CPRAL: Collaborative Panoptic-Regional Active Learning for Semantic Segmentation

Yu Qiao[1]*, Jincheng Zhu[1]*, Chengjiang Long[2]†
Zeyao Zhang[1], Yuxin Wang[1], Zhenjun Du[3], Xin Yang[1]†

[1]Dalian University of Technology
[2]JD Finance America Corporation
[3]SIASUN Robot & Automation CO.,Ltd
{qiaoyu2020,zjccool951111,zhangzey}@mail.dlut.edu.cn,
cjfykx@gmail.com, {wyx,xinyang}@dlut.edu.cn, duzhenjun@siasun.com

Abstract

Acquiring the most representative examples via active learning (*AL*) can benefit many data-dependent computer vision tasks by minimizing efforts of image-level or pixel-wise annotations. In this paper, we propose a novel Collaborative Panoptic-Regional Active Learning framework (*CPRAL*) to address the semantic segmentation task. For a small batch of images initially sampled with pixel-wise annotations, we employ panoptic information to initially select unlabeled samples. Considering the class imbalance in the segmentation dataset, we import a Regional Gaussian Attention module (*RGA*) to achieve semantics-biased selection. The subset is highlighted by vote entropy and then attended by Gaussian kernels to maximize the biased regions. We also propose a Contextual Labels Extension (*CLE*) to boost regional annotations with contextual attention guidance. With the collaboration of semantics-agnostic panoptic matching and region-biased selection and extension, our *CPRAL* can strike a balance between labeling efforts and performance and compromise the semantics distribution. We perform extensive experiments on Cityscapes and BDD10K datasets and show that *CPRAL* outperforms the cutting-edge methods with impressive results and less labeling proportion.

Introduction

Active learning frameworks (Cohn, Ghahramani, and Jordan 1996) resort to well-designed acquiring functions to gradually capture representative samples from the dataset, and the generalized model can benefit from the final low-cost annotations with comparable performance. Many active learning algorithms have been developed to mitigate the dependency of deep-learning-based models on the finely annotated dataset. Although active learning has contributed greatly to image classification (Beluch et al. 2018; Long, Hua, and Kapoor 2013; Hua et al. 2013; Li and Guo 2013; Long and Hua 2015; Long, Hua, and Kapoor 2016; Hua et al. 2018; Sinha, Ebrahimi, and Darrell 2019; Choi et al. 2021; Caramalau, Bhattarai, and Kim 2021), there is a further margin for exploring incremental annotations, especially for pixel-wise computer vision tasks, like semantic segmentation.

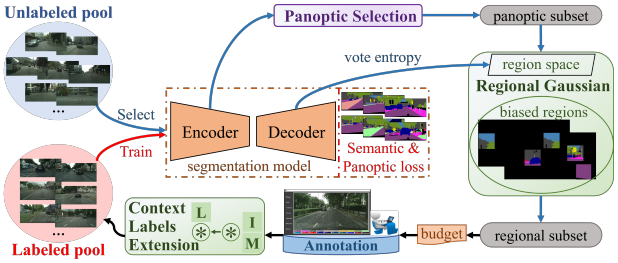


Figure 1: Red streams-model training, blue streams-samples selection, black streams-sampled images annotation.

Some active learning algorithms have been developed to alleviate the data dependency in semantic segmentation (Górriz et al. 2017; Yang et al. 2017; Mackowiak et al. 2018; Cai et al. 2021). According to the different granularity of sample annotations, active learning-based semantic segmentation can be divided into panoptic labels guided methods (Dai et al. 2020; Yoo and Kweon 2019; Sinha, Ebrahimi, and Darrell 2019; Kim et al. 2021) and regional information supervised ones (Mackowiak et al. 2018; Casanova et al. 2019; Colling et al. 2021). The former provides the whole images for oracles, and each annotation increment is based on the image size, while regional annotations consider region-based label boosting. Flexible shape or size variety can contribute to the region-based selection. Thus the performance of regional selections usually has an advantage over image-based annotations. However, more regional selections require enormous acquisition execution and slow down the annotation time for the entire active learning framework.

To design an active learning model that can handle large-scale semantic segmentation datasets with compatible accuracy and sampling cost, we observe that an initial panoptic selection is feasible to narrow the acquisition to relatively few regions. A distillation module after the panoptic selec-

*These authors contributed equally.
†This work was co-supervised by Chengjiang Long and Xin Yang. Xin Yang is the corresponding author.

tion is also necessary to select representative regions from the narrowed subset. Besides, there are essential contextual associations between regions, and existing region-based selection methods ignore the similarity of context-related image areas. If some appropriate guidance is investigated, regional annotations can be extended to related contexts.

In this paper, we propose a collaborative panoptic-regional active learning framework (*CPRAL*) to strike a balance between labor efforts and prediction performance. With panoptic information as initial selection, some images can represent the general distribution of the dataset. Then Regional Gaussian Attention (*RGA*) can consolidate discrete pixels and decide semantics-biased regions, which can alleviate the class imbalance between different semantic distributions. The final queried selection is provided for annotators to label different labels, and then we move them from unlabeled pool to labeled set. Considering the relevance of regional space, we import the Contextual Labels Extension module (*CLE*) to extend region-based annotations. Such a cycle can populate the number of labeled examples and improve the performance of the segmentation network. We repeat the cycle until the budget is exceeded (Fig. 1).

The contributions of this paper are three-folds:

- We propose a Collaborative Panoptic-Regional Active Learning framework to achieve partial-annotated semantic segmentation. Panoptic information can select the initial subset with limited budgets, and regional acquisition can decide representative semantic-biased regions.
- Considering the semantics-agnostic essence of data selection, we employ the Regional Gaussian Attention (*RGA*) to mitigate the class imbalance of sample distribution. We also propose a Contextual Labels Extension module (*CLE*) to boost regional annotations to related context, further enlarging the labeling proportion.
- We perform extensive experiments to demonstrate the performance of *CPRAL*, and our model outperforms state-of-the-art methods on Cityscapes and BDD100K. We also design an interactive GUI tool to support pixel-wise semantic annotations and verify our CPRAL.

Related Work

Active Learning (AL). Except for some synthesized-based query methods (Mahapatra et al. 2018; Mayer and Timofte 2020), most AL researches focus on selecting informative samples from the unlabeled pool, including three major acquisition functions: uncertainty, representation and their integration. Uncertainty-based methods (Ebrahimi et al. 2019; Kapoor et al. 2007; Wang and Ye 2015) explore Gaussian, entropy or decision to estimate uncertainty and are always struggled in the dataset scale. Monte Carlo Dropout (MC Dropout) architecture is introduced in (Gal and Ghahramani 2016) with a Bayesian approximation. Many follow-up approaches (Gal, Islam, and Ghahramani 2017; Kirsch, Van Amersfoort, and Gal 2019) incorporate MC Dropout into their algorithms to refine active learning. (Kuo et al. 2018; Beluch et al. 2018) employ ensembles to regress uncertainty attributes, which may influence the class diversity (Melville and Mooney 2004). The representation-based methods (Sener and Savarese 2018; Jain and Grauman 2016) consider categories for selection, and the computations may explode as the number of classes increases.

Graph Convolutional Network (GCN) for active learning is first imported in (Kipf and Welling 2017), and later (Wu et al. 2019; Caramalau, Bhattarai, and Kim 2021) also encode features as graph nodes to bridge correlations. Recent variational autoencoders (VAE) attract a lot of attention in active learning. The learned features in VAE-based methods (Sinha, Ebrahimi, and Darrell 2019; Choi et al. 2021; Kim et al. 2021; Zhang et al. 2020) can describe the uncertainty and representation simultaneously. They can all predict competent results with additional training efforts and time as sacrifices. There are also AL researches focusing on other computer vision tasks, like object detection (Yuan et al. 2021; Aghdam et al. 2019), person re-Identification (Liu et al. 2019), image matting (Yang et al. 2018, 2020) and 3D segmentation (Siddiqui, Valentin, and Nießner 2020).

Semantic Segmentation. Many of the methods derived from (Lin et al. 2017; Chen et al. 2017) are used for semantic segmentation, and high-accuracy annotations are required by them to provide essential supervision. Many other computer vision tasks (Mei et al. 2020, 2021; Liu et al. 2021; Qiao et al. 2020) also require high-precision annotation. However, pixel-wise adaptions are expensive and intractable, especially for some areas of expertise, like medical segmentation. Some researchers (Ahn and Kwak 2018; Lee et al. 2019) exploit weakly supervised solutions to mitigate the dependency of dense annotations, which always require additional labels and are vulnerable to the scale of the dataset. There are also many approaches developed based on active learning to release the cost of labeling efforts. AL-based methods can effectively select representative samples and alleviate the disturbance of redundant images.

Semantic Segmentation and Active Learning. Active learning-based methods mainly fall into two major categories, panoptic-guided and region-supervised methods, according to different annotation types. The former queries the next batch in the images unit (Dai et al. 2020; Sinha, Ebrahimi, and Darrell 2019; Yang et al. 2017). Although the selection of images is heuristic and firsthand, they will sample many redundant pixels. Region-supervised methods (Casanova et al. 2019; Colling et al. 2021; Mackowiak et al. 2018; Cai et al. 2021) annotate regions for model training, which can provide more effective labels at the expense of regional execution time. These region-supervised approaches are implemented without consideration of the class imbalance in local patches. Besides, the context correlations are also ignored in the regional annotations.

In this paper, we incorporate panoptic selection and regional annotations as an integration. Panoptic sampling can decide some representative images and regional sampling can eliminate redundant pixels from the panoptic subset. Considering labeling efforts and semantics completeness (Mackowiak et al. 2018), we employ regular regions as samples for annotations instead of superpixels. Besides, regional Gaussian attention is introduced to mitigate the class imbalance in local patches, and contextual labels extension is proposed to generalize annotations in adjacent areas.

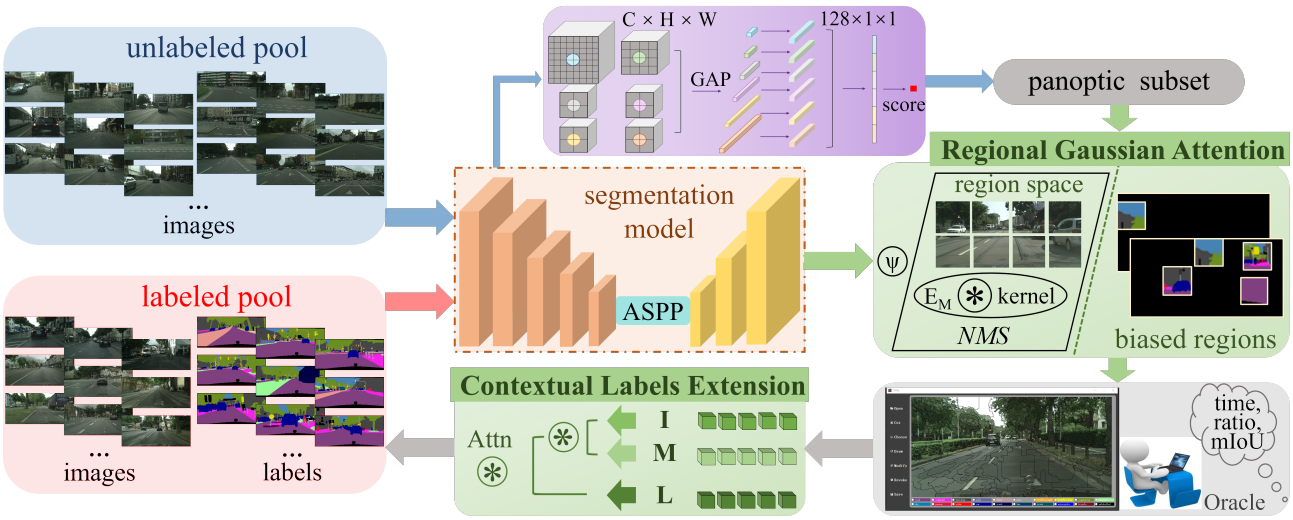


Figure 2: The diagram outlines our *CPRAL* pipeline. Red streams-training, blue and green streams-subset acquiring, gray streams-annotating. The initial labeled set can roughly train the segmentation network, and the next batch of images is predicted from the unlabeled pool according to the panoptic and regional selection.

Methodology

Panoptic and Regional Motivation Statement

Let define some descriptors first, (x, y)-images and ground truths in the labeled pool, $(\tilde{x}, \tilde{y})$-unlabeled pairs, corresponding to the labeled pool $\mathcal{D}$ and unlabeled pool $\tilde{\mathcal{D}}$. At each iteration t, the acquisition function $\mathcal{R}$ can collect a potential subset $\mathcal{C}$ and then move the samples from $\mathcal{D}$ to $\tilde{\mathcal{D}}$. Then according to KL divergence, we can formulate the active learning optimization during samples selection (Gudovskiy et al. 2020):

$$\mathcal{R}_{opt}(t, \mathcal{C}) = \underset{\mathcal{R}(t,\mathcal{C})}{\arg\min}\, D_{KL}(\mathcal{D}||\tilde{\mathcal{D}}), \tag{1}$$

If the distribution difference between $\mathcal{D}$ and $\tilde{\mathcal{D}}$ is smaller, the trained model on $\mathcal{D}$ can fully represent the complete training dataset and predict competent results on the validation set. Considering the parameters attributes of deep learning models and the optimization of the training procedure, Eqn. 1 is equivalent to optimize:

$$D_{KL}(\mathcal{D}||\tilde{\mathcal{D}}) \approx D_{KL}(P_{x,y}(\theta)||P_{\tilde{x},\tilde{y}}(\theta)), \tag{2}$$

where θ is the model parameters, $P(\cdot)$ describes the sample distribution on the learned model. The approximate distribution of $\mathcal{D}$ and $\tilde{\mathcal{D}}$ means given parameters θ they share comparable model performance. Vice versa, for a deep learning model, the close panoptic feature attributes can suggest approximate distribution for different data pools.

For the regional selection, due to the inevitable class imbalance of semantic segmentation (road, vegetation $vs.$ person, motorcycle), few classes may occupy most regions. Thus the acquisition of x^r, y^r should maintain the diversity to cover all potential semantics in $\tilde{\mathcal{D}}$ and validation set. Then the semantic segmentation can summarize different classes:

$$\mathcal{L}(y_i, \hat{y}_i) = \sum \mathcal{L}(y_i^r, \hat{y}_i^r), \tag{3}$$

here i means different classes.

Overall Pipeline

The pipeline of the proposed active learning framework is unfolded in (Fig. 2), and the initial $\mathcal{D}$ is fixed. The purple, green and gray stages correspond to the panoptic selection, regional selection and annotation phase. There are four phases in an iteration to finish the samples selection and annotation process. (1): Take unlabeled examples as input and extract multi-scale features to regress a matching rating. (2) Regional Gaussian Attention (*RGA*): Perform vote entropy, kernel filter, and non-maximum suppression (NMS) on panoptic samples to decide semantics-biased regions. (3) Oracle or our designed label tool can annotate the selected subset with high accuracy and then move them to the labeled pool. (4) Contextual Labels Extension (*CLE*): Take images, regional labels and masks as input, extract patches to boost annotations with contextual attention as guidance. We adopt MobileNet (Sandler et al. 2018) and DRN (Yu, Koltun, and Funkhouser 2017) as two different segmentation encoders. The former has impressive efficiency, while the latter shows high performance. The ASPP module and decoder phase are implemented referring DeepLab (Chen et al. 2017).

According to the above analysis, we first use panoptic information to compare the subset from $\tilde{\mathcal{D}}$ with initial $\mathcal{D}$. Here we modify the loss prediction in (Yoo and Kweon 2019) to achieve a matching rating. The features from 6 different encoder stages are involved and averagely pooled to $C \times 1 \times 1$ to regress a panoptic rating. The expanded features can better capture the distribution variation on the learned model, and for the images share the approximate distribution and similar panoptic information, $P_{x,y}(\theta)$ and $P_{\tilde{x},\tilde{y}}(\theta)$ will predict an equivalent rating. Therefore, we select the samples in reverse order to enrich the diversity and representation of $\mathcal{D}$.

After panoptic selection, we import Regional Gaussian Attention module (*RGA*) to decide semantics-biased regions. Although the panoptic information can capture abun-

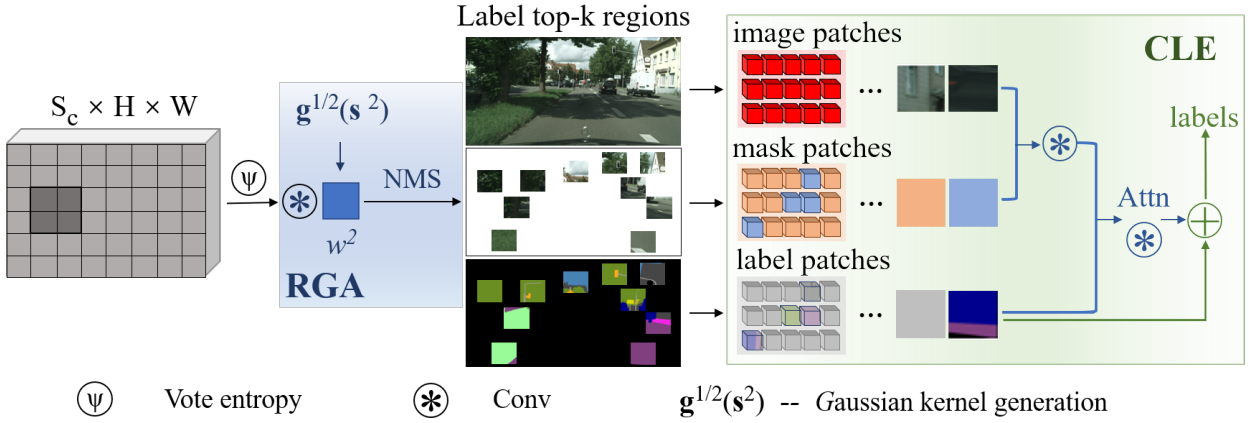


Figure 3: Selection of top-K biased regions and labels extension with contextual guidance. For a logical output with S_c classes, we can generate an entropy map. Then an s-steps cumulative Gaussian distribution can produce a w^2 kernel to perform a region-wise filter on the entropy map, and we select the top-K regions to provide labels. Finally, we split images, labels, and masks into patches and perform contextual extension with region annotations as attention guidance to boost labels.

Figure 4: The visualization of our label tool.

dant feature attributions, the final rating is essentially a semantics-agnostic regression, and the average pooling operation will ignore the non-dominant classes. Such class imbalance can result in the absence of critical information in semantic segmentation, like persons. The *RGA* module can pay attention to discrete semantics in local areas and increase the class variety. The final selected regions are displayed for oracle to provide pixel-wise annotations. We also design an interactive label tool for novice users to achieve high-quality annotations. Considering the context correlations between adjacent regions, contextual semantics can also influence the model training. Thus we propose a Contextual Labels Extension module (*CLE*) to extend regional annotations with contextual attention guidance.

Regional Gaussian Attention

To maintain the semantic attributes for all pixel locations, we take the panoptic subset as input pool and convert the output of segmentation model with $n\_classes$-channels into 1-channel entropy map with the same resolution (Mackowiak et al. 2018). The generation formula of entropy map is:

$$\mathcal{E}_M = \sum_{i \in C} -\frac{\sum \tilde{y}_i}{m} \cdot \log \frac{\sum \tilde{y}_i}{m}, \tag{4}$$

where i is different classes and m describes the number of model iterations and set as 20 in our framework. The negative accumulation of Eqn. 4 can select the semantics that occupies fewer areas. However, the dominant labels (road, vegetable) may dramatically influence the model convergence direction in the initial training phase and mislead the gradients for rare classes. Therefore we import Gaussian attention instead of a direct convolution layer to filter redundant semantics and distill representative ones:

$$\mathcal{K} = sqrt(\frac{\mathcal{N}^2(s,\sigma)}{\sum \mathcal{N}^2(s,\sigma)}) \cdot s, \tag{5}$$

σ is 2 in our experiments and s equals the region size-w. The entropy map can be refined by Gaussian kernel $\mathcal{K}$ by a convolution filter: $\mathcal{E}_M = \mathcal{E}_M \otimes \mathbf{K}$. Then we can vectorize the entropy map and decide the highest pixel index by non-maximum suppression (NMS):

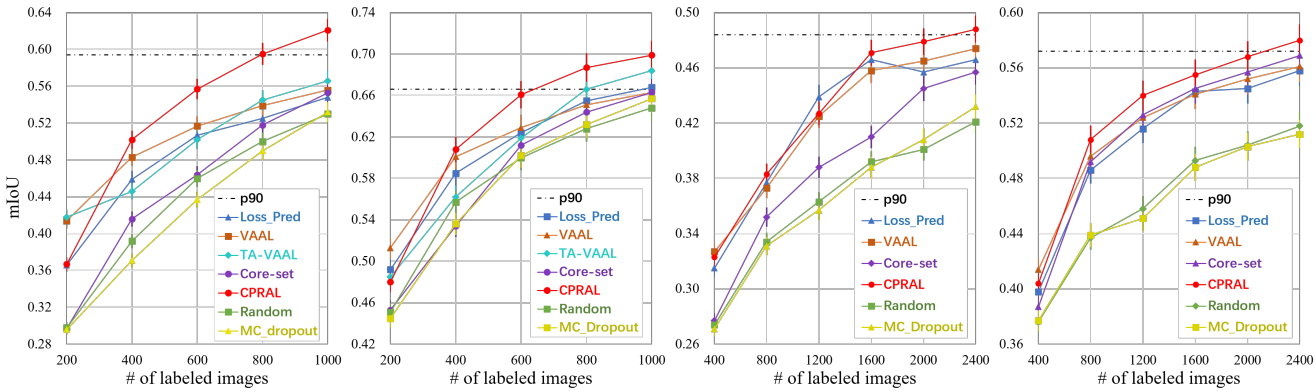
$$\mathcal{Z}_{high} = \arg\max_z Vector(\mathcal{E}_M), \tag{6}$$

With the highest entropy pixel index as center coordinates, we can crop a target region with size w^2 from unlabeled images. The final regions are suggested by entropy map and s^2 Gaussian kernels, both can highlight the semantics-biased areas. The biased regions will obtain their semantic annotations after oracle or our label tool.

Contextual Labels Extension

The contextual attention has been explored in (Yu et al. 2018) to integrate attributes from relative context. After oracle or tool labeling for each iteration, some regions can obtain exact annotations, while others in the panoptic subset remain unlabeled. Adjacent image patches have strong context relations and the labeled regions can provide critical references for unlabeled ones. Therefore, we separate the images in the panoptic subset into patches and take contextual attention as guidance to extend regional annotations. For the image $\tilde{I}$ and label $\tilde{L}$ in the panoptic subset, we can generate mask $\tilde{M}$ for labeled regions and divide them into patches. We can define regional attention score as follow:

$$\mathcal{S} = \sum Conv(I, \Omega_{I_k}) \odot \Omega_{M_k}, \tag{7}$$



(a) Cityscapes&MobileNet (b) Cityscapes&DRN (c) BDD100K&MobileNet (d) BDD100K&DRN

Figure 5: The quantitative performance compared with SOTA on Cityscapes (Cordts et al. 2016) and BDD100K (Yu et al. 2020) datasets with different backbones. "p90" represents 90% of the performance trained on the complete dataset. The TA-VAAL on the BDD100K is omitted here due to the convergence trouble in the implementation.

where k is patch index and Ω represents the patch set. Eqn. 7 can condense the patch correlations on images and mask them with labeled information. Then the $softmax$ operation can handle $\mathcal{S}$ as attention guidance. And the final labels information can be calculated from:

$$\mathcal{L}_{last} = \sum \mathcal{C}onv(\mathcal{S}, \Omega_{L_k}). \tag{8}$$

CLE integrates context relations to extend regional annotations, and the extension insights consider the distribution similarity of adjacent areas. With *CLE* as a cascaded procedure following the annotation phase, we can further enrich the labeled pixels and provide more supervised information for segmentation training.

Smart Segmentation Tool

For convenient annotation and practical application, we design an interactive segmentation tool based on the proposed *CPRAL*. Fig. 4 displays a visualized annotation process. We employ superpixels to label large image areas (road, vegetables) quickly. For biased regions selected by *CPRAL*, participants can draw lines to shape them. We exploit some raw datasets to validate the Smart Segmentation Tool and may provide large-scale annotations in the future.

Experiments

Here we demonstrate the performance on two public semantic segmentation datasets: Cityscapes (Cordts et al. 2016) and BDD100K (Yu et al. 2020). We use MobileNet (Sandler et al. 2018) and DRN (Yu, Koltun, and Funkhouser 2017) for feature extraction, respectively, one with efficiency advantage and the other with accuracy. We first compare *CPRAL* with the cutting-edge methods. Then we show the superiority for the cooperation of panoptic and regional information. We also analyze the robustness and ablation study.

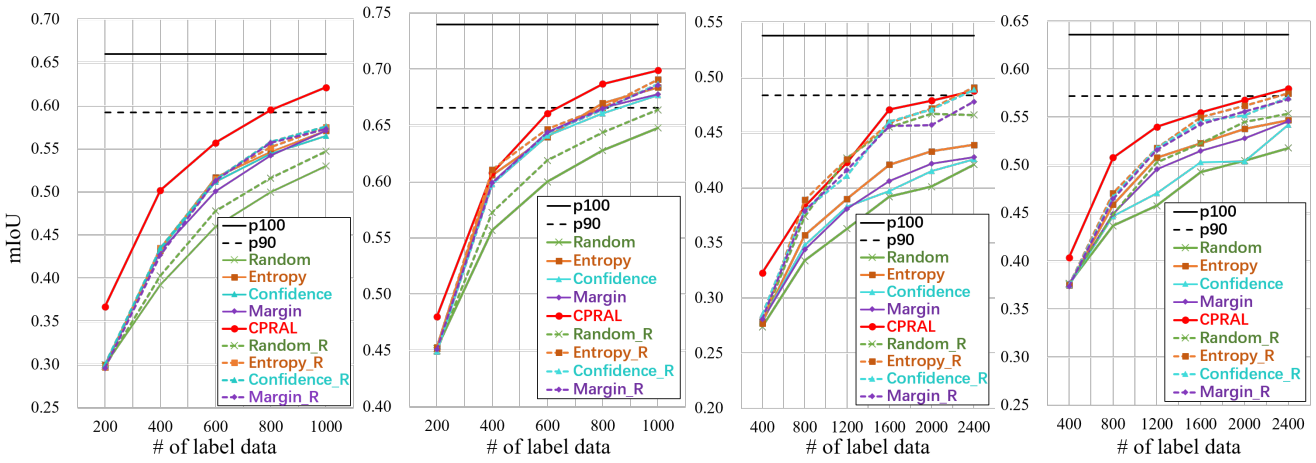
Experiments Datasets and Details

Datasets. Both Cityscapes and BDD100K are driving video datasets with 19 semantic classes, collected from Europe and the United States, respectively. There are 5000 examples in Cityscapes, 2975 for training, 500 for validation, and 1525 for testing. We use the training set to train different models and verify their performance on the validation set. Each image has a resolution of 1024×2048, and we reduce the size by half as input. BDD100K consists of 7000 training images, 1000 validation images, and 2000 testing images. We also train the models on the training set and verify them on the validation set. The images in the BDD100K are fed into the network at their original resolution of 720×1280.

Implementation details. We implement *CPRAL* using PyTorch and Tesla P100 graphics cards. Random horizontal flip and Gaussian blur are employed to augment the sampling diversity. The segmentation and loss prediction modules adopt the (SGD) optimizer with a momentum of 0.9 and a weight decay of 0.0005. For each sampling iteration, there are 50 epochs for training. The initial learning rate is 0.001 and drops to 0.0001 at epoch 35. The sampling will iterate five times on Cityscapes with initial 200 examples and six times on BDD100K with initial 400 examples. The panoptic subset size is 400 for Cityscapes and 800 for BDD100K, and the final selection is 200 and 400. The batch size for MobileNet and DRN are 4 and 2, separately. The loss function for the segmentation model is cross-entropy, and the panoptic loss is the same as (Yoo and Kweon 2019).

Comparisons with SOTA

We compare our model with five SOTA methods, TA-VAAL (Kim et al. 2021), VAAL (Sinha, Ebrahimi, and Darrell 2019), Loss Prediction (Yoo and Kweon 2019), Coreset (Sener and Savarese 2018), and MC Dropout (Gal and Ghahramani 2016). Random selection is also involved as the baseline. The quantitative comparison is demonstrated in Fig. 5. The four groups of results correspond to two different datasets and two backbone models. The evaluation metric is the Mean Intersection over Union (mIoU). The ninety percent of the performance trained on the full dataset is also shown (p90). The TA-VAAL has an unstable convergence on BDD100K, and we ignore their results.



(a) Cityscapes&MobieNet (b) Cityscapes&DRN (c) BDD100K&MobileNet (d) BDD100K&DRN

Figure 6: The comparison of panoptic sampling methods. "_R"-with our regional selection, "p100"-trained on the full dataset.

On the Cityscapes dataset, our *CPRAL* shows obvious superiority, especially as sampling increases after the initial iteration, which can prove the effectiveness of valid regional selection and contextual extension. Given the fixed sampling amount, the Gaussian filter can distill semantics-biased regions, balancing the selection for non-dominant classes and reducing the annotations for redundant pixels. Simultaneously, *CLE* can extend labels to context-related areas, enlarging the coverage of each annotated region. After the last iteration, *CPRAL* can achieve more than 90% performance with only 33.6% data proportion and has a clear margin over the TA-VAAL. Before the first annotation phase, the distribution of the selected samples is uniform for the segmentation model. However, the inductive prediction or VAE module in Loss_Pred, VAAL, and TA-VAAL can extract panoptic information and jointly optimize the segmentation module in a self-adapted manner. Thus their performance for the first iteration is superior to Core-set and MC Dropout. Our panoptic selection can also utilize the image-based features and benefit from the joint optimization. However, the adversarial training between VAE and discriminator in VAAL and TA-VAAL have better contributions at the expense of additional convergence consumption.

As for the BDD100K, most observations are similar. BDD100K has more diversity in the cities, weather, and driving image types. Nevertheless, *CPRAL* can also achieve better than 90% performance with only 34.3% annotated images. The promotion of *CPRAL* after the first iteration is prominent, except for the third phase with MobileNet, which may arise from the stability of the backbone. The VAAL and Loss_Pred also have better behaviors for the initial phase, proving the generalization of panoptic information.

For the cross-comparisons between Cityscapes and BDD100K, we can observe that *CPRAL* has better adaptation with different segmentation backbone models. If 50% of mIoU is the essential requirement for the segmentation annotation task, *CPRAL* demands approximately 13.4%, 7.9%, 38.5%, and 11% of labeled data for each of the combinations in Fig. 5. By contrast, the Core-set has impressive predictions with DRN as the backbone model while showing sub-optimal performance with the features from MobileNet. As the annotated images accumulate, the improvement trend of Loss_pred decreases dramatically. Panoptic features can quickly capture dominant semantics and lose the class balance control when acquiring many redundant pixels.

Performance of Regional Selection

There are many panoptic sampling methods to replace the image-based selection in our framework. Like the mentioned above, the loss prediction can benefit the panoptic selection for the initial model training. In this section, we combine the proposed regional selection with different panoptic sampling methods to demonstrate the accuracy promotion.

Entropy (Ozdemir et al. 2018), Confidence (Li and Sethi 2006), and Margin (Balcan, Broder, and Zhang 2007) are representative active learning sampling methods. We adopt them to replace the panoptic selection in our pipeline. The panoptic and regional sampling sizes are the same as *CPRAL*. The quantitative results are reported in Fig.6. We can observe that all methods have a noticeable improvement after embedding with regional selection. On the Cityscapes dataset, our extended loss prediction module can capture general semantics, contributing to the panoptic selection (8% and 2% performance promotion with MobileNet and DRN). On the BDD100K dataset, due to the data diversity and scene complexity, semantic-agnostic panoptic loss fails to summary the features of large-scale samples. The performance of Entropy, Confidence, and Margin are very close or even better than *CPRAL*, which proves the compatibility and efficacy of regional selection. By the way, our collaborative sampling can save much more time than region-only selection. *CPRAL* requires 8.28 (MobileNet) and 25.17 (DRN) minutes for each iteration on Cityscapes, while region-only selection takes about 23.2 and 131.57 minutes. The 64.3% and 80.8% time savings from the panoptic selection are significant for the active learning process.

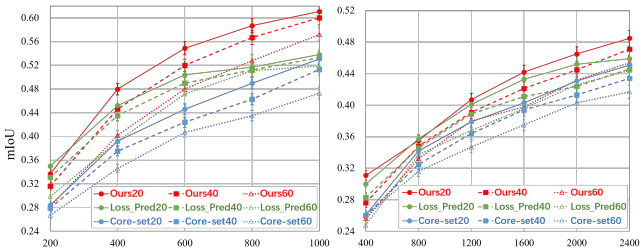


Figure 7: The performance of *CPRAL*, Loss_Pred, and Core-set under 3 proportions of noise levels (20%, 40%, 60%).

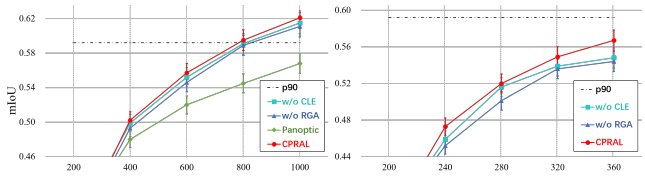


Figure 8: Ablation study on Cityscapes with MobileNet. Left-regional subset size 200, right-regional subset size 40.

Robustness Analysis

Noisy labels are a potential distraction for human-in-loop annotation procedures. Especially for novice users, label missing and label errors are prevalent in practical applications. Here we explore three noise levels according to the proportion of wrong labels (20%, 40%, and 60%) and assume three potential types of noisy labels to investigate the robustness of *CPRAL*. (1) $symmetry\_noise$: replace the ground truth with the subtraction of n-classes, $c = n - c$. (2) $missing\_noise$: replace the ground truth with ignoring index 255. (3) $asymmetric\_noise$: replace the ground truth with a random label. The ratio of the three types is $1:2:1$ for all noise levels. We compare *CPRAL* with Loss_Pred and Core-set on the Cityscapes and BDD100K datasets with MobileNet backbone. Fig. 7 shows how noisy label accumulation affects data sampling. The results reflect the error tolerance and robustness of the models. A model with better results has lower requirements for annotators.

At noise level-20, there is a slight degradation in the performance of *CPRAL*, and after five iterations, mIoU decreases by 1.6 and 0.6 percent. By contrast, Loss_Pred drops 2.7% and 1.5%, Core-set drops 4% and 1.3%. For the intractable noise-60 level, *CPRAL* can also achieve the mIoU of 0.57 and 0.45, and Core-set is greatly hampered under such widespread misleading. Furthermore, $missing\_noise$ is the most common situation in real-time annotations, and half of the noise labels here are from $missing\_noise$, which can demonstrate the competent robustness of *CPRAL*.

Ablation Study

Here we remove the contextual labels extension module (*CLE*) and regional Gaussian attention module (*RGA*) from *CPRAL* to make an ablation study. The results are displayed in Fig. 8. "Panoptic" means the panoptic subset is directly acquired for labeling. "w/o CLE" means removing *CLE* from *CPRAL*, and "w/o RGA" means removing *RGA* based on "w/o CLE". Intuitively, the performance of "Panoptic" drops a lot, suggesting the significance of region information. The Gaussian attention and contextual extension built on regional selection can also bring improvement for pixel sampling. *RGA* can distill class-biased regions and complement the semantic-agnostic panoptic selection, while *CLE* can connect relevant context to extend regional annotations. The visual ablation is shown in Fig. 9, and many tiny areas are refined (we crop the patch for better display).

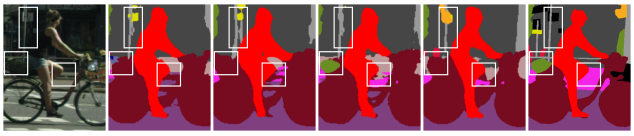


Figure 9: Visual comparison on Cityscapes&MobileNet. "Full" means the performance trained on the full dataset.

To further prove the effect of CLE and RGA, we also perform an ablation on with regional subset size 40 in Fig. 8 (right). For a fair comparison with most existing methods, *CPRAL* selects 200 images from 400 panoptic samples as a regional subset and achieves around 90% full-trained performance (Fig. 8, left). RGA and CLE are also calculated based on 400 panoptic samples. However, if we reduce the regional subset size to 40, their promotion is more obvious.

Discussions

The performance of *CPRAL* is contributed from panoptic selection, *RGA* and, *CLE*. Panoptic selection is essentially a regressed rating, ignoring the primary semantics of unlabeled images, and it will fade into mediocrity as data increment. *RGA* can capture discrete semantics, while complex scenes in BDD100K may produce unbridled pixels to discount attention filter. The biggest threat to *CLE* is the wrong label, which can affect adjacent regions after contextual extension.

Conclusion

This paper exploits the Collaborative Panoptic-Regional Active Learning (*CPRAL*) for semantic segmentation. Panoptic information is responsible for selecting fully supervised images and summarizing the unlabeled pool to a representative subset. Then Regional Gaussian Attention module (*RGA*) can decide semantics-biased areas to eliminate redundant pixels and acquire final queried regions. The Contextual Labels Extension module (*CLE*) can extend annotations to relevant context with attention guidance. The cooperation of panoptic and regional information can strike a balance between samples acquisition and model performance. Extensive experiments on Cityscapes and BDD100K datasets can demonstrate the annotation accuracy of *CPRAL*. In future work, we will combine the sequential relations and optical flow variation between frames to optimize the video segmentation in active learning. Some domain adaption-related researches may also be considered to transfer different representations between video clips.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61972067 U1908214, Natural Science Foundation of Liaoning under Grant 20180520032, and the Innovation Technology Funding of Dalian (2020JJ26GX036).

References

Aghdam, H. H.; Gonzalez-Garcia, A.; Weijer, J. v. d.; and López, A. M. 2019. Active learning for deep detection neural networks. In *IEEE ICCV*, 3672–3680.

Ahn, J.; and Kwak, S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *IEEE CVPR*, 4981–4990.

Balcan, M.-F.; Broder, A.; and Zhang, T. 2007. Margin based active learning. In *International Conference on Computational Learning Theory*, 35–50. Springer.

Beluch, W. H.; Genewein, T.; Nürnberger, A.; and Köhler, J. M. 2018. The power of ensembles for active learning in image classification. In *IEEE CVPR*, 9368–9377.

Cai, L.; Xu, X.; Liew, J. H.; and Foo, C. S. 2021. Revisiting Superpixels for Active Learning in Semantic Segmentation With Realistic Annotation Costs. In *IEEE CVPR*, 10988–10997.

Caramalau, R.; Bhattarai, B.; and Kim, T.-K. 2021. Sequential Graph Convolutional Network for Active Learning. In *IEEE CVPR*, 9583–9592.

Casanova, A.; Pinheiro, P. O.; Rostamzadeh, N.; and Pal, C. J. 2019. Reinforced active learning for image segmentation. In *ICLR*.

Chen, L.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *CoRR*, abs/1706.05587.

Choi, J.; Yi, K. M.; Kim, J.; Choo, J.; Kim, B.; Chang, J.; Gwon, Y.; and Chang, H. J. 2021. VaB-AL: Incorporating Class Imbalance and Difficulty with Variational Bayes for Active Learning. In *IEEE CVPR*, 6749–6758.

Cohn, D. A.; Ghahramani, Z.; and Jordan, M. I. 1996. Active learning with statistical models. *Journal of artificial intelligence research*, 4: 129–145.

Colling, P.; Roese-Koerner, L.; Gottschalk, H.; and Rottmann, M. 2021. MetaBox+: A New Region based Active Learning Method for Semantic Segmentation using Priority Maps. In *ICPRAM*, 51–62.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *IEEE CVPR*, 3213–3223.

Dai, C.; Wang, S.; Mo, Y.; Zhou, K.; Angelini, E.; Guo, Y.; and Bai, W. 2020. Suggestive Annotation of Brain Tumour Images with Gradient-Guided Sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 156–165. Springer.

Ebrahimi, S.; Elhoseiny, M.; Darrell, T.; and Rohrbach, M. 2019. Uncertainty-guided Continual Learning with Bayesian Neural Networks. In *ICLR*.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 1050–1059.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *ICML*, 1183–1192.

Górriz, M.; Giró Nieto, X.; Carlier, A.; and Faure, E. 2017. Cost-effective active learning for melanoma segmentation. In *NeurIPS Workshop*, 1–5.

Gudovskiy, D.; Hodgkinson, A.; Yamaguchi, T.; and Tsukizawa, S. 2020. Deep active learning for biased datasets via fisher kernel self-supervision. In *IEEE CVPR*, 9041–9049.

Hua, G.; Long, C.; Yang, M.; and Gao, Y. 2013. Collaborative Active Learning of a Kernel Machine Ensemble for Recognition. In *ICCV*, 1209–1216. IEEE.

Hua, G.; Long, C.; Yang, M.; and Gao, Y. 2018. Collaborative Active Visual Recognition from Crowds: A Distributed Ensemble Approach. *T-PAMI*, 40(3): 582–594.

Jain, S. D.; and Grauman, K. 2016. Active image segmentation propagation. In *IEEE CVPR*, 2864–2873.

Kapoor, A.; Grauman, K.; Urtasun, R.; and Darrell, T. 2007. Active learning with gaussian processes for object categorization. In *IEEE ICCV*, 1–8.

Kim, K.; Park, D.; Kim, K. I.; and Chun, S. Y. 2021. Task-aware variational adversarial active learning. In *IEEE CVPR*, 8166–8175.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Kirsch, A.; Van Amersfoort, J.; and Gal, Y. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *NeurIPS*, 32: 7026–7037.

Kuo, W.; Häne, C.; Yuh, E.; Mukherjee, P.; and Malik, J. 2018. Cost-sensitive active learning for intracranial hemorrhage detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 715–723. Springer.

Lee, J.; Kim, E.; Lee, S.; Lee, J.; and Yoon, S. 2019. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *IEEE CVPR*, 5267–5276.

Li, M.; and Sethi, I. K. 2006. Confidence-based active learning. *TPAMI*, 28(8): 1251–1261.

Li, X.; and Guo, Y. 2013. Adaptive active learning for image classification. In *IEEE CVPR*, 859–866.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *IEEE CVPR*, 2117–2125.

Liu, Y.; Xie, J.; Shi, X.; Qiao, Y.; Huang, Y.; Tang, Y.; and Yang, X. 2021. Tripartite Information Mining and Integration for Image Matting. In *ICCV*, 7555–7564.

Liu, Z.; Wang, J.; Gong, S.; Lu, H.; and Tao, D. 2019. Deep reinforcement active learning for human-in-the-loop person re-identification. In *IEEE ICCV*, 6122–6131.

Long, C.; and Hua, G. 2015. Multi-class multi-annotator active learning with robust gaussian process for visual recognition. In *ICCV*, 2839–2847.

Long, C.; Hua, G.; and Kapoor, A. 2013. Active Visual Recognition with Expertise Estimation in Crowdsourcing. In *ICCV*, 3000–3007. IEEE.

Long, C.; Hua, G.; and Kapoor, A. 2016. A Joint Gaussian Process Model for Active Visual Recognition with Expertise Estimation in Crowdsourcing. *IJCV*, 116(2): 136–160.

Mackowiak, R.; Lenz, P.; Ghori, O.; Diego, F.; Lange, O.; and Rother, C. 2018. CEREALS-Cost-Effective REgion-based Active Learning for Semantic Segmentation. In *BMVC*.

Mahapatra, D.; Bozorgtabar, B.; Thiran, J.-P.; and Reyes, M. 2018. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 580–588. Springer.

Mayer, C.; and Timofte, R. 2020. Adversarial sampling for active learning. In *WACV*, 3071–3079.

Mei, H.; Dong, B.; Dong, W.; Peers, P.; Yang, X.; Zhang, Q.; and Wei, X. 2021. Depth-Aware Mirror Segmentation. In *IEEE CVPR*, 3044–3053.

Mei, H.; Yang, X.; Wang, Y.; Liu, Y.; He, S.; Zhang, Q.; Wei, X.; and Lau, R. W. 2020. Don't Hit Me! Glass Detection in Real-World Scenes. In *IEEE CVPR*, 3687–3696.

Melville, P.; and Mooney, R. J. 2004. Diverse ensembles for active learning. In *ICML*, 74.

Ozdemir, F.; Peng, Z.; Tanner, C.; Fuernstahl, P.; and Goksel, O. 2018. Active learning for segmentation by optimizing content information for maximal entropy. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 183–191. Springer.

Qiao, Y.; Liu, Y.; Zhu, Q.; Yang, X.; Wang, Y.; Zhang, Q.; and Wei, X. 2020. Multi-scale Information Assembly for Image Matting. *CGF*, 39(7): 565–574.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE CVPR*, 4510–4520.

Sener, O.; and Savarese, S. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *ICLR*.

Siddiqui, Y.; Valentin, J.; and Nießner, M. 2020. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *IEEE CVPR*, 9433–9443.

Sinha, S.; Ebrahimi, S.; and Darrell, T. 2019. Variational adversarial active learning. In *IEEE ICCV*, 5972–5981.

Wang, Z.; and Ye, J. 2015. Querying discriminative and representative samples for batch mode active learning. *TKDD*, 9(3): 1–23.

Wu, Y.; Xu, Y.; Singh, A.; Yang, Y.; and Dubrawski, A. 2019. Active learning for graph neural networks via node feature propagation. *arXiv preprint arXiv:1910.07567*.

Yang, L.; Zhang, Y.; Chen, J.; Zhang, S.; and Chen, D. Z. 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, 399–407. Springer.

Yang, X.; Qiao, Y.; Chen, S.; He, S.; Yin, B.; Zhang, Q.; Wei, X.; and Lau, R. W. 2020. Smart Scribbles for Image Matting. *ACM TOMM*, 16(4): 1–21.

Yang, X.; Xu, K.; Chen, S.; He, S.; Yin, B.; and Lau, R. W. 2018. Active matting. In *NeurIPS*, 4595–4605.

Yoo, D.; and Kweon, I. S. 2019. Learning loss for active learning. In *IEEE CVPR*, 93–102.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE CVPR*, 2636–2645.

Yu, F.; Koltun, V.; and Funkhouser, T. 2017. Dilated residual networks. In *IEEE CVPR*, 472–480.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018. Generative image inpainting with contextual attention. In *IEEE CVPR*, 5505–5514.

Yuan, T.; Wan, F.; Fu, M.; Liu, J.; Xu, S.; Ji, X.; and Ye, Q. 2021. Multiple instance active learning for object detection. In *IEEE CVPR*, 5330–5339.

Zhang, B.; Li, L.; Yang, S.; Wang, S.; Zha, Z.-J.; and Huang, Q. 2020. State-relabeling adversarial active learning. In *IEEE CVPR*, 8756–8765.