Discriminative Re-ranking of Diverse Segmentations

Payman Yadollahpour
TTI-Chicago

Dhruv Batra
Virginia Tech

Gregory Shakhnarovich
TTI-Chicago

Abstract

This paper introduces a two-stage approach to semantic image segmentation. In the first stage a probabilistic model generates a set of diverse plausible *segmentations. In the second stage, a discriminatively trained re-ranking model selects the best segmentation from this set. The re-ranking stage can use much more complex features than what could be tractably used in the probabilistic model, allowing a better exploration of the solution space than possible by simply producing the most probable solution from the probabilistic model. While our proposed approach already achieves state-of-the-art results (48.1%) on the challenging VOC 2012 dataset, our machine and human analyses suggest that even larger gains are possible with such an approach.*

1. Introduction

Perception problems are hard. Consider the task of semantic segmentation[1] – *i.e.* recognizing and segmenting objects – in the picture shown in Fig. 1. A semantic segmentation algorithm must deal with tremendous amount of uncertainty – from inter and intra object occlusion and varying appearance, lighting & pose. Unfortunately, idealized models that reason about (the distribution over) all possible segmentations jointly with all confounding factors in a fully probabilistic setting are typically computationally intractable.

This results in a major formal divide – we can either build:

- **Restrictive Probabilistic Models** that reason about the full posterior and make a joint prediction over all variables of interest at the expense of performance-limiting independence assumptions, OR
- **Expressive Feed-Forward Models** that abandon the probabilistic joint-prediction framework altogether in favor of richer modelling but then mismanage uncertainty by only making feed-forward predictions.

In the context of semantic segmentation, the former class includes Conditional Random Field (CRF) models like [3, 17, 18] that make optimal (or provably approximate) joint predictions over all pixels, but require restrictive assumptions that might not be true in reality, like assuming that all interactions between variables are purely associative (or attractive) [18]. The latter class includes expressive feed-forward pipelines like [1, 5, 14] that first find regions in the image, score these region and then combine these scores into a segmentation. This feed-forward process captures rich dependencies between pixels and regions, but errors are accumulated and propagated from one stage to the next.

In this paper, we propose a hybrid approach that leverages the best of both worlds. We propose a two-stage

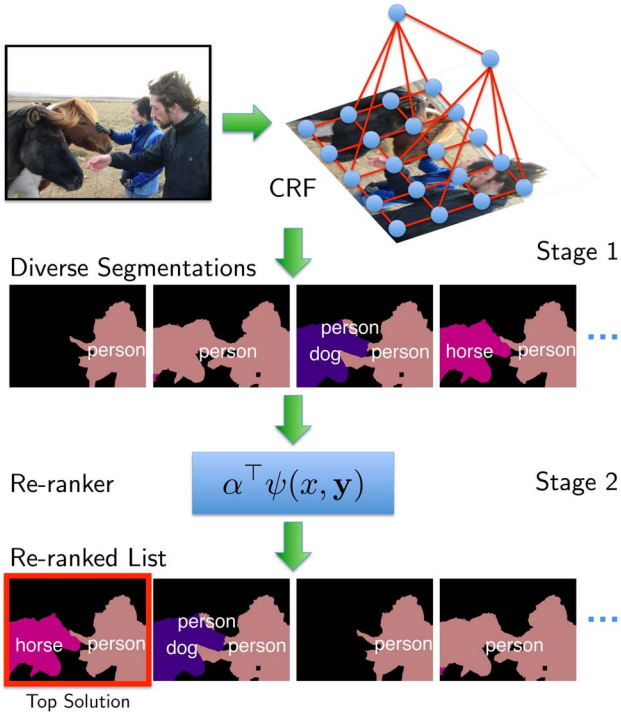


Figure 1: An overview of our *Div*MBEST+RERANK approach. In *Stage 1* diverse segmentations are computed from a tractable probabilistic model. These are fed to a large-margin re-ranker in *Stage 2*. The top re-ranked segmentation is returned as the final solution. Even though the most probable segmentation from Stage 1 is incorrect, the set of segmentations does contain an accurate solution, which the re-ranker is able to score to the top.

[1] We use 'semantic segmentation' or simply 'segmentation' to mean a labelling of an image, *i.e.* an assignment of a category label to each pixel.

model where the first stage is a tractable probabilistic model that reasons about an exponentially large output-space and makes a joint prediction – but crucially outputs a *diverse set of plausible segmentations*, not just a single one. The second stage in our approach is a discriminative re-ranker that is free to exploit arbitrarily complex features, and attempts to pick out the best segmentation from this set. We refer to this two-stage process as *Div*MBEST+RERANK. Fig. 1 illustrates the idea.

Thinking about semantic segmentation as a two-stage *Div*MBEST+RERANK process has several key advantages:

- **Global Optimization over a Simple Model.** The first stage of this approach is able to perform global optimization over all variables of interest, in a tractable albeit imperfect model to find a small set ($\simeq$ 10) of plausible hypotheses. We find that typically at least one of these solutions is *highly* accurate.
- **Rich (Higher-Order) Features in Re-ranker.** Since the re-ranker works with only a small set of segmentations, we do not need to worry about tractability when designing re-ranker features. The re-ranker is free to compute arbitrarily complex features that are not amenable to tractable inference and could not be added to the probabilistic model in the first stage. This is because the re-ranker does not need to *optimize* over all possible segmentations, it merely needs to *evaluate* these features on a small set of solutions.
- **Discrimination only within the Set.** The re-ranker can utilize feature that need not be *globally* discriminative, rather only *locally* discriminative within the set. Specifically, for the re-ranker the goal is no longer to use features than can identify generic good segmentations, rather to use features that can help it discriminate good solutions from bad ones *within* a small set.

Contributions. The key contribution of this paper is a novel approach to difficult perception problems with a *Div*MBEST+RERANK paradigm. While this paradigm is broadly applicable, we pick semantic segmentation as a case study in this paper.

Our main technical contribution is a discriminative re-ranking formulation for semantic segmentation. Our algorithm takes as input a set of labellings $\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}\}$ for an image and predicts the most accurate labelling from this list. We formulate this learning task as a Structured SVM (SSVM) [26], where the task loss penalizes the re-ranker for deviating from the most accurate solution in this set. In order to generate this set of segmentations, we build on our previous work [2], which produces diverse M-Best solutions from any probabilistic model. For the first stage of our approach, we analyze two different semantic segmentation probabilistic models – Automatic Labelling Environment (ALE) [18, 20] and Second Order Pooling (O_2P) [4] – and find that *Div*MBEST+RERANK results in significant improvements for both of them. Specifically, we achieve state-of-art performance in the challenging Pascal Visual Object Classes (VOC) 2012 segmentation dataset [12].

Fundamentally, we believe our work brings into focus a different way of thinking about difficult perception problems. Instead of attempting to directly answer the completely general question: 'What makes a good segmentation of an image?', perhaps more progress can be made by answering a simpler question – 'Given two plausible segmentations for an image, can we tell a good segmentation from a bad segmentation?' Our discriminative re-ranking algorithm is a first step towards answering the latter question. We also perform human and machine diagnoses to analyze the inherent difficulty of this task. From the human analyses, we find that people are surprisingly good at picking a good segmentation from a bad segmentation *by looking at the segmentations alone*. From our machine analyses, we find that significant gains in performance are possible by developing algorithms for this "pick-one-out-of-10" task. While our proposed ranker already achieves state of the art, even larger gains in performance are possible.

2. Related Work

At a high-level, the core set of ideas in our approach have been around for a long time [7]. It is common working wisdom to "delay making a hard decision till the last step of the pipeline". Our two-stage approach can be considered an instantiation of the same principle.

Relation to cascades. The idea of pruning possible solutions in successive stages has been central to many vision systems, including the seminal cascaded architecture of Viola and Jones [28] and more recent work [24, 29]. These techniques can be thought of as *deep* cascades, consisting of many (weak) stages, each incrementally pruning away some part of the search space. Our approach on the other hand is a *shallow* cascade, with a powerful first stage that performs an *exponentially large* pruning: from all possible segmentations to a small list of size M ($\simeq$ 10). Our results suggest that this is an effective approach since the first stage is computationally efficient and successful at producing a very small list with at least one high-quality solution.

Relation to proposal-generation works. Category-independent segmentation has long been thought of as a preprocessing stage for higher vision tasks. In this spirit, a number of works [6, 11, 16, 23] produce *intermediate* proposals, *i.e.* a pool of category-independent segments on which recognition tasks can be performed. Segment proposals can be combined & labelled in exponentially many ways. In contrast, stage 1 in our work produces *holistic* proposals, *i.e.* a small set of ($\simeq$ 10) complete labellings.

Multiple solutions and diversity. Stage 1 of our approach is related to a problem studied in the graphical mod-

els literature called M-Best MAP [13, 22, 30], which involves finding the top M most probable solutions in a probabilistic model. Unfortunately, since there is no emphasis on diversity, such solutions are typically minor perturbations of each other. This paper builds on our recent work, called *Div*MBEST [2], which produces *diverse* M-Best solutions. Diversity in solutions is crucial in re-ranking because we don't want to pick from a set of solutions that are simply minor perturbations of each other but rather ones that present whole alternative explanations. Our previous work [2] mostly focused on interactive applications where these diverse M-Best solutions could simply be shown to a user/expert. The key contribution of this paper is the automatic re-ranking of these multiple solutions.

Discriminative re-ranking in other domains. Discriminative re-ranking of multiple solutions is a dominant paradigm in domains like speech [9, 10] and natural language processing [8, 25]. In fact, the title of this paper is a reference to such speech & NLP papers. To the best of our knowledge, this is the first application of this paradigm for vision tasks.

3. Approach

We now describe our proposed two-stage *Div*MBEST+RERANK approach. Recall that the first stage is a Conditional Random Field (CRF) that produces a diverse set of segmentations and the second stage re-ranks this set and then picks the top scoring segmentation.

Notation. For any positive integer n, let $[n]$ be shorthand for the set $\{1, 2, \ldots, n\}$. The input to our system is a training dataset of (image, ground-truth segmentation) pairs $\{(x_i, \mathbf{y}_i^{gt}) \mid i \in [n]\}$, where x_i is the i^{th}-image and $\mathbf{y}_i^{gt}$ is the corresponding ground-truth segmentation. A segmentation $\mathbf{y}$ is a set of discrete random variables, representing the category assigned to each labelling unit (pixel or superpixel or region), *i.e.* $\mathbf{y} = \{y_1 \ldots y_k\}$. Each variable y_u can take value in a finite label set, *e.g.* $y_u \in \mathcal{Y}_u = \{\text{aeroplane, bicycle, bird, bottle, car, } \ldots\}$.

The quality of the predicted segmentation is measured by a loss function $\ell(\mathbf{y}_i^{gt}, \hat{\mathbf{y}})$ that denotes the cost of predicting $\hat{\mathbf{y}}$ when the ground-truth is $\mathbf{y}_i^{gt}$. In Pascal VOC [12], this loss would be the standard $1 - \frac{\text{intersection}}{\text{union}}$ measure, averaged over masks of all categories.

Stage 1: Producing Diverse Segmentations

Let us first describe how we generate multiple segmentations from the CRF in stage 1. For ease of notation, this subsection is described for a single training image pair $(x, \mathbf{y})$.

CRF Model. Let $G = (\mathcal{V}, \mathcal{E})$ be a graph defined over the segmentation variables $\mathbf{y}$, *i.e.* $\mathcal{V} = [k]$, $\mathcal{E} \subseteq \binom{[k]}{2}$. Each variable y_u corresponds to vertex u, and two vertices (u, v) are connected by an edge if the pixels/superpixels u and v are adjacent in the image.

Let $\theta_u(y_u)$ be the unary term expressing the local confidence for label y_u, and $\theta_{uv}(y_u, y_v)$ be the pairwise term expressing compatibility of labels y_u and y_v at adjacent vertices. The score for any segmentation $\mathbf{y}$ is given by the sum $S(\mathbf{y}) = \sum_{u \in \mathcal{V}} \theta_u(y_u) + \sum_{(u,v) \in \mathcal{E}} \theta_{uv}(y_u, y_v)$, and its probability is given by the Gibbs distribution: $P(\mathbf{y}) = \frac{1}{\mathcal{Z}} e^{S(\mathbf{y})}$, where $\mathcal{Z}$ is the partition function.

These unary and pairwise terms are typically derived from a weighted combination of features extracted at vertices and edges, *i.e.*, $\theta_u(y_u) = w_u^\intercal \phi(x, y_u)$ (and θ_{uv} is defined analogously). The weights w_u, w_{uv} are typically learnt from data or sometimes set by hand.

MAP Segmentation. The goal of MAP inference is to find the highest scoring labeling, *i.e.* $\operatorname{argmax}_{\mathbf{y}} S(\mathbf{y})$.

Diverse M-Best Segmentations. To generate a set of segmentations, we utilize our previous work called *Div*MBEST [2], which produces diverse M-Best solutions from any probabilistic model that allows for efficient MAP computation. For the sake of completeness, we give a brief overview of *Div*MBEST below; details can be found in [2].

*Div*MBEST finds diverse M-best solutions incrementally. Let $\mathbf{y}^1$ be the best solution (or MAP), $\mathbf{y}^2$ be the second solution found and so on. At each step, the next best solution is defined as the highest scoring state with a minimum degree of "dissimilarity" w.r.t. previously chosen solutions, where dissimilarity is measured under a function $\Delta(\cdot, \cdot)$:

$$\mathbf{y}^{(M)} = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} \sum_{u \in \mathcal{V}} \theta_u(y_u) + \sum_{(u,v) \in \mathcal{E}} \theta_{uv}(y_u, y_v) \quad \text{(1a)}$$

$$s.t. \quad \Delta(\mathbf{y}, \mathbf{y}^{(m)}) \geq k_m \quad \forall m \in [M-1]. \quad \text{(1b)}$$

In general, this problem is NP-hard and Batra *et al.* [2] proposed to use the Lagrangian relaxation formed by dualizing the dissimilarity constraints $\Delta(\mathbf{y}, \mathbf{y}^{(m)}) \geq k_m$:

$$f(\boldsymbol{\lambda}) = \max_{\mathbf{y} \in \mathcal{Y}} \; S_\Delta(\mathbf{y}) \doteq \sum_{u \in \mathcal{V}} \theta_u(y_u) + \sum_{(u,v) \in \mathcal{E}} \theta_{uv}(y_u, y_v) + \sum_{m=1}^{M-1} \lambda_m \left(\Delta(\mathbf{y}, \mathbf{y}^{(m)}) - k_m \right) \quad \text{(2)}$$

Here $\boldsymbol{\lambda} = \{\lambda_m \mid m \in [M-1]\}$ is the set of Lagrange multipliers, which determine the weight of the penalty imposed for violating the diversity constraints.

Following [2], we use Hamming diversity, *i.e.* $\Delta(\mathbf{y}, \mathbf{y}^{(m)}) = \sum_{u \in \mathcal{V}} [\![y_u \neq y_u^{(m)}]\!]$, where $[\![\cdot]\!]$ is 1 if the input condition is true, and 0 otherwise. This function counts the number of nodes that are labeled differently between two solutions. For Hamming dissimilarity, the

Δ-augmented scoring function (2) can be written as:

$$S_\Delta(\mathbf{y}) = \sum_{u\in\mathcal{V}} \underbrace{\left(\boldsymbol{\theta}_u(y_u) + \sum_{m=1}^{M-1} \lambda_m [\![y_u \neq y_u^{(m)}]\!]\right)}_{\text{Perturbed Unary Score}} + \sum_{(u,v)\in\mathcal{E}} \boldsymbol{\theta}_{uv}(y_{uv}). \quad (3)$$

Thus, the maximization in (2) can be performed simply by feeding a perturbed unary term to the algorithm used for MAP inference (*e.g.* α-expansion or TRW-S). This makes it really efficient to produce *Div*MBEST solutions in stage 1.

Stage 2: Re-ranking Diverse Segmentations

We now describe our proposed approach for re-ranking the diverse set of segmentations produced by stage 1.

Let $\mathbf{Y}_i = \{\mathbf{y}_i^{(1)} \dots \mathbf{y}_i^{(M)}\}$ denote the set of M segmentations for image i. The input to stage 2 at train-time is a set of (image, ground-truth, segmentation-set) triplets: $\{x_i, \mathbf{y}_i^{gt}, \mathbf{Y}_i \mid i \in [n]\}$. Note that the ground-truth $\mathbf{y}_i^{gt}$ typically will not be a part of the segmentation-set $\mathbf{Y}_i$. Let $\mathbf{y}_i^*$ denote the most accurate segmentation in the set, *i.e.* $\mathbf{y}_i^* = \operatorname{argmin}_{\mathbf{y}\in\mathbf{Y}_i} \ell(\mathbf{y}_i^{gt}, \mathbf{y})$. The accuracy of solution $\mathbf{y}_i^*$ forms an upper-bound on the re-ranker performance since we are committed to picking one solution from $\mathbf{Y}_i$. We refer to this as the *oracle* accuracy.

Re-ranker Model. The goal of the re-ranker is to predict the best segmentation in the set. We formulate this problem as a Structured SVM (SSVM) [26]. The re-ranker assigns a score to each segmentation in the set, *i.e.* $S_r(\mathbf{y}) = \boldsymbol{\alpha}^\intercal \boldsymbol{\psi}(x, \mathbf{y})$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}(x, \mathbf{y})$ are the re-ranker parameters and features respectively. Inference in the re-ranker consists of finding the highest scoring solution, $\hat{\mathbf{y}}_i = \operatorname{argmax}_{\mathbf{y}\in\mathbf{Y}_i} S_r(\mathbf{y})$.

The re-ranking features $\boldsymbol{\psi}$ need not be the same as the CRF features ϕ, and can be quite complex, because inference in the re-ranker merely involves extracting the features on a small set of solutions, taking a dot-product with the weights and sorting according to the resulting score. Also notice that the features are a function of both the image x_i *and* the segmentation $\mathbf{y}_i$. Thus, we can compute features like size of various categories, connectivity of the label masks, relative location of label masks and other such quantities that are functions of *global* statistics of the segmentation and thus intractable to include in the first stage. We describe the features used in detail in Section 5.

Re-ranker Loss. In order to measure re-ranker performance, we use a re-ranker loss $\mathcal{L}(\mathbf{y}_i^{gt}, \hat{\mathbf{y}}_i)$, which is different from the task loss ℓ. Specifically, we use the *relative* loss:

$$\mathcal{L}(\mathbf{y}_i^{gt}, \hat{\mathbf{y}}_i) = \ell(\mathbf{y}_i^{gt}, \hat{\mathbf{y}}_i) - \ell(\mathbf{y}_i^{gt}, \mathbf{y}_i^*),$$

i.e. the task loss of segmentation $\hat{\mathbf{y}}_i$ relative to the best segmentation in this set $\mathbf{y}_i^*$.

This forces the re-ranker to focus its effort on training instances where it is under-performing *relative* to the set and not globally. For instance, consider two images i, j with two segmentations each, whose accuracies are $Acc(\mathbf{Y}_i) = \{95\%, 75\%\}$ and $Acc(\mathbf{Y}_j) = \{40\%, 35\%\}$ respectively. If we use the task loss as the re-ranking loss, *i.e.*, $\mathcal{L}(\mathbf{y}_i^{gt}, \mathbf{y}) = \ell(\mathbf{y}_i^{gt}, \mathbf{y})$, the re-ranker will focus its attention on set j and ignore set i because both solutions in $\mathbf{Y}_j$ have high loss w.r.t. ground-truth $\{100\text{-}40\%, 100\text{-}35\%\} = \{60\%, 65\%\}$, while both solutions in $\mathbf{Y}_i$ have significantly lower loss $\{5\%, 25\%\}$. This is not desirable because j is already performing close to the best it can, given that we are committed to the set. Set i, on the other hand, has significant room for improvement. Using the relative loss correctly shifts the focus to i because an incorrect choice in that set is much costlier (difference of 20%) than an incorrect choice in set j (difference of 5%). Empirically, we found this choice to play an important role in the performance of the re-ranker.

Re-ranker Training. We learn the re-ranker parameters by solving the following Structured SVM Quadratic Program:

$$\min_{\boldsymbol{\alpha},\xi_i} \quad \|\boldsymbol{\alpha}\|_2^2 + C\sum_{i\in[n]} \xi_i \quad (4a)$$

$$s.t. \quad \boldsymbol{\alpha}^\intercal\Big(\boldsymbol{\psi}(x_i, \mathbf{y}_i^*) - \boldsymbol{\psi}(x_i, \mathbf{y})\Big) \geq 1 - \frac{\xi_i}{\mathcal{L}(\mathbf{y}_i^{gt}, \mathbf{y})} \quad (4b)$$

$$\xi_i \geq 0 \qquad \forall \mathbf{y} \in \mathbf{Y}_i \setminus \mathbf{y}_i^*, \quad (4c)$$

Intuitively, we can see that the constraint (4b) tries to maximize the (soft) margin between the score of the oracle solution and all other solutions in the set. Importantly, the slack (or violation in the margin) is scaled by the loss of the solution. Thus if in addition to $\mathbf{y}_i^*$ there are other good solutions in the set, the margin for such solutions will not be tightly enforced. On the other hand, the margin between $\mathbf{y}_i^*$ and bad solutions will be very strictly enforced. We solve (4) via the 1-slack cutting-plane algorithm of Joachims [15].

At test-time, the evaluation of our *Div*MBEST+RERANK pipeline is simple – we compute stage 1 CRF features ϕ on an image and parameters θ_u, θ_{uv}, run the *Div*MBEST algorithm [2] to produce a set of diverse segmentations $\mathbf{Y}$, compute re-ranker features $\boldsymbol{\psi}$ on this set and output the highest scoring solution.

We now provide a detailed analysis of both stages of our *Div*MBEST+RERANK approach – Section 4 analyzes stage 1 and Section 5 analyzes stage 2.

4. Analyzing Diverse Segmentations

In this section, we provide details of the CRFs used to produce multiple segmentations and characterize the diversity achieved in these segmentations. Specifically, we investigate the sources of diversity, and attempt to quantify the extent to which diversity enables potential gain in accuracy over the MAP solution.

4.1. CRFs: ALE and O_2P

We used two different models for semantic segmentation – the Associative Hierarchical CRF [18] (implemented as the Automatic Labeling Environment, ALE) and the Second-Order Pooling (O_2P) model of Carreira *et al*. [4]. Both models have publicly available implementations.

Both of these models incorporate high-order information in the image, albeit in different ways. ALE defines cliques over pixels and superpixels, and incorporates many different potentials such as unary potentials based on textonboost features, P^n Potts terms between pixels and superpixels and a global co-occurrence potential [19]. Much of the complex dependency between regions of the image is captured by the graph structure of the CRF and high-order cliques. In contrast, O_2P incorporates high-order dependencies between regions[2] (not pixels) in the image using second-order pooling of local descriptors such as SIFT and local binary patterns (LBP) to form global region descriptors.

For both models, *Div*MBEST is able to reuse the respective MAP inference algorithms to produce a diverse set of segmentations. Note that inference in O_2P was originally proposed as a greedy procedure. In our work, we formulate their approach as a CRF constructed on overlapping CPMC segments in the image.

4.2. Diversity and Oracles

For the analysis reported in this subsection, we used the VOC 2012 `train` and `val` sets. ALE and O_2P models were trained on VOC2012 `train`, and the models were used to produce 10 segmentations for each image in `val`. Following [2], we tuned the Lagrangian multipliers via cross-val ($\lambda_{ALE} = 1.25$ and $\lambda_{O_2P} = 0.08$).

Oracle Accuracies. Since ground-truth is known for VOC `val` images, we can find the oracle accuracy, *i.e*. the accuracy of the best solution in the set. This accuracy, shown in Fig. 3 (lines with circles), is striking: with only 10 solution on O_2P, it reaches 60.12%, which is 15%-points higher the accuracy of MAP! O_2P MAP is approximately the same as the winning entry in the 2012 Pascal VOC challenge. Winning accuracies typically improve 3-4%-points each year, so this is a *significant* potential. Oracle accuracy with ALE solutions show a similar increase.

To put these oracle numbers in context, we ask what is the the *best possible* segmentation that could be constructed with the 150 CPMC segments. We implemented a greedy algorithm that tries to find the subset of CPMC segments that best cover ground-truth segments and then simply copies labels over from the ground-truth. This achieves an accuracy of 80.78%. Notice that this procedure takes the supremum of accuracy of *exponentially* many solutions, and we only need 10 *Div*MBEST solutions to reach to 60.12%.

Diversity of solutions. We now turn to empirical analysis that quantifies the amount of diversity in these solutions, and how that affects the oracle performance. Details can be found in the supplementary material. Here we briefly describe the methodology and outline the main conclusions.

The first question we address is: how much diversity do the *Div*MBEST solutions contain over MAP? To answer this, we can look at the solution in the set that is most different from MAP, as measured by average region overlap. With just 10 solutions, this number drops to about 0.3 for O_2P and 0.1 for ALE. Thus, on average at least one out of 10 *Div*MBEST solutions for O_2P overlaps MAP by only 10%.

Of course, diversity is useful only if it brings in improved quality, and our next goal is to assess this. We computed the covering of MAP by the oracle for every image, and found that on average this covering is less than 61% for O_2P and 55% for ALE. If we constrain the covering to be category-consistent, these numbers drop to 58% and 45% respectively. Thus, we can conclude that the oracle segmentations are not simply minor perturbations of the MAP.

Gain from diversity. Diving a bit deeper we can investigate the *modes* of this diversity: how is the oracle different from the MAP? The analysis above tells us that the set of regions in the oracle tends to be very different from MAP. But perhaps the oracle simply contains a better set of masks for the same categories present in the MAP? We show in the supplementary materials that this is not the case: if we find the best labeling of any of the *Div*MBEST solutions restricted to the set of categories present in MAP, we obtain performance significantly inferior to that of the true oracle.

Thus, we can conclude that there are clear differences in both the labels and segments of the oracle segmentations compared to the MAP.

5. Re-ranking experiments

We now describe implementation details of our re-ranker, and report the results on VOC2012 `val` and `test` sets.

5.1. Re-ranker features and training

Our re-ranker uses a number of features that we separate into a few groups. In the discussion, below we say a label c is present in $\mathbf{y}$ if at least one pixel in $\mathbf{y}$ is labeled c.

Model features rely on properties derived from the model that produced $\mathbf{y}$– model score of $\mathbf{y}$, average pixel score, number of CPMC masks used to construct foreground, the final background threshold at the end of the greedy foreground assembly, and the rank of $\mathbf{y}$ among the M diverse hypotheses for the given input image. (5 dimensions)

Diversity features measure average per pixel agreement of $\mathbf{y}$ with the majority vote by the diverse set (weighted or

[2]Specifically, O_2P uses 150 segments generated by Constrained Parametric Min-Cuts (CPMC) algorithm [6].

unweighted by the model scores). (2 dimensions)

Recognition features. We use outputs of object detectors from [21] to get detector-based segmentations $\mathcal{D}_1, \mathcal{D}_2$, where each pixel is assigned by majority vote on detection scores (thresholded & un-thresholded). Then we compute the agreement matrix: for every c_1, c_2 we count pixels assigned to c_1 by $\mathbf{y}$ and to c_2 by $\mathcal{D}_1$, yielding a 441-dimensional feature. We compute max/median/min of the detection score (with and without thresholding) for every category in $\mathbf{y}$ (120 dims); the average overlap between category masks in $\mathcal{D}_1, \mathcal{D}_2$ and in $\mathbf{y}$ (2 dims); and pixelwise average detector scores for categories in $\mathbf{y}$ (2 dims). We also use the the estimated posterior for each category present in $\mathbf{y}$, using the classifier from [27] (20 dimensions).

Segment features measure the geometric properties of the segments in $\mathbf{y}$: perimeter, area, and the ratio of the two; computed separately for segments in every class and for the entire foreground (63 dimensions). Relative location of the centroids of masks for each category pair (420 dimensions).

Label features rely on information regarding the labels assigned to masks in $\mathbf{y}$, but not the geometry of these masks. For every pair of labels c_1, c_2 we compute the binary co-occurrence (1 if both categories are present in $\mathbf{y}$) and the percentage of pixels assigned to c_1 & c_2. (420 dimensions)

All the features above are independent of the image x; the following features rely on image measurements as well as properties of the solution $\mathbf{y}$.

Boundary features. We compute the total gPb probability of boundary response in a band along the category boundaries; for 3 widths of the band, this produces a 3-dimensional feature (with 3 more for normalized versions). We also compute recall by the gPb map of the category boundaries in the $\mathbf{y}$; this produces a 10 dimensional feature for ten equally spaced precision values. Finally, we compute the histogram (6 bins) of Chamfer distance between the boundaries in $\mathbf{y}$ and the thresholded gPb, and vice versa; with 10 thresholds this produces a 120 dimensional feature. For each category, we also computed normalized histogram of gPb responses in the non-boundary regions (210 dims).

Entropy features. For every category (and the combined foreground) we measure the entropy of color histograms, computed per color channel with two binning resolutions, yielding 126 dimensions. We do the same for textons, with a single binning, for another 21 features.

We stress that most of these features rely on higher-order information that would be intractable to incorporate into the CRF model used in stage 1. For instance, using features that refer to segment boundaries is hard in CRF. However, evaluating these features on M segmentations is easy, which allows us to use them at the re-ranking stage.

Training the re-ranker. The combined feature vector per solution $\mathbf{y}$ has 1988 dimensions. The only hyper-parameter for the re-ranker is the regularization parameter C (4a), which is chosen via cross-validation on the `val` set[3].

Re-ranker Results. In Fig. 3 we show the *Div*MBEST+RERANK results (`Re-rank`) for ALE and O_2P models on VOC 2012 `val`. As a baseline, we report the results of a binary classifier (`Classifier`) that is trained to discriminate between the best and the worst segmentations, and used at test-time to re-rank via the classification score. We also compared against randomly selecting one-out-of-M solutions (`Rand`).

For VOC 2012 validation *Div*MBEST+RERANK achieves an accuracy of 29.27% on ALE and 48.2% on O_2P. This is in contrast to a MAP accuracy of 24.3% for ALE and 45.1% for O_2P, which is an increase of more than 5% and 3% points respectively. The O_2P re-ranker, trained on the entire `val` set, was applied to VOC 2012 `test` with results summarized in Tbl. 1. *Div*MBEST+RERANK outperforms O_2P-MAP by 1.6% points and achieves state-of-the-art performance on this challenging dataset. In Fig. 5 we show a few examples where *Div*MBEST+RERANK outperforms O_2P-MAP.

Re-ranker Behavior Fig. 2 shows the behavior of the re-ranker on VOC 2012 `val`: (a) shows the number of images in which the oracle solution was originally at rank M. We can see that there is a heavy tail in the distribution, indicating that high-quality solutions are often found near the bottom of the list; (b) shows the number of images where the re-ranker predicts solution M. We can see a much lighter tail, suggesting that the re-ranker 'plays it safe' and predicts MAP very frequently; (c) shows a scatter plot of re-ranker score vs solution accuracy. We can see that the re-ranker score is quite well correlated with the solution quality.

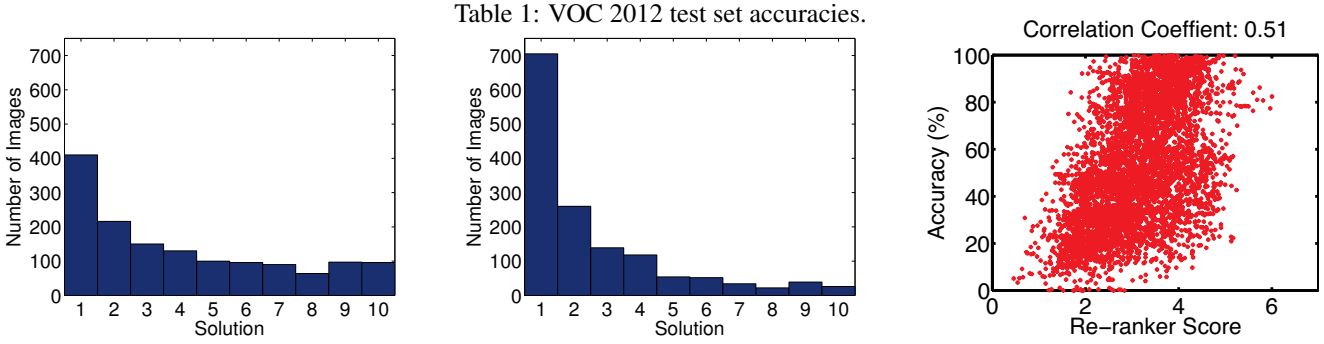
5.2. Human Experiments

Finally, to characterize the difficulty of the re-ranking problem we performed human-studies on Amazon Mechanical Turk to investigate how well people perform the task of picking a good segmentation. We selected 150 images from the VOC 2012 validation set where the MAP segmentation was neither the worst nor the best segmentation. For each image we constructed three tasks, comparing best-vs-worst, best-vs-MAP and MAP-vs-worst *Div*MBEST segmentations. Subjects were not shown the image and had to pick the better segmentation simply by looking at labelings with category names annotated. Subjects were also presented with a text box to give optional feedback on reasons for their choice. An example of the interface is shown in Fig. 4, with actual feedback received from AMT workers. The comments are instructive because they show that people are remarkably accurate at discriminating between

[3]We also used cross-validation to evaluate the feature set, rejecting some additional features not listed here that did not contribute to re-ranking accuracy.

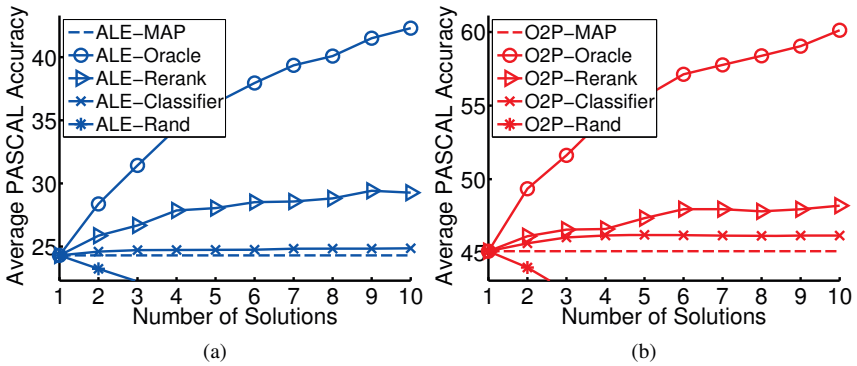
	Backgr.	Plane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	D.Table	Dog	Horse	M.bike	Person	Plant	Sheep	Sofa	Train	TV.Mo.	Average
O_2P-MAP	84.8	63.7	23.4	44.9	40.8	45.1	58.0	58.8	57.6	12.1	43.8	31.0	44.8	56.2	56.8	52.3	37.1	44.0	29.5	48.6	42.9	46.5
*Div*MBEST+RERANK	**85.7**	62.7	**25.6**	**46.9**	**43.0**	**54.8**	**58.4**	58.6	55.6	**14.6**	**47.5**	**31.2**	44.7	51.0	**60.9**	**53.5**	36.6	**50.9**	**30.1**	**50.2**	**46.8**	**48.1**

Table 1: VOC 2012 test set accuracies.



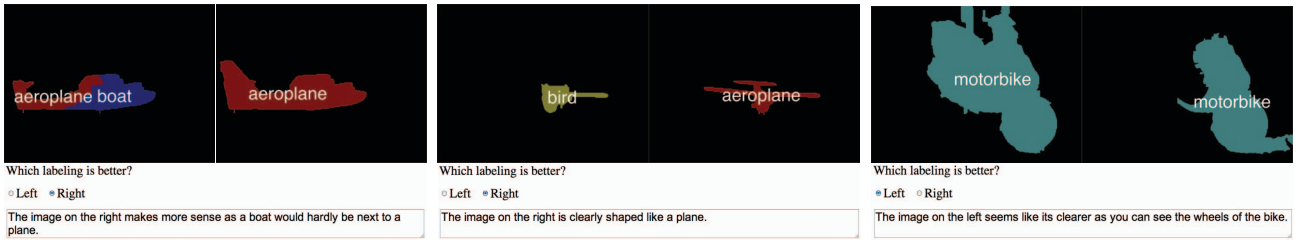
(a) Oracle Solution Rank Histogram. (b) Predicted Solution Rank Histogram. (c) Re-ranker Score vs Solution Accuracy.

Figure 2: Statistics on VOC 2012 `val` with O_2P model: (a),(b) show the number of images in which the oracle / top-re-ranked solution was originally at rank M. We can see that there is a heavy tail in the oracle distribution, but a much lighter tail in the re-ranker, suggesting that the re-ranker "plays it safe" and predicts MAP very frequently; (c) shows a scatter plot of re-ranker score vs solution accuracy.



(a) (b)

Figure 3: *Div*MBEST+RERANK performance on PASCAL VOC 2012 `val` using (a) ALE and (b) O_2P models vs. the the number of solutions.



(a) (b) (c)

Figure 4: Example MTurk tasks along with user-provided responses which were instructive in the creation of segmentation-specific features.

segmentations using cues such as category co-occurrence (Fig. 4a), category specific silhouettes (Fig. 4b), and part-whole relationships (Fig. 4c) - all of which provide support for our choice of features. The results for subject performance are shown in Tbl. 2. We can see that the binary tasks of picking between the best-vs-worst and MAP-vs-worst are easier than picking between best-vs-MAP solutions (which often tend to be both of high-quality). For the case of O_2P, picking between best and MAP is substantially more difficult - understandably so given that the MAP solutions are relatively good. Most notable is that the segmentations picked by humans achieve a significant increase in accuracy over MAP, which is impressive considering that the subjects *never saw the original image*.

Figure 5: Cases where *Div*MBEST+RERANK outperforms O_2P-MAP. In each group of images, the first column shows the original image followed by the ground-truth, MAP, and top re-ranked solution returned by *Div*MBEST+RERANK. PASCAL intersection-over-union accuracy is shown below the segmentations.

	Binary Task Accuracies			Pascal VOC Avg. Acc.			
	B-vs-W	M-vs-W	B-vs-M	Best	MAP	Worst	HR
ALE	71.9	64.4	61.7	38.0	19.1	3.2	20.5
O_2P	73.9	73.1	56.3	62.8	43.6	24.5	49.0

Table 2: (*left*) Human accuracy in predicting (B)est-vs-(W)orst, (M)AP-vs-(W)orst, and (B)est-vs-(M)AP solutions. (*right*) Pascal VOC accuracies over 150 images for best, MAP, worst, and human response (HR) solutions.

6. Conclusions

We have presented a two-stage hybrid approach to segmentation: produce a set of diverse solutions from a generative model, then re-rank them using a discriminative re-ranker. Our detailed analysis, applied to two models (ALE and O_2P) shows that the set of solutions obtained in stage 1 contains segmentations dramatically more accurate than the single MAP solution, and that the sources of diversity are non-trivial. With the re-ranker trained using a novel structured SVM formulation, we obtain state of the art results on VOC 2012 segmentation `test` set.

Chief among our future work directions is to continue closing the gap between what is achieved by the re-ranker and what is possible based on our oracle analysis of the diverse solution sets. The gap and the actual values of the oracle suggest that efforts of CRF modelling community may be misguided – the bottleneck is not optimization algorithms for probabilistic models, rather the bottleneck is the absence of rich features that can tell a dog from a cat.

This intuition, and the specific *Div*MBEST+RERANK approach developed in this paper, are applicable to vision problems beyond segmentation, and indeed to other domains such as natural language processing. Our current work includes exploration of such applications.

References

[1] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *CVPR*, pages 3378 –3385, june 2012. 1

[2] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse M-Best Solutions in Markov Random Fields. In *ECCV*, 2012. 2, 3, 4, 5

[3] X. Boix, J. M. Gonfaus, J. van de Weijer, A. D. Bagdanov, J. S. Gual, and J. Gonzàlez. Harmony potentials - fusing global and local scale for semantic image segmentation. *IJCV*, 96(1):83–102, 2012. 1

[4] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, pages 430–443, 2012. 2, 5

[5] J. Carreira, F. Li, and C. Sminchisescu. Object recognition by sequential figure-ground ranking. *IJCV*, 98(3):243–262, 2012. 1

[6] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010. 2, 5

[7] Y.-L. Chow and R. Schwartz. The n-best algorithm: an efficient procedure for finding top n sentence hypotheses. In *Proceedings of the Workshop on Speech and Natural Lang.*, pages 199–202, 1989. 2

[8] M. Collins. Discriminative reranking for natural language parsing. In *ICML*, pages 175–182, 2000. 3

[9] M. Collins. Discriminative syntactic language modeling for speech recognition. In *In Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 507–514, 2005. 3

[10] M. Dinarelli, A. Moschitti, and G. Riccardi. Discriminative reranking for spoken language understanding. *Trans. Audio, Speech and Lang. Proc.*, 20(2):526–539, Feb. 2012. 3

[11] I. Endres and D. Hoiem. Category independent object proposals. In *ECCV*, 2010. 2

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, June 2010. 2, 3

[13] M. Fromer and A. Globerson. An LP view of the m-best MAP problem. In *NIPS*, 2009. 3

[14] C. Gu, J. J. Lim, P. Arbelaez, and J. Malik. Recognition using regions. In *CVPR*, pages 1030–1037, 2009. 1

[15] T. Joachims, T. Finley, and C.-N. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009. 4

[16] J. Kim and K. Grauman. Shape sharing for object segmentation. In *ECCV*, pages 444–458, 2012. 2

[17] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008. 1

[18] L. Ladickỳ, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. *ICCV*, 2009. 1, 2, 5

[19] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010. 5

[20] L. Ladicky and P. H. Torr. The automatic labelling environment. http://cms.brookes.ac.uk/staff/PhilipTorr/ale.htm. 2

[21] R. Mottaghi. Augmenting deformable part models with irregular-shaped object patches. In *CVPR*, 2012. 6

[22] D. Nilsson. An efficient algorithm for finding the m most probable configurations in probabilistic expert systems. *Statistics and Computing*, 8:159–173, 1998. 10.1023/A:1008990218483. 3

[23] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 2

[24] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010. 2

[25] L. Shen, A. Sarkar, and F. J. Och. Discriminative reranking for machine translation. In *HLT-NAACL*, pages 177–184, 2004. 3

[26] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005. 2, 4

[27] J. Uijlings, K. van de Sande, A. Smeulders, T. Gevers, N. Sebe, and C. Snoek. The most telling window for image classification. In *ICCV Pascal VOC Workshop*, 2011. 6

[28] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004. 2

[29] D. Weiss, B. Sapp, and B. Taskar. Sidestepping intractable inference with structured ensemble cascades. In *NIPS*, 2010. 2

[30] C. Yanover and Y. Weiss. Finding the m most probable configurations using loopy belief propagation. In *NIPS*, 2003. 3