Dense Semantic Image Segmentation with Objects and Attributes

Shuai Zheng[1] Ming-Ming Cheng[1] Jonathan Warrell[2] Paul Sturgess[2]
Vibhav Vineet[2] Carsten Rother[3] Philip H. S. Torr[1]
[1]University of Oxford [2]Oxford Brookes University [3]TU Dresden
http://www.robots.ox.ac.uk/~tvg/ http://tu-dresden.de/inf/cvld

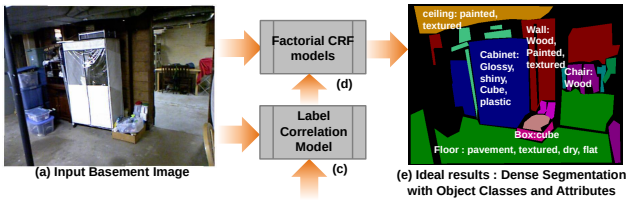
Abstract

*The concepts of objects and attributes are both important for describing images precisely, since verbal descriptions often contain both adjectives and nouns (e.g. 'I see a shiny red chair'). In this paper, we formulate the problem of joint visual attribute and object class image segmentation as a dense multi-labelling problem, where each pixel in an image can be associated with both an object-class and a set of visual attributes labels. In order to learn the label correlations, we adopt a boosting-based piecewise training approach with respect to the visual appearance and co-occurrence cues. We use a filtering-based mean-field approximation approach for efficient joint inference. Further, we develop a hierarchical model to incorporate region-level object and attribute information. Experiments on the a*PASCAL, CORE *and attribute augmented* NYU *indoor scenes datasets show that the proposed approach is able to achieve state-of-the-art results.*

1. Introduction

Using objects and attributes jointly provides a much more precise way to describe the content of a scene than using only one alone. *e.g.*, the image description *a shiny red chair* is more precise than the description *chair* on its own. Motivated by this fact, we introduce the problem of joint attribute-object image segmentation, where each image pixel is labelled with (i) an object label, such as car or road, (ii) visual attribute labels such as materials (wood, glass), and (iii) surface properties (shiny, glossy). We also make the distinction between things and stuff; where objects with a well defined shape and centroid are called things, and amorphous objects are refered to as stuff [13, 14, 21]. This problem is well suited for being solved in a joint hierarchical model, as the attributes can help with the object predictions and vice versa in both region and pixel levels.

In semantic image segmentation for object classes, existing approaches, *e.g.* [20, 31], treat the problem as a multi-class classification problem, where the goal is to associate each pixel with one of the object class labels. Recent works have also shown the advantages of using *visual attributes* [9, 11, 23, 29] and *relative visual attributes* [24] in object recognition, object localization [23, 27, 39], and scene classification [25, 40]. However, few of these works have been proposed to address the problem of dense image segmentation for things and stuff using attributes, and it is not yet clear whether visual attributes improve the performance of object segmentation.



	Objects	Attributes
Region-Level	Things(e.g. Chair, TV monitor, Box, etc.)	Shape (e.g. Cube) , Material, Surface ...
Pixel-Level	Things and stuff (e.g. Chair, TV monitor, Box, Wall, ceiling, floor, etc.)	Material (e.g. plastic, wood), Surface (e.g. glossy), ...

(b) Semantic Space

Figure 1. **Illustration of the proposed approach.** (a) shows the input image, a scene image from NYU dataset. (b) represents the semantic label space including pixel-level objects and attributes, region-level objects and region attributes. (e) shows conceptual ideal results for dense semantic segmentation with objects and attributes. Best view in color.

In this paper, we model scene images using a fully-connected multi-label conditional random field (CRF) with joint learning and inference. In our framework each image pixel is associated with both a set of attributes and a single object-class label, as illustrated in Fig. 1. In order to efficiently tackle the multi-labelling problem, we break it down into manageable multi-class and binary subproblems using a factorial CRF framework [15, 22, 34]. The structure of the factorial CRF we propose includes links between object and attribute factors that explicitly allow us to model correlations between these output variables. In order to handle the use of attributes at different levels, we also propose a

hierarchical model in which both objects and attributes are labelled at two levels, pixels and regions. Using the regions provided by the efficient object detector [1, 5, 10, 37] and the segmentation methods [2, 3, 4, 26], we can predict attributes such as shape, which apply to object instances as a whole. This allows us to deal with attributes both for objects of fixed spatial extent, *i.e*. things that can be detected with deformable part based detector (*e.g*. chair, etc) as well as amorphous objects (stuff), *i.e*. ones that are more ambiguous (*e.g*. floor, etc). Previous works [8, 9] have only focused on one of these forms and have not attempted to solve both types. To learn the correlations between factors we employ a boosting framework [28, 30] that exploits both the visual similarity and co-occurrence relations between object and attributes labels. This provides an effective piecewise learning strategy to train the model. To perform joint inference we use a mean field based algorithm [18, 38, 19]. This allows us to use a fully-connected graph topology for both object and attribute factor CRFs, whilst maintaining efficiency through filtering.

Our work is different from previous works [12, 35] in several ways. Both these approaches deal only with a very limited set of spatial attributes. While Tighe *et al*. [35] consider a region MRF with only adjacent pairwise connections, we propose a hierarchical model with both pixel and region levels, which is fully-connected at the pixel level. We also use mean-field inference rather than graph-cuts to handle the dense topology. Gould *et al*. [12] only consider pixel labelling for object classes and spatial attributes. In contrast, our approach can deal with a much more general problem. Furthermore, we also differ substantially from [6]. They have also considered the task of estimating objects and attributes in images. However the focus of that work is to analyse the use of verbal interactions, performed by the user, in order to verbally guide image editing. They have not explored a hierarchical formulation, as done in this work, which is important to achieve a higher level of accuracy. Also, they have not considered learning the attribute-object relationship using a boosting-based piecewise training.

Our contributions in this paper are as follows:

- We present an efficient hierarchical fully-connected multi-label CRF based framework, which involves assigning pixels with object class and attributes labels.
- We explore a piecewise boosting-based training strategy to learn the label correlations based on visual appearance similarity and label co-occurrence statistics.
- We augment the NYU dataset [32] with attribute labels (*attribute* NYU *dataset*, aNYU) to provide a benchmark to encourage alternative approaches.

2. Factorial Multi-Label CRF Model

We address the problem of joint semantic image segmentation for objects and attributes using a multi-label CRF, which we factor into multi-class and binary CRFs.

2.1. Multi-class CRF for Objects

We first review a general multi-class CRF model, which we will use as a factor in the joint model for the object classes, and which we generalize below to form the multi-label CRF for attribute labels. We define the CRF over a set of random variables, $\mathcal{X} = \{X_1, X_2, ..., X_n\}$, where each variable will take values from a set of *object labels*, $x_i \in \mathcal{O}$, where $\mathcal{O} = \{l_1, l_2, ..., l_k\}$. We denote by $\mathbf{x}$ a joint configuration of these random variables, and write $\mathbf{I}$ for the observed image data. The random field is defined over a graph $G(\mathcal{V}, \mathcal{E})$ with the i-th vertex being associated with a corresponding X_i and $(i, j) \in \mathcal{E}$ representing the i-th vertex and the jth vertex are connected by an edge. A pairwise multi-class CRF model can be defined in terms of an energy function:

$$E^{\mathcal{O}}(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i^{\mathcal{O}}(x_i) + \sum_{\{i,j\} \in \mathcal{E}} \psi_{ij}^{\mathcal{O}}(x_i, x_j), \qquad (1)$$

where $\psi_i^{\mathcal{O}}$ and $\psi_{ij}^{\mathcal{O}}$ are potential functions discussed below. The probability of a configuration $\mathbf{x}$ under the CRF distribution is found by normalizing the exponential of its negative energy, $P(\mathbf{x}|\mathbf{I}) \propto \exp(-E^{\mathcal{O}}(\mathbf{x}))$. Although it is generally computationally infeasible to calculate $P(\mathbf{x}|\mathbf{I})$ exactly due to the partition function, various approximate methods for inference exist, such as approximate *maximum a posteriori* methods (*e.g*. graph-cuts) which minimize Eq. 1, or variational methods, such as mean-field approximate $P(\mathbf{x}|\mathbf{I})$ [18], and allow us to approximately estimate a *maximum posteriori marginals* solution (MPM), $x_i^\star = \arg\max_l \sum_{\{\mathbf{x}'|x_i=l\}} P(\mathbf{x}')$.

Typical graph topologies for object class segmentation consider $\mathcal{V}$ to correspond to the pixels of an image, and $\mathcal{E}$ as a 4 or 8-connected neighborhood relation. Recently, mean-field inference methods have also made it possible to use a fully connected graph, where $\mathcal{E}$ connects every pair of pixels, i.e. $\mathcal{E} = \{(i, j)|i, j \in \mathcal{V}, i \neq j\}$ (see [18]) given certain forms of pairwise potential, and we shall follow this approach in our models. Further, a hierarchical topology may be used, as in [21], which is discussed below.

We set $\psi_i^{\mathcal{O}}(x_i) = -\log(\Pr(X_i = x_i))$, where the probability is derived from a discriminatively trained pixel classifier, TextonBoost [20, 31] [1]. The potential $\psi_{ij}^{\mathcal{O}}(x_i, x_j)$ takes the form of a Potts model:

$$\psi_{ij}^{\mathcal{O}}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ g(i, j) & \text{otherwise.} \end{cases} \qquad (2)$$

[1] TextonBoost in this paper means the unary potential in ALE library http://www.robots.ox.ac.uk/~phst/ale.htm.

For a fully connected graph topology as in [18] $g(i,j)$ is defined as:

$$g(i,j) = w^{(1)} \exp(-\frac{|p_i - p_j|^2}{2\theta_\mu^2} - \frac{I_i - I_j}{2\theta_\nu^2}) + w^{(2)} \exp(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}), \quad (3)$$

where p_i indicates the location of the ith pixel, I_i indicates the intensity of the ith pixel, and θ_μ,θ_ν, and θ_γ are the parameters.

2.2. Multi-label CRF for Attributes

We define a *multi-label* CRF for attributes similarly to the multi-class CRF above, but where the random variables take sets of labels instead of single labels. These sets represent the set of attributes present in a pixel. Formally, we have a set of random variables $\mathcal{Y} = \{Y_1, Y_2, ..., Y_n\}$, and a set of *attribute labels*, $\mathcal{A} = \{a_1, a_2, ..., a_m\}$. Rather than taking values directly in $\mathcal{A}$ though, the Y_i's take values in the *power-set* of the attributes, *i.e.* $y_i \in \mathcal{P}(\mathcal{A})$, where $\mathcal{P}$ is the power-set operator. As in the multi-class case, $\mathbf{y}$ is a joint assignment of these random variables. If we ignore the object labels for now, we can define a multi-label CRF distribution by an energy over $\mathcal{Y}$ as:

$$E^{\mathcal{A}}(\mathbf{y}) = \sum_{i \in \mathcal{V}} \psi_i^{\mathcal{A}}(y_i) + \sum_{\{i,j\} \in \mathcal{E}} \psi_{ij}^{\mathcal{A}}(y_i, y_j), \quad (4)$$

and we imply that $P(\mathbf{y}|\mathbf{I}) \propto \exp(-E^{\mathcal{A}}(\mathbf{y}))$. In general, since $|\mathcal{P}(\mathcal{A})|$ grows exponentially with $|\mathcal{A}|$, the number of parameters in $\psi_i^{\mathcal{A}}$ and $\psi_{ij}^{\mathcal{A}}$ will also grow exponentially if we allow arbitrary potential forms. Below, we describe how we factorize these terms, leading to a tractable model at inference time.

We express $\psi_i^{\mathcal{A}}(y_i)$ as follows:

$$\psi_i^{\mathcal{A}}(y_i) = \sum_a \psi_{i,a}^{\mathcal{A}}(y_{i,a}) + \sum_{a_1 \neq a_2} \psi_{i,a_1,a_2}^{\mathcal{A}}(y_{i,a_1}, y_{i,a_2}). \quad (5)$$

Here we use auxiliary binary indicator variables $y_{i,a}$, where $y_{i,a} = [a \in y_i]$ (where [.] is the Iverson bracket), which is 1 for a true condition and 0 otherwise (i.e. $y_{i,a}$ indicates whether attribute a is present in the set at pixel i). We set $\psi_{i,a}^{\mathcal{A}}(y_{i,a})$ based on the output of a probabilistic classifier, $\psi_{i,a}^{\mathcal{A}}(b) = -\log(\Pr(y_{i,a} = b))$, $b \in \{0, 1\}$. For this purpose, we train m independent binary TextonBoost classifiers [20], one for each attribute. Further, we set:

$$\psi_{i,a_1,a_2}^{\mathcal{A}}(y_{i,a_1}, y_{i,a_2}) = \begin{cases} 0 & \text{if } y_{i,a_1} = y_{i,a_2}, \\ R^{\mathcal{A}}(a_1, a_2) & \text{otherwise,} \end{cases} \quad (6)$$

where $R^{\mathcal{A}}(a_1, a_2) \in [-1, 1]$ is a learnt *correlation* between a_1 and a_2. Hence, for highly correlated attributes, we pay a high cost if their indicators do not match. We discuss how to learn $R^{\mathcal{A}}$ in Sec. 3.

We define $\psi_{i,j}^{\mathcal{A}}(y_i, y_j)$ as follows:

$$\psi_{i,j}^{\mathcal{A}}(y_i, y_j) = \sum_a \psi_{i,j,a}^{\mathcal{A}}(y_{i,a}, y_{j,a}). \quad (7)$$

Here, we define $\psi_{i,j,a}^{\mathcal{A}}$ as a Potts model over binary indicators:

$$\psi_{i,j,a}^{\mathcal{A}}(y_{i,a}, y_{j,a}) = \begin{cases} 0 & \text{if } y_{i,a} = y_{j,a}, \\ g(i,j) & \text{otherwise,} \end{cases} \quad (8)$$

where, as above, we take $g(i,j)$ as in Eq.3 for the fully connected model, allowing us to use filter-based inference.

2.3. Factorial CRF for Objects and Attributes

We now describe our combined CRF model for objects and attributes. We define the CRF over random variables $\mathcal{Z} = \{Z_1, Z_2, ..., Z_n\}$, where we take $Z_i = (X_i, Y_i)$, i.e. a combination of an object label and an attribute set. Hence, $z_i \in \mathcal{J} = \mathcal{O} \times \mathcal{P}(\mathcal{A})$, where we write $\mathcal{J}$ for joint label set. We then define a joint CRF in terms of a pairwise energy over the Z_i's as above:

$$E^{\mathcal{J}}(\mathbf{z}) = \sum_{i \in \mathcal{V}} \psi_i^{\mathcal{J}}(z_i) + \sum_{\{i,j\} \in \mathcal{E}} \psi_{ij}^{\mathcal{J}}(z_i, z_j), \quad (9)$$

and let $P(\mathbf{z}|\mathbf{I}) \propto \exp(-E^{\mathcal{J}}(\mathbf{z}))$.

Note that, equivalently, we could think of Eq. 9 as defining a single multi-label CRF over both object and attribute label sets, i.e. $z_i \in \mathcal{P}(\mathcal{O} \cup \mathcal{A})$. The factorization into multi-class object and multi-label attribute components makes the assumption that any configuration $\mathbf{z}$ has infinite energy (or zero probability) for some i and object labels $l_1 \neq l_2$, $l_1 \in z_i$ and $l_2 \in z_i$, or $l \notin z_i$ for all l. Indeed, it may be appropriate in certain cases to allow multiple object labels at each pixel, for instance if we have a semantic hierarchy including labels such as animal, mammal, dog etc., or a hierarchy of parts such as bicycle, wheel, spoke etc. In this case we would form a product of two multi-label CRF.

We define the joint unary potential as follows:

$$\psi_i^{\mathcal{J}}(z_i) = \psi_i^{\mathcal{O}}(x_i) + \psi_i^{\mathcal{A}}(y_i) + \sum_{l,a} \psi_{i,l,a}^{\mathcal{O}\mathcal{A}}(x_i, y_{i,a}), \quad (10)$$

where $\psi_i^{\mathcal{O}}$ and $\psi_i^{\mathcal{A}}$ are defined as above, and the final term takes the form:

$$\psi_{i,l,a}^{\mathcal{O}\mathcal{A}}(x_i, y_{i,a}) = \begin{cases} 0 & \text{if } y_{i,a} = [x_i = l] \\ R^{\mathcal{O}\mathcal{A}}(l, a) & \text{otherwise,} \end{cases} \quad (11)$$

where, as before $R^{\mathcal{O}\mathcal{A}}(l, a) \in [-1, 1]$ is a learnt *correlation* between l and a. The first condition in Eq. 11 is satisfied if $x_i = l$ holds, and $y_{i,a} = 1$ is also satisfied.

Our joint pairwise term simply combines the individual object and attribute pairwise terms:

$$\psi_{ij}^{\mathcal{J}}(z_i, z_j) = \psi_{ij}^{\mathcal{O}}(x_i, x_j) + \psi_{ij}^{\mathcal{A}}(y_i, y_j). \quad (12)$$

2.4. Hierarchical Model

In addition to a fully connected CRF over a pixel variable set, we also consider a two-level hierarchial model, where, in addition to labelling object classes and attributes at the *pixel* level, we also label objects and attributes at a *region* level, as shown in Fig. 2. We thus consider that our vertex set is partitioned into disjoint sets $\mathcal{V}_{\text{pix}}$ and $\mathcal{V}_{\text{reg}}$, each associated with its own set of attributes, $\mathcal{A}_{\text{pix}}$, $\mathcal{A}_{\text{reg}}$. We maintain dense connectivity over all variables at the pixel level, i.e. $(i,j) \in \mathcal{E}$ for all $i \neq j$ and $i,j \in \mathcal{V}_{\text{pix}}$. For each $j \in \mathcal{V}_{\text{reg}}$, we assume that we have a subset of pixels $\mathcal{S}_j \subset \mathcal{V}_{\text{pix}}$ (which represent the region), and that the edge set contains an edge joining each region variable to all the pixels in its subset, $(i,j) \in \mathcal{E}$ for all $i \in \mathcal{S}_j$. This gives rise to the energy:

$$\begin{aligned} E^{\mathcal{H}}(\mathbf{z}) &= \sum_{i \in \mathcal{V}_{\text{pix}}} \psi_i^{\mathcal{J}}(z_i) + \sum_{\substack{(i,j)\in\mathcal{E},\\ i,j\in\mathcal{V}_{\text{pix}}}} \psi_{ij}^{\mathcal{J}}(z_i,z_j) \\ &+ \sum_{i \in \mathcal{V}_{\text{reg}}} \psi_i^{\mathcal{J}'}(z_i) + \sum_{\substack{(i,j)\in\mathcal{E},\\ i\in\mathcal{V}_{\text{pix}}, j\in\mathcal{V}_{\text{reg}}}} \psi_{ij}^{\mathcal{J}'}(z_i,z_j), \quad (13) \end{aligned}$$

where we implicitly take $\psi_i^{\mathcal{J}}(z_i) = \infty$ if $a \in y_i$ with $i \in \mathcal{V}_{\text{pix}}$ and $a \in \mathcal{A}_{\text{reg}}$, and vice versa for region variables and object attributes.

Similar to [21], we use the efficient object detector [10, 5] and binary segmentation methods [4] to get regions $\mathcal{S}_j$. We thus assume that we have a proposed object class for each region, $o_j \in \mathcal{O}$, $j \in \mathcal{V}_{\text{reg}}$, and an associated score from the detector, s_j. Also, we train a classifier to produce probabilistic outputs for all attributes $\mathcal{A}_{\text{reg}}$ at the region level, and estimate a correlation matrix $R^{\mathcal{O}\mathcal{A}_{\text{reg}}}$ between objects and region level attributes. The joint unary terms of a region $\psi_i^{\mathcal{J}'}(z_i)$ then take the same form as Eq. 10, except that we set $\psi_i^{\mathcal{O}}(x_i) = 0$ for all x_i, and $\psi_{i,l,a}^{\mathcal{O}\mathcal{A}_{\text{reg}}}(x_i, y_{i,a}) = 0$ for all $x_i \neq o_i$. Our region-pixel pairwise terms take the form:

$$\psi_{ij}^{\mathcal{J}'}(z_i,z_j) = \begin{cases} -s_j & \text{if } x_i = o_j \text{ and } x_j = o_j \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

where, s_j is the score from the jth region associated object detector.

2.5. Inference in the Joint CRF

Following Krahenbuhl *et al.* [18], we adopt a mean field approximation approach for inference. This involves finding a mean field approximation $Q(\mathbf{z})$ that minimizes the KL-divergence $D(Q||P)$ among all distributions Q that can be expressed as a product of independent marginals, $Q(\mathbf{z}) = \prod_i Q_i(z_i)$. Given the form of our factorial model, we can factorize Q further into a product of marginals over multi-class object and binary attribute variables. Hence we take $Q_i(z_i) = Q_i^{\mathcal{O}}(x_i) \prod_a Q_{i,a}^{\mathcal{A}}(y_{i,a})$, where $Q_i^{\mathcal{O}}$ is a multi-class

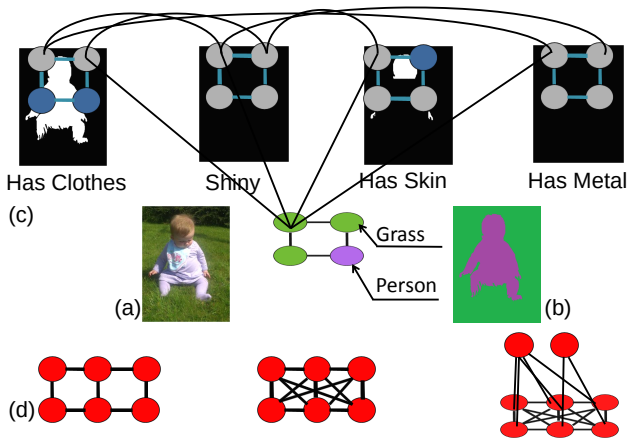


Figure 2. **Illustration of Factorial-CRF-based Semantic Segmentation for object classes and Attributes**. (a) shows the input image. (b) shows the ground truth mask image for object classes. (c) shows the attributes masks. (d) compares various CRF topologies including a grid CRF, a fully-connected CRF, and a hierarchial fully connected CRF. Best view in color.

distribution over the object labels, and $Q_{i,a}^{\mathcal{A}}$ is a binary distribution over $\{0,1\}$.

Given this factorization, we can express the required mean field updates (see [17]) for the non-hierarchical model as:

$$\begin{aligned} Q_i^{\mathcal{O}}(x_i = l) &= \frac{1}{Z_i^{\mathcal{O}}} \exp\{-\psi_i^{\mathcal{O}}(l) \\ &- \sum_{j \neq i} Q_j^{\mathcal{O}}(x_j = l)(-g(i,j)) \\ &- \sum_{a,b \in \{0,1\}} Q_{ja}^{\mathcal{A}}(y_{ja} = b)\psi_{i,x_i,a}^{\mathcal{O}\mathcal{A}}(l,b)\}, \end{aligned} \quad (15)$$

and

$$\begin{aligned} Q_{i,a}^{\mathcal{A}}(y_{i,a} = b) &= \frac{1}{Z_{ia}^{\mathcal{A}}} \exp\{-\psi_{ia}^{\mathcal{A}}(b) \\ &- \sum_{j \neq i} Q_{ja}^{\mathcal{A}}(y_{ja} = b)(-g(i,j)) - \\ &\sum_{a' \neq a, b' \in \{0,1\}} Q_{ia'}^{\mathcal{A}}(y_{ia'} = b')\psi_{i,a,a'}^{\mathcal{A}}(b,b') \\ &- \sum_{l} Q_i^{\mathcal{O}}(x_i = l)\psi_{i,l,a}^{\mathcal{O}\mathcal{A}}(l,b), \end{aligned} \quad (16)$$

where $Z_i^{\mathcal{O}}$ and $Z_{ia}^{\mathcal{A}}$ are per-pixel normalization factors, and $b \in \{0,1\}$. As in [18], we can efficiently evaluate the pairwise summations in Eq. 15 and Eq. 16 using $n+m$ Gaussian convolutions given that our pairwise factors take Potts forms as described. Updates for the hierarchical model take a similar form.

2.6. Learning parameters for the CRF

For the low-level feature descriptors (LBP, SIFT, HOG, Texton, Color SIFT), we fixed the parameters for the datasets according to the setting for the best results on PASCAL VOC 2010 dataset using AHCRF [20]. Regarding the parameters of the CRFs, we use cross-validation [16, 31] to learn the weights for the objects unary responses, attributes unary responses, pairwise, and region-level responses.

3. Label Correlation Discovery

In this section, we describe a piecewise method for training the label correlation matrices, $R^{\mathcal{A}}$, $R^{\mathcal{OA}}$ and $R^{\mathcal{OA}_{\text{reg}}}$ in the model described. We train all matrices simultaneously by learning an $(n+m)^2$ correlation strength matrix (hence treating the problem as a purely multi-label problem) and then extracting the relevant sub-matrices.

Specifically, we use the modified Adaboost framework of [28, 36] with multiple hypothesis reuse as described in [30]. In training, we denote by $\mathcal{D} = \{(\mathbf{f}_1, \bar{\mathbf{z}}_1), (\mathbf{f}_2, \bar{\mathbf{z}}_2), \ldots, (\mathbf{f}_N, \bar{\mathbf{z}}_N)\}$ a training dataset of N instances (i.e. pixels or regions), where $\mathbf{f}_i$ is a feature vector for the i-th instance derived from the image $\mathbf{I}$ (*e.g.* a bag of words vector) and $\bar{\mathbf{z}}_i = [\bar{\mathbf{x}}_i; \bar{\mathbf{y}}_i]$ is an indicator vector of length $n+m$, where $\bar{\mathbf{x}}_i(l) = 1$ implies object l is associated with instance i, and $\bar{\mathbf{x}}_i(l) = -1$ implies it is not, and similarly for $\bar{\mathbf{y}}_i(a) = 1$ for attribute a. $\bar{\mathbf{z}}_i$ is thus a vector representation of a set of objects/attributes present at i.

In the description below, we focus on deriving the attribute-attribute correlations, but the same approach is used for deriving object-attribute correlations. The boosting approach of [30] generates strong classifiers $H_{t,a}(\mathbf{f})$ for each attribute a and each round of boosting, $t = 1...T$. These strong classifiers have the form:

$$H_{t,a} = \sum_{t=1,\ldots,T} \alpha_{t,a} h_{t,a}(\mathbf{f}), \tag{17}$$

where $h_{t,a}$ are weak classifiers, and $\alpha_{t,a}$ are the non-negative weights set by the boosting algorithm. Further, the joint learning approach of [30] generates a sequence of *reuse weights* $\beta_{t,a_1}(H_{t-1,a_2})$ for each pair of attributes a_1, a_2 at each iteration t. These represent the weight given to the strong classifier for attribute a_2 in round $t-1$ in the classifier for a_1 at round t. Further, [30] show how these quantities can be used to estimate the label correlation by calculating:

$$\begin{aligned} R(a_1, a_2) \quad = \quad & \sum_{t=2...T} \alpha_{t,a_1}(\beta_{t,a_1}(H_{t-1,a_2}) \\ & -\beta_{t,a_1}(-H_{t-1,a_2})). \end{aligned} \tag{18}$$

Learning the correlations this way incorporates both information about visual appearance similarities and co-occurrence relationships between attributes and objects.

4. Datasets

We evaluate our approach using three datasets: the Attribute Pascal (aPASCAL) dataset [9], the Cross-category Object REcognition (CORE) dataset [8], and the NYU indoor V2 dataset [32]. In this paper we only use the RGB images from the NYU dataset.

aNYU Dataset. Our first set of experiments is on the RGB images from the NYU V2 dataset [32] [2]. As shown in Fig 3, we added 8 additional attribute labels, *i.e.Wood, Painted, Cotton, Glass, Glossy, Plastic, Shiny, and Textured*. We asked 3 annotators to assign material, surface property attributes on each segmentation ground truth region. We then adopted the majority votes from 3 workers as our 8 additional attribute labels. We call this extended dataset the attribute NYU (aNYU) dataset. This dataset has 1449 images collected from 28 different indoor scenes. In our experiments, we select 15 object classes and 8 attributes that have sufficient numbers of instances to train the unary potential. Further, we randomly split the dataset, into 725 images for the training set, 100 for the validation set, and 624 for the testing set.

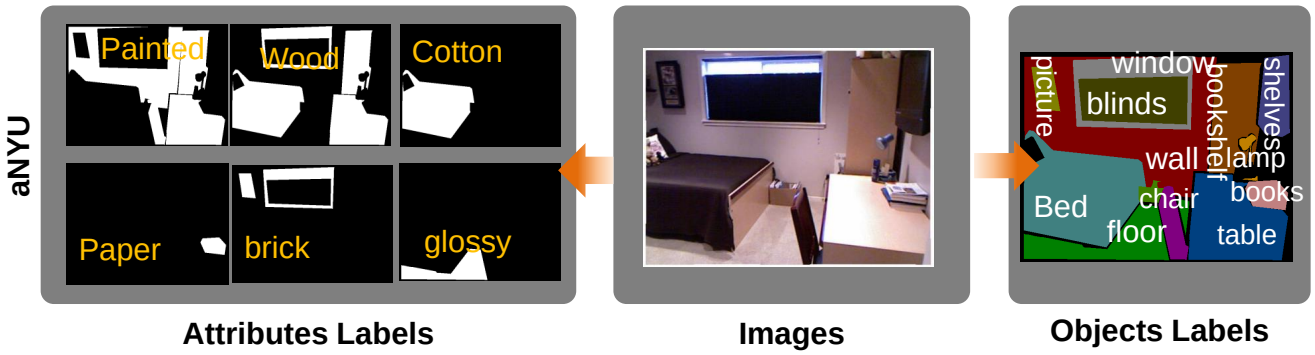
CORE Dataset. Our second set of experiments is conducted on the Cross-Category Object Recognition (CORE) dataset [8]. This dataset comes with 1059 images and ground truth segmentations for 27 object classes and 9 material attributes. The "objects" set has 27 labels, of which 14 are animals and 13 are vehicles. The "material" set contains nine different materials. Other images in the original CORE dataset are not used because they contain no pixel-level labels. In our experiments, we use 465 images to form the training set, and the remaining 594 images to form a test set. In the original CORE dataset experimental setting [8], some object classes have no training samples. Hence, we move some instances of those objects from test set to the training set.

aPASCAL Dataset. The existing aPASCAL dataset [9] is designed for bounding box level attributes. We transfer the existing 64 bounding-box-level attribute labels to our region-level attributes by finding the closest region segments from the image segmentation ground truth. We select 8 material attributes from 64 as pixel-level attributes, as other attributes are not well-defined on the pixel-level. Among the images in aPASCAL dataset, there are 639 having segmentation ground-truth annotation for both object classes and attributes. We use 313 for testing, and 326 for training.

5. Experiments

Our approach is a hierarchical fully-connected CRF model (HI). We compare our approach against the other

[2] http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html



Dataset	Object Labels		Pixel-level Labels		Region-level Labels	
	Number	Names	Number	Names	Number	Names
aNYU	15	Wall, Floor, Picture, Cabinet,...	8	Wood, Painted, Cotton, Glass,...	8	Wood, Painted, Cotton, Glass,...
CORE	27	Airplane, Alligator, Bat,...	9	Bare Metal, Feathers,...	9	Bare Metal, Feathers,..
aPASCAL	20	Aeroplane, Person, Bird, Cat,...	8	Skin, Metal, Plastic, Wood,...	64	2DBoxes, Round, Occluded,...

Figure 3. **Annotation illustration.** Extra annotation example and statistics on aNYU, CORE, and aPASCAL datasets. Best view in color.

state-of-the-art image segmentation approaches, including per pixel TextonBoost unary potential [20, 31] (Texton), Pairwise CRF semantic image segmentation approach (AHCRF [20]), Fully-connected CRF with detection and super-pixel higher orders (Full-C [18, 38]), and Joint attributes-objects Pixel-level fully-connected CRF (JP). JP has the same setting with the proposed approach, but the region-level terms are disabled. The problem of semantic image segmentation for attributes is a multi-label problem and these methods are not designed for dealing with it, so we we treat each as a binary one-vs-all label problem, with no pairwise terms between them, in contrast to our method in which we learn the important correlations between attributes. We also conduct experiments to understand the effect of each term in the proposed full model.

We choose the average intersection/union score as the evaluation measure. This measure is adopted from VOC [7], defined as TP / (TP + FP + FN). TP represents the true positive, and FP means false positive, and FN indicates the false negative. We compute the average intersection/union score across the attribute classes via summing up the intersection/union score for all the binary attribute segmentations and then dividing by the number of attributes.

We have conducted comprehensive evaluation on three datasets including aNYU, CORE, and aPASCAL. Compared with 5 other methods, we observe that HI outperforms the other approaches across all datasets, as illustrated in Fig. 4. In Fig. 4, HI achieves higher performance than JP, indicating that exchanging information between attributes and objects at both levels helps to predict both types of variable. Moreover, we observe a significant qualitative improvement, and we believe that a higher percentage increase would be archived if the datasets had more finely labelled data in the test set.

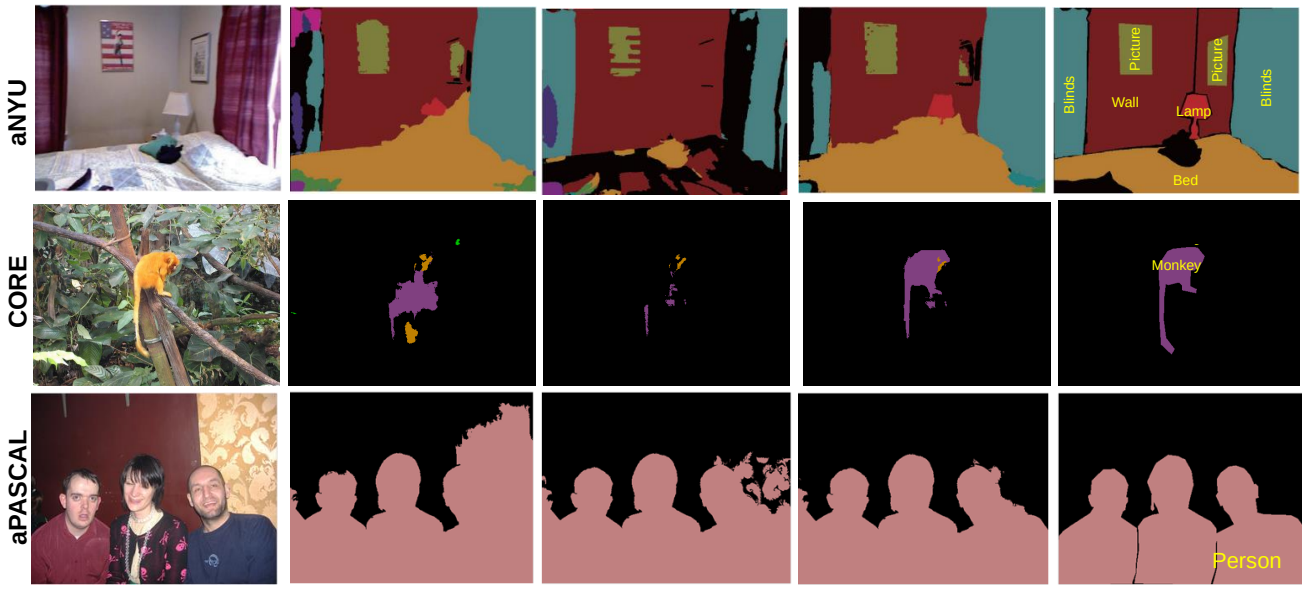
Effect of attribute terms. To clarify the effect of each attribute term in Eq. 13, we report the performance of object segmentation, using HI with different components being disabled. We take the learned models and remove, in turn, each type of attribute term (i.e. the joint attributes-objects term, the joint attributes-attributes term, the attributes in region level, and the attributes in pixel level), and report the performance in Table 1. When we remove the per-pixel attribute assignment, the object segmentation accuracy reduces by 5%, but when we remove the region-level attributes, the accuracy reduces by 4.4%. This suggests per-pixel attribute assignment is important to achieve higher accuracy and finer segmentation.

Dataset	Average label-accuracy(%) for object segmentation			
	full model	W/O pix-att	W/O region-att	W/O att
aNYU	61.4	56.4	57.0	51.3

Table 1. **Effect of different terms in our model.** We compare the average object label-accuracy(%) of our full model without (W/O) different components. "Full model" means the proposed approach, the hierarchical semantic image segmentation for both objects and attributes. "W/O pix-att" indicates the one without pixel-level attribute terms, "W/O region-att" represents the one without region-level attribute terms, and "W/O att" is the one without attribute terms.

In addition, to understand the potential of using attributes in helping semantic image segmentation, we evaluate the performance improvement of HI by setting the attribute factors to the ground truth labels (as if we had a perfect attribute CRF). Result shows 42% average label accuracy improvement on the object class segmentation, compared against the results of the proposed joint inference approach. This suggests that there is still great potential in using attributes towards semantic image segmentation.

Joint Inference Timings. All the experiments are carried out on a machine with a Intel Xeon E5 – 2687W



Dataset	Average intersection-union[7](%)									
	Object segmentation					Attribute segmentation				
	Texton [20, 31]	AHCRF [20]	Full-C [18, 38]	JP	HI	Texton [20, 31]	AHCRF [20]	Full-C [18, 38]	JP	HI
aNYU	17.4	20.0	18.9	20.8	**22.0**	8.90	10.1	10.0	14.4	**15.1**
CORE	17.5	17.6	17.5	19.1	**20.1**	15.6	17.0	17.4	17.5	**17.8**
aPASCAL	27.0	30.3	36.9	36.4	**37.1**	15.0	16.5	16.5	16.9	**17.6**

Figure 4. **Quanlitative and quantitative results.** Results on the aNYU, CORE [8] and aPASCAL [9] datasets. We compare 5 different approaches: TextonBoost classifier (Texton [20, 31]), Pairwise CRF with detection and super-pixel higher orders (AHCRF [20]), Fully-connected CRF with detection and super-pixel higher orders (Full-C [18, 38]), Joint Pixel-level CRF (JP), and Hierarchical CRF (HI). The results are reported as average intersection-union [7]. We obtain the attribute unary potentials with multiple binary segmentation, using the AHCRF [20] library. The attribute segmentation results for the method Full-C are obtained using Dense CRF inference based on these attribute unary potentials. Best view in color.

(3.1GHz, 1600MHz) and 64.0GB. For the hierarchical model, the straightforward implementation of the inference takes on average 11 seconds per image on the aNYU dataset, where the image size is 620×460. This inference can easily be parallelized. By enabling OpenMP and optimizing the implementation, the inference part can achieve 1.2 seconds per 620×460 image, on all 16 cores of the same machine. Further speed boost can be achieved with GPU implementation.

6. Conclusions and Future Work

In this paper, we have proposed a joint approach to simultaneously predict the attribute and object class labels for pixels and regions in a given image. The experiments suggest that combining information from attributes and objects at region and pixel-levels helps semantic image segmentation for both object classes and attributes. Further experiments also show that per-pixel attribute segmentation is important in achieving higher accuracy and finer semantic segmentation results. In order to encourage future work on the problem of semantic image segmentation with objects and attributes, we expand the aNYU dataset by adding per-pixel attribute annotation [3].

In the future work, we intend to consider allowing multi-label object predictions as well as attributes, and combining our piecewise learning approach to jointly learn all the parameters. We also plan to achieve the GPU implementation for the proposed approach and generalize current approach for 3D scenes understanding. It is possible to extend the set of object and attribute labels and maintain efficiency by following Sturgess *et al*. [33]. We will continue expanding the annotations and the data in the aNYU dataset.

Acknowledgments

We would like to thank all the anonymous reviewers for their valuable suggestions. This research is supported by the EPSRC (EP/I001107/1), ERC grant ERC-2012-AdG (321162-HELIOS). Philip H.S. Torr is in receipt of Royal Society Wolfson Research Merit

[3] http://www.robots.ox.ac.uk/~tvg/projects.php

Award.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. In *CVPR*, 2012. 2

[2] A. Blake, C. Rother, M. Brown, P. Perez, and P. H. S. Torr. Interactive image segmentation using an adaptive gmmrf model. In *ECCV*, 2004. 2

[3] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook. Efficient salient region detection with soft image abstraction. In *ICCV*, 2013. 2

[4] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011. 2, 4

[5] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, 2014. 2, 4

[6] M.-M. Cheng, S. Zheng, W.-Y. Lin, J. Warrell, V. Vineet, P. Sturgess, N. Crook, N. Mitra, and P. H. S. Torr. ImageSpirit: Verbal Guided Image Parsing. *ACM TOG*, 2014. 2

[7] M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2010. 6, 7

[8] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, 2010. 2, 5, 7

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 2, 5, 7

[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010. 2, 4

[11] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2007. 1

[12] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009. 2

[13] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008. 1

[14] B. Kim, M. Sun, P. Kohli, and S. Savarese. Relating things and stuff by high-order potential modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014. 1

[15] J. Kim and R. Zabih. Factorial markov random fields. In *ECCV*, 2002. 1

[16] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008. 5

[17] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009. 4

[18] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 2, 3, 4, 6, 7

[19] P. Krähenbühl and V. Koltun. Parameter learning and convergent inference for dense random fields. In *ICML*, 2013. 2

[20] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009. 1, 2, 3, 5, 6, 7

[21] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What,where & how many? combining object detectors and crfs. In *ECCV*, 2010. 1, 2, 4

[22] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001. 1

[23] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1

[24] D. Parikh and K. Grauman. Relative attributes. *ICCV*, 2011. 1

[25] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012. 1

[26] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. pages 309–314, 2004. 2

[27] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 1

[28] R. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 1999. 2, 5

[29] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Augmented attribute representations. In *ECCV*, 2012. 1

[30] Y. Y. Sheng-Jun Huang and Z.-H. Zhou. Multi-label hypothesis reuse. In *KDD*, 2012. 2, 5

[31] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multiclass object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81:2–23, 2009. 1, 2, 5, 6, 7

[32] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 5

[33] P. Sturgess, L. Ladicky, N. Crook, and P. H. S. Torr. Scalable cascade inference for semantic image segmentation. In *BMVC*, 2012. 7

[34] C. Sutton, K. Rohanimanesh, and A. McCallum. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In *ICML*, 2004. 1

[35] J. Tighe and S. Lazebnik. Understanding scenes on many levels. In *ICCV*, 2011. 2

[36] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: Efficient boosting procedures for multiclass object detection. In *CVPR*, 2004. 5

[37] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *IJCV*, pages 154–171, 2013. 2

[38] V. Vineet, J. Warrell, and P. H. S. Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. *IJCV*, 2014. 2, 6, 7

[39] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV*, 2010. 1

[40] J. Xiao, J. Hays, B. C. Russell, G. Patterson, K. A. Ehinger, A. Torralba, and A. Oliva. Basic level scene understanding: categories, attributes and structures. *Frontiers in psychology*, 2013. 1