Noisy Boundaries: Lemon or Lemonade for Semi-supervised Instance Segmentation?

Zhenyu Wang Yali Li [*] Shengjin Wang
Beijing National Research Center for Information Science and Technology (BNRist)
Department of Electronic Engineering, Tsinghua University
wangzy20@mails.tsinghua.edu.cn, {liyali13, wgsgj}@tsinghua.edu.cn

Abstract

Current instance segmentation methods rely heavily on pixel-level annotated images. The huge cost to obtain such fully-annotated images restricts the dataset scale and limits the performance. In this paper, we formally address semi-supervised instance segmentation, where unlabeled images are employed to boost the performance. We construct a framework for semi-supervised instance segmentation by assigning pixel-level pseudo labels. Under this framework, we point out that ***noisy boundaries*** *associated with pseudo labels are double-edged. We propose to exploit and resist them in a unified manner simultaneously: 1) To combat the negative effects of noisy boundaries, we propose a noise-tolerant mask head by leveraging low-resolution features. 2) To enhance the positive impacts, we introduce a boundary-preserving map for learning detailed information within boundary-relevant regions. We evaluate our approach by extensive experiments. It behaves extraordinarily, outperforming the supervised baseline by a large margin, more than 6% on Cityscapes, 7% on COCO and 4.5% on BDD100k. On Cityscapes, our method achieves comparable performance by utilizing only 30% labeled images.*

1. Introduction

"When life gives you lemons, make lemonade."

– Elbert Hubbard

The performance of instance segmentation has been improved significantly with the development of deep learning [19, 24, 5, 50, 52]. However, current instance segmentation methods require pixel-level labeled images for fully-supervised training, which are prohibitively expensive to annotate. Statistically, segmenting one object instance requires 79s on average [36]. In some cases, annotating a single image with high quality even costs more than 1.5h [16]. This severely restricts the scale of datasets and further limits the performance of models. Researches in cognition science [18, 31] have demonstrated that human concept learning involves large amounts of unlabeled experience without feedback. Works in object detection [47, 25, 38] or semantic segmentation [41, 39, 21] have seeked for semi-supervised learning to alleviate the huge expense of human labeling. However, utilizing nearly labor-free unlabeled images is still unexplored for instance segmentation, partially because of its intrinsic difficulty. These motivate us to use unlabeled images to break through the upper bound of fully-supervised instance segmentation. We call this task *semi-supervised instance segmentation*.

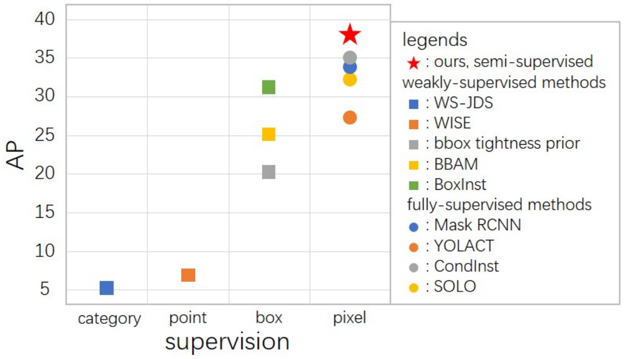


Figure 1: Semi-supervised instance segmentation (on the COCO dataset), which explores to utilize unlabeled images, is a novel problem that has not been formally defined and addressed so far. Compared with weakly-supervised and fully-supervised methods, it excavates existing data sufficiently and seeks to use a large number of unlabeled resources, making instance segmentation more practical.

The difficulty to collect pixel-level annotated data in instance segmentation has been recognized by many works. Most of them attempt to deal with this problem by weakly-supervised instance segmentation [22, 51, 45]. The main benefit of semi-supervised instance segmentation, compared to fully-supervised and weakly-supervised ones, is

[*] Corresponding author
Codes are available at https://github.com/zhenyuw16/noisyboundaries.

that it exploits existing resources sufficiently and allows to pursue larger-scale learning. Pixel-level annotated images have been provided in several existing datasets [36, 16, 56]. Semi-supervised instance segmentation can utilize these data, which are necessary for high-quality segmentation masks. Unlabeled images are enormous, and obtaining them is easy. As a result, the scale of learning is not restricted by datasets and can be as large as possible. This endows semi-supervised instance segmentation the potential to achieve better performance continuously.

Stimulated by the importance of semi-supervised instance segmentation, it is natural to ask: *what's the core issue of semi-supervised instance segmentation?* The core is to excavate information within unlabeled data. Progress in fully-supervised or weakly-supervised methods cannot be applied to the semi-supervised task, as supervision clues are necessary for them. To tackle this issue, we adopt the idea of pseudo labels and propose a semi-supervised instance segmentation framework. With this framework, the noise, especially included in masks from pseudo labels, is essential for the effective exploitation of unlabeled images. Considering that a high proportion of pixel-level noise lies in boundary regions, we focus on noisy boundaries. They provide incorrect supervision signals, but also contain many details that contribute to the model performance. This contradiction makes noisy boundaries a challenging problem.

In a word, noisy boundaries are double-edged (both “lemon” and “lemonade”), including useful and harmful information together. *How to learn from noisy boundaries for semi-supervised instance segmentation?* We need to exploit and combat them jointly. Specifically, we propose a noise-tolerant mask head (NTM) and a boundary-preserving map (BPM). Our NTM introduces a mask prediction branch for low-resolution segmentation output. With a low-resolution ground-truth for supervision, the details from boundaries are eliminated, where most of the noise exists. This contributes to noise-resistant learning. Meanwhile, our proposed BPM facilitates boundary learning. Different from previous approaches which preserve boundaries at the cost of enlarging pixel-level noise, our BPM strongly corresponds with the boundary regions but is irrelevant to noise. This leads to more precise results. With the help of our NTM and BPM, our method benefits from valuable features within noisy boundaries and discards the detrimental ones, thus mining unlabeled information more effectively.

Our main contributions can be summarized as follows:

- We formally address the semi-supervised instance segmentation task and construct a framework to exploit unlabeled data, which empowers us ability to break through the fully-supervised upper bound.
- We demonstrate the negative correlation between mask resolution and pixel-level noise, then propose a noise-tolerant head by interweaving low and high resolution features, which can resist noise in boundary regions.
- We propose a boundary-preserving map, which enriches boundary-relevant regions and suppresses narrow noise-excessive regions simultaneously. This produces more accurate segmentation boundaries.

Extensive experiments on Cityscapes [16], COCO [36] and BDD100K [56] demonstrate the effectiveness of our method. It obtains comparable results with only 30% amount of labeled images and surpasses its fully-supervised counterpart with only 40% labeled data. The performance is even better than approaches using human-annotated coarse labels or extra box-level annotations. We provide a simple and effective framework, which we believe will facilitate future research towards this direction.

2. Related Work

Instance Segmentation. Most of instance segmentation methods can be categorized into detection-based methods. Mask RCNN [19] extends Faster RCNN [44] to instance segmentation by adding an FCN based mask prediction branch. PANet [37] introduces bottom-up path augmentation for better feature learning. Cascade Mask RCNN [7] extends Cascade RCNN [6] to instance segmentation. HTC [9] further interweaves feature learning and adopts semantic knowledge to facilitate instance segmentation learning. Following works [24, 28, 15, 58] continue to improve the performance of instance segmentation. Recently, one-stage methods [5, 50, 52, 8, 53] also develop rapidly and achieve satisfying results with faster speed. They aim to predict masks directly, instead of generating proposals first. However, all of these methods require pixel-level annotated images, which are expensive to obtain.

Instance Segmentation with Incomplete Supervision. Considering the difficulty of obtaining pixel-level annotated images, some recent works aim to use incomplete annotations for instance segmentation. Weakly-supervised methods perform instance segmentation using either box-level labels [22, 51, 33] or image-level tags [45, 17]. However, they do not utilize existing pixel-level annotations, thus hard to obtain satisfying results compared to fully-supervised ones. Partially supervised approaches [23, 29, 59] adopt the setting where a small number of categories are pixel-level annotated and others have only box-level annotations. They aim to utilize box labels to expand the number of categories. Different from them, our target is to improve the performance of fully-supervised networks by using extra unlabeled data.

Semi-supervised Learning. Training models with both labeled and unlabeled images as semi-supervised learning has been widely used in image classification to boost the performance of fully-supervised learning. The prevailing methods include consistency regularization [30, 40, 49],

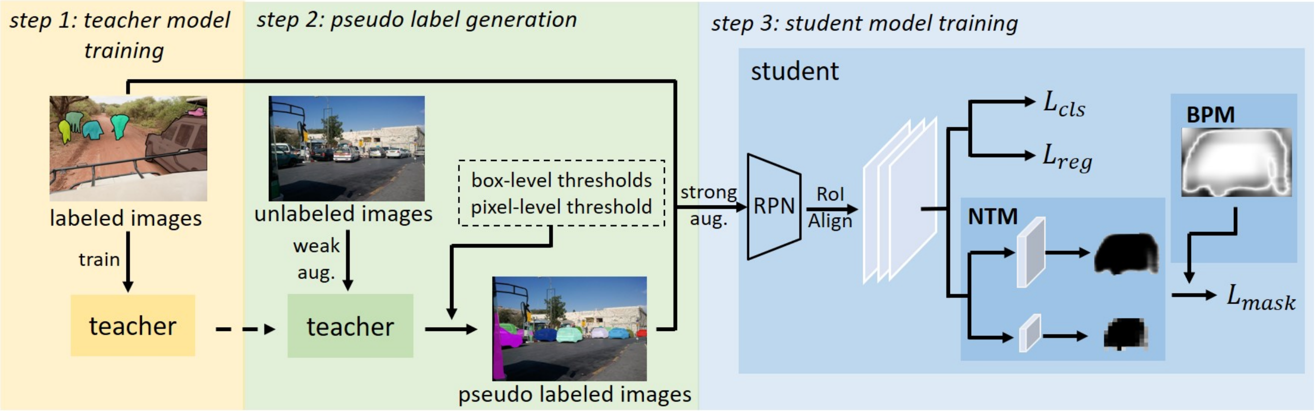


Figure 2: **Framework for semi-supervised instance segmentation.** A teacher model is trained with labeled images, then extracts pseudo labels for unlabeled images. After data aug., these images serve for training the student model. Our noise-tolerant mask head (NTM head) and boundary-preserving map (BPM) helps the student better learn from noisy boundaries.

pseudo labeling [32, 4, 3, 46], data augmentation [55, 4, 46], or label propagation [60, 2]. Recent works have extended semi-supervised learning to object detection and semantic segmentation. For example, [25, 41, 26] adopts the idea of consistency regularization and [39, 38, 48, 54, 21] utilizes pseudo labels. Recently, self-supervised learning [11, 12, 13] also utilizes unlabeled images. The difference is that self-supervised learning trains pretext tasks and is agnostic from downstream tasks, while semi-supervised learning targets at the specific task. In this work, we adopt pseudo labels to solve semi-supervised learning for instance segmentation, a naturally more difficult task.

3. Method

3.1. Semi-supervised Instance Segmentation

Our goal is to address the semi-supervised instance segmentation task. Specifically, we have a set of pixel-labeled images and aim to utilize easily obtained unlabeled data to boost the performance of instance segmentation.

Our basic framework consists of three steps:

Step 1: Teacher Model Training. We first train a teacher model with only labeled data as the common supervised learning. The teacher model will be applied to generate pixel-level pseudo labels for training the student model in the later steps. We choose Mask RCNN [19] as our teacher model but do not restrict to it.

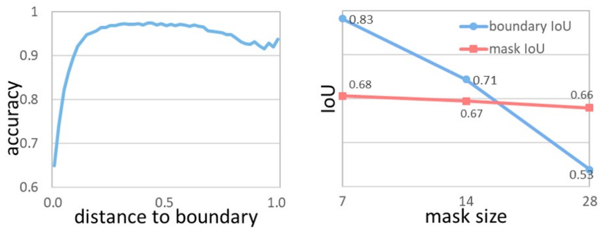
Step 2: Pseudo Label Generation. With the pre-trained teacher model, we perform inference on unlabeled images to produce instance segmentation masks. Data augmentation as scaling and horizontal flipping is conducted to improve the mask quality and reduce the miscalibration of neural networks [1]. We refer to this as weak augmentation.

To acquire pseudo labels, the raw inference masks need to be processed by two kinds of thresholds: the box-level and the pixel-level ones. At the box level, a large number of bounding boxes are predicted to guarantee a high recall. We thus need to filter low-quality boxes with a confidence threshold. At the pixel level, instance segmentation methods usually calculate foreground probability with *sigmoid*. A probability threshold is required to separate foreground and background pixels for creating masks.

Existing methods usually set the thresholds in a straightforward way. The box-level threshold is usually fixed to 0.7 or 0.9 [47, 38, 48], and the pixel-level threshold is generally taken as 0.5. However, this setting way is inappropriate. For the detection branch, current models with *softmax* for category probability are prone to be biased and predict the dominant classes. For the mask branch, the imbalance between foreground and background pixels also affects the prediction. In such a situation, a single threshold is easy to amplify the imbalance problem in pseudo labels.

To tackle this issue, we set the thresholds to match the distribution between labeled and unlabeled images. At the box level, we follow [43] and apply a per-category threshold. For each category, the principle is to keep the average number of instances per image matched for the labeled and unlabeled dataset. Similarly, after filtering low-quality boxes, we set the pixel-level threshold to keep the ratio of foreground to background pixels equivalent. Since mask prediction is acted for RoIs, we only count pixels in the bounding boxes. Also, this threshold is class-agnostic since mask and class prediction is usually decoupled. Note that in the test phase, we still adopt the 0.5 threshold, as we cannot access the distribution of test datasets.

Step 3: Student Model Training. With thresholding at the box-level and pixel-level, we obtain pseudo labels with mask annotations (pseudo masks). They are treated as the ground-truth labels for training the student model. According to previous works on semi-supervised learning, the di-



(a) (b)

Figure 3: **Illustration of pixel-level noise in instance segmentation.** (a): the mean accuracy of pixels *v.s.* their relative distance to the boundary. (b): the mean IoU between pseudo mask labels and their ground-truth ones *v.s.* their sizes. Pixels that are closer to the boundary are more likely to be noisy, and reducing size can suppress noise.

versity of the student model is crucial [49] and data augmentation is important [46, 47, 38]. We thus conduct data augmentation for images when training the student model, mainly including color transformation and cutout. We call this strong augmentation, to differ from that in the pseudo label generation step. Note that we do not adopt any augmentation strategy in the test phase for a fair comparison.

3.2. Noise-tolerant Mask Head

The above framework enables us to train a semi-supervised instance segmentation model. However, noise inherently exists in pseudo masks, which impedes the performance. We need a noise-resistant learning to combat it.

When training the student model, each proposal generated by RPN [44] will be assigned a mask from pseudo labels. After RoI-Align and a mask head, the mask prediction is generated. The assigned mask supervises this learning process. In the pseudo mask circumstance, assigned labels are not always accurate. The incorrect labels mislead the learning and deteriorate the performance. We design a noise-tolerant mask head (NTM) to help our model better resist noise in pseudo labels.

To resist noise in the learning process, we need to investigate which pixels are more likely to be noisy. The answer is: pixels that are closer to the boundaries. Boundary-relevant regions are noisier because they usually correspond with the decision boundary, where category features are not salient. Also, they contain detailed information that is difficult to learn. To further verify this, we conduct an empirical study on the Cityscapes dataset [16] and plot it in Fig. 3a. The mean accuracy of pixels is high, more than 90%. However, for pixels that are extremely close to the boundaries, the mean accuracy is significantly lower. Therefore, to resist noise in pseudo masks, the key lies in boundary-related regions. Their details and features are only visible when the mask resolution is high enough. In Mask RCNN [19], the mask ground-truth is usually downsized to 28 × 28. If the size turns smaller and the resolution is lowered, the image details can be implicit, where noise mainly lies. This is demonstrated in numerical analysis from Fig. 3b. As the mask size decreases, the overall mask IoU between pseudo labels and their corresponding ground-truth labels increases a bit, and the boundary IoU [14] improves significantly. We conclude that downsizing masks benefits the quality of pseudo labels, especially for regions near boundaries.

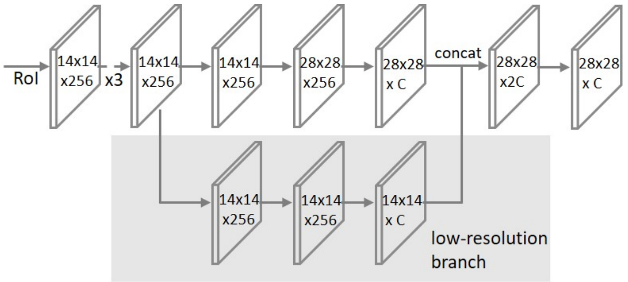


Figure 4: **The structure of our noise-tolerant mask head.** We add a branch for predicting results with low resolution. The low resolution mask better resists noise thus makes the network more noise-tolerant. Arrows, unless otherwise specified, denote conv or deconv layers. The conv kernel sizes are all the same as that in Mask RCNN. C denotes the number of categories.

Motivated by the above analysis, we propose our noise-tolerant mask head. We add a branch for the low-resolution mask prediction, and the structure is illustrated in Fig 4. This branch is supervised by a smaller size mask (we adopt 14 in practice). With a small size and low resolution, its features are cleaner and more noise-resistant. Consequently, this branch is able to utilize more accurate information, which contributes to learning in the semi-supervised setting. However, since the resolution is low, the predicted segmentation results are coarse and hard to reserve details. Therefore, the original high-resolution mask head is still retained. Specifically, the original high-resolution branch aims to learn fine-grained information, which is more likely to be affected by noise, while the low-resolution branch targets at learning coarse but clean information. Features from the low-resolution branch are fused into the high-resolution branch to pass clean messages. With this structure, we achieve more noise-tolerant learning. In the test phase, we only apply the high-resolution branch.

3.3. Boundary-preserving Map

With the noise-tolerant mask head, our model better resists noise from boundary-related regions. However, boundaries are also essential for instance segmentation, since detailed information within them is necessary for the quality of the predicted masks. In this subsection, we propose a

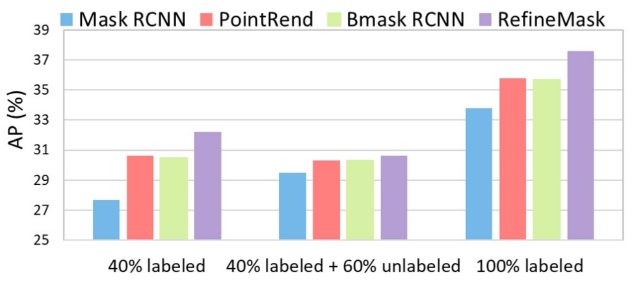
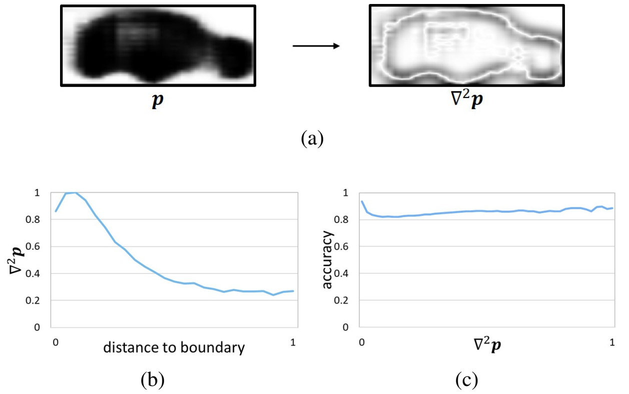


Figure 5: **Performance of existing boundary-preserving models on fully-supervised and semi-supervised tasks.** These methods improve the performance significantly for the fully-supervised task, but is limited in the semi-supervised setting because of the noise in boundary regions.

boundary-preserving map (BPM) to assist boundary learning in the semi-supervised task.

Facilitating boundary learning has been discussed in recent works such as PointRend [28], BMask RCNN [15] and RefineMask [58]. These methods are effective for fully-supervised learning but limited to the semi-supervised task. To corroborate this, we perform experiments on the Cityscapes dataset with 40% randomly selected images as labeled ones, and plot the acquired mask AP in Fig. 5. It is observed that these methods improve more than 2% compared to the Mask RCNN baseline in the fully-supervised task, but less than 1% for semi-supervised learning. The reason lies in noisy boundaries. In the semi-supervised task, traditional methods promote boundary learning at the cost of increasing the adverse effects of boundary-aware noise. Consequently, these methods are not suitable for semi-supervised segmentation.

In semi-supervised learning, the importance is thus to preserve boundaries but not to amplify noise. To promote boundary learning, the model should focus more on pixels that are closer to the boundaries. To reduce noise, pixels that are most likely to be noisy should be suppressed during training. Fig. 3a indicates that noise is excessive for those pixels whose distances to boundaries are extremely small. So we need to repress these pixels. Based on the above analysis, we present our boundary-preserving map. In BPM, the value of a pixel is negatively correlated with its distance to the boundaries. The only exception is pixels that are extremely close to boundaries, whose values should be small to suppress noise. Distance calculation for all pixels is effective but computationally complex, significantly decreasing the training speed. Denote the mask probability output by *sigmoid* function as $\boldsymbol{p} = [p_{ij}]$. We find that the laplace operation of the probability map, $\nabla^2 \boldsymbol{p}$, well meets the above requirement and is computationally efficient. As a result, we adopt $\nabla^2 \boldsymbol{p}$ as our BPM. We directly use our BPM to re-weight mask loss for different pixels, which is a simple but effective strategy.



(a)

(b)

(c)

Figure 6: **Illustration for our boundary-preserving re-weighting.** (a): illustrative examples, (b): $\nabla^2 \boldsymbol{p}$ *v.s.* the distance of the pixel to the boundary, (c): the mean of accuracy of pixels *v.s.* their values of $\nabla^2 \boldsymbol{p}$.

We show illustrative examples of our BPM in Fig. 6. For pixels that belong to boundary-relevant regions but do not lie in the narrow band along the boundaries, their values are the highest. These pixels usually contain detailed information and are relatively clean, thus should be paid attention to. Also, because of this design, our BPM is somewhat irrelevant to noise. With this property, our BPM benefits boundary learning and does not increase the effect of noise. This makes it appropriate for the semi-supervised task.

4. Experiments

We evaluate our proposed method on Cityscapes [16], MS COCO [36] and BDD100K [56]. Cityscapes provides 2,975 images for the training set. Besides, it consists of 20,000 images with coarse annotations. COCO includes 118,287 images. It also provides 123,403 unlabeled images. BDD100K is a diverse dataset about visual driving scenes. Only a subset of BDD100K is pixel-level annotated: about 7k images with mask annotations and 70k images with box annotations. Among them, 67k images have box-level annotations but no pixel-level labels. Our method is implemented with Pytorch [42] and MMDetection [10]. Unless otherwise specified, we use Mask RCNN [19] with ResNet50 [20] and FPN [34].

4.1. Experiments on Cityscapes

Experiments with a varying percentage of labeled images. We evaluate our method on the Cityscapes validation set. We randomly select a certain percentage of images from the training set as labeled images and treat the rest as unlabeled ones. Since semi-supervised instance segmentation is a new task, we extend methods about the two most relevant tasks - semi-supervised object detection and semantic segmentation for comparison. Results from Tab. 1 suggest

Table 1: **Experimental Results on Cityscapes with a varying percentage of labeled images.** † denotes adopting the same data augmentation in the semi-supervised training. § denotes using focal loss for the detection branch.

Method	5%	10%	20%	30%	40%
supervised	11.8	16.8	22.3	26.3	27.7
supervised †	11.3	16.4	22.6	26.6	28.3
semi-supervised object detection methods					
DD [43]	13.7	19.2	24.6	27.4	29.5
STAC [47]	11.9	18.2	22.9	29.0	29.8
CSD [25]	14.1	17.9	24.6	27.5	28.9
Ubteacher [38]	16.0	20.0	27.1	28.0	29.6
semi-supervised semantic segmentation methods					
CCT [41]	15.2	18.6	24.7	26.5	28.4
Dual-branch [39]	13.9	18.9	24.0	28.9	28.9
semi-supervised instance segmentation methods					
baseline	15.7	20.2	25.5	28.3	29.5
ours	**17.1**	**22.1**	**29.0**	**32.4**	**33.0**
ours §	**21.2**	**23.7**	**30.8**	**33.2**	**34.1**

that these methods are strong baselines. The pseudo label method without data augmentation, NTM and BPM, is our semi-supervised instance segmentation baseline.

The results are listed in Tab. 1. Our approach behaves consistently better under different degrees of supervised data. Compared to unbiased teacher [38], the state-of-the-art detector in semi-supervised object detection, our method outperforms it by a large margin under various settings. When the labeled ratio is 30% and 40%, the AP improvement reaches 4.4% and 3.4%. For CCT [41], one recent effective semi-supervised semantic segmentation method, our approach surpasses it by almost 6% in the 30% setting. Compared with the semi-supervised instance segmentation baseline, our method basically enhances the mask AP by 3%. When the labeled ratio is moderate, the increase is more: 3.5% and 4.1% for the 20% and 30% labeled ratio respectively. This demonstrates that our method learns better from noisy boundaries. Compared with the supervised counterpart, we achieve a more than 6% improvement. The importance of unlabeled images and the necessity of semi-supervised instance segmentation is validated.

Our method aims to learn from noisy boundaries for semi-supervised instance segmentation, thus targeting at the mask prediction branch. Focal loss [35] has been proved beneficial to semi-supervised object detection. We apply focal loss for the detection branch and the segmentation accuracy can be further improved. In particular, when the labeled percentage is 40%, the mask AP is 34.1%, which is higher than the fully-supervised method where all images are pixel-annotated (33.8%). When labeled images are 30%, the 33.2% AP is also comparable. This substantiates the great potential of semi-supervised learning.

Experiments with coarse-annotated images. We also conduct experiments with all images from the training set and utilize the extra coarse-annotated images as unlabeled ones. We design the following experiments for comparison. Coarse GT: directly using the given coarse annotations; coarse finetune: firstly training with coarse-annotated images, then finetuning with fine-annotated images; fine → coarse → fine: firstly training the model with fine-annotated images, then learning with coarse-annotated images, finally finetuning with fine-annotated images, just as [57].

Table 2: **Experimental Results on Cityscapes with coarse-annotated images.** † denotes adopting the same data augmentation in the semi-supervised training. § denotes using focal loss for the detection branch.

Method	AP	AP_{50}	AP_{75}
supervised	33.8	61.8	31.4
supervised †	34.7	61.8	33.7
coarse GT	23.3	49.4	18.3
coarse finetune	34.2	59.9	32.3
fine → coarse → fine	35.8	62.9	35.3
ours	**39.3**	**65.6**	**38.9**
ours §	**41.1**	**68.2**	**42.1**

From Tab. 2, we notice that with the 20,000 images, our method achieves the 39.3% AP, which exceeds supervised learning by 5.5%. This demonstrates that our semi-supervised method helps the model get rid of the limitation of the dataset scale. Our method behaves better than designed approaches using human-labeled coarse annotations - more than 3.5% higher than them. **Utilizing unlabeled images obtain even better performance than using human-labeled images!** This proves the high effectiveness of our method to exploit unlabeled information. With focal loss, we obtain a 41.1% AP. This remarkable performance corroborates the capability of semi-supervised instance segmentation for practical application.

4.2. Ablation Study

We perform ablation study on Cityscapes using 30% percent of images as labeled ones. The results are in Tab. 3. We adopt the general mask AP and the boundary AP [14] to evaluate the quality of masks and boundaries separately.

Data augmentation. We first evaluate the effect of data augmentation in the student model training step. Data augmentation increases the diversity of input samples, hence helping improve the performance. However, it is limited in fully-supervised learning, only bringing a 0.3% improvement. Even when images are all labeled, the AP increase is still less than 1%. In comparison, for semi-supervised learning, data augmentation increases the segmentation AP from 28.3% to 30.2%. The improvement is more significant, almost 2%. This corresponds with the conclusion in previous works [46, 47, 38] that data augmentation is crucial for the student model in semi-supervised learning.

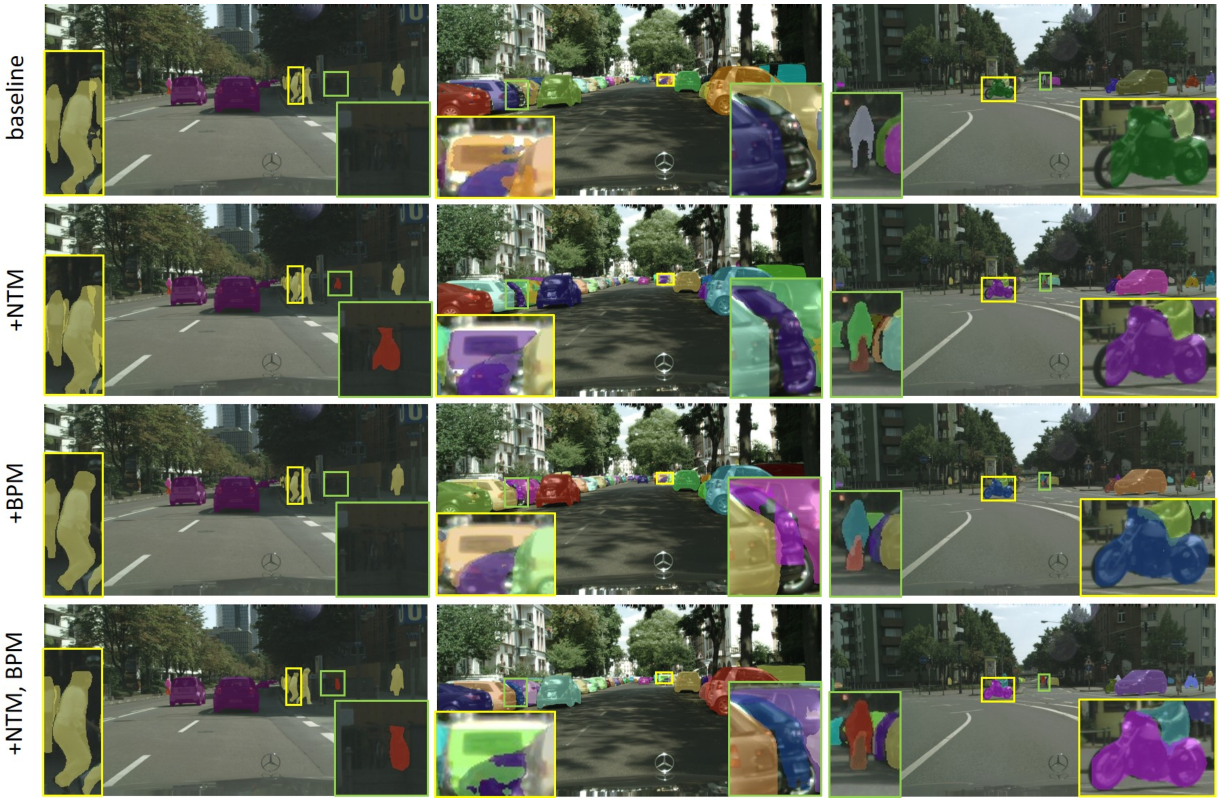


Figure 7: **Illustrative results on Cityscapes to show the effectiveness of our NTM and BPM.** NTM helps more correct detected instances (zoomed in green boxes) and BPM helps more precise boundary (zoomed in yellow boxes).

Table 3: **Ablation study on Cityscapes.** DA: data augmentation,, NTM: noise-tolerant mask head, BPM: boundary-preserving map. We evaluate the mask AP and boundary AP, abbreviated as AP_{bd}.

annotations	DA	NTM	BPM	AP	AP_{bd}
30% labeled				26.3	8.2
	✓			26.6	7.8
30% labeled 70% unlabeled				28.3	10.0
	✓			30.2	10.4
	✓	✓		31.1	10.9
	✓		✓	31.0	11.6
	✓	✓	✓	32.4	11.6
100% labeled				33.8	12.7
	✓			34.7	12.9

Noise-tolerant mask head. Results in Tab. 3 show that our NTM helps improve the mask AP by 0.9%, while the boundary AP improvement is not salient. This indicates that NTM helps semi-supervised learning mainly because it benefits the overall segmentation performance, like more correct detected instances or the holistic masks. Since noise in pseudo labels misleads the network learning, the overall discrimination ability of the network is hurt. Our NTM alleviates this problem, thus contributing to the mask AP. Illustrative results in Fig. 7 also confirm our analysis. In the first and the third images, the missed bicycles are segmented because of the NTM. In the second image, the middle car is detected. The visualized results correspond with the numerical results and our analysis above.

Boundary-preserving map. From Tab. 3, we observe that the mask AP is boosted by 0.8% with the help of our BPM. Different from NTM, BPM also improves the boundary AP significantly, from 10.9% to 11.6%. This indicates that BPM helps the performance mainly because it assists in the quality of boundaries. Such fact is strongly related to the function of BPM: it helps the model focus on boundary regions and learn more detailed information. This is also confirmed by illustrative results in Fig. 7. In the first image, with BPM, the contour of the person, especially at the head part, is more realistic. The same thing also occurs at the top part of the car in the second image and the front wheel of the motor in the third image. This demonstrates the effectiveness of our BPM to boundary learning.

4.3. Experiments on COCO

We also perform experiments on the challenging COCO dataset. Similarly, we randomly select a certain ratio of images from the COCO 2017 training set as labeled data, and

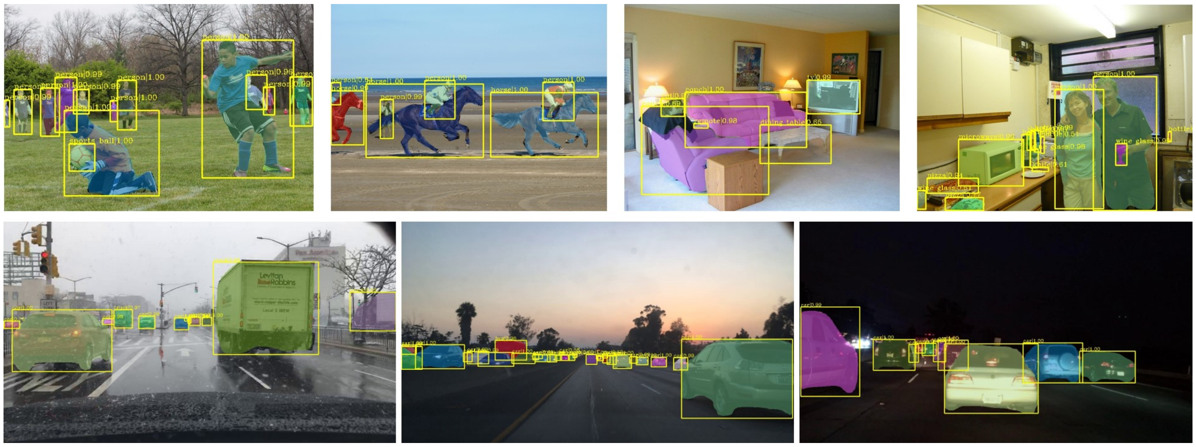


Figure 8: **Instance segmentation results of our method** on COCO (the first row) and BDD100K (the second row).

Table 4: **Results on COCO with a varying percentage of labeled images.** † denotes data augmentation. We use COCO 120k unlabeled images for the 100% experiment.

Method	1%	2%	5%	10%	30%	100%
supervised	3.5	9.4	17.3	22.0	28.9	34.5
supervised †	3.5	9.5	17.4	21.9	29.0	37.1
DD [43]	3.8	11.8	20.4	24.2	30.5	35.7
ours	**7.7**	**16.3**	**24.9**	**29.2**	**32.8**	**38.6**

Table 5: **Experimental Results on BDD100K.**

annotations	Method	AP
7k w/ masks	Mask RCNN	21.6
7k w/ masks 67k w/ boxes	Mask RCNN	24.5
	Grabcut Mask RCNN [27]	21.0
	Progressive Mask RCNN [59]	24.8
	Mask RCNN w/ ShapeProp [59]	26.2
7k w/ masks 67k w/o labels	semi baseline	24.4
	ours	**26.3**

the rest of them are unlabeled ones. For the 30% setting, we simply use the 35k subset of the COCO 2014 validation set as labeled images. For the 100% setting, we use all images from the COCO 2017 training set as labeled ones and 120k COCO unlabeled images as unlabeled ones. We list the mask AP in Tab. 4. Our method continues to be better than the supervised baseline. When the labeled images are 5% and 10%, our semi-supervised learning boosts the supervised learning by more than 7%, which is quite prominent. For the 100% experiment, where all pixel-annotated images are adopted and utilized data are more, our method achieves a 38.6% AP. The above experiments verify the value of our semi-supervised instance segmentation.

4.4. Experiments on BDD100K

We further benchmark on the BDD100K dataset. We use the 7k images with mask annotations as labeled images and the 67k images with only box-level annotations as unlabeled ones. Their box annotations do not participate in training. We compare our results with methods in [59], where box annotations for the 67k images are utilized for the partially-supervised learning. As Tab. 5 shows, our method obtains a 26.3% AP, which outperforms its supervised baseline by 4.7%. Our method performs better than Mask RCNN w/ ShapeProp [59], where box annotations are utilized. This further indicates that our method utilizes the information within unlabeled images quite sufficiently, so that utilizing unlabeled images outperform previous approaches that adopt box-level annotated images.

5. Conclusion

Considering the huge expense of labeling mask annotations, we propose the semi-supervised instance segmentation task. It enables the model to fully excavate available information and explore more extensive resources. With pseudo labels, unlabeled images participate in training and help improve the performance. By further learning from noisy boundaries, we alleviate the negative effects brought by noisy pseudo labels and exploit more valuable information within boundary-relevant regions. The extraordinary performance on benchmark datasets demonstrates the great ability of our method. Semi-supervised instance segmentation is a challenging but interesting problem. We hope that our simple and effective framework will stimulate future research along this direction.

Acknowledgement

This work was supported by the state key development program in 14th Five-Year under Grant Nos.2021QY1702, 2021YFF0602103, 2021YFF0602102. We also thank for the research fund under Grant No. 2019GQG0001 from the Institute for Guo Qiang, Tsinghua University.

References

[1] Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *MIDL*, 2018. 3

[2] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. label propagation and quadratic criterion. 2006. 3

[3] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *ICLR*, 2020. 3

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 3

[5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *CVPR*, 2019. 1, 2

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 2

[7] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *TPAMI*, 2019. 2

[8] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *CVPR*, 2020. 2

[9] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 2

[10] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[13] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 3

[14] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *CVPR*, 2021. 4, 6

[15] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. In *ECCV*, 2020. 2, 5

[16] 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 *CVPR*, 2016. 1, 2, 4, 5

[17] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *ECCV*, 2018. 2

[18] Bryan R Gibson, Timothy T Rogers, and Xiaojin Zhu. Human semi-supervised learning. *Topics in cognitive science*, 2013. 1

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2, 3, 4, 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[21] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *ICCV*, 2021. 1, 3

[22] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. In *NeurIPS*, 2019. 1, 2

[23] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *CVPR*, 2018. 2

[24] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *CVPR*, 2019. 1, 2

[25] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 2019. 1, 3, 6

[26] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *ECCV*, 2020. 3

[27] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017. 8

[28] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, 2020. 2, 5

[29] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *CVPR*, 2019. 2

[30] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 2

[31] Alexander LaTourrette and Sandra R Waxman. A little labeling goes a long way: Semi-supervised learning in infancy. *Developmental science*, 2019. 1

[32] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, 2013. 3

[33] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *CVPR*, 2021. 2

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 5

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 6

[36] 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 *ECCV*, 2014. 1, 2, 5

[37] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018. 2

[38] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021. 1, 3, 4, 6

[39] Wenfeng Luo and Meng Yang. Semi-supervised semantic segmentation via strong-weak dual-branch network. In *ECCV*, 2020. 1, 3, 6

[40] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *TPAMI*, 2018. 2

[41] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 2020. 1, 3, 6

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[43] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, 2018. 3, 6, 8

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2, 4

[45] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *CVPR*, 2019. 1, 2

[46] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 3, 4, 6

[47] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1, 3, 4, 6

[48] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *CVPR*, 2021. 3

[49] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2, 4

[50] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, 2020. 1, 2

[51] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *CVPR*, 2021. 1, 2

[52] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *ECCV*, 2020. 1, 2

[53] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. In *NeurIPS*, 2020. 2

[54] Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, and Shengjin Wang. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In *CVPR*, 2021. 3

[55] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. In *NeurIPS*, 2020. 3

[56] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 2, 5

[57] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 6

[58] Gang Zhang, Xin Lu, Jingru Tan, Jianmin Li, Zhaoxiang Zhang, Quanquan Li, and Xiaolin Hu. Refinemask: Towards high-quality instance segmentation with fine-grained features. In *CVPR*, 2021. 2, 5

[59] Yanzhao Zhou, Xin Wang, Jianbin Jiao, Trevor Darrell, and Fisher Yu. Learning saliency propagation for semi-supervised instance segmentation. In *CVPR*, 2020. 2, 8

[60] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002. 3