Region-Based Active Learning for Efficient Labeling in Semantic Segmentation

Tejaswi Kasarla[1], G Nagendar[1], Guruprasad M. Hegde[2], V. Balasubramanian[3] and C.V Jawahar[1]
[1]Center for Visual Information Technology, IIIT Hyderabad
[2]Bosch Research and Technology Centre, India
[3] Indian Institute of Technology, Hyderabad, India

Abstract

As vision-based autonomous systems, such as self-driving vehicles, become a reality, there is an increasing need for large annotated datasets for developing solutions to vision tasks. One important task that has seen significant interest in recent years is semantic segmentation. However, the cost of annotating every pixel for semantic segmentation is immense, and can be prohibitive in scaling to various settings and locations. In this paper, we propose a region-based active learning method for efficient labeling in semantic segmentation. Using the proposed active learning strategy, we show that we are able to judiciously select the regions for annotation such that we obtain 93.8% of the baseline performance (when all pixels are labeled) with labeling of 10% of the total number of pixels. Further, we show that this approach can be used to transfer annotations from a model trained on a given dataset (Cityscapes) to a different dataset (Mapillary), thus highlighting its promise and potential.

1. Introduction

The need for large, annotated datasets has seen increasing significance with the growing capabilities of deep neural network models in solving real-world vision tasks, and their subsequent absorption in usable technologies. Autonomous systems, such as self-driving vehicles, have further underscored this need, in order to scale to various geographies and settings around the world. Existing efforts to create large-scale vision datasets have largely been localized, and there is a need to provide better methods to create datasets more efficiently. This work is an effort in this direction.

Among vision tasks, semantic segmentation has recently attracted a lot of attention, where the objective is to classify each pixel into its corresponding semantic class. Fully supervised methods for this problem require annotation of all pixels of all given images/videos. This is a tedious task, and requires a huge amount of time and need human effort [24]. While semi-supervised methods (where part of the data used to train a machine learning model is unlabeled) can be used to offset the annotation load, fully supervised methods have continued to maintain state-of-the-art performance on tasks, especially in automation. Our objective in this work is to reduce the annotation load for semantic segmentation tasks.

Active learning methods have been known in machine learning for a couple of decades now. They operate in a setting where given a learned model, a learning algorithm chooses data points intelligently (using the given model) and subsequently queries an oracle for the labels of the chosen data points. The model is then updated using the newly obtained labeled data, with a view to increase the model's performance on unseen data. There have been limited related efforts in the past on active learning in semantic segmentation [27, 28]; however, their focus was different and algorithm-driven, and did not show tangible improvements in dataset development for real-world tasks. We instead focus on using active learning to contribute to dataset development for tasks based on contemporary datasets such as Cityscapes and Mapillary, with the main objective to reduce annotation cost on unlabeled data.

In this paper, we propose an entropy-based active learning approach for efficient labeling in semantic segmentation. While entropy has been extensively used for active learning in the past [13] for classification [29], we use this exclusively for semantic segmentation in this work. Furthermore, we propose a region-based active learning methodology, where annotation at the superpixel level in images, along with the use of fully connected Conditional Random Fields (CRF) [17] for label propagation, provides significant benefits in reducing annotation cost.

Our key contributions can be summarized as follows. Convolutional Neural Networks (CNNs) have shown to be very effective for the semantic segmentation task in recent years [3, 19, 31, 30]. To the best of our knowledge, none of the earlier methods for active learning in semantic segmentation were based on CNNs, and this is the first such effort. Recently, fully connected Conditional Random Fields have been used with CNNs [6] for improving the model performance in semantic segmentation task too. We lever-

age this development in our work to use the dense connectivity in fully connected CRF at the pixel level, along with CNN-based models, to achieve semantic segmentation with limited labeling effort. We apply our proposed method for the road scene understanding problem, which has significant applications in autonomous driving. We evaluate our method on the Cityscapes dataset, and our approach achieves 93.4% percent of the accuracy of the corresponding fully annotated model while querying just 10% of the pixel labels. We also demonstrate the effect of transfer learning over the Mapillary dataset, where the initial model is learned on the Cityscapes data, and this model is used to label the Mapillary dataset with minimal annotation effort.

The remainder of this paper is organized as follows. Section 2 briefly reviews related efforts on active learning methods and region-based segmentation. The proposed active learning methodology and our strategies for entropy-based and region-based active learning are discussed in Section 3. Experiments and results are discussed in Section 4, followed by concluding remarks in Section 5.

2. Related Work

Active learning methods have been proposed for many years now, and many criteria have been proposed for selecting the most informative data points to query for labels from an oracle. In [15, 18], uncertainty-based measures are used for choosing the queries. In [27], expected change (EC) in the labeling is used for selecting the informative data points, based on which points induce the largest expected change in the current model. In [16], Epistemic and Aleatoric uncertainty measures are studied in Bayesian deep learning models for vision tasks. The epistemic uncertainty measure is used for active learning in [9]. In [9], a Cost effective Active Learning approach using dropout at test time as Monte Carlo sampling is proposed to model the pixel-wise uncertainty. It estimates the uncertainty based on the stability of the pixel-wise predictions when a dropout is applied to a deep neural network. An earlier method proposed in [26] achieves semantic segmentation by active learning query for foreground object and Sparse Gaussian Process to obtain segmentation. The other commonly used active learning methods include query-by-committee [25], expected error reduction [10, 22], expected model change [8], variance reduction [11, 23] and Min-max view active learning [12, 14]. The proposed method can be viewed as an uncertainty-based sampling method.

Region-based methods [2, 4] have been popular for image segmentation in the past, and have only recently been integrated into deep neural network models for semantic segmentation. In typical region-based methods, the image is first divided into small coherent regions, until some stopping criterion is satisfied. These small regions form seed regions for the given image, and can capture complete objects or its canonical parts. Labels over these seed regions are used to classify its neighboring pixels/regions. We leverage this approach for label propagation in active learning. We now present the proposed methodology.

3. Region Based Active Learning for Semantic Segmentation

Given a (small) labeled dataset and a deep learning model trained on this data, our objective is to reduce the cost of annotations on available unlabeled data, so as to obtain a new model which provides better validation accuracy than the original model. Let $X = P \cup Q$ be the given data, where P is a set of labeled data and Q is the set of unlabeled data. Let Θ be a deep learning model trained on the labeled data P. Our objective is to reduce the number of annotations on Q using the trained model Θ such that the model's performance (in terms of semantic segmentation) on the un-annotated remainder of Q improves. We now describe the general approach for active learning in semantic segmentation using uncertainty measures based on entropy.

Our overall active learning methodology starts after training the initial network Θ over the labeled data P. We present active learning strategies to reduce annotations at both image-level and pixel-level. At an image-level, using uncertainty-based measures (which we describe later in this section), we rank images in the unlabeled set Q for annotation. The most uncertain images in Q are then selected as candidate images for annotation. The images are then annotated by an oracle O in groups, i.e. the first group of images are the most uncertain images and these are selected first for annotation. Given a pixel x_i^j, where x_i is the unlabeled image from Q, we can query its label using a given oracle O. $O(x_i^j)$ gives the true label of j^{th} pixel in x_i i.e. x_i^j. We assume group size to be l and the number of groups to be b.

At a pixel-level, for each candidate image in Q, we identify the most uncertain pixels for annotations. For a given image x_i, our objective is to find a given $m \in [0, 1]$ portion of candidate pixels for annotations, i.e. for the image x_i, where $|x_i|$ is the number of pixels, we only annotate $m|x_i|$ number of pixels. These pixels are identified using uncertainty measures computed at a pixel-level (described later in this section). The oracle O is then queried for the true labels of the identified pixels. Finally, the given network represented by Θ is retrained on the selected data B using the annotations obtained using the oracle O.

The overview of our proposed active learning method is summarized in Algorithm 1. In the final step, the model is retrained using the selected data and the labels obtained from O, to obtain the updated model Θ. We also update the current unlabeled set Q as $Q - B$. The new retrained model and the updated unlabeled set are further provided as input to the next iteration of the algorithm, and the process is continued for the given number of groups b. The uncertainty at

image and pixel level is computed using entropy measures, described in Section 3.1.

Algorithm 1 Active learning for Semantic Segmentation

Input: (i) Given data $X = P \cup Q$, where P is labeled, and Q is unlabeled; (ii) Oracle O for providing labels on unlabeled data; (iii) Deep learning model Θ trained on P; (iv) Proportion of annotations m on Q; (v) Number of groups b and group size l

Output: Updated model Θ

$i = 1$

while $i < b$ **do**

(1) Select a group $B \subset Q$, $|B| = l$ such that $uncertainty(B) > uncertainty(B^{'}) \forall B^{'} \subset Q, |B^{'}| = l$

(2) Query oracle O for relevant m portion of pixels of images in B

(3) Retrain the model Θ on B using the new labels obtained from Step (2) and update Θ

(4) $i = i + 1$; $Q = Q - B$

end while

Return updated model Θ

3.1. Computation of Uncertainty

We describe four different strategies for computing uncertainty, to be used for our active learning methodology described in Algorithm 1: *Pixel-level Entropy*, *Image-level Entropy*, *Edge Pixel-based Entropy*, and finally, the *Region-based Entropy* (which constitutes our proposed region-based active learning strategy and is the most effective of the proposed strategies).

3.1.1 Image-level Entropy:

The entropy at an image level is obtained by summing the uncertainties over all the pixels in the image x_i, as below:

$$H_i = \sum_{j=1}^{|x_i|} H_i^j \tag{1}$$

Entropy for an image x_i gives the uncertainty present in the prediction for the model Θ over the entire image x_i. In Step 1 of Algorithm 1, we compute the most uncertain images for the current model Θ (by ranking images in Q based on their respective entropies) and, using the oracle O, we only select a group of l images for annotation. We hence annotate a total of $l \times b$ number of images only in this strategy (which is much lesser than the size of Q in our experiments).

3.1.2 Pixel-level Entropy

Given an unlabeled image x_i, we compute its probability score map $p(c_k/x_i)$, where $k \in \{i, \ldots, C\}$ and C is the number of classes. This gives the probability scores $p(c_k/x_i^j)$ for $j \in \{i, \ldots, |x_i|\}$, where $|x_i|$ is the number of pixels in x_i. $p(c_k/x_i^j)$ gives the probability for the pixel x_i^j belonging to the class c_k. This probability distribution is obtained using the available deep learned model Θ, with no additional annotation effort on the unlabeled set. The entropy (uncertainty estimate), H_i^j at each pixel, is then computed using Eqn 2 below.

$$H_i^j = \sum_{k=1}^{C} p(c_k|x_i^j, \Theta) \log(p(c_k|x_i^j, \Theta)) \tag{2}$$

This entropy is computed for each image x_i separately. Our active learning methodology first ranks all images in the unlabeled set Q according to their entropies (as in Section 3.1.1), and the m portion of each image is then selected based on pixel-level entropy for labeling.

3.1.3 Edge Pixel-based Entropy

In general, the misclassification rate for pixels at object boundaries/edges is more when compared to the other pixels in the image. This suggests that edge pixels inherently have high uncertainty. However, not all edge pixels have high uncertainty like small edges inside an object. Also, few boundary pixels have a high chance of being misclassified, but their uncertainty is not as high as other uncertainty pixels. To consider edge pixels for annotation, we modify the pixel-level entropy strategy to give a higher weightage to edge pixels. We use a Canny edge detector to identify edge pixels, and the weighted entropy computed for edge pixels in a given image x_i is obtained as follows:

$$H_i = \sum_{k=1}^{|x_i|} \sum_{l=1}^{C} w_e p(c_l|x_i^k, \theta) \log(p(c_l|x_i^k, \theta)) \tag{3}$$

where $w_e > 1$ is the weight given to the edge pixels. For other pixels it is set to 1.

3.1.4 Region-based Entropy

In semantic segmentation, the neighboring pixels are highly likely to have a close relationship and share similar information. Therefore, they are likely to belong to the same semantic class. However, in all the aforementioned strategies, the entropy of each pixel is calculated independently without considering this relationship. In order to take advantage of the spatial correlation in images, we propose a region-based strategy that is applied at the level of superpixels (SP) in an image. We use (SLIC) [1] to computing the superpixels in a given image, and define the entropy at the superpixel level as the sum of its pixel entropies. To further leverage this strategy, we apply fully connected Conditional Random

Field (CRF) model over the segmentation output (probability map) obtained using a deep learned model. This gives the probability score maps for all the pixels. In other words, instead of computing the uncertainty measure over the probability obtained from the current network, we compute our uncertainty measure over the probability map obtained from CRF in this case. We find that this region-based strategy with the propagation obtained using a CRF is immensely useful in obtaining promising results with little annotation. Overview of the Region-based Entropy method is given in Figure 1. Here, we take a single image for illustrating each step. The final segmentation image is used for retraining the previously trained network.

3.1.5 Class Specific Selection of Pixels/SuperPixels

In general road scene images, few frequent classes like road, trees, and buildings dominate other classes like traffic signal and signboards. So the number of pixels considered for labeling in these classes dominates other classes. It is desirable to take equal number of pixels from each class to avoid the dominance of frequent classes. In this method, we first take the deep network trained on the initial labeled data. Now, we represent all the pixels in the labeled images using the feature vector obtained from the network. We construct a feature space using these feature vectors, where each category is denoted with a feature vector calculated using the feature vectors of the pixels in the same category. We represent each category with the mean of feature vectors of all the pixels in that category. Now, for the given unlabeled image, we calculate the similarity (cosine similarity) for each of its pixels with all the class means as follows,

$$sim^k_{x_i^j} = \frac{F_{x_i^j} m_k}{||F_{x_i^j}||\,||m_k||}$$

where $F_{x_i^j}$ is the feature vector for the pixel $x_i^{j\,th}$ obtained from the network and m_k is the feature vector representing k^{th} category. $sim^k_{x_i^j}$ gives the similarity between the $x_i^{j\,th}$ pixel and the feature vector representing the k^{th} category. Next, each pixel is assigned to its most similar category. In this method, instead of computing the entropy independent of its class, we compute the entropy of the pixels in each category separately and pick the high entropy pixels from each class independently.

4. Experimental Results

In this section, we evaluate our entropy-based active learning methods for semantic segmentation.

4.1. Datasets and Experimental settings

We evaluate our proposed method on two large widely used datasets for semantic segmentation: Cityscapes [7] and Mapillary [21]. Both Cityscapes and Mapillary datasets have emerged as the most popular choices for road scene understanding and autonomous driving. Cityscapes contains 2975 finely annotated training images along with 500 validation images. To evaluate the performance of our proposed method on Cityscapes, we divide the training data into 2 sets. The first set contains 1175 images and the second set contains 1800 images. A deep learning network is trained over the set containing 1175 images using original ground truth, and this model is used as the initial network. Using this model, we try to reduce the number of annotations required on the remaining 1800 images and the performance is evaluated over a hold-out validation dataset. We use ICNet [30] as our deep learning network. ICNet is a popular network for real-time semantic segmentation on high-resolution images. Since all the images in Cityscapes dataset are high-resolution images (1024×2048), we use ICNet as our deep learning network.

The Mapillary Vistas Dataset [21] is a large scale street-level image dataset. It contains 25000 high-resolution images annotated into 66 object categories. For this dataset, the network trained over cityscapes training data (2975 images) using original ground truth is considered as the initial model. For consistency in the results, we use 19 common classes in both cityscapes and mapillary datasets. The performance over these datasets is measured in terms of mean of class-wise intersection over union (mIoU).

For all the experiments, we take the number of groups as 6. The group size for cityscapes is 300 and for mapillary is 3000. For both cityscapes and mapillary data, we report the results on the validation data. We only select 10% of pixels for annotation. For computing superpixels (SP), in SLIC, we take the size of the superpixels as 1400. For training ICNet [30], we use SGD solver. For Cityscapes, we resized the training images to 512×1024 and while testing is done on the original images without resizing. On the other hand, for Mapillary dataset, we take the image size as 1920×1080. For both the datasets, we take the base learning rate as 0.01 and the poly learning rate policy is adopted with a power of 0.9. For cityscapes, we train the network for 30K iterations and for mapillay, we train it for 90k iterations. For both the datasets, we fix the momentum as 0.9, weight decay as 0.0001 and we take the batch size as 8. We use Caffe framework for the implementation of ICNet. For fully connected CRF, we take the similar settings as given in Deeplab [6]. We take the default values for $w_2(= 3)$ and $\sigma_y(= 3)$. The values for the parameters w_1, σ_α and σ_β are computed using cross-validation. For cross-validation, we use a small subset of 100 images.

4.2. Results on Cityscapes dataset

4.2.1 Image Level Annotations

First, we demonstrate how the incremental selection of images/groups is effective. In this experiment, we label all the

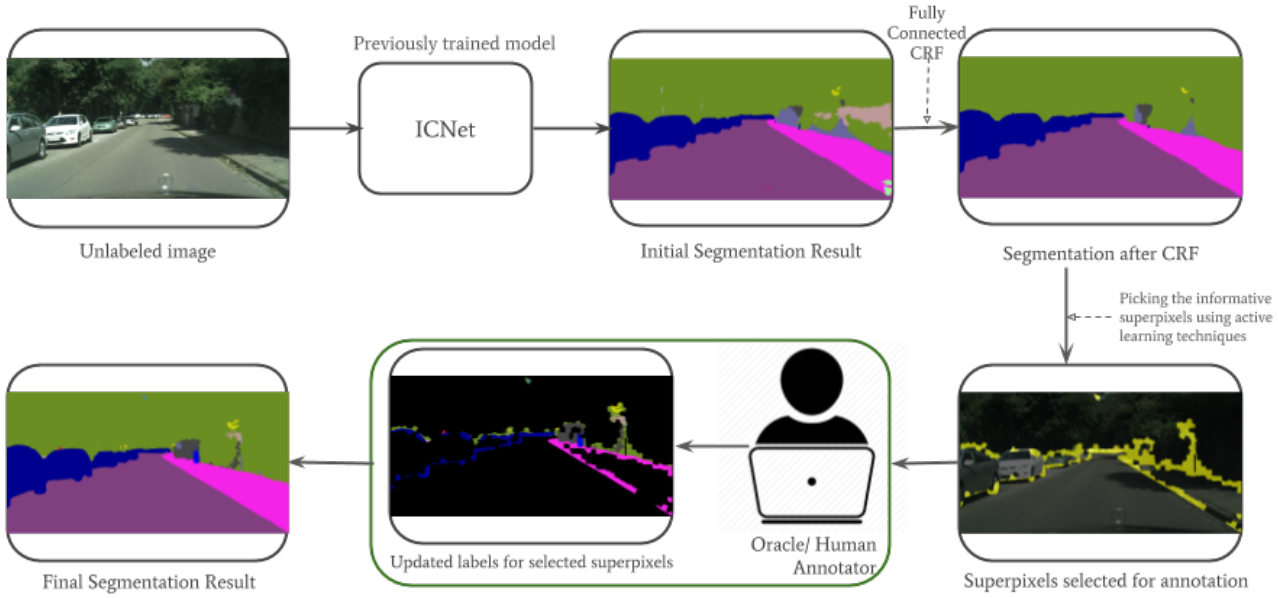

Figure 1. Overview of the Region-based Entropy method. Here, the final segmentation image is used for retraining the previously trained model.

	Baseline	Random 10% GT	Entropy	Entropy + Edge pixels	SP	SP + CRF	Class-specific SP+CRF
# Training images	100% GT	10% GT					
1175	55.6	55.6	55.6	55.6	55.6	55.6	55.6
1475	57.9	55.9 (96.5%)	56.1 (96.8%)	56.4 (97.4%)	56.5 (97.5%)	56.9 (98.2%)	**57.0 (98.4%)**
1775	59.7	56.2 (94.1%)	56.5 (94.6%)	57.0 (95.4%)	57.1 (95.6%)	57.8 (96.8%)	**57.9 (96.9%)**
2075	61.5	56.3 (91.5%)	56.9 (92.5%)	57.9 (94.1%)	58.0 (94.3%)	58.5 (95.1%)	**58.7 (95.4%)**
2375	62.7	56.5 (90.1%)	57.4 (91.5%)	58.7 (93.6%)	58.8 (93.7%)	59.4 (94.7%)	**59.7 (95.2%)**
2675	63.8	56.4 (88.4%)	57.8 (90.5%)	59.4 (93.1%)	59.3 (92.9%)	60.2 (94.3%)	**60.4 (95.2%)**
2975	65.3	56.5 (86.5%)	58.1 (88.9%)	59.8 (91.5%)	60.0 (91.8%)	61.0 (93.4%)	**61.3 (93.8%)**

Table 1. Comparison of our various active learning techniques over incremental training data for cityscapes data. Performance is given in meanIoU. Here, SP means SuperPixels and GT means Groundtruth. Results over 1175 images are obtained using initial trained network over full ground truth. The values in the bracket indicates the ratio to the baseline. It means the percentage of baseline accuracy the method is able achieve.

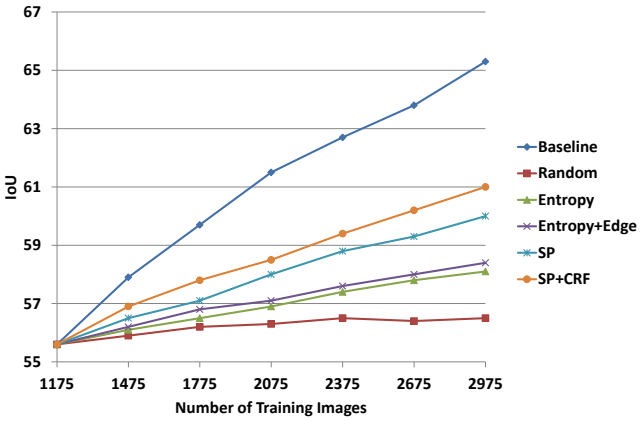

Figure 2. Performance of the proposed active learning methods over incremental selection of groups on cityscapes data.

pixels in the images using original ground truth (all the pixels in the images are annotated). We only try to reduce the number of images for annotation. We start with the model trained over 1175 images using the full ground truth of pixels. The results for group selection are given in Table 1. The results are given under "Baseline". Here, for each group, the performance is evaluated on the validation set. We also show the performance of the network over these incremental groups in Figure 2. From the figure (Baseline), we can observe a steep increase in IoU for the initial groups of images compared to the final groups. This is mainly due to the high uncertainty images present in the initial groups compared to the final groups. This means, the entropy-based selection allows us to identify the most informative images.

4.2.2 Pixels/Region Level Annotations

In all the "Pixels/Region Level Annotations" experiments, we only annotate 10% of the pixels in each image obtained from the corresponding methods and study their performance for the given segmentation task. For all other pixels, we take labels from the trained model.

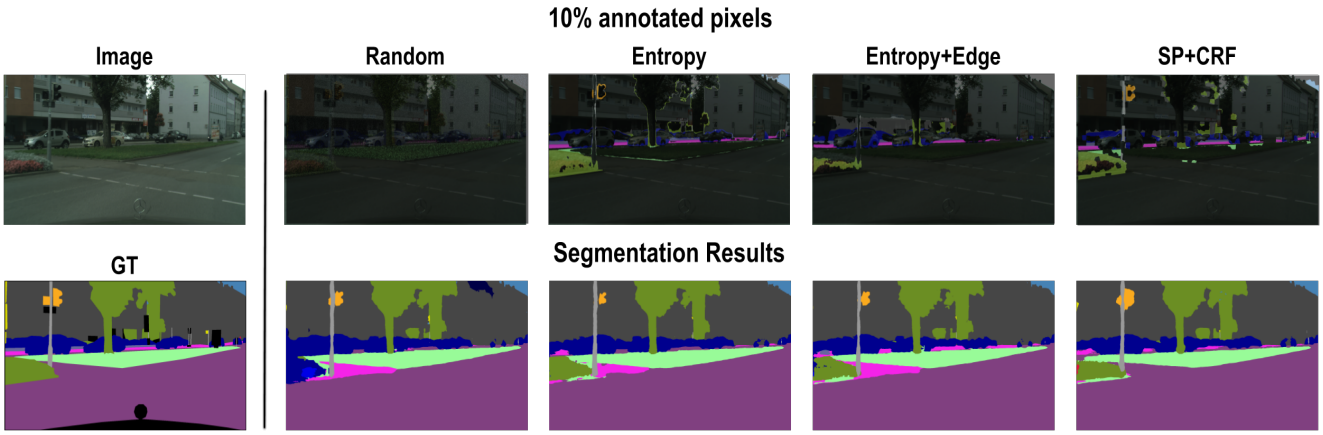

Figure 3. Semantic segmentation results on Cityscapes data. Here, we show the 10% of annotated pixels for different methods and their segmentation results. The first column is the original image and its ground truth (GT). In row 1, from column 2-5, shows the selected 10% of pixels in different methods for annotation. In row 2, from column 2-5, shows their corresponding segmentation results.

Random Selection of Pixels The results for random selection of pixel selection are given in Table 1 under Random 10% GT. In this experiment, we randomly select 10% of pixels for annotation. From the results, we can observe that this is not performing well compared to the baseline. Here, the results under the baseline are obtained using full ground truth (annotating all the pixels). The drop in the performance is high compared to the baseline. To get the better performance, we need to select the most uncertain pixels for annotation. However, in random selection, we are annotating both right prediction pixels as well as wrong prediction pixels for the current network.

Pixel-level Entropy based Selection In this experiment, instead of selecting random 10% of pixels for annotation in each image, we use the proposed pixel-level entropy for selecting the pixels for annotation. The results are given under Entropy in Table 1. From the results, we can observe that with only 10% of pixel annotations, the network is giving comparable results compared to the baseline. This suggests that our pixel-level entropy based active learning methods are able to pick the more informative pixels (high uncertain pixels). From the results, we can also observe that the pixel-level entropy based selection is outperforming the random selection of 10% of pixels for annotation. Using pixel-level entropy based selection, we are able to achieve 88.9% of baseline performance.

Edge-Pixels Based Selection In this experiment, while computing the entropy for each pixel, we give higher weightage to the edge pixels. The results are given under Entropy + Edge pixels in the Table 1. From the results we can observe that it further improved the performance of pixel-level entropy based selection. In our experiments, mostly the selected edge pixels are coming from the class boundaries. Few segmentation results for Edge based selection are given in Figure 3. From the results, we can observe that compared to the entropy based segmentation results, the class boundaries in Entropy + Edge method are improved. We can also observe that in Entropy+Edge selection, the most uncertain class boundaries are selected for annotation. Using Edge-Pixels based selection, we are able to achieve 89.4% of baseline performance.

Region based Selection In region-based selection, we conducted the experiments at the superpixel (SP) level. Here, we select the most uncertain super pixels for annotation. The results are given under SP (Superpixels) in Table 1. From the results, it can be observed that the uncertainty computed at the superpixel level is performing well compared to the pixel level. In superpixels, it forces the neighboring pixels which share the similar information to get the same class label. Using superpixels, we are able to achieve 91.8% of baseline accuracy. In the next experiment, we use CRF for improving the super pixel based selection. The results for CRF are given in Table 1. Here, we apply CRF at superpixel level. The results are given under SP+CRF. From the results, we can clearly observe that it improved the performance of SP based selection. Using our SP+CRF based selection, we are able to achieve 93.4% of baseline performance.

Class specific based Selection In this experiment, we select the 10% of pixels for annotation from each category. To evaluate its performance, we apply it over SP + CRF based method. The results are given in Table 1 under Class-specific SP + CRF. From the results, we can observe that it further improved the performance of SP + CRF based

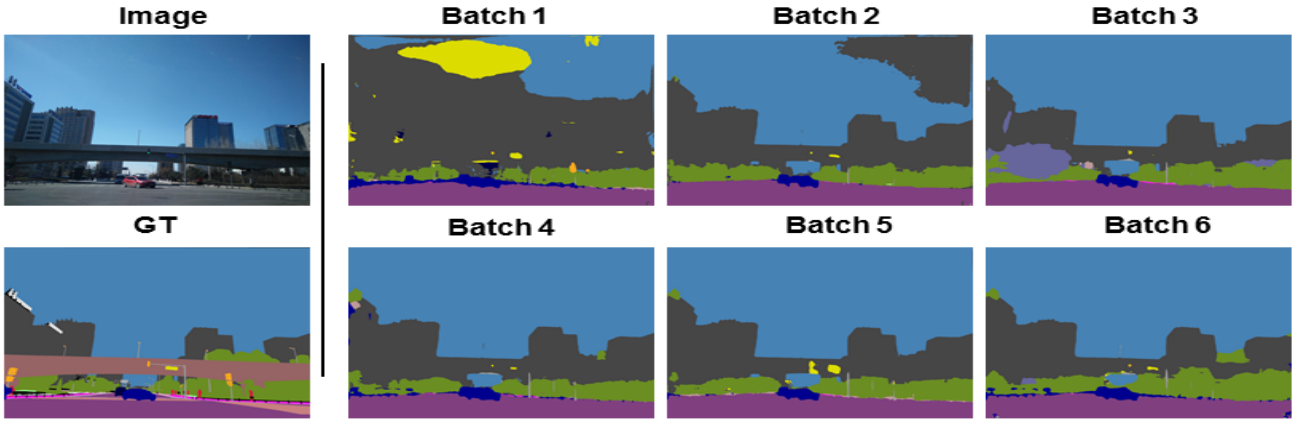

Figure 4. Semantic segmentation results on Mapillary data. Here, we show the segmentation results over incremental selection of groups. The first column is the input image and its ground truth (GT). In row 1, from column 2-4, shows the segmentation results over groups 1-3. In row 2, from columns 2-4, shows the segmentation results over groups 4-6.

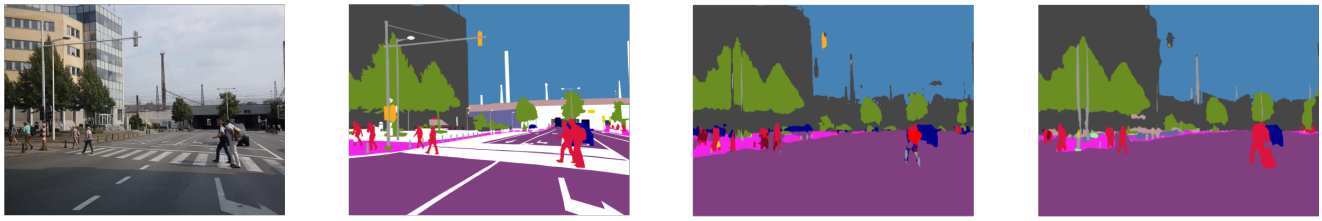

Image | Groundtruth | No additional labels | Using 10% labelling

Figure 5. Segmentation results using transfer learning on mapillary data. In No Additional labeling, the segmentation result is obtained from the network with out any annotations. In using 10% labeling, the segmentation result is obtained using our active learning method using 10% of pixel annotations.

method and it outperformed all other methods. Using this method, we are able to achieve 93.8% of baseline performance.

We also compared the performance of all the methods in Figure 2. From the Figure, we can observe that Class-specific SP+CRF based selection outperforms all other methods. We have also compared the performance of our active learning method over different percentages of pixel level annotations in Table 2. Here, we take SP+CRF method for demonstrating the results. From the Table, we can observe that the segmentation performance is improving with the addition of more annotations. Using 40% of annotations, we are able to achieve 97.4% of baseline performace. This also gives us the trade-off between the performance and annotation cost. Here, we can choose, is it better to have 93.4% of peak performance with 10% of annotation cost or 97.4% of peak performance with 40% of annotation cost. We also show the selected 10% of pixels for annotations in different methods in Figure 3. From the Figure, we can observe that the regions where both Entropy and Entropy+ Edge methods are given wrong labels are selected for annotation in SP+CRF. We can also observe that the segmentation results are improved in SP+CRF.

Baseline	SP+CRF			
100% GT	10% GT	20% GT	30% GT	40% GT
65.3	61.0 (93.4%)	61.8 (94.6%)	62.4 (95.5%)	**63.6 (97.4%)**

Table 2. Performance of SP+CRF on cityscapes data over different percentage of pixel level annotations. GT = ground truth.

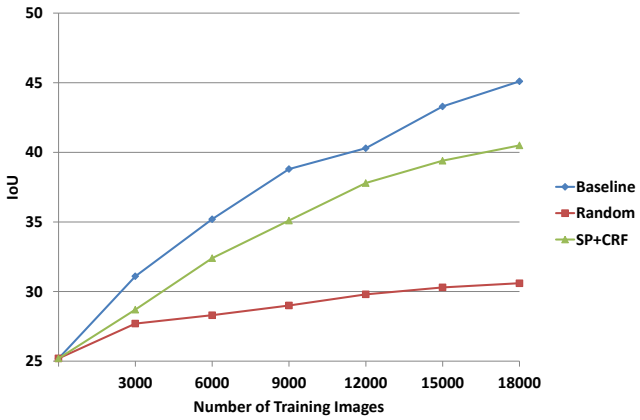

Figure 6. Performance of the proposed active learning methods over incremental selection of groups on mapillary data.

# Training images	Baseline	Random	SP + CRF
	100% GT	10% GT	
Cityscapes - 2975	25.2	25.2	25.2
3000	31.1	27.7 (89.0%)	**28.7 (92.2%)**
6000	35.2	28.3 (80.3%)	**32.4 (92.0%)**
9000	38.8	29.0 (74.7%)	**35.1 (90.4%)**
12000	40.3	29.8 (73.9%)	**37.8 (93.7%)**
15000	43.3	30.3 (69.9%)	**39.4 (90.9%)**
18000	45.1	30.6 (67.8%)	**40.5 (89.8%)**

Table 3. Results on Mapillary data. The results are given over incremental groups. The percentage values in the bracket indicates the relation to the baseline.

4.3. Results on Mapillary Dataset

The results on transfer learning are demonstrated over mapillary dataset. The results are given in Table 3 under Baseline. Here, the trained model on cityscapes data (2975 images) is used for propagating the labels to mapillary data. We also show the performance of the network over these incremental groups in Figure 6. From the results, we can observe that, the maximum improvement in the accuracy is obtained in the initial groups. We also presented the segmentation results over these groups in Figure 4. We can observe that the segmentation results are improving with the addition of more groups.

The results over Mapillary data for SP+CRF are given in Table 3. On complete training data with 18000 images, we obtain 89.8% of baseline performance by querying only 10% of pixels for labeling. It outperformed the random selection of 10% of pixels for annotation. In Figure 5, we show the segmentation results obtained using transfer learning. Here, we show the segmentation result obtained from the Cityscapes model without any additional labeling (No additional labels) and the result obtained using 10% of pixel annotations using the proposed method (Using 10% labeling). For annotation of regions by an oracle, we use a similar annotation tool to the one in [5]. The larger regions are annotated through the watershed algorithm. For annotating smaller/narrower regions, we selected the regions through a magnetic lasso tool proposed in [20]. Few examples for annotation are given in Figure 7. The regions marked in green are suitable for annotation by [5] and the regions marked in red are annotated by [20].

Internal experiments showed us that using different cameras with different imaging sensor characteristics (e.g. dynamic range, signal to noise ratio etc) for recording same dataset also leads to failure of current model even for the same classes. In this context, we refer to similar dataset captured with different imaging sensors as different target domains. In this work, we use Cityscapes as source domain and Mapilliary as target domain. Cityscapes dataset is obtained from a single camera source and Mapillary consists of similar dataset captured using various imaging sensors such as mobile cameras, automotive cameras etc. The predictions on Mapillary from Cityscapes trained model also proves that a different camera plays a significant role in performance of a model and how active learning is advantageous to improve the performance of that model without heavy annotation load. So, in this work, we define domain as similar datasets captured using imaging sensors with varied characteristics.

	Baseline	Entropy	[9]
	100% GT	10% GT	
Accuracy (mIoU)	56.0	49.9 (89.1%)	**50.0 (89.2%)**
Computational Time	-	**0.09** Sec	0.7 Sec

Table 4. Comparison of the proposed entropy based uncertainty measure with the uncertainty measure given in [9]. The computational time is given in seconds.

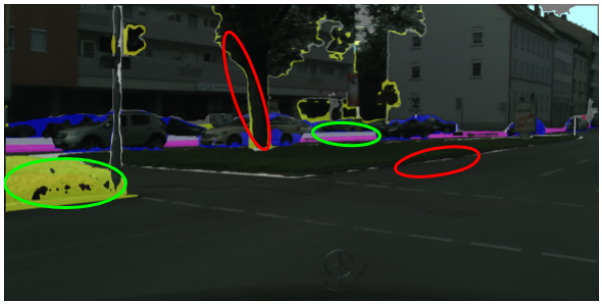

Figure 7. Few examples of region selection for human annotation. The regions marked green are labeled by watershed algorithm similar to [5] and the red regions are annotated by [20]

To further evaluate the proposed entropy based uncertainty measure, we compare it with the uncertainty measure given in [9]. The results are given in Table 4. In terms of accuracy, the proposed entropy based uncertainty measure is comparable with [9]. However, the proposed entropy based method is 8 times faster compared to [9]. In terms of class-wise IoU, entropy based method is performing well for the person and vehicle classes (truck, bus, train, motorcycle, bicycle). On the other hand, the uncertainty measure defined in [9] is performing well for the classes like traffic light, traffic sign, terrain, sky and rider. Class-wise IoUs are given in the supplementary material.

5. Conclusion

In this paper, we introduced an active learning methodology based on entropy and a region-based strategy for efficient labeling in semantic segmentation. The proposed method was evaluated using contemporary state-of-the-art deep learning models on the road scene understanding problem. In particular, our proposed method obtained 93.8% of baseline performance with only 10% annotations on the Cityscapes dataset, and $\sim$90% of baseline performance on the Mapillary dataset.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, Nov 2012.

[2] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3378–3385. IEEE, 2012.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):2481–2495, 2017.

[4] H. Caesar, J. Uijlings, and V. Ferrari. Region-based semantic segmentation with end-to-end training. In *European Conference on Computer Vision*, pages 381–397. Springer, 2016.

[5] H. Caesar, J. R. R. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. *CoRR*, 2016.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. 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, 2018.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] A. Freytag, E. Rodner, and J. Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision*, pages 562–577. Springer, 2014.

[9] M. Gorriz Blanch, A. Carlier, E. Faure, and X. Giro I Nieto. Cost-Effective Active Learning for Melanoma Segmentation (poster). In *NIPS ML4H Workshop*, 2017.

[10] Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In *Advances in neural information processing systems*, pages 593–600, 2008.

[11] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424. ACM, 2006.

[12] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Semisupervised svm batch mode active learning with applications to image retrieval. *ACM Transactions on Information Systems (TOIS)*, 27(3):16, 2009.

[13] A. Holub, P. Perona, and M. C. Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, June 2008.

[14] S.-J. Huang, R. Jin, and Z.-H. Zhou. Active learning by querying informative and representative examples. In *Advances in neural information processing systems*, pages 892–900, 2010.

[15] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[16] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017.

[17] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

[18] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning Proceedings 1994*, pages 148–156. Elsevier, 1994.

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[20] E. N. Mortensen and W. A. Barrett. Intelligent scissors for image composition. SIGGRAPH '95, 1995.

[21] G. Neuhold, T. Ollmann, S. R. Bul, and P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5000–5009, Oct 2017.

[22] N. Roy and A. Mccallum. Toward optimal active learning through sampling estimation of error reduction. In *In Proc. 18th International Conf. on Machine Learning*, pages 441–448. Morgan Kaufmann, 2001.

[23] A. I. Schein and L. H. Ungar. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265, 2007.

[24] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[25] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[26] R. Triebel, J. Stühmer, M. Souiai, and D. Cremers. Active online learning for interactive segmentation using sparse gaussian processes. In *German Conference on Pattern Recognition*, pages 641–652. Springer, 2014.

[27] A. Vezhnevets, J. M. Buhmann, and V. Ferrari. Active learning for semantic segmentation with expected change. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3162–3169, June 2012.

[28] S. Vijayanarasimhan and K. Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2262–2269. IEEE, 2009.

[29] Y. Yang and M. Loog. Active learning using uncertainty information. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 2646–2651. IEEE, 2016.

[30] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. Icnet for real-time semantic segmentation on high-resolution images. *arXiv preprint arXiv:1704.08545*, 2017.

[31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.