enables us to use exploratory data analysis [37] to explicitly specify a parametric error distribution of the statistics, as well as proposing a simple convex relaxation for the inference algorithm.

The drawback of the framework lies in the sampling of variable subsets. Since variable subsets are overlapping, it is impossible to assume that the variable subsets are sampled i.i.d.. Thus it becomes difficult to develop theoretical results on learning rates. Besides, in practice one needs to control the weights of the sampled variable subsets to alleviate biased sampling. We will also propose a practical method to deal with this in semantic segmentation.

The rest of the paper is organized as follows: in section 2, we review previous literature on higher order graphical model inference and semantic segmentation. In section 3, we propose the CSLI framework and prove theoretical consistency results. In section 4, we develop a CSI algorithm for the spatial inference in semantic segmentation. In section 5, implementation details are presented, as well as the technique for alleviating biased sampling. Results on the popular MSRC-21 and PASCAL VOC datasets are shown in section 6, and the conclusion is presented in section 7.

2 Related Work

2.1 Higher Order CRF Inference

The inference part of our approach can be viewed as a higher-order CRF inference method, albeit very different from traditional ones that work directly with explicit clique potentials for the conditional distribution. A well-known earlier work by Kohli

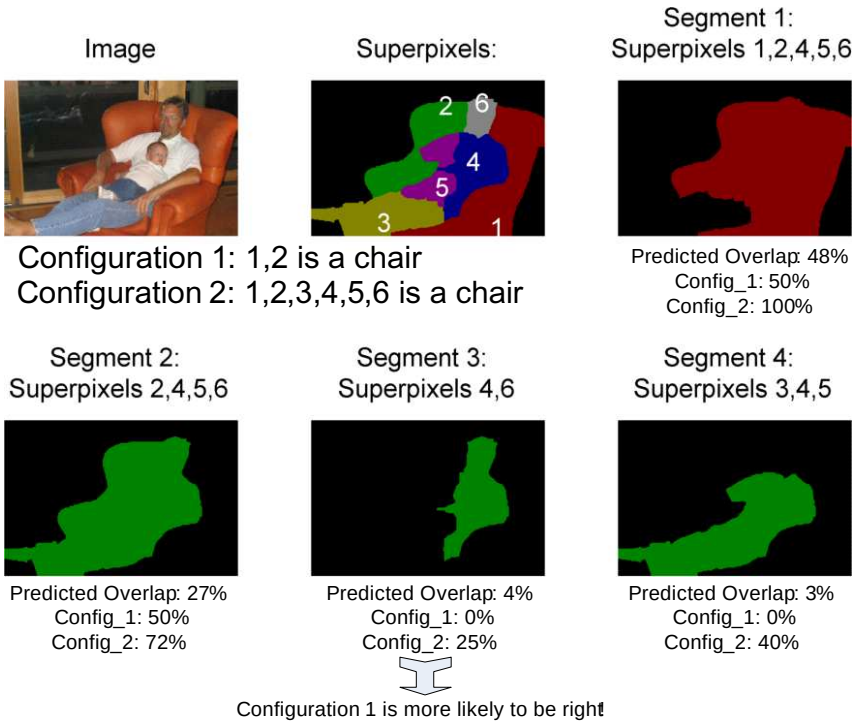

Figure 1: (Best viewed in color) The goal of our inference can be intuitively thought as finding the superpixel configuration which best explains most of the predicted segment statistics, here spatial overlap with the chair object. This formulation allows discovering objects that are cut into disconnected components, such as the chair. Instead of find such a superpixel configuration using a search algorithm, we formulate it as a continuous maximum composite likelihood problem with a convex relaxation, where a near-optimal solution can be found via mathematical optimization.

et al. [20] proposes the robust P^n potential where the potential is defined on a set of superpixels, and the penalty is linearly increased if more superpixels are different from the predicted label, the confidence of the prediction controls the rigidity (slope) of the potential. This is very different from our approach where region statistics are predicted and the loss function is determined by the predicted region statistics. Especially, in our approach if the spatial overlap between the segment and the ground truth category is predicted to be 75%, then no loss will be incurred if there are exactly 75% of the pixels in the segment assigned to the ground truth category. In the case of [20], there will still be a loss incurred corresponding to the 25% region, and hence our approach is more suitable with segments that only partially overlap the object. Kohli and Kumar [19] proposes to represent higher-order potentials using lower envelopes of linear functions, which unfortunately also cannot represent the aforementioned phenomena. The marginal probability field approach [41] attempted to solve this problem by incor-

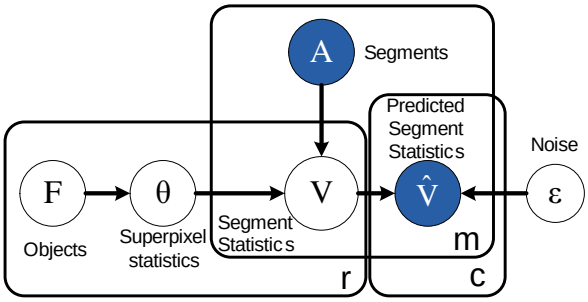

Figure 2: The conceptual graphical model. Superpixel statistics are generated from the ground truth objects. Segment statistics are generated from superpixel statistics and the segments. The observations are predicted segment statistics on each category. They are the maximal segment statistic for all ground truth objects in the same category, perturbed with noise ϵ. During inference, we first solve for the superpixel statistics θ, then output full object segmentations given θ.

porating dependent statistics estimates, however it does not include our idea of making these statistics based on overlapping regions (segments).

(less relevant) Other higher-order inference models involve ones based on patterns [35, 21], which is different from the current ones that do not form a specific pattern. Komodakis and Paragios [21] proposed to use dual decomposition for optimizing higher-order potentials, where each clique turns into one subproblem. Lempitsky et al. [26] includes higher-order potentials of bounding box priors. Label counts were used in [40, 30].

2.2 Composite Likelihood

The composite likelihood method has been proposed as a generalization of the pseudo-likelihood in the 1980s [31], [38] summarizes the earlier work in this direction. Dillon and Lebanon [10] proposes a stochastic version that involves random draws on the variable subsets instead of fixed ones, and proved consistency of this version.

On the theoretical side, Liang and Jordan proposed an asymptotic analysis of certain pseudo likelihood/composite likelihood methods [29]. But what they call as composite likelihood is a model that always model the probability of all the variables, but condition on a subset of them. This is significantly different from our settings. Bradley and Guestrin has developed PAC bounds for learning CRF using composite likelihood [3], and claimed that the only PAC-learning methods for CRF can be re-casted into pseudo-likelihood approach.

2.3 Semantic Segmentation

The common approach to use CRF in semantic segmentation is as a hierarchical model [24, 25, 45] , where the higher-order potentials are decomposed into pixel/superpixel level layers and segment-level layers, with different layers connected based on overlap and compositionality. However, the interactions in these models are complex and involve

Figure 3: (Best viewed in color). The need for an efficient inference procedure given multiple object segmentation proposals. Overlap predictions between each segment and all categories are given as input to our algorithm - however only the category with maximal score is shown on the figure. Identifying the correct object layout from the overlapping segment predictions is a nontrivial task. Simply performing non-maximum suppression would discard all the `person` segments, which have lower scores because they all overlap the first `bike` segment.

different types of pairwise potentials (between pixels, between pixels and segments and between segments) which limits the range of potential functions for which tractable approximate inference is feasible. Another recently proposed variation using latent topics, the Pottics model [9], sidesteps the need for high-order cliques but still requires approximate inference.

This hypothesis-testing scheme is similar to the popular sliding window detection which test multiple rectangular window hypotheses [39, 8]. However, by exploiting low-level cues, segments align better with object boundaries. There are usually fewer candidate object regions than the set of all possible bounding boxes.

Conditional random field approaches have been proposed by modelling the probability of the labels conditioned on the image, so that local potentials can take full advantage of the entire image. However, the higher order interactions among subsets of labels are still important even within such a CRF framework. In practice models that incorporate higher order knowledge consistently outperform local approaches.

Other approaches search for configurations of non-overlapping segment hypotheses [15, 23] by using non-maxima suppression and maximum clique random field models [17]. They can be tractable since the decision space that has to be searched is limited to the initial segments (normally < 200 in practice). However, these are likely

to encounter difficulties when multiple objects touch or interact with each other. Examples are: people riding bicycles or horses, interacting with other people, sitting on sofas or chairs, etc. In such cases, segments often occlude and cut through each other and the initial mid-level proposals may not be entirely accurate. In such situations, a high-precision approach should be able to refine the initial hypotheses.

Non-probabilistic methods have also been developed to produce an average [16, 34] or weighted average [7, 27] of the predicted scores on each pixel/superpixel, then output the highest scoring labels. Arbelaez *et al*. learn to classify superpixels using class predictions from all enclosing segments as input features [1]. This strategy would typically allow for the refinement of a semantic segmentation, but in a heuristic manner, by *e.g*. thresholding pixel or superpixel scores.

3 Composite Statistical Learning and Inference

Throughout the paper we denote $p(x)$ the probability of random variable x, $\mathbb{I}$ the indicator function. $\mathcal{N}(x;\mu,\sigma^2)$ the density function of the normal distribution with mean μ and variance σ^2, $\tilde{\mathcal{N}}(x;\mu,\sigma^2)$ a truncated normal distribution to the domain $(0,1]$. $Ber(\alpha)$ a Bernoulli distribution with parameter α, $\mathrm{Exp}(x;\lambda)$ the density of an exponential distribution with parameter λ, and $\delta(x)$ the Dirac function. When x is a vector, $x \geq 0$ means that all dimensions of x are larger or equal to 0. For a set A, let $|A|$ denote its cardinality. A segment is considered a set whose cardinality is the area, i.e., the amount of pixels within the segment.

3.1 Pseudo Likelihood and Composite Likelihood

The composite likelihood approach that we are going to use is a generalization of the pseudo-likelihood approach [2]. In the pseudo-likelihood approach, for a random vector X, one maximizes the pseudo-likelihood

$$p(X) = \prod_i p(x_i|x_i) \tag{1}$$

instead of the conventional likelihood. This has been shown as asymptotically consistent but not as efficient as maximum likelihood.

A maximum composite likelihood (MCL) approach [31, 38] drops the independence assumptions typical in maximum likelihood. For us, this is important, in order to be able to leverage overlapping higher-order observations (on segments) that are strongly inter-dependent. We adopt a version in [10] with some simplifications.

Definition 1. *Suppose we have a dataset $D = \{X^{(1)},\ldots,X^{(n)}\}$, where each $X^{(i)}$ is a m-dimensional vector. Consider a finite sequence of variable subset pairs (called m-pairs) $(A_1,B_1),\ldots,(A_k,B_k)$, where $A_j,B_j \subset \{1,\ldots,m\},\forall j \in 1,\ldots,k$ with $A \neq \emptyset = A \cap B$. Given vector $\beta \geq 0$, the composite likelihood is*

$$cl(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{k} \beta_j \log p_\theta(X^{(i)}_{A_j}|X^{(i)}_{B_j}). \tag{2}$$

MCL is the approach to solve for θ by maximizing the composite likelihood (2). When β has stochastic components, this is called stochastic composite likelihood (SCL)[10]. The MCL/SCL approach is statistically consistent given an identifiability assumption:

Definition 2. *A sequence of m-pairs $(A_1, B_1), \ldots, (A_k, B_k)$ is identifiable of p_θ if the map $\{p_\theta(X_{A_j}|X_{B_j})\}$ is injective. In other words, there exists only a single collection of conditionals $\{p_\theta(X_{A_j}|X_{B_j})\}$ that does not contradict the joint $p_\theta(x)$.*

3.2 Composite Statistical Learning

Unlike pseudo-likelihood, MCL/SCL is intractable in most cases because of the need to model a high-dimensional distribution $p_\theta(X_{A_j}|X_{B_j})$, for which close forms exist for only very limited types of distributions. We propose to extend the MCL framework to distributions on statistical estimates. This makes us work with low-dimensional distributions which are much easier to model and estimate.

Definition 3. *With the same conditions as in Definition 1 for $D, X^{(i)}, A_j, B_j$ and β, let us further assume that $f(X^{(i)}, A_j, B_j)$ is an observed statistic from $X^{(i)}, A_j$ and B_j. We define the composite f-likelihood as $p_\theta(f(X^{(i)}), A_j, B_j)$, and composite statistical learning (CSL) problem as maximizing the composite f-likelihood*

$$\max_\theta \sum_{i=1}^{n} \sum_{j=1}^{k} \beta_j \log p_\theta(f(X^{(i)}, A_j, B_j)). \tag{3}$$

This new CSL problem recovers the model parameters θ from the composite f-likelihood $\log p_\theta(f(X^{(i)}, A_j, B_j))$ for all the random variables on multiple different subsets. It seeks to find a parameter vector θ that best explains all the observed statistics from $X^{(i)}$ and the two given subsets A_j and B_j. The distribution $p_\theta(f(X^{(i)}, A_j, B_j))$ is modeled as a 1-dimensional distribution. The identifiability given function f is similarly defined as:

Definition 4. *A sequence of m-pairs $(A_1, B_1), \ldots, (A_k, B_k)$ is f-identifiable of p_θ if the map $\{p_\theta(f(X, A_j, B_j))\}$ is injective. In other words, there exists only a single collection of conditionals $\{p_\theta(f(X, A_j, B_j))\}$ that does not contradict the joint distribution $p_\theta(X)$.*

An example of CSL is a linear subset regression model with Gaussian errors. Suppose $X^{(i)}$ is an image, A_j is a subset of its pixels (a segment) and B_j is a background segment non-overlapping with A_j. Then a fixed-length feature vector Z_{ij} can be extracted from each segment and the distribution of f_{ij} can be modeled as $p_\theta(f_{ij}) = \mathcal{N}(\theta^\top Z_{ij}, \sigma^2)$, with θ the regression weights. Given observed values of f_{ij} for many different $X^{(i)}, A_j$ and B_j, the CSL problem in this case becomes a weighted least squares regression of solving for θ.

Intuitively, as the number of observations goes to infinity, the true model parameters θ should give the best performance for each individual segment, hence converge to the optimal solution of the MCL problem (8), given a suitably chosen β vector. We proceed to prove the consistency of this newly-defined CSL problem, which mimics the proof in [10].

Theorem 1. *Let $\Omega \in \mathbb{R}^r$ be an open set, let $-C \leq f(x, A, B) \leq C$ be a bounded function, $p_\theta(f(x, A, B)) > 0$ and continuous and smooth in Ω, $(A_1, B_1), \ldots, (A_k, B_k)$ be a sequence of m-pairs which ensures f-identifiability. Then the sequence of maximizers of the CSL problem (8) is strongly consistent, that is, suppose θ_0 is the true parameter value, we have*

$$p\left(\lim_{n \mapsto \infty} \hat{\theta}_n = \theta_0\right) = 1. \tag{4}$$

Proof. The CSL objective function can be modified slightly by a linear combination with a constant term:

$$cl'_f(\theta; n) = \sum_{i=1}^{n} \sum_{j=1}^{k} \beta_j \left(\log p_\theta(f(X^{(i)}, A_j, B_j)) - \log p_{\theta_0}(f(X^{(i)}, A_j, B_j))\right). \tag{5}$$

By the strong low of large numbers, the above expression converges as $n \mapsto \infty$ to its expectation

$$\mu(\theta) = -\sum_{j=1}^{k} \beta_j D(p_{\theta_0}(\mathbb{E}_X(f(X, A_j, B_j)))||p_\theta(\mathbb{E}_X(f(X, A_j, B_j)))) \tag{6}$$

We can always restrict ourselves to compact set $S : \{c_1\|\theta - \theta_0\| \leq c_2\}$ so that both $cl'_f(\theta; n)$ and $\mu(\theta)$ remain bounded, so that the conditions for uniform strong law of large numbers hold [13], which leads to:

$$P\{\lim_{n \mapsto \infty} \sup_{\theta \in S} |cl'_f(\theta; n) - \mu(\theta)| = 0\} = 1 \tag{7}$$

However the identifiability condition specifies that the KL-divergence term $\mu(\theta) = 0$ iff $\theta = \theta_0$. And since KL-divergence is non-negative and $\theta_0 \notin S$, $\sup_{\theta \in S} \mu(\theta) < 0$. Combine this with (7) there exists N so that $\sup_{X, \theta \in S, n > N} cl'_f(\theta; n) < 0$ with probability 1. However, since $cl'_f(\theta; n)$ can always be arbitrarily close to 0 when $\theta = \theta_0$, we have $\max_\theta cl'_f(\theta; n) \notin S$ when $n > N$. Since c_1, c_2 are chosen arbitrarily $\max_\theta cl'_f(\theta; n) \to \theta_0$ with probability 1. □

3.3 Composite Statistical Inference

Given learned parameters θ, it is often needed to find the data configuration that maximizes the likelihood or posterior of the composite f-likelihood (8), which is defined as the MLE or MAP composite statistical inference problem, depending on whether there is a prior or not.

Definition 5. *We define the composite statistical inference (CSI) problem as finding the input X that maximizes the composite f-likelihood*

$$\max_X \sum_{j=1}^{k} \beta_j \log p_\theta(f(X, A_j, B_j)). \tag{8}$$

Suppose in the CSL stage one has taken a linear subset regression model and has obtained regression weights as θ, then in the CSI stage one could compute $p_\theta(f(X, A_j, B_j))$ for every (X, A_j, B_j) by applying the regression weights and assuming that $p_\theta \sim \mathcal{N}(\theta^\top Z_j, \sigma^2)$. Then the problem becomes finding the X that are consistent with all the least squares predictions. Note the statistic $f(X, A_j, B_j)$ is assumed to be directly computable from X, A_j and B_j without θ. The intuition of the whole learning/inference process is as follows: in the CSL stage previous knowledge are encoded in the form of a predictor of certain statistics; during inference, $p_\theta(f(X), A_j, B_j)$ taps such previous knowledge to verify whether X is consistent in all the subset statistics.

A more practical example would be in the form of classification, where the knowledge could be, e.g., about whether the object is a `bottle`, then during learning such knowledge is encoded in a `bottle` classifier. And during inference, the pixels corresponding to the `bottle` should be collected so as to maximize the classifier output for `bottle`.

A major difference of inference and learning is that in inference one do not sum over all i.i.d. samples of X. Instead, one work on the same X and seeks to evaluate it using many different subsets. This makes the consistency harder as no i.i.d assumptions can be made and statistics on different overlapping subsets are inherently correlated to each other. The consistency of the inference procedure is still an open problem.

One certain thing though, is that CSI is usually not identifiable beyond the limit that the subsets specify. Suppose there exists two indices $i_1, i_2 \in \{1, \ldots, m\}$, so that $i_1 \in A_j$ iff $i_2 \in A_j$ and $i_1 \in B_j$ iff $i_2 \in B_j$, $\forall A_j, B_j$, then for many potential statistic f such as the mean or variance, $f(X)$ would not change if one swaps X_{i_1} with X_{i_2}. Therefore, we define the *identifiable partition* of CSI as:

Definition 6. *The identifiable partition is defined as the coarsest partition P of $\{1, \ldots, m\}$ so that for every set $P_i \in P$, $\forall A_j$, either $P_i \subset A_j$ or $P_i \cap A_j = \emptyset$. The same holds for B_j.*

If we sample enough subsets, we can identify all the dimensions. However normally with limited number of subsets we can only identify up to the level of the identifiable partition with the CSI approach. This makes CSI undesirable for certain text processing problems where identification needs to be at the per-word level. However, for image applications, per-pixel level accuracy is not always needed and CSI is well-suited for superpixel-level prediction problems that are abundant in computer vision.

Obviously, the inference is not usually as trivial as the learning. Choices of f and p_θ are crucially important to the success of the inference and will vary greatly from problem to problem. In the remaining of this paper, we show a successful application of this framework in the semantic segmentation problem and derive a detailed inference algorithm for it, in order to inspire future applications of the same framework.

4 CSI for Semantic Segmentation

In this section we present the main models of applying the proposed framework to semantic segmentation. We will first present the problem setup (Section 4.1), then the probabilistic model (Section 4.2), followed by the EM algorithm to estimate parameters

(Section 4.3). We must convert the per-category scores to per-object scores in order to properly maximize the likelihood. To do so, we need to estimate the number of objects in each category and assign the score of each segment belonging to a particular object. We postpone the relevant discussion to Section 4.4 because it uses the same probabilistic model and EM formulation introduced in Section 4.2 and 4.3. Implementation details are presented in Section 5. A discussion on how to output final segmentations given the superpixel statistics estimated from MCL is deferred to Section 6.

4.1 Semantic Segmentation from Figure-Ground

First, suppose there are c object categories $C_1, C_2, \ldots, C_c$ to be predicted. Let I represents the image, as a lattice of pixels. Only one image is concerned since we mainly deals with the inference problem in this section. Suppose in the image I there are r objects $F_1, \ldots, F_r$, presented as subsets of pixels in I. Each object belongs to a particular category, denoted as $F_k \in C_j$. Each pixel p in the image should either belong to a single object or to the background, i.e. $\sum_{k=1}^{r} \mathbb{I}(p \in F_k) \leq 1$. An *object segmentation proposal* (or simply *segment*) $A_i \subset I$ is a subset of I. Suppose we have extracted m segments $A_1, A_2, \ldots, A_m$. Normally, segments can be obtained by an unsupervised multiple segmentation approach such as [6, 11].

Given segments $A_1, A_2, \ldots, A_m$, we partition image I into the identifiable partition $P = \{S_1, S_2, \ldots, S_n\}$ and call each $S_j \in P$ a *superpixel*. In practice we consider only segments that have non-negligible predicted overlap (over a loose threshold) with at least one category. Therefore, in many cases, the superpixels have finer granularity inside objects of interest (fig. 5) and coarser granularity on the background. In practice, we filter out superpixels that are very small (< 50 pixels in our experiments) and assign the corresponding pixels to nearby superpixels. After filtering, most images can be represented using only $20 - 300$ superpixels.

For each segment A_i, its overlap with an object F_k is defined by

$$V_j(F_k, A_i) = \frac{|F_j \cap A_i|}{|F_j \cup A_i|} \tag{9}$$

This intersection-over-union overlap metric is commonly used to evaluate image segmentations since it discriminates a good segment from either a too small segment or a too large one. However for the CSL purpose, this metric is hard to estimate since it directly works on individual objects. Therefore, we define:

$$V_{ik}^0 = V(C_k, A_i) = \max_{F_j \in C_k} \frac{|F_j \cap A_i|}{|F_j \cup A_i|}. \tag{10}$$

as the class-specific overlap that is defined on every category, which is much easier to be learned by a learner.

As the CSL step, we collect a training set of images with pixel-level annotations so that we can observe V_{ik}^0 of each segment A_i on each category. Then, from each segment in each image, a feature vector is extracted using standard appearance features such as SIFT [32], HOG [8], LBP [**?**], pooled on the segment and the ambient context and concatenated. Any regression algorithm can then be used to learn category-specific regressors of V_{ik}^0, with support vector regression normally used for robustness. For details on possible training methods one can consult e.g. [27, 7, 4].

During the CSI step, the test image is first segmented as in the training, all the regressors are then tested on all segments in the test image I, obtaining estimates of the class-specific overlap $\hat{V}^0_{ik}$. The rest of the paper focuses on inference, i.e. recovering the pixel-level labels from these observed statistic estimates.

4.2 The Probabilistic Model for Predicted Overlap

We use θ_{kj} to model the percentage of pixels within a superpixel S_k that belongs to object F_j. Then, the overlap between a segment A_i and F_j can be computed as

$$V_{ij}(\theta) = \frac{|F_j \cap A_i|}{|F_j \cup A_i|} = \frac{\sum_{S_k \in A_i} \theta_{kj}|S_k|}{\sum_{S_k \in A_i} |S_k| + \sum_{S_k \notin A_i} \theta_{kj}|S_k|} \tag{11}$$

Importantly, V_{ij} is computable with θ as the only variables, since each $|S_k|$ is a constant. The idea is that if one parameterizes the ground truth object with θ, then its overlap with each segment can be computed (fig. 1). Now, given the observed overlaps $\hat{V}^0_{ij}$, one can optimize θ by maximizing the composite likelihood of $\hat{V}^0_{ij}$, given the overlap $V_{ij}(\theta)$ computed from θ:

$$\max_{\theta} \sum_{i=1}^{m} \sum_{k=1}^{c} \max_{F_j \in C_k} \log p(\hat{V}^0_{ik}|V_{ij}(\theta)) \tag{12}$$

where the inside $\max$ operation represents the fact that $\hat{V}^0_{ik}$ is an estimate of $\max_{F_j \in C_k} V_{ij}(\theta)$, instead of any $V_{ij}(\theta)$. If we know the number of objects in each category and their rough locations, this can be solved by assigning each $\hat{V}^0_{ik}$ to one of the objects in C_k, so that likelihood is maximized. In order to simplify the presentation of the graphical model, we assume for a moment that this assignment has been resolved, so that each $\hat{V}_{ij}$ has been properly assigned from a corresponding $\hat{V}^0_{ik}$, if $F_j \in C_k$. The CSI problem becomes:

$$\max_{\theta} \sum_{i=1}^{m} \sum_{j=1}^{r} \beta_{ij} \log p(\hat{V}_{ij}|V_{ij}(\theta)) \tag{13}$$

where θ is an $n \times r$ matrix, $\beta_{ij} = 1$ if segment A_i has been assigned to object F_j and 0 otherwise. Note that the assignment is performed within each category, hence a segment can be assigned to many objects, but at most 1 per category. The resolution of the assignment problem within each category will be described in Sec. 4.3.

We assume that the estimated overlap $\hat{V}_{ik}$ is generated from the true overlap V_{ik} plus noise. In order to determine the form of $p(\hat{V}|V)$, we resort to histograms. Fig. 6 shows histograms of $V|\hat{V}$ on data collected from PASCAL VOC training set. The distribution of $V|\hat{V}$ can easily be interpreted as a combination of two components: a bump at $V = 0$, which apparently corresponds to false positive detections, and a centered distribution with $V \neq 0$. As $\hat{V}$ increases, the chance of misclassification is reduced.

Motivated by these observations, we introduce an additional Bernoulli random variable z_{ij} for each predicted score $\hat{V}_{ij}$ (fig. 4). The outcome of z_{ij} informs whether the

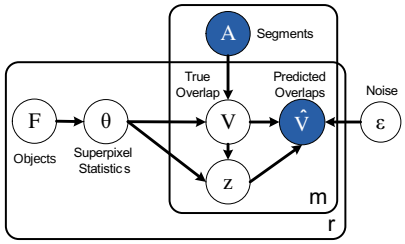

Figure 4: The graphical model used. We separate objects within each category (Sec. 4.4) so that the categorical predictions are mapped to each object. Also, θ and V generate a Bernoulli random variable z, which determines whether the predicted overlap would be a false positive.

prediction $\hat{V}_{ij}$ is a false positive. We make the following distributional assumptions:

$$\begin{aligned} V_{ij}|\hat{V}_{ij}, z_{ij} &\sim \begin{cases} Exp(\lambda)/(1-\exp(-\lambda)), & z_{ij}=0 \\ \tilde{\mathcal{N}}(\hat{V}_{ij}, \sigma^2), & z_{ij}=1 \end{cases} \\ z_{ij}|\hat{V}_{ij} &\sim Ber(\alpha(\hat{V}_{ij})) \\ p(z_{ij}=1|\hat{V}_{ij}, V_{ij}, \theta) &= p(z_{ij}=1|V_{ij}, \hat{V}_{ij}) f(V_{ij}, \theta_{-j}) \end{aligned}$$

where $\theta_{-j} = [\theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_r]$ represents all the θ columns without the j-th.

The three conditional assumptions are in line with our observations: if $z_{ij} = 1$, then $\hat{V}_{ij}$ should be centered around the true overlap, with some noise; an observation is more likely to be a false positive ($z_{ij} = 0$), if the predicted overlap $\hat{V}_{ij}$ is small; when it is indeed a false positive, then the observation $\hat{V}_{ij}$ is independent of V_{ij} given $z_{ij} = 0$, and V_{ij} follows a exponential distribution truncated at 1.

The final assumption is a 'mutual exclusion' assumption. We observe that in categories which are hard to distinguish, e.g. `cat` and `dog`, `horse` and `cow`, a segment often has significant predicted overlaps on multiple categories, but only one of them is correct (see Fig. 9 for an example). In such cases when we have evidence from θ_{-j} that an object in another category might exist, the probability of $z_{ij} = 1$ is diminished by a factor f. For each segment A_i, we compute the segment inside A_i that has the maximal overlap with object F_k (using the algorithm in Section 4), denoted as B_{ik}. Then we take f to be the following function:

$$f(V_{ij}, \theta_{-j}) = \exp\left(-\beta\left(\max_{k \neq j}(V(F_k, B_{ik}) - V_{ij}, 0)\right)\right) \tag{14}$$

Therefore if a part of the current segment A_i has superior overlap with the optimal segment of another object, rather than the current object F_j, then the chance that $\hat{V}_{ij}$ is a true positive is sharply reduced by an exponential term.

Note that the introduction of z_{ij} is only needed when there are significant false positives in the system. In cases where the predictions are accurate enough, z would not need to be estimated as a latent variable. In the "ideal" case experiment we conducted in Section 7.1, we did not employ z and only make the simple assumption that $V|\hat{V} \sim$

$\tilde{\mathcal{N}}(\hat{V}, \sigma^2)$. The EM machinery discussed in the next section is also not needed in this case and the MCLE problem is solved in one run.

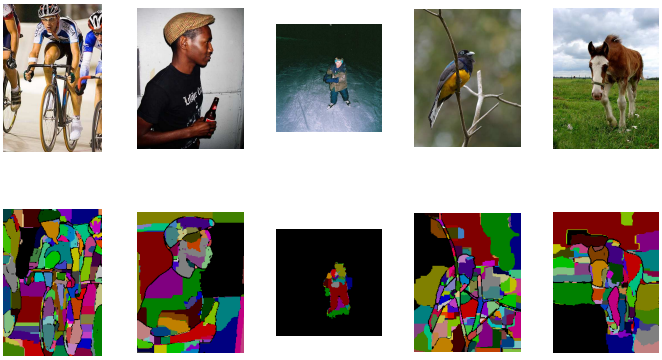

Figure 5: (Best viewed in color) Refined superpixels obtained by multiple intersection from original mid-level segments. Each different color represents a different superpixel (black identifies the largest one). Note that the partitions are, automatically, finer-grained, on the objects of interest.

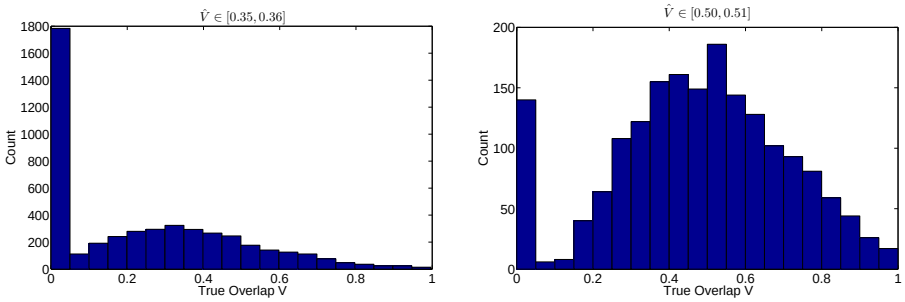

Figure 6: Histograms of true overlap given predicted overlap across the VOC validation set. One can easily identify two components: a probability mass at 0 and a centered distribution to the right. The 0 mass corresponds to misclassifications, where the object does not belong to the predicted category. Also note that with higher predicted overlap $\hat{V}$, there is less chance for $V = 0$.

4.3 Generalized EM Estimation

To maximize the likelihood with latent variable z_{ij}, we adopt an expectation maximization (EM) approach. In the E-step, we will average over choices of z_{ij}, and then in the M-step maximize the expected log-likelihood. Formally, we would like to optimize the

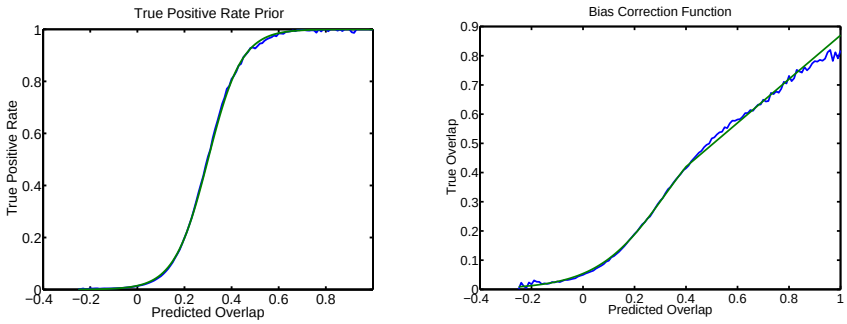

Figure 7: True positive rate prior α and bias correction g fitted from the VOC validation set. α adopts a sigmoid shape. The fitted function is $\alpha(\hat{V}) = 1 - \frac{1}{1+0.015\exp(14\hat{V})}$. For g, the true overlap increases slower when the predicted overlap is high.

composite likelihood with latent variables $Z = [z_{ij}]$:

$$\max_{\theta,Z} \sum_{i=1}^{m}\sum_{j=1}^{r} \beta_{ij} \log p(\hat{V}_{ij}|V_{ij}(\theta), z_{ij}) \tag{15}$$

In the E-step, $\mathbb{E}(z_{ij}|\hat{V}, V, \theta)$ is computed from existing estimates using the Bayes formula:

$$\begin{aligned}\mathbb{E}(z_{ij}|\hat{V}, V, \theta) &= p(z_{ij}=1|\hat{V}, V, \theta) = f(V_{ij}, \theta_{-j})p(z_{ij}=1|\hat{V}_{ij}, V_{ij}) \\ &= \frac{f(V_{ij},\theta_{-j})p(V_{ij}|z_{ij}=1,\hat{V}_{ij})p(z_{ij}=1|\hat{V}_{ij})}{p(V_{ij}|\hat{V}_{ij},z_{ij}=1)p(z_{ij}=1|\hat{V}_{ij})+p(V_{ij}|\hat{V}_{ij},z_{ij}=0)p(z_{ij}=0|\hat{V}_{ij})} \\ &= f(V_{ij},\theta_{-j})\frac{\tilde{\mathcal{N}}(V_{ij};\hat{V_{ij}},\sigma^2)\alpha_{ij}}{\tilde{\mathcal{N}}(V_{ij};\hat{V_{ij}},\sigma^2)\alpha_{ij}+Exp(V_{ij};\lambda)/(1-\exp(-\lambda))(1-\alpha_{ij}))}.\end{aligned} \tag{16}$$

For the M-step, we factorize the joint likelihood function:

$$\begin{aligned}p(\hat{V}_{ij}, z_{ij}=1|V,\theta) &= p(\hat{V}_{ij}, z_{ij}=1|V_{ij},\theta_-) \\ &= p(\hat{V}_{ij}|z_{ij}=1,V_{ij})p(z_{ij}=1|V_{ij},\theta_-) \\ &= \frac{p(V_{ij}|z_{ij}=1,\hat{V}_{ij})p(\hat{V}_{ij}|z_{ij}=1)}{p(V_{ij}|z_{ij}=1)}p(z_{ij}=1|V_{ij})f(V_{ij},\theta_-) \\ &= p(V_{ij}|z_{ij}=1,\hat{V}_{ij})p(\hat{V}_{ij}|z_{ij}=1)f(V_{ij},\theta_-)\frac{p(z_{ij}=1)}{p(V_{ij})}\end{aligned} \tag{17}$$

$$\begin{aligned}p(\hat{V}_{ij}, z_{ij}=0|V,\theta) &= p(\hat{V}_{ij}, z_{ij}=0|V_{ij},\theta_-) \\ &= p(\hat{V}_{ij}|z_{ij}=0)p(z_{ij}=0|V_{ij},\theta_-) \\ &= p(\hat{V}_{ij}|z_{ij}=0)\left(1-p(z_{ij}=1|V_{ij})f(V_{ij},\theta_-)\right)\end{aligned} \tag{18}$$

Take a uniform prior on $p(V_{ij})$ so that it is independent on V_{ij} and ignoring the factors that are independent of θ, and making a further simplification assuming $f(V_{ij}, \theta_-)$

is fixed, we obtain the MAP problem

$$\begin{aligned}
\min_{\theta} \quad & \sum_{i,j} \beta_{ij} \left(\mathbb{E}(z_{ij} = 1|V, \hat{V}, \theta) \left((\hat{V}_{ij} - V_{ij}(\theta))^2 \right) \right) \\
& -2\sigma^2 \mathbb{E}(z_{ij} = 0|V, \hat{V}, \theta) \log(1 + a_1 f(V_{ij}, \theta_{-j}) \exp(-\lambda_1 V_{ij}) - f(V_{ij}, \theta_{-j})) \\
& + \sum_{k=1}^{n} \sum_{j=1}^{c} \lambda_2 |S_k| \theta_{kj}^2 \\
\text{s.t.} \quad & 0 \leq \theta_{kj} \leq 1, k = 1, \ldots, n, j = 1, \ldots, C; \\
& \sum_{j=1}^{C} \theta_{kj} \leq 1, k = 1, \ldots, n
\end{aligned} \tag{19}$$

where a_1 is a constant depending on the various priors. The second part (when $z_{ij} = 0$) is still hard to optimize, but note that when $f(V_{ij}, \theta_{-j}) = 1$, then $\log(1 - f(V_{ij}, \theta_{-j}) + a_1 f(V_{ij}, \theta_{-j}) \exp(-\lambda_1 V_{ij})$ can be simplified to $-\lambda_1 V_{ij}$ in the optimization, we use this approximation and present the final problem for the M-step:

$$\begin{aligned}
\min_{\theta} \quad & \sum_{i,j} \beta_{ij} \left(\mathbb{E}(z_{ij} = 1|V, \hat{V}, \theta) \left((\hat{V}_{ij} - V_{ij}(\theta))^2 \right) + 2\mathbb{E}(z_{ij} = 0|V, \hat{V}, \theta) \sigma^2 \lambda_1 V_{ij}(\theta) \right) \\
& + \sum_{k=1}^{n} \sum_{j=1}^{c} \lambda_2 |S_k| \theta_{kj}^2 \\
\text{s.t.} \quad & 0 \leq \theta_{kj} \leq 1, k = 1, \ldots, n, j = 1, \ldots, C; \\
& \sum_{j=1}^{C} \theta_{kj} \leq 1, k = 1, \ldots, n
\end{aligned} \tag{20}$$

After the M-step, objects F_j with $\max_k \theta_{kj} < 0.05$ or $\max_i z_{ij} < 0.05$ are pruned to improve speed. Therefore the algorithm will process less and less variables with the advance of the EM procedure. Normally the number of iterations won't exceed 20.

The above M-step optimization has a simple convex relaxation. One can simply multiply the denominator in the true overlap V_{ij} to both V_{ij} and $\hat{V}_{ij}$ and divide them by the size of the segment only – which is a constant. This yields the following convex relaxation – a quadratic program:

$$\begin{aligned}
\min_{\theta} \quad & \sum_{i=1}^{m} \sum_{j=1}^{c} \beta_{ij} \left(\mathbb{E}(z_{ij} = 1|V, \hat{V}, \theta) \left(\frac{\sum_{S_k \in R_i} \theta_{kj} |S_k|}{\sum_{S_k \in R_i} |S_k|} - \hat{V}_{ij} \left(1 + \frac{\sum_{S_j \notin R_i} \theta_{kj} |S_k|}{\sum_{S_k \in R_i} |S_k|} \right) \right)^2 \right. \\
& \left. + 2\mathbb{E}(z_{ij} = 0|V, \hat{V}, \theta) \sigma^2 \lambda_1 \frac{\sum_{S_k \in R_i} \theta_{kj} |S_k|}{\sum_{S_k \in R_i} |S_k|} \right) \frac{\sum_{S_k \in R_i} \theta_{kj} |S_k|}{\sum_{S_k \in R_i} |S_k|} + \lambda_2 \sum_{k=1}^{n} |S_k| \left(\sum_{j=1}^{c} \theta_{kj}^2 \right) \\
\text{s.t.} \quad & \sum_{j=1}^{c} \theta_{kj} \leq 1, \theta_{kj} \geq 0, k = 1, \ldots, n; j = 1, \ldots, c
\end{aligned} \tag{21}$$

The solution of this problem is used as a starting point for the solution of optimization (20). The relaxation is in general quite accurate, the only problem is that it biases toward small segments when they are part of a bigger ground truth (then the segment size $|R_i|$ is not a good estimate of $R_i \cup GT$). Therefore we run the relaxation only on segments with size at least 10% of the biggest segment in the image to prevent bad local

minima when later optimizing the true underlying objective. This strategy works well in our case. In the M-step of each EM iteration we first solve the convex relaxation, then use the solution to warm start the optimization (20). A spectral projected gradient method from `minConf`[1] is used to solve both optimization problems.

Two illustrations are shown (Fig. 8 and Fig. 9) to illustrate the procession of the EM algorithm. In the first one (Fig. 8), the algorithm first figures out the person is not a motorbike in the first few iterations, then struggles a bit until finally determining the layout of the person and the horse. In the second one (Fig. 9), the dog also has very strong predicted scores therefore it takes the algorithm a while to gradually wash it out.

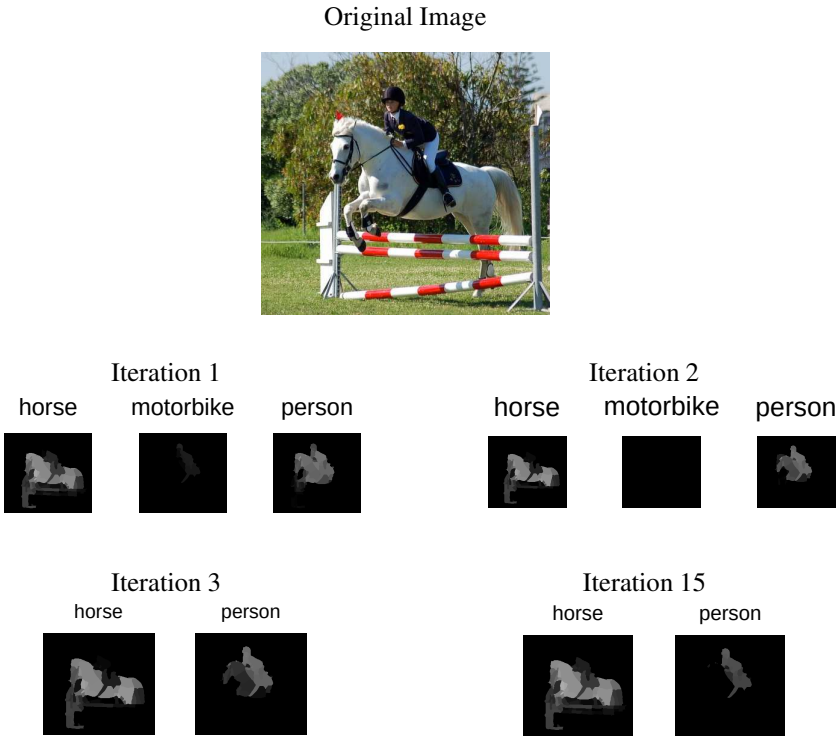

Figure 8: In this image, the first few EM iterations have the `motorbike` potential suppressed. Then, note that in the 1st iteration, both `person` and `horse` have strong potential on the frontal part of the horse. And the horse mask also has some potential on the person. However, the horse finally wins its frontal part and the person wins its body when the algorithm converges.

[1] http://www.di.ens.fr/ mschmidt/Software/minConf.html

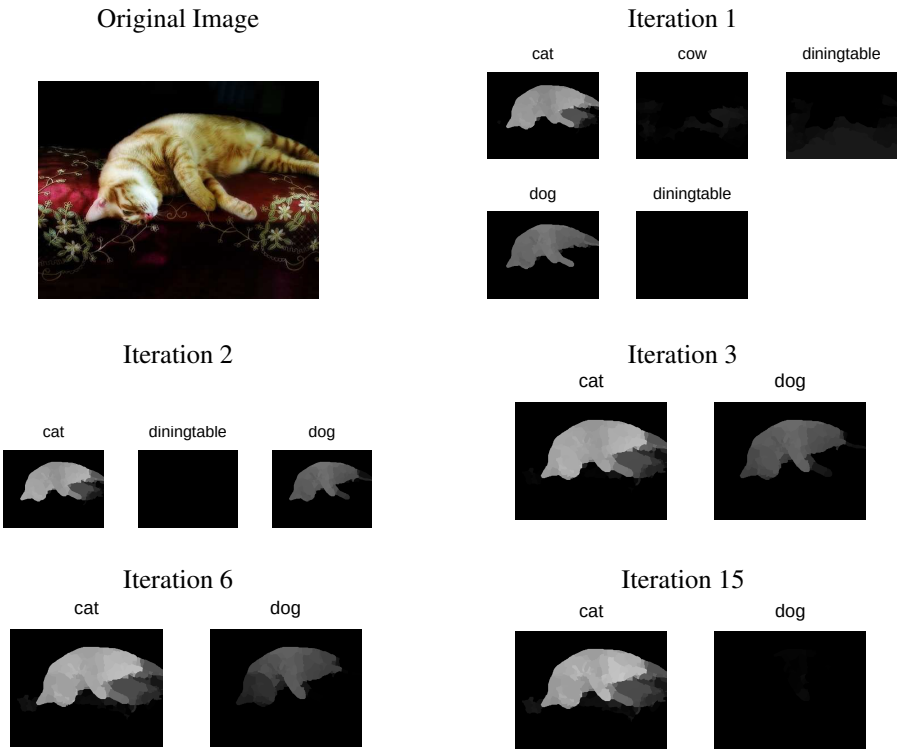

Figure 9: In this image, the main difficulty is the confusion between the `cat` and `dog` categories. The algorithm takes several iterations to suppress the strong prediction score on the `dog` object.

4.4 Locating Multiple Objects within Each Category

To locate multiple objects in one category and in order to assign the predicted overlaps to each object, we adopt the above EM estimation with a hypothesis-testing MAP framework to find the number of objects in each category, before engaging in the full EM estimation described in the previous subsection. Namely, we solve (8) for each category C_k independently, with an additional geometric prior on the number of objects r_k: $p(r_k = j) = (1-q)^j q$, where $q > 0$ is a parameter. For each of $r_k = 1, 2, 3$, etc., the following posterior is computed:

$$L_{r_k} = \max_{\theta, Z} \sum_{i=1}^{m} \max_{j \in 1,\ldots,r} \log p(\hat{V}_{ik}^0 | V_{ij}(\theta), z_{ij}) + r_k(1-q) \tag{22}$$

by maximizing over θ and Z. The posteriors L_{r_k} are computed iteratively. First L_1 is computed by setting all $\mathbb{E}(z_{i1}) = \alpha_{ik}$ and running the M-step (20) only. Then, suppose L_{r_k} is computed with the optimized parameters as θ_{r_k}, Z_{r_k}, L_{r_k+1} is inductively

computed by adding one object with an initialization of:

$$\mathbb{E}(z_{i,r_k+1}) = 1 - \frac{\max_{j\in\{1,\ldots,r_k\}} p(\hat{V}^0_{ik}|V_{ij}(\theta_r))}{p(\hat{V}^0_{ik}|V = \hat{V}^0_{ik})} \tag{23}$$

and running the EM steps (20) and (17) until convergence. In (23), the denominator represents the maximum likelihood from any configuration, and the nominator represents the likelihood of the best explanation of $\hat{V}^0_{ik}$ by any of the current j objects. The logic behind (23) is that, if $\hat{V}^0_{ik}$ has already been explained perfectly, adding an object cannot improve the likelihood thus $\mathbb{E}(z_{i,r_k+1})$ is initialized to 0. If none of the objects has been able to explain $\hat{V}^0_{ik}$ so far, then a new object is likely present, thus $\mathbb{E}(z_{i,r_k+1})$ is initialized to 1.

In addition, we employ a k-means initialization to escape potential local optima using the previous optimization. With this initialization, k-means on all the segments with significant predicted overlap are performed, using the hamming distance on the segments (as a binary vector of include/exclude on all superpixels). Then $\mathbb{E}(z_{i,k})$ are set to 0.75 if segment A_i belongs to cluster k, and 0.25 otherwise. The algorithm optimizes with both initializations and choose the one with the best objective value.

At any point, if $L_{r_k+1} < L_{r_k}$, the computation is stopped and r_k is decided to be the number of objects. Then, each segment is assigned to the object F_j that maximizes $\mathbb{E}(z_{i,j})$ in the final Z_{r_k}. The joint inference on all categories is subsequently performed, by treating each object as a different category with separately assigned predictions.

4.5 The Full Procedure

The full inference procedure involves two steps:

- Determining the number of objects within each category by the within-class object separation routine in Sec. 4.4.
- Performing joint inference by iterating (17) and (20) across all categories and objects.

Notice that we choose to perform the within-class object separation routine before the joint inference, because within each category the enumeration of object counts is tractable. If one enumerates in the joint inference phase, then hypotheses like "1 object in c_1, 2 objects in c_2" need to be tested and could lead to exponential blowup when there are many categories. Whereas, even if the within-class object separation can make mistakes, the erroneous object hypotheses can still be suppressed during the joint inference.

In fig. 10 we show the result of running the within-class object separation routine on the segments in fig. 3. One can see that in both the bicycle and the person categories, two objects are generated instead of one. Although both categories improve the likelihood by predicting 2 objects, the second bicycle object is erroneous whereas the second person object is correct. After detecting two objects for each category and running joint inference with these 4 objects, the algorithm is able to correct that mistake, as shown in fig. 11.

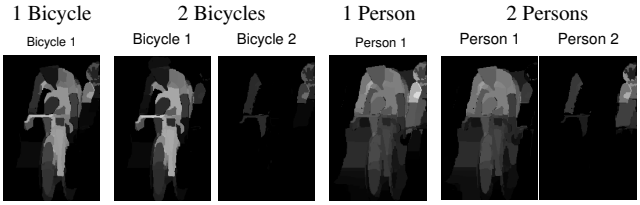

Figure 10: Different θ computed for 1 bicycle/2 bicycles, and 1 person/2 persons hypotheses for the same set of predicted segment overlaps. The second bike represents spurious predictions from noise, whereas separating two people indeed improves the solution.

Figure 11: Joint optimization on 4 objects. One can see that potentials for Bicycle 2 have been suppressed due to similar spatial layout and lower scores to Person 2.

5 Implementation Details

5.1 Segment Weighting

A major shortcoming of the composite statistical inference approach is the need to re-weight the sampled subsets (segments) during inference. Because i.i.d. sampling is not assumed and the fact that we use a deterministic prediction function (instead of multiple random samples of the statistics), significant sampling biases could occur in the segments and a meticulous effort must be made to correct those biases. As a simple example of the sampling bias in the semantic segmentation problem, imagine we sample a same wrongly predicted (in terms of category label) segment 10 times and the correctly predicted segment only once. Without proper weighting, the composite likelihood inference problem will favor the wrong segment simply because it occurs more often, although that is an artifact of non i.i.d. sampling.

It is obvious how to avoid completely identical samples occurring multiple times, but usually one should retain multiple segments that have significant amount of overlap (e.g. 70%) since those could provide information from different facets that are all helpful to the inference. However, care still needs be taken to avoid segments from a certain category outscore others merely by outnumbering them in the (biased) sampling.

Another bias that is introduced is the regularization. Since we regularize equally on each superpixel in the image, segments that have more superpixels are receiving more regularization than segments with less superpixels. Therefore they ought to be given a higher weight to counter-balance such an effect.

It is hard to design weight choices on multiple overlapping segments with predic-

tions on many objects. Therefore we explore achieving our design goals from simple cases. The first design goal of the weighting is to ensure that segments predicted as different categories and of different size would incur the same loss if their predicted overlap is the same. Suppose one have segments A_1 and A_2, all containing exactly one superpixel (but of different size $|S_1|$ and $|S_2|$ respectively), and predicted to 2 objects with the same overlap score $\hat{V}$. Then suppose the CSI optimization problem is:

$$\min_{\theta_1,\theta_2} f(\theta_1,\theta_2) = w_1(\theta_1 - \hat{V})^2 + w_2(\theta_2 - \hat{V})^2 + \lambda_{21}|S_1|\theta_1^2 + \lambda_{22}|S_2|\theta_2^2 \tag{24}$$

where we have splitted λ_2 into λ_{21} for the regularization on the first object and λ_{22} for the regularization on the second object to have more degrees of freedom in weighting.

Our design goal for the weighting is to have

$$f(\hat{\theta}, 0) = f(0, \hat{\theta}), \forall \hat{\theta} \geq 0 \tag{25}$$

because according to the information we have, there should be no difference in likelihood predicting to these 2 objects. It is easy to see that the solution to achieve the identity is to have $\frac{\lambda_{21}|S_1|}{w_1} = \frac{\lambda_{22}|S_2|}{w_2}$, which leads to our principle #1: **The regularization parameters should be inversely proportional to the size of the objects**.

Now suppose there are n_1 copies of A_1 and n_2 copies of A_2, in order to still satisfy the above identity, we can set w_1 and w_2 to different values so that $n_1 w_1 = n_2 w_2$ to negate this sampling bias. This leads to design principle #2, **The total amount of weights on each object should be about the same**. In order to achieve this on multiple overlapping segments, we designate that **the total weight on each superpixel is to be 1** and let the segment weight be a sum of all the superpixel weights inside the segment. One can verify in the above case that our design choice achieves the desired goal, because multiplicities in sampling have been negated by spreading the weight evenly across all the copies.

One heuristic design choice we made is to make smaller segments take a larger share when a superpixel is shared by many segments. The rationale behind this is that in a smaller segment that contains the superpixel, it contributes more to the segment statistic, while in a larger segment (e.g. the full image), the small superpixel still contributes, but to a much less extent. For accordance with the Gaussian assumption (and square loss) we chose, we designate $v_{ij} = \mathbb{I}(S_j \in A_i)\left(\frac{|S_j|}{|A_i|}\right)^2$ and $\bar{v}_{ij} = \frac{v_{ij}}{\sum_i v_{ij}}$ be its normalized version. The weight of a segment is thus:

$$w_i = \sum_{j, S_j \in A_i} \bar{v}_{ij} = \sum_{j, S_j \in A_i} \frac{v_{ij}}{\sum_i v_{ij}} \tag{26}$$

Finally, the weight is renormalized so that the average weight of each segment is 1, to keep accordance with the other parameters in the system.

After weighting, we also need to renormalize the λ_2 for different objects, this is done by summing up the sizes of all superpixels that are supposedly within the object. Ideally, this should change as we are progressing the optimization, but that will lead to convergence issues of the algorithm. We heuristically fix the λ_2 for each object F_k to be

$$\lambda_{2k} = \frac{1}{\sum_{j, \sum_{S_j \in A_i} \alpha_{ik} \geq n_m} |S_j|}. \tag{27}$$

where the summand in the denominator means that we only consider superpixels which belong to multiple detections that has a sum of true positive probabilities more than n_m (a parameter). n_m is set to 7 in our experiments.

6 Optimal Full Image Labeling

Given the inferred real-valued parameters θ (e.g. fig. 11), we still need to produce a consistent segment for each object. A graph-cut algorithm can be used on a potential map like fig. 11, but because θ has different magnitudes in different images, a uniform cut parameter choice across a dataset is unlikely to be successful. We propose an algorithm to produce optimal segments that maximizes the overlap with ground truth, without the need to re-segment. First, note that the overlap formula (11) can also be written as:

$$V(F_j, A) = \frac{\sum_{S_k \in A} \theta_{kj} |S_k|}{\sum_{k=1}^{n} \theta_{kj} |S_k| + \sum_{S_k \in A} (1 - \theta_{kj}) |S_k|} \tag{28}$$

where in the denominator we first count all the ground truth pixels in F_j by $\sum_{k=1}^{n} \theta_{kj} |S_k|$, then sum all the pixels inside segment A_i that do not belong to F_j. This reformulation leads to a simple approach to grow A optimally. Suppose we have A with $V(F_j, A) = V_0$, then V can be increased if and only if we add a superpixel to A with $\frac{\theta_{kj}}{1-\theta_{kj}} > V$, because $\frac{a+c}{b+d} > \frac{a}{b}$ iff $\frac{c}{d} > \frac{a}{b}$. Therefore, when the image contains only an object in a single category, the optimal segment can be found by starting from $A = \emptyset$ and $V = 0$; sort $\frac{\theta_{kj}}{1-\theta_{kj}}$ corresponding to all superpixels in descending order; and keep adding superpixels from the top of the list until $V \geq \frac{\theta_{kj}}{1-\theta_{kj}}$ for all remaining superpixels.

In case the optimal segments in multiple categories conflict on some superpixels, one can run a branch-and-bound search on all the conflicting superpixels to maximize the sum of overlaps on each object. The overall goal is to optimize the joint overlap of all non-overlapping objects:

$$\max_{\substack{A_1, \ldots, A_r \\ A_i \cap A_j = \emptyset, \forall i,j}} \sum_{i=1}^{r} V_\theta(F_i, A_i). \tag{29}$$

The best-first search starts from the configuration $A_1, \ldots, A_r$ so that the superpixels are assigned to the optimal segment of each individual object (but may overlap), and try to make moves by removing superpixels. The first move is performed by removing a conflicting superpixel S_k from object F_j that minimizes the quality function

$$Q_{kj} = V_\theta(F_j, A_j) - \max_{A, S_k \notin A} V_\theta(F_j, A) \tag{30}$$

which means after removal of S_k, we recompute the overlap of the best segment for F_j without S_k, and choose the move which loses minimal overlap. The search is performed then in a depth-first manner, first obtaining a full solution without conflicts, then backtrack and search for potential improvements. It is easy to prune the search space because the overlap upper bound

$$\overline{V^j_{-k}} = \max_{A, S_k \notin A} V_\theta(F_j, A) \tag{31}$$

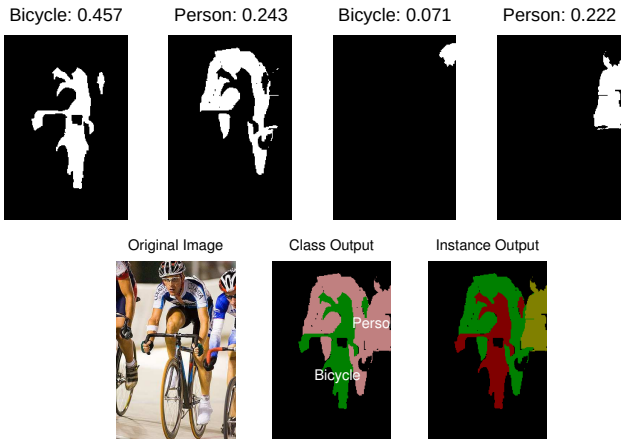

Figure 12: Final masks and final output of the algorithm. Bike 2 is filtered out because of very low score. Not all superpixels with non-zero potentials are in the final mask, because adding some more would be suboptimal according to the procedure in Sec. 6. It is interesting to see that the first person has his right leg correctly cut through by the bicycle, a solution that was not available in any of the initial object segmentation proposals.

can be computed in constant time, given sorted lists of $\frac{\theta_{kj}}{1-\theta_{kj}}$. If after a move of removing S_k from A_j, the upper bound

$$\overline{V^j_{-k}} + \sum_{i=1, i\neq j}^{r} \max_A V_\theta(F_i, A) \tag{32}$$

is smaller than the current objective value, then the branch consisting of the move removing S_k from A_j can be directly pruned.

The search can be performed quickly because: 1) Since θ from all categories are optimized jointly, one superpixel is likely to be assigned to a single category and only a limited number of superpixels will be simultaneously present in the optimal segment of many categories (see e.g. Fig. 8); 2) The bounds obtained with the above procedure are usually quite tight In many cases, a greedy approach using the quality function achieves the optimal solution. Fig. 12 shows the search results for the 4 objects in fig. 11 as well as the final output.

7 Experiments

The main experiments are conducted on the PASCAL VOC Segmentation dataset [12], a widely used benchmark for semantic segmentation. This dataset defines 20 object categories and provides around $3,000$ training images with pixelwise ground truth annotations. This set, named `trainval`, was further divided into half in the `train` and half in the `val` set. In addition, around $9,000$ images annotated with bounding box

information can be used for training. The final benchmark of performance is a held out `test` set, for which the ground truth is not available and evaluation can only be done by submitting results to an online evaluation server. Performance is evaluated as the average pixel precision, computed on all the pixels of each class and then averaged over the 20 classes plus background.

7.1 Experiment using Noise-Free Predictions

In order to examine the consistency of the approach empirically, we perform an experiment by supplying the ground truth maximal class-specific overlap to the algorithm. In this experiment, we do not use EM iterations except in the within-class object separation routine (Sec. 4.4). We also do not use segment weighting (Sec. 5.1) since any weighting scheme would generate the same results under zero noise. Besides, λ is always set to 0 since there is no false positives. Ideally, we should also set λ_2 to 0, however, empirically we found out that setting it too small greatly slows down convergence. Therefore we set it to 0.005. Since δ is irrelevant without EM iterations, the algorithm in this case has no parameters except λ_2, which is nonzero purely for numerical convergence reasons. One should note how simple the algorithm is, if we do not have to deal with the complicated noise in the practical scenario.

We compare on the PASCAL VOC `val` set against the heuristic sequential approach SVRSEGM, and another approach which greedily selects the best non-overlapping segments given the ground truth objects. Another upper bound is the `Superpixel Max`, which measures the performance by classifying each superpixel to the best category label. The result is shown in Table 1. One can see that our multi-intersection superpixels lose about 5% from the full 100% due to errors in boundary detection and segmentation. The performance of CSI given the ground truth maximal class-specific overlap is only less than 6% worse than the optimal superpixel performance, and outperforms `CPMC Max` by a margin of 6%. This shows the need for recombining initial noisy segments in order to fully solve the segmentation problem. It also shows the effectiveness of the overlap statistic in encoding higher-order information in noisy segments.

As a class-specific analysis, note that the `bicycle` category is the hardest where even the best superpixels can only obtain less than 80% accuracy. Then `CSI Max` further drops this to 63%, due to the difficulty of finding good segments. The next two difficult categories are `Potted Plant` and `Chair` where the `Superpixel Max` is about 90% and `CSI Max` is about 80%. For most other categories, `Superpixel Max` has over 95% accuracy and `CSI Max` is about 5% worse than the best superpixels. Given that we have only used 150 segments and the combinatorial possibilities are far from fully explored, we conclude that these results are reasonable and show the vast potential of the CSI framework.

7.2 PASCAL VOC Results

For this experiment, we tune the parameters λ, λ_2 and δ and the α function on the `val` set using the regressor output trained on `train` and the additional images with bounding box annotations. Then, evaluation is performed on the `test` set with the

Table 1: Upper bound results on the VOC 2012 `val` set. `SVRSEGM Max` and `CSI Max` represents the performance SVRSEGM and CSI could obtain by supplying them the maximal class-specific ground truth overlaps on each category. `CPMC Max` represents the performance that can be obtained by selecting the best original CPMC segments, note that since these segments could overlap, this oracle accuracy is not attainable in reality. `Superpixel Max` represents the performance obtained by classifying each of the refined superpixels to the right category.

Method	SVRSEGM MAX	CPMC MAX	CSI MAX	Superpixel MAX
Mean	79.0	81.8	**90.2**	95.2
Background	93.3	91.2	**97.2**	98.8
Airplane	84.6	83.5	**93.3**	96.4
Bike	47.3	48.8	**62.9**	79.4
Bird	87.6	82.5	**95.0**	97.0
Boat	80.1	82.0	**88.8**	94.3
Bottle	77.8	82.4	**90.9**	95.7
Bus	83.2	85.3	**92.3**	96.7
Car	79.4	78.8	**89.8**	94.4
Cat	89.4	91.6	**96.2**	98.2
Chair	63.2	71.6	**82.2**	92.3
Cow	86.2	89.2	**94.8**	98.0
Dining Table	75.0	82.8	**90.9**	96.5
Dog	88.0	90.6	**95.9**	98.1
Horse	82.5	82.3	**92.6**	96.2
Motorbike	78.3	78.1	**91.0**	95.9
Person	74.9	77.0	**89.7**	96.2
Potted Plant	72.2	79.0	**79.4**	90.2
Sheep	87.8	84.2	**94.5**	98.1
Sofa	67.2	85.0	**93.7**	97.3
Train	80.8	84.6	**92.8**	95.9
TV/Monitor	79.2	88.2	**89.6**	94.5

tuned parameters and fitted functions. The overlap predictions $\hat{V}$ used in our system are obtained by combining the regressors from [27] and [4], with linear weights learned on the `trainval` set. The parameters λ, λ_2 and δ are tuned on the `val` set.

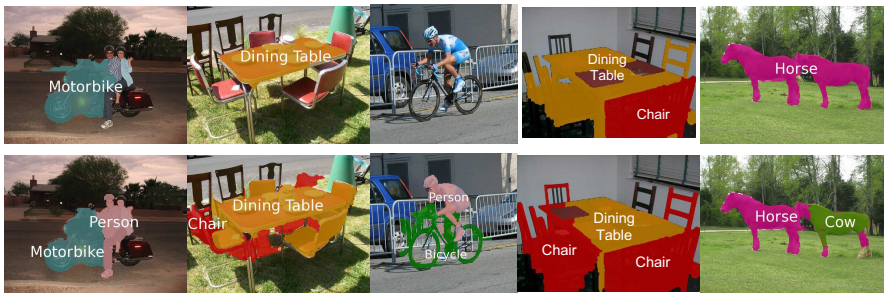

Figure 13: Example of semantic segmentations. The first row shows results using the post-processing algorithm of [27], the second row shows results of the proposed CSI algorithm. Areas of the image labeled as background are depicted with their original appearance. The first four images show cases where our algorithm is more accurate, mainly involving relatively complex scenes with multiple interacting objects. The last image, on the right, shows a typical failure case: segments covering part of one of the horses are strongly confused and assigned to 'cow'. The algorithm of [27] typically oversmooths the predictions, which is advantageous in some cases, like in this image.

On the VOC `test` set, we compare the proposed CSI approach against other methods on the 2012 challenge using the same set of category prediction scores, which includes SVRSEGM [7] and JSL [17]. The JSL entry to VOC 2012 is different from the paper [17] in that it also employed pixel-level averaging to improve performance. It can be seen from Table 2 that the method performs slightly better than the others, especially for object categories involved in interactions such as `Bike`, `Chair`, `Person` and `Sofa`. It does less well in the animal categories where interactions are less likely to happen. The 47.5% overall result for CSI is the best reported on *comp5* of the VOC 2012 challenge so far [12].

In order to gain more insights, we partition the 20 PASCAL object categories into bigger subgroups. We make 2 types of partitions to the 20 object categories. In the first partition, we divide the categories into 3 almost equal subsets: **Indoor objects + Person**, **Animals**, **Outdoor Objects**. `Bottle`, `Chair`, `Dining Table`, `Potted Plant`, `Sofa`, `TV/Monitor` and `Person` belongs to the class **Indoor objects + Person**. The 6 animal categories `Bird`, `Cat`, `Cow`, `Dog`, `Horse`, `Sheep` which are partitioned to **Animals**. The other 6 non-animal categories are placed in **Outdoor Objects**. The second split is into 2 subgroups: **High Interaction Categories** and **Low Interaction Categories**. **High Interaction Categories** include "ride-able" ones such as `Bike`, `Motorbike`, `Horse`, "sit-able" ones such as `Chair`, `Sofa`; "hold-able" ones such as `Bottle`; `Dining Table` because it is often in some configuration with chairs and persons, as well as `Person` which interact with all other categories. The rest are classified as **Low Interaction Categories**. From Table 3 one can see that CSI obviously improves on **Indoor Objects**, where it is 2.4% and 2.8% better than

Table 2: VOC 2012 test results

Method	SVRSEGM	JSL	CSI
Mean	46.8	47.0	**47.5**
Background	84.9	85.1	**85.2**
Airplane	63.8	**65.4**	64.0
Bike	22.1	29.3	**32.2**
Bird	50.5	**51.3**	45.9
Boat	**38.9**	33.4	34.7
Bottle	44.8	44.2	**46.3**
Bus	**61.3**	59.8	59.5
Car	**63.3**	60.3	61.6
Cat	48.8	**52.5**	49.4
Chair	9.8	13.6	**14.8**
Cow	**57.2**	53.6	47.9
Dining Table	**35.6**	32.6	31.2
Dog	**43.0**	40.3	42.5
Horse	51.1	**57.6**	51.3
Motorbike	**58.8**	57.3	**58.8**
Person	53.7	49.0	**54.6**
Potted Plant	29.7	33.5	**34.9**
Sheep	49.8	53.5	**54.6**
Sofa	30.3	29.2	**34.7**
Train	47.0	47.6	**50.6**
TV/Monitor	38.0	37.6	**42.2**

SVRSEGM and JSL, respectively. On **High Interaction Categories**, it also shows 2.2% improvement over SVRSEGM and 1.4% improvement over JSL. It works worst in **Animals**, where it is 1.5% worse than SVRSEGM and 2.9% worse than JSL. This indicates that the same parameters for interaction handling might not work uniformly for both **Indoor Objects** and **Animals**, especially because confusion between categories is a much more significant problem in **Animals** than **Indoor Objects** (e.g. confusion between `Cat` and `Dog`, `Horse` and `Cow`, `Horse` and `Dog`, `Cow` and `Sheep` are all pretty significant). In future work, we will try to work on injecting different priors in different categories hoping to make the system better all around.

We show some images on the VOC test set in fig. 13. It can be seen that CSI handles object interactions very well in many cases.

7.3 Validation Set Results

We show results on the validation set of the 2012 challenge in Table 4. We compare in detail the results of non-maximum suppression, SVRSEGM and CSI in different conditions. It can be seen that both SVRSEGM and CSI are much better than non-maximum suppression, with CSI slightly better than SVRSEGM, especially on interaction categories such as `Bike`, `Chair`, `Dining Table`, `Potted Plant` and `Person`.

Method	Indoor Objects + Person	Animals	Outdoor Objects	High Interaction Categories	Low Interaction Categories
SVRSEGM	34.6	50.1	50.7	38.3	**49.3**
JSL	34.2	**51.5**	50.4	39.1	49.1
CSI	**37.0**	48.6	**51.6**	**40.5**	49.0

Table 3: Comparison of 3 inference methods on VOC 2012 test set when the predictions are grouped by categories. CSI obviously outperforms SVRSEGM and JSL in **Indoor Objects + Person** as well as **High Interaction** objects.

Similar to test set results, CSI is worse than SVRSEGM on animal categories such as `Cat` and `Cow`. `Sheep` is a particular animal category in which CSI does better than SVRSEGM, upon investigation on the dataset we find that when `Sheep` appears, there are very often multiple sheep in one image. Thus CSI's better handling of multiple objects might have helped it.

We have also tried to run CSI without the EM (as **No EM**) or without the mutual exclusion prior (as **No ME**). The similarly dismal performances there show the crucial importance of the mutual exclusion prior as giving the approach some non-maximum suppression capabilities. However, it can be seen for certain categories such as `Bike`, `Person`, `Motorbike`, `Boat` where false positives are unlikely to happen, the performance without mutual exclusion is on par or slightly better than the full system. This shows that perhaps false positives are unlikely in these categories, and without the mutual exclusion, the interaction between objects is still handled correctly to give good performance.

Note that validation set results are significantly higher than the test set results ($50.5 - 50.9\%$ vs. $46.8 - 47.5\%$). There are two potential reasons of this, first is when we are combining the scores of two sets, we used ground truth overlaps to fit a regressor on the `trainval` set, which can overfit to `val` quite a bit. Another potential reason is that the test set of 2012 is harder than `trainval`, as well as the test sets in previous years. For example, with the same set of images that obtain 47.5% in the 2012 test, we have obtained 48.8% and 49.6% in the 2011 and 2010 test sets, respectively.

7.4 MSRC Results

In order to test full-image annotation capabilities of the algorithm, we have also included tests on the MSRC-21 segmentation dataset. The MSRC-21 dataset contains 21 object categories in 591 photographs, separated into training (45%), validation (10%) and test (45%) sets. The main difference between MSRC and PASCAL VOC is the inclusion of "stuff" classes, which are background classes that are mainly defined by texture and cannot be counted, such as `Grass`, `Sky`, `Water`, etc. Some other MSRC classes are also of the textured nature although it might not look so from the definition, such as `Flower` and `Book`, where the books are usually stacked on a bookshelf and all the books on the shelf are annotated as one book (Fig. 14). Even `Bicycles` are always a lot of bikes stacked together. Furthermore, in the standard benchmark, background is not counted in the final score. All these seem to be favoring CRF-type

	NMS	SVRSEGM	CSI		
			No EM	No ME	Full
Mean	47.53	50.51	45.66	45.73	50.86
Background	84.58	85.44	82.49	82.58	85.38
Airplane	66.80	70.19	60.65	60.26	67.21
Bike	17.72	23.89	28.14	28.36	27.61
Bird	50.11	50.58	43.22	41.52	50.17
Boat	38.89	42.45	45.89	46.01	44.57
Bottle	41.74	44.61	38.65	36.88	43.96
Bus	62.19	66.93	61.95	61.78	66.16
Car	66.22	67.85	61.07	61.46	65.36
Cat	60.62	67.01	54.20	54.45	62.36
Chair	9.20	20.04	20.00	19.99	22.95
Cow	47.70	51.92	40.82	42.06	50.97
Dining Table	26.71	27.95	24.24	26.12	29.38
Dog	50.66	50.22	40.44	41.09	50.19
Horse	50.65	56.76	44.66	43.82	56.43
Motorbike	48.68	51.29	49.04	50.20	49.88
Person	52.85	55.04	54.55	55.35	56.27
Potted Plant	33.07	37.14	32.83	33.28	40.90
Sheep	58.01	60.61	47.03	45.65	62.76
Sofa	28.72	29.80	27.59	28.81	30.23
Train	54.11	53.51	54.36	54.99	56.14
TV/Monitor	48.94	47.39	46.97	45.67	49.12

Table 4: Validation results. **No EM** shows the setting when the algorithm is stopped after the first iteration and **No ME** shows the setting when the mutual exclusion prior is not enforced (see Section 4.1). One can see CSI without either of these two would not perform very well.

approaches that work on pixel-level classification and smooth it over the entire image. With certain parameter settings, CRF is capable of oversmoothing to the entire image, beneficial for several categories such as `Flower` and `Book`, as shown in the figure.

Our goal is to test the sliding segment-based approach on this dataset while acknowledging its heavy design bias toward CRF-type texture recognition and smoothing models. This has not been done in previous sliding segment approaches due to the lack of a consistent inference method. As in PASCAL VOC, we extract 150 segments per image, extract the same features as in PASCAL VOC, and train the regressors with O2P features and color SIFT. The parameters of the regressor and CSI are tuned with the MSRC validation set. There are two types of annotations in MSRC-21 that has been commonly used. The original annotation and the more accurate "clean" annotation from [33]. We train with the precise annotation and test on both the original and clean annotations.

The results are shown in Table 5. The "Average" column shows the per-category averages and the "Global" Column shows the per-pixel averages. One can see the segment-based approach is comparable with CRF approaches. On the Clean annota-

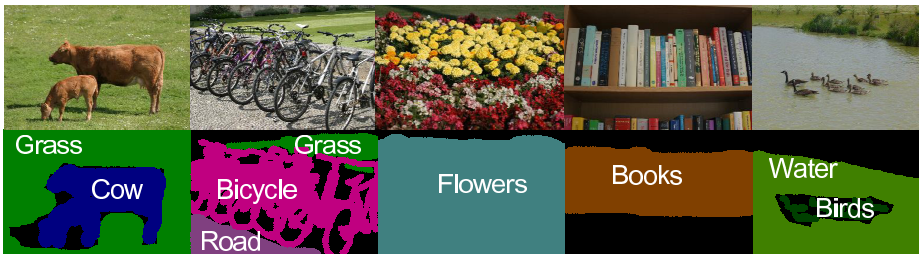

Figure 14: MSRC images and annotations. One can see background categories and texture categories.

tions which is more accurate, an obvious pattern is that CSI greatly outperform CRF-based methods on objects, but CRF-based methods excel on textures. The Global accuracy of CSI is lower because there are more pixels in the textures but the average accuracy of CSI is higher because there are more categories that are objects, as compared with texture categories. One notable thing is CSI, as a superpixel-level approach, still does a reasonably good job on textures, such as `Grass`, `Sky`, `Water`.

Taking a closer look, on the clean annotations in 4 object categories CSI improved over 10% over the competitors: `Cow`, `Bird`, `Chair`, `Boat`. The only classes that CSI is more than 10% worse than the top method is `Book` and `Road`, the former is mainly due to the lack of oversmoothing capability in CSI, the latter due to the fact that road pixels are often mixed with other objects, therefore some accuracy has already been lost in the superpixel stage.

On the original annotations, we compare with more approaches due to more have reported results using these annotations. The average accuracy of CSI is still comparable with the top methods here. It only wins in two categories, `Sign` and `Bird`, but outperformed the next competitor at least 9% in each of these categories. Note one can clearly see that some deficiencies of CSI are mainly due to annotation errors, since two competitors Yao et al. [45] and HCRF [25] dropped 2% moving from original to clean annotations, but CSI gained 1% instead. Especially, in categories such as `Cow` and `Sheep`, with clean annotations CSI improved by a lot. Therefore, the results here against HCRF and Yao et al. are not meant to be trusted. But one can see that CSI outperforms earlier methods such as TextonBoost [36], Jiang and Tu, and Harmony potential [14] by a comfortable margin. Dense CRF [22] on the other hand, has better pixel-level accuracy than the other approaches thus would benefit more if the annotation is clean [22] (however they pixel-level accuracy is so high that they need even cleaner annotations than the "Clean" one used here) and might outperform CSI with the clean annotation.

If one groups the object categories and texture categories (defined as `Grass`, `Sky`, `Water`, `Road`, `Flowers`, `Book`, the reason the last two are texture categories can be seen in Fig. 14) separately, it can be seen that CSI wins the object categories by a large margin but is losing on the texture categories. A potential future work is to connect the CSI with Dense CRF in order to gain both object recognition capabilities and high

Table 5: MSRC-21 results

Method	Average	Global	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat
Clean MSRC Annotations																							
CSI	79.2	81.8	68	87	79	92	92	94	84	75	89	84	89	84	89	68	83	80	78	79	60	71	39
HCRF [25]	75.8	85.9	73	93	82	81	91	98	81	83	88	74	85	97	79	38	96	61	90	69	48	67	18
Yao et al. [45]	77.4	84.4	67	92	80	82	89	97	86	83	86	79	94	96	85	35	98	70	86	78	55	62	23
Original MSRC Annotations																							
CSI	78.2	83.1	69	95	83	82	86	95	76	76	89	81	89	83	86	66	82	78	83	78	60	70	36
TextonBoost [36]	67	72	49	88	79	97	97	78	82	54	87	74	72	74	36	24	93	51	78	75	35	66	18
Jiang and Tu [18]	68	78	53	97	83	70	71	98	75	64	74	64	88	67	46	32	92	61	89	59	66	64	13
Harmony Potential [14]	75	77	60	78	77	91	68	88	87	76	73	77	93	97	73	57	95	81	76	81	46	56	46
HCRF [25]	77.8	86.5	74	98	90	75	86	99	81	84	90	83	91	98	75	49	95	63	91	71	49	72	18
Dense CRF [22]	78.3	86.0	75	99	91	84	82	95	82	71	89	90	94	95	77	48	96	61	90	78	48	80	22
Yao et al. [45]	79.3	86.2	71	98	90	79	86	93	88	86	90	84	94	98	76	53	97	71	89	83	55	68	17

pixel-level accuracies.

Some sample images are shown in Fig. 15, where one can see the strength and weaknesses of CSI for interpreting the full image. CSI is capable of recognizing object categories and arrange them for a full-image interpretation. Contrary to some popular beliefs on methods based on unsupervised figure-ground segmentations, the performance on certain texture categories, such as `grass`, are not bad. However, CSI does suffer when it sometimes attempts to separate different objects for a texture category, which leads to a deficiency in certain texture categories such as `water` or `book`.

8 Conclusion

This paper proposes a composite statistical inference approach to semantic segmentation. The composite likelihood methodology is generalized to model one-dimensional error distributions of statistical estimates. Based on this generalization, superpixel-level inference is performed based on a set of mutually overlapping object segmentation proposals and their predicted overlaps with object categories. The generative process underlying overlap prediction is modeled using a graphical model and an EM algorithm is proposed to solve the maximum composite likelihood inference in two steps: the number of objects in each category is first determined, then a joint optimization is performed for all objects across categories. Once superpixel-level parameters have been estimated, the optimal pixel-level segmentation can be computed efficiently by best-first search. Experiments demonstrate the effectiveness of the approach, especially in scenes with multiple objects and interactions.

Table 6: MSRC-21 results grouped by objects or texture categories. CSI is even in both, while the CRF approaches usually does texture categories way better than object categories.

Method	Object Categories	Texture Categories
Clean MSRC Annotations		
CSI	77.5	83.5
HCRF [25]	69.0	92.8
Yao et al. [45]	71.4	92.0
Original MSRC Annotations		
CSI	75.3	85.7
TextonBoost	62.8	77.5
Jiang and Tu	61.3	84
Harmony Potential	71.1	85
HCRF	71.1	94.2
Dense CRF	73.4	91.0
Yao et al.	73.7	93.5

Acknowledgements: The authors thank Joshua Dillon for helpful discussions. This work was supported in part by NSF project IIS-1016772 and FCT project PTDC/EEA-CRO/122812/2010.

References

[1] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *CVPR*, 2012. 2, 7

[2] J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24:179–195, 1975. 7

[3] J. K. Bradley and C. Guestrin. Sample complexity of composite likelihood. 2012. 5

[4] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 11, 26

[5] J. Carreira, F. Li, and C. Sminchisescu. Object Recognition by Sequential Figure-Ground Ranking. *IJCV*, 2012. 2

[6] J. Carreira and C. Sminchisescu. Constrained parametric min cuts for automatic object segmentation. In *CVPR*, 2010. 11

[7] J. Carreira and C. Sminchisescu. CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts. *PAMI*, 2012. 2, 7, 11, 26

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 6, 11

[9] C. Dann, P. V. Gehler, S. Roth, and S. Nowozin. Pottics the potts topic model for semantic image segmentation. In *DAGM*, 2012. 6

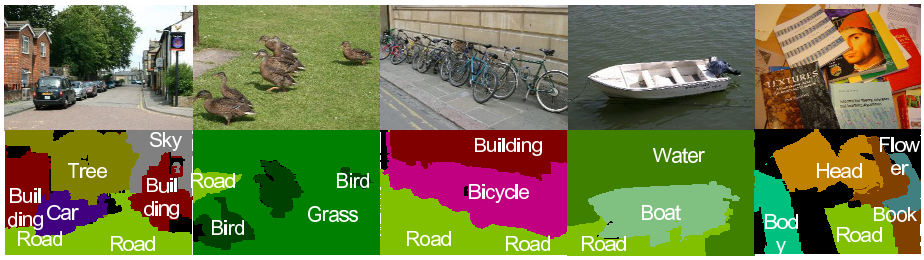

Figure 15: CSI results on MSRC. The ones with two `Roads` indicate that CSI have discovered two difference road objects in the scene. One can see that CSI successfully recovered a semantic depiction of the whole image with both object categories and background texture categories. The last two images show (partial) failure cases. The fourth image is a common failure case for texture categories, where CSI believed that the textured background (water) should be separated into two objects, and then misclassified one of them. In the fifth image, CSI tries to interpret each book as a different object, but ended up misclassifying most of them.

[10] J. V. Dillon and G. Lebanon. Stochastic composite likelihood. *J. Mach. Learn. Res.*, pages 2597–2633, 2010. 3, 5, 7, 8

[11] I. Endres and D. Hoiem. Category independent object proposals. In *ECCV*, pages 575–588, 2010. 11

[12] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2012. www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 23, 26

[13] T. Ferguson. *A Course in Large Sample Theory*. Chapman and Hall, 1996. 9

[14] J. Gonfaus, X. Boix, J. V. de Weijer, A. Bagdanov, J. Serrat, and J. Gonzalez. Harmony potentials for joint classification and segmentation. In *CVPR*, 2010. 2, 30, 31

[15] S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In *NIPS*, 2009. 2, 6

[16] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007. 7

[17] A. Ion, J. Carreira, and C. Sminchisescu. Probabilistic joint segmentation and labeling. In *NIPS*, 2011. 2, 6, 26

[18] J. Jiang and Z. Tu. Efficient scale space auto-context for image segmentation and labeling. In *CVPR*, 2009. 31

[19] P. Kohli and M. P. Kumar. Energy minimization for linear envelope mrfs. In *CVPR*, 2010. 4

[20] P. Kohli, L. Ladicky, and P. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, pages 1–8, 2008. 4

[21] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order mrfs. In *CVPR*, 2009. 5

[22] P. Krahenbuhl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 2, 30, 31

[23] M. P. Kumar, H. Turki, D. Preston, and D. Koller. Learning specific-class segmentation from diverse data. In *ICCV*, 2011. 2, 6

[24] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009. 5

[25] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. Torr. What,where and how many? combining object detectors and crfs. In *ECCV*, 2010. 2, 5, 30, 31, 32

[26] V. S. Lempitsky, P. Kohli, C. Rother, and T. Sharp. Image segmentation with a bounding box prior. In *ICCV*, 2009. 5

[27] F. Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *CVPR*, 2010. 2, 7, 11, 26

[28] Z. Li, E. Gavves, K. E. A. van de Sande, C. G. M. Snoek, and A. W. M. Smeulders. Codemaps segment, classify and search objects locally. In *ICCV*, 2013. 2

[29] P. Liang and M. I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *ICML*, 2008. 5

[30] Y. Lim, K. Jung, and P. Kohli. Energy minimization under constraints on label counts. In *ECCV*, 2010. 5

[31] B. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 1988. 2, 5, 7

[32] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 11

[33] T. Malisiewicz and A. A. Efros. Recognition by association via learning per-exemplar distances. In *CVPR*, 2008. 29

[34] C. Pantofaru, C. Schmid, and M. Hebert. Object recognition by integrating multiple image segmentations. In *ECCV*, 2008. 2, 7

[35] C. Rother, P. Kohli, W. Feng, and J. Jia. Minimizing sparse higher order energy functions of discrete variables. In *CVPR*, 2009. 5

[36] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 2, 30, 31

[37] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977. 3

[38] C. Varin and P. Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92, 2005. 5, 7

[39] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *CVPR*, 2001. 6

[40] T. Werner. High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimization (map-mrf. In *CVPR*, 2009. 5

[41] O. Woodford, C. Rother, and V. Kolmogorov. A global perspective on map inference for low-level vision. In *ICCV*, 2009. 4

[42] W. Xia, C. Domokos, J. Dong, L. F. Cheong, and S. Yan. Semantic segmentation without annotating segments. In *ICCV*, 2013. 2

[43] W. Xia, Z. Song, J. Feng, L.-F. Cheong, and S. Yan. Segmentation over detection by coupled global and local sparse representations. In *ECCV*, 2012. 2

[44] P. Yadollahpour, D. Batra, and G. Shakhnarovich. Discriminative re-ranking of diverse segmentations. In *CVPR*, 2013. 2

[45] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 5, 30, 31, 32