How to Save your Annotation Cost for Panoptic Segmentation?

Xuefeng Du[1], **ChenHan Jiang**[2], **Hang Xu**[*2], **Gengwei Zhang**[3], **Zhenguo Li**[2]
[1] Xi'an Jiaotong University
[2] Huawei Noah's Ark Lab
[3] Sun Yat-Sen University
{xuefengdu1,jchcyan,chromexbjxh,zgwdavid}@gmail.com, li.zhenguo@huawei.com

Abstract

How to properly reduce the annotation cost for panoptic segmentation? How to leverage and optimize the cost-quality trade-off for training data and model? These questions are key challenges towards a label-efficient and scalable panoptic segmentation system due to its expensive instance/semantic pixel-level annotation requirements. By closely examining different kinds of cheaper labels, we introduce a novel multi-objective framework to automatically determine the allocation of different annotations, so as to reach a better segmentation quality with a lower annotation cost. Specifically, we design a Cost-Quality Balanced Network (CQB-Net) to generate the panoptic segmentation map, which distills the crucial relations between various supervisions including panoptic labels, image-level classification labels, bounding boxes, and the semantic coherence information between the foreground and background. Instead of ad-hoc allocation during training, we formulate the optimization of cost-quality trade-off as a Multi-Objective Optimization Problem (MOOP). We model the marginal quality improvement of each annotation and approximate the Pareto-front to enable a label-efficient allocation ratio. Extensive experiments on COCO benchmark show the superiority of our method, e.g. achieving a segmentation quality of 43.4% compared to 43.0% of OCFusion while saving 2.4x annotation cost.

Introduction

Panoptic segmentation unifies foreground instance segmentation (named *thing*) and semantic segmentation on amorphous background regions (named *stuff*) to generate rich and coherent masks. It poses a challenging problem on holistic image understanding of all foreground objects and background contents simultaneously. State-of-the-art methods (Yang et al. 2019b) are trained in a fully-supervised fashion, for which per-pixel background class and foreground instance labels are required. Due to the data-hungry nature of deep networks, these approaches demand an enormous number of training images with curated groundtruth labels, which are given by hand in general. However, manual annotation of such labels is extremely labor-intensive and time-consuming, e.g. taking around 19 minutes for one single image in COCO to be fully annotated (Caesar, Uijlings, and Ferrari 2018; Lin et al. 2014). Thus, developing a customized panoptic segmentation model from a large dataset is restricted by both the budget of recruiting annotators and limited class diversity due to the annotation difficulty.

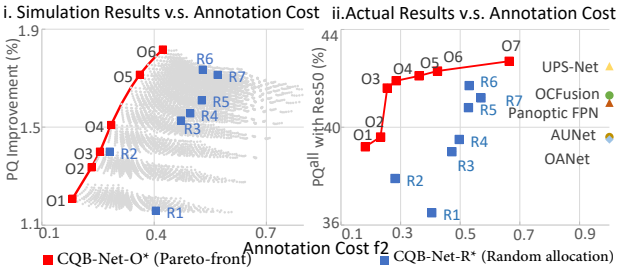

Figure 1: Annotation Cost vs. COCO Panoptic Segmentation Quality. Our CQB-Net obtains a better allocation ratio for different labels by approximating the Pareto-front (left) and in practice achieves a better tradeoff between two essential objectives for panoptic segmentation, which outperforms models trained with random ratios and fully-supervised baselines (right, on ResNet50).

One possible solution to alleviate this issue is adopting weaker labels for training. As shown in Fig.2(a), while the panoptic label is the most expensive (1104s), cheaper labels, such as image labels (7s) and bounding box (32s), can be obtained more easily or even readily available by an image search engine or boundary click (Papadopoulos et al. 2017). Verified by the success of weakly supervised learning (Zhang et al. 2018a; Song et al. 2019; Wan et al. 2019; Ahn and Kwak 2018; Khoreva et al. 2017), these labels and their mutual relationship should provide coarse but useful information for our task. In Fig.2(b), five kinds of annotations carry different semantic/spatial information which may compensate and fill up the missing information during training. However, directly ensembling weakly supervised approaches to the *thing* and *stuff* branches in a hand-crafted fashion, such as (Li, Arnab, and Torr 2018), cannot guarantee to achieve an optimal trade-off between the segmentation quality and the corresponding annotation cost while sometimes certain weak supervision will exert negative impacts on others if trained in a unified framework.

Therefore, a natural question emerges: *Assume we start*

*Corresponding Author

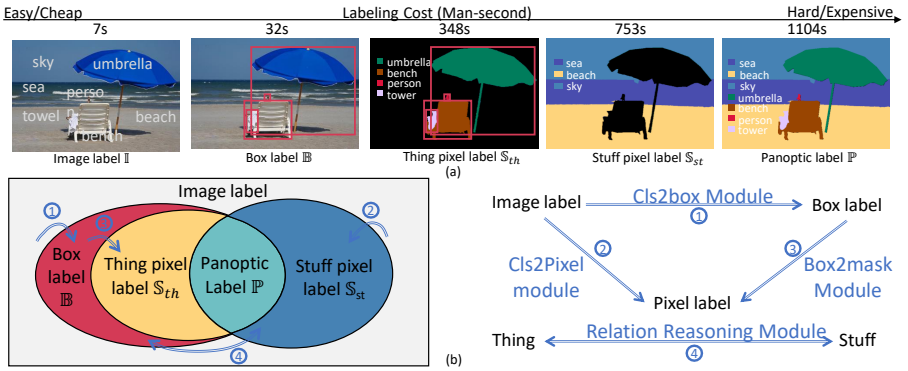

Figure 2: (a) different weaker annotations and their cheaper labeling cost. (b) illustrates the relationship between possible annotations in a Venn graph (left) while CQB-Net incorporates four modules to learn from different labels (right).

with a dataset with partial images annotated in a panoptic way [1]*, how to optimally annotate a dataset from scratch for both a better segmentation quality and a lower annotation cost? Will a curated allocation ratio for different supervisions benefit the trade-off between these two objectives?* In this paper, we propose a Cost-Quality Balanced Network (CQB-Net) for a possible solution. Specifically, CQB-Net formulates this natural question as a Mulit-Objective Optimization Problem (MOOP) with the two objectives to be a high segmentation quality and low annotation cost. Mathematically, the trade-off among different objectives is typically captured by the Pareto front, i.e., the set of Pareto-optimal ratios with the property that no objective can be improved without harming the other objectives.

Obtaining the full Pareto front accurately and efficiently is challenging for panoptic segmentation which requires a large number of network evaluations under different annotation ratios. Randomly sampling allocation ratios for such evaluation is thus a sub-optimal solution. In the CQB-Net, we first assume the sensitivity of the panoptic segmentation quality with respect to each supervision, including panoptic labels, image labels, bounding boxes and the semantic coherence between the foreground and the background annotations (i.e., how changing the ratio of a certain supervision influences the segmentation quality) is different and independent. Then we collect several statistics that reflect such sensitivity and obtain an explicit function between the ratio of each supervision and the marginal segmentation quality improvement by regression. Given such a regression function and the cost for different annotations, we approximately obtain the Pareto-front and pick the Pareto-dominate ratios for training a panoptic segmentation model.

Except for the panoptic labels, we instantiate the benefits of different weak supervisions by weakly-supervised approaches. For image label, we expect it to provide potentially useful information for object detection in the *thing* branch and the semantic segmentation in the *stuff* branch. For bounding boxes, we argue they may provide coarse instance segmentation cues through the box-tightness prior (Hsu et al. 2019). Additionally, we introduce a new bidirectional weak supervision between the *stuff* segmentation masks and the *thing* bounding boxes by a semantic coherence prior (Wu et al. 2020). For instance, if an image contains foreground objects, *person*, *fork*, *cup*, *spoon*, it is more likely to predict the nearby background to be *dining hall*, *house* and vice versa. Thus, *thing* and *stuff* labels can be regarded as weak annotation for each other. Specifically, we develop a relational reasoning network to model their semantic correlation by intra and inter-modular information propagation such that when bounding boxes are missing, the detection head is optimized implicitly by the supervision signal of another branch and vice versa. For simplicity, we denote the four modules as Cls2box, Box2mask, Cls2pixel and relation reasoning module (Fig.2).

Our contributions are summarized as follows:

- We propose a multi-objective framework for label-efficient panoptic segmentation, which is able to predict the Pareto-optimal allocation ratios for different supervisions in order to obtain a trade-off between the segmentation quality and the annotation cost.
- We design a CQB-Net, which is able to incorporate various supervisions in a unified way that performs better than fully-supervised baselines with much lower cost, e.g. 42.7% PQ^{all} (on ResNet-50) compared to 42.5% by UPS-Net and 41.3% by OCFusion.
- The Pareto-dominate ratios help the CQB-Net achieve better segmentation results under similar annotation cost on COCO compared to random ratios (Fig.1).

Related Works

Panoptic Segmentation is a novel task proposed by (Kirillov et al. 2019b). Starting from simple combination of two branches, techniques such as shared backbone (Porzi et al. 2019; Yang et al. 2019a; Li et al. 2018), attention (Li et al. 2019), spatial ranking (Liu et al. 2019), deformable convolutions (Xiong et al. 2019) and overlapping resolving (Yang et al. 2019b) have been utilized to boost the performance. However, few study focuses on reducing the annotation burden except for (Li, Arnab, and Torr 2018; Hou et al. 2020). Both of them only consider limited label scenarios and ig-

[1] It is practical in real-world applications if researchers desire to use a dataset for panoptic segmentation

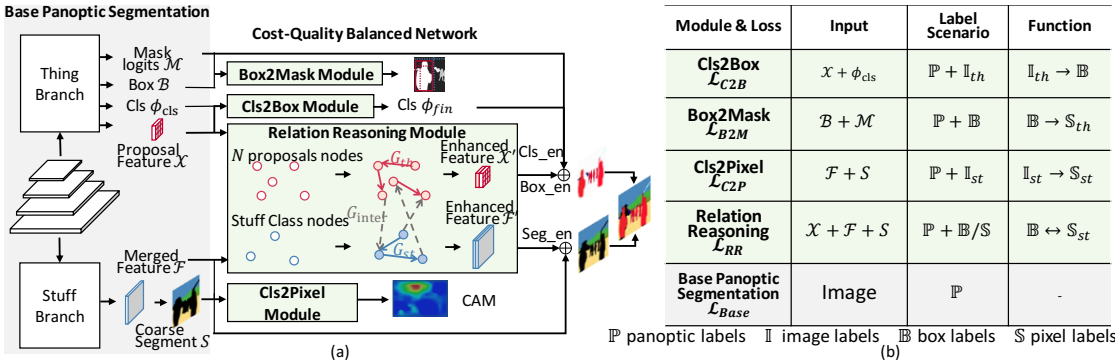

Figure 3: Illustration of the CQB-Net. Based on a panoptic segmentation network, we adopt four modules to unify different supervisions. (a) shows the Input/Output flow of the proposed modules. (b) summarizes the configurations of the four modules, including their input in (a), label scenario and function in terms of weakly-supervised learning.

nore the cost-quality trade-off for different supervisions.

Weakly/Semi-Supervised Learning can be categorized into three types, namely for object detection, semantic segmentation and instance segmentation. For object detection, such approaches train a detector using image labels by Multiple Instance Learning (MIL) (Zhang et al. 2018a; Jie et al. 2017; Tang et al. 2017, 2020). Another direction mines pseudo labels to learn a supervised detector (Zhang et al. 2018b; Shen et al. 2018). The others explore semi-supervised object detection (Fang et al. 2020; Yan et al. 2017) utilizes partial boxes. They are sometimes unstable and require heavy hyperparameter tuning.

Weakly/Semi-supervised approaches for semantic segmentation (Wang et al. 2018; Wei et al. 2017; Singh and Lee 2017; Hung et al. 2018; Ouali, Hudelot, and Tami 2020; Papandreou et al. 2015) and instance segmentation (Hu et al. 2018; Kuo et al. 2019; Khoreva et al. 2017; Hsu et al. 2019; Ge et al. 2019) are also well investigated, which usually require multi-stage training.

In summary, all of them do not take the optimality of allocation ratio for different labels into account, but are able to serve as the weakly-supervised modules in our framework.

Optimization using Pareto-front Pareto-front based approaches are usually used to solve multi-objective optimization, such as for Neural Architecture Search (NAS) (Chen et al. 2020), multi-task learning (Ma, Du, and Matusik 2020) or reinforcement learning (Moffaert and Nowé 2014). In this paper, we solve our MOOP by selecting the annotation ratios that Pareto-dominate others for achieving a trade-off between the segmentation quality and the annotation cost.

Proposed Approach

Preliminaries

Without loss of generality about $\min$ or $\max$, given m objective functions, $f_1 : \mathcal{X} \rightarrow \mathbb{R}, ..., f_m : \mathcal{X} \rightarrow \mathbb{R}$, a Multi-Objective Optimization Problem (MOOP) is formulated as:

$$\max f_1(\boldsymbol{x}), \ldots, \max f_m(\boldsymbol{x}) \text{ s.t. } \boldsymbol{x} \in \mathcal{X}, \quad (1)$$

where $\mathcal{X}$ is a set of samples. Under such condition, a sample $\boldsymbol{x}_1 \in \mathcal{X}$ dominates another sample $\boldsymbol{x}_2 \in \mathcal{X}$ if $f_i(\boldsymbol{x}_1) \geq f_i(\boldsymbol{x}_2), \forall i \in \{1, \ldots, m\}$ and if $f_j(\boldsymbol{x}_1) > f_j(\boldsymbol{x}_2), \exists j \in \{1, \ldots, m\}$. We denote such dominance as $\boldsymbol{x}_1 \succ \boldsymbol{x}_2$. A sample $\boldsymbol{x}^*$ that is Pareto-optimal if there is no sample $\boldsymbol{x} \in \mathcal{X}$ that satisfies $\boldsymbol{x} \succ \boldsymbol{x}^*$. The Pareto-front P_f is a set of the Pareto-optimal samples. As there will hardly be solutions that simultaneously optimize all m objectives, in order to trade-off between different objective functions, the primary solutions are those which are in the Pareto-front.

Problem Formulation

Take it into the context of panoptic segmentation, the samples $\boldsymbol{x}$ are the allocation ratio ρ for different annotations. The two objective functions f_1, f_2 are the segmentation quality PQ^{all} (Panoptic Quality in both the *thing* and *stuff* region) and the annotation cost $\mathcal{C}$, respectively. Denote the annotation cost of the panoptic labels $\mathbb{P}$ as 1, the cost for other weak labels can be calculated as $\mathcal{C}_{\mathbb{I}} = 0.006$ (image labels), $\mathcal{C}_{\mathbb{B}} = 0.03$ (bounding boxes), and $\mathcal{C}_{\mathbb{S}_{st}} = 0.68$ (semantic segmentation maps in the *stuff* branch) following (Caesar, Uijlings, and Ferrari 2018; Li, Arnab, and Torr 2018). Formally, the studied MOOP is formulated similarly as follows:

$$\max PQ^{all}(\rho), -\max \mathcal{C}(\rho) \text{ s.t. } \rho \in \mathcal{X}, \sum \rho = 1, \quad (2)$$

where $\sum \rho = 1$ means $\rho_{\mathbb{P}} + \rho_{\mathbb{B}} + \rho_{\mathbb{I}} + \rho_{\mathbb{S}_{st}} = 1$. Compared to conventional panoptic segmentation models which merely maximize $PQ^{all}(\rho)$ with only panoptic labels $\mathbb{P}$ that leads to an expensive annotation cost $\mathcal{C}(\rho)$, our multi-objective panoptic segmentation framework fully utilizes various annotations to trade-off between the segmentation quality and the cost by finding the optimal allocation ratios ρ.

Overview

An overview of the proposed CQB-Net is shown in Fig.3, which is stacked on a base panoptic segmentation network (Panoptic FPN (Kirillov et al. 2019a)) and includes four modules to solve this problem. When image labels $\mathbb{I}$ are provided, we use Cls2Box module inspired by (Bilen and Vedaldi 2016) to predict a confident score of proposals and refine the RCNN head. Meanwhile, the Cls2Pixel module improves the *stuff* semantic segmentation branch by CAM

Method	Backbone	PQ^{all}	PQ^{th}	PQ^{st}	Cost f_2
JSIS-Net	Res50	26.9	29.3	23.3	1.0
Panoptic FPN		41.0	47.5	31.6	
AUNet		39.6	49.1	25.2	
UPS-Net		42.5	48.6	**33.4**	
OCFusion		41.3	49.4	29.0	
OANet		39.0	48.3	24.9	
AdaptIS		35.9	40.3	29.3	
SpatialFlow		40.9	46.8	31.9	
Panoptic FPN	Res101	41.9	48.9	32.9	
OANet		40.7	50.0	26.6	
OCFusion		43.0	51.1	30.7	
AdaptIS		37.0	41.8	29.9	
SpatialFlow		42.9	49.5	33.0	
CQB-Net-R4	Res50	39.5	47.6	27.3	0.496
CQB-Net-R7		41.2	48.7	31.2	0.530
CQB-Net-O1		39.2	46.6	28.2	0.182
CQB-Net-O6		42.3	49.8	31.0	0.423
CQB-Net-O7		42.7	49.5	32.3	0.664
CQB-Net-O1	Res101	40.7	48.2	29.3	0.182
CQB-Net-O6		**43.4**	**51.8**	32.3	**0.423**

CQB-Net	Ratio				Actual	Improvement	Cost
	$\rho_{\mathbb{P}}$	$\rho_{\mathbb{I}}$	$\rho_{\mathbb{B}}$	$\rho_{\mathbb{S}_{st}}$	PQ^{all}	f_1	f_2
O1	0.067	0.733	0.067	0.133	39.2	1.208	0.182
O2	0.125	0.687	0.063	0.125	39.6	1.336	0.233
O3	0.083	0.457	0.375	0.083	41.6	1.400	0.255
O4	0.120	0.440	0.360	0.080	41.9	1.509	0.285
O5	0.214	0.394	0.321	0.071	42.1	1.714	0.362
O6	0.290	0.355	0.290	0.065	42.3	1.817	0.423
O7	0.500	0.166	0.167	0.167	**42.7**	-	0.664
R1	0.1	0.4	0.1	0.4	36.5	1.159	0.404
R2	0.15	0.55	0.2	0.1	37.9	1.463	0.281
R3	0.2	0.15	0.45	0.2	39.0	1.527	0.472
R4	0.25	0.25	0.25	0.25	39.5	1.556	0.496
R5	0.3	0.2	0.3	0.2	40.8	1.610	0.527
R6	0.4	0.1	0.45	0.05	41.7	1.713	0.570
R7	0.3	0.24	0.23	0.23	41.2	1.735	0.530

Table 1: Left: Main results on COCO. Our CQB-Net achieves comparable performance with fully-supervised methods even using random annotation ratios (CQB-Net-R*). The CQB-Net-O6 outperforms SOTA methods with less than 50% annotation cost on Res101. "CQB-Net-O*" indicates Pareto-optimal ratios on Pareto-front. Right: Specific ratios and performance (estimated improvement and actual segmentation quality) of CQB-Net with ResNet50. The comparison is also shown in Fig.1.

and energy regularization (Zhang et al. 2019). Given box labels $\mathbb{B}$, the Box2Mask module helps estimate instance masks via the box tightness prior (Hsu et al. 2019). Moreover, a relation reasoning module (Wu et al. 2020) incorporates mutual correlations between boxes $\mathbb{B}$ and *stuff* pixel labels $\mathbb{S}_{st}$, such that two supervisions complement each other. All of these modules are used to collect the final segmentation quality PQ^{all} under different allocation ratios. Although there are certainly alternative approaches, we argue that these weakly supervised approaches are simple and effective enough, which does not require iterative training and heavy hyperparamter tuning.

We firstly introduce the key modules for weakly supervised learning and then explain our solution in detail.

Cost-Quality Balanced Network

In this section, we discuss four modules in CQB-Net to handle additional weak supervisions, including image labels, bounding boxes and semantic coherence between foreground and background. Panoptic labels are trained only through the base panoptic segmentation network.

Image Labels The image labels are commonly used in weakly-supervised approaches to help object detection and semantic segmentation. Although several works adopted image labels for instance segmentation (Ge et al. 2019), the multi-stage training and iterative refinement strategies are contradictory with the simplicity requirement of the entire system, let alone their mediocre performance.

For object detection, we include the strategies (Bilen and Vedaldi 2016) in the Cls2box module by coupling a Multiple Instance Detection (MID) head with the common two-stage detector. This MID head obtains a output $\phi_{MID}(x) \in \mathbb{R}^{N \times C_{th}}$ with N, C_{th} to be the number of proposals and foreground classes, which denotes the probability that an image label falls into those proposals. The final classification score is obtained by multiplying $\phi_{MID}(x)$ with the classification output from the common detector for training with only the cross entropy loss $\mathcal{L}_{C2B}$.

For semantic segmentation, the Cls2pixel module adopts (Zhang et al. 2019) as the weakly-supervised approach, which is based on CAM from an extended classification branch to generate pseudo semantic maps for the *stuff* classes. Additionally, an energy regularization term (Joy et al. 2019) is added to ensure color and spatial coherence on the pseudo label. The final loss for the Cls2Pixel module is calculated as $\mathcal{L}_{C2P} = \mathcal{L}_{pseudo} + \mathcal{L}_{cls} + \mathcal{L}_{energy}$ which corresponds to three different tasks.

Bounding Boxes Given the bounding boxes, the Box2mask module utilizes the box tightness prior (Hsu et al. 2019) to approximate the instance segmentation maps, which assumes the the box is the smallest rectangle that encloses the whole instance so that the instance touches four sides of the box. The region outside the box has no overlap with instances. Then, it uses the multi-instance learning based heuristics to constrain the instance map. Furthermore, to avoid segmenting only discriminative regions, a regularization is added to propagate the information from those regions to their neighborhood. The module is end-to-end trainable through the loss $\mathcal{L}_{B2M} = \mathcal{L}_{MIL} + \mathcal{L}_{structure}$. where the former one is the loss to constrain the instance map and the latter one is the structure regularization loss.

Semantic Coherence between Foreground and Background While the weakly-supervised literature overlooks the underlying relations between foreground and background, the proposed Relation Reasoning module (Wu et al. 2020) propagates the label information between $\mathbb{B}$ and $\mathbb{S}_{st}$ via a bidirectional scheme. Different from (Wu et al. 2020) that tries to exploit the relationship between two branches in a fully-supervised way, we adopt this module to distill bene-

ficial information for supporting weakly-supervised learning in a bidirectional way.

To instantiate the semantic coherence prior, we first build two learnable image-specific graphs G_{th} for the RCNN head and G_{st} for the *stuff* segmentation head to enable flexible reasoning in the instance and class level. Then we develop a super graph $\mathbf{G}$ to exploit the diverse relationship between these two branches. Afterwards, we project the diffused graph node features back for intra-modular reasoning, which enhances the features in both branches so that the supervision from one branch benefits the other. Therefore, the loss for the relation reasoning module over the enhanced predictions is formulated as: $\mathcal{L}_{RR} = \mathcal{L}_{cls\_en} + \mathcal{L}_{reg\_en} + \beta\mathcal{L}_{seg\_en}$, where β is the loss weight for the *stuff* segmentation branch. When *stuff* pixel labels are missing, the semantic segmentation branch is optimized by $\mathcal{L}_{cls\_en}$ and $\mathcal{L}_{reg\_en}$ through relation reasoning modules. Similarly, $\mathcal{L}_{seg\_en}$ improves the detection performance when boxes are unavailable.

Overall End-to-End Training

We train the above four modules through corresponding weak annotations and also the panoptic label $\mathbb{P}$ as Fig.3(b). For the base panoptic network, recall that the predictions of the regression layer in RCNN head are highly noisy since we have limited data to train it. For progressive refinement, we adopt a cascade structure with increasing IoU threshold for the *thing* branch. Denote λ_i as the loss weight at the i-th cascade stage and $\mathcal{L}_{base}$ as the conventional loss of the base panoptic network, our CQB-Net is end-to-end trainable via:

$$\mathcal{L} = \mathcal{L}_{base} + \beta\mathcal{L}_{C2P} + \sum_i \lambda_i(\mathcal{L}_{C2B} + \mathcal{L}_{B2M} + \mathcal{L}_{RR}). \quad (3)$$

We empirically found training with such a multi-task loss enables easy convergence and we can train with the default setting of a Panoptic FPN (Kirillov et al. 2019a) and Cascade RCNN (Cai and Vasconcelos 2018). More specifically, we use three cascade heads with increasing IoU threshold to be $0.5, 0.6, 0.7$. The loss weights λ_i are set to $1, 0.5, 0.25$, respectively. β is set to 0.5.

MOOP by Pareto-front

One naive idea for MOOP is to regress the relationship between the value PQ^{all} and the allocation ratio ρ such that a trade-off can be found between the cost $\sum_i \mathcal{C}_i\rho_i$ and PQ^{all} to reach Pareto-optimality. However, such approximation is hard because of the randomness caused by different groups of data for annotation or different trials, thus we resort to the marginal segmentation quality improvement $\nabla PQ^{all}_{\rho_i}$ w.r.t. the allocation ratio of different supervisions. Specifically, we desire to observe how changing the ratio of a certain supervision leads to the change of PQ^{all} by ablative experiments on each supervision. Recall that $PQ^{all} = r_{th}PQ^{th} + r_{st}PQ^{st}$ where r_{th}, r_{st} denotes the class ratio in the *thing* and *stuff* branch (i.e., $\frac{80}{133}$ and $\frac{53}{133}$ in COCO), we further disentangle the $\nabla PQ^{all}_{\rho_i}$ as:

$$\nabla PQ^{all}_{\rho_i} = r_{th}\nabla PQ^{th}_{\rho_i} + r_{st}\nabla PQ^{st}_{\rho_i}, \quad (4)$$

For the panoptic labels $\mathbb{P}$, we did a series of experiments in which QCB-Net only has access to $\mathbb{P}$ with different ratios. For instance, for two ratios $\rho^1_{\mathbb{P}}, \rho^2_{\mathbb{P}}$ ($\rho^2_{\mathbb{P}} \geq \rho^1_{\mathbb{P}}$), we regress such two functions:

$$\begin{aligned} \nabla PQ^{st}_{\rho_{\mathbb{P}}} &= PQ^{st}_{\rho^2_{\mathbb{P}}} - PQ^{st}_{\rho^1_{\mathbb{P}}} = f^{st}_{\mathbb{P}}(\rho^1_{\mathbb{P}}), \\ \nabla PQ^{th}_{\rho_{\mathbb{P}}} &= PQ^{th}_{\rho^2_{\mathbb{P}}} - PQ^{th}_{\rho^1_{\mathbb{P}}} = f^{th}_{\mathbb{P}}(\rho^1_{\mathbb{P}}), \end{aligned} \quad (5)$$

where we assume the segmentation quality improvement is a function of $\rho^1_{\mathbb{P}}$. Note that $\rho^1_{\mathbb{P}}$ and $\rho^2_{\mathbb{P}}$ can be used interchangeably. The empirical results used for regression are shown in the next section (Tab.2).

For other weak supervisions, we approximate the function in a different way from Eqn.5 since training the CQB-Net with only one weak supervision will lead to highly noisy segmentation results, which cannot be stably used to capture the relationship between the PQ^{all} improvement and the allocation ratio. Instead, we accompany each weak supervision with partial panoptic labels so that the CQB-Net is trained with a hybrid supervision. For instance, when regressing $\nabla PQ^{th}_{\rho_{\mathbb{B}}}$, we conduct two groups of experiments, the first one is baseline experiment when the CQB-Net is trained with $\rho_{\mathbb{P}}$ panoptic labels and $\rho_{\mathbb{B}} = 1 - \rho_{\mathbb{P}}$ box labels without the help of weak supervision (i.e. the Box2mask module), whose segmentation quality is $PQ^{th}_{\rho^1_{\mathbb{B}}}$. Then we did another experiment with the same annotation ratio but with the weakly-supervised loss from the Box2mask module and obtain a segmentation quality $PQ^{th}_{\rho^2_{\mathbb{B}}}$ (These two experiments correspond to the "Baseline" and "Box2Mask Module" block in Tab.2). Then we regress such a function:

$$\nabla^2 PQ^{th}_{\rho_{\mathbb{B}}} = \frac{PQ^{th}_{\rho^2_{\mathbb{B}}}}{PQ^{th}_{\rho^1_{\mathbb{B}}}} - 1 = f^{th}_{\mathbb{B}}(\rho_{\mathbb{B}}), \quad (6)$$

which can be regarded as the second derivative of the segmentation quality w.r.t. the ratio of bounding boxes. For $\nabla^2 PQ^{st}_{\rho_{\mathbb{B}}}$, the regression procedure is similar which also requires two groups of experiments, but the weak supervision comes from the relation reasoning module. Afterwards, we obtain the relationship between the segmentation quality improvement and the annotation ratio through integration:

$$\nabla PQ^k_{\rho_{\mathbb{B}}}(\rho^j_{\mathbb{B}}) = \int_0^{\rho^j_{\mathbb{B}}} f^k_{\mathbb{B}}(\rho_{\mathbb{B}})d\rho_{\mathbb{B}}, k \in [th, st], \quad (7)$$

where $\rho^j_{\mathbb{B}}$ is a random ratio with which we want to calculate the segmentation quality improvement. Similarly, this approximation procedure can be applied to other weak supervisions. The regressed functions are shown as follows:

$$\begin{aligned} \nabla PQ^{th}_{\rho_{\mathbb{P}}} &= 0.3101 - 1.407\rho_{\mathbb{P}} + 2.420\rho^2_{\mathbb{P}} - 1.439\rho^3_{\mathbb{P}}, \\ \nabla PQ^{th}_{\rho_{\mathbb{I}}} &= 0.0043 + 0.0422\rho_{\mathbb{I}} - 0.2969\rho^2_{\mathbb{I}} + 0.5146\rho^3_{\mathbb{I}}, \\ \nabla PQ^{th}_{\rho_{\mathbb{B}}} &= 0.0047 + 0.0910\rho_{\mathbb{B}} - 0.4602\rho^2_{\mathbb{B}} + 0.8664\rho^3_{\mathbb{B}}, \\ \nabla PQ^{th}_{\rho_{\mathbb{S}_{st}}} &= 0.0096 - 0.1837\rho_{\mathbb{S}_{st}} + 0.7641\rho^2_{\mathbb{S}_{st}} - 0.7982\rho^3_{\mathbb{S}_{st}}, \\ \nabla PQ^{st}_{\rho_{\mathbb{P}}} &= 0.3324 - 1.592\rho_{\mathbb{P}} + 2.927\rho^2_{\mathbb{P}} - 1.850\rho^3_{\mathbb{P}}, \\ \nabla PQ^{st}_{\rho_{\mathbb{I}}} &= -0.0089 + 0.0439\rho_{\mathbb{I}} + 0.19\rho^2_{\mathbb{I}} - 3.148\rho^3_{\mathbb{I}}, \\ \nabla PQ^{st}_{\rho_{\mathbb{B}}} &= 0.024 - 0.1548\rho_{\mathbb{B}} + 0.4131\rho^2_{\mathbb{B}} - 0.3491\rho^3_{\mathbb{B}}, \end{aligned} \quad (8)$$

where we do not include $\nabla PQ^{st}_{\rho_{\mathbb{S}_{st}}}$ since using *stuff* semantic pixel labels to help the background semantic segmentation is not a valid weak supervision. Then, one objective

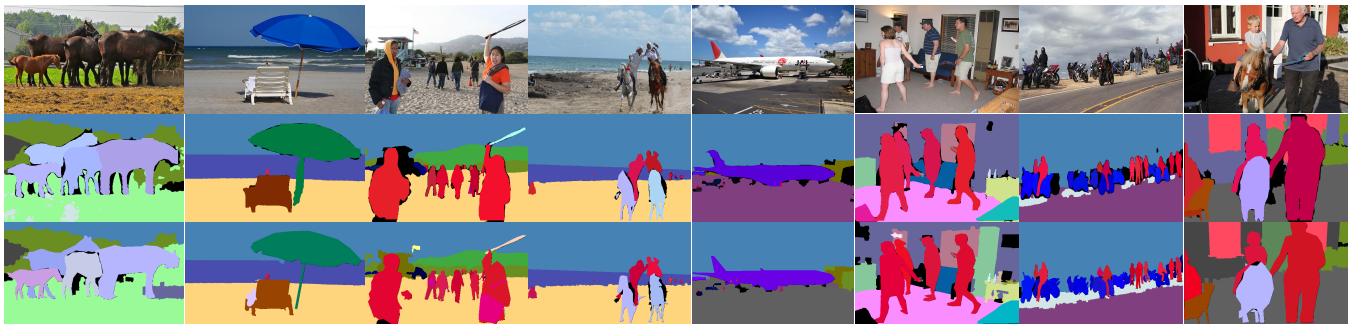

Figure 4: Panoptic segmentation results. The first row shows the original images. The second row shows the CQB-Net segmentation results with $\rho_{\mathbb{P}} = 0.290, \rho_{\mathbb{I}} = 0.355, \rho_{\mathbb{B}} = 0.290, \rho_{\mathbb{S}_{st}} = 0.065$ on ResNet-50. The third shows the ground truth. Additional visualization results are attached in the appendix. The corresponding annotation cost is approximately 0.423 for the second row while obtaining the groundtruth requires a cost of 1.0.

function in the Pareto-optimal problem is the absolute segmentation quality improvement value:

$$f_1(\rho) = \prod(1 + \nabla PQ^k_{\rho_i}), i \in [\mathbb{P}, \mathbb{I}, \mathbb{B}, \mathbb{S}_{st}], k \in [st, th], \quad (9)$$

while the other objective function is the annotation cost $f_2(\rho) = \sum_i \rho_i \mathcal{C}_i$. With such approximated functions, we randomly sample 7500 different allocation ratios with $\sum_i \rho_i = 1$ and obtain their Pareto-front ratios for training the CQB-Net, which is expected to achieve a better trade-off between the annotation cost and the segmentation quality. The sampled ratios and the approximated Pareto-front are shown on the left of Fig.1.

Experiments and Results

Dataset and Implementation Details

We conduct experiments on MS-COCO 2017 (Lin et al. 2014). It has 80 *thing* categories and 53 *stuff* categories, which is divided into train set (118K images), val set (5K images) and test set (20K unannotated images). We train our model on the train set and evaluate on the val set. Panoptic FPN (Kirillov et al. 2019a) with ResNet-50/101 (He et al. 2016) pretrained on ImageNet is regarded as our base panoptic network. We implement our model using MMDetection (Chen et al. 2019) and train with 8 GPUs. We train for 24 epochs with a batch size of 16, weight decay of 1e-4, learning rate of 0.02 with step decay at epoch 18 and 23 by 0.1. We use SGD optimizer with momentum of 0.9. We use multi-scale training between 1333×400 and 1333×900 pixels with random flipping. The test image scale is 1333×800. We use 2 inter-graph reasoning layers and the dimension of $\tilde{\mathbf{f}}$, f^{en}_{th} and f^{en}_{st} are 128. we average the extra box classification and semantic segmentation outputs in the relation reasoning module with those from the base panoptic network as in Fig.3. All the other hyperparameters are kept the same as the original papers. We run 50 ablative experiments for each supervision to approximate Eqn.8.

Comparison with Non Pareto-optimal Ratios

In Tab.1 right, we report *Panoptic Quality* on both *thing* branch PQ^{th}, *stuff* branch PQ^{st} and PQ^{all} of the model trained with the Pareto-optimal ratios. For simplicity, we select 7 ratios from them. Also, we randomly select 7 other ratios with similar annotation cost and compare the segmentation quality with our CQB-Net. As can be seen, our CQB-Net trained with the Pareto-optimal ratios on the ResNet-50 achieves a better segmentation quality compared to randomly sampled ratios (i.e. 42.3% PQ^{all} of $O6$ vs. 41.2% of $R6$), which is also shown in Fig.1 and aligns well with our statistical findings in the MOOP. The qualitative results are shown in Fig.4, where CQB-Net can segment tiny and overlapping objects with comparable visual quality to ground truth. With the ResNet-50 as backbone, the inference time (on V100) for CQB-Net is 283ms/image while that of Panoptic FPN (3 cascade stages) is 256ms. The increased FLOPs are 14.02G in CQB-Net.

Comparison with Fully Supervised Approaches

In order to further demonstrate the effectiveness of the CQB-Net, we compare its segmentation quality with fully-supervised panoptic segmentation baselines in Tab.1. As can be seen, the CQB-Net trained with the Pareto-optimal ratio $O7$ is better than UPS-Net (Xiong et al. 2019) with less than half of annotation cost (i.e. 42.7% PQ^{all} vs. 42.5%). With ResNet-101 as backbone, the performance is boosted by around 1% for all the metrics. In addition, we note that even the CQB-Net trained with the random allocation ratio $R6$ can outperform fully-supervised segmentation models (i.e. 41.2% PQ^{all} of $R6$ vs. 41.0% of Panoptic FPN).

Ablation Study for Pareto-front Approximation

To approximate the relationship between the marginal segmentation quality improvement and the ratio of each weak supervision, each time we run the panoptic segmentation experiments on CQB-Net with $\rho_{\mathbb{P}}$ fully-annotated data and $1 - \rho_{\mathbb{P}}$ data with only one type of weak annotation. With the backbone of ResNet-50, we report the panoptic metrics, i.e. $PQ^{th}, SQ^{th}, RQ^{th}$ for Cls2Box module, Box2Mask module and Relation Reasoning module (*st*→*th*). $PQ^{st}, SQ^{st}, RQ^{st}$ are reported for the Cls2Pixel module and Relation Reasoning module (*th*→*st*) (Tab.2).

For each module, we compare with a baseline model and

$\rho_{\mathbb{P}}$	Baseline		Cost	Cls2Box		Cost	Upper bound		Cost
0.1	40.1	79.1	0.10	44.0	79.4	0.10	47.3	79.9	0.13
0.3	47.7	81.4	0.30	49.2	82.3	0.30	50.7	82.7	0.32
0.5	50.4	82.1	0.50	51.4	82.5	0.50	51.7	82.6	0.51
$\rho_{\mathbb{P}}$	Baseline		Cost	Box2Mask		Cost	Upper bound		Cost
0.1	47.3	79.9	0.13	48.8	80.0	0.13	49.4	81.6	0.38
0.3	50.7	81.7	0.32	51.0	81.9	0.32	51.2	82.2	0.52
0.5	51.7	82.6	0.51	52.0	82.6	0.51	52.1	82.6	0.66
$\rho_{\mathbb{P}}$	Baseline		Cost	GR $_{st\rightarrow th}$		Cost	Upper bound		Cost
0.1	39.1	76.6	0.71	39.7	77.3	0.71	47.5	80.0	0.74
0.3	47.6	81.0	0.78	48.3	81.3	0.78	50.6	82.5	0.80
0.5	49.9	81.3	0.84	50.6	81.5	0.84	51.8	82.5	0.86

$\rho_{\mathbb{P}}$	Baseline		Cost	Cls2Pixel		Cost	Upper bound		Cost
0.1	25.3	70.9	0.10	26.5	72.4	0.10	31.6	75.8	0.71
0.3	30.3	73.5	0.30	31.5	74.3	0.30	33.2	75.0	0.78
0.5	32.0	77.3	0.50	32.5	77.7	0.50	33.3	78.3	0.84
$\rho_{\mathbb{P}}$	Baseline		Cost	GR $_{th\rightarrow st}$		Cost	Upper bound		Cost
0.1	24.4	66.6	0.13	26.1	68.0	0.13	30.8	73.2	0.74
0.3	29.9	72.3	0.32	31.3	**75.9**	0.32	33.4	75.8	0.80
0.5	31.3	76.3	0.51	**32.8**	**79.9**	0.51	32.7	78.0	0.86

Table 2: Ablation study results on different weakly-supervised modules with different allocation ratios ρ. These statistics are used for Pareto-front approximation. The baseline model has available weak labels but is trained without related module while upper bound model accepts fine-grained annotation. Results are reported in the fashion of $PQ^{th}SQ^{th}|f_2(\rho)$ (left) and $PQ^{st}SQ^{st}|f_2(\rho)$ (right). For simplicity, three experiments for each scenario are reported where we conduct 50 experiments per label to approximate Eqn.8.

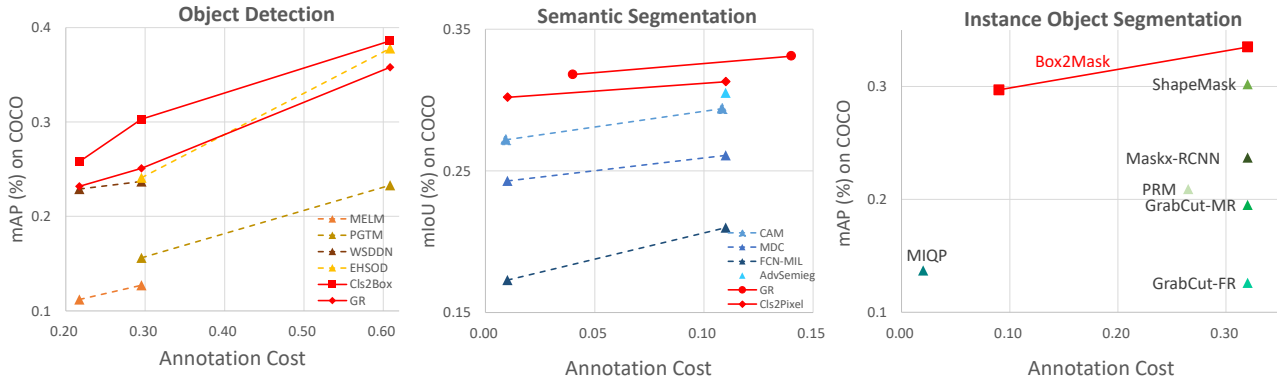

Figure 5: Comparative results on three sub-tasks with weakly/semi-supervised baselines under different annotation costs. We report the metrics on baselines with the same backbone. GR refers to the graph relation reasoning module.

an upper bound model. The former one is trained with the same weak labels but not with the related weakly-supervised module. The latter one accepts fine-grained annotations, which is regarded as the upper bound. For instance, for the Cls2Box module, the baseline model is trained with $\rho_{\mathbb{P}}$ panoptic labels and $1-\rho_{\mathbb{P}}$ image labels. And the upper bound model is trained with the same $\rho_{\mathbb{P}}$ panoptic labels but $1-\rho_{\mathbb{P}}$ bounding boxes. As can be seen, all of the four modules are able to lift the panoptic segmentation quality from the baseline model with a considerable margin. Using such statistics, we then approximate Eqn.8 for obtaining the optimal allocation ratios in the Pareto-front.

Comparison on Sub-tasks

With the same setting above, we report mAP over Cls2Box module and Relation Reasoning module (*st→th*) for object detection and over Box2Mask module for instance segmentation. mIoU is reported for Cls2Pixel and Relation Reasoning module (*th→st*) for semantic segmentation (Fig.5). We compare our Cls2Box module and relation reasoning module (*st→th*) with weakly-supervised approaches, namely, EHSOD (Fang et al. 2020), PGTM (Zhang et al. 2020), MELM (Wan et al. 2019), WSDDN (Bilen and Vedaldi 2016) trained with the same ratio of box and image labels. The Cls2Pixel and the relation reasoning module (*th→st*) are compared with CAM (Zhou et al. 2016), MDC (Wei et al. 2018), FCN-MIL (Pathak et al. 2015), AdvSeg (Hung et al. 2018) trained with the same ratio of *stuff* pixel labels and image labels. Besides, the Box2Mask module is compared to ShapeMask (Kuo et al. 2019), $Mask^X$RCNN (Hu et al. 2018), GrabCut-FasterRCNN (Rother, Kolmogorov, and Blake 2004), GrabCut-MaskRCNN (Hu et al. 2018), PRM (Laradji, Vázquez, and Schmidt 2019). Our four modules show outperforming results in all three tasks, e.g. 30.3% mAP for object detection compared to 24.1% on EHSOD given only 10% bounding boxes and 90% image labels. More comparative results are in the appendix.

Conclusion

We propose a multi-objective framework for panoptic segmentation where a CQB-Net is designed to fully utilize the available cheaper annotations in a dataset. In order to trade-off between the segmentation quality and the annotation cost instead of ad-hoc ensembling different weakly-supervised approaches, we formulate a Multi-Objective Optimization Problem and select the allocation ratio for different supervisions from the approximated Pareto-front. These allocation ratios empirically enable CQB-Net to perform better than models trained with random ratios and even outperform fully supervised methods with $\leq$50% annotation cost.

References

Ahn, J.; and Kwak, S. 2018. Learning Pixel-Level Semantic Affinity With Image-Level Supervision for Weakly Supervised Semantic Segmentation. In *CVPR*.

Bilen, H.; and Vedaldi, A. 2016. Weakly Supervised Deep Detection Networks. In *CVPR*, 2846–2854.

Caesar, H.; Uijlings, J. R. R.; and Ferrari, V. 2018. COCO-Stuff: Thing and Stuff Classes in Context. In *CVPR*.

Cai, Z.; and Vasconcelos, N. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. In *CVPR*.

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; Zhang, Z.; Cheng, D.; Zhu, C.; Cheng, T.; Zhao, Q.; Li, B.; Lu, X.; Zhu, R.; Wu, Y.; Dai, J.; Wang, J.; Shi, J.; Ouyang, W.; Loy, C. C.; and Lin, D. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *CoRR* abs/1906.07155.

Chen, Z.; Zhou, F.; Trimponias, G.; and Li, Z. 2020. Multi-objective Neural Architecture Search via Non-stationary Policy Gradient. *CoRR* abs/2001.08437.

Fang, L.; Xu, H.; Liu, Z.; Parisot, S.; and Li, Z. 2020. EHSOD: CAM-Guided End-to-end Hybrid-Supervised Object Detection with Cascade Refinement. *CoRR* abs/2002.07421.

Ge, W.; Huang, W.; Guo, S.; and Scott, M. R. 2019. Label-PEnet: Sequential Label Propagation and Enhancement Networks for Weakly Supervised Instance Segmentation. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hou, R.; Li, J.; Bhargava, A.; Raventos, A.; Guizilini, V.; Fang, C.; Lynch, J. P.; and Gaidon, A. 2020. Real-Time Panoptic Segmentation From Dense Detections. In *CVPR*.

Hsu, C.; Hsu, K.; Tsai, C.; Lin, Y.; and Chuang, Y. 2019. Weakly Supervised Instance Segmentation using the Bounding Box Tightness Prior. In *NeurIPS*.

Hu, R.; Dollár, P.; He, K.; Darrell, T.; and Girshick, R. B. 2018. Learning to Segment Every Thing. In *CVPR*.

Hung, W.; Tsai, Y.; Liou, Y.; Lin, Y.; and Yang, M. 2018. Adversarial Learning for Semi-supervised Semantic Segmentation. In *BMVC*.

Jie, Z.; Wei, Y.; Jin, X.; Feng, J.; and Liu, W. 2017. Deep Self-Taught Learning for Weakly Supervised Object Localization. In *CVPR*.

Joy, T.; Desmaison, A.; Ajanthan, T.; Bunel, R.; Salzmann, M.; Kohli, P.; Torr, P. H. S.; and Kumar, M. P. 2019. Efficient Relaxations for Dense CRFs with Sparse Higher-Order Potentials. *SIAM J. Imaging Sciences* 12(1): 287–318.

Khoreva, A.; Benenson, R.; Hosang, J. H.; Hein, M.; and Schiele, B. 2017. Simple Does It: Weakly Supervised Instance and Semantic Segmentation. In *CVPR*.

Kirillov, A.; Girshick, R. B.; He, K.; and Dollár, P. 2019a. Panoptic Feature Pyramid Networks. In *CVPR*.

Kirillov, A.; He, K.; Girshick, R. B.; Rother, C.; and Dollár, P. 2019b. Panoptic Segmentation. In *CVPR*.

Kuo, W.; Angelova, A.; Malik, J.; and Lin, T. 2019. ShapeMask: Learning to Segment Novel Objects by Refining Shape Priors. In *ICCV*.

Laradji, I. H.; Vázquez, D.; and Schmidt, M. 2019. Where are the Masks: Instance Segmentation with Image-level Supervision. In *BMVC*.

Li, J.; Raventos, A.; Bhargava, A.; Tagawa, T.; and Gaidon, A. 2018. Learning to Fuse Things and Stuff. *CoRR* abs/1812.01192.

Li, Q.; Arnab, A.; and Torr, P. H. S. 2018. Weakly- and Semi-supervised Panoptic Segmentation. In *ECCV*.

Li, Y.; Chen, X.; Zhu, Z.; Xie, L.; Huang, G.; Du, D.; and Wang, X. 2019. Attention-Guided Unified Network for Panoptic Segmentation. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, H.; Peng, C.; Yu, C.; Wang, J.; Liu, X.; Yu, G.; and Jiang, W. 2019. An End-To-End Network for Panoptic Segmentation. In *CVPR*.

Ma, P.; Du, T.; and Matusik, W. 2020. Efficient Continuous Pareto Exploration in Multi-Task Learning. *CoRR* abs/2006.16434.

Moffaert, K. V.; and Nowé, A. 2014. Multi-objective reinforcement learning using sets of pareto dominating policies. *JMLR* .

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-Supervised Semantic Segmentation with Cross-Consistency Training. *CoRR* abs/2003.09005.

Papadopoulos, D. P.; Uijlings, J. R. R.; Keller, F.; and Ferrari, V. 2017. Extreme Clicking for Efficient Object Annotation. In *ICCV*.

Papandreou, G.; Chen, L.; Murphy, K. P.; and Yuille, A. L. 2015. Weakly-and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation. In *ICCV*.

Pathak, D.; Shelhamer, E.; Long, J.; and Darrell, T. 2015. Fully Convolutional Multi-Class Multiple Instance Learning. In *ICLR*.

Porzi, L.; Bulò, S. R.; Colovic, A.; and Kontschieder, P. 2019. Seamless Scene Segmentation. In *CVPR*.

Rother, C.; Kolmogorov, V.; and Blake, A. 2004. "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.* 23(3): 309–314.

Shen, Y.; Ji, R.; Zhang, S.; Zuo, W.; and Wang, Y. 2018. Generative Adversarial Learning Towards Fast Weakly Supervised Detection. In *CVPR*.

Singh, K. K.; and Lee, Y. J. 2017. Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-Supervised Object and Action Localization. In *ICCV*.

Song, C.; Huang, Y.; Ouyang, W.; and Wang, L. 2019. Box-Driven Class-Wise Region Masking and Filling Rate Guided Loss for Weakly Supervised Semantic Segmentation. In *CVPR*.

Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; and Yuille, A. L. 2020. PCL: Proposal Cluster Learning for Weakly Supervised Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 42(1): 176–191.

Tang, P.; Wang, X.; Bai, X.; and Liu, W. 2017. Multiple Instance Detection Network with Online Instance Classifier Refinement. In *CVPR*.

Wan, F.; Wei, P.; Han, Z.; Jiao, J.; and Ye, Q. 2019. Min-Entropy Latent Model for Weakly Supervised Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(10): 2395–2409.

Wang, X.; You, S.; Li, X.; and Ma, H. 2018. Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features. In *CVPR*.

Wei, Y.; Liang, X.; Chen, Y.; Shen, X.; Cheng, M.; Feng, J.; Zhao, Y.; and Yan, S. 2017. STC: A Simple to Complex Framework for Weakly-Supervised Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(11): 2314–2320.

Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; and Huang, T. S. 2018. Revisiting Dilated Convolution: A Simple Approach for Weakly- and Semi-Supervised Semantic Segmentation. In *CVPR*.

Wu, Y.; Zhang, G.; Gao, Y.; Deng, X.; Gong, K.; Liang, X.; and Lin, L. 2020. Bidirectional Graph Reasoning Network for Panoptic Segmentation. *CoRR* abs/2004.06272.

Xiong, Y.; Liao, R.; Zhao, H.; Hu, R.; Bai, M.; Yumer, E.; and Urtasun, R. 2019. UPSNet: A Unified Panoptic Segmentation Network. In *CVPR*.

Yan, Z.; Liang, J.; Pan, W.; Li, J.; and Zhang, C. 2017. Weakly- and Semi-Supervised Object Detection with Expectation-Maximization Algorithm. *CoRR* abs/1702.08740.

Yang, T.; Collins, M. D.; Zhu, Y.; Hwang, J.; Liu, T.; Zhang, X.; Sze, V.; Papandreou, G.; and Chen, L. 2019a. DeeperLab: Single-Shot Image Parser. *CoRR* abs/1902.05093.

Yang, Y.; Li, H.; Li, X.; Zhao, Q.; Wu, J.; and Lin, Z. 2019b. SOGNet: Scene Overlap Graph Network for Panoptic Segmentation. *CoRR* abs/1911.07527.

Zhang, B.; Xiao, J.; Wei, Y.; Sun, M.; and Huang, K. 2019. Reliability Does Matter: An End-to-End Weakly Supervised Semantic Segmentation Approach. *CoRR* abs/1911.08039.

Zhang, X.; Feng, J.; Xiong, H.; and Tian, Q. 2018a. Zigzag Learning for Weakly Supervised Object Detection. In *CVPR*.

Zhang, Y.; Bai, Y.; Ding, M.; Li, Y.; and Ghanem, B. 2018b. W2F: A Weakly-Supervised to Fully-Supervised Framework for Object Detection. In *CVPR*.

Zhang, Y.; Ding, M.; Bai, Y.; Xu, M.; and Ghanem, B. 2020. Beyond Weakly Supervised: Pseudo Ground Truths Mining for Missing Bounding-Boxes Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology* 30(4): 983–997.

Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. In *CVPR*.