Inferring M-Best Diverse Labelings in a Single One

Alexander Kirillov[1], Bogdan Savchynskyy[1], Dmitrij Schlesinger[1], Dmitry Vetrov[2], Carsten Rother[1]
[1] TU Dresden, Dresden, Germany; [2] Skoltech, Moscow, Russia; vetrovd@yandex.ru
{alexander.kirillov,bogdan.savchynskyy,dmytro.shlezinger,carsten.rother}@tu-dresden.de

Abstract

We consider the task of finding M-best diverse solutions in a graphical model. In a previous work by Batra et al. an algorithmic approach for finding such solutions was proposed, and its usefulness was shown in numerous applications. Contrary to previous work we propose a novel formulation of the problem in form of a single energy minimization problem in a specially constructed graphical model. We show that the method of Batra et al. can be considered as a greedy approximate algorithm for our model, whereas we introduce an efficient specialized optimization technique for it, based on alpha-expansion. We evaluate our method on two application scenarios, interactive and semantic image segmentation, with binary and multiple labels. In both cases we achieve considerably better error rates than state-of-the art diversity methods. Furthermore, we empirically discover that in the binary label case we were able to reach global optimality for all test instances.

1. Introduction

A large variety of computer vision tasks can be formulated in the form of an energy minimization problem, known also as *maximium a posteriori* (MAP) inference in an undirected graphical models (related to Markov or conditional random fields). Its modeling power and importance are well-recognized, which recently resulted into specialized benchmarks for its solvers [31, 17]. This underlines the importance of finding *the most probable* variable configuration. Following [4] we argue, however, that finding $M > 1$ *diverse* configurations with low energies is also of importance in a number of scenarios, such as: (a) Expressing uncertainty of the found solution [27]; (b) Faster training of model parameters [14]; (c) Ranking of inference results [35]; (d) Empirical risk minimization [26].

It is important to note that in many application scenarios, such as [35], the diverse solutions are computed for an energy that has been trained in a discriminative fashion such that the solution with lowest energy should correspond to the most accurate result. We will also use such pre-trained energies. However, we observe that at test time a result which is different to the lowest energy solution, but still has a low energy, can achieve higher accuracy. This has also been observed in previous works. The task of training one or more energies for producing optimal M diverse solutions, as e.g. in [13, 15], is not the subject of this work.

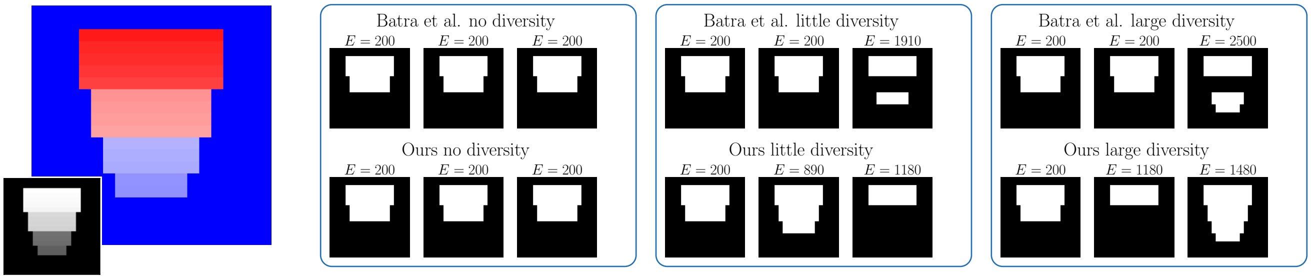
In this work we propose *a novel* formulation for the problem of finding *M-best-diverse* solutions as a MAP-inference problem in a specially constructed graphical model. Any variable configuration in this model corresponds to M solutions of the original problem and the best configuration then corresponds to *the M best diverse* solutions. We introduce an efficient, specialized solver for our model, although other standard MAP-inference techniques are potentially applicable as well. In fact, we empirically observe that with this solver is able to reach global optimality for all test instances of a binary labelling problem.

Related work. The importance of the considered problem is demonstrated by the number of works addressing it from different perspectives.

A procedure of computing *M-best solutions* to discrete optimization problems was proposed in [21], which dates back to 1972. Later, more efficient specialized procedures were introduced for MAP-inference on a tree [30, Ch. 8], junction-trees [22] and general graphical models [36, 11, 3]. These methods are well-suited for certain scenarios, however in typical structured computer vision problems (like e.g. pixel-level image segmentation) M-best solutions differ from each other only by a small number of variables (pixels) and from an application point of view are all equivalent and hence practically useless.

Sampling methods allow to approximate marginal probabilities and therefore can be used for estimating solutions uncertainty. Though the methods like [24, 33] are designed to address different modes of the underlying distribution, they do not enforce diversity explicitly, hence can hardly be used for faster discriminative training of model parameters

This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 647769). D. Vetrov was supported by RFBR proj. (No. 15-31-20596) and by Microsoft (RPD 1053945).



(a) **(b)** **(c)** **(d)**

Figure 1: **Comparing our approach to Batra et al. [4] for a toy image.** (a) the unary terms of the energy (red/blue means more likely fore-/background). The Inlet shows the original image. (b-d) Results for three different levels of diversity. For each diversity level we ask for three output images, where E is the associated energy. (b) Both methods produce three times the same MAP solution. (c) When forcing diversity both methods give different results. Note that the sum of the pairwise Hamming distances between the results is the same for both methods. However, our result is visually better, since it contains solutions that are more coherent. This is also reflected in the energy. The sum of the energy of the three solutions is lower with our method (2270 compared to 2310). Our method directly optimizes for this energy. (d) When forcing strong diversity, the set of solutions becomes more diverse for both methods. Again, our result is superior, visually and in terms of total energy, while both results have the same overall Hamming distance.

or ranking of inference results. Local techniques like e.g. Gibbs sampling [12] may take prohibitively long to transfer between modes of the underlying distribution. *Perturb-and-map* [23] does not have these drawbacks, but is limited to the cases where *an exact* MAP-solution can be obtained relatively fast due to the need of its multiple computation. Indeed, this method can be seen as the closest probabilistic counterpart to the deterministic ones considered in this work.

Structured Determinant Point Processes [20] is a tool to model probabilistic distributions over structured models. Unfortunately an efficient sampling procedure is feasible for tree-structured graphical models only. The recently proposed algorithm [7] to find M *best modes* of a distribution is limited to the same narrow class of problems.

Training of M *independent* graphical models to produce diverse solutions was proposed in [13, 15]. In contrast, we assume *a single fixed* model supporting reasonable MAP-solutions.

The most relevant for us is the work [4], addressing the problem of *the* M *best diverse* solutions to energy minimization. It proposes an algorithm, which starts with finding a MAP-solution. On each iteration it penalizes already found labelings and obtains the next optimal one. The penalization enforces diversity of the found solutions. The greedy character of the method is its main disadvantage, which leads (as we show in Section 5) to suboptimal results. The recent follow-up work [25] proposes a subclass of new diversity penalties, for which the greedy nature of the algorithm can be substantiated.

Contribution. We formulate the problem of finding M *best diverse* solutions to energy minimization as a problem that has the same format as the energy minimization itself. In other words, *a single* labeling in our specially constructed graphical model corresponds to M labelings in the initial model. Based on this formulation we show that

(i) the algorithm proposed in [4] (and used in [14, 27, 35, 26]) can be viewed as an approximate greedy energy minimization to our model;

(ii) if the initial MAP-inference problem was (approximately) solvable with α-expansion or α-β-swap our model, delivering M best diverse solutions, maintains this property. Furthermore, we empirically found that in case the original energy was binary and submodular, we were always able to minimize the energy of our model *exactly*.

We demonstrate superiority of our approach in terms of the quality of found solutions on several computer vision datasets published in [4] and [25].

Paper structure. In Section 2 we briefly describe the diversity method of Batra et al. [4]. Section 3 is devoted to *an explicit* formulation of our novel diversity model. Here we also provide an overview of existing diversity measures and show that the method [4] can be seen as a greedy inference for it. In Section 4 we present a reformulation of our model, which allows for efficient inference with graph-cuts and LP-relaxation based techniques. Finally, Sections 5 and 6 are devoted to the experimental evaluation and conclusions.

2. DivMBest Method of Batra et al. [4]

Preliminaries. Let $2^{\mathcal{A}}$ denote the powerset of a set $\mathcal{A}$. The pair $\mathcal{G} = (\mathcal{V}, \mathcal{F})$ is called a *factor graph* and has $\mathcal{V}$ as a finite *set of variable nodes* and $\mathcal{F} \subseteq 2^{\mathcal{V}}$ as *a set of factors*. Each variable node $v \in \mathcal{V}$ is associated with a *variable* y_v taking its values in a finite *set of labels* L_v. The set $L_{\mathcal{A}} = \prod_{v \in \mathcal{A}} L_v$ denotes a Cartesian product of sets of labels corresponding to the subset $\mathcal{A} \subseteq \mathcal{V}$ of variables. Functions $\theta_f\colon L_f \to \mathbb{R}$, associated with factors $f \in \mathcal{F}$, are called *potentials* and define local costs on values of variables and their combinations. The set $\{\theta_f\colon f \in \mathcal{F}\}$ of all potentials is described by $\boldsymbol{\theta}$. For any factor $f \in \mathcal{F}$ the corresponding set of variables $\{y_v\colon v \in f\}$ will be denoted by y_f. *The energy minimization* problem then consists of finding *a labeling* $\boldsymbol{y}^* = \{y_v\colon v \in \mathcal{V}\} \in L_\mathcal{V}$ which minimizes the total sum of corresponding potentials:

$$\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}\in L_\mathcal{V}} E(\boldsymbol{y}) = \arg\min_{\boldsymbol{y}\in L_\mathcal{V}} \sum_{f\in\mathcal{F}} \theta_f(y_f)\,. \qquad (1)$$

Problem (1) is also known as *MAP-inference*. Labeling $\boldsymbol{y}^*$ satisfying (1) will be later called *a solution of the energy-minimization* or *MAP-inference problem*, shortly *MAP-labeling* or *MAP-solution*. Finally, a *model* is defined by the triple $(\mathcal{G}, L_\mathcal{V}, \boldsymbol{\theta})$, i.e. the underlying graph, the sets of labels and the potentials.

Diversity Method [4]. We will refer to this method as *DivMBest*. In order to find the M diverse, low energy, labellings $\boldsymbol{y}^1, \dots, \boldsymbol{y}^M$, the method proceeds by solving a sequence of problems of the form

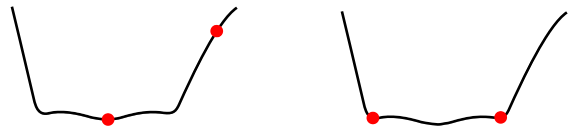
$$\boldsymbol{y}^m = \arg\min_{\boldsymbol{y}} \left[E(\boldsymbol{y}) - \lambda \sum_{i=1}^{m-1} \Delta(\boldsymbol{y}, \boldsymbol{y}^i) \right] \qquad (2)$$

for $m = 1, 2\dots, M$, where $\lambda > 0$ determines a trade-off between diversity and energy, $\boldsymbol{y}^1$ is the MAP-solution and the function $\Delta\colon L_\mathcal{V} \times L_\mathcal{V} \to \mathbb{R}$ defines the *diversity* of two labelings. In other words, $\Delta(\boldsymbol{y}, \boldsymbol{y}')$ takes a large value if $\boldsymbol{y}$ and $\boldsymbol{y}'$ are diverse, in a certain sense, and a small value otherwise. This problem can be seen as an energy minimization problem, where additionally to the initial potentials $\boldsymbol{\theta}$ the potentials $-\lambda\Delta(\cdot, \boldsymbol{y}^i)$, associated with an additional factor $\mathcal{V}$, are used. In the simplest and most commonly used form, $\Delta(\boldsymbol{y}, \boldsymbol{y}')$ is represented by a sum of node-wise diversities $\Delta_v\colon L_v \times L_v \to \mathbb{R}$,

$$\Delta(\boldsymbol{y}, \boldsymbol{y}') = \sum_{v\in\mathcal{V}} \Delta_v(y_v, y'_v)\,, \qquad (3)$$

and the potentials are split to a sum of *unary* potentials, i.e. those associated with additional factors $\{v \in \mathcal{V}\}$. This implies that in case efficient graph-cut based inference methods (including α-expansion [6], α-β-swap [6] or their generalizations [2]) are applicable to the initial problem (1) then they remain applicable to the augmented problem (2), which assures efficiency of the method.

Although the DivMBest method (2) shows impressive results in a number of computer vision applications, we argue that it suffers from its greedy nature. Each new labeling is obtained based on previously found solutions only, and is not influenced by upcoming labelings. As we show in this work, optimization for all M labelings *jointly* allows to improve the resulting solutions. A toy example illustrating our claim is presented in Fig. 1. Another scenario is sketched in Fig. 2. Note that with our approach we do not enforce that the MAP solution is part of the set of solutions. This is in contrast to the DivMBest [4] method. If this is a requirement then we can run a MAP solver and add the solution to our set.



(a) Sequentially inferred solutions **(b)** Jointly inferred solutions

Figure 2: Energy landscape with two different couples of solutions depicted by red points. (a) Corresponds to the DivMBest algorithm (2), which finds solutions sequentially. (b) Joint inference of diverse solutions may lead to lower total energy.

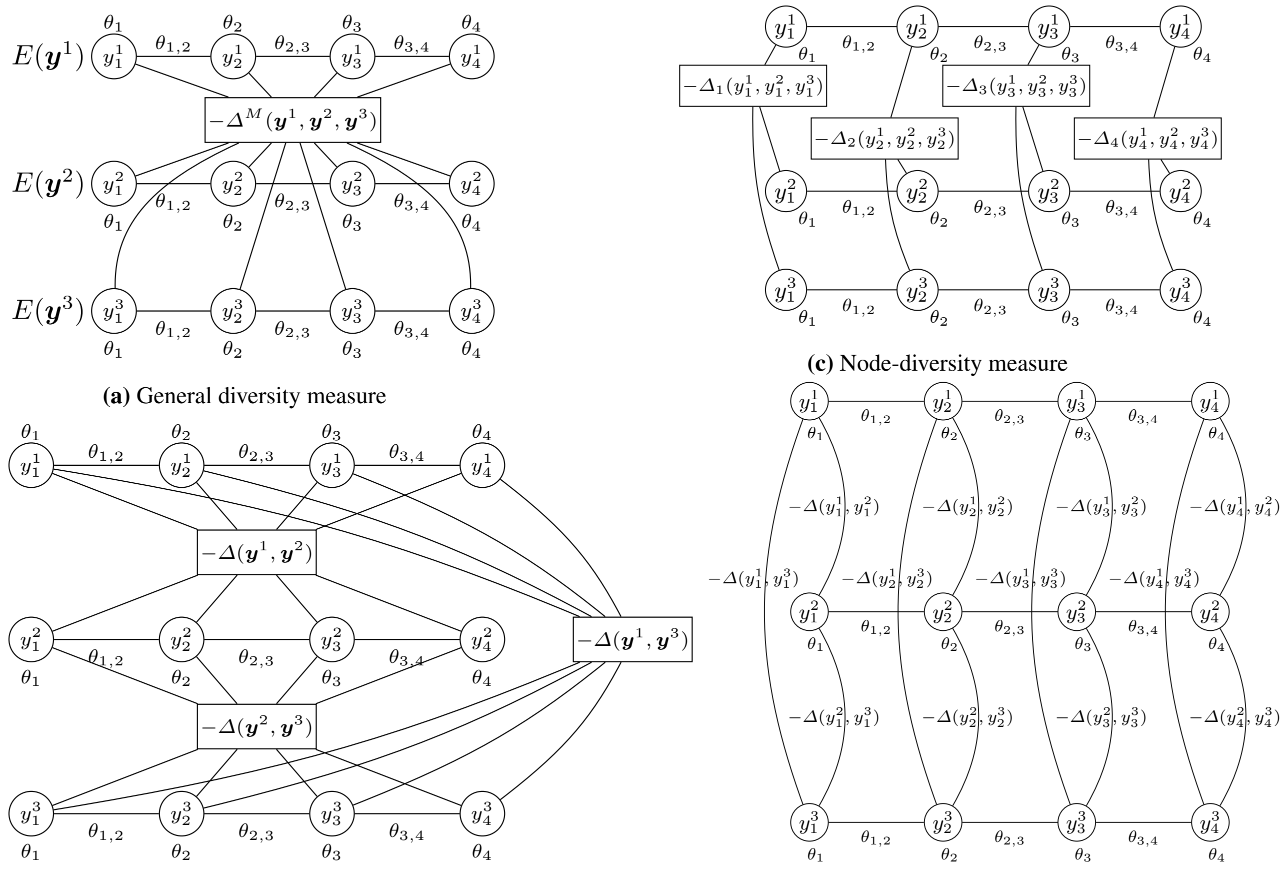
3. Diversity Model - Explicit Representation

In the following, we use brackets to distinguish between upper index and power, i.e. $(\mathcal{A})^n$ means the n-th power of $\mathcal{A}$, whereas n is an upper index in the expression $\mathcal{A}^n$. The notation $f^M(\{\boldsymbol{y}\})$ will be used as a shortcut for $f^M(\boldsymbol{y}^1, \dots, \boldsymbol{y}^M)$, for any function $f^M\colon (L_\mathcal{V})^M \to \mathbb{R}$. Instead of the greedy sequential procedure (2) we suggest to infer all M labelings *jointly*, by minimizing

$$E^M(\{\boldsymbol{y}\}) = \sum_{i=1}^{M} E(\boldsymbol{y}^i) - \lambda\Delta^M(\{\boldsymbol{y}\}) \qquad (4)$$

for $\boldsymbol{y}^1, \dots, \boldsymbol{y}^M$. Function Δ^M defines *the total* diversity of any M labelings. Though the expression (4) looks complicated we will show that it can be nicely represented in the form (1) and hence constitutes an energy minimization problem. To achieve this, let us first create M copies $(\mathcal{G}^i, \mathcal{L}^i_\mathcal{V}, \boldsymbol{\theta}^i) = (\mathcal{G}, \mathcal{L}_\mathcal{V}, \boldsymbol{\theta})$ of the initial model $(\mathcal{G}, \mathcal{L}_\mathcal{V}, \boldsymbol{\theta})$. We define the factor-graph $\mathcal{G}^M_1 = (\mathcal{V}^M_1, \mathcal{F}^M_1)$ for the new task as follows. The set of nodes in the new graph is the union of the node sets from the considered copies $\mathcal{V}^M_1 = \bigcup_{i=1}^M \mathcal{V}^i$. Factors are $\mathcal{F}^M_1 = \mathcal{V}^M_1 \cup \bigcup_{i=1}^M \mathcal{F}^i$, i.e. again the union of the initial ones extended by a special factor corresponding to the diversity penalty. Each node $v \in \mathcal{V}^i$ is associated with the label set $L^i_v = L_v$. The corresponding potentials $\boldsymbol{\theta}^M_1$ are defined as $\{-\lambda\Delta^M, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^M\}$, see Fig. 3a for illustration. The model $(\mathcal{G}^M_1, \mathcal{L}_{\mathcal{V}^M_1}, \boldsymbol{\theta}^M_1)$ corresponds to the energy (4). An optimal M-tuple of these labelings, corresponding to a minimum of (4), is a trade-off between low energy of individual labelings $\boldsymbol{y}^i$ and their total diversity.

Diversity measures. We now discuss three specific different diversity measures which are illustrated in Fig. 3. The *split-diversity* measure is written as the sum of pairwise diversities, i.e. those penalizing pairs of labelings



(a) General diversity measure

(b) Split-diversity measure

(c) Node-diversity measure

(d) Node-split-diversity measure

Figure 3: Examples of factor graphs for 3 diverse solutions of the original MRF (1) with different diversity measures. The circles represent the nodes of the original model that are copied 3 times. For clarity only the newly introduced factors of order higher than 2 are shown as squares. Pairwise factors are depicted by edges connecting the nodes. We omit λ for readability. (a) The most general diversity measure (4), (b) the split-diversity measure (5), (c) the node-diversity measure (6), (d) the node-split-diversity measure (7). Note that (b-d) are special cases of (a). Also, note that (d) is a special case of (c) and also of (b).

$$\Delta^M(\{\boldsymbol{y}\}) = \sum_{i=2}^{M} \sum_{j=1}^{i-1} \Delta(\boldsymbol{y}^i, \boldsymbol{y}^j)\,. \tag{5}$$

This means that $\mathcal{V}_1^M$ splits into $M(M-1)/2$ factors of the form $\mathcal{V}^i \cup \mathcal{V}^j$, $1 \le i < j \le M$, as shown in Fig. 3b.

We define the *node-diversity* measure as

$$\Delta^M(\{\boldsymbol{y}\}) = \sum_{v \in \mathcal{V}} \Delta_v(y_v^1, \dots, y_v^M) \tag{6}$$

where $\Delta_v \colon (L_v)^M \to \mathbb{R}$ are arbitrary *node-wise* diversity functions (see Fig. 3c).

Finally the special case of the split-diversity and node-diversity measures is the *node-split-diversity* measure

$$\Delta^M(\{\boldsymbol{y}\}) = \sum_{v \in \mathcal{V}} \sum_{i=2}^{M} \sum_{j=1}^{i-1} \Delta_v(y_v^i, y_v^j)\,, \tag{7}$$

which is a sum of pairwise factors, as illustrated in Fig. 3d. The special case of this diversity measure is the Hamming distance, i.e.

$$\Delta_v(y, y') = [\![y \neq y']\!]\,, \tag{8}$$

where expression $[\![A]\!]$ equals 1 if A is true and 0 otherwise.

In the recent work [25] the three alternative diversity measures of general form Fig. 3a were used in combination with DivMBest method (2):

- *Label Cost* diversity enforces the upcoming m-th labeling to contain labels that were not present in the already obtained $m-1$ labelings. In each iteration of the DivMBest algorithm (2) the α-expansion with label cost potentials was used (see [8]) for efficient inference.
- *Label Transitions* enforces m-th labeling to contain previously unseen pairs of labels of adjacent variables. Cooperative cuts [16] were used for inference in each iteration.

• *Hamming Ball* greedily enforces the volume of a union of Hamming balls around current m solutions to be as big as possible. The HOP-MAP [32] algorithm was applied for the MAP-inference in each iteration.

In the following we concentrate on the node-diversity measure. We show that when using the Hamming distance, which is its special case, we empirically outperform the methods introduced in [4] and [25].

DivMBest [4] as Greedy Minimization of the Split-diversity measure. Plugging (5) into (4) gives

$$E^M(\{\boldsymbol{y}\}) = \sum_{i=1}^{M} E(\boldsymbol{y}^i) - \lambda \sum_{i=2}^{M}\sum_{j=1}^{i-1} \Delta(\boldsymbol{y}^i, \boldsymbol{y}^j)\,. \tag{9}$$

Comparing function minimized in (2) and (9) one can see that the algorithm DivMBest (2) performs a greedy approximate minimization of (9). On the i-th iteration it optimizes over variables corresponding to $\boldsymbol{y}^i$, given fixed states of variables corresponding to $\boldsymbol{y}^1,\dots,\boldsymbol{y}^{i-1}$. Connections to variables corresponding to $\boldsymbol{y}^j$, $j > i$, are ignored and will be taken into account only later, on the j-th iteration.

Analyzing the optimization problem. Let us consider a specific form of our model (4) with the Hamming distance (8) as a diversity measure. The diversity constraint adds many pairwise potentials which are all of *repulsive* form (see also Fig. 3d), i.e. they penalize equal labels and do not penalize different ones. This makes efficient graph-cut based methods inapplicable and moreover, as shown in Section 5, the bounds delivered by LP-relaxation [34] based solvers are practically very bad as well. Indeed, solutions delivered by such solvers are significantly inferior even to the results of the greedy DivMBest method (2) (see Table 2). This motivates an alternative representation of the problem (4), which we discuss next.

4. Diversity Model - Clique Encoding

We now present an alternative representation of the model (4) with the node-diversity measure. This representation has fewer number of nodes but at the same time a larger label space. We will see that this representation is easier to optimize. With the node-diversity measure (6) the energy (4) can be rewritten as

$$E^M(\{\boldsymbol{y}\}) = \sum_{i=1}^{M} E(\boldsymbol{y}^i) - \lambda \sum_{v\in\mathcal{V}} \Delta_v(y_v^1,\dots,y_v^M)$$

$$= \sum_{i=1}^{M}\left[\sum_{\substack{f\in\mathcal{F}\\|f|=1}} \theta_f(y_f^i) + \sum_{\substack{f\in\mathcal{F}\\|f|>1}} \theta_f(y_f^i)\right] - \lambda \sum_{v\in\mathcal{V}} \Delta_v(y_v^1,\dots,y_v^M)\,.$$

Assume w.l.o.g. that $\{v\} \in \mathcal{F}$ for all $v \in \mathcal{V}$. Then we denote *unary* potentials θ_f for $|f| = 1$ as θ_v and regrouping terms, the above equation can be written as

$$= \sum_{v\in\mathcal{V}}\left[\sum_{i=1}^{M}\theta_v(y_v^i) - \lambda\Delta_v(y_v^1,\dots,y_v^M)\right] + \sum_{\substack{f\in\mathcal{F}\\|f|>1}}\sum_{i=1}^{M}\theta_f(y_f^i)\,.$$

Let us introduce the new variables $\boldsymbol{z}_v = (y_v^1,\dots,y_v^M)$, $v \in \mathcal{V}$ and the respective label sets $\hat{L}_v = (L_v)^M$. Informally, each label of a new variable $\boldsymbol{z}_v$ in a node v corresponds to an M-tuple of labels from the original task. In other words, we simply enumerate all possible label combinations in each node v, that are possible by M solutions. The new potentials $\hat{\theta}_v \colon \hat{L}_v \to \mathbb{R}$, $v \in \mathcal{V}$ and $\hat{\theta}_f \colon (L_f)^M \to \mathbb{R}$, $f \in \mathcal{F}\colon |f| > 1$ are defined as

$$\hat{\theta}_v(\boldsymbol{z}_v) = \sum_{i=1}^{M}\theta_v(y_v^i) - \lambda\Delta_v(y_v^1,\dots,y_v^M)\,, \tag{10}$$

$$\hat{\theta}_f(\boldsymbol{z}_f) = \sum_{i=1}^{M}\theta_f(y_f^i)\,. \tag{11}$$

In this notation the energy is given as

$$E^M(\{\boldsymbol{y}\}) = \sum_{v\in\mathcal{V}}\hat{\theta}_v(\boldsymbol{z}_v) + \sum_{\substack{f\in\mathcal{F}\\|f|>1}}\hat{\theta}_f(\boldsymbol{z}_f)\,. \tag{12}$$

Special Case: Pairwise Model For second order models (i.e. the cardinality of factors is two at most) equation (12) is written as

$$E^M(\{\boldsymbol{y}\}) = \sum_{v\in\mathcal{V}}\hat{\theta}_v(\boldsymbol{z}_v) + \sum_{uv\in\mathcal{F}}\hat{\theta}_{uv}(\boldsymbol{z}_u,\boldsymbol{z}_v)\,. \tag{13}$$

The following Theorem 1 basically states that in case the original MAP-inference problem is (approximately) solvable with α-β-swap [6] (α-expansion [6]) then minimization of $E^M(\{\boldsymbol{y}\})$ in (13) can be performed with α-β swap (α-expansion) as well.

Definition 1. *For any set L the function $f\colon L\times L\to\mathbb{R}$ is called* a semi-metric *if for all $x, x' \in L$ there holds: (i) $f(x,x') \ge 0$; (ii) $f(x,x') = 0$ iff $x = x'$; (iii) $f(x,x') = f(x',x)$.*

Definition 2. *Function $f\colon L\times L\to\mathbb{R}$ is called* a metric *if it is a semi-metric and additionally there holds: $f(x,x') + f(x',x'') \ge f(x,x'')$, $\forall x,x',x'' \in L$.*

Theorem 1. *Let $L_v = L_u$, $uv \in \mathcal{F}$ and functions θ_{uv} be semi-metrics (metrics). Then functions $\hat{\theta}_{uv}(\boldsymbol{z}_u,\boldsymbol{z}_v)$ defined as in (11) are semi-metrics (metrics) as well.*

We refer to the supplementary material for the proof.

For instance, in the special case of *Potts model* $\theta_{uv}(y,y') = [\![y \ne y']\!]$ the pairwise factors defined by (11) constitute the Hamming distance between vectors $\boldsymbol{z}_v$ representing the new labels:

$$\hat{\theta}_{uv}(\boldsymbol{z}_u,\boldsymbol{z}_v) := \sum_{i=1}^{M}\theta_{uv}(y_u^i,y_v^i) = \sum_{i=1}^{M}[\![y_u^i \ne y_v^i]\!]\,. \tag{14}$$

Both Potts potentials and Hamming distance are metrics, which defines a special case of Theorem 1.

K-truncated Clique Encoding The disadvantage of the clique encoding representation (12) is an exponential growth of cardinality of the label set $\hat{L}_v = (L_v)^M$, which implies inefficiency for inference with large L_v and especially a large M. For these cases we propose an efficient approximative Algorithm 1 combining clique encoding (12) and greedy minimization for the energy (4). Though it can be used with the node-diversity measures (3) we describe it for the special case of the node-split-diversities (7), as it is used in our experiments.

Algorithm 1 K-truncated Clique Encoding

Require: $(\mathcal{G},\ L_{\mathcal{V}}, \boldsymbol{\theta})$ – original model,
$\lambda \in \mathbb{R}$ – diversity parameter,
$M \in \mathbb{N}$ – total number of diverse labelings,
$K < M$ – num. of processed labelings in each step.

1: **for** $i = 0, \dots, \lfloor \frac{M}{K} \rfloor$ **do**
2: $\quad s = iK + 1; \quad t = \min\{M, (i+1)K\}$
3: $\quad \{\boldsymbol{y}^s, \dots, \boldsymbol{y}^t\} = \arg \min\limits_{\{\boldsymbol{x}^s,\dots,\boldsymbol{x}^t\}} \Big[E^K(\boldsymbol{x}^s, \dots, \boldsymbol{x}^t) - \lambda \sum\limits_{v\in\mathcal{V}} \sum\limits_{l=s}^{t} \sum\limits_{m=1}^{s-1} \Delta_v(x_v^l, y_v^m) \Big]$
4: **end for**
5: **return** $\{\boldsymbol{y}^1, \dots, \boldsymbol{y}^M\}$

In each iteration Algorithm 1 performs optimization with respect to at most K labelings $\{\boldsymbol{y}^s, \dots, \boldsymbol{y}^t\}$, $t-s+1 = K$, (less than K in the last iteration, if M is not dividable by K) given already computed labelings $\{\boldsymbol{y}^1, \dots, \boldsymbol{y}^{s-1}\}$. Diversity of $\{\boldsymbol{y}^s, \dots, \boldsymbol{y}^t\}$ with respect to $\{\boldsymbol{y}^1, \dots, \boldsymbol{y}^{s-1}\}$ is provided by taking into account the sum of corresponding diversity terms $\lambda \sum\limits_{v\in\mathcal{V}} \sum\limits_{l=s}^{t} \sum\limits_{m=1}^{s-1} \Delta_v(x_v^l, y_v^m)$ playing the role of addition to unary potentials in line 3 of Algorithm 1. Minimization (possibly approximate) in line 3 is done with the clique encoding approach (12).

Overall, algorithm performs a greedy optimization similar to DivMBest (2) with the difference that in each iteration K labelings are inferred *jointly* instead of a single one. The method coincides with DivMBest (2) for $K = 1$ and with clique encoding for $K = M$.

As we show in Section 5, Algorithm 1 significantly outperforms DivMBest (2) already for $K = 2$. Larger values of K lead to further improvements.

5. Experimental evaluation

We show benefits of our approach in **two applied scenarios**: (a) interactive foreground/background segmentation for images with provided scribbles annotations [4] and (b) Category level segmentation on PASCAL VOC 2012 data [10].

Notation	Appr.	MAP Inference	Div. measure
DivMBest	(2)	α-expansion [6]	HD
ADSal	(4)	ADSal [28]	HD
CE	(13)	α-expansion	HD
CE-TRWS	(13)	TRW-S[18]	HD
$\mathtt{CE}_K$	Alg. 1	α-expansion	HD
LC*	(2)	α-expansion [8]	LC
LT*	(2)	Coop. cuts [16]	LT
HB*	(2)	HOP-MAP [32]	HB

Table 1: Diversity methods used in our experiments. Column *Appr.* corresponds to the selected approach: either it is a greedy optimization of DivMBest (2) or direct optimization of the energy (4), its clique encoding representation (13) or the mixed K-truncated clique encoding Algorithm 1. (*)- methods were not run by us and the results were taken from [25] directly.

Diversity measures used in experiments are: the Hamming distance (8) HD, Label Cost LC, Label Transitions LT and Hamming Ball HB. The last three measures were introduced in [25] and are briefly described in Section 3. Following [25] we use $D_1 \otimes D_2 \cdots \otimes D_n$ to denote that the diversity measures $D_1, D_2, \dots D_n$ were sequentially applied to obtain the next $\frac{M}{n}$ solutions within DivMBest algorithm (2), e.g. HD⊗LC for $M = 4$ means the first 2 labelings were found with HD diversity measure and the following two – with LC. Notation ⊕ means that diversity measures were linearly combined. We refer to [25] for a detailed description.

MAP-Inference Algorithms In our experiments we used α-expansion [6], which turns into the max-flow algorithm in case of two labels. To estimate accuracy of inference we used TRW-S [18] and ADSal [28], which provide lower bounds. We used ADSal because contrary to TRW-S it guarantees convergence to a solution of the LP-relaxation [34] of the energy minimization problem and moreover provides accuracy of the found LP solution. We used implementations of TRW-S and ADSal provided with OpenGM2 [1] library.

We summarized notation of the **compared diversity methods** in Table 1.

5.1. Interactive segmentation

Interactive image segmentation is a possible application scenario for diversity techniques. Instead of returning a single segmentation corresponding to a MAP-solution, diversity methods return a small number of possible low-energy results. Following [4] we model only the first iteration of such an interactive procedure, i.e. we consider user scribbles to be given and compare the sets of segmentations returned by the compared diversity methods.

Authors of [4] kindly provided us their 50 graphical

	ADSal	DivMBest	CE-TRWS	CE
$M=2,\ \lambda=0.45$	0.009	0.005	0.0	0.0
$M=2,\ \lambda=0.5$	0.013	0.008	0.0	0.0
$M=6,\ \lambda=0.15$	0.074	0.002	0.0	0.0
$M=6,\ \lambda=0.25$	0.301	0.034	0.0	0.0

Table 2: Interactive segmentation: comparison of attained relative precisions $(E^M(\{\boldsymbol{y}\}) - D)/D$, where D is a dual bound obtained by CE-TRWS. The first two methods are applied to (4), the second two used representation (13). The LP relaxation was solved by ADSal with the relative accuracy of 0.001.

model instances, corresponding to the MAP-inference problem (1). They are based on a subset of the PASCAL VOC 2010 [9] segmentation challenge with manually added scribbles. Pairwise potentials constitute contrast sensitive Potts terms [5], which implies that the MAP-inference is submodular and therefore solvable by min-cut/max-flow algorithms [19].

Energy comparison. Table 2 provides comparison of different inference methods for the diversity model (4) and its clique encoding representation (13), for the interactive segmentation dataset (λ is the same in all cases). It can be seen that LP-relaxation of the explicit formulation (4) is far from being LP-tight and moreover returns even worse results than the greedy DivMBest method. As mentioned in Section 3 we believe that the reason is the repulsive diversity potentials. However the same problem in its clique encoding representation *empirically turns to be LP-tight* though the problem (13) is not (permuted [29]) submodular. Moreover, α-expansion found its optimal solutions in *all considered cases*. We believe that inference results improved due to moving the repulsive potentials to unary costs.

Quantitative and Qualitative Comparison. Table 3 and Fig. 5 show comparison of several techniques for this dataset. As a quality measure we used per pixel accuracy of the best solution for each sample averaged over all test images. Parameter λ has been chosen for each method separately via cross-validation. In Fig. 4 we show accuracy of the CE method against DivMBest for a range of different values of λ and number M of diverse solutions. In all these experiments, our CE method shows significantly better accuracy than its competitors.

Running time of our CE method is, as expected, higher than those for DivMBest, however it still can be considered as practically useful: for $M = 2$ the average DivMBest time is 0.45 ms whereas CE runtime is 2.9 ms, for $M = 6$ times are 2.4 ms and 47.6 ms per image respectively.

Energy of Labelings for a Given Diversity Level. We also compared the total energy for $M = 6$ labelings for

	M=1	M=2	M=6
DivMBest*[4]	91.57	93.16	95.02
HB[25]*	91.57	93.95	94.86
DivMBest*⊗HB*[25]	-	-	95.16
DivMBest*⊕HB*[25]	-	-	95.14
CE	91.57	**95.13**	**96.01**

Table 3: Interactive segmentation: averaged pixel accuracies.

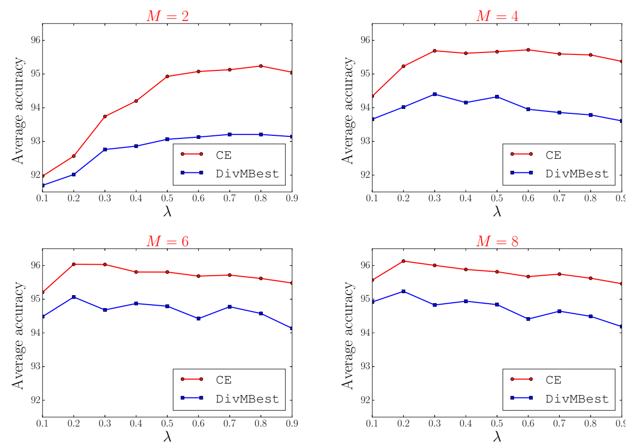


Figure 4: Pixelwise accuracy comparison for interactive segmentation for different values of λ and M.

DivMBest and CE methods. Parameter λ for DivMBest was fit for each sample to get equal or less diversity than those provided by CE algorithm. We managed to attain almost equal diversities for 44 out of 50 instances and in all these cases the total energy $\sum_{i=1}^{m} E(\boldsymbol{y}^i)$ of obtained labelings $\boldsymbol{y}^i$ was smaller for CE compared to DivMBest. This shows practical superiority of our approach.

5.2. Category level segmentation

The category level segmentation from PASCAL VOC 2012 challenge [10] contains 1449 validation images with known ground truth, which we used for evaluation of diversity methods. Corresponding pairwise models with contrast sensitive Potts terms were used in [25] and kindly provided us by authors. Contrary to interactive segmentation label sets contain 21 elements and hence the respective MAP-inference problem (1) is not submodular anymore. However it still can be approximatively solved by α-expansion.

Because of a significant number of labels we were unable to use CE approach for $M > 5$ and resorted to CE_2 and CE_3. Results of the quantitative evaluation are presented in Table 4, where each method was used with parameter λ optimally tuned via cross-validation on validation set in PASCAL VOC 2012. Exemplary comparison of CE and

Figure 5: Comparison for samples from interactive segmentation dataset. Number above each solution is a corresponding per pixel accuracy.

	M=1	M=5	M=15	M16
`DivMBest`*[4]	43.43	51.21	52.90	-
`HB`*[25]	-	51.71	55.32	-
`LC`*[25]	-	46.28	50.39	-
`LT`*[25]	-	45.92	46.89	-
`DivMBest`*⊕`HB`*[25]	-	-	55.89	-
`HB`*⊗`LC`*⊗`LT`*[25]	-	-	56.97	-
`DivMBest`*⊗`HB`*⊗`LC`*⊗`LT`*[25]	-	-	-	57.39
`CE`	-	**54.22**	-	-
$\texttt{CE}_2$	-	**53.08**	**57.46**	**57.76**
$\texttt{CE}_3$	-	**54.14**	**57.76**	**58.36**

Table 4: PASCAL VOC 2012. Intersection over union quality measure. The best segmentation out of M is considered. Notation '-' correspond to absence of result due to computational reasons or inapplicability of the method.

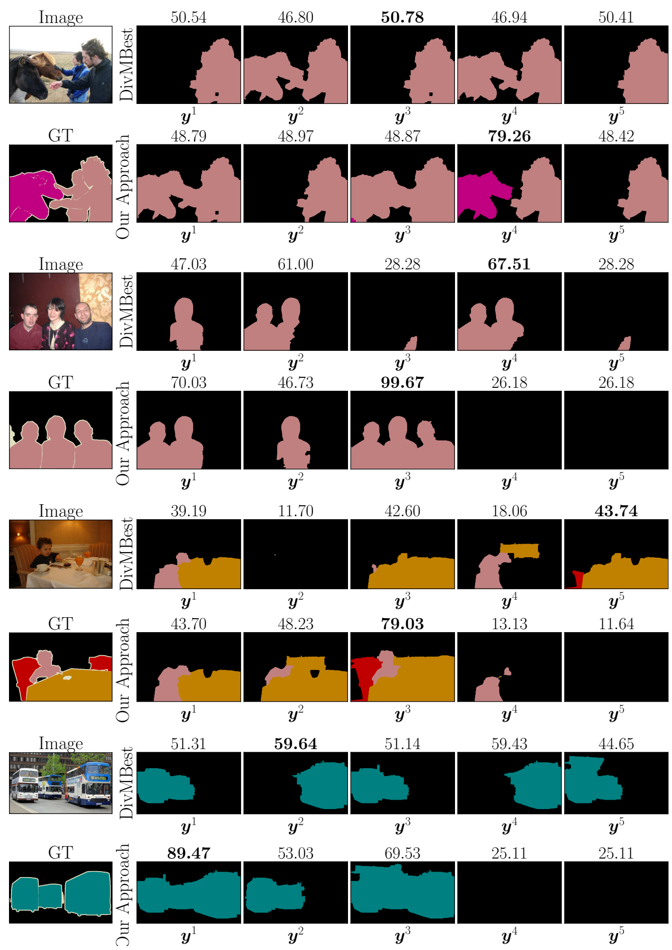


Figure 6: Comparison for samples from Pascal VOC 2012 dataset. Number above each solution is a corresponding intersection over union quality measure.

`DivMBest` is shown in Fig. 6. It turns out that even the suboptimal optimization method $\texttt{CE}_2$ outperforms *all* competitors, except $\texttt{CE}_3$ and `CE`, which show even better segmentation accuracy.

Average running times per image for $M = 5$ for `DivMBest`, $\texttt{CE}_2$, $\texttt{CE}_3$ and `CE` methods are 0.01, 0.14, 2.28 and 733 seconds, respectively. For $M = 15$ times are 0.03, 0.39 and 5.87 seconds for `DivMBest`, $\texttt{CE}_2$, and $\texttt{CE}_3$, respectively. We observe approximately linear growth of running time wrt number of nodes in the original problem for for `DivMBest`, $\texttt{CE}_2$, $\texttt{CE}_3$ and `CE`.

6. Conclusions and Outlook

We proposed a novel non-greedy approach for the problem of finding M diverse low energy labelings. This is done by solving an energy minimization in a specially constructed graphical model. We show that inference in this model can be addressed by graph-cut based methods like α-expansion if the MAP-inference in the original model was solvable by these methods. Our experiments suggest that even with a Hamming distance as diversity measure our method qualitatively and quantitatively outperforms competing diversity techniques using more involved measures.

In future we plan to improve the computational efficiency of our method.

References

[1] B. Andres, J. H. Kappes, U. Köthe, C. Schnörr, and F. A. Hamprecht. An empirical comparison of inference algorithms for graphical models with higher order factors using OpenGM. pages 353–362. 2010. 6

[2] C. Arora, S. Banerjee, P. Kalra, and S. Maheshwari. An efficient graph cut algorithm for computer vision problems. In *ECCV*. Springer, 2010. 3

[3] D. Batra. An efficient message-passing algorithm for the M-best MAP problem. *ArXiv preprint arXiv:1210.4841*, 2012. 1

[4] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse M-best solutions in markov random fields. 2012. 1, 2, 3, 5, 6, 7, 8

[5] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *ICCV*, 2001. 7

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *TPAMI*, 23(11):1222–1239, 2001. 3, 5, 6

[7] C. Chen, V. Kolmogorov, Y. Zhu, D. N. Metaxas, and C. H. Lampert. Computing the M most probable modes of a graphical model. In *AISTATS*, 2013. 2

[8] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. *IJCV*, 96(1):1–27, 2012. 4, 6

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. 7

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. 6, 7

[11] M. Fromer and A. Globerson. An lp view of the m-best map problem. 2009. 1

[12] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *TPAMI*, (6):721–741, 1984. 2

[13] A. Guzman-Rivera, D. Batra, and P. Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *NIPS 25*. 2012. 1, 2

[14] A. Guzman-Rivera, P. Kohli, and D. Batra. DivMCuts: Faster training of structural SVMs with diverse M-best cutting-planes. In *AISTATS*, 2013. 1, 2

[15] A. Guzman-Rivera, P. Kohli, D. Batra, and R. A. Rutenbar. Efficiently enforcing diversity in multi-output structured prediction. In *AISTATS*, 2014. 1, 2

[16] S. Jegelka and J. Bilmes. Submodularity beyond submodular energies: coupling edges in graph cuts. In *CVPR*, 2011. 4, 6

[17] J. H. Kappes, B. Andres, F. A. Hamprecht, C. Schnörr, S. Nowozin, D. Batra, S. Kim, B. X. Kausler, T. Kröger, J. Lellmann, N. Komodakis, B. Savchynskyy, and C. Rother. A comparative study of modern inference techniques for structured discrete energy minimization problems. *IJCV*, pages 1–30, 2015. 1

[18] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *TPAMI*, 28(10):1568–1583, 2006. 6

[19] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *TPAMI*, 26(2):147–159, 2004. 7

[20] A. Kulesza and B. Taskar. Structured determinantal point processes. 2010. 2

[21] 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(7), 1972. 1

[22] D. Nilsson. An efficient algorithm for finding the m most probable configurationsin probabilistic expert systems. *Statistics and Computing*, 8(2):159–173, 1998. 1

[23] G. Papandreou and A. Yuille. Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models. In *ICCV*, 2011. 2

[24] J. Porway and S.-C. Zhu. C^ 4: Exploring multiple solutions in graphical models by cluster sampling. *TPAMI*, 33(9):1713–1727, 2011. 1

[25] A. Prasad, S. Jegelka, and D. Batra. Submodular meets structured: Finding diverse subsets in exponentially-large structured item sets. In *NIPS 27*, 2014. 2, 4, 5, 6, 7, 8

[26] V. Premachandran, D. Tarlow, and D. Batra. Empirical minimum bayes risk prediction: How to extract an extra few % performance from vision models with just three more parameters. In *CVPR*, 2014. 1, 2

[27] V. Ramakrishna and D. Batra. Mode-marginals: Expressing uncertainty via diverse M-best solutions. 2012. 1, 2

[28] B. Savchynskyy, S. Schmidt, J. Kappes, and C. Schnörr. Efficient MRF energy minimization via adaptive diminishing smoothing. *arXiv preprint arXiv:1210.4906*, 2012. 6

[29] D. Schlesinger. Exact solution of permuted submodular minsum problems. In *EMMCVPR*, 2007. 7

[30] M. I. Schlesinger and V. Hlavac. *Ten lectures on statistical and structural pattern recognition*, volume 24. Springer Science & Business Media, 2002. 1

[31] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. F. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *TPAMI*, 30(6):1068–1080, 2008. 1

[32] D. Tarlow, I. E. Givoni, and R. S. Zemel. Hop-map: Efficient message passing with high order potentials. In *AISTATS*, 2010. 5, 6

[33] Z. Tu and S.-C. Zhu. Image segmentation by data-driven markov chain monte carlo. *TPAMI*, 24(5):657–673, 2002. 1

[34] T. Werner. A linear programming approach to max-sum problem: A review. *TPAMI*, 29(7):1165–1179, 2007. 5, 6

[35] P. Yadollahpour, D. Batra, and G. Shakhnarovich. Discriminative re-ranking of diverse segmentations. 2013. 1, 2

[36] C. Yanover and Y. Weiss. Finding the M most probable configurations using loopy belief propagation. 2004. 1