Active Learning for Lane Detection: A Knowledge Distillation Approach

Fengchao Peng, Chao Wang, Jianzhuang Liu, Zhen Yang
Noah's Ark Lab, Huawei Technologies
{pengfengchao, wangchao165, liu.jianzhuang, yang.zhen}@huawei.com

Abstract

Lane detection is a key task for autonomous driving vehicles. Currently, lane detection relies on a huge amount of annotated images, which is a heavy burden. Active learning has been proposed to reduce annotation in many computer vision tasks, but no effort has been made for lane detection. Through experiments, we find that existing active learning methods perform poorly for lane detection, and the reasons are twofold. On one hand, most methods evaluate data uncertainties based on entropy, which is undesirable in lane detection because it encourages to select images with very few lanes or even no lane at all. On the other hand, existing methods are not aware of the noise of lane annotations, which is caused by heavy occlusion and unclear lane marks. In this paper, we build a novel knowledge distillation framework and evaluate the uncertainty of images based on the knowledge learnt by the student model. We show that the proposed uncertainty metric overcomes the above two problems. To reduce data redundancy, we explore the influence sets of image samples, and propose a new diversity metric for data selection. Finally we incorporate the uncertainty and diversity metrics, and develop a greedy algorithm for data selection. The experiments show that our method achieves new state-of-the-art on the lane detection benchmarks. In addition, we extend this method to common 2D object detection and the results show that it is also effective.

1. Introduction

Lane detection is a crucial task for autonomous driving. Recently, great advances have been made by deep learning to improve the lane detection performance [33, 31, 8]. However, a deep model requires a huge amount of training data in order to yield a satisfying result. Due to the large aspect ratio and the special shape of lanes, it is highly expensive and cumbersome to annotate a sufficiently large dataset.

Active learning is a well-known technique to reduce the annotation cost [34, 42, 25]. It is proposed to select the most informative data items from the unlabeled dataset according to some policy. The selected data items are then annotated manually and added to the training set. Various selection policies have been proposed. Compared to the random selection, these policies manage to reduce the annotation cost by a large margin, and in the meanwhile, they are able to achieve a competitive, or even better, training performance.

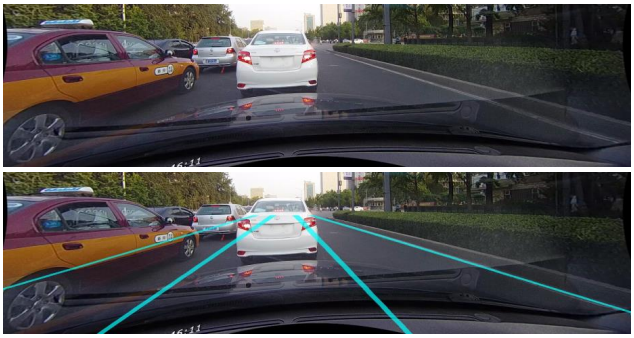


Figure 1: Examples of noisy annotations. All the lanes except the rightmost one are invisible (occluded or unclear) in the original image. Their locations are annotated by guessing and can be misleading to a lane detection model.

However, though active learning has been fruitful in image classification [3, 42, 39], object detection [2, 6, 19], semantic segmentation [38, 41], and other non-computer-vision areas [30, 15, 32], we find that for the lane detection task, the existing methods are not so effective. The reasons are twofold. On one hand, entropy is widely used to estimate the uncertainty of images. The images with highest entropy values are considered informative. But in practice, we observe that entropy-based methods are prone to selecting images with very few lanes. These images provide less useful information than normal ones, and therefore a model trained using them does not perform well. On the other hand, lane annotations are often noisy. For example, on the CULane dataset [29], many annotations are made in regions where there are no visible lane marks at all. An example image is shown in Fig. 1. In these regions, annotators decide the location of an invisible lane just by guessing. The guessed annotations are often incorrect and can bring heavy noise. Existing methods do not model the noisy lane annotations, and are therefore easily disturbed.

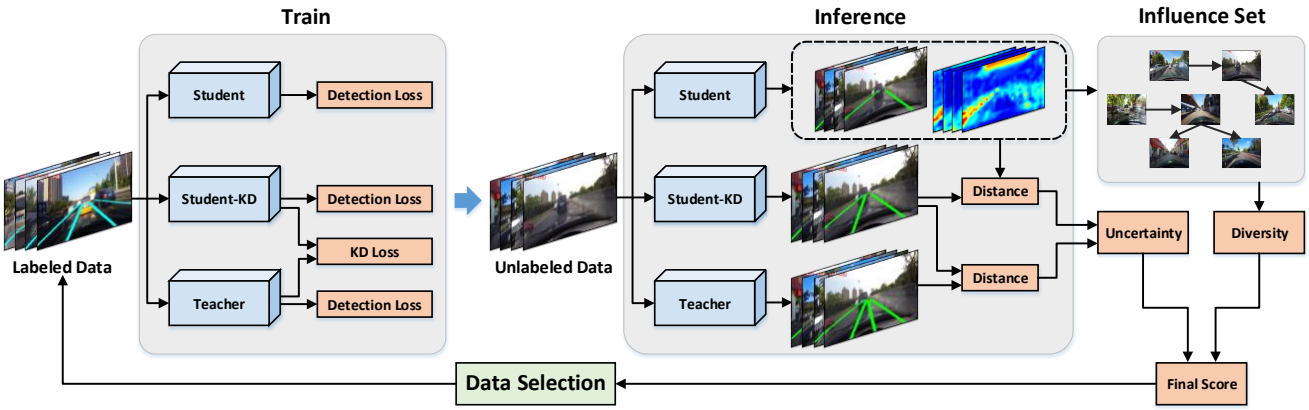


Figure 2: Framework of our proposed method. We build three models, the large teacher model, the small student model, and the student model distilled by the teacher (Student-KD in the figure). The prediction gaps of three models on each unlabeled image are used to estimate the uncertainty. The diversity score of an image is estimated based on its influence set, extracted from the outcome of the student model. The uncertainty and diversity scores are combined as the final score for data selection.

In this paper, we propose the first active learning method for lane detection, as shown in Fig. 2. It is able to solve the two above-mentioned problems (i.e., unsuitable entropy metric and label noise). To get rid of estimating the entropy, we propose to use Knowledge Distillation (KD) to explore uncertain samples. We regard the lane detection model to be deployed as the student (denoted as Student-KD in Fig. 2), and train it together with a large teacher model. We use their prediction gap as the basic estimation of uncertainty.

In addition, we also use KD to solve the label noise problem. We find that useful knowledge can be transferred from the teacher to the student, but label noise is difficult to transfer. On images with noisy labels, the prediction gaps between the teacher and the student are generally larger than those on normal images. However, a large prediction gap between the teacher and the student does not necessarily indicate high label noise. There can be knowledge, i.e., label with no noise, which is naturally difficult for the student to learn. To distinguish noise from hard-to-learn knowledge, we train another student model (denoted as Student in Fig. 2) that has the same structure as Student-KD. The difference is that we train it independently without knowledge distillation from the teacher. We also measure the prediction gap between the two students. Since label noise is random, on a noisy image, the prediction gap between any pair of the three models is likely to be large. On the contrary, a small prediction gap between any pair of the models indicates that the label noise is likely to be low. Based on these observations, we propose a novel uncertainty metric that is able to capture both the knowledge and the noise. Images with more knowledge and less noise are selected.

Uncertainty is not the only factor to decide the informativeness of a sample. Data redundancy can lead to waste in annotation cost [37]. Therefore, diversity is also a key factor for efficient data selection. We present a new diversity metric that uses influence sets [23] to estimate the diversity of a selected set. In the data selection phase, we calculate the similarity between unlabeled images based on their feature maps. Given the pair-wise similarity, we build the influence set for each image based on its reverse nearest neighbors and estimate the diversity score. Then, the uncertainty score and diversity score are combined as the final score for data selection (see Fig. 2). We formulate the data selection as a set cover problem, and use a greedy algorithm to solve it.

We perform extensive experiments on the most widely used benchmarks [29, 4]. The results show that our method achieves state-of-the-art performance on all the datasets. In addition, we adapt our method to 2D object detection and test it on a benchmark. The results show that our method outperforms a recent active learning method specifically designed for 2D object detection, and is therefore extendable to other visual recognition tasks.

Our contributions are summarized as follows:

1. We propose the first active learning method for lane detection. A knowledge distillation framework is built to solve the two specific problems in lane detection, the unsuitable entropy and label noise problems. We are also the first to explore knowledge distillation in the data selection of active learning.

2. We propose a novel uncertainty metric that is able to capture both knowledge and noise. Besides, we present a diversity metric based on reverse nearest neighbors to solve the data redundancy problem. The combination of the two metrics is not only effective but also extendable to other visual recognition tasks.

3. The experiments on two widely used lane detection benchmarks show that our method outperforms recent active learning methods. We also demonstrate the effectiveness of our method for object detection.

2. Related Work

Lane Detection. The lane detection problem has been studied for decades. Early methods rely on hand-crafted features and their capability is limited to detecting lanes in easy cases [40, 5]. In recent years, deep learning improves the lane detection performance by a large margin. Pan *et al*. [29] proposed a spatial CNN model to pass information across rows and columns inside a neural layer, and used a segmentation head to predict lanes. PINet [21] first predicted point clouds in the lane regions and then performed clustering in post-processing. PointLaneNet [8] and FastDraw [31] detected lanes in an object-detection manner. They both enumerated anchors on feature maps and built multi-task headers to perform classification and regression, respectively. The UFLD [33] first built row anchors on an image and formulated the lane detection problem as to select certain pixels in each anchor. A very recent work [26] used a transformer to predict the shape parameters of lanes. Different from these methods, we focus on active learning for lane detection and our method is model agnostic.

Active Learning. Active learning aims at selecting most informative data items to form the training set and improve the training performance at a very low annotation cost. The informativeness of a data item is studied from two perspectives, uncertainty and diversity. A variety of methods were proposed to estimate the uncertainty, such as cross entropy [16, 30], best-vs-second-best [18, 36], expected model change [7, 9, 22], etc. Gal *et al*. [20, 10, 24] proposed a series of Bayesian methods, drawing samples from the dropout distribution of a stochastic neural network. The uncertainty was evaluated as the mutual information of the sample outputs. Yoo *et al*. [42] proposed to directly estimate the uncertainty using a header of the model. Gao *et al*. [12] and Zhou *et al*. [45] added augmentations to input images and evaluated the uncertainty as the consistency of model predictions. Instead of solving general problems, Liu *et al*. [25] incorporated spatial information to the active learning for human pose estimation. Similar idea was used by Aghdam *et al*. [2] in 2D object detection.

A common drawback of the above methods is that they ignore data redundancy. To deal with this problem, Nguyen *et al*. [28] extracted clusters from the unlabeled dataset and prevented the model from repeatedly selecting samples from the same cluster. Sener *et al*. [37] defined the problem as a core-set selection problem and proposed a k-center greedy algorithm to solve it. Agarwal *et al*. [1] combined this idea with a contextual diversity measurement, encoding spatial context variations in sample selection. Sinha *et al*. [39] directly searched for the most representative samples using an adversarial learning framework.

None of the previous methods is designed specifically for lane detection. They are either inapplicable to lane detection or able to solve the problem only partially. In comparison, we propose a framework that works very well for this task, and it is extendable to other visual recognition tasks.

Knowledge Distillation. Knowledge Distillation (KD) [13] was firstly proposed to transfer knowledge from a large model to a small model for model compression. Recently, researchers started to exploit the competence of KD in semi-supervised learning and active learning. Gao *et al*. [11] used KD in semi-supervised learning to improve the tolerance to data noise. Yun *et al*. [44] were the first to employ KD in active learning. However, they used KD only in the training phase, while we also use KD in the data selection phase (see Section 3.2).

3. Proposed Method

In this section, we introduce our method in detail. We first describe the knowledge distillation method, which is used to train the models and perform prediction on the unlabeled dataset. Then we present the calculation of prediction gaps, and propose a novel uncertainty metric to estimate valuable knowledge as well as label noise. After that, we design a diversity metric to reduce data redundancy. Finally, we combine the uncertainty and diversity metrics and develop an algorithm to select most informative samples.

3.1. Knowledge Distillation

We first build three models, a teacher (M_T) and two students (M_S and M_{S-KD}). The teacher model uses a larger backbone network, while the two students are of the same structure. In the training phase, we train M_S and M_T independently on an initial training set. In this step, no information is passed between them.

We choose PointLaneNet (PLN) [8] as the primary model for its simplicity. Other models are also applicable. PLN has a backbone network to extract visual features, and builds two headers on the feature map. One header predicts the class of each feature pixel. The other predicts the coordinates of all the points of the lane passing through this pixel. The detection loss of PLN consists of a classification loss and two regression losses, which is defined as:

$$L_{det} = \sum_{i=1}^{w}\sum_{j=1}^{h}(\lambda CE_{ij} + \mu \mathbb{1}_{ij} L_{ij}^{loc} + \nu \mathbb{1}_{ij} L_{ij}^{pos}), \quad (1)$$

where CE denotes the cross entropy, L^{loc} and L^{pos} are two L2 losses for regression, w and h are respectively the width and height of the feature map, λ, μ and ν are weights, and $\mathbb{1}_{ij}$ is the indicator function which is 1 if the pixel (i, j) is selected and otherwise 0. The definitions of these losses are based on the model predictions and the ground truth. More details about Eq. (1) refers to [8].

After the initial independent training, we obtain the trained student model M_S and teacher model M_T. Then,

(a) Input **(b)** GT **(c)** M_T **(d)** M_{S-KD} **(e)** M_S

Figure 3: An example of case 4. (a) The original image. (b) The ground truth. (c) The prediction of the teacher. (d) The prediction of the distilled student. (e) The prediction of the student without distillation. Obviously, the difference D_{ST} between (c) and (d) is large, and D_{SS} between (d) and (e) is also large. The reason is that, as shown in (a), most lane marks in this image are occluded and it is difficult for the models to detect the lanes. However, though this image presents a difficult case, we do not want to select it, because in this case, annotators are prone to guessing the locations of the occluded lanes. As shown in (b), the middle and right lanes are guessed by the annotator. Guessed annotations are often incorrect and thus misleading the models. Therefore, our uncertainty metric penalizes this by the fraction between D_{SS} and D_{ST} in Eq. (5).

we train the other student M_{S-KD} by distilling the knowledge from M_T using the same initial training set. Following [8], we design a distillation loss L_{dis} that also consists of one classification loss and two regression losses. These losses are similar to those in [8], and the only difference is that the ground truth in [8] is replaced by the soft prediction of the teacher M_T. Thus, the training loss of M_{S-KD} (i.e., KD Loss in Fig. 2) is defined as:

$$Loss = L_{det} + \alpha L_{dis}, \tag{2}$$

where α is a weighting factor. The difference between M_{S-KD} and M_S is that M_{S-KD} is trained to learn from both the ground truth and the teacher. It is usually stronger than M_S. Here we also call M_S a student because it has the same structure as the real student M_{S-KD}.

3.2. Uncertainty

Uncertainty is a natural criterion for data selection. Cross entropy is a commonly used uncertainty measurement. However, we will show in the experiments that this method is prone to selecting images with few lanes, which therefore provide less information for training. Label noise also causes uncertainty. Fitting to noisy labels reduces the model performance. In this subsection, we propose a novel uncertainty metric to solve these two problems.

We first define the prediction gap between two models. Given an image p and two models M_1 and M_2, denote the sets of their predicted lanes as $M_1(p)$ and $M_2(p)$, respectively. For each lane $l_1 \in M_1(p)$, we find its closest lane in $M_2(p)$ with:

$$l_2 = \arg\min_{l \in M_2(p)} Dist(l_1, l). \tag{3}$$

The distance $Dist()$ between two lanes is calculated as the segment-wise Euclidean distance. Then the prediction gap between M_1 and M_2 is defined as:

$$D_{12}(p) = \max_{l_1 \in M_1(p)} Dist(l_1, l_2). \tag{4}$$

To simplify the notation, we use $D_{12}(p)$ and D_{12} interchangeably in the following.

We now have three trained models, the student M_S, the distilled student M_{S-KD}, and the teacher M_T. To model the uncertainty, we calculate the gap D_{SS} between M_S and M_{S-KD}, and the gap D_{ST} between M_{S-KD} and M_T. D_{SS} captures the knowledge learned by the distilled student from the teacher. D_{ST} captures the knowledge that is difficult for the student M_{S-KD} to learn. Based on D_{SS} and D_{ST}, we divide uncertain samples into four typical cases:

1. **Small D_{SS} and small D_{ST}.** This means the model predictions are stable and consistent. The image is likely to be an easy sample. There is no need to annotate it.

2. **Small D_{SS} and large D_{ST}.** There is a large gap between the teacher and the distilled student. This means the student M_{S-KD} cannot well learn from the teacher for this image, which can be caused either by difficult knowledge or noise. The small D_{SS} indicates that the image is unlikely to cause label noise, because an image easy to cause wrong annotations usually has unclear lane marks, leading two different models to guessing randomly with a large gap between their predictions. Therefore, this image hopefully contains useful knowledge and is valuable to annotate.

3. **Large D_{SS} and small D_{ST}.** A small gap between the teacher and the distilled student indicates the knowledge is easy to learn. However, a large D_{SS} means there is a risk to trust the teacher. The teacher can transfer incorrect knowledge to the student. Therefore, this image is also valuable to annotate.

4. **Large D_{SS} and large D_{ST}.** The three models cannot provide consistent predictions on this image. Noisy images usually cause this problem and we do not want to select them. Even if this is not due to noise, it is not an easy sample for all the three models. Considering the two reasons, though this image can be valuable to annotate, we do not treat it as the highest priority.

An example of case 4 is shown in Fig. 3. Other cases are given in the appendix.

Figure 4: Comparison of the distance-based [25] and our RNN-based strategies. We randomly select five video segments from the CULane dataset, and in each segment we randomly choose two frames. Then the two strategies are used to select five from these ten images. The distance-based method chooses images 3, 5, 6, 9, and 10, while ours obtains images 2, 3, 6, 8, and 9. Our selection naturally covers all the five segments, but the distance-based method ignores the first and forth segments. This indicates that the average distance is not an appropriate metric to evaluate the influence among images.

Combining the above four cases, we propose a simple yet effective uncertainty metric for image p:

$$Uncr(p) = (D_{SS} + D_{ST}) \cdot \max\{\frac{D_{ST}}{D_{SS}}, \frac{D_{SS}}{D_{ST}}\}. \quad (5)$$

This metric encourages a large difference between D_{SS} and D_{ST}, so that images of cases 2 and 3 will get large scores. For easy samples, D_{SS} and D_{ST} are both small, and this metric is small. If D_{SS} and D_{ST} are both large (case 4), $(D_{SS} + D_{ST})$ is also large, but it is penalized by $\max\{\frac{D_{ST}}{D_{SS}}, \frac{D_{SS}}{D_{ST}}\}$. In this way, the images with potential noise will not get very large uncertainty scores. During data selection, we simply sort the images in the decreasing order of their uncertainty values and select those with top values.

3.3. Diversity

In addition to uncertainty, diversity is another important factor in selecting informative samples. It encourages the selected samples to be representative of, or in other words, to be able to influence, a variety of other unlabeled samples.

A recent method [25] evaluates the diversity of a sample by its average feature distance to other unlabeled samples. An image with the minimum average distance to all other unlabeled images is considered as the most influential sample. However, for lane detection, we can illustrate that this average distance is not an appropriate metric of the influence (diversity). In a lane dataset, images are sampled from video segments, and therefore, there are many similar images from a scene. A natural way to find an influential subset is to select one image from every video segment and then annotate them for training. However, the distance-based metric [25] cannot achieve this, as shown in the simplified experiment in Fig. 4. The reason is that the distance between two images calculated in a high-dimensional feature space is often not equivalent to the perceptual dissimilarity of them. Another recent work [3] uses the K-Means++ method for data selection, but its effectiveness on lane detection has not been validated. In fact, our experiments (Section 4.1 and Section 4.2) show that our diversity metric proposed below is a superior one.

In this work, we explore the influence set of each image, and define the diversity of a selected subset as the number of unlabeled samples it can influence. The influence set is extracted based on reverse nearest neighbors (RNNs). If a sample p is the nearest neighbor of a sample q, then reversely, q is called the reverse nearest neighbor of p. Given a sample p, a dataset S, a distance function $d()$, and an integer k, the reverse k nearest neighbors of p is defined as:

$$RNN_k(p) = \{q \in S - \{p\} | p \in NN_k(q)\}, \quad (6)$$

where $NN_k(q)$ denotes the k nearest neighbors of q. $RNN_k(p)$ means that the sample p is closer to all the samples in $RNN_k(p)$ than most of the other samples in the entire dataset S. Therefore, p is likely to be the most influential sample for all the samples in $RNN_k(p)$.

Given the unlabeled dataset S_U, the current subset of selected samples $V \subset S_U$, and an image $p \in S_U$, we define the diversity of p as the number of its reverse k nearest neighbors in S_U:

$$Div(p|V, S_U) = |RNN_k(p) - V|. \quad (7)$$

Note that different from nearest neighbors, for different p's, the sizes of $RNN_k(p)$ can be very different and even much larger than k. In this work, k is considered as a hyper-parameter, so it is dropped on the left side of Eq. (7). The experiments in Fig. 4 and Section 4 show that for lane detection, our RNN-based strategy performs better than the previous diversity-based methods.

3.4. Active Learning Algorithm

We combine the uncertainty and diversity metrics, and define the data selection problem as follows:

$$\begin{aligned} \max_{V \subset S_U} \quad & \sum_{p \in V} (Uncr(p) + \beta Div(p|V, S_U)), \\ \text{s.t.} \quad & |V| = b, \end{aligned} \quad (8)$$

where β is a weighting factor and b is the annotation budget (number of selected samples). The target of this problem is to select a subset of samples that are of high uncertainty, and at the same time, can influence a large subset of samples in the remaining unlabeled dataset.

This optimization is a set cover problem [14]. Though it is NP-Hard, the objective function is non-decreasing and submodular, and an $O(N^2)$ greedy algorithm is able to ensure a $(1 - \frac{1}{e})$-approximation to the optimal solution [27].

Algorithm 1 Active Learning with Knowledge Distillation

Input: Labeled dataset S_L, unlabeled dataset S_U, number of rounds r, budget b per round;
Output: Selected dataset $V \subset S_U$, with annotations;

1: $M_S, M_T \leftarrow Train(S_L)$;
2: $M_{ST} \leftarrow Train_{KD}(S_L)$;
3: $V \leftarrow \emptyset$;
4: **while** $|V| < r \cdot b$ **do**
5: **for** $p \in S_U$ **do**
6: $P_S, P_{ST}, P_T \leftarrow Predict(M_S, M_{ST}, M_T, p)$;
7: Compute D_{SS} and D_{ST} with P_S, P_{ST}, P_T;
8: $uncr \leftarrow (D_{SS} + D_{ST}) \cdot \max\{\frac{D_{ST}}{D_{SS}}, \frac{D_{SS}}{D_{ST}}\}$;
9: $div \leftarrow Div(p|V, S_U)$;
10: $S_{score}(p) \leftarrow uncr + \beta \cdot div$;
11: **end for**
12: $Q \leftarrow Greedy(S_U, S_{score}, b)$;
13: $S_U \leftarrow S_U - Q$;
14: $Q \leftarrow Annotation(Q)$;
15: $V \leftarrow V \cup Q$;
16: $M_S, M_T \leftarrow Train(S_L \cup V)$;
17: $M_{ST} \leftarrow Train_{KD}(S_L \cup V)$;
18: **end while**

The complete active learning algorithm is shown in Alg. 1. We start by training the student and teacher models in the knowledge distillation framework (lines 1 and 2). Then given the label budget b of each round, we iteratively calculate the uncertainty and diversity scores of samples (lines 5–11), select b images from the unlabeled dataset using the greedy algorithm (line 12), annotate them (line 14), and use the updated training set to train the models (lines 15–17). The iteration (lines 4–18) repeats until $|V| \geq r \cdot b$.

3.5. Extension to 2D Object Detection

Though our method is initially designed for lane detection, it is easy to extend it to other active learning tasks. In this paper, we explore 2D object detection.

We use the same KD framework, and the same definitions of uncertainty and diversity. We only slightly change the definition of the prediction gap between two models. Given an image p and two models M_1 and M_2, we first match the two sets of bounding boxes detected by M_1 and M_2, respectively. That is, for each predicted box $b_1 \in M_1(p)$, find $b_2 \in M_2(p)$ with the largest $IoU(b_1, b_2)$. Then the prediction gap is defined as:

$$D_{12}(p) = \sum_{\{b_1, b_2\}} (1 - IoU(b_1, b_2)) \cdot (1 + \gamma \mathbb{1}(c_1 \neq c_2)), \tag{9}$$

where $\gamma > 0$ is a hyper-parameter, and c_1 and c_2 are the predicted classes of b_1 and b_2, respectively.

4. Experiments

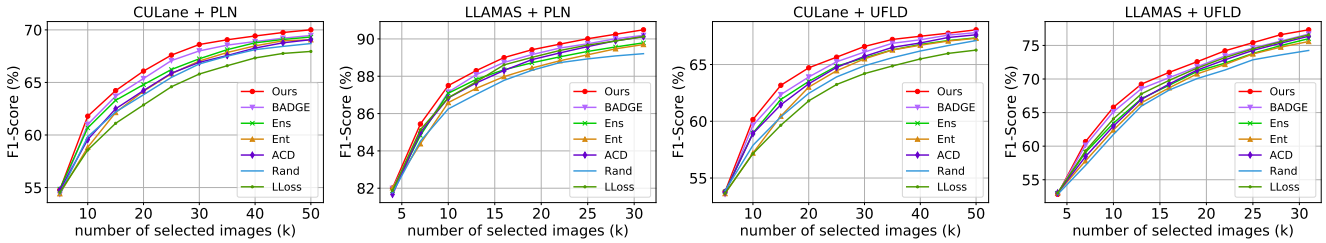
Datasets. We perform experiments on two most popular datasets, CULane [29] and LLAMAS [4]. CULane contains 88880 training images and 34680 test images. These images are annotated by humans. The LLAMAS dataset is annotated automatically using Lidar maps. It contains 58269 training images and 20844 validation images.

Models. We test our method using two different lane detection models, PointLaneNet (PLN) [8] and UFLD [33]. For PLN, we use a pruned ResNet-122 as the backbone network for the student model, and a SENet-154 [17] for the teacher model. SGD is used to train the models with the batch size of 32. The learning rate is set to 0.02. The parameters λ, μ, and ν in Eq. (1) are respectively set to 1, 0.01, and 0.1; α in Eq. (2) is set to 1; k in Eq. (7) is set to 3; β in Eq. (8) is set to 5. We use the output of the last convolutional layer of the backbone network as the feature of an image. Either the students or the teacher can be used to extract features, and there is almost no difference between their effects. Here we use the student model without distillation (M_S). On each dataset, the models are firstly trained for 30 epochs on a randomly selected training set, containing about 5% samples of the entire training set. Then we iteratively select a subset of the unlabeled images for annotation, and fine-tune the models for 10 epochs. We repeat these until the budget is used up. To make a fair comparison with other methods, we use M_S to perform evaluation, though we prefer to deploy M_{S-KD} in practice. The experiment is repeated for five times and the curve of the mean evaluation results is reported. For UFLD, we use ResNet-18 for the student model and ResNet-101 for the teacher model. The same parameter values are set as recommended by the authors of UFLD. All the experiments are performed on a GPU server with 8 NVIDIA Tesla V100 cards.

Evaluation metrics. We use F1-Score for evaluation. Given a fixed lane width, a predicted lane and an annotated lane are considered matched if their IoU is greater than 0.5. Then the F1-Score is defined as: $F_1 = \frac{2 \times P \times R}{P + R}$, where P is the precision and R is the recall.

Compared methods. We compare our work with the following six methods:

1. Random (Rand). The baseline random selection.
2. Entropy (Ent). This method selects samples based on the cross entropy.
3. Ensemble (Ens). We use the student and teacher models, M_S and M_T, to predict lanes on each unlabeled image. Then we select the images with largest prediction gaps between M_S and M_T.
4. ACD [2]. This method is designed specifically for object detection. It incorporates the spatial information to estimate the entropy.



(a) **(b)** **(c)** **(d)**

Figure 5: Results of the seven methods on the CULane and LLAMAS datasets. (a) and (b) show the results with PointLaneNet. (c) and (d) show the results with UFLD.

5. LLoss [42]. This method adds a header to the network to estimate the loss of each sample. Samples with largest predicted losses are selected.

6. BADGE [3]. This method combines an uncertainty metric (gradient norm) and a diversity metric (K-Means++) to select samples.

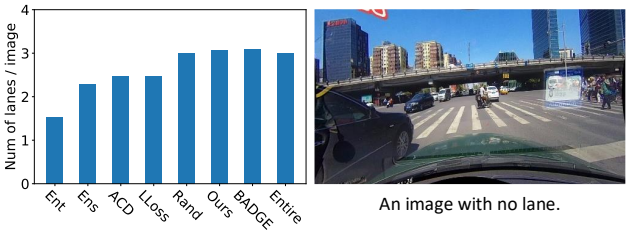
The methods 2–5 consider only uncertainty.

4.1. Results

The results are shown in Fig. 5. Figures (a) and (b) show the results with PLN. We find that our method achieves the best performance on the two datasets. Given the same label budget, the training sets selected by our method are the most informative, making the model trained with these sets yield the highest F1-Score on the test datasets. The BADGE method achieves the second best performance. The reason is that it also takes both uncertainty and diversity into consideration in the data selection process. The other methods only focus on uncertainty, and therefore perform worse.

In particular, the CULane dataset contains a variety of scenes such as crowd, night, shadow, etc. We observe that on this dataset, the methods with only uncertainty achieve a mild improvement, compared with the baseline Rand. One important reason is that these methods are prone to selecting samples with few lanes, as shown in Fig. 6. In the first iteration, the samples selected by Ent contain only 1.5 lanes per image, while for the entire dataset, the average number of lanes per image is 3.0. Other uncertainty-only methods, i.e., Ens, LLoss, and ACD, also have this problem. These selected samples provide too few positive annotations, leading the model to not being trained sufficiently. By comparison, the average lane number per image selected by our method is 3.1. However, this does not mean that we should simply select images with the most lanes. We have tested this idea but the result is worse than Rand.

ACD extends the calculation of entropy by taking the spatial neighborhood of each pixel into consideration. This idea is useful for object detection because it provides a more accurate estimation of uncertainty for each pixel. However, it is not so effective for lane detection because lanes are too thin, and a foreground pixel does not have as many foreground neighbor pixels as those in object detection. Therefore, the estimation of pixel-level uncertainty cannot be enhanced and ACD does not obtain the best performance. The LLoss method performs even worse than Rand on CULane. The reason is that in the first 4 or 5 iterations, this method focuses on selecting night images; 70% of the selected images are in night scenes. This leads to a high redundancy in the selected training set. Though in the later iterations it starts to focus on other scenes, this cannot resolve the redundancy problem caused by so many night images selected already. Therefore, the performance of LLoss is worst.



(a) **(b)**

Figure 6: (a) Average number of lanes per image in the dataset selected by each method. The last bar (Entire) denotes the average number of lanes per image in the entire dataset. (b) An example of images with no lanes that are likely to be selected by the uncertainty-only methods.

In the LLAMAS dataset, the average number of lanes per image is 5.2. Most of the images each contains at least two lanes. The uncertainty-only methods are not likely to select images with few lanes, as those on the CULane dataset. Therefore, though performing worse than ours, they are able to outperform Rand by a large margin. In this dataset, most images are collected from scenes in daytime without crossing and heavy shadow, so that LLoss does not suffer from selecting too many night images. Its performance is much better than that on CULane.

In Fig. 5, (c) and (d) show the results with UFLD. Similar to those with PLN, our method achieves the best performance on the two datasets. The methods considering both uncertainty and diversity outperform those using only

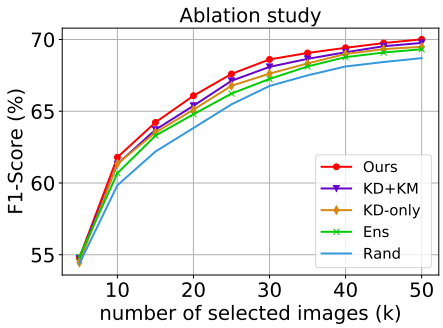


Figure 7: Results of the ablation study.

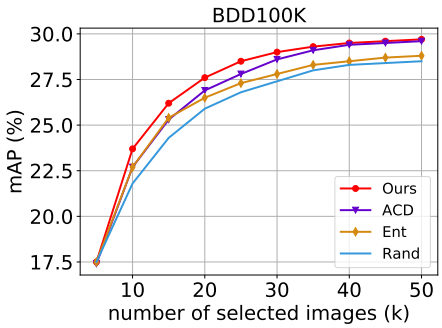


Figure 8: Results of 2D object detection.

uncertainty, on both datasets. On CULane, LLoss still focuses on selecting night images, and therefore performs the worst. Ent is prone to selecting images with few lanes in the beginning, so on the CULane dataset, its performance is worse than Rand in the first two iterations. On LLAMAS, all methods manage to outperform Rand.

These comparisons show that our method is effective in reducing the annotation cost for lane detection. The PLN and UFLD make different formulations to the lane detection problems. But our method achieves good performance with both of them. This shows that our method is model agnostic and is suitable for a variety of lane detection models.

4.2. Ablation Study

We now validate the effectiveness of our uncertainty metric and diversity metric. We first build a KD-only version of our method, i.e., we perform data selection only based on our uncertainty metric. Then to validate the diversity metric, we build a combination of the KD-only version and the widely used K-Means++ method [3], denoted as KD+KM. Due to limited space, we only perform experiments on CULane with PointLaneNet. The results are shown in Fig. 7. We find that even if we only use our uncertainty metric for data selection, its performance is still better than Ens, which performs best overall on CULane among the four uncertainty-only methods. The KD+KM method achieves a better performance. But this performance is worse than the combination of KD-only with our diversity metric based on reverse nearest neighbors (Ours).

These results show that both our uncertainty metric and diversity metric are effective for data selection. The combination of them obtains the best performance.

4.3. Extension to Object Detection

In this subsection, we extend our method to 2D object detection. We use the Faster-RCNN [35] as the base model, with ResNet-18 as the backbone for the student model, and ResNet-101 for the teacher model. The experiments are conducted on the BDD100K dataset [43]. It contains 70000 images in the training set, and 10000 images in the validation set. Our method is compared with three methods, Rand, Ent, and ACD [2]. ACD is a recent active learning method specifically designed for 2D object detection. The pipeline of the experiments is similar to that of the lane detection. Each method starts from training the models using an initial training set, which is formed via randomly sampling 5000 images from the original training set. Then in each iteration, each method selects 5000 images, adds them to the training set, and fine-tunes the models. The initial training takes 12 epochs and the fine-tuning takes 8 epochs, with SGD optimization and a learning rate of 0.02. The batch size is set to 32; γ in Eq. (9) is set to 5; α in Eq. (2) is set to 10. The iteration repeats until the annotation budget is used up. This experiment is conducted for five times. We use the average mAP of the five experiments to evaluate the performance.

The results are shown in Fig. 8. Ent obtains a good improvement over Rand in the first three iterations. But its improvement decreases significantly in the later iterations. ACD outperforms Ent by a large margin because it considers the spatial information when calculating the entropy. It provides a more accurate pixel-level estimation of the uncertainty. Our method achieves an improvement that is even larger than that of ACD. These results show that our method is extendable to 2D object detection.

5. Conclusion

In this paper, we present the first active learning method for lane detection. We find that two problems restrict the performance of existing methods, namely the unsuitable entropy and label noise. To solve these problems, we propose to employ knowledge distillation to evaluate both the data uncertainty and the potential label noise. We also propose a diversity metric based on reverse nearest neighbors. This metric can help reduce the redundancy of the selected dataset. The experiments show that both the metrics are able to improve the lane detection performance, and the combination of them achieves the best results on two popular benchmarks. In addition, our method is extendable to other visual recognition tasks, and in this paper, we show its effectiveness on 2D object detection. In the future, we will extend this active learning framework to a wider range of recognition tasks to further examine its capability.

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *ECCV*, 2020.

[2] Hamed H Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M López. Active learning for deep detection neural networks. In *ICCV*, 2019.

[3] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2019.

[4] Karsten Behrendt and Ryan Soussan. Unsupervised labeled lane marker dataset generation using maps. In *ICCV*, 2019.

[5] Amol Borkar, Monson Hayes, and Mark T Smith. A novel lane detection system with efficient ground truth generation. *IEEE Transactions on Intelligent Transportation Systems*, 2011.

[6] Clemens-Alexander Brust, Christoph Käding, and Joachim Denzler. Active learning for deep object detection. *arXiv preprint arXiv:1809.09875*, 2018.

[7] Wenbin Cai, Ya Zhang, and Jun Zhou. Maximizing expected model change for active learning in regression. In *ICDM*, 2013.

[8] Zhenpeng Chen, Qianfei Liu, and Chenfan Lian. Pointlanenet: Efficient end-to-end cnns for accurate real-time lane detection. In *IEEE Intelligent Vehicles Symposium*, 2019.

[9] Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *ECCV*, 2014.

[10] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. *arXiv preprint arXiv:1703.02910*, 2017.

[11] Jiyang Gao, Jiang Wang, Shengyang Dai, Li-Jia Li, and Ram Nevatia. Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In *ICCV*, 2019.

[12] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *ECCV*, 2020.

[13] Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey. *arXiv preprint arXiv:2006.05525*, 2020.

[14] Andrew Guillory and Jeff Bilmes. Interactive submodular set cover. *arXiv preprint arXiv:1002.3345*, 2010.

[15] Dilek Hakkani-Tür, Giuseppe Riccardi, and Allen Gorin. Active learning for automatic speech recognition. In *ICASSP*, 2002.

[16] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *CVPR Workshops*, 2008.

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[18] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, 2009.

[19] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *ACCV*, 2018.

[20] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017.

[21] Yeongmin Ko, Jiwon Jun, Donghwuy Ko, and Moongu Jeon. Key points estimation and point instance segmentation approach for lane detection. *arXiv preprint arXiv:2002.06604*, 2020.

[22] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. In *NeurIPS*, 2017.

[23] Flip Korn and Suresh Muthukrishnan. Influence sets based on reverse nearest neighbor queries. *ACM Sigmod Record*, 2000.

[24] Yingzhen Li and Yarin Gal. Dropout inference in bayesian neural networks with alpha-divergences. *arXiv preprint arXiv:1703.02914*, 2017.

[25] Buyu Liu and Vittorio Ferrari. Active learning for human pose estimation. In *ICCV*, 2017.

[26] Ruijin Liu, Zejian Yuan, Tie Liu, and Zhiliang Xiong. End-to-end lane shape prediction with transformers. In *IEEE Winter Conference on Applications of Computer Vision*, 2020.

[27] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 1978.

[28] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *ICML*, 2004.

[29] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. In *AAAI*, 2018.

[30] Fengchao Peng, Qiong Luo, and Lionel M Ni. Acts: an active learning method for time series classification. In *ICDE*, 2017.

[31] Jonah Philion. Fastdraw: Addressing the long tail of lane detection by adapting a sequential prediction network. In *CVPR*, 2019.

[32] Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. Bayesian batch active learning as sparse subset approximation. In *NeurIPS*, 2019.

[33] Zequn Qin, Huanyu Wang, and Xi Li. Ultra fast structure-aware deep lane detection. *arXiv preprint arXiv:2004.11757*, 2020.

[34] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *arXiv preprint arXiv:2009.00236*, 2020.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *TPAMI*, 2016.

[36] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *ECML*, 2006.

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

[38] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *CVPR*, 2020.

[39] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *ICCV*, 2019.

[40] Luo-Wei Tsai, Jun-Wei Hsieh, Chi-Hung Chuang, and Kuo-Chin Fan. Lane detection using directional random walks. In *IEEE Intelligent Vehicles Symposium*, 2008.

[41] Alexander Vezhnevets, Joachim M Buhmann, and Vittorio Ferrari. Active learning for semantic segmentation with expected change. In *CVPR*, 2012.

[42] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *CVPR*, 2019.

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

[44] Juseung Yun, Byungjoo Kim, and Junmo Kim. Weight decay scheduling and knowledge distillation for active learning. In *ECCV*, 2020.

[45] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *CVPR*, 2017.