Region-aware Contrastive Learning for Semantic Segmentation

Hanzhe Hu[1,2], Jinshi Cui[1]*, Liwei Wang[1]*
[1]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University [2]Zhejiang Lab
huhz@pku.edu.cn {cjs, wanglw}@cis.pku.edu.cn

Abstract

Recent works have made great success in semantic segmentation by exploiting contextual information in a local or global manner within individual image and supervising the model with pixel-wise cross entropy loss. However, from the holistic view of the whole dataset, semantic relations not only exist inside one single image, but also prevail in the whole training data, which makes solely considering intra-image correlations insufficient. Inspired by recent progress in unsupervised contrastive learning, we propose the region-aware contrastive learning (RegionContrast) for semantic segmentation in the supervised manner. In order to enhance the similarity of semantically similar pixels while keeping the discrimination from others, we employ contrastive learning to realize this objective. With the help of memory bank, we explore to store all the representative features into the memory. Without loss of generality, to efficiently incorporate all training data into the memory bank while avoiding taking too much computation resource, we propose to construct region centers to represent features from different categories for every image. Hence, the proposed region-aware contrastive learning is performed in a region level for all the training data, which saves much more memory than methods exploring the pixel-level relations. The proposed RegionContrast brings little computation cost during training and requires no extra overhead for testing. Extensive experiments demonstrate that our method achieves state-of-the-art performance on three benchmark datasets including Cityscapes, ADE20K and COCO Stuff.

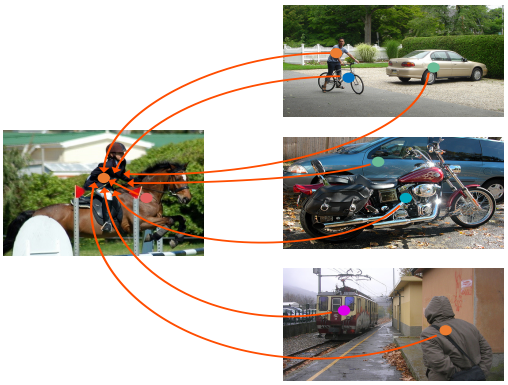
1. Introduction

Semantic segmentation, which aims to assign a category label to each pixel in an image, is a fundamental and challenging problem in computer vision. It has been widely applied to many applications, such as autonomous driving, scene understanding and image editing.

In the past few years, benefiting from the availability of large-scale datasets such as ImageNet [11] and Cityscapes [10], semantic segmentation has achieved significant progress. In particular, based on the fully convolutional network (FCN) [30], many state-of-the-art methods emerge, which focus on exploiting contextual information. DeepLabV3 [5] proposes ASPP module which aggregates spatial regularly sampled pixels at different dilated rates while PSPNet [52] proposes pyramid pooling module which partitions the feature maps into multiple regions before pooling. Non-local Network [40] adopts self-attention mechanism to enable every pixel to receive information from all other pixels, resulting in a much more complete pixel-wise representation.

Aforementioned methods, though achieving satisfactory

(a) Most Current Methods



(b) Proposed RegionContrast

Figure 1. Main difference between our method and previous ones. 1(a) Most existing methods only focus on intra-image relations. 1(b) Our proposed RegionContrast, apart from solely focusing on intra-image information, also considers inter-image correlations in a region level.

* Corresponding authors.

segmentation results in most occasions, are still faced with critical drawback. Concretely, most current methods focus on mining contextual information in all kinds of ways within the image, neglecting the potential relation information from other images. As illustrated in Fig. 1, inter-class relations are also worth exploring. For one region of an image, dilation convolution or self-attention mechanism can only enable it to receive information from some specific features of surrounding categories, while in reality, this kind of region could get in touch with much more kinds of features. Hence, only exploring intra-image relations is not comprehensive enough, which results in a demand for feature learning from the holistic view of the whole dataset.

Recently, unsupervised contrastive learning has gained much attention in pre-training a strong feature extractor for downstream tasks such as image classification or object detection. In brief, most works perform contrastive learning in the image level where all other images from the dataset are considered as negative samples while the very image with random augmentations is treated as the postive sample. Benefiting from the utilization of memory bank, large amount of negative samples can be brought in to assist the contrastive learning for better feature representations. Note that most unsupervised contrastive learning methods focus on the classification problem, while semantic segmentation, on the other hand, requires much more semantic information than classification. Intuitively, to adapt to segmentation problem, instead of performing image-level contrastive learning, we can adjust to a pixel-level one where pixels inside and outside the very image get contrasted, as depicted in [41]. However, this kind of formulation suffers from a serious issue: pixels from different images may belong to the same category, which would deteriorate the subsequent feature learning. Hence, instead of sticking to unsupervised settings, we explore contrastive learning in the fully supervised manner to obtain abundant category information.

In this work, we propose a new contrastive learning paradigm in a fully supervised way, targeting at semantic segmentation problem. With the corresponding categories of pixels as prior knowledge, the contrastive learning can be performed in a much more efficient way. We will first describe the most straightforward approach. Specifically, when the category of each pixel is known which arises from the prediction of the model, different memory banks that conforms to each class are built to store different classes of pixel embeddings. And for each pixel of an image, its corresponding positive and negative samples can be retrived from the memory banks, which would complete the contrastive learning process. Though simple and effective, this method would result in heavy memory burden since the number of pixels of an entire dataste is too large, which will also severely slow down the training speed.

To tackle the above issue while restoring enough embeddings, we propose the region-aware contrastive learning (RegionContrast). In particular, since the region features for one class in an image are composed of all the pixel features belonging to this category, we can construct the region centers for different categories within one image. In that way, instead of pushing all the pixel embeddings into memory banks, we just push several region centers from different categories into the banks. Although an image may contain multiple regions that belong to the same category, the features in the embedding space are similar. Thus for simplicity, we generate one region center for each class of one single image. To facilitate the feature learning for hard-to-classify pixels, we further propose a dynamic sampling method when generating the regions centers to allocate more attention to hard samples. After building the memory banks for different classes, region-aware contrastive learning can be performed. Specifically, for one region center of an image, its corresponding positive samples come from the embeddings in the memory bank of the same class, while its negative samples are embeddings from other memory banks. With positive and negative samples provided, contrastive learning procedure can be implemented.

The overall framework of our RegionContrast is shown in Fig. 2, where conventional cross entropy loss works as a pixel-wise supervision and RegionContrast focuses on inter-image relation learning. Most importantly, the proposed RegionContrast can be easily applied into any segmentation models, and only demands little computation resources during training while requiring no extra overhead for testing.

To sum up, our contributions are summarized as follows:

- We propose a new contrastive learning setting in the fully supervised manner and target at the specific semantic segmentation problem.
- To adapt to segmentation scenario in a memory-efficient way, we design an effective region-aware contrastive learning (RegionContrast) to explore semantic relations from the holistic view of the whole dataset.
- We conduct extensive experiments on several public datasets, and obtain state-of-the-art performance on three semantic segmentation benchmarks, including Cityscapes, ADE20K and COCO Stuff.

2. Related Work

2.1. Semantic Segmentation

With the success of deep neural networks [25, 35, 19], semantic segmentation has achieved great progress. FCN [30] is the first approach to adopt fully convolutional network for semantic segmentation. Later, many FCN-baed works are proposed, such as UNet [33], RefineNet [29],

PSPNet [52], DeepLab series[3, 4, 5, 6]. Chen *et al.* [4] and Yu *et al.* [45] remove the last two downsample layers to obtain a dense prediction and utilized dilated convolutions to enlarge the receptive field. We choose DeepLabV3 as the basic segmentation network for convenience. And we also adopt the above paradigm to get a better feature map and thus improve the performance of the model. However, most of the previous methods utilize a typical pixel-wise cross-entropy loss to supervise the training of the model, neglecting the intrinsic correlations between different pixels.

2.2. Context

Contextual information is critical for semantic segmentation to generate better feature representations. From the local perspective, DeepLabV3 [5] employs multiple atrous convolutions with different dilation rates to capture contextual information, while PSPNet [52] utilizes pyramid pooling over sub-regions to harvest information. While from the global perspective, Wang *et al.* [40] apply the idea of self-attention from transformer [37] into vision problems and propose the non-local module where correlations between all pixels are calculated to guide the dense contextual information aggregation. These methods, though effectively harness contextual information within the image, all suffer from the drawback that inter-image relations are neglected. Hence, to learn a more comprehensive feature representation, we propose to further explore inter-image relations on a region level.

2.3. Contrastive Learning

Recently, contrastive learning [36, 42, 7, 18, 8] which is based on Siamese networks [1] has achieved great progress in unsupervised learning problem, and significantly outperform the previous pretext task based methods [26, 16, 13, 31]. SimCLR [7] proposes a simple framework to perform contrastive learning, where positive pairs are generated with two random augmented views of the same image and negative ones are obtained with different images, forming a image-level discrimination task. Contrastive learning aims to increase the instance discriminative power between different images and mainly benefits from the large number of negative samples. Furthermore, MoCo [18] maintains a queue of negative samples and turns one branch of Siamese network into a momentum encoder to improve consistency of the queue. Moreover, targeting at specific semantic segmentation problem, DenseCL [41] performs dense contrastive learning at the level of pixels. Besides, a few previous methods [23, 22, 51] propose to perform unsupervised clustering for segmentation problem using contrastive losses. However, above works using contrastive learning mainly focus on unsupervised pre-training task or clustering. Without the guidance of labels, some serious problems may occur. In particular, these methods treat instances or pixels from different images as negative pairs, which may come from the same category. With contrastive learning pushing away these features, the final performance for downstream tasks would get jeopardized. To overcome this issue, we choose to explore contrastive learning from the view of fully supervised manner. Based on the available segmentation labels, deeper level of specific semantic relations can be explored and contrastive learning can help enhance the feature similarity within the same class and increase the discrimination power between different classes.

3. Method

In this section, we will describe the proposed Region-aware Contrastive Learning (RegionContrast) in detail. We will first revisit the background knowledge about conventional contrastive learning in unsupervised representation learning. Then we will present the details of our proposed RegionContrast which is in a supervised manner.

3.1. Background

Unsupervised Contrastive Learning. Recently, unsupervised (self-supervised) representation learning has gained considerable progress. Breakthrough approaches such as SimCLR [7], MoCo-v1/v2 [18, 8] utilize contrastive learning to obtain good representations from unlabeled data, which aims to learn a CNN encoder to transform the input images to feature representations. Given an unlabeled dataset, an instance discrimination pretext task is employed where the feature of each image in the training set is pulled away from that of other images. For each input image, random 'views' are generated by random data augmentations. Each view is fed into an encoder to extract high-dimensional feature that holistically encodes the input view. The encoder consists of a backbone network and a projection head. The backbone network is the model to be used for downstream tasks after pre-training and the project head will be discarded afterwards. The parameters of encoders for different views can be shared or momentum-updated for the other. Encoders are optimized by a pairwise contrastive loss which measures the similarity of different feature vectors generated from the projection heads.

Memory Bank. To better optimize the encoder, positive and negative samples are indispensable for contrastive learning. While different views of the same image generated from augmentations are considered as positive samples, other images can be treated as negative ones. Since the size of mini-batch is limited, a large memory bank which stores embeddings of training images is adopted in [42, 18, 8]. During training, negative samples can be effectively retrieved from the memory bank to construct the complete contractive loss function.

Loss Function. Following MoCo [18, 8], the contrastive learning can be considered as a dictionary look-up task.

Formally, for each encoded query q, a set of encoded keys $\{k_0, k_1, ...\}$ can be retrieved from the memory bank, among which a single positive key k_+ corresponds to query q while other negative keys k_- encode views of other images. A contrastive loss function InfoNCE [36] is employed to pull q close to k_+ while pushing it away from negative keys k_-:

$$L_q^{NCE} = -\log \frac{\exp(q \cdot k_+/\tau)}{\exp(q \cdot k_+/\tau) + \sum_{k_-} \exp(q \cdot k_-/\tau)}, \tag{1}$$

where τ denotes a temperature hyper-parameter. All the embeddings in the loss function are L_2-normalized.

3.2. Region-aware Contrastive Learning

In this work, we explore contrastive learning in a supervised manner. Benefiting from the available labeled data, the contrastive learning is exployed in a category level rather than instance level as in previous methods, hence better enhancing the feature representation.

3.2.1 Overall Framework

As illustrated in Fig. 2, we present the overall framework of our proposed RegionContrast. We choose DeepLabV3 [5] as the basic segmentation network. We use the ResNet pretrained on ImageNet dataset as the backbone, replace the last two down-sampling operations and employ dilation convolutions in the subsequent convolutional layers, enlarging the resolution and receptive field of the feature map, so the output stride becomes 8 instead of 16. The model is supervised with conventional pixel-wise cross entropy loss together with the proposed region-aware contrastive loss.

Specifically, to perform region-level contrastive learning, region features need to be extracted. Given an input image $I \in \mathbb{R}^{C\times H\times W}$, we feed it through the backbone and ASPP module to obtain the feature map $F \in \mathbb{R}^{C\times H\times W}$. Region features will be further extracted from the feature map F under the guidance of predictions of the network, which are achieved by adding a segmentation head onto the feature map. For simplicity, we choose to represent each class in one image with one region center that encodes the most essential information about the very class. In practice, the generated region center is a vector $R_i \in \mathbb{R}^C$ for class i. Subsequently, region centers from the same category are positive samples while those from other categories become negative samples. The key of the proposed supervised region-aware contrastive learning is to pull region features of the same class together while keeping discriminative power between different classes.

3.2.2 Region Center

Intuitively, the region center of class i can be defined as the average of features of all pixels belonging to class i in a single image. Formally, given the feature map $F \in \mathbb{R}^{C\times H\times W}$, where C, H and W denote the dimension, height and width of the feature map respectively, the straightforward definition of a region center of class i can be defined as,

$$R_i = \frac{\sum_{x,y} F_{(x,y)} \mathbb{1}[L_{(x,y)} = i]}{\sum_{x,y} \mathbb{1}[L_{(x,y)} = i]}, \tag{2}$$

where $L_{(x,y)}$ is the label of the pixel which is predicted by the basic segmentation network and $\mathbb{1}(\cdot)$ is the binary indicator denoting whether the pixel belongs to class i.

However, with the formulation depicted above, the constructed region center covers ambiguous features of pixels since the network prediction contains false prediction, which would mislead the learning process of region centers. In order to allocate more attention to hard-to-classify pixels, we further propose a **dynamic sampling** method to construct the region centers.

Apart from feature map $F \in \mathbb{R}^{C\times H\times W}$ and prediction map $P \in \mathbb{R}^{N\times H\times W}$, groundtruth map $G \in \mathbb{R}^{H\times W}$ is incorporated to mine out hard pixels, where C is the feature dimension and N is the number of classes. With the guidance of groudtruth, hard negative pixels can be filtered out while hard positive ones can be retrieved. To pay more attention to hard samples, the weights of different pixel features when producing the region center should be different, where pixels of hard samples require higher weights than easy positive ones. To this end, we harness the predicted confidence map to allocate weights to easy positive samples, where the value on each position of confidence map $c_{i,j} \in [0, 1]$. Hence, the weight of easy positive pixel (i, j) is $1 - c_{i,j}$, while the weight of hard positive pixel is 1. Formally, easy positive samples for class i can be denoted as $EP_i = \sum_{x,y} \mathbb{1}[L_{(x,y)} = i] \cap \mathbb{1}[G_{(x,y)} = i]$, and hard positive samples can be denoted as $HP_i = \sum_{x,y} \mathbb{1}[G_{(x,y)} = i] - \mathbb{1}[L_{(x,y)} = i] \cap \mathbb{1}[G_{(x,y)} = i]$. The final definition of regions centers of class i can be defined as,

$$R_i = \frac{\sum_{x,y} F_{(x,y)}((1 - c_{(x,y)}) \cdot EP_i + HP_i)}{\sum_{x,y} \mathbb{1}[G_{(x,y)} = i]}, \tag{3}$$

3.2.3 RegionContrast

After constructing the region centers of all categories for each image, region-aware contrastive loss for a region center from class i can be directly defined as,

$$L_i^{NCE} = \frac{1}{|M_i|} \sum_{k_+ \in M_i} -\log \frac{\exp(q \cdot k_+/\tau)}{\exp(q \cdot k_+/\tau) + \sum_{k_-} \exp(q \cdot k_-/\tau)}, \tag{4}$$

where M_i denotes the memory bank collecting region centers from category i of the whole training set and k_- comes from the memory banks for other categories.

As stated in Section. 3.1, memory bank that contains negative samples is vital in learning good feature representations. Previous methods [41, 54, 39] that apply contrastive

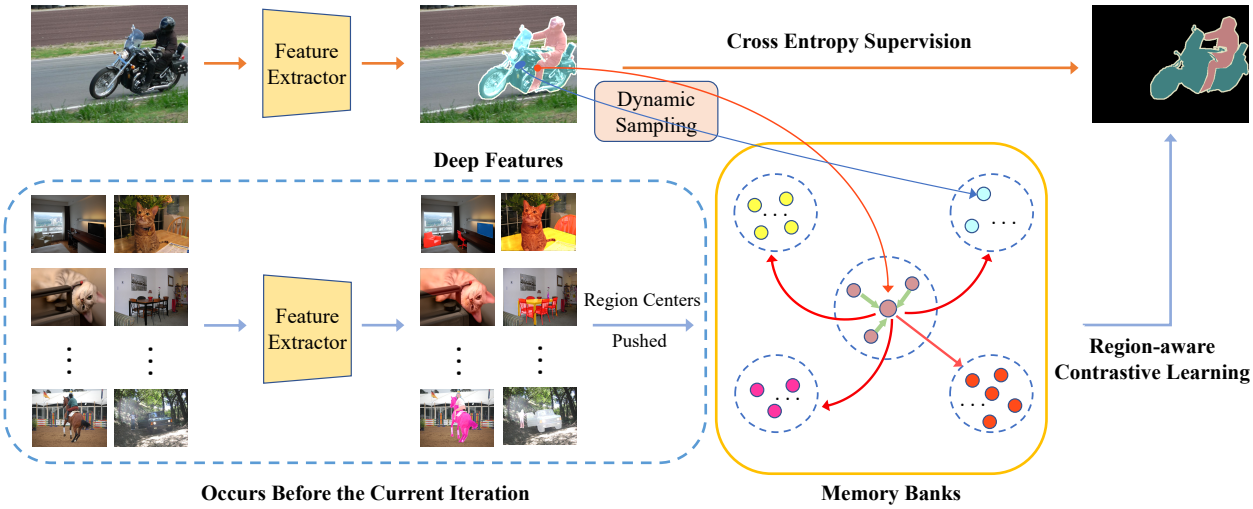


Figure 2. The overall framework of our proposed region-aware contrastive learning (RegionContrast). The memory banks for all the categories contain region centers generated from images before the current iteration. Given an input image for the current iteration, we first feed it into the feature extractor (Backbone + ASPP module) to obtain its deep features. With the proposed dynamic sampling method, we construct the region centers from the image and push them into the memory banks according to the corresponding category. Within the memory banks, region-aware contrastive learning is performed, where red arrows denote pushing force and green ones denote pulling force (different colors of dots denote region centers of different classes). The model is jointly supervised with cross entropy loss and the proposed RegionContrast.

learning in semantic segmentation have to maintain a large memory bank to store embeddings in a pixel level, which leads to a serious demand for large capacity of memory bank and slow training speed due to the heavy memory burden. While our method, on the other hand, benefits from the introduction of region centers and thus requires much fewer memory during training. Formally, for a training set with D images from N categories, N memory banks are constructed, each with the size of $D \times C$ where C is the feature dimension of the embeddings which are also known as region centers. Specifically, we maintain these N memory banks as different queues during training. For each mini-batch, region centers of different categories are generated and pulled into the corresponding queue and get updated in the next training epoch.

To sum up, the final supervision for semantic segmentation is summarized as follows. The proposed region-aware contrastive loss is adopted. Conventional pixel-wise cross entropy loss is also employed together with the auxiliary loss as in previous state-of-the-art works [52, 47, 20]. Specifically, the output of the third stage of our backbone ResNet is further fed into a auxiliary layer to produce a prediction supervised by the auxiliary loss which is also cross entropy loss. In a word, the loss can be formulated as follows,

$$L = L_{CE} + \alpha L_{AUX} + \beta L_{RC}, \tag{5}$$

where α, β are used to balance the main segmentation loss L_{CE}, auxiliary loss L_{AUX} and region-aware contrastive loss L_{RC}.

4. Experiments

To evaluate the performance of our proposed RegionContrast, we carry out extensive experiments on three benchmark datasets including Cityscapes [10], ADE20K [55] and COCO Stuff [2]. Experimental results demonstrate that the proposed RegionContrast can effectively boost the performance of the state-of-the-art methods. In the following section, we will first introduce the datasets and implementation details, and then present detailed ablation study on Cityscapes dataset. Finally, we will report the results on ADE20K dataset and COCO Stuff dataset.

4.1. Datasets and Evaluation Metrics

Cityscapes. The Cityscapes dataset [10] is tasked for urban scene understanding, which contains 30 classes and only 19 classes of them are used for scene parsing evaluation. The dataset contains 5000 finely annotated images and 20000 coarsely annotated images. The finely annotated 5000 images are divided into 2975/500/1525 images for training, validation and testing.

ADE20K. The ADE20K dataset [55] is a large scale scene parsing benchmark which contains dense labels of 150 stuff/object categories. The dataset includes 20K/2K/3K images for training, validation and testing.

COCO Stuff. The COCO Stuff dataset [2] is a challenging scene parsing dataset containing 59 semantic classes and 1 background class. The training and test set consist of 9K and 1K images respectively.

Method	mIoU(%)
CE Baseline	76.4
RegionContrast (intra-image)	77.5
RegionContrast (inter-image)	79.6

Table 1. Performance comparisons of our proposed RegionContrast on Cityscapes validation set.

Evaluation Metric. In our experiments, the mean of class-wise Intersection over Union (mIoU) is used as the evaluation metric.

4.2. Implementation Details

We choose the ImageNet pretrained ResNet as our backbone, remove the last two down-sampling operations and employ dilated convolutions in the subsequent convolution layers, making the output stride equal to 8. For training, we use the stochastic gradient descent (SGD) optimizer with initial learning rate 0.01, weight decay 0.0005 and momentum 0.9 for Cityscapes dataset. Moreover, we adopt the 'poly' learning rate policy, where the initial learning rate is multiplied by $(1 - \frac{iter}{max\_iter})^{power}$ with power=0.9. For Cityscapes dataset, we adopt the crop size as 769×769, batch size as 8 and training iterations as 30K. For ADE20K dataset, we set the initial learning rate as 0.004, weight decay as 0.0001, crop size as 480×480, batch size as 16 and training iterations as 150K. For COCO Stuff dataset, we set initial learning rate as 0.001, weight decay as 0.0001, crop size as 520×520, batch size as 16, and training iterations as 60K. The loss weights for L_{AUX} and L_{RC} are 0.4 and 0.1 respectively.

4.3. Ablation Study

In this subsection, we conduct extensive ablation experiments on the validation set of Cityscapes dataset with different settings for our proposed RegionContrast. For all the experiments in this subsection, we use the DeepLabV3 as our segmentation network with dilated ResNet-50 as the backbone network.

The Impact of RegionContrast. We carry out experiments to evaluate the effectiveness of the proposed RegionContrast. Different levels of RegionContrast are adopted. Concretely, we choose cross entropy (CE) loss function as the pixel-wise supervision, which is also our baseline method. RegionContrast (intra-image) denotes that region-aware contrastive learning only takes place inside the image, where much fewer positive and negative samples are used. RegionContrast (inter-image) denotes that region-aware contrastive learning takes place in the whole training set, which induce adequate positive and negative samples to ensure the contrastive learning. As shown in Table 1, RegionContrast can achieve consistent improvement over the baseline. Moreover, with inter-image contrastive learning bringing sufficient negative samples, the performance is further boosted.

Method	mIoU(%)
CE Baseline	76.4
RegionContrast (Direct Average)	78.2
RegionContrast (EP + HP)	78.8
RegionContrast (Weighted EP + HP)	79.6

Table 2. Performance comparisons of different construction methods of region centers on Cityscapes validation set. 'EP' and 'HP' denote easy positive and hard positive samples respectively.

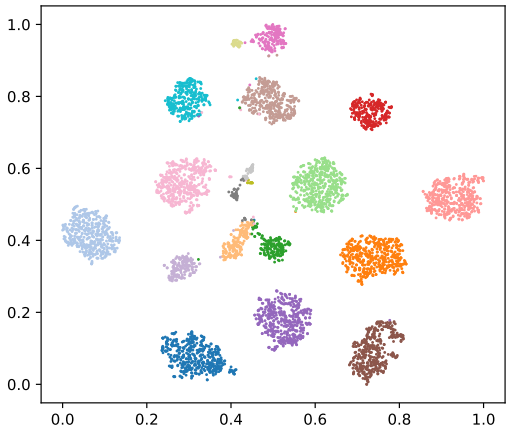
Method	Bank_Size	mIoU(%)
CE Baseline	0	76.4
RegionContrast w/o MB	0	77.5
RegionContrast w/MB	1000	78.5
RegionContrast w/MB	2000	79.1
RegionContrast w/MB	2975	79.6

Table 3. Performance effect of memory banks on Cityscapes validation set. 'MB' denotes the memory bank. '2975' is the size of training data of Cityscapes dataset.

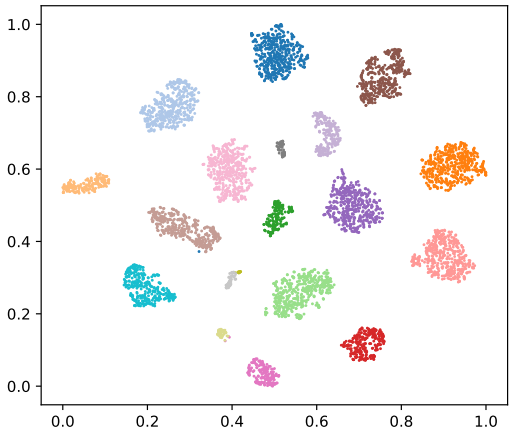
The Impact of Dynamic Sampling. We further perform expeirments to validate the effectiveness of dynamic sampling method when constructing the region centers. As illustrated in Table 2, we apply different construction methods, where 'Direct Average' corresponds to the plain way defined in Eq. 2 and 'Weighted EP + HP' denotes the final dynamic sampling construction manner. From line 2 and line 3 of the table, it can be seen that hard positive samples are more critical than hard negative ones. Moreover, the results demonstrate that the dynamic sampling method can effectively deal with hard samples and provoke a more robust region representation.

The Impact of Memory Bank. In this subsection, we implement extensive experiments to evaluate the significance of memory bank. As shown in Table 3, several experimental settings are adopted. Specifically, when RegionContrast is performed without memory bank, it becomes the same as RegionContrast (intra-image) as in Table 1. It can be induced that as the size of memory bank grows, the performance can be further improved, which validate the effectiveness of the memory bank. Larger memory bank is capable of containing more features, providing richer information for later contrastive learning.

The impact on Different Models. We carry out experiments to assess the effectiveness of the proposed Re-



(a) CE Supervision



(b) Joint Supervision of CE and RegionContrast

Figure 3. Visualization results of region centers. The model is supervised by cross entropy loss and cross entropy loss together with RegionContrast, respectively. The points with different colors denote region centers from different classes

gionContrast with different models. Different segmentation models and backbones are adopted, including DeepLabV3 with ResNet [19] and HRNetV2 [38]. As shown in Table 4, RegionContrast improves DeepLabV3 with ResNet-50 by 3.2% in mIoU, DeepLabV3 with ResNet-101 by 2.3% and HRNetV2 by 1.5%, indicating that the proposed RegionContrast can be applied into any segmentation models.

Visualizations of the effect of RegionContrast. To further comprehend the effect of RegionContrast, qualitative results are shown in Fig. 3. Concretely, we calculate the region centers for all the classes for every image in the validation set of Cityscapes and visualize all the features using t-SNE visualization in Fig. 3, where each point corresponds to one region center. In Fig. 3(a), with only cross entropy loss supervision, several region centers get mixed together, which severely increases the ambiguity among categories

Method	Backbone	mIoU(%)
DeepLabV3	ResNet-50	76.4
DeepLabV3 + RegionContrast	ResNet-50	79.6
DeepLabV3	ResNet-101	79.0
DeepLabV3 + RegionContrast	ResNet-101	81.3
HRNetV2	HRNetV2-W48	80.4
HRNetV2 + RegionContrast	HRNetV2-W48	81.9

Table 4. Performance effect of RegionContrast with different models on Cityscapes validation set.

Method	Backbone	mIoU(%)
RefineNet [29]	ResNet-101	73.6
GCN [32]	ResNet-101	76.9
PSPNet [52]	ResNet-101	78.4
AAF [24]	ResNet-101	79.1
DFN [44]	ResNet-101	79.3
PSANet [53]	ResNet-101	80.1
GloRe [9]	ResNet-101	80.9
CPNet [43]	ResNet-101	81.3
CCNet [21]	ResNet-101	81.4
DANet [14]	ResNet-101	81.5
OCR [46]	ResNet-101	81.8
RegionContrast(Ours)	ResNet-101	**82.3**

Table 5. Comparisons with state-of-art on the Cityscapes test set.

and raises the classification difficulty for the model. With the introduction of RegionContrast, as shown in Fig. 3(b), the discriminative power between region centers from different categories gets significantly enhanced. Therefore, the joint supervision of CE and RegionContrast remarkably benefits the feature learning in amplifying the discrimination between classes.

We further provide comparisons of visualization results on validation set of Cityscapes dataset in Fig. 4. It can be seen that our proposed RegionContrast can effectively improve the consistency of predictions with region-level inter-image relation explorations.

4.4. Comparisons with State-of-the-Arts

Cityscapes. Furthermore, we also train the proposed method using both training and validation set of Cityscapes dataset and make the evaluation on the test set by submitting our test results to the official evaluation server. For a fair comparison, we use ResNet-101 as backbone, OHEM loss as the pixel-level loss, and our proposed RegionContrast to supervise the learning process. Moreover, we use the multi-scale and flipping strategies while testing. From Table 5, it

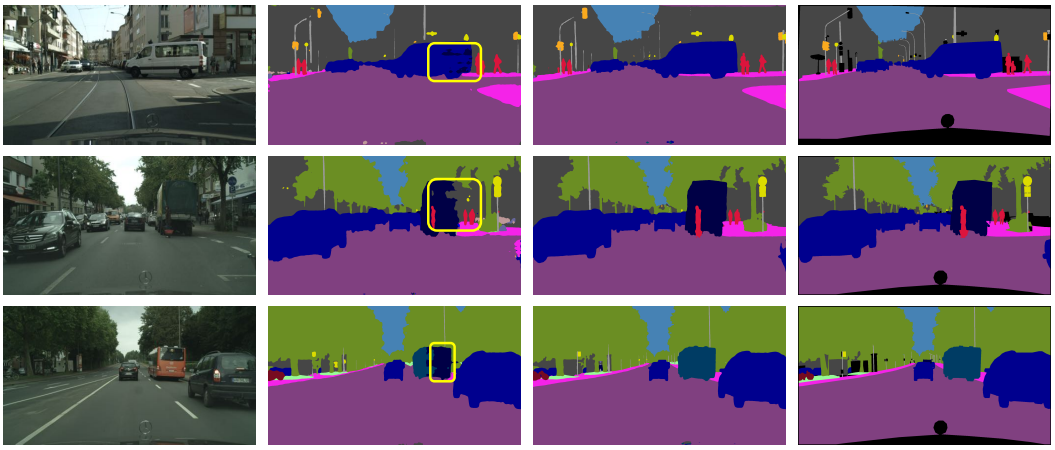


Figure 4. Visualization results on Cityscapes validation set. From left to right: input image, predictions made by the baseline method supervised by cross entropy loss, predictions made by our proposed RegionContrast and groudtruth map.

Method	Backbone	mIoU(%)
RefineNet [29]	ResNet-152	40.70
PSPNet [52]	ResNet-101	43.29
DSSPN [28]	ResNet-101	43.68
PSANet [53]	ResNet-101	43.77
SAC [50]	ResNet-101	44.30
EncNet [48]	ResNet-101	44.65
CFNet [49]	ResNet-101	44.89
APCNet [17]	ResNet-101	45.38
CPNet [43]	ResNet-101	46.27
RegionContrast(Ours)	ResNet-101	**46.85**

Table 6. Comparisons with state-of-art on the ADE20K validation set.

Method	Backbone	mIoU(%)
FCN-8s [30]	VGG-16	22.7
DAG-RNN [34]	VGG-16	31.2
RefineNet [29]	ResNet-101	33.6
CCL [12]	ResNet-101	35.7
DSSPN [28]	ResNet-101	38.9
DANet [14]	ResNet-101	39.7
EMANet [27]	ResNet-101	39.9
ACNet [15]	ResNet-101	40.1
RegionContrast (Ours)	ResNet-101	**40.7**

Table 7. Comparisons with state-of-art on the COCO Stuff test set.

can be observed that our proposed RegionContrast achieves state-of-the-art performance on Cityscapes test set.

ADE20K. We also conduct experiments on the ADE20K dataset. Performance results on the validation set are reported in Table 6. Our method achieves state-of-the-art result on the validation set of ADE20K dataset.

COCO Stuff. We also conduct experiments on the COCO Stuff dataset and report the results in Table 7. Results show that our model achieves 40.7% in mean IoU which is the highest record. Hence our method can effectively collect useful long-range contextual information and obtain better feature representation in semantic segmentation.

5. Conclusions

In this paper, we have presented the Region-aware Contrastive Learning (RegionContrast) to incorporate contrastive learning into semantic segmentation problem. Different from previous unsupervised contrastive learning methods, we propose a new contrastive learning setting in the fully supervised manner and target at segmentation problem. With the availability of labels, we are capable of exploring more semantic relations. Moreover, we propose the concept of region centers for different categories which are stored in the memory and participate in the subsequent contrastive learning procedure. With region-level embeddings instead of pixel-level embeddings to store the information of the holistic training set, contrastive learning can be implemented in a memory-efficient way. The ablation experiements demonstrate the effectiveness of each component of RegionContrast. Our proposed RegionContrast achieves state-of-the-art results on three benchmark datasets, *i.e.* Cityscapes, ADE20K and COCO Stuff.

Acknowledgment

This work was supported by the National Key R&D Program of China under grant 2017YFB1002804.

References

[1] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, pages 737–737, 1994.

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

[8] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020.

[9] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2019.

[10] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[12] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2393–2402, 2018.

[13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[15] Jun Fu, Jing Liu, Yuhang Wang, Yong Li, Yongjun Bao, Jinhui Tang, and Hanqing Lu. Adaptive context network for scene parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6748–6757, 2019.

[16] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[17] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7519–7528, 2019.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. IEEE, 2020.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Hanzhe Hu, Deyi Ji, Weihao Gan, Shuai Bai, Wei Wu, and Junjie Yan. Class-wise dynamic graph convolution for semantic segmentation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, volume 12362 of *Lecture Notes in Computer Science*, pages 1–17. Springer, 2020.

[21] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. *arXiv preprint arXiv:1811.11721*, 2018.

[22] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7334–7344, 2019.

[23] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.

[24] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X Yu. Adaptive affinity fields for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 587–602, 2018.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural net-

works. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[26] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016.

[27] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[28] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 752–761, 2018.

[29] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[31] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[32] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[34] Bing Shuai, Zhen Zuo, Bing Wang, and Gang Wang. Scene segmentation with dag-recurrent neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1480–1493, 2017.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[38] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[39] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *CoRR*, abs/2101.11939, 2021.

[40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[41] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. *CoRR*, abs/2011.09157, 2020.

[42] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3733–3742. IEEE Computer Society, 2018.

[43] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12416–12425, 2020.

[44] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1857–1866, 2018.

[45] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[46] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019.

[47] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding. Acfnet: Attentional class feature network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6798–6807, 2019.

[48] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.

[49] Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 548–557, 2019.

[50] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2031–2039, 2017.

[51] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[52] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[53] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018.

[54] Xiangyun Zhao, Raviteja Vemulapalli, Philip Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label-efficient semantic segmentation. *CoRR*, abs/2012.06985, 2020.

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.