Scribble Supervised Annotation Algorithms of Panoptic Segmentation for Autonomous Driving

Ruobing Shen, Thomas Guthier
Technical Center Europe
Hyundai Mobis
65936 Frankfurt, DE
ruobing.shen@gmobis.com

Bo Tang *
Department of Mathematics
Noetheastern University
02115 Boston, USA
tang.bo@gmobis.com

Ismail Ben Ayed
Ecole de Technologie Superieure
H3C 1K3 Montreal, CA
ismail.benayed@etsmtl.ca

Abstract

Large-scale ground truth dataset is of crucial importance for deep learning based segmentation models, but annotating per-pixel masks is extremely time consuming. In this paper, we investigate semi-annotated graph based segmentation algorithms that enforce connectivity. To be more precise, we introduce a class-agnostic heuristic of a discrete Potts model, and a class-aware Integer Linear Programming (ILP) that ensures global optimum. Both algorithms are able to generate panoptic segmentation supervised by scribbles, and can take RGB, or utilize the feature maps from any DCNN, whether trained on the target dataset or not, as input. We present competitive semantic segmentation results on the PASCAL VOC dataset, as well as report panoptic segmentation result on the more challenging Cityscapes dataset. Our algorithms show superior results that makes them suitable for weakly supervised segmentation on new dataset, or interactive semi-automated ground truth generation by human annotators on existing dataset.

1 Introduction

Deep Convolutional Neural Networks (DCNNs) excel at a wide range of image recognition tasks [16, 33, 39], such as semantic segmentation [10, 39, 51] and panoptic segmentation [20, 48, 47, 19]. Semantic segmentation studies the tasks of assigning a class label to each pixel of an image, where instance segmentation [11] detects and segment each object instance. Panoptic segmentation unifies both tasks that investigate to segment both *things* (such as person, cars) and *stuff* (such as road, sky) classes, which is more relevant in the application towards autonomous driving and parking [38].

While DCNNs show outstanding results for semantic and panoptic segmentation, they have two conceptional problems. First, they require huge amounts of annotated data. Annotating segmentation masks is a very time consuming and labor extensive task. For example, annotating a semantic image mask took "more than 1.5h on average" on the Cityscapes dataset [12]. For autonomous parking, there are no public surround-view fish-eye cameras [4] dataset available. In addition, DCNNs rely on their implicitly learned generalization probability and most of the state of the art architectures do not make use of any domain specific knowledge, such as neighborhood relations and connectivity priors

*Bo Tang is the second author and this work was done when he was a research intern at Hyundai Mobis.

Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

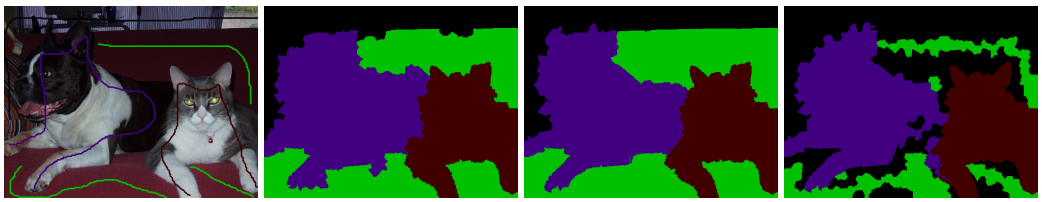

Figure 1: Left: image with scribbles from Pascal VOC dataset. Mid-left: semantic segmentation result of our heuristic using layer 3 of ResNet 101 as input information. Mid-right: result of our ILP using probability map of DeepLab V2 as input. Right: ILP without connectivity prior (MRF).

for segments. In contrary, classical graph based segmentation models [6, 7, 49] do not require any learning data and can incorporate specific domain knowledge. Their major drawback is they relay on human designed similarity features and require complex optimization algorithms or solvers, which are mainly CPU based and non-suitable for real time applications.

In this work, we explore the combination of DCNNs and graph based segmentation algorithms for the task of annotation ground truth panoptic segmentation. Specifically, we design a heuristic region growing method based on the Potts model [31], as well as an optimization solver based integer linear programming (ILP) model. We utilize scribble based annotations from human annotators as an initialized hard constraints for our optimization algorithm, which is typical in a human-in-the-loop (HITL) annotation process. We explore two different scenarios. In the first scenario, we assume that for the segmentation task, there is already a pre-trained neural network with the same class mapping. Here, we use the probability map of the DCNNs, as well as scribbles, as input to our algorithms and show significantly improved results for semantic segmentation on the Pascal and Cityscape datasets [12]. In the second scenario, we assume that no pre-trained DCNN for the same objective is available. This is true for a lot of existing datasets, e.g., Cityscapes does not contain any class labels for lane marking, which is crucial information for an autonomous vehicle. In this case, we cannot use the class specific probability map, but the more generic low level features of a DCNNs can be utilized as feature description for the optimization algorithm.

To highlight the conceptual limitation of DCNNs, we propose two interactive graph based segmentation algorithms, that both enforces *connectivity* of pixels that belong to the same region (to be more precise in Sec. 2.1) of a class. We first present a class-agnostic heuristic algorithm that efficiently approximates the Potts model. We then study and propose a novel ILP of the markov random field (MRF) by introducing pseudo edges in case of multi-instances (or regions) of the same class, that greatly reduces the complexity. Furthermore, we add probability maps and lower level feature maps of DCNNs as unary and pairwise priors for the heuristic and ILP, which improves the accuracy by a large margin. The connectivity prior [46] improves the quality of the segmentation and gives the annotator more control over per scribble instance (or regions) segmentation. See Fig. 1 for a visualization of our result for just one loop of drawing scribbles.

We prototype our algorithms for the task of ground truth segmentation, and present competitive semantic segmentation on the PASCAL VOC dataset, as well as report panoptic segmentation on the more challenging Cityscapes dataset, both supervised by one single round of scribbles. While there is public scribble dataset [25] for Pascal but none for Cityscapes, we create a artificial scribble dataset [2] for Cityscapes based on the ground truth. To investigate the general purpose of low level DCNN feature, we compare DCNN trained on ImageNet with and without fine tuning on the target dataset. Our experiments show that incorporating the connectivity prior as well as the neural network feature maps greatly improve the algorithm performance.

Summarized our key contributions are

- an in depth analysis of a combination of neural network and graph based scribble supervised segmentation algorithms with connectivity prior,
- improved heuristic algorithm and novel formulation of the ILP via pseudo edges which superior performance,

[2]We will make the code that generates the Cityscapes artificial scribble dataset open-source.

- extensive evaluation of the scribble based weakly supervised algorithm for semantic and panoptic segmentation on two challenging datasets.

Our proposed algorithms have multiple use cases in annotating datasets for panoptic segmentation. Firstly, when no training data is available, one can use our algorithm to generate a good quality panoptic segmentation baseline. Secondly, if training data is available, one can use any trained DCNN and its probability as the input feature to our algorithm, then improve the inference result of the DCNN. These can be used inside any HITL annotation tool, as the annotator can interact with the image in forms of scribbles until satisfaction. Finally, it can be used inside any weakly or semi-supervised learning framework for panoptic segmentation. To the best of our knowledge, our methods are the first "non-DCNN" panoptic segmentation algorithms on Cityscapes with competitive results, which shows the potential improvement gained by a combined approach.

Related Work. The procedure of annotating per pixel segmentation masks is similar to interactive image segmentation, which is widely studied in the past decade. The method using bounding boxes [34] is suitable for instance segmentation, which requires the user to draw the box as tight as possible. Similarly, 4 extreme points [28, 30] can be used. In both cases only the thing classes can be annotated. On the other hand, polygon based methods, such as LabelMe [36], require users to carefully click the extreme points of things and stuff, and the accuracy heavily depends on the number of clicks. On the contrast, scribbles are recognized as a more user-friendly way among various forms of user interactions [8, 23]. Moreover, it is also easier to annotate stuff classes using scribbles.

Modern segmentation annotation tools often adopts deep learning based methods, including Polygon-RNN++ [2] and Curve-GCN [27]. Besides, they also take advantage of deep learning based ensemble learning [52], to combine several inference results to produce better segmentation. However, these methods requires existing ground truth dataset for DCNNs to learn on the first hand, which may not be available when unknown domains or new classes are introduced.

As a cheaper alternative, weakly supervised segmentation has drawn a lot of attention recently. [13, 18, 24] claim that weekly iteratively trained by just bounding boxes and image tags, the DCNN can get 95% score compared to fully supervised on Pascal VOC. Since this method emphasizes on thing classes, it has worse score (or even none) on stuff classes. Instead, [25, 9] claim that iteratively training a DCNN by scribble annotations alone suffers only a small degradation in performance on both thing and stuff classes.

For graph based methods, the (discrete) Potts model [31] is widely used for denoising and segmentation. The author of [43] formulate the problem as an ILP and tries to solve the global optimal solution, but only to a reduced image size, while [29] proposed an efficient region fusion based heuristic algorithm. The author of [40] extend their work by incorporating scribbles and enforcing connectivity of pixels with the same scribble at every iteration (which we call $\ell_0 H_g$).

The MAP-MRF (maximizing a posterior in an markov random field) has been well studied for image segmentation. Previous methods focus on local priors[6, 21], and efficient approximate algorithms exists, e.g., graph cut [7] and belief propagation [49]. Recently, [32, 41] has looked into the global connectivity prior of MRF, and formulate the problem as an ILP, which uses the cutting plane method and are solved by an ILP solver. Solving an ILP is in general $\mathcal{NP}$-hard, and the branch and bound [22] is a fundamental method to solve an ILP inside any modern ILP solver [3].

Our proposed algorithms are scribble based with global connectivity prior, which enforces pixels of the same region to be connected, and allows the annotator to better control the final segmentation.

2 Proposed approach

2.1 Prerequisite

Given an image, we build an undirected graph $G = (V, E)$ where V represents a set of pixels (or superpixels) and E a set of edges consisting of un-ordered pairs of nodes. Image segmentation can thus be transformed into a graph labeling problem, where the label set C is pre-defined.

When talking about segmentation, we need to first distinguish between class, instance, and region (associated with a scribble) ID of a node. In semantic segmentation, the task is to assign a class label to each node in a graph. In panoptic segmentation, one has to further assign an instance ID to the

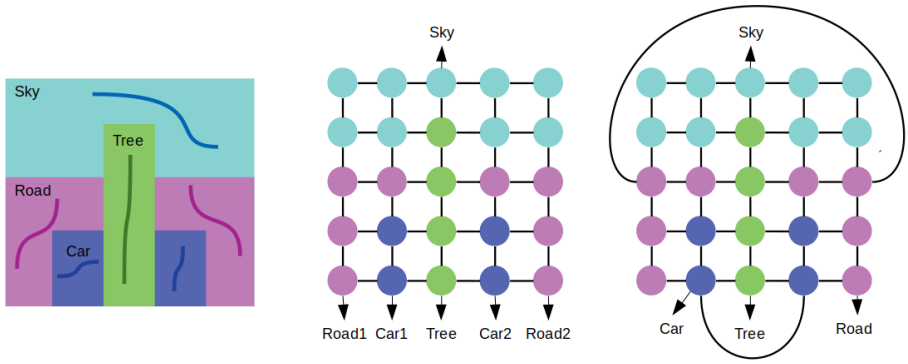

Figure 2: Left: our scribble policy to draw as many scribbles as there are separated regions. Middle: ILP-D that introduces dummy class variables on each region. Right: ILP-P that introduces pseudo edges which connect multiple regions of the same class, which does not increase number of variable.

node that belong to the "thing" class. In this paper, our algorithms require an additional region ID, which is linked to a scribble and we assume nodes within the same region must be connected (to be explained in Sec. 2.4). This is to deal with the case where an object of the 'thing" or "stuff" class is separated into several connected regions, e.g., the car in Fig. 2 is separated into two regions by a tree. Afterwards, the class and region labels can be combined to generate a panoptic segmentation.

2.2 Overall workflow

We first describe the pre-processing steps before running the two proposed segmentation algorithms from Sec. 2.3 through 2.5, i.e., applying superpixel algorithm as a dimension reduction step, the scribble policy needed to follow in order to produce panoptic segmentation, and extracting different levels of image features to pass to the algorithm. Thereafter, in Sec. 2.6, we discuss the formal definition of connectivity, and the two proposed algorithms in details, i.e., the class-agnostic heuristic and the class-aware ILP with connectivity constraints.

2.3 Superpixels as dimension reduction

Superpixels have long been used for image segmentation [44, 14, 42, 1], as they can greatly reduce the problem size while not sacrificing much of the accuracy. In this paper, we adopt SEEDS [14] to generate superpixels on the PASCAL VOC dataset [15], while using a deep learning based method [45] on the more challenging Cityscapes dataset [12]. We then build a region adjacency graph (RAG) $G(V, E)$ of the superpixels, where each superpixel forms a node (vertex) and edges connect two adjacent superpixels. The RAG is then processed by our graph based algorithms.

2.4 Extracting features using scribbles

Our segmentation algorithms are scribble supervised, which are two folded. On the one hand, the node labels, such as class, instance and region IDs are fixed if the nodes are covered by the corresponding scribbles. On the other hand, for the ILP algorithm, if no high level image information exists, the superpixels covered by scribbles will be used to extract information for the class, i.e., one can use the average color of the scribbled superpixels to represent the corresponding class [41].

Scribbles generation policy. First of all, the scribble itself must be connected. Second, one has to draw as many scribbles as there are connected regions (both "thing" and "stuff" class) presented in the image. For instance, if an object is cut into separated regions, one has to draw a scribble on each region. One sample image with scribbles is shown in Fig. 2.

2.5 Extracting features from images by DCNN

Although the input to the algorithms can be as simple as image RGB information, one can also take advantage of the modern DCNN to extract deeper features. We distinguish two scenarios:

- No previous training data is available – one starts annotating images in a new dataset.
- Training data available – one continues annotating more images of an existing dataset.

In the former case, other than RGB, one can also adopt any base network (i.e., ResNet 50 [16]) pre-trained on other datasets (i.e., ImageNet) and use the output of the low level features that extracts image edges, textures, etc. In the later case, one can fully utilize any modern DCNN trained on the existing dataset, and use the output of the final layer (i.e., probability map). We conduct detailed experiments on adopting different levels of feature maps as input to our algorithms in Sec. 3.

2.6 Our proposed optimization algorithms

Given a region adjacency graph $G(V, E)$, we use graphical models to propagate information from labeled (scribbles) to unlabeled nodes. We present a local greedy class-agnostic clustering algorithm of the discrete Potts model, and a class-aware ILP formulation of MRF with connectivity prior.

2.6.1 The connectivity prior

Two nodes u, v in a graph G are *connected* if there is a (u, v)-path in G. G is called *connected* if every pair of nodes are connected in G, otherwise it is *disconnected*. Let $\bar{G}_\ell \subseteq G$ be a connected subgraph where every node is labeled $\ell \in C$. Then, the image segmentation with connectivity constrains corresponds to find a partition of G into k ($k = |C|$) connected (and disjoint) subgraphs $\{\bar{G}_1, \bar{G}_2, \ldots, \bar{G}_k\}$. Enforcing connectivity constrains itself is proven to be $\mathcal{NP}$-hard even in the rooted case [41], where at least one node of each subgraph is fixed (fulfilled by our scribble policy).

2.6.2 The ℓ_0 region fusion based heuristic

Given a graph $G(V, E)$, Let y_i be the information (either RGB or from any DCNN) of node i, and w_i be its estimated value, the discrete Potts model [31] has the following form:

$$\min_{\mathbf{w}} \sum\nolimits_{i \in V} \|w_i - y_i\|_2 + \sum\nolimits_{(i,j) \in E} \lambda \|w_i - w_j\|_0 , \tag{1}$$

where λ is the regularization parameter. Here, the first term is the data fitting and the second is the regularization term. We recall that the ℓ_0 norm of a vector gives its number of nonzero entries.

In this paper, we introduce an iterative scribble based region fusion heuristic algorithm (which we call $\ell_0 H_l$) with the "class" and "region" ID for each node. In the beginning, the nodes covered by the same scribble are grouped together and labeled with the same IDs, while all other nodes are unlabeled and in their individual group. Note that different regions can share the same class ID. Then the procedure of merging groups follows that of [29], which iterates over each group and its neighboring, except that it needs to first check the region ID. If both groups have region IDs and are different, then they cannot be merged. In all other cases, i.e., if both groups have no region ID or have the same region ID, the following merging criteria of [29] are checked:

$$\sigma_i \cdot \sigma_j \cdot \|Y_i - Y_j\|_2 \leq \beta \cdot \gamma_{ij} \cdot (\sigma_i + \sigma_j). \tag{2}$$

where σ_i denotes the number of pixels in group i, Y_i the mean of image information (e.g., RGB color) of group i, and γ_{i_j} denotes the number of neighboring pixels between groups i and j. Here, β is the regularization parameter, and it increases over the iteration number.

If (2) is satisfied, two groups are merged, and their labels are updated according to the following rule. If both groups have no "region" ID, the merged group still have none, hence unlabeled. If only one group has region ID (by our policy also class ID), the merged group inherit the label, hence labeled.

After one iteration of all groups, β increases (follows the exponential growing strategy of [29]), i.e., $\beta = (\frac{\text{iter}}{50})^{2.2} * \eta$, where "iter" is the current iteration number and η is a parameter. The procedure continues until all groups are labeled, and the complexity of $\ell_0 H_l$ is $\mathcal{O}(n)$ (n is the nodes number).

Note that the above algorithm is approximate to (1), and connectivity of each region is enforced at every step. Given desired scribbles, $\ell_0 H_l$ is able to generate panoptic segmentation (and also semantic segmentation). Also note that the class ID does not play any role in the algorithm, it inherits from the scribble and propagates with the region ID. Hence, this algorithm is class-agnostic.

2.6.3 The ILP formulation with connectivity constraints

The MRF with pairwise data term can be formulated as the following ILP:

$$\min_x \sum_{\ell \in C} \sum_{i \in V} c_i^\ell x_i^\ell + \lambda \sum_{\ell \in C} \sum_{(i,j) \in E} d_{ij} |x_i^\ell - x_j^\ell| \tag{3}$$

$$\sum_{\ell \in C} x_i^\ell = 1, \ \ \forall i \in V, \tag{3a}$$

$$x_i^\ell \in \{0, 1\}, \ \ \forall i \in V, \ \ \ell \in C, \tag{3b}$$

where c_i^ℓ denotes the unary data term for assigning class label ℓ to node i (hence class-aware), d_{ij} the simplified pairwise term for assigning i, j different labels, and λ is the regularization parameter. Constraint (3a) enforces that each node is assigned exactly one label, i.e., $x_i^\ell = 1$ if and only if node i is labeled ℓ. Note that the absolute term can be easily transformed into linear terms by introducing additional continuous variables [41]. The complexity for solving (3) is $\mathcal{O}(2^n)$ ($\mathcal{NP}$-hard), where n is the number of binary variables.

Connectivity constraints with root node Let r (the first node of scribble) denote the root node of the scribbled subgraph (region) in $G(V, E)$. Then, the following constraints suffice to characterize the set of all connected subgraphs that contain r

$$x_i \leq \sum_{s \in S} x_s, \ \ \forall i \in V : (i, r) \notin E, \ \forall S \in \mathcal{S}(i, r), \tag{4}$$

where S, recall from [37], is the *vertex-separator set* of $\{i, r\}$, i.e., if the removal of S from G disconnects i and r. And $\mathcal{S}(i, r)$ is the the collection of all vertex-separator sets of $\{i, r\}$.

The number of constraints (4) is exponential with respect to the number of nodes in G, and in practice, they are added iteratively when needed (called cutting planes method [17]). For the simplest case where the region (scribble) ID coincide with the class ID, i.e., the number of regions equals that of classes, (3) with connectivity prior is solved as follows. We first solve (3) alone and then check if all subgraphs G_i are connected. If not, we iteratively adds constraints of type (4) on the fly [41], and then solve the resulting ILP again. This procedure continues until all subgraphs are connected. This method ensures global optimality if no time limit is restricted.

In the case where k regions share the same class ID, [41] adds $k - 1$ "dummy classes" to retain the connectivity property (we call it ILP-D, and is illustrated in Fig. 2). This has two drawbacks. First of all, the number of binary variables is increased by $(k - 1)|V|$ and thus the complexity increases dramatically. Second, if one uses the probability map of any DCNN, then all the k dummy classes share the same unary term (c_i^ℓ). In the extreme case where no pairwise term d_{ij} exists in (3), there exists symmetry on assigning which of the k dummy classes to one node. Note that, similar to $\ell_0 H_l$, ILP-D is able to directly output panoptic segmentation.

In this paper, we introduce a much lighter formulation (we call it ILP-P). Instead of adding dummy classes, we add $k - 1$ pseudo edges that "connect" all separated regions of the same class (illustrated in Fig. 2), which fixes both issues. In particular, it does not increase the number of variables and is class-aware. But ILP-P is only designed for semantic segmentation, i.e., it is not region-aware. Post-processing methods are needed to generate panoptic segmentation.

3 Semantic and Panoptic Segmentation Experiments

3.1 Experimental setup

In this session, we conduct extensive experiments on the public Pascal VOC [15] and Cityscapes [12] dataset. In all our experiments, when we mention base network, we refer to the publicly released ResNet 101 [16] that is pre-trained on ImageNet [35] and COCO [26]. We adopt DeepLab V2 [10] (without CRF as post-processing) and DRN [50] as our DCNN to get the probability maps, trained on their corresponding training sets. We adopt IBM Cplex [5] version 12.8 to solve the ILP.

All computational experiments are performed on a Intel(R) Xeon(R) CPU E5-2620 v4 machine, with 64 GB memory. Other than using GPU to extract feature maps from DCNN as input to our algorithms, we use CPU to run our algorithms. If not otherwise mentioned, the input data term y_i and parameter η for the heuristic are reported directly in the tables. The regularization parameter λ for the ILP is set

Table 1: Comparison of different models when no training data is available.

Model	Dim	Time	mIoU
$\ell_0 H_g$-RGB	3	1.1	57.7
$\ell_0 H_l$-RGB	3	2.2	69.8
$\ell_0 H_l$-layer 1	64	2.9	70.8
$\ell_0 H_l$-layer 3	256	3.9	71.6
ILP-P-10	–	9.7	**71.9**

Table 2: Comparison when training data is available, using DeepLab V2 [10] as baseline.

Model	Time (s)	mIoU (%)
DeepLab V2 [10]	–	70.5
MRF-prob	0.2	80.8
$\ell_0 H_l$-prob	0.8	80.8
ILP-D-5	4.5	80.9
ILP-P-5	4.2	81.3
ILP-P-10	7.9	**81.9**

to 1, and the pairwise term $d_{ij} = e^{\|y_i - y_j\|_2}$. We have two scenarios. When training data is available, we can use the probability map p_i and $c_i^\ell = \left\|\mathbb{1}^\ell - p_i\right\|_2$, where $\mathbb{1}^\ell$ is an k (k being the number of classes) dimensional vector with $\left\|\mathbb{1}^\ell\right\|_1 = 1$ and the ℓ's position equals 1. When there is no training data, we compute the average of the nodes information (y_i) covered by scribbles of the same class (i.e., class ℓ), and use this to represent class ℓ (denote as Y_ℓ). Then $c_i^\ell = \|y_i - Y_\ell\|_2$.

We report the semantic and panoptic segmentation scores, where the mean intersection over union (mIoU) is commonly used for semantic segmentation, and the panoptic quality (PQ) metric is newly introduce in [20] and is a combination of segmentation quality (SQ) and recognition quality (RQ).

3.2 Results on Pascal VOC 2012

Pascal VOC has 20 "thing" classes and a single "background" class for all other classes. We evaluate our algorithms on the 1449 validation images. We first apply [14] to produce around 700 superpixels, and use the public available scribble set of Pascal provided by [25]. Since the scribbles do not meet our policy for panoptic segmentation, we only report the semantic segmentation results.

No training data When no previous training data is available, one can either use RGB or the output of lower level features of a base network as input (y_i) to our algorithms. We compare our class-agnostic heuristic ($\ell_0 H_l$) to the more greedy one ($\ell_0 H_g$) in [41] using RGB as input. We also compare using different low levels of features from ResNet 101, against using just RGB, as well as ILP-P with $\ell_0 H_l$. We do not set any time or step limit for both heuristics, but a time limit of 10 seconds for ILP-P (denoted ILP-P-10). We report in Table 1 the detailed comparison, where we use the RGB, first and third layer of ResNet 101 as input to $\ell_0 H_l$, and "Dim" is the dimension of the input feature map. After several trials, the growing parameter η is set to 0.1, 20 and 100 for RGB, layer 1 and layer 3 of ResNet 101.

We can see in Table 1 the advantage of $\ell_0 H_l$ over $\ell_0 H_g$, that improves mIoU by over 12%, despite the time increases. We could also see the improvement by incorporating lower level features maps of ResNet 101, that improves mIoU by almost 2%, even though it is pre-trained on completely different dataset. ILP-P adopts $\ell_0 H_l$-layer 3 as initial solution, and further increase the mIoU by 0.3%.

With training data In addition, if training data is available, we use the probability map of DeepLab V2 (with base line mIoU 70.5%) that is trained on Pascal as input information to our algorithms, $\ell_0 H_l$ get another huge boost of 9.2% to 80.8% compared to using layer 3. We compare MFR (3) (without connectivity but solved to global optimum) with $\ell_0 H_l$ (heuristic with connectivity) and found out they perform the same, which shows the importance of the connectivity prior. An example of visual comparison is illustrated in Fig. 1.

We then compare ILP-D (with dummy classes) and ILP-P (with pseudo edges) with $\ell_0 H_l$ as initial solutions (baseline 80.8%), and set the time limit to 5 and 10 seconds. Table 2 suggests that both ILPs improve the baseline within the time limit, where ILP-P outperforms ILP-D in both accuracy and efficiency. Also, because of the $\mathcal{NP}$-hardness, many problems remain non-optimum and it is in general beneficial to allow the ILP solver to run for more time. The best mIoU is reported by ILP-P-10 at 81.9%, which improves the baseline of DeepLab V2 by 11.4% and the initial solution of $\ell_0 H_l$ by 1.1%. Finally, since ILP is class-aware and encodes pairwise term d_{ij}, further boost on the performance is expected, given better baseline DCNN and an edge detector.

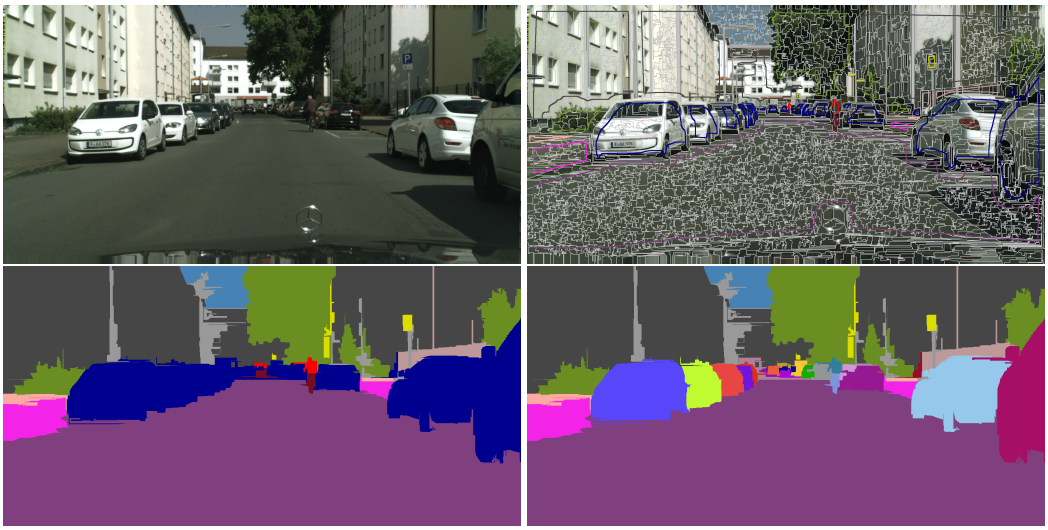

Figure 3: Top: Cityscapes original image and image with superpixels and artificial scribbles, Bottom: semantic and panoptic segmentation of $\ell_0 H_l$-prob, with DCNN's probability map as input.

3.3 Results on Cityscapes

Cityscapes has 8 "thing" classes, and 11 "stuff" class labels. While there is no public scribble set for Cityscapes, we hack the ground truth instance segment and apply erosion and skeleton algorithms to generate artificial scribbles (see Fig. 3) that meet our request in Sec. 2.4. We evaluate our algorithms on the 500 validation images with these scribbles. We first apply [45] to produce around 2000 superpixels, then apply our heuristic and ILP algorithms to get semantic and panoptic segmentation.

No training data When no training data is available, we use RGB and output of layer 3 of Resnet 101 as input to $\ell_0 H_l$, and report in Table 3 both mIoU and PQ. We also compare $\ell_0 H_l$ with one recent weekly supervised methods [24]. It uses a more powerful PSPNet [51] supervised by bounding boxes and image tags, and requires end to end training. $\ell_0 H_l$ shows superior results on both semantic and panoptic segmentation, both around 10% boost of performance compared to [24]. Note that it is not a fair comparison since our algorithm requires scribbles at inference time while [24] does not. The running time is 6.5 and 7.2 seconds for RGB and layer 3.

With training data We use the public full-supervised (trained on Cityscapes) DRN [50] as our baseline (71.4% mIoU), and run $\ell_0 H_l$ and ILP-P using its probability map. Table 4 shows $\ell_0 H_l$-prob improves the baseline by 4.4%, and by 1.5% compared to $\ell_0 H_l$-layer 3 . Besides, PQ also increases from 49.6% to 51.2%. Because of the rich class and instance information (more than 15 classes per image) of Cityscapes and also that we use around 2000 superpixels, which results in nearly 30k binary variables, ILP struggles to find any better solution within the time limit of 10 seconds. Hence, the score remains unchanged compared to $\ell_0 H_l$ in our experiment, which is not shown in the table.

Finally, note that the performance of our algorithms on both Pascal and Cityscapes are upper bounded by the accuracy of superpixels. Hence, using better superpixel algorithms or increasing the superpixel number may further influence our performance.

Table 3: Comparison of different models when no training data is available. Our heuristic is around 10% better than [24] on mIoU and PQ.

Model	mIoU	PQ	SQ	RQ
$\ell_0 H_l$-RGB	74.2	**49.6**	74.3	63.8
$\ell_0 H_l$-layer 3	**74.3**	**49.6**	74.5	63.7
Weakly [24]	63.6	40.5	–	–

Table 4: Comparison when training data is available, based on DRN [50]'s probability map. Our result improves the baseline by 4.4%.

Model	Time	mIoU	PQ	SQ	RQ
DRN [50] (Baseline)	–	71.4	–	–	–
$\ell_0 H_l$-prob	7.2	**75.8**	**51.2**	75.6	64.8

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[2] D. Acuna, H. Ling, A. Kar, and S. Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. 2018.

[3] Rimmi Anand, Divya Aggarwal, and Vijay Kumar. A comparative analysis of optimization solvers. *Journal of Statistics and Management Systems*, 20(4):623–635, 2017.

[4] H. Banzhaf, D. Nienhüser, S. Knoop, and J. M. Zöllner. The future of parking: A survey on automated valet parking with an outlook on high density parking. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1827–1834, 2017.

[5] Christian Bliek, Pierre Bonami, and Andrea Lodi. Solving mixed-integer quadratic programming problems with ibm-cplex : a progress report. 2014.

[6] Y. Boykov and O. Veksler. *Graph Cuts in Vision and Graphics: Theories and Applications*, pages 79–96. Springer US, Boston, MA, 2006.

[7] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, Nov 2001.

[8] Y. Y. Boykov and M. . Jolly. Interactive graph cuts for optimal boundary amp; region segmentation of objects in n-d images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 105–112 vol.1, July 2001.

[9] Yigit B. Can, Krishna Chaitanya, Basil Mustafa, Lisa M. Koch, Ender Konukoglu, and Christian Baumgartner. Learning to segment medical images with scribble-supervision alone. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, volume 11045 of *Lecture Notes in Computer Science*, pages 236 – 244. Springer, 2018.

[10] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, April 2018.

[11] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2018.

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1635–1643, 2015.

[14] M. V. den Bergh, X. Boix, G. Roig, B. de Capitani, and L. Van Gool. Seeds: Superpixels extracted via energy-driven sampling. *European Conference on Computer Vision (ECCV)*, pages 13–26, 2012.

[15] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[17] J. Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960.

[18] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic Feature Pyramid Networks. *arXiv e-prints*, page arXiv:1901.02446, Jan 2019.

[20] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic Segmentation. *arXiv e-prints*, page arXiv:1801.00868, Jan 2018.

[21] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 109–117. Curran Associates, Inc., 2011.

[22] A. H. Land and A. G. Doig. An automatic method for solving discrete programming problems. *ECONOMETRICA*, 28(3):497–520, 1960.

[23] A. Levin, D. Lischinski, and Y. Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):228–242, Feb 2008.

[24] Qizhu Li, Anurag Arnab, and Philip H.S. Torr. Weakly- and semi-supervised panoptic segmentation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[25] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3159–3167, 2016.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[27] Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, and Sanja Fidler. Fast Interactive Object Annotation with Curve-GCN. *arXiv e-prints*, page arXiv:1903.06874, Mar 2019.

[28] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[29] R. M. H. Nguyen and M. S. Brown. Fast and effective l0 gradient minimization by region fusion. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 208–216, Dec 2015.

[30] Dimitrios Papadopoulos, Jasper Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *International Conference on Computer Vision (ICCV 2017)*, pages 4940–4949. IEEE, 12 2017.

[31] R. B. Potts. Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(1):106–109, 1952.

[32] Markus Rempfler, Bjoern Andres, and Bjoern H. Menze. The minimum cost connected subgraph problem in medical image analysis. In *MICCAI*, 2016.

[33] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(06):1137–1149, jun 2017.

[34] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, August 2004.

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[36] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1):157–173, May 2008.

[37] Nowozin S. and CH. Lampert. Global connectivity potentials for random field models. In *CVPR 2009*, pages 818–825, Piscataway, NJ, USA, June 2009. Max-Planck-Gesellschaft, IEEE Service Center.

[38] Timo Sämann, Karl Amende, Stefan Milz, Christian Witt, Martin Simon, and Johannes Petzold. Efficient semantic segmentation for visual bird's-eye view interpretation. In *Proceedings of the 15th International Intelligent Autonomous Systems*, 2018.

[39] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, April 2017.

[40] R. Shen. *MILP Formulations for Unsupervised and Interactive Image Segmentation and Denoising*. PhD thesis, Heidelberg University, 2018.

[41] R. Shen, E. Kendinibilir, I. B. Ayed, A. Lodi, A. Tramontani, and G. Reinelt. An ILP solver for multi-label MRFS with connectivity constraints. *CoRR*, abs/1712.06020, 2017.

[42] Ruobing Shen, Xiaoyu Chen, Xiangrui Zheng, and Gerhard Reinelt. Discrete potts model for generating superpixels on noisy images. *CoRR*, abs/1803.07351, 2018.

[43] Ruobing Shen, Gerhard Reinelt, and Stéphane Canu. A first derivative potts model for segmentation and denoising using milp. In *OR*, 2017.

[44] D. Stutz, A. Hermans, and B. Leibe. Superpixels: an evaluation of the state-of-the-art. *Computer Vision and Image Understanding*, pages 1–32, 2017.

[45] Wei-Chih Tu, Ming-Yu Liu, Varun Jampani, Deqing Sun, Shao-Yi Chien, Ming-Hsuan Yang, and Jan Kautz. Learning superpixels with segmentation-aware affinity loss. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[46] S. Vicente, V. Kolmogorov, and C. Rother. Graph cut based image segmentation with connectivity priors. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[47] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. UPSNet: A Unified Panoptic Segmentation Network. *arXiv e-prints*, page arXiv:1901.03784, Jan 2019.

[48] Tien-Ju Yang, Maxwell D. Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. DeeperLab: Single-Shot Image Parser. *arXiv e-prints*, page arXiv:1902.05093, Feb 2019.

[49] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Exploring artificial intelligence in the new millennium. chapter Understanding Belief Propagation and Its Generalizations, pages 239–269. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

[50] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[51] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[52] Zhi-Hua Zhou. *Ensemble Learning*, pages 270–273. Springer US, Boston, MA, 2009.