The Name of the Title is Hope

Ben Trovato*
G.K.M. Tobin*
trovato@corporation.com
webmaster@marysville-ohio.com
Institute for Clarity in Documentation
Dublin, Ohio, USA

Lars Thørväld
The Thørväld Group
Hekla, Iceland
larst@affiliation.org

Valerie Béranger
Inria Paris-Rocquencourt
Rocquencourt, France

Aparna Patel
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India

Huifen Chan
Tsinghua University
Haidian Qu, Beijing Shi, China

Charles Palmer
Palmer Research Laboratories
San Antonio, Texas, USA
cpalmer@prl.com

John Smith
The Thørväld Group
Hekla, Iceland
jsmith@affiliation.org

Julius P. Kumquat
The Kumquat Consortium
New York, USA
jpkumquat@consortium.net

ABSTRACT

A clear and well-documented LaTeX document is presented as an article formatted for publication by ACM in a conference proceedings or journal publication. Based on the "acmart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Ben Trovato, G.K.M. Tobin, Lars Thørväld, Valerie Béranger, Aparna Patel, Huifen Chan, Charles Palmer, John Smith, and Julius P. Kumquat. 2018. The Name of the Title is Hope. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Semantic segmentation is a crucial task in computer vision, which aims to assign a semantic label to each pixel in an image. With the popularity of deep learning, semantic segmentation tasks have also made a lot of progress. More and more application scenarios in reality need to infer relevant knowledge or semantics from images (that is, the process from concrete to abstract). Examples of applications are autonomous driving and environment understanding[6][43], biomedical analyses [33], remote sensing[3], robot manipulation[31] and further computer visions tasks. A key ingredient to the success of deep learning is the availability of large corpora of annotated training data. The performance of semantic segmentation is highly dependent on the availability and quality of labeled data, which can be costly and time-consuming to obtain.

Given enough labeled data, incredibly good semantic segmentation systems can be trained using deep learning [66, 4, 45, 42, 12, 91, 63, 78]. However, obtaining pixel-wise labels for semantic segmentation is incredibly time-consuming and expensive. For the COCO dataset, this required over 85,000 annotator hours[24] and on average more than 1.5 hours of annotation and quality control was required for each image in the CITYSCAPES segmentation dataset[9].Our objective in this work is to reduce the annotation load for semantic segmentation tasks.

Active learning (AL) can help achieve this goal. Active learning is a machine learning paradigm that enables a model to iteratively identify the most informative data samples for annotation, thus improving its performance with reduced reliance on labeled data. By identifying the most informative samples and requesting annotations from human experts, active learning can effectively reduce the annotation cost while achieving similar or even better performance compared to conventional supervised learning methods.

In an active learning setup, a model is initially trained on a small labeled dataset. An acquisition function uses the trained model to identify the most informative samples from a pool of unlabeled data points, which are then labeled by an expert. The newly labeled samples are incorporated into the training dataset and a new model is trained using the updated training dataset. This iterative process continues until the desired level of performance is achieved or the predetermined annotation cost limit is reached.

*Both authors contributed equally to this research.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

The objective of this work is to propose a simple yet effective approach for training a good semantic segmentation model at minimal annotation cost.

However, selecting informative samples for semantic segmentation is challenging due to the high dimensionality of the image data and the complex correlations between pixels. In light of the characteristics of semantic segmentation, we raise the following inquiries. How to fully leverage the information within individual image pixels and between adjacent pixels; how to effectively extract feature information from high-dimensional spaces; and how to minimize labeling costs. To address this issue, we propose a novel active learning method for semantic segmentation that utilizes superpixel-based contrastive learning to select informative samples.

During the process of model training, our proposed method extracts feature vectors from individual pixels using a contrastive learning approach, which maximizes the similarity between pixels belonging to the same class while minimizing the similarity between pixels from different classes. Due to the fact that semantic segmentation generates a feature vector for each pixel, a large number of feature vectors are produced in each iteration. We have employed the memorybank strategy to selectively store the most informative feature vectors for later usage in order to overcome this. This method has the advantage of not only avoiding exorbitant storage costs, but also enabling the preservation of previous data, which improves the model's robustness.

During the process of active learning sample selection. By using a pre-trained model to make predictions for each pixel, we can obtain the corresponding feature vectors and then calculate the feature vector for each superpixel region. The extracted feature vectors are then reduced in dimension using the Uniform Manifold Approximation and Projection (UMAP)[28] method, which is a non-linear dimensionality reduction technique that preserves the local structure of the data. Finally, the reduced feature vectors are clustered using the KMeans algorithm, and the most informative samples are selected for annotation.

The proposed method has several advantages over existing active learning methods. First, by extracting features from individual pixels within superpixels, we can capture the fine-grained details and the spatial context of the image data, which can improve the selection of informative samples. Second, by reducing the dimensionality of the feature vectors using UMAP, we can effectively remove the redundant and irrelevant features, which can improve the clustering performance. Third, by clustering the reduced feature vectors using KMeans, we can efficiently select the most informative samples without requiring a large computational cost.

We evaluate the proposed method on two widely used datasets for semantic segmentation, Camvid and Cityscapes, and compare our method with several state-of-the-art active learning methods. The experimental results demonstrate that our method can achieve comparable or even better performance with significantly fewer labeled data, thus reducing the annotation cost while maintaining high segmentation accuracy.

The contributions of this paper can be summarised as follows:

- We proposed a novel active learning approach based on contrastive learning for semantic segmentation. By utilizing contrastive learning, we aim to reinforce the similarity between pixel embeddings that belong to the same semantic class, while making them more dissimilar to those from different classes. This approach allows us to explicitly explore the structures of labeled pixels thereby facilitating the selection of samples for active learning.
- We update the UMAP algorithm iteratively during the training process to dynamically uncover high-level features of the data. By training the UMAP algorithm on the vectors prestored in memorybank, and applying it to the feature vectors of unlabeled superpixel blocks, we obtained the high-level feature space representation and the reduced-dimensional coordinates of each superpixel block. We then apply the k-means algorithm to cluster the reduced coordinates and select samples to label using an information-based approach.
- We conducted experiments on two datasets, CamVid and Cityscapes, and compared our proposed active learning method with previous state-of-the-art methods. The results showed that our method achieved state-of-the-art performance, reaching up to 95% of the performance of full-data training with only 8% of the dataset.

2 RELATED WORK

2.1 Active learning

Existing method in active learning can broadly be categorized into 3 branches from the perspective of querying strategy: uncertainty-based, diversity-based and combined strategies.

Uncertainty-based methods aim to detect the unlabeled samples that pose the greatest challenge to the current model, which has been trained on the labeled set according to the target objective function. These methods are designed to identify the most ambiguous data points, in order to improve the overall accuracy and robustness of the model. Maximum Entropy[36] selects data that maximize the predictive entropy. Margin[29] selects data whose two most likely labels have smallest difference in posterior probabilities. Least Confidence[42] selects data whose most likely label has lowest posterior probability. Bayesian Active Learning by Disagreements (BALD)[12][16] chooses data points that are expected to maximize the information gained from the model parameters. The assumption here is that having these uncertain samples labelled will add the most value to the next model training round.

The use of diversity-based strategies in batch selection entails the selection of samples that are representative of the unlabeled set, guided by the assumption that labeled representatives can serve as a surrogate for the entire dataset.[34] define active learning tasks as core set selection problems, sequentially selecting budget samples from the unlabeled data to add to the set s and the newly added points u needs to satisfy the distance to the set s and the set s. Variational Adversarial AL[39] learns a distribution of labeled data in latent space using a VAE and an adversarial network trained to discriminate between unlabeled and labeled data. The max-min game between VAE and the adversarial network works like this: VAE tries to trick the adversarial network into predicting that all

data points are from the labelled pool; the adversarial network learns how to distinguish dissimilarities in the potential space.

Generative Adversarial Active Learning (GAAL)[53] synthesizes queries via Generative Adversarial Networks (GANs). In contrast to regular AL that selects points from the unlabeled data pool, GAAL generates images from GAN for querying human annotators. [49] suggests incorporating a loss prediction module into the target network, which determines the top K most valuable data (i.e., the K most lossy data) depending on the loss value of the model. [50] is an improved version of VAAL. In VAAL, the discriminator is trained with only two states, labeled/unlabeled, and the authors of SRAAL argue that this ignores the information that sometimes the task model can already predict an unlabeled sample with a high degree of confidence, and that it should lower the priority of selecting that sample. To implement this idea, the authors give a function for calculating the prediction uncertainty of the task model, and use the output of this function as a label for unlabeled samples during the training of the discriminator of the generative adversarial network, instead of simply a binary variable. [18] proposes Temporal Output Discrepancy (TOD), which is the difference in predictions given by the model across iteration processes. Their theoretical work suggests that TOD is a lower bound on cumulative sample loss and can therefore be used to select information-rich unlabeled samples.

2.2 Active learning for semantic segmentation

The issue of active learning is less studied in the field of semantic segmentation than it is in image classification. Based on the level of sample granularity, AL for semantic segmentation can be divided into image-level approaches and region-level approaches. Whereas the latter separates an image into non-overlapping patches and treats each patch as a sample, the former treats the full image as a sample.

For image-level approaches. [48] selected a group of samples that were the most representative and uncertain to annotate by using model predictions and feature descriptors generated from the trained CNN model. [10] also utilized Variational Autoencoder (VAE) to acquire knowledge of a latent space, which was then utilized for performing gradient-guided sampling. The Minimax Active Learning (MAL) framework [11]employs a discriminator to classify the most diverse samples in comparison to the labeled set. This discriminator is paired with class prototypes to identify the samples with the highest entropy. The Difficulty-awarE Active Learning (DEAL) architecture [46] extends the standard semantic segmentation framework by adding a probability attention branch. This enables the model to learn to prioritize pixels belonging to the same semantic category before calculating metrics for sample acquisition.

For region-level approaches. [27] introduced regional Monte-Carlo Dropout by fusing information extraction module with cost extraction module to select regions that were informative yet cheap to annotate. [5] proposes a new modification of the deep Q-network (DQN) formulation for active learning, that an agent learns a policy to select a subset of small informative image regions. The region selection decision is made based on predictions and uncertainties of the segmentation model being trained. [8] train a meta regression model to estimate the segment-wise Intersection over Union of each predicted segment of unlabeled images. [20] performed selection and annotation of data at the superpixel level and utilized CRF to improve the accuracy of the superpixel labels. [38] proposed a formulation for viewpoint entropy that leverages prediction consistency among different views in multi-view datasets. It also performs uncertainty computation and selection at the superpixel level. Recently, PixelPick [37]significantly reduces the cost of labeling by training networks solely with sparse pixel annotations. EquAL[13] combines a straightforward but efficient semi-supervised technique. It integrates self-consistency on the image and its horizontally flipped equivalent. The labeling of regions within images from the unlabeled pool is then queued up using the same constraint as the acquisition metric.

The closest approach to ours is MEAL [40]. MEAL formulates the problem within an exploration-exploitation framework by combining a Uniform Manifold Approximation-based embedding to model representativeness, with entropy as an uncertainty measure to model informativeness. However, MEAL requires unsupervised pre-training to obtain the feature vectors for the image patches, which is computationally intensive and time consuming for new datasets. Furthermore, the resulting feature vectors and model are task-agnostic and cannot be updated with model iterations. Our approach iteratively updates the task-specific features that fit the model and the task through contrastive learning, which is then used to train the UMAP algorithm. We prioritize sample diversity by initially clustering samples, thereby avoiding the potential sacrifice of diversity in favor of informative samples. This approach allows us to tailor the features to the specific model and task while maintaining the diversity of the samples.

2.3 Contrastive learning for semantic segmentation

Contrastive learning explores a similarity function that brings similar data closer together while pushing dissimilar data further apart in the representation space[25]. Numerous outstanding contrastive learning methods exist for semantic segmentation, each with different design strategies that primarily focus on the selection of feature extraction techniques and loss design. [1, 17, 23, 44] focus on constructing high-quality and representative feature vectors for contrastive learning. [1, 21, 51] design and use different loss to improve performance. As for semi-supervised semantic segmentation, more attention is paid to the quality of pseudo-labels. [47, 51, 52] work to improve the quality of pseudo-labels for self-training using contrastive learning. [45] employs unreliable pseudo-labels as negative samples in contrastive learning to make sufficient use of unlabeled data. In addition, [51] performs contrastive learning via pre-training by using a pixel-wise, label-based contrastive loss.

3 METHOD

In this section, we present the details of our proposed method. Our training procedure is shown in Algorithm 1. We first go through how to combine contrastive learning with semantic segmentation networks to enhance network performance and the precision of feature vectors in 3.2. Then, we introduce a dimensionality reduction and clustering method based on UMAP in 3.3

3.1 Preliminaries

We consider this problem as finding the most significant and illustrative subset of data by iteratively selecting samples from a given partition for annotation by an oracle or a human expert. Before presenting our method, we divide images into superpixels as the fundamental sample unit of the labeled/unlabeled set. Given an annotation budget b at each active learning cycle, 1) we first select a small number of the most informative and representative superpixels from the unlabeled set $D_u$and annotate them until the annotation budget is exhausted; 2) We use the labeled superpixels to train our semantic segmentation model M_s; 3) Repeat 1) and 2) until the model achieves the performance of 95% fully supervised learning. To summarize, our aim is to minimize the annotation effort required to attain a certain level of segmentation performance by sampling and labeling the most informative and representative superpixels.

3.2 Semantic Segmentation combing Contrastive Learning

Pixel-Wise Loss. In the context of semantic segmentation, each pixel i of an image I has to be classified into a semantic class $c \in C$. For cross-entropy Loss, we cast this task as a pixel-wise classification problem. Specifically, let f_{FCN} be an FCN encoder, that produces a dense feature $I \in \mathbb{R}^{H\times W\times D}$ for I. Then a segmentation head f_{SEG} maps I into a categorical score map $Y = f_{SEG}(I) \in \mathbb{R}^{H\times W\times C}$. Further let $y = [y_1, , y_C] \in R_C$be the unnormalized score vector(termed as logit) for pixel i, derived from Y. The pixel-wise cross-entropy loss is optimized as Eq.1.

$$\mathcal{L}_i(y_i, \hat{y_i}) = -\sum_{c=1}^{C} y_{ic} \log \hat{y}_{ic} \tag{1}$$

Here, C is the number of classes, y_{ic} represents whether the i-th pixel belongs to class c, and $\hat{y}_{ic}$ represents the predicted probability of pixel i belonging to class c.

To compute the overall cross-entropy loss for the labels dataset D_l, we need to average the loss of each pixel over all the pixels in the image as Eq.2:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} \log \hat{y}_{ic} \tag{2}$$

Here, N is the total number of labeled pixels in the image.

Feature-Wise Loss. In this work, we developed a feature-wise contrastive learning approach to regularize the embedding space and explore the global patterns of the training data. Through optimization and learning in the feature space, it can complement the pixel-wise cross-entropy loss that only takes into account individual pixels. In addition, for a pixel i with its groundtruth semantic label $\bar{c}$, the positive samples are other pixels also belonging to the class $\bar{c}$, while the negatives are the pixels belonging to the other classes. We used a popular loss function for contrastive learning, called InfoNCE [30] as Eq.3.

$$\mathcal{L}_{NCE} = -\frac{1}{N}\sum_{i^+\in P} \log \frac{\exp(i \cdot i^+)/\tau)}{\exp(i \cdot i^+)/\tau) + \sum_{i^-\in N} \exp(i \cdot i^-/\tau)} \tag{3}$$

Here, N is the size of the labeled set, and τ is a temperature parameter. P and N denote pixel embedding collections of the positive and negative samples, respectively, for pixel i. note that the positive and negative samples, as well as the anchor i, are not necessarily required to be sourced from the same image.

As Eq.3 shows, the aim of designing a loss function based on pixel-to-pixel contrast is to facilitate the learning of an embedding space, where pixels belonging to the same class are brought closer together, while those from different classes are pushed further apart.

Memory Bank Design. As revealed by recent studies [7, 14], a large set of negatives is critical in contrastive representation learning. Since the number of negatives is constrained by the mini-batch size, modern contrastive approaches employ vast external memories to serve as a repository for storing additional navigable samples.

To this end, we choose to keep a separate pixel queue for each category. From each image in the latest batch, only a limited set of distinctive features are selected for each category and added to the corresponding queue. Specifically, for a segmentation dataset with a total of N training images, C semantic classes, the size of the whole memory is $CxTxD$, where D is the dimension of pixel features, T represents the total number of features that are stored in each category.

Given the memory bank's limited capacity, it is necessary to prioritize the selection of the most informative and representative features. As it complements the cross-entropy loss, we regard the pixel entropy[35] as an indicator of sample difficulty as Eq.4. The pixel entropy effectively measures the quantity of information held by the pixel, whereby higher entropy values indicate greater uncertainty about its class and informational content. Conversely, during the initial stages of network training, suboptimal classification can lead to inadequate feature representation. Considering the interdependent relationship between the classification and representation heads of the network, the accurate classification of selected features is paramount from the outset. Specifically, at each iteration, we add the top-K features with both correct classification and highest entropy for each category into the memory bank.

$$H(x) = -\sum_{c=1}^{C} p_c(x) \log_2 p_c(x) \tag{4}$$

where x is the output of the network prediction after softmax, $p_c(x)$ is the probability that pixel x belongs to category c, and C is the number of categories.

Challenging Instance Selection. Previous studies [19, 21, 32] have indicated that, beyond loss design and sample size, the discriminative quality of the training data plays a critical role in metric learning. We utilize cosine distance between negative samples and anchor as an indicator of challenging samples. We view the negatives with cosine distance closer to 0 to be challenging, *i.e.*, negatives which are similar to the anchor i. Studies conducted by [19, 44] suggest that as the training advances, a larger proportion of negatives may become less informative and thus make an insignificant contribution to the contrastive loss. To remedy this problem, We adopt the method proposed by [44], for each anchor feature i, we select a total of K_n negative samples and initially gather the $K_n/2$ nearest negative features from the memory bank, from which

we then randomly select $K_n/2$ negatives for our contrastive loss calculation. We use the mean of its corresponding class in the memory bank as the positive sample for each anchor. For the choice of anchor, contrary to the assertion that accurate categorization is necessary for negative sample selection. In particular, we consider pixels that were incorrectly labeled as hard samples. During the selection process, we select a total of K_a anchors and initially gather the $K_n/2$ hard samples in the current batch, from which we then randomly select $K_n/2$ anchors. Our anchor sampling strategy enables contrastive learning to concentrate on pixels that are more challenging to classify at the level of feature space. And by doing so, we obtain features that are more conscious of the segmentation task.

By combining the advantages of pairwise metric loss and unary cross-entropy loss, our segmentation network can produce more distinct features and produce better results. Thus the total training loss is Eq.5.

$$\mathcal{L}_{Total} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{NCE} \tag{5}$$

As our dense prediction setting involves an extensive number of pixel samples, the majority of which are redundant, it's crucial to identify valid positive and negative samples, as well as anchors, to achieve optimal results.

3.3 Performing active learning with UMAP

superpixels segmentation Given the success of querying specific regions of an image, rather than the entire image, in active learning methods designed for semantic segmentation [5, 13], our approach is likewise designed to select regions of interest within images. Unlike rectangular regions, we opted to use superpixel areas, as they are better suited for preserving object boundaries and ensuring that pixels within the same superpixel tend to belong to the same class, thus facilitating subsequent region dimensionality reduction and clustering operations. And we employed the traditional superpixel segmentation method Seeds[41] as our region partitioning approach.

UMAP-based dimensionality reduction and clustering.. Uniform Manifold Approximation and Projection(UMAP)[28] is a non-linear technique used for downsizing high-dimensional data. It is capable of mapping high-dimensional data into lower dimensions while maintaining the local structure between the data points. UMAP stands out from other dimensionality reduction techniques due to its fast computation time and ability to preserve the local structure of the data while achieving a global layout.

We trained the UMAP using feature vectors saved in the memory bank. In this way, UMAP can learn the mapping from high-dimensional vectors to low-dimensional space. Given that the memory bank is constantly updated, the features contained therein are both highly informative and correctly predicted, while retaining historical information. Therefore, the UMAP trained on these features is more robust and better reflects the feature space of the corresponding class. And the UMAP algorithm is updated at each iteration of active learning.

During the process of selecting samples in active learning, first, we obtain the feature representation of each pixel in the unlabeled dataset through the feature representation head. Next, we obtain the feature representation of each region by averaging the feature representations of pixels within the same region based on the superpixel segmentation. As most pixels within the same superpixel region belong to the same class, the averaged vector can approximately represent the feature of that region. After obtaining the feature representations of superpixel regions, we further reduce their dimensionality using the trained UMAP algorithm to obtain the low-dimensional feature representations of superpixel regions.

As a result, we obtain a low-dimensional feature representation for each superpixel region in the unlabeled dataset. Then we employ the K-means algorithm[2] to cluster the low-dimensional feature space, with the number of clusters being equal to the number of samples to be selected in each iteration. By using this method of dimensionality reduction followed by clustering, we ensure that the selected samples are distributed across different clusters, which guarantees that each class is well-represented and that the selected samples are diverse.

Entropy-based uncertainty sampling After obtaining the clusters, we further select the samples to be labeled based on their information content or uncertainty. Here, we compute entropy as an uncertainty measure. We use the segmentation model which is trained using labeled dataset to compute the prediction of all images from the unlabeled dataset.

$$H(S) = \sum_{x \in S} H(x) \tag{6}$$

Here, S represents a superpixel region, x is pixels belonging to superpixel S.

Algorithm 1: Active learning for Semantic Segmentation

Input: Given data $D_t = D_l$ and D_u, where D_l is labeled, and D_u is unlabeled; Oracle O for providing labels on unlabeled data; Initialized model M_s, M_t; Budget B; Initialized dimensionality reduction method $UMAP$

Output: Trained M_s, M_t, labeled dataset D_l

Perform superpixel segmentation on the dataset to obtain superpixel regions. $b = 0$. Initialize $Memorybank$.

while $b < B$ **do**
- Obtain the feature f for each pixel through the model Mt.
- Obtain the feature F for each region by taking the average f of the pixels within each superpixel region.
- Reduce the dimensionality of F using $UMAP$, and then perform clustering using Kmeans.
- Select a group $b \subset D_u$ with the maximum entropy in each cluster for labeling
- Query oracle O to label dataset b
- $D_l = D_l - b$, $D_u = D_u + b$
- Compute L_{CE} by using Eq.1
- Select anchors i, positive samples P and negative samples N
- Update $Memorybank$
- Compute L_{Contra} by using EQ.2
- $L_{Total} = L_{CE} + \alpha x L_{Contra}$
- Update the model M_s on Q through L_{Total}
- $M_t = \alpha M_{t-1} + (1 - \alpha) M_s$
- Update $UMAP$ through $Memorybank$

end while

4 EXPERIMENTS

We conducted a series of experiments to evaluate the effectiveness of our proposed method in comparison with several active learning strategies reported in the literature. The evaluation criterion we use is the ratio of mean Intersection over Union (mIoU) obtained by training on the selected data to that obtained by training on the entire dataset.

4.1 Datasets

We conducted experiments using the publicly available semantic segmentation datasets CamVid [4] and Cityscapes [9]. To simulate the acquisition process, instead of obtaining real annotations from an oracle, we masked the already available labels of the query dataset and only revealed them once the AL algorithm selects those patches.

CamVid [4] is a dataset consists of street scene view images that are sized 360 × 480 and annotated with 11 semantic classes. It consists of 367 training, 101 validation, and 233 testing images with pixel-level annotations, such as road, building, sky, car, and pedestrian.

Cityscapes [9] is a widely used urban semantic segmentation dataset that consists of high-quality images captured from various cities. The dataset includes 5,000 finely annotated images of dimension 1024 × 2048 with 19 semantic classes. It is split into three subsets, with 2,975 images in the training set, 500 images in the validation set, and 1,524 images in the test set. Given its large size, we resized the images to 512 × 1024 during the training process.

4.2 Implementation Details

Network Module: We employ the same segmentation network (ResNet50 [15] + FCN [26]) across all the methods. We follow the experimental setup of [13]. We apply random horizontal flipping as data augmentation. During the final retraining stage, we run 60 epochs with a batch size of five for CamVid and four for Cityscapes. The adaptive moment estimation optimizer (ADAM) [22] with β_1 = 0.9, β_2 = 0.999 is used for optimization, with the initial learning rate and weight decay set to 5e-4 and 2e-4, respectively.

Project Head: It projects each high-dimensional pixel embedding onto a 128-dimensional feature space, facilitating the computation of the contrastive loss. The linear classification module is composed of a 3x3 convolutional layer, followed by a batch normalization step, a rectified linear unit (ReLU) activation function, and a dropout operation with a probability of 0.1. Lastly, a 1x1 convolutional layer is applied. Note that the project head is exclusively employed during the training phase and excluded during the inference phase.

Memry Bnak: For each training image, we sample P_i = 10 pixels per class. For each class, we set the size of the pixel queue as P_t = $100 \times P_i$. The memory bank is also discarded after training.

Total Loss $\mathcal{L}_{Total}$: We empirically set the coefficient λ to 0.1. For $\mathcal{L}_{NCE}$ in Eq.3, we set the temperature τ as 0.1. We set the number of sampled instances to 512 and 1024 for positive and negative, respectively, during sampling. During each mini-batch, we sample 50 anchors per category, half of which are randomly sampled while the other half are segmentation-hard ones.

4.3 Experimental Results

4.4 Ablation Study

REFERENCES

[1] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. 2021. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8219–8228.

[2] David Arthur and Sergei Vassilvitskii. 2007. K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 1027–1035.

[3] Seyed Majid Azimi, Corentin Henry, Lars Sommer, Arne Schumann, and Eleonora Vig. 2019. Skyscapes fine-grained semantic understanding of aerial scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7393–7403.

[4] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. 2009. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters* 30, 2 (2009), 88–97.

[5] Arantxa Casanova, Pedro O Pinheiro, Negar Rostamzadeh, and Christopher J Pal. 2020. Reinforced active learning for image segmentation. *arXiv preprint arXiv:2002.06583* (2020).

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

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

[8] Pascal Colling, Lutz Roese-Koerner, Hanno Gottschalk, and Matthias Rottmann. 2020. Metabox+: A new region based active learning method for semantic segmentation using priority maps. *arXiv preprint arXiv:2010.01884* (2020).

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.

[10] Chengliang Dai, Shuo Wang, Yuanhan Mo, Kaichen Zhou, Elsa Angelini, Yike Guo, and Wenjia Bai. 2020. Suggestive annotation of brain tumour images with gradient-guided sampling. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*. Springer, 156–165.

[11] Sayna Ebrahimi, William Gan, Dian Chen, Giscard Biamby, Kamyar Salahi, Michael Laielli, Shizhan Zhu, and Trevor Darrell. 2020. Minimax active learning. *arXiv preprint arXiv:2012.10467* (2020).

[12] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*. PMLR, 1183–1192.

[13] S Alireza Golestaneh and Kris M Kitani. 2020. Importance of self-consistency in active learning for semantic segmentation. *arXiv preprint arXiv:2008.01860* (2020).

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

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

[16] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745* (2011).

[17] Hanzhe Hu, Jinshi Cui, and Liwei Wang. 2021. Region-aware contrastive learning for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16291–16301.

[18] Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan, and Dejing Dou. 2021. Semi-supervised active learning with temporal output discrepancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3447–3456.

[19] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems* 33 (2020), 21798–21809.

[20] Tejaswi Kasarla, Gattigorla Nagendar, Guruprasad M Hegde, Vineeth Balasubramanian, and CV Jawahar. 2019. Region-based active learning for efficient labeling in semantic segmentation. In *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1109–1117.

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems* 33 (2020), 18661–18673.

[22] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[23] Junnan Li, Caiming Xiong, and Steven CH Hoi. 2021. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9475–9484.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[25] Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew J Davison. 2021. Bootstrapping semantic segmentation with regional contrast. *arXiv preprint arXiv:2104.04465* (2021).

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

[27] Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. 2018. Cereals-cost-effective region-based active learning for semantic segmentation. *arXiv preprint arXiv:1810.09726* (2018).

[28] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[29] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. (2011).

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[31] En Yen Puang, Peter Lehner, Zoltan-Csaba Marton, Maximilian Durner, Rudolph Triebel, and Alin Albu-Schäffer. 2019. Visual repetition sampling for robot manipulation planning. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 9236–9242.

[32] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592* (2020).

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

[34] Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489* (2017).

[35] Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.

[36] Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review* 5, 1 (2001), 3–55.

[37] Gyungin Shin, Weidi Xie, and Samuel Albanie. 2021. All you need are a few pixels: semantic segmentation with pixelpick. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1687–1697.

[38] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. 2020. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9433–9443.

[39] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. 2019. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5972–5981.

[40] Deepthi Sreenivasaiah, Johannes Otterbach, and Thomas Wollmann. 2021. Meal: Manifold embedding-based active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1029–1037.

[41] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin de Capitani, and Luc Van Gool. 2012. SEEDS: Superpixels extracted via energy-driven sampling. *ECCV (7)* 7578 (2012), 13–26.

[42] Dan Wang and Yi Shang. 2014. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*. IEEE, 112–119.

[43] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* 43, 10 (2020), 3349–3364.

[44] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. 2021. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7303–7313.

[45] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. 2022. Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4248–4257.

[46] Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. 2020. Deal: Difficulty-aware active learning for semantic segmentation. In *Proceedings of the Asian conference on computer vision*.

[47] Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. 2022. Class-Aware Contrastive Semi-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14421–14430.

[48] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. Springer, 399–407.

[49] Donggeun Yoo and In So Kweon. 2019. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 93–102.

[50] Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. 2020. State-relabeling adversarial active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8756–8765.

[51] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. 2021. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10623–10633.

[52] Zhen Zhao, Luping Zhou, Lei Wang, Yinghuan Shi, and Yang Gao. 2022. LaSSL: Label-guided Self-training for Semi-supervised Learning. (2022).

[53] Jia-Jie Zhu and José Bento. 2017. Generative adversarial active learning. *arXiv preprint arXiv:1702.07956* (2017).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009