Diverse M-Best Solutions in Markov Random Fields

Dhruv Batra[1], Payman Yadollahpour[1],
Abner Guzman-Rivera[2], and Gregory Shakhnarovich[1]

[1]TTI-Chicago [2]UIUC

Abstract. Much effort has been directed at algorithms for obtaining the highest probability (MAP) configuration in probabilistic (random field) models. In many situations, one could benefit from additional high-probability solutions. Current methods for computing the M most probable configurations produce solutions that tend to be very similar to the MAP solution and each other. This is often an undesirable property. In this paper we propose an algorithm for the *Diverse M-Best* problem, which involves finding a diverse set of highly probable solutions under a discrete probabilistic model. Given a dissimilarity function measuring closeness of two solutions, our formulation involves maximizing a linear combination of the probability and dissimilarity to previous solutions. Our formulation generalizes the M-Best MAP problem and we show that for certain families of dissimilarity functions we can guarantee that these solutions can be found as easily as the MAP solution.

1 Introduction

The introduction of sophisticated discrete optimization tools for inference in Markov Random Fields (MRFs) over the last two decades has allowed optimal or provably approximate solutions to certain vision problems previously deemed intractable. For instance, using max-flow/min-cut methods [7, 13], one can find the globally optimal solution for a (submodular) foreground-background segmentation problem on a 2 Megapixel image within seconds, effectively searching $2^{2,000,000}$ possible segmentations.

However, *optimization error* is only one component of generalization error of a learning algorithm [4]. Even when exact inference in MRFs is efficient, the maximum a posteriori (MAP) solution could be far from the ground truth. The source of this discrepancy may be *approximation error*, due to the limitations of the model class (*e.g.* pairwise binary submodular MRFs), or *estimation error* *i.e.* error made because parameters are learnt from a finite training set.

Indeed, recent empirical studies [20, 32] have repeatedly found that MAP solutions of existing models are of much poorer quality than the ground-truth on vision problems like segmentation, stereo, optical flow, denoising, etc. Equivalently, the ground-truth has lower probability than the MAP solution under existing models. This discrepancy has been the driving force behind the search for more accurate (higher-order) models; [16, 23, 27] are just a few examples.

In addition to this quest for better models, we believe that another way to mitigate this problem is to look beyond obtaining a single MAP solution – to look

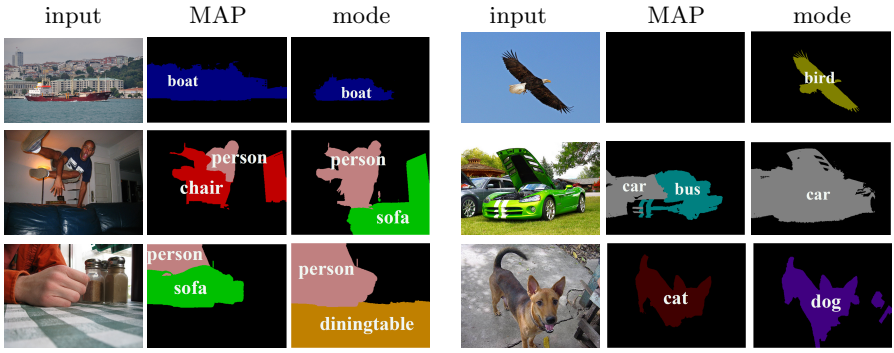


Fig. 1: Examples of category-level segmentation results on test images from VOC 2010. For each image, "mode" above is the best of 10 solutions obtained with *Div*MBEST.

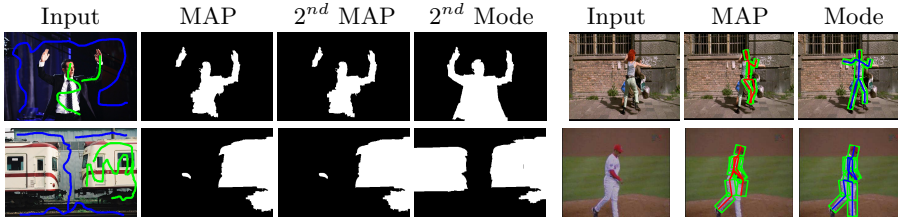


Fig. 2: (Left) Interactive segmentation setup; (Right) Pose tracking. "Mode" above refers to the solution obtained with *Div*MBEST.

for a *diverse set* of highly probable solutions instead from our existing models. Even if the MAP solution alone is of poor quality, a diverse set of highly probably hypotheses might still enable accurate predictions. Note that in contrast to the M-Best MAP problem [9, 22, 38] that involves finding the top M most probable solutions under a probabilistic model, our approach emphasizes *diversity*. We seek to produce highly probable solutions that are qualitatively different from the MAP and from each other. This is an important distinction because the literal definition of M-best MAP is not expected to work well in practice. In a large state-space problem (*e.g.* $2^{2,000,000}$) any reasonable setting of M $(10 - 50)$ would return solutions nearly identical to the MAP. Ideally, we would like to find the modes of the distribution learnt by our probabilistic model. Figs. 1, 2 show examples of these diverse solutions extracted for different tasks.

Overview. In this paper, we introduce the *Diverse M-Best* problem, which involves finding a diverse set of highly probable solutions under a discrete probabilistic model. Our formulation assumes access to a dissimilarity function $\Delta(\cdot, \cdot)$ measuring the difference between two solutions. We present an integer programming formulation for the Diverse M-Best problem, study its Lagrangian relaxation. We show that this relaxation has an interesting interpretation as the Δ-augmented energy minimization problem, which minimizes a linear combination of the energy and similarity to previous solutions, thereby producing solutions with low energy and high diversity. For simplicity we sometimes will refer to these solutions as modes, although they are not guaranteed to be actual modes of the Gibbs distribution.

Contributions. Our principal contribution is the *first* principled optimization formulation for extracting diverse high-probability solutions in MRFs. It includes the familiar M-Best MAP problem as a special case. Perhaps most crucial for a practitioner, we show that for certain families of Δ-functions (that we discuss in Section 5), the Δ-augmented energy minimization problem is *as easily solvable* as the original MAP problem. Thus if exact or provably approximate algorithms exist for finding the MAP solution in a model, (*e.g.* graph-cuts [7, 13]), those *same* algorithms are applicable for finding Diverse M-Best solutions.

Applications. Our algorithm could be applicable whenever multiple hypotheses can be effectively used to infer a solution to the problem. In this paper, we show applications on three vision tasks: (i) an interactive application (object-cutout) where multiple diverse cutouts are presented to a user to minimize interaction time (ii) category-level image segmentation, where multiple highly-probable solutions could be ranked using a secondary mechanism [19] not amenable to inference, and finally (iii) human pose tracking in video, where multiple solutions per frame can be used for pose-tracking via a Viterbi-like decoding scheme that exploits temporal context [25]. In all three of these tasks we demonstrate that Diverse M-Best solutions can be used to significantly improve performance.

2 Related Work

M-Best MAP. Most directly relevant to our work is literature on the M-Best MAP problem. Lawler [18] proposed a general algorithm to compute the top M solutions for a large family of discrete optimization problems, and the ideas used in Lawler's algorithm form the basis of most algorithms for the M-Best MAP problem. The first family of algorithms for M-Best MAP [22, 30] were junction-tree based exact algorithms, thus feasible only for low-treewidth graphs. Dechter and colleagues [8, 21] have recently provided dynamic-programming algorithms for M-Best MAP, but these are exponential in treewidth as well. Yanover and Weiss [38] proposed an algorithm that requires access only to max-marginals. Thus, for certain classes of MRFs that allow efficient exact computation of max-marginals, *e.g.* binary pairwise supermodular MRFs [12], M-Best solutions can be found for arbitrary treewidth graphs. Moreover, *approximate* M-Best solutions may be found by approximating the max-marginal computation, *e.g.* via max-product BP. More recently, Fromer and Globerson [9] provided a Linear Programming (LP) view of the M-Best MAP problem by studying the assignment-excluding marginal polytope. We show in Section 4 that M-best MAP is a special case of our formulation, with a particular kind of $\Delta(\cdot,\cdot)$.

Sampling based Approaches. A common alternative is to use sampling [1, 26, 35] to produce multiple solutions, which may then be refined or checked for diversity. Such approaches typically exhibit long wait times to transition from one mode to another. Moreover, in contrast to our work, there is no direct mechanism to require the multiple solutions (samples) obtained in this way to be diverse. An interesting approach by Papandreou and Yuille [24] shows that approximate samples may be drawn from the Gibbs distribution by injecting a certain noise into the parameters and solving for MAP on these perturbed

parameters. Our approach can be seen as a deterministic counterpart to their stochastic perturbation – we always perturb the parameters in a fixed way, and the perturbation does not produce iid samples, rather Diverse M-Best solutions.

Diverse Solutions. The need for diverse solutions arises in a number of problems in computer vision and more broadly in machine learning. Yu and Joachims [39] proposed to learn a predictor that selects a topic-diverse subset of documents. Diversity is also a key goal in the context of non-maximal suppression for object detection [2, 3]. Recently, Park and Ramanan [25] applied the max-marginal algorithm of [38] to decode multiple solutions from a deformable parts model, with an added constraint forcing at least one non-overlapping part. Their approach is also contained in our formulation as a special case (see Section 4).

3 Preliminaries: MAP Inference in MRFs

Notation. For any positive integer n, let $[n]$ be shorthand for the set $\{1, 2, \ldots, n\}$. We consider a set of discrete random variables $\mathbf{x} = \{x_i \mid i \in [n]\}$, each taking value in a finite label set, $x_i \in X_i$. For a set $A \subseteq [n]$, we use x_A to denote the tuple $\{x_i \mid i \in A\}$, and X_A to be the cartesian product of the individual label spaces $\times_{i \in A} X_i$. For ease of notation, we use x_{ij} as a shorthand for $x_{\{i,j\}}$.

MAP. Let $G = (\mathcal{V}, \mathcal{E})$ be a graph defined over these variables, *i.e.* $\mathcal{V} = [n]$, $\mathcal{E} \subseteq \binom{\mathcal{V}}{2}$, and let $\theta_A : X_A \to \mathbb{R}, \quad (\forall A \in \mathcal{V} \cup \mathcal{E})$ be functions defining the energy at each node and edge for the labeling of variables in scope. The goal of MAP inference is to find the labeling $\mathbf{x}$ of the variables that minimizes this real-valued energy function:

$$\min_{\mathbf{x} \in X_{\mathcal{V}}} \sum_{A \in \mathcal{V} \cup \mathcal{E}} \theta_A(x_A) = \min_{\mathbf{x} \in X_{\mathcal{V}}} \sum_{i \in \mathcal{V}} \theta_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(x_i, x_j). \tag{1}$$

The techniques proposed in this paper are naturally applicable to higher-order MRFs. However, to simplify the exposition we only consider pairwise energies.

MAP Integer Program. MAP inference is typically set up as an integer programming problem over boolean variables. For each node and edge $A \in \mathcal{V} \cup \mathcal{E}$, let $\boldsymbol{\mu}_A = \{\mu_A(s) \mid s \in X_A,\ \mu_A(s) \in \{0, 1\}\}$, be a vector of indicator variables encoding all possible configurations of x_A. If $\mu_A(s)$ is set to 1, this implies that x_A takes label s. Moreover, let $\boldsymbol{\theta}_A = \{\theta_A(s) \mid s \in X_A\}$ be a vector holding energies for all possible configurations of x_A, and $\boldsymbol{\mu} = \{\boldsymbol{\mu}_A \mid A \in \mathcal{V} \cup \mathcal{E}\}$ be a vector holding the entire configuration. Using this notation, the MAP inference integer program can be written as:

$$\min_{\boldsymbol{\mu}_i, \boldsymbol{\mu}_{ij}} \quad \sum_{i \in \mathcal{V}} \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{(i,j) \in \mathcal{E}} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij} \tag{2a}$$

$$s.t. \quad \sum_{s \in X_i} \mu_i(s) = 1 \qquad \forall i \in \mathcal{V} \tag{2b}$$

$$\sum_{s \in X_i} \mu_{ij}(s,t) = \mu_j(t), \quad \sum_{t \in X_j} \mu_{ij}(s,t) = \mu_i(s) \qquad \forall \{i,j\} \in \mathcal{E} \tag{2c}$$

$$\mu_i(s), \mu_{ij}(s,t) \in \{0, 1\}. \tag{2d}$$

Here (2b)and (2c) enforce that exactly one label is assigned to a variable, and that assignments are consistent across edges. To be concise, we will use $\mathcal{L}(G)$ to

denote the set of $\boldsymbol{\mu}$ that satisfy these two constraints. Thus, the above problem (2) can be written concisely as: $\min_{\boldsymbol{\mu}\in\mathcal{L}(G),\ \mu_A(s)\in\{0,1\}} \sum_{A\in\mathcal{V}\cup\mathcal{E}} \boldsymbol{\theta}_A \cdot \boldsymbol{\mu}_A$.

Problem (2) is known to be NP-hard in general. A number of techniques solve a Linear Programming (LP) relaxation of this problem [36], which is given by relaxing the boolean constraints (2d) to the unit interval, *i.e.* $\mu_i(s), \mu_{ij}(s,t) \geq 0$.

4 Diverse M-Best: Formulation

We now describe our proposed Diverse M-Best formulation. Recall that the goal is to produce a diverse set of low-energy solutions. We approach this problem with a greedy algorithm, where the next solution is defined as the lowest energy state with at least some minimum dissimilarity from the previously chosen solutions. To do so, we assume access to a dissimilarity function $\Delta(\boldsymbol{\mu}^1, \boldsymbol{\mu}^2)$ between solutions. Let $\boldsymbol{\mu}^m$ denote the m^{th}-best mode. Thus $\boldsymbol{\mu}^1$ is the MAP, $\boldsymbol{\mu}^2$ is the second-best mode and so on.[1] Let us first search for the second mode. We propose the following straightforward yet fairly general formulation:

$$\boldsymbol{\mu}^2 = \operatorname*{argmin}_{\boldsymbol{\mu}\in\mathcal{L}(G),\ \mu_A(s)\in\{0,1\}} \sum_{A\in\mathcal{V}\cup\mathcal{E}} \boldsymbol{\theta}_A \cdot \boldsymbol{\mu}_A \tag{3a}$$

$$s.t. \quad \Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^1) \geq k. \tag{3b}$$

We refer to the above formulation as $Div2\textsc{Best}(\Delta, k)$. Note that it is parametrized by the two design choices Δ and k (both of which we will discuss in detail). Intuitively, we can see that the above formulation searches for the lowest energy solution that is at least k-units away from the MAP solution under $\Delta(\cdot,\cdot)$. The extension from $Div2\textsc{Best}(\Delta, k)$ to $Div\textsc{MBest}\ (\Delta, \boldsymbol{k})$ is fairly simple: we search for the lowest energy solution at least k_m-units away from each of the previously found (M-1) solutions, *i.e.* $\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^m) \geq k_m, \quad \forall m \in \{1, \ldots, M-1\}$, where $\boldsymbol{k} = \{k_m \mid m \in [M-1]\}$ is the vector of different distance thresholds.

This formulation is general enough to contain existing ones as special cases:

Special Case: M-Best MAP is obtained when Δ is a $0-1$ dissimilarity (*i.e.* $\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^1) = [\![\boldsymbol{\mu} \neq \boldsymbol{\mu}^1]\!]$, where $[\![\cdot]\!]$ is an indicator function), and $k = 1$. Thus (3b) simply forces the next best solution to not be identical to MAP. Other choices of $\Delta(\cdot,\cdot)$ are also possible to express the M-Best MAP problem.

Special Case: N-Best Maximal Decoding of Park and Ramanan [25] corresponds to $\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^1) = \max_{i\in\mathcal{V}} \Delta_i(\boldsymbol{\mu}_i, \boldsymbol{\mu}_i^1)$ and $k = 1$. Intuitively, their approach defines local dissimilarity functions at each node, and forces at least one node label to be non-identical (under Δ_i) to the MAP label at that node.

5 Diverse M-Best: Lagrangian Relaxation

In general, $Div\textsc{MBest}\ (\Delta, \boldsymbol{k})$ is at least as hard to solve the MAP inference problem, which is NP-hard. Moreover, the dissimilarity constraints obfuscate

[1] Whenever we refer to MAP we mean exact or approximate MAP, as produced by the inference engine for the model at hand.

some of the structure in the problem typically exploited by MAP inference algorithms. Thus we study the Lagrangian relaxation of *Div*MBest $(\Delta, \boldsymbol{k})$, formed by dualizing the dissimilarity constraints $\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^m) \geq \boldsymbol{k}$:

$$f(\boldsymbol{\lambda}) \quad = \quad \min_{\boldsymbol{\mu} \in \mathcal{L}(G),\ \mu_A(s) \in \{0,1\}} \quad \sum_{A \in \mathcal{V} \cup \mathcal{E}} \boldsymbol{\theta}_A \cdot \boldsymbol{\mu}_A - \sum_{m=1}^{M-1} \lambda_m \left(\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^m) - k_m \right). \quad (4)$$

Here $\boldsymbol{\lambda} = \{\lambda_m \mid m \in [M-1]\}$ is the set of Lagrange multipliers, that determine the weight of the penalty imposed for violating the constraints by a solution. Intuitively, we can see that the Lagrangian relaxation minimizes a linear combination of the energy of the MRF and similarity (negative dissimilarity) to the MAP solution, with the weighting given by the Lagrange multipliers. Formally, we can state the following theorem, proven in [5]:

Proposition 1. *For all values of* $\boldsymbol{\lambda} \geq 0$, $f(\boldsymbol{\lambda})$ *is a lower-bound on the value of the primal problem Div*MBest $(\Delta, \boldsymbol{k})$. *Moreover,* $f(\boldsymbol{\lambda})$ *is a piece-wise linear function and concave in* $\boldsymbol{\lambda}$.

5.1 Supergradient Ascent on the Lagrangian Dual

The tightest lower-bound is obtained by setting up the Lagrangian dual problem: $\max_{\boldsymbol{\lambda} \geq 0}\ f(\boldsymbol{\lambda})$. Since f is a non-smooth *concave* function, this can be achieved by the *supergradient ascent* algorithm, analogous to the subgradient descent for minimizing non-smooth convex functions [31]. Since $\boldsymbol{\lambda}$ is a constrained variable, we follow the projected supergradient ascent algorithm: iteratively updating the Lagrange multipliers according to the following update rule: $\boldsymbol{\lambda}^{(t+1)} \longleftarrow \left[\boldsymbol{\lambda}^{(t)} + \alpha_t \nabla f(\boldsymbol{\lambda}^{(t)})\right]_+$, where $\nabla f(\boldsymbol{\lambda}^{(t)})$ is the supergradient of f at $\boldsymbol{\lambda}^{(t)}$, α_t is the step-size and $[\cdot]_+$ is the projection operator that projects a vector onto the positive orthant. If the sequence of multipliers $\{\alpha_t\}$ satisfies $\alpha_t \geq 0$, $\lim_{t \to \infty} \alpha_t = 0$, $\sum_{t=0}^{\infty} \alpha_t = \infty$, then projected supergradient ascent converges to the optimum of the lagrangian dual [31].

To find the supergradient of $f(\boldsymbol{\lambda})$, note that f is a point-wise minimum of linear functions: *i.e.* $f(\boldsymbol{\lambda}) = \min_{\boldsymbol{\mu}} \boldsymbol{a}_{\boldsymbol{\mu}} \cdot \boldsymbol{\lambda} + b_{\boldsymbol{\mu}}$. It can be easily shown that the supergradient of f is given by $\nabla f(\boldsymbol{\lambda}) = \boldsymbol{a}_{\hat{\boldsymbol{\mu}}(\boldsymbol{\lambda})}$, where $\hat{\boldsymbol{\mu}}(\boldsymbol{\lambda}) = \operatorname{argmin}_{\boldsymbol{\mu}} \boldsymbol{a}_{\boldsymbol{\mu}} \cdot \boldsymbol{\lambda} + b_{\boldsymbol{\mu}}$. We omit the proof due to space constraints.

Mapping this definition to (4), we can see that the supergradient of f for our formulation is given by (5):

$$\nabla f(\boldsymbol{\lambda}) = -\begin{bmatrix} \Delta\left(\hat{\boldsymbol{\mu}}(\boldsymbol{\lambda}), \boldsymbol{\mu}^1\right) - k_1 \\ \cdots \\ \cdots \\ \Delta\left(\hat{\boldsymbol{\mu}}(\boldsymbol{\lambda}), \boldsymbol{\mu}^{M-1}\right) - k_{M-1} \end{bmatrix} \quad (5)$$

where $\hat{\boldsymbol{\mu}}(\boldsymbol{\lambda})$ is the optimal primal solution of (4) for the current setting of $\boldsymbol{\lambda}$. This supergradient (and the update procedure) has an intuitive interpretation. Recall that the Lagrangian relaxation minimizes a linear combination of the energy and similarity to the MAP solution, with the weighting given by $\boldsymbol{\lambda}$. If $\hat{\boldsymbol{\mu}}(\boldsymbol{\lambda}^{(t)})$ violates one of the diversity constraints, *i.e.* is less than k_m units away from a previous solution $\boldsymbol{\mu}^m$, then the supergradient w.r.t. $\lambda_m^{(t)}$ will be positive and the cost for violating the constraint will increase after the update, thus

encouraging the next solution $\hat{\boldsymbol{\mu}}(\boldsymbol{\lambda}^{(t+1)})$ to satisfy the constraints. Conversely, if the diversity constraints are satisfied, the supergradient is negative indicating that $\lambda_m^{(t)}$ may be over-penalizing for violations and may be reduced to allow lower energy solutions.

5.2 Computing the Supergradient

Note that computing the supergradient requires solving the Δ-augmented energy minimization problem (4). At a high-level, this setup is a similar to cutting-plane methods for training Structured-SVMs [34], where the generation of each additional cutting plane requires solving the loss-augmented energy minimization problem. As we discuss next, for some classes of Δ-functions, we can solve the Δ-augmented energy minimization problem simply by reusing the *same algorithms* used for finding the MAP, by modifying the energy of the MRF. This allows all the developments in the MAP inference literature to be directly translated to the Diverse M-Best problem, without any changes.

Example: Dot-Product Dissimilarity: $\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^1) = -\sum_{i \in \mathcal{V}} \boldsymbol{\mu}_i^T W \boldsymbol{\mu}_i^1$, *i.e.* the sum of bilinear forms of labellings at nodes. For discrete solutions $\boldsymbol{\mu}(s), \boldsymbol{\mu}^1(s) \in \{0, 1\}$, setting W to the identity matrix (I) makes this dissimilarity function equivalent to the Hamming distance between the two solutions. In the presence of non-identity W, this is a weighted-Hamming distance, where W incorporates cross-label similarity. Moreover, note that $\sum_{A \in \mathcal{V} \cup \mathcal{E}} \boldsymbol{\theta}_A \cdot \boldsymbol{\mu}_A - \sum_{m=1}^{M-1} \lambda_m (\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^m) - k_m) = \sum_{i \in \mathcal{V}} (\boldsymbol{\theta}_i + \sum_{m=1}^{M-1} \lambda_m W \boldsymbol{\mu}_i^m) \cdot \boldsymbol{\mu}_i + \sum_{(i,j) \in \mathcal{E}} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij}$. Thus, $f(\boldsymbol{\lambda})$ becomes the same as the MAP problem with modified unary energies. When $W = I$, the modification simply increases the local cost of each of the previous states $\boldsymbol{\mu}_i^m$ by λ_m. Non-identity W "smears" the affect of $\boldsymbol{\mu}_i^m$. For this Δ-function, we can use *any* existing MAP inference algorithm to solve this problem. Perhaps the most attractive feature is that the edge-energies are left unaffected. For instance, if they were submodular in the original model, they continue to be submodular. This allows us to use efficient graph-cut algorithms [7, 13].

Example: Higher-order Dissimilarity. Another example of a useful dissimilarity function is one that decomposes into functions of subsets of variables, *i.e.* $\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^1) = \sum_{A \in \mathcal{V}} \Delta_A(\boldsymbol{\mu}_A, \boldsymbol{\mu}_A^1)$. If each of the these terms $\Delta_A(\cdot, \cdot)$ has some structure, *e.g.* cardinality potentials [33] or lower linear-envelope potentials [10] or sparse (pattern-based) higher-order potentials [14, 28], that allows for *messages* to be efficiently computed, this Δ-augmented energy minimization can be performed via dual-decomposition based message-passing algorithms. The details can be found in the supplementary material (from the authors' webpages).

Finally, we note that successive supergradient computations (at $\boldsymbol{\lambda}^{(t)}$ and $\boldsymbol{\lambda}^{(t+1)}$) require solving fairly similar inference problems, which may be warm-started from the solutions of the previous iterations – either by re-using the search trees in graph-cuts [11] or by reusing messages in dual-decomposition.

5.3 Tightness of the Lagrangian Dual

Recall that the Lagrangian dual involves finding the tightest lower-bound for the primal, *i.e.* $\max_{\boldsymbol{\lambda} \geq 0} f(\boldsymbol{\lambda}) \leq \mathit{Div}\text{MBest}(\Delta, \boldsymbol{k})$. One important question to ask is

– when is the relaxation tight? To answer this question, we state the following theorem (see supplementary material from the authors' webpages for the proof):

Theorem 1. *First, the Lagrangian dual* $\max_{\boldsymbol{\lambda}\geq 0} f(\boldsymbol{\lambda})$ *is equivalent to solving (*i.e. *has zero duality gap with) the following primal relaxation of Div*MBEST $(\Delta, \boldsymbol{k})$*:*

$$\min_{\boldsymbol{\mu}} \quad \sum_{A\in\mathcal{V}\cup\mathcal{E}} \boldsymbol{\theta}_A \cdot \boldsymbol{\mu}_A \tag{6a}$$

$$s.t. \quad \boldsymbol{\mu} \in Co\Big\{\mu_A(s) \in \{0,1\} \mid \boldsymbol{\mu} \in \mathcal{L}(G)\Big\} \tag{6b}$$

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^m) \geq k_m \qquad \forall m \in \{1,\ldots,M-1\} \tag{6c}$$

where $Co\{\cdot\}$ *denotes the convex hull of the discrete solutions.*

Second, for certain specific Δ*-functions (*e.g. *the 0-1 function of M-Best MAP), the convex hull can be replaced by the discrete solutions themselves,* i.e. *(6b) may be replaced by* $\mu_A(s) \in \{0,1\}$, $\boldsymbol{\mu} \in \mathcal{L}(G)$*. Thus for such specific families, the Lagrangian relaxation* is *tight.*

However, if no assumptions are made on Δ*, the Lagrangian relaxation is not guaranteed to be tight and may leave a duality gap.*

Although the dot-product dissimilarity does not lead to a tight Lagrangian relaxation, we show in our experiments that even the relaxed solutions achieved by the Lagrangian dual could be very useful for the applications.

5.4 How Much Diversity?

Our *Div*MBEST formulation provides a principled way to trade off diversity vs. optimizing the energy. One issue of practical concern is how the tradeoff parameter $\boldsymbol{k}$ is chosen. The answer is related to the topology of the energy landscape. If $\boldsymbol{k}$ is set too low the next solution may not be able to escape the energy valley (and thus will not be a real "mode"). If $\boldsymbol{k}$ is set too high, then several modes will be ignored. An appropriate value of $\boldsymbol{k}$ is problem-dependent and must be tuned as a free parameter. Moreover, $\hat{\boldsymbol{\lambda}}(\boldsymbol{k}) = \operatorname{argmin}_{\boldsymbol{\lambda}\geq 0} f(\boldsymbol{\lambda})$ the optimum setting of lambda is different for different values of $\boldsymbol{k}$. Thus, instead of performing grid search on $\boldsymbol{k}$ and running supergradient ascent for each value of $\boldsymbol{k}$, we can directly perform grid search on $\boldsymbol{\lambda}$. Intuitively, this corresponds to learning from data a linear weighting between the energy of the model and a diversity-inducing term. This is computationally much more efficient, both at the validation and testing stage. For our experiments, we learn the appropriate degree of diversity by tuning $\boldsymbol{\lambda}$ on a validation data set. This is similar in principle to the treatment of, say, a regularizer in learning. In the primal form, the regularization parameter limits the norm of the solution, but in practice the Lagrangian multiplier is treated as a norm penalty coefficient, and directly tuned via cross-validation.

6 Experiments

We apply our Diverse M-Best formulation, with (uniformly weighted) Hamming (dot product) dissimilarity Δ, to three scenarios:

1. An interactive segmentation (object cutout) setup in Section 6.1.

2. Category-level segmentation on PASCAL VOC 2010 data in Section 6.2.
3. Human articulated pose tracking in video in Section 6.3.

These scenarios use very different models (a binary pairwise submodular flat CRF, a multi-label hierarchical CRF with global factors, and a multi-label tree-CRF), with different MAP inference algorithms (max-flow/min-cut, α-expansion, dynamic programming). Despite the differences, our approach can naturally use the inference algorithm in each model to find the Diverse M-Best solutions.

We can compare *Div*MBest against two alternatives for producing multiple solutions: **M-Best-MAP**, that produces low energy solutions without a focus on diversity, and perturbation-based techniques that simply produce diverse solutions without optimization within the original model. Specifically, given a MAP solution, we can produce additional solutions by changing assignments of a subset of nodes. The subset can be chosen either randomly (denoted **Random**), or based on the estimated confidence in the MAP labels; a natural measure of confidence is the entropy of the estimated min-marginals (**Confidence**). To make for a fair comparison, for each solution produced by *Div*MBest that differs from MAP in S nodes, we generate a perturbation-based solution that changes exactly S nodes as well, thus achieving equal amount of diversity with *Div*MBest.

We show that both these alternatives fall short in all scenarios. The M-Best-MAP produces redundant solutions with little extra information over the MAP, despite significant computational effort. The sampling-based methods produce solutions that are diverse, but improbable and inaccurate. In contrast, *Div*MBest provides a principled way to trade-off both goals.

6.1 Interactive segmentation

Interactive segmentation is a task where a user is interested in cutting out a foreground object of interest from an image via annotations like scribbles [6] or a coarse bounding box [29]. This is typically treated as a figure/ground segmentation task, with the MAP solution presented to the user. The user then provides additional supervision, leading to updated MAP solutions, till the MAP solution is acceptable. In order to minimize user interactions, the interface could show not a single cutout, but a set of possible cutouts for the user to simply select from. To make the most of this setup, this list of solutions must be small, diverse, and the algorithm that generates it must be efficient.

Dataset, Features, Energies, Inference. We simulate this scenario on 100 images from the PASCAL VOC 2010 dataset, and manually provided scribbles on objects contained in them; Fig. 2 shows examples. For each image, we set up a 2-label pairwise CRF, with a node term for each superpixel in the image and an edge term for each adjacent pair of superpixels. At each superpixel, we extract colour and texture features, and given the foreground/background scribbles on a single input image, train a Transductive SVM on these features. The node terms for the image are derived from the output of these TSVMs. The edge terms are contrast-sensitive Potts. A detailed description of the pipeline, the features and the parameter setting is in the supplementary materials. Fifty of the images were used for tuning the parameters, and the other 50 for reporting testing accuracies.

	MAP	*Div*MBEST-dot prod.	*Div*MBEST-HOP	M-Best	Random	Confidence
Acc.(%)	91.542	**95.16**	93.82	91.59	91.68	93.17

Table 1: Interactive segmentation: pixel accuracies averaged over 50 test images.

Baselines. The 2-label contrast-sensitive Potts model results in a submodular energy function so we can efficiently compute the exact MAP solution using graph-cuts implementation [13]. Moreover, with Hamming dissimilarity, we can efficiently and optimally solve the Δ-augmented energy minimization problem to compute the Diverse M-Best solutions using graph-cuts as well. As a first baseline, we implemented the M-best MAP algorithm of Yanover and Weiss [38], which requires repeated computation of min-marginals. We computed exact min-marginals using the dynamic graph-cuts algorithm of Kohli and Torr [12]; these were used to produce the **Confidence** baseline as well. Finally, we tested *Div*MBEST with a higher-order potential (HOP) dissimilarity; we include the results in Table 1 and give additional details in supplementary material.

Results For each of the 50 test images in our dataset we generated MAP and 5 additional solutions using the methods described above. Table 1 shows the max accuracy of these 6 solutions for each method, averaged over 50 images. We can see that *Div*MBEST solutions result in the highest improvement over MAP. Figure 2 shows some example segmentations. Notice that the 2^{nd}-best MAP solution is nearly identical to the MAP solution whereas the solution from *Div*MBEST is qualitatively different, and could be significantly closer to the ground-truth labeling when MAP makes a mistake. In one case, the 2^{nd} mode found another instance of the object that MAP had missed, and in another, it completed a thin long structure (the arm of the person). These results suggest that the segmentation model is not completely accurate but still contains useful segmentations as other modes of the CRF distribution. *Div*MBEST provides a principled way of extracting these other segmentations.

6.2 Category level Segmentation

Problem. In the second experiment we consider the problem of category-level segmentation, *i.e.* labeling each pixel in an image with one of 20 object categories or the background. This task is part of the PASCAL VOC `comp5` challenge.

Model. The model we consider for this problem is the Associative Hierarchical CRF of Ladicky *et al.* [15], which achieved competitive results in recent years' VOC challenges. We used the publicly available implementation by the authors – the Automatic Labeling Environment (ALE) [17]. ALE includes many kinds of potentials: unary potentials based on textonboost features, P^n Potts terms (between superpixel nodes and pixel nodes) and a global co-occurence potential [16]. In a sequence of papers [15,16], the authors developed a MAP inference algorithm for this model, and we are directly able to utilize this algorithm to compute *Div*MBEST solutions, by rerunning it with modified node energies.

Baselines. The same three baselines as in Section 6.1 are applicable here, however we found the M-best MAP algorithm [38] to be infeasibly slow for this model. Unlike the previous application, energy in this model is not submodular and does not allow for efficient computation of min-marginals. According to

	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
MAP	78.5	35.1	5.2	20.3	20.8	11.8	39.4	38.2	25.8	8.9	14.1	30.2	10.0	12.3	37.6	33.5	10.3	24.2	16.2	28.7	20.5	24.8
Conf.	78.5	35.1	5.3	20.1	20.7	12.6	39.4	37.9	26.8	8.9	14.1	30.2	10.3	12.2	39.5	33.4	10.6	24.2	17.3	28.4	20.5	25.1
Random	74.9	32.4	6.4	16.1	14.7	12.3	34.3	32.6	22.6	8.0	13.2	21.1	8.7	10.4	32.9	28.9	7.8	20.6	10.8	23.5	17.3	21.4
*Div*10Best	85.6	53.9	14.6	36.9	33.6	33.2	64.2	56.3	47.7	16.1	30.3	46.8	29.1	28.7	59.0	50.0	32.5	46.7	31.2	52.9	39.0	42.3

Table 2: VOC 2010 Validation set accuracies.

	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
MAP	73.7	44.0	14.2	15.3	21.0	23.2	41.3	37.0	27.6	6.1	23.9	25.2	12.8	24.3	51.0	27.8	20.0	28.2	17.1	36.5	23.9	28.3
*Div*10Best	83.4	54.4	19.6	22.4	34.5	22.2	60.8	55.5	45.8	14.0	45.5	35.1	34.8	40.1	53.6	48.7	28.0	48.7	31.2	50.5	33.9	41.1

Table 3: VOC 2010 Test set accuracies.

our estimates, it would take 10 years of CPU-time to compute each additional M-Best MAP solution for each image. On the other hand, computation of each additional solution for *Div*MBest takes the same time as MAP. We also report the **Random** (averaged over ten runs) and **Confidence** baselines.

Results. We evaluated all methods on the VOC 2010 dataset, consisting of train, validation and test partitions with 964 images each. We use the standard PASCAL pixelwise "intersection / union" performance measure for each category. Since labels for test are not directly available, we first compare different methods for obtaining multiple solutions on validation (having trained the model on train), and then report accuracy on test (having trained on trainval).

Given a set of candidate segmentations, one could *select* a single solution from the set, or *combine* them, obtaining a solution not equal to any of the candidates. We discuss means of achieving this automatically below; here we consider an "oracle" evaluation protocol designed to measure the upper bound on performance of any eventual selection mechanism.

On validation data, for which we have ground truth, we obtained the oracle accuracy automatically by selecting, for each image, the solution with the highest pixel label accuracy averaged over categories. Table 2 shows the result of this evaluation for *Div*MBest and the two sampling baselines, with $M = 10$. On test we manually selected the solution that was visually perceived to be the best among the M solutions. Since the baselines are clearly inferior, we only evaluated the oracle on test with *Div*MBest, $M = 10$, with the results in Table 3. With larger values of M on validation (Fig. 3) the oracle bound goes further up and with $M = 30$ the average accuracy on validation reaches about 48%, a near 7%-point improvement[2] over state of the art!

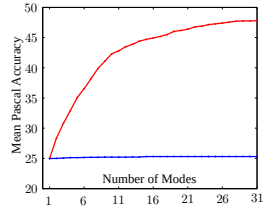


Fig. 3: Oracle accuracy vs. M on VOC2010 validation. Red: *Div*MBest, blue: **Confidence**.

[2] The winning entry of VOC2010 competition achieved 40.1% on test. Accuracies on validation and test aren't directly comparable, but performance of most techniques is higher on test than val, and the purpose of the comparison here is to give a qualitative idea of relative performance.

Ranking. The oracle results above indicate that the small set of diverse solutions from *Div*MBest has a lot of potential; to realize this potential one needs a ranking and/or combination mechanism. A method recently proposed in [19] is designed to rank and combine a very large pool of segments (single label masks) most of which are of poor quality. This is different from the situation here (very small number of full image segmentations, some of which are of high quality) but a similar method may be applicable, and we are investigating this direction.

6.3 Pose Estimation

Problem. We applied our formulation to the problem of tracking and estimating the pose of people in video sequences – a challenging problem due to appearance variation and articulation. We follow the setup of Park and Ramanan [25], where M candidate human poses are generated for every frame, and a single smooth track is selected using temporal context. We compare different methods that produce M-best pose estimates with respect to the quality of the tracks.

Model. The model we consider for this problem is the articulated part-based model of Yang and Ramanan [37], which has demonstrated competitive performance on various benchmarks. The variables in the model are part (head, body, etc) locations and type. The graph-structure is a tree and (exact) inference is performed by dynamic programming. We used the authors' implementation, and modified node potentials to produce Diverse M-Best solutions.

The tracking model [25] is a chain-CRF, where each frame is a node whose label is the choice of the mode. Node potentials prefer low energy modes and edge-potentials prefer smooth transitions between successive frames. Exact MAP inference on the chain-CRF was also performed via dynamic-programming.

Dataset, Baselines. We used the dataset of [25] which consists of four video sequences (walking, pitching, lola1, lola2) of varying lengths for which a few frames have been manually annotated with ground-truth limb locations. We compared *Div*MBest against the NBest method of [25], which is a special case of our formulation (as we describe in Section 4). Note that in [25] NBest was shown to outperform a number of sampling-based algorithms. We also compare to a **Confidence** baseline, where parts are repositioned to the next best position in the min-marginal table that hasn't been used for previously generated candidates. We omit **Random** since it is clearly inferior.

Results. Each algorithm was used to generate M candidate poses for all frames in every sequence. A track was computed using the chain-CRF to select a *single* pose for each frame. Algorithms were evaluated on the recovered tracks by computing *Percentage of Correct Parts* (PCP) scores (avg-CRF-PCP), which is shown in Fig. 4. *Div*MBest produced the best tracks for the vast majority of sequence-M combinations. For sequences lola1 and lola2 *Div*MBest significantly outperforms NBest by 10% and 4% points respectively.

We also evaluate "oracle" accuracies by selecting for each frame the most accurate (w.r.t. the ground-truth) pose among M (avg-Oracle-PCP), which is shown in Fig. 5. In general, we observed that *Div*MBest was able to produce better sets of hypotheses than NBest and Confidence across all values of M.

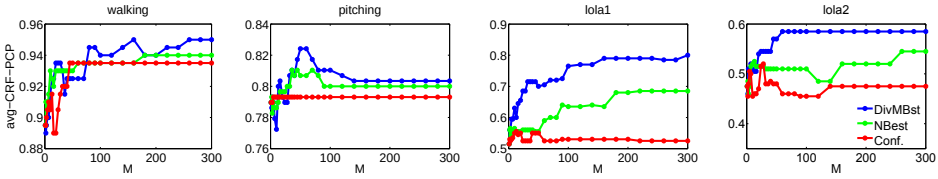


Fig. 4: Average PCP scores for the pose tracks vs. M (avg-CRF-PCP).

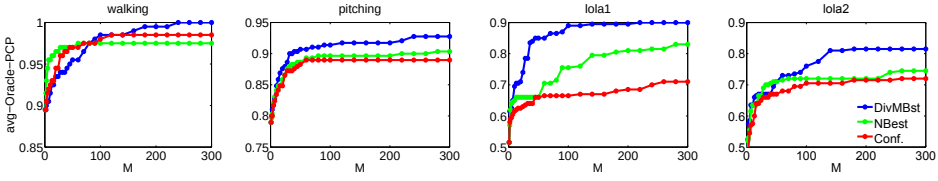


Fig. 5: Oracle PCP scores averaged across frames vs. M(avg-Oracle-PCP).

Finally, additional analysis, including average Hamming distance of the modes to the MAP solution, is included in the supplementary materials.

7 Discussions and Conclusion

We have presented the first algorithm for the Diverse M-Best MAP problem, which involves finding a diverse set of highly probable solutions under a discrete probabilistic model. Our Lagragian relaxation formulation involves solving the Δ-augmented energy minimization problem, minimizing a linear combination of the energy and similarity to previous solutions. We showed that this formulation is a generalization of the M-best MAP problem and that for certain classes of the Δ-function, *Div*MBest can be computed using the *same algorithms* as those developed for computing the MAP solution. With some of the models and inference algorithms commonly used in vision, this can be tremendously useful.

Currently researchers have to design sophisticated high-order models for images *and* clever optimization methods to allow for reasonably efficient inference under such models. Our work suggests a different paradigm: use simpler models in which exact or approximate MAP inference, and thus *Div*MBest inference with sufficiently nice Δ, is tractable, and obtain a set of diverse solutions. Then merely *evaluate* the more complex high-order model on these solutions to rank or otherwise combine them to provide the final output.

In this vein, our results bring into focus the problem of ranking a small set of diverse highly probably solutions. This is seen most dramatically in the results on VOC segmentation data; if one could only pick the most accurate segmentation in a pool of 30 segmentations for an image, one would improve the state of the art on VOC 2010 by 10%-points.

As future work we would like to investigate the performance and implications of other Δ-functions, apply them to higher-order energy functions and also apply this method to speed up cutting-plane methods for training Structural SVMs.

Acknowledgements. We thank Pushmeet Kohli for helpful discussions; Joao Carreira, Fuxin Li, Lubor Ladicky and Deva Ramanan for making their respective implementations publicly available and answering our questions.

References

1. A. Barbu and S.-C. Zhu. Generalizing swendsen-wang to sampling arbitrary posterior probabilities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:1239–1253, August 2005.
2. O. Barinova, V. Lempitsky, and P. Kohli. On detection of multiple object instances using hough transforms. In *CVPR*, pages 2233–2240, 2010.
3. M. Blaschko. Branch and bound strategies for non-maximal suppression in object detection. In *EMMCVPR*, pages 385–398, 2011.
4. L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Adv. in NIPS*, pages 161–168, 2008.
5. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.
6. Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. *ICCV*, 2001.
7. Y. Boykov, O. Veksler, and R. Zabih. Efficient approximate energy minimization via graph cuts. *PAMI*, 20(12):1222–1239, 2001.
8. N. F. Emma Rollon and R. Dechter. Inference schemes for m best solutions for soft csps. In *Proceedings of Workshop on Preferences and Soft Constraints*, 2011.
9. M. Fromer and A. Globerson. An LP view of the m-best MAP problem. In *NIPS*, 2009.
10. P. Kohli and M. P. Kumar. Energy minimization for linear envelope mrfs. In *CVPR*, pages 1863–1870, 2010.
11. P. Kohli and P. H. S. Torr. Effciently solving dynamic markov random fields using graph cuts. In *ICCV*, pages 922–929, 2005.
12. P. Kohli and P. H. S. Torr. Measuring uncertainty in graph cut solutions. *CVIU*, 112(1):30–38, 2008.
13. V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *PAMI*, 26(2):147–159, 2004.
14. N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order MRFs. In *CVPR*, 2009.
15. L. Ladickỳ, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. *ICCV*, 2009.
16. L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, pages 239–253, 2010.
17. L. Ladicky and P. H. Torr. The automatic labelling environment. http://cms.brookes.ac.uk/staff/PhilipTorr/ale.htm.
18. E. L. Lawler. A procedure for computing the k best solutions to discrete optimization problems and its application to the shortest path problem. *Management Science*, 18:401–405, 1972.
19. F. Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *CVPR*, 2010.
20. T. Meltzer, C. Yanover, and Y. Weiss. Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. In *ICCV*, pages 428–435, 2005.
21. E. R. Natalia Flerova and R. Dechter. Bucket and mini-bucket schemes for m best solutions over graphical models. In *IJCAI Workshop on Graph Structures for Knowledge Representation and Reasoning*, 2011.
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.
23. S. Nowozin and C. Lampert. Global connectivity potentials for random field models. In *CVPR*, 2009.
24. G. Papandreou and A. Yuille. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *ICCV*, pages 193–200, Nov. 2011.
25. D. Park and D. Ramanan. N-best maximal decoders for part models. In *ICCV*, 2011.
26. J. Porway and S.-C. Zhu. C^4: Exploring multiple solutions in graphical models by cluster sampling. *PAMI*, 33(9):1713 –1727, 2011.
27. S. Roth and M. Black. Fields of experts. *IJCV*, 82(2), April 2009.
28. C. Rother, P. Kohli, W. Feng, and J. Jia. Minimizing sparse higher order energy functions of discrete variables. In *CVPR*, pages 1382–1389, 2009.
29. C. Rother, V. Kolmogorov, and A. Blake. "Grabcut": interactive foreground extraction using iterated graph cuts. *SIGGRAPH*, 2004.
30. B. Seroussi and J. Golmard. An algorithm directly finding the k most probable configurations in bayesian networks. *Int. J. of Approx. Reasoning*, 11(3):205 – 233, 1994.
31. N. Shor. *Minimization methods for non-differentiable functions*. Springer series in computational mathematics. Springer-Verlag, 1985.
32. R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *PAMI*, 30(6):1068–1080, 2008.
33. D. Tarlow, I. E. Givoni, and R. S. Zemel. Hop-map: Efficient message passing with high order potentials. In *AISTATS*, pages 812–819, 2010.
34. I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005.
35. Z. Tu and S.-C. Zhu. Image segmentation by data-driven markov chain monte carlo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:657–673, May 2002.
36. T. Werner. A linear programming approach to max-sum problem: A review. *PAMI*, 29(7):1165–1179, 2007.
37. Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392, 2011.
38. C. Yanover and Y. Weiss. Finding the m most probable configurations using loopy belief propagation. In *NIPS*, 2003.
39. Y. Yue and T. Joachims. Predicting diverse subsets using structural SVMs. In *ICML*, pages 271–278, 2008.

Supplementary Materials: Diverse M-Best Solutions in Markov Random Fields

Dhruv Batra[1], Payman Yadollahpour[1],
Abner Guzman-Rivera[2], and Gregory Shakhnarovich[1]

[1]TTI-Chicago [2]UIUC

Abstract. In this supplementary document, we describe how approximate supergradient computation may be performed via message-passing (Section 1), provide a proof of Theorem 1 (Section 2), and give experimental details that could not be accommodated in the main manuscript (Section 3).

1 Higher-Order Potential (HOP) Dissimilarity and Supergradient Computation by Message-Passing

In this section, we describe how the Δ-augmented energy minimization problem may be solved, when Δ contains higher-order (or even global) potentials.

To simplify the exposition, let us focus only on the *Div*2Best case. Recall from the main manuscript (Sec. 5, Eqn. 4), that the supergradient computation involves solving the following Δ-augmented energy minimization problem:

$$\min_{\boldsymbol{\mu}\in\mathcal{L}(G),\ \boldsymbol{\mu}(\cdot)\in\{0,1\}} \quad \sum_{A\in\mathcal{V}\cup\mathcal{E}} \boldsymbol{\theta}_A\cdot\boldsymbol{\mu}_A - \lambda_1\Big(\Delta(\boldsymbol{\mu},\boldsymbol{\mu}^1) - k_1\Big) \tag{1a}$$

$$= \min_{\boldsymbol{\mu}\in\mathcal{L}(G),\ \boldsymbol{\mu}(\cdot)\in\{0,1\}} \quad \sum_{A\in\mathcal{V}\cup\mathcal{E}} \boldsymbol{\theta}_A\cdot\boldsymbol{\mu}_A - \lambda_1\,\Delta(\boldsymbol{\mu},\boldsymbol{\mu}^1) - \lambda_1 k_1. \tag{1b}$$

Higher-Order Potential (HOP) Dissimilarity. Let us consider the case when $\Delta(\boldsymbol{\mu},\boldsymbol{\mu}^1)$ does not decompose into unary terms like the Hamming dissimilarity. Moreover, let $\theta^1_{hop}(\boldsymbol{\mu}) \triangleq -\lambda_1\Delta(\boldsymbol{\mu},\boldsymbol{\mu}^1)$. Also notice that $\lambda_1 k_1$ is a constant w.r.t. $\boldsymbol{\mu}$ and may be ignored for this discussion. With this notation, we can write the Δ-augmented energy minimization problem as:

$$\min_{\boldsymbol{\mu}\in\mathcal{L}(G),\ \boldsymbol{\mu}(\cdot)\in\{0,1\}} \quad \sum_{A\in\mathcal{V}\cup\mathcal{E}} \boldsymbol{\theta}_A\cdot\boldsymbol{\mu}_A + \theta^1_{hop}(\boldsymbol{\mu}) \tag{2}$$

Now, even if MAP inference was tractable in the original model, this problem is difficult to solve due to the higher-order term θ^1_{hop}. However, if θ^1_{hop} has some structure, *e.g.* cardinality potentials [8,16], or lower linear-envelope potentials [10] or sparse (pattern-based) higher-order potentials [12,14], that allows for *messages* to be efficiently computed, this Δ-augmented energy minimization can be performed via dual-decomposition based message-passing algorithms.

Approximate Supergradient via Dual-Decomposition. In order to set solve (2) via message-passing, we follow the dual-decomposition approach [2,7]. The following steps are a straighforward adaptation from Komodakis *et al.* [13].

Let us first assign to the HOP term in (2), it's own copy of the variables:

$$\min_{\boldsymbol{\mu},\, \boldsymbol{\mu}^{hop}} \quad \sum_{A \in \mathcal{V} \cup \mathcal{E}} \boldsymbol{\theta}_A \cdot \boldsymbol{\mu}_A + \theta^1_{hop}(\boldsymbol{\mu}^{hop}) \tag{3a}$$

$$s.t. \quad \boldsymbol{\mu} \in \mathcal{L}(G) \tag{3b}$$

$$\boldsymbol{\mu}^{hop}_i = \boldsymbol{\mu}_i \qquad \forall i \in \mathcal{V} \tag{3c}$$

$$\boldsymbol{\mu}(\cdot), \boldsymbol{\mu}^{hop}(\cdot) \in \{0,1\} \tag{3d}$$

Notice that although the optimization is now performed over two different variables $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^{hop}$, the constraint (3c) forces them to always agree. Thus this formulation is equivalent to problem (2). Now, we can study the Lagrangian relaxation of this problem, formed by dualizing constraint (3c):

$$g(\boldsymbol{\delta}) \quad = \quad \min_{\boldsymbol{\mu},\, \boldsymbol{\mu}^{hop}} \quad \sum_{A \in \mathcal{V} \cup \mathcal{E}} \boldsymbol{\theta}_A \cdot \boldsymbol{\mu}_A + \theta^1_{hop}(\boldsymbol{\mu}^{hop}) - \sum_{i \in \mathcal{V}} \boldsymbol{\delta}_i \cdot \left(\boldsymbol{\mu}^{hop}_i - \boldsymbol{\mu}_i \right) \tag{4a}$$

$$s.t. \quad \boldsymbol{\mu} \in \mathcal{L}(G) \tag{4b}$$

$$\boldsymbol{\mu}_A(\cdot), \boldsymbol{\mu}^{hop}(\cdot) \in \{0,1\}, \tag{4c}$$

where $\boldsymbol{\delta}_i$ is the vector of Lagrangian multipliers for each node $i \in \mathcal{V}$. Notice that dualizing the constraint (3c) decouples the above problem into two independent problem, one over the energy of the MRF and another over the HOP terms:

$$g(\boldsymbol{\delta}) = \min_{\boldsymbol{\mu} \in \mathcal{L}(G),\ \boldsymbol{\mu}(\cdot) \in \{0,1\}} \left\{ \sum_{i \in \mathcal{V}} (\boldsymbol{\theta}_i + \boldsymbol{\delta}_i) \cdot \boldsymbol{\mu}_i + \sum_{e \in \mathcal{E}} \boldsymbol{\theta}_e \cdot \boldsymbol{\mu}_e \right\} \tag{5a}$$

$$+ \min_{\boldsymbol{\mu}^{hop}(\cdot) \in \{0,1\}} \left\{ \theta^1_{hop}(\boldsymbol{\mu}^{hop}) - \sum_{i \in \mathcal{V}} \boldsymbol{\delta}_i \cdot \boldsymbol{\mu}^{hop}_i \right\} \tag{5b}$$

Now, just as we describe in the main manuscript, we can search for the tightest Lagrangian relaxation: $\max_{\boldsymbol{\delta}} g(\boldsymbol{\delta})$, which may be solved via supergradient ascent.

Notice that the first term (5a) involves solving MAP with perturbed node energies and can be solved with the same algorithm (*e.g.* graph-cuts) used for computing MAP. Moreover, the second term (5b) is efficiently computable for HOPs with structure, *e.g.* Gupta *et al.* [8] and Tarlow *et al.* [16] showed how messages may be computed for cardinality potentials.

In our interactive segmentation experiments, we used a cardinality-based HOP (describe in Section 3.1) and used the HOPMAP implementation of Tarlow *et al.* [16] to solve the Δ-augmented inference problem.

2 Proof of Theorem 1

In this section, we provide the proof of Theorem 1 from the main manuscript. Recall that the *Div*MBEST primal problem is:

$$\min_{\boldsymbol{\mu}} \quad \sum_{A\in\mathcal{V}\cup\mathcal{E}} \boldsymbol{\theta}_A \cdot \boldsymbol{\mu}_A \tag{6a}$$

$$s.t. \quad \boldsymbol{\mu} \in \mathcal{L}(G) \tag{6b}$$

$$\mu_A(\cdot) \in \{0,1\} \tag{6c}$$

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^m) \geq k_m \qquad \forall m \in \{1,\dots,M-1\}. \tag{6d}$$

Also recall that the Lagrangian relaxation is given by following:

$$f(\boldsymbol{\lambda}) \quad = \quad \min_{\boldsymbol{\mu}\in\mathcal{L}(G),\ \mu_A(s)\in\{0,1\}} \quad \sum_{A\in\mathcal{V}\cup\mathcal{E}} \boldsymbol{\theta}_A \cdot \boldsymbol{\mu}_A - \sum_{m=1}^{M-1} \lambda_m \Big(\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^m) - k_m\Big), \tag{7}$$

where $\boldsymbol{\lambda} = \{\lambda_m \mid m \in [M-1]\}$ is the vector of Lagrange multipliers. Finally, recall that the dual problem is given by $\max_{\boldsymbol{\lambda}\geq 0} f(\boldsymbol{\lambda})$.

Theorem 1. *i) The Lagrangian dual* $\max_{\boldsymbol{\lambda}\geq 0} f(\boldsymbol{\lambda})$ *is equivalent to solving (*i.e. *has zero duality gap with) the following primal relaxation of Div*MBEST $(\Delta, \boldsymbol{k})$*:*

$$\min_{\boldsymbol{\mu}} \quad \sum_{A\in\mathcal{V}\cup\mathcal{E}} \boldsymbol{\theta}_A \cdot \boldsymbol{\mu}_A \tag{8a}$$

$$s.t. \quad \boldsymbol{\mu} \in Co\Big\{\mu_A(s) \in \{0,1\} \mid \boldsymbol{\mu} \in \mathcal{L}(G)\Big\} \tag{8b}$$

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^m) \geq k_m \qquad \forall m \in \{1,\dots,M-1\}, \tag{8c}$$

where $Co\{\cdot\}$ *denotes the convex hull of the discrete solutions.*

ii) For certain specific Δ*-functions (*e.g. *the 0-1 function of M-Best MAP), the convex hull can be replaced by the discrete solutions themselves,* i.e. *(8b) may be replaced by* $\mu_A(s) \in \{0,1\}$, $\boldsymbol{\mu} \in \mathcal{L}(G)$*. Thus for such specific families, the Lagrangian relaxation* is *tight.*

iii) However, if no assumptions are made on Δ*, the Lagrangian relaxation is not guaranteed to be tight and may leave a duality gap.*

Proof. i) This is a direct application of the result of Geoffrion [6]. Geoffrion [6] showed that the following two programs are LP-duals:

$$(Dual) \qquad \max_{\lambda\geq 0} \quad \min_{x\geq 0} \quad c^T x - \lambda^T (Ax - b) \tag{9a}$$

$$s.t. \quad Bx \geq d \tag{9b}$$

$$x_j \ \text{ integer}, \ j \in \mathcal{I}, \tag{9c}$$

and

$$(Primal) \qquad \min_{x} \quad c^T x \tag{10a}$$

$$s.t. \quad Ax \geq b \tag{10b}$$

$$x \in Co\left\{x \geq 0,\ Bx \geq d,\ x_j \text{ integer},\ j \in \mathcal{I}\right\}, \tag{10c}$$

where $Co\{\cdot\}$ is the convex hull operator.

We can now map the variables in Geoffrion's result to the *Div*MBest formulation to get the desired result.

ii) In order to know when the Lagrangian relaxation is tight, we need to study when the convex hull operator $Co\{\cdot\}$ may be removed in (8b). First, note that $Co\left\{\mu_A(s) \in \{0,1\} \mid \boldsymbol{\mu} \in \mathcal{L}(G)\right\}$ is precisely the marginal polytope [17], *i.e.* the set of marginal distributions realizable over the graphical model.

Moreover, recall that minimizing a linear objective over the convex hull of discrete points is equivalent to directly minimizing over the discrete solution. Thus, in the absence of diversity constraints, we could easily remove the $Co\{\cdot\}$ operator.

In the presence of diversity constraints, one *sufficient* condition for removing the convex hull operator $Co\{\cdot\}$ is the following – if the set of feasible solutions in the Lagrangian primal relaxation (8), *i.e.* $P = \Big\{\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^m) \geq k_m \cap Co\left\{\mu_A(s) \in \{0,1\} \mid \boldsymbol{\mu} \in \mathcal{L}(G)\right\}\Big\}$, is a polytope with integral vertices, then the convex hull operator $Co\{\cdot\}$ may be removed.

When is this sufficient condition true? Since the marginal polytope always has integral vertices, we just need to check if the diversity constraints $\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^m) \geq k_m$ introduce any fractional vertices. For the M-Best MAP diversity function, Fromer and Globerson [5] presented *spanning-tree inequalities* that were guaranteed to not introduce any fractional vertices for the case of tree models with $M = 2$.

iii) In the absence of any assumptions, the Lagrangian relaxation not guaranteed to be tight. Specifically, the (unweighted) Hamming dissimilarity used in our experiments does not result in a tight relaxation. Fromer and Globerson [5] presented a simple counter-example which we describe here.

Counter-example. Consider a 2-node 1-edge MRF, where each node takes 2 labels. Let the node energies be non-informative, *i.e.* $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = (0,0)$. Let the edge energy prefer the $(0,0)$ configuration, *i.e.* $\boldsymbol{\theta}_{1,2} = (0,10,10,10)$. Clearly the MAP state is $(0,0)$. Now, the search for second best mode would add the following inequality: $-\mu_1(0) - \mu_2(0) \geq -1 \implies \mu_1(0) + \mu_2(0) \leq 1$. Solving with this additional constraint produces the fractional solution $(0.5, 0.5)$ (we can easily verify that $(0.5, 0.5)$ achieves an energy of 5, while all integer states other than MAP achieve an energy of 10). Thus the Lagrangian relaxation not tight in this case.

3 Experimental Details

In this section, we provide additional experimental details that could not be accommodated in the main manuscript.

3.1 Interactive Segmentation

The dataset consists of 100 images from Pascal VOC 2010 [4], with corresponding annotations ("scribbles") indicating foreground/background superpixels[1] in each image. Fifty of the images were used for tuning the regularization weight, λ, and the other 50 for testing. Grid search was performed over values of λ in the range $[0, 1]$. The best results on the training images was achieved with $\lambda = .18$, which we fixed for experiments on the test set.

Energy. Consider an image-scribble pair $(\mathbf{x}, \mathcal{S})$, where each image is a collection of n superpixels to be labelled as either foreground or background, *i.e.* $\mathbf{x} = \{x_i \mid i \in [n]\}$ where each $x_i \in \{fg, bg\}$, and $\mathcal{S} \subset [n]$ is a subset of superpixels for whom labels are known. We chose to use superpixels as opposed to pixels for computational efficiency and in order to produce segmentations that were better aligned to boundaries in the image. We build a graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, over the superpixels with edges connecting adjacent superpixels. The pairwise MRF energy for this application is given by:

$$E(\mathbf{x} : \mathcal{A}) = \sum_{i \in \mathcal{V}} \theta_i(x_i : \mathcal{A}) + \lambda \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(x_i, x_j), \tag{11}$$

where the first (data) term is the cost for assigning a superpixel to the foreground or background classes, while the second (smoothness) term is the penalization for having different labelings of neighboring superpixels. The data term depends on an appearance model $\mathcal{A}$ learned from the user scribbles ($\mathcal{S}$) and is defined below.

Data Term. The unary appearance model is based on the output of a linear Transductive SVM (TSVM). Specifically, we extract feature vectors $\phi(x_i)$ from both labeled and unlabeled superpixels and train a TSVM [15] to learn the appearance model. We extract colour features (C1-C4 as proposed by [9]), a histogram of gradients (HOG features), and a histogram over SIFT codewords. Let $\mathbf{w}$ be the learnt weight vector from the TSVM and $s_i = \mathbf{w}^T \phi(x_i)$ be the score for each superpixel. The resulting data term for the foreground label is of the following form:

$$\theta_i(x_i = fg) = \begin{cases} \eta, & \text{if } s_i \geq 0 \\ 1 - \eta, & \text{otherwise'} \end{cases} \tag{12}$$

[1] We used SLIC [1] to extract superpixels, with the desired number of superpixels in an image set to 3000. The images contained roughly 150K-200K pixels.

(correspondigly $\theta_i(x_i = bg) = 1 - \theta_i(x_i = fg)$), and $\eta = 0.5\, e^{\frac{-|s_i|^2}{\alpha\sigma^2}}$. We set $\sigma^2 = var\left(\{s_j \,|\, j \in \mathcal{S}\}\right)$, and α is set to a hand tuned constant fixed for all images.

Smoothness Term. The pairwise smoothness term we used is the contrast sensitive Potts model:

$$\theta_{ij}(x_i, x_j) = [\![x_i \neq x_j]\!] \cdot \beta_1 \cdot e^{-\beta_2 d_{ij}}, \tag{13}$$

where $[\![\cdot]\!]$ is an indicator function that is 1 when its argument evaluates to true and 0 otherwise, d_{ij} is the distance between feature vectors at superpixels i and j, and β_1, β_2[2] are scale parameters. The effect of this smoothness term is to penalize for label transitions between neighboring superpixels, but the penalization decreases as the distance in feature space between the corresponding feature vectors increases.

Inference. The contrast-sensitive Potts model results in a submodular energy function for a two-label problem so we can efficiently compute the MAP solution using the publicly available graph-cut implementation of [3, 11]. Moreover, with negative dot-product dissimilarity, we can efficiently and optimally compute the Diverse M-Best solutions using graph-cuts as well.

Higher-Order Potential (HOP) Dissimilarity. In additional to the (unweighted) Hamming dissimilarity, we also experimented with a HOP dissimilarity. Let $\#\boldsymbol{\mu} \triangleq \sum_{i \in \mathcal{V}} \mu_i(1)$ denote the number of nodes that are set to label 1. Note that this is the size of the foreground object cut out by the inference algorithm. The HOP dissimilarity is given by:

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu^{(1)}}) = \begin{cases} \left(\#\boldsymbol{\mu} - \#\boldsymbol{\mu^{(1)}}\right)^2 & \text{if } \#\boldsymbol{\mu} \geq \#\boldsymbol{\mu^{(1)}} \\ 0 & \text{else} \end{cases} \tag{14}$$

Intuitively, we can see that the above Δ function is 0 as long as the size of the foreground in the second mode is *smaller* than the size of the foreground in the MAP, and increases quadratically if it is larger. Thus, this dissimilarity function prefers the foreground object in the second solution to be *larger* than its size in the MAP. Notice that this is a global potential that depends on all nodes in the CRF and does not decompose into subsets. However, since this is a cardinality potential [8, 16], messages can be efficiently computed from this potential and an approximate solution can be computed for the Δ-augmented energy minimization problem. We use the HOPMAP implementation of Tarlow *et al.* [16] to solve this problem.

Fig. 1 shows the segmentations extracted from this size-based HOP dissimilarity, where we extract larger foregrounds and smaller foregrounds (by changing the sign of inequality in the definition of Δ) than the MAP.

[2] In the experiments $\beta_1 = 2$ and $\beta_2 = \sqrt{.05 \max\{d_{ij}\}}$.

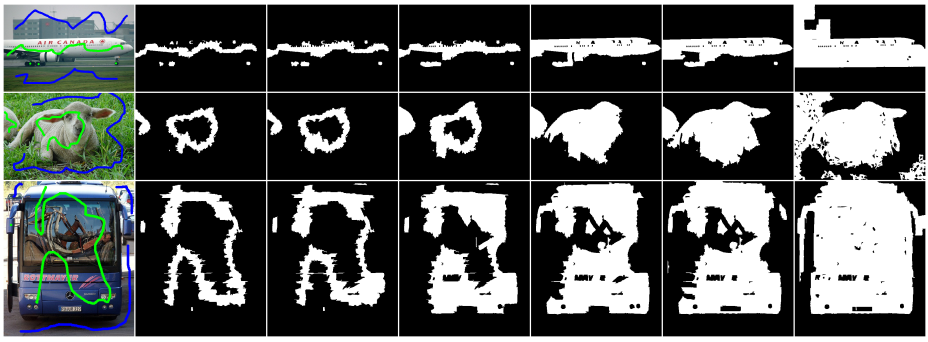
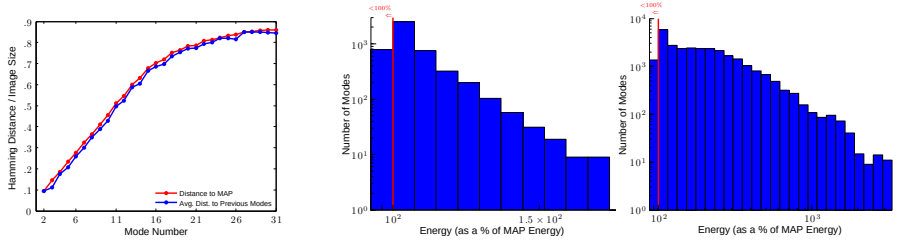


Fig. 1: Size-based HOP Diversity: Figure shows modes extracted from the cardinality-based HOP described in text. We can see that the segmentations are ordered according to the size of the foreground object.

3.2 Category level segmentation

In this section we report additional results on the category level segmentation experiment. In Figure 2 we show the distance of the modes to MAP (and to previous modes), along with the energy of the modes (as % of MAP) on validation images of PASCAL VOC 2010. The first plot shows that on average the Hamming distance of the modes does in fact monotonically increase. The second plots confirm our initial hypothesis that most of the additional modes have significantly higher energy than the MAP. Note that since this model allows only approximate inference, a very small percentage of the modes do have lower energy than the MAP, something that would never happen if the inference was exact.



(a) (b) (c)

Fig. 2: (*a*) Mean hamming distances between each mode and the MAP solution (red), and average to previous modes (blue), normalized by image size on PASCAL VOC 2010 validation set. Also show, histogram of energies (as % of MAP) over (*b*) 6 modes, (*c*) 31 modes, on validation set. The bar to the left of red vertical lines indicate number of modes with energy less than or equal to MAP.

Name	length	gt
walking	125	20
pitching	300	29
lola1	80	20
lola2	80	20

Table 1: Dataset for Pose Estimation.

Below in Figures 5 through 10 we show a sample of results for validation images of PASCAL VOC 2010 set. For each image, the figure includes the input image, the ground truth, the MAP and additional 9 modes obtained with *Div*MBest. For each solution we show the accuracy: pixelwise intersection over union accuracy for each class, averaged over classes present in the ground truth and/or in the solution. Refer to Fig. 3 to interpret color labels for the categories.

Fig. 3: Color map for reading VOC segmentation results.

3.3 Pose Estimation

Table 1 reports the number of total and annotated frames in our dataset, and Fig. 11 shows the average Hamming distance to the MAP solution for modes 2 to 300. As in the previous experiment, we can see that the average hamming distances monotically increase. Note that for these experiments, the Hamming distance between two solutions cannot be larger than 26, which is the number of nodes in the graphical model.

References

1. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC Superpixels. Technical report, EPFL, 2010. 5
2. D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, September 1999. 2
3. Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:359–374, 2001. 6
4. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html. 5
5. M. Fromer and A. Globerson. An LP view of the m-best MAP problem. In *NIPS*, 2009. 4
6. A. M. Geoffrion. Lagrangean Relaxation for Integer Programming. *Mathematical Programming Study*, 2:82–114, 1974. 3
7. M. Guignard. Lagrangean relaxation. *TOP: An Official Journal of the Spanish Society of Statistics and Operations Research*, 11(2):151–200, 2003. 2

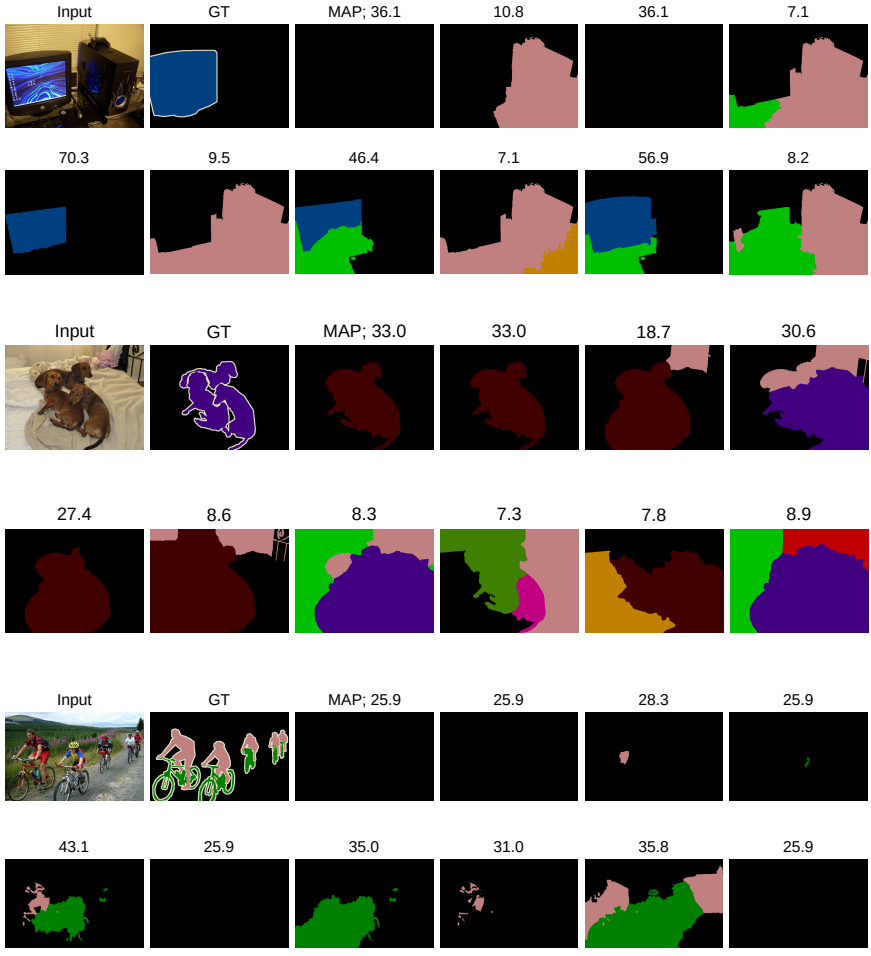


Fig. 4: Example segmentations on validation, PASCAL VOC 2010

8. R. Gupta, S. Sarawagi, and A. A. Diwan. Collective inference for extraction mrfs coupled with symmetric clique potentials. *J. Mach. Learn. Res.*, 11:3097–3135, 2010. 1, 2, 6
9. D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *In ICCV*, pages 654–661, 2005. 5
10. P. Kohli and M. P. Kumar. Energy minimization for linear envelope mrfs. In *CVPR*, pages 1863–1870, 2010. 1
11. P. Kohli and P. H. S. Torr. Effciently solving dynamic markov random fields using graph cuts. In *ICCV*, pages 922–929, 2005. 6
12. N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order MRFs. In *CVPR*, 2009. 1
13. N. Komodakis, N. Paragios, and G. Tziritas. MRF optimization via dual decomposition: Message-passing revisited. In *ICCV*, 2007. 2
14. C. Rother, P. Kohli, W. Feng, and J. Jia. Minimizing sparse higher order energy functions of discrete variables. In *CVPR*, pages 1382–1389, 2009. 1
15. V. Sindhwani and S. S. Keerthi. Large scale semi-supervised linear svms. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 477–484, New York, NY, USA, 2006. ACM. 5

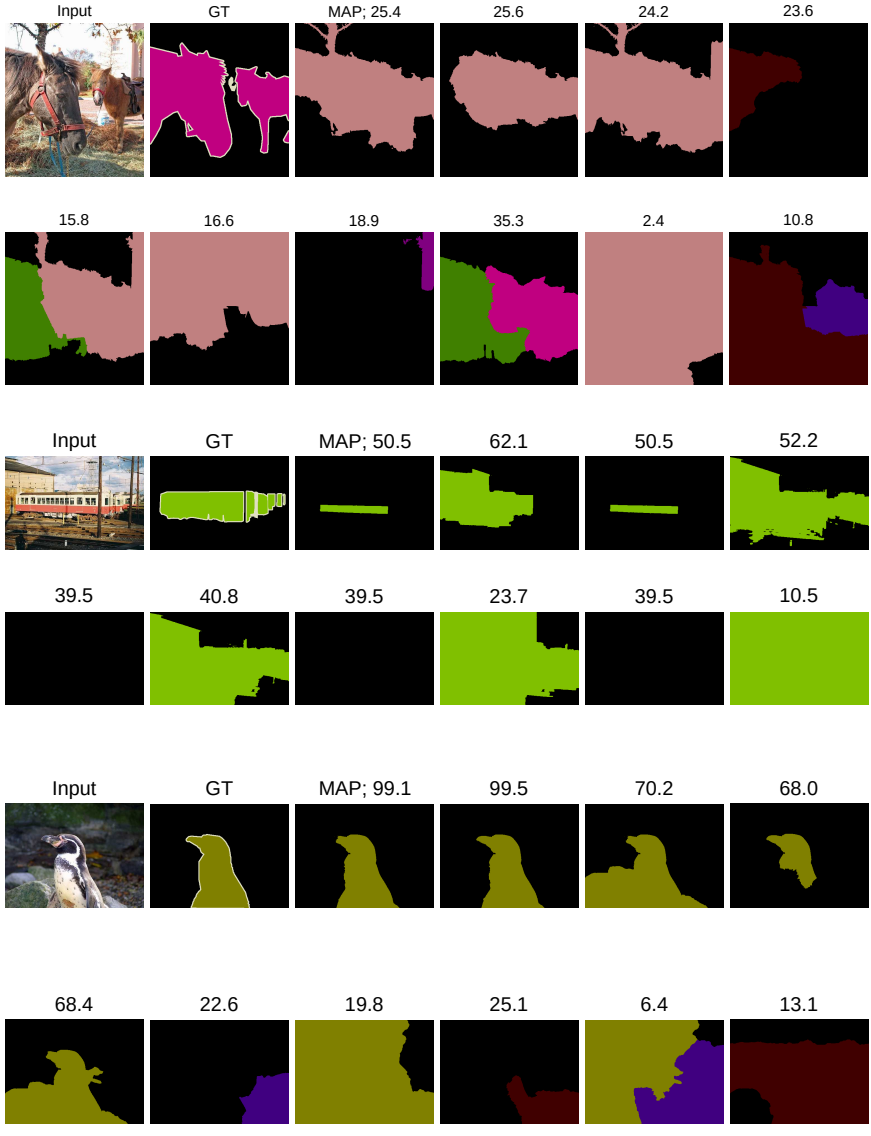


Fig. 5: More example segmentations on validation, PASCAL VOC 2010

16. D. Tarlow, I. E. Givoni, and R. S. Zemel. Hop-map: Efficient message passing with high order potentials. In *AISTATS*, pages 812–819, 2010. 1, 2, 6
17. M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008. 4

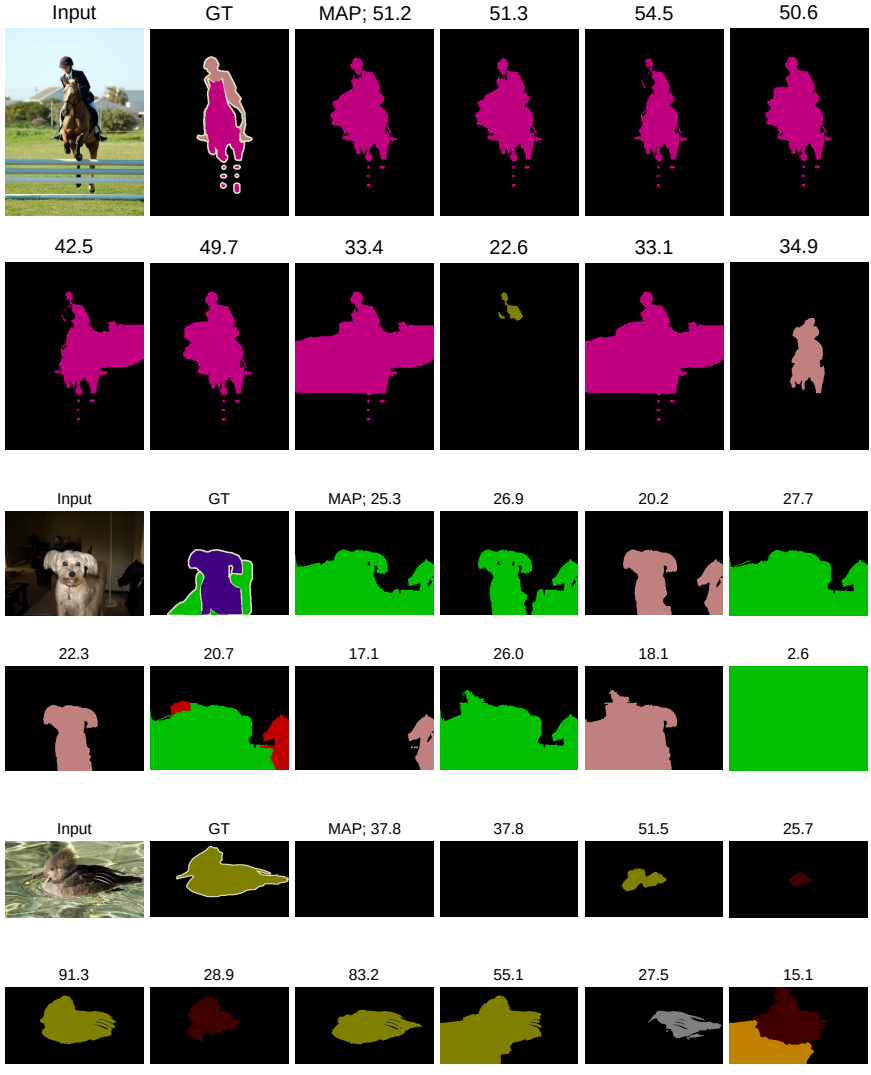


Fig. 6: More example segmentations on validation, PASCAL VOC 2010

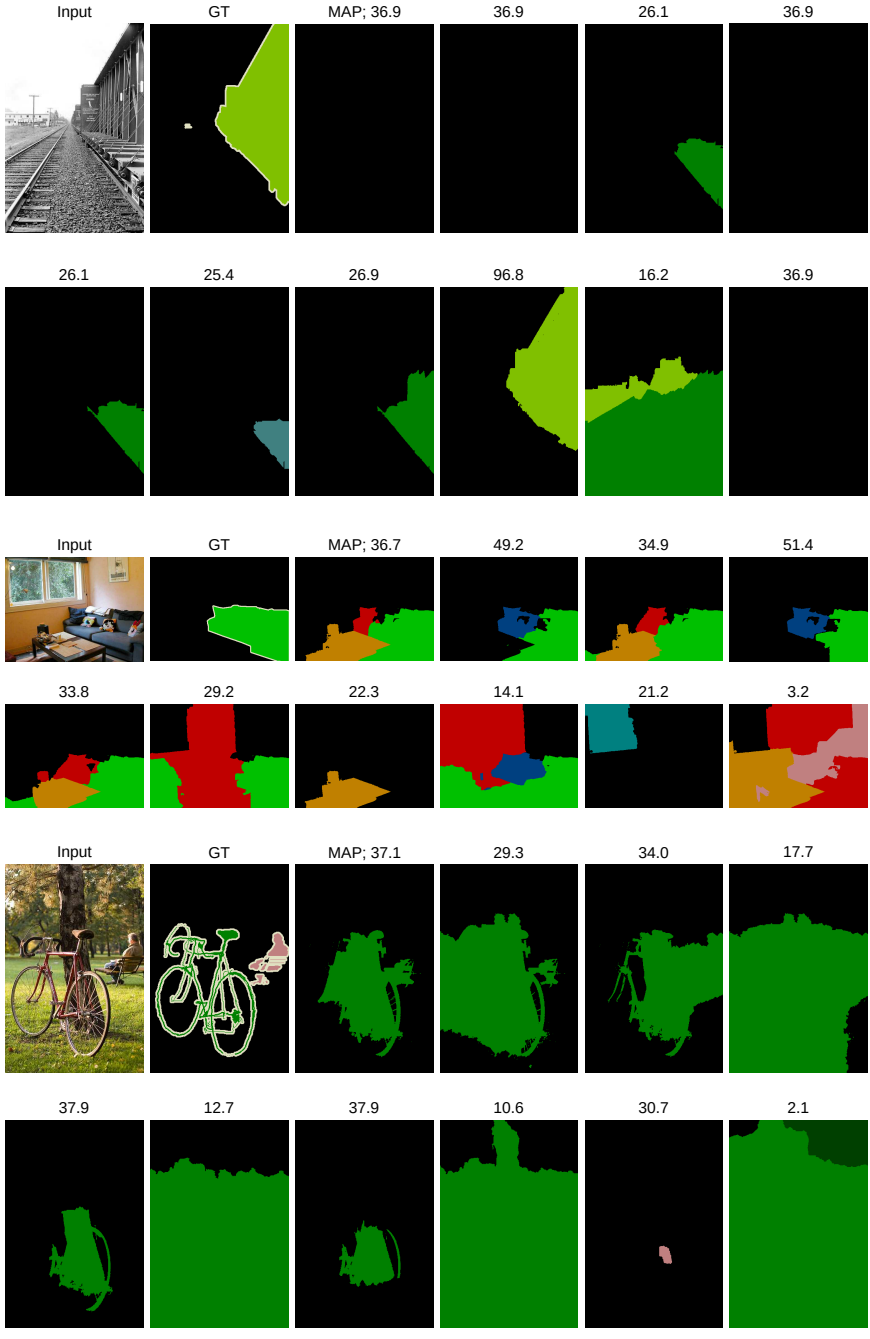


Fig. 7: More example segmentations on validation, PASCAL VOC 2010

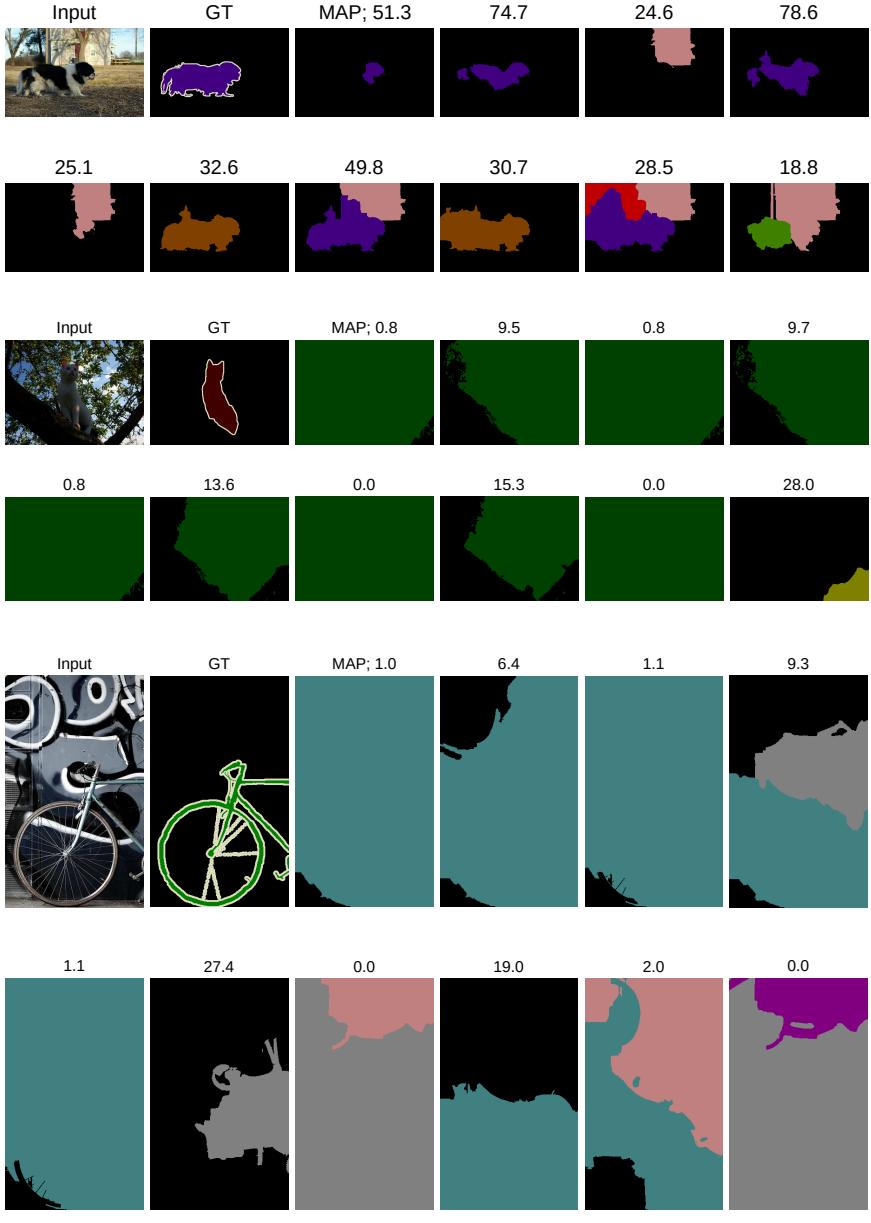


Fig. 8: More example segmentations on validation, PASCAL VOC 2010

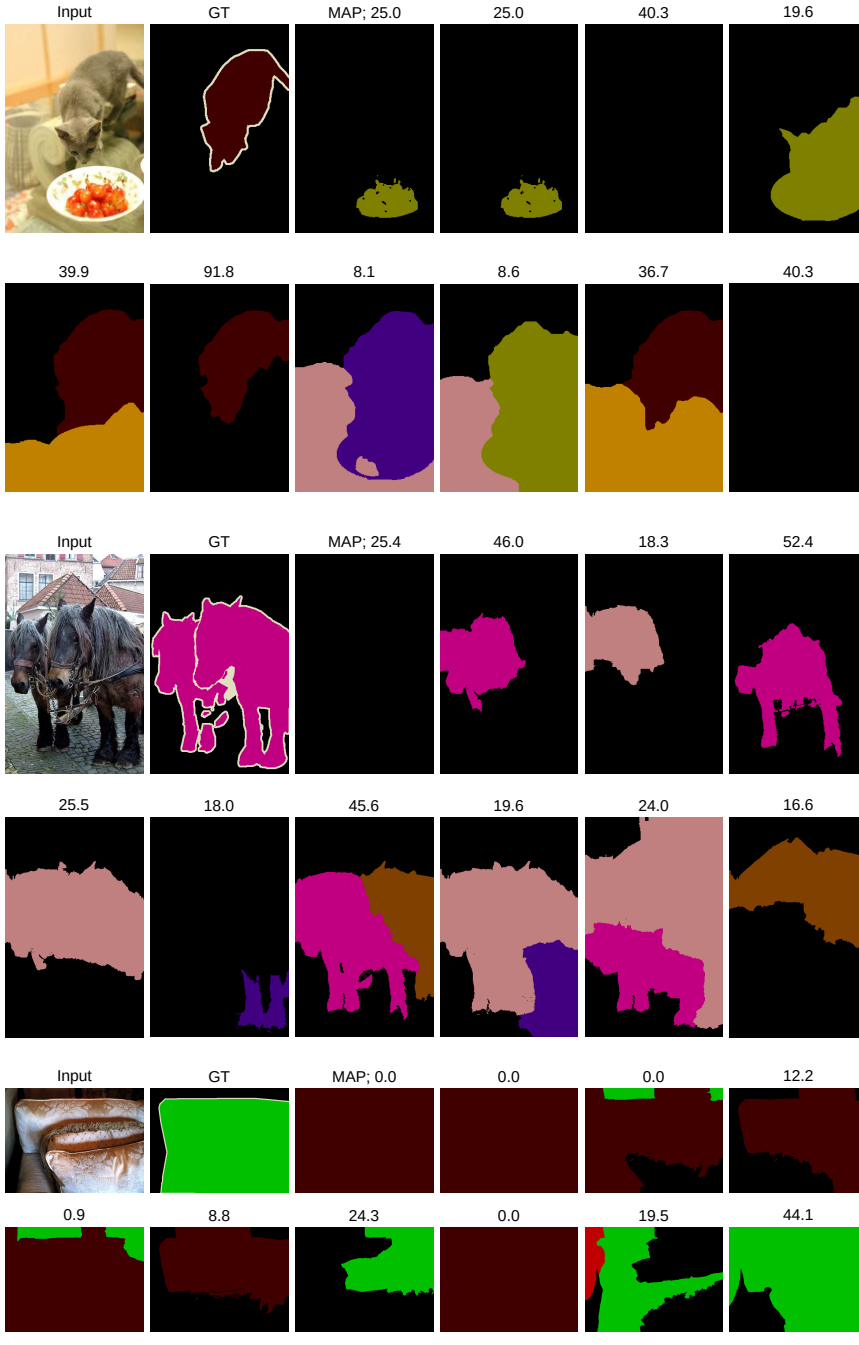


Fig. 9: More example segmentations on validation, PASCAL VOC 2010

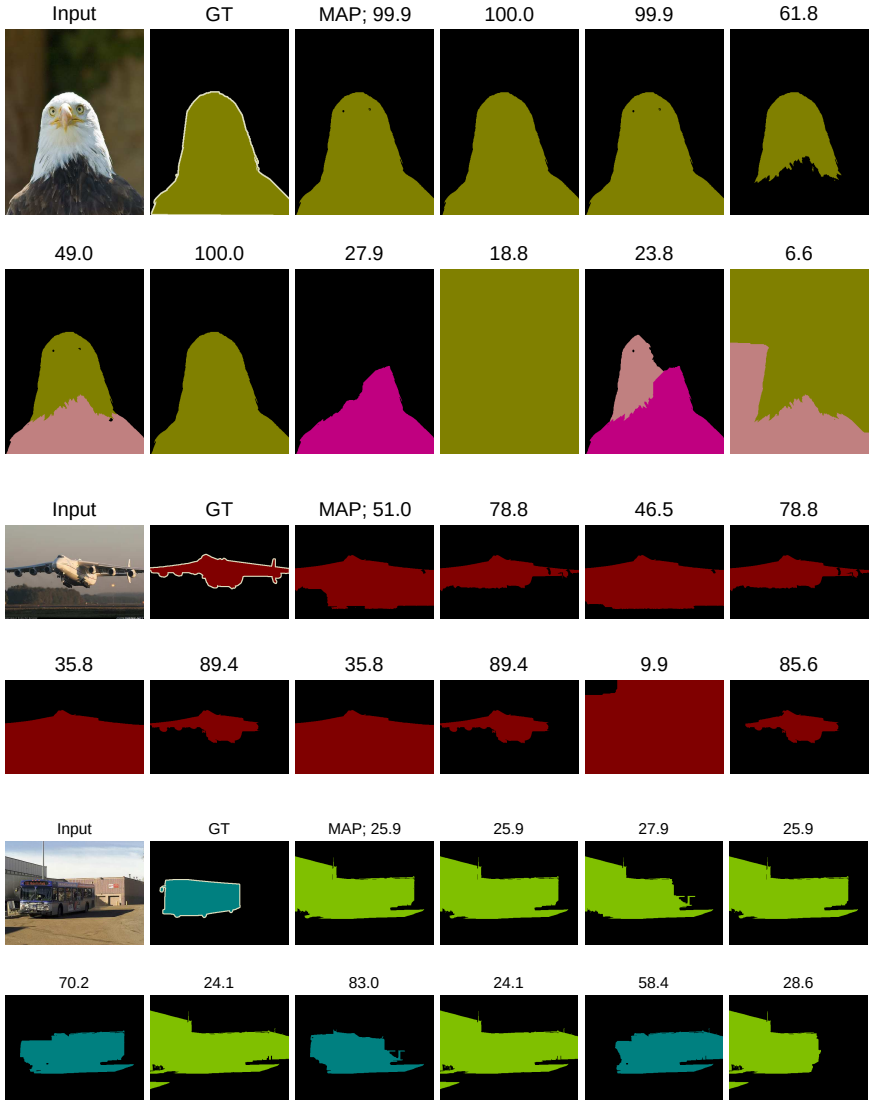


Fig. 10: More example segmentations on validation, PASCAL VOC 2010

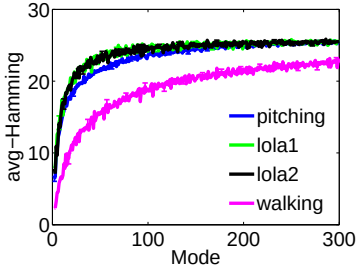


Fig. 11: Average Hamming distance to MAP for each mode.