International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 14, No.1 (Sep-2023)

http://dx.doi.org/10.12785/ijcds/140165

A Cost-effective Deep Active Learning for Object Detection in Automated Driving Systems

Mohib Eddine Khebbache[1], Salim Bitam[2] and Abdelhamid Mellouk[3]

[1]*Department of Computer Science,University Of El Oued, El Oued, Algeria*
[2]*Department of Computer Science,University of Biskra, Biskra, Algeria*
[3]*EPISEN, THINCNET,University of Paris-Est Creteil, Vitry sur Seine, France*

Received 29 Nov. 2022, Revised 22 Jul. 2023, Accepted 10 Aug. 2023, Published 01 Sep. 2023

Abstract: In Automated Driving Systems (ADS), the function of detecting objects on the road assists vehicle traffic and improves road safety. Deep Active Learning (DAL) is an advanced training method suitable for building robust Convolutional Neural Network (CNN)-based on road object detection models. This method automatically selects and manually labels training samples that are significantly less noisy, non-redundant and more useful. Depending on the complexity of detection task and the characteristics of urban scenes, the batch selection in the conventional batch mode DAL can suffer from the impact of the correlation between frame labels and batch size as well as variable labeling costs. This paper introduces a novel cost-effective-based training approach suitable for CNN-based on-road object detector, where frames labeling and batch size are considered in the sample selection process. We propose a batch sampling strategy that leverages the model prediction uncertainty along with dynamic programming to alleviate the selection batch size issue. Additionally, we investigate the effects of classification uncertainty, regression uncertainty and batch size during sample selection. Our approach was extensively validated on the Caltech Pedestrian dataset to fine-tune a pre-trained Tiny-YOLOv3 for performing pedestrian detection task. Results showed that our approach, compared to basic methods, can build robust detection model that keeps the detection error less than 57%, saving up 50% of the labeling effort and alleviating batch size dependency.

Keywords: autonomous driving, object detection, visual similarity, deep active learning, cost-effective training, pedestrian detection

1. Introduction

Over the past two decades, autonomous driving has become a reality with the advent of the first self-driving car projects. As can be seen today, Automated Driving Systems (ADS) have emerged as future innovations in research and industry fields. Although taking a step further on the safety guarantee is already an important promise, the large-scale deployment of such systems depends critically on the effectiveness and robustness of their perceptual capability to understand the surrounding scene in a non-deterministic urban environment [1].

Several recent works have tackled object detection, among other perception functions, focusing on deep learning approaches since their performance has significantly improved real-world autonomous driving benchmarks [2]. However, these approaches are prone to detection errors as they fail to identify and recognize the surrounding object.

As a functional safety concept for deep learning approaches, it is crucial to assess the safety risk by interpreting the behavior of a DNN-based model [3]. In practice, building an efficient DNN-based object detector largely depends on fully supervised training of such a model while assuming the availability of a large, finely labeled training dataset.

Considering the dynamics, scalability and visual pattern similarity of the real-world vehicular environment, the fully supervised learning can easily become a drain on cost, time, and computing resources. On the other hand, with the presence of noisy samples in this learning paradigm, the detection model can suffer from loss of robustness when it comes to guarantee an accurate model. Moreover, high-quality annotation of such a complex environment can only be guaranteed if enough human effort and labor costs are spent while exceeding the limited budget. In light of this, delivering more safer ADS is problematic as the deployment of the DNN-based on-road object detectors on embedded hardware is hampered.

The easiest way to overcome the aforementioned issues is to focus on training the deep detector networks using a random subset of labeled data. Considering both scale and visual similarity issues, a limited random subset may lead to lower accuracy comparing to training the network on the entire dataset [4]. Alternatively, other efficient strate-

E-mail address: mohibeddine-khabache@univ-eloued.dz , s.bitam@univ-biskra.dz, mellouk@u-pec.fr

gies, such as Active Learning (AL), can reduce the heavy annotation burden and the amount required of training data.

Active learning is a commonly-used paradigm designed for shallow machine learning models to overcome the limitations of traditional "passive" supervised learning algorithms. An active learning algorithm interactively selects, based on the model's current knowledge, the data from which the model learns new knowledge that contributes to achieving significant accuracy without extra annotation and training costs [5], [6]. Recently, DAL has emerged to deal with high-dimensional data and complex deep neural networks (DNN) while retaining the powerful capabilities of deep learning. Such a strategy explores DNN's architecture or their prediction outputs as useful cues to select informative samples for the labeling process [7], [8]. In contrast, DAL algorithms for object detection are likely to become a computational bottleneck. Especially when considering related details regarding DNN-based object detector models and public autonomous driving benchmarks, which are: (1) architectural design, nature and granularity levels (coarse to fine-grained) of outputs, and (2) high-dimensionality, visual similarity, labeling cost variability, multi-modality (images, videos, optical flow, LiDAR), intra-class variance and imbalanced problems of data. In this context, a cost-effective approach is necessary to save cost as much as possible while considering the general deficiency related to the excessive redundant examples query, limited annotation budget, aggregation of multiple DNN outputs and bridge the performance gap with other learning methods that could deliver satisfactory performance.

This paper proposes a new Cost-Effective Deep Batch Mode Active Learning framework (CEDBMAL) that consists of a label-efficient learning algorithm. This labeling algorithm aims to incrementally improve the performance of a CNN-based object detector with less excessive retraining on a small amount of labeled urban area frames.

Instead of fully supervised training on the entire sequence of frames, CEDBMAL provides a subset of useful labeled training data, by performing a batch sampling strategy and optimization technique, yet saving labeling costs. More specifically, CEDBMAL implies an annotation cost-based active selection strategy for querying manual annotation of an optimal batch of more valuable unlabeled frames, with minimal risk of redundant samples (as noisy samples), imbalanced class distribution and annotation cost wasted.

Accordingly, the main contributions of our work are summarized as follows:

- We design an efficient and scalable batch sampling strategy that deals with a large-scale sequence of highly similar unlabeled data to fine-tune a CNN-based object detector.
- We propose an annotation cost-based approach for determining the optimal batch with an optimal size that ensures a few outliers, balanced class distribution and low similar frames that meet the annotation cost.
- For future deployment, we use a tiny version of an object detector and estimate the annotation time in an autonomous driving context.
- We evaluate our framework on the Caltech Pedestrian dataset by performing pedestrian detection in static images.

The rest of this paper is organized as follows: Section 2 presents the problem to be solved and the inadequacies of the current approaches to solve the current problem. Section 3 reviews the related work in deep active learning for general object detection and specifically on-road object detection. Section 4 illustrates in detail our proposed framework by describing each component. Section 5 provides the experiment setup and delivers the results. Finally, Section 6 presents the conclusion and the suggested future work.

2. Problem Statement and Motivation

In general, the main challenge facing deep learning approaches for object detection in autonomous driving is the requirement of a massive amount of training samples. The usual approach to create such datasets consists of collecting, from public traffic scenes, as many frames as possible and manually drawing bounding boxes (labeling) for all objects of interest in all frames. However, training samples collected from non-deterministic urban scenes suffer from data imbalance problems, such as class and scale imbalances, as well as similar visual patterns (trees, cars, sky, etc.). These issues are due to the diversity of classes (background and foreground) and density (sparse and dense) distribution of on-road objects across frames.

For a detection task in such a context, the labeling and training costs are notably high and unequal especially when similar samples are considered in training and samples selection process. In order to reduce the aforementioned costs in such a task, the primary hypothesis is to make usefull annotation budget and training resources by selecting a subset of informative samples that maximize performance gain and minimize human labeling effort, with an increased concern about capturing diverse visual patterns.

Using random sampling approach, the selected subset does not guarantee that it will capture diverse visual patterns [4]. Alternatively, the selection of one sample (frame) at each DAL cycle, performed by traditional one-by-one DAL query methods, for deep detection model retraining imposes a higher annotation and training costs. In contrast, batch mode query strategies select a batch of samples that provide a balance between most informative and diverse samples within a batch based on uncertainty and diversity measures. Furthermore, labeling and then using such a batch as a mini-batch can lead to retrain efficiently detection model without extra burden. As a result, the batch mode DAL is more

efficient and cost-effective for building CNN-based object detection models [7], [8].

In recent years, existing works on batch mode DAL have focused on the step of identifying useful measures to select informative and representative batches while assuming a static batch size and a fixed labeling cost. In a real-world applications such as autonomous driving, the static specification of the batch size without looking at the correlation between frames' objects amount, estimated by model's output, and labeling cost may not lead to good generalization accuracy, effective management of labeling cost variability, practical batch size determination and low redundancy ensuring. In other words, there is still a considerable research gap when it comes to fulfilling the need for cost-effective batch selection strategy in batch mode DAL algorithms where batch samples are selected in an adaptive and dynamic manner, at each DAL cycle, according to the distribution of the most promising predicted objects and the cost of labeling each batch sample given this distribution. The objective of our paper is to fill this gap.

3. Related Work

A. Active Learning for Deep Architectures

With the advent of DNN models, various scenarios have been widely explored in the field of deep active learning, where CNNs are well-studied deep learning models. Key differences between these approaches reside in different aspects, including learning tasks, data and embeddings distribution, deep model predictions, query strategy design, selection criteria quantification (uncertainty, diversity, inconsistency, and label correlation), scoring level (pixel, box, region or image) and metric measurement, aggregation techniques, sampling granularity, labeling source, and data accessibility.

Unlike the typical AL methods, advanced DAL-related researches are not rich in terms of properly approaching different aspects of various DAL problems. In most heuristic-based DAL approaches, Uncertainty Sampling (US) [4], [9], [10], [11], [12], Query-by-Committee (QBC) [13], mutual information [14], and expected model change [15] are most heuristics used, as a single criterion, throughout the query strategy to select a single instance at a time. However, several researches have reported the invalidity of applying a one-to-one query strategy to support the batch training principle inherent in DL approaches. Furthermore, in the supervised training scenario of deep model, access is only allowed to the labeled data through the cycles of DAL, without any assistance from the remaining unlabeled data.

To deal with these limitations, multiple selection criteria are considered for enabling efficient CNN-based model training across AL cycles. Therefore, hybrid-criteria, mixture-criteria, multi-criteria [16] or batch-based sampling [17], [18] query were proposed to select a substantial amount of samples to be labeled at a time while attempting to find a balance between the considered strategies.

Moreover, promising research directions have been explored to extend DAL algorithms regarding the integration of different annotation granularity, abundant unlabeled data and related supervision setting into active learning pipeline, including multi-label [19], multi-view [20], multi-instance [10], multi-instance multi-label (M2AL)[21], multi-view multi-instance multi-label (M3AL) [22], and unsupervised [23], [14] AL schemes. Among them, more attention has been paid to address two aspects: the automatic design of selection samples strategy [24] and the alleviation of various problems, namely data-related problems such as confidence and insufficient labeled sample, model-related problems such as generalization ability, and domain-specific problems such as domain shift, cold-start problem and class imbalance.

Besides, serious researches have been conducted in recent years to properly design cost-effective DAL frameworks. The main idea is to adopt both query-driven and data-driven cost-saving strategies. The query-driven approaches were based on gaining support from complementary techniques to perform query improvement, such as optimization techniques, metrics learning [25], and alternative learning paradigms (one-shot, contrastive, federated, goal-driven, domain adaptive...) [26], [27], [28], [29], [30]. On the other side, data-driven approaches were attempted to address several data-level perspectives in terms of data labeling supervision (weak, self, semi...) [11], [31], [32], [33], labeling setting (open-set recognition) [34], [35] and granularity [18], [21]. For further details please refer to the survey papers [7], [8], [36]. In this paper, we describe related work on the latest DAL approaches that employ CNNs for object detection tasks in general and autonomous driving applications in particular.

B. Active Learning for Deep Object Detection

An uncertainty-based active learning approach for object detection in remote sensing images is presented in [9]. The authors argue that an efficient weighted combination of classification and regression uncertainty could overcome class imbalance and object densities variation difficulties. Based on predictions (bounding box and classification probability) of a CNN-based detector on unseen, unlabeled images, the high ranked image could be selected according to the image-level uncertainty score aggregated by summing each object uncertainty within unlabeled image. With the low granularity level, the authors in [10] explored the instance-level for object detection. Throughout a multiple instance unsupervised active learning approach, the unlabeled images are treated as instance bags and feature anchors in images as instances where the image uncertainty is estimated using an instance uncertainty learning and instance uncertainty re-weighting modules. As result, the high ranked images are used to train a constructed detector based on RetinaNet. By adopting query by committee, Roy et al. [13] formed a committee of classifiers by leveraging extra detection head layers of the deep network architecture (SSD). As selection criterion, the disagreement is mea-

sured and aggregated by introducing the 'margin' for each bounding box. By considering mAP improvement and class imbalance between background and object categories, li et al. [17] proposed WBetGS that enhance typical diversity and uncertainty based batch sampling for batch mode active learning in object detection. Nevertheless, inefficient training of CNN-based detector, redundant data selection, scalability handling and heavy burden of convergence time are main shortcomings.

A review of existing research on cost-effective DAL for object detection is relatively sparse. Most of these works are built upon mixed supervised learning methods. Leveraging access to both labeled and unlabeled data, a supervised signal is provided which optimizes iterative DAL cycles and reduces human annotator [11], [31], [32], [33]. Wang et al. [31] proposed an active sample mining (ASM) framework for cost-effective training of object detectors. Focusing on switchable sample selection mechanism, a number of unlabeled samples are selected, according to deep detector predictions, to automatically pseudo-label via novel self-learning process. However, the remaining samples are manually annotated via active learning process. For cost-effective panicle detection in cereal crops, the authors in [11] proposed an uncertainty-based active learning approach suitable for two-stage models. Only strong labels (tight bounding boxes) are queried by considering high uncertainty images picked from a constructed low-cost weak labeled (object centre clicking) subset driven by the oracle labeling knowledge. Alternatively, some works focus on exploring other metrics, such as consistency and entropy, to evaluate model predictions between the original and augmented data [37], [38].

Prior to the success of deep active learning in computer vision tasks for autonomous driving, numerous proposals of active learning methods, involving hand-crafted features and shallow classifiers, have targeted vehicle [39] and pedestrian [40] detection. Recently, few works have described deep active learning for on-road object detection. Aghdam et al.[4] addressed pedestrian detection in images and video. Based on CNN-based object detector predictions, pixel-level scores are computed and aggregated as a single image-level score. Thus, a fixed number of high ranked unlabeled images is selected for querying. With the introduction of temporal selection rules, the selection of high visual similar video frames could be avoided. Furthermore, the authors in [12] investigated LiDAR data and deep active learning for 3D object detection task. For training a LiDAR 3D object detector, an uncertainty-based method queries informative unlabeled samples from point cloud data, with the help of the 2D region proposals in RGB images. Using the same data format, the authors in [41] explored localization-based uncertainty metric for selecting samples from feature space extracted without any additional 2D input information. The proposed DAL method is built upon a specific object matching process and is suitable for a specific anchor-based object detection architecture. Besides, Liang et al. [42] tailored the diversity metric by proposing a novel spatio-temporal diversity-based acquisition function that selects frames from multimodal data pool. To ensure multi-view vehicle detection, the authors in [43] proposed an active learning algorithm to enhance the typical deformable part model by selecting B more effective part samples for query labeling by human annotator from multi-view vehicle images. Consequently, labeled part samples are considered as positive samples to retrain SVM model as learning part model. However, the desired performance could not be achieved unless the best samples to be labeled is determined in terms of computational efficiency, low redundancy and query and annotation cost saving.

Using the inherently more efficient and scalable batch sampling strategy, we argue that the uncertainty of CNN-based detector predictions, the diversity of learned representations, and the adaptive selection of the best batch size can reduce the selection of redundant (noisy) samples, handle the variable cost, speed up the annotation process and hence constitute an effective training set for building a competitive object detector while relaxing human supervision.

4. Cost-effective Deep Batch Mode Active Learning for Object Detection: Our Proposal

In this section, we describe the proposed framework in general, the underlying detection model and query strategy in detail.

A. Overall Framework

CEDBMAL framework focuses on a pool-based setting that consists of iterative selection/annotation process as depicted in Figure 1. Given a large pool of unlabeled images U_{unann} and a labeling budget, CEDBMAL firstly employs the underlying pre-trained detection model to examine each unlabeled image and then selects, using query strategy, a batches of more valuable examples based on uncertainty as informativeness measure and diversity as representativeness measure. Next, a batch with the best size value, amongst the selected batches in the first step, is picked out to be labeled manually by leveraging the information from resolved 0-1 Knapsack problem according to the existence of low redundant instances with more objects of interest and less estimated annotation time. Once labeled by an oracle(e.g., human annotator), labeled images pool U_{ann} is enlarged with this labeled subset which is retired from U_{unann}. These accumulated actively-labeled images are considered as training set to fine-tune the detector while getting an updated model in the result. The cycle of above steps is performed on the remaining unlabeled images, till the process exhausts the labeling budget or the required performance is achieved in a cost-effective manner. In the next subsections, we describe the detection model followed by the cost-effective active learning strategy performed by our framework.

B. Detection Model

In this work, we focus on single stage CNN-based object detector as the state-of-the-art object detector. Such

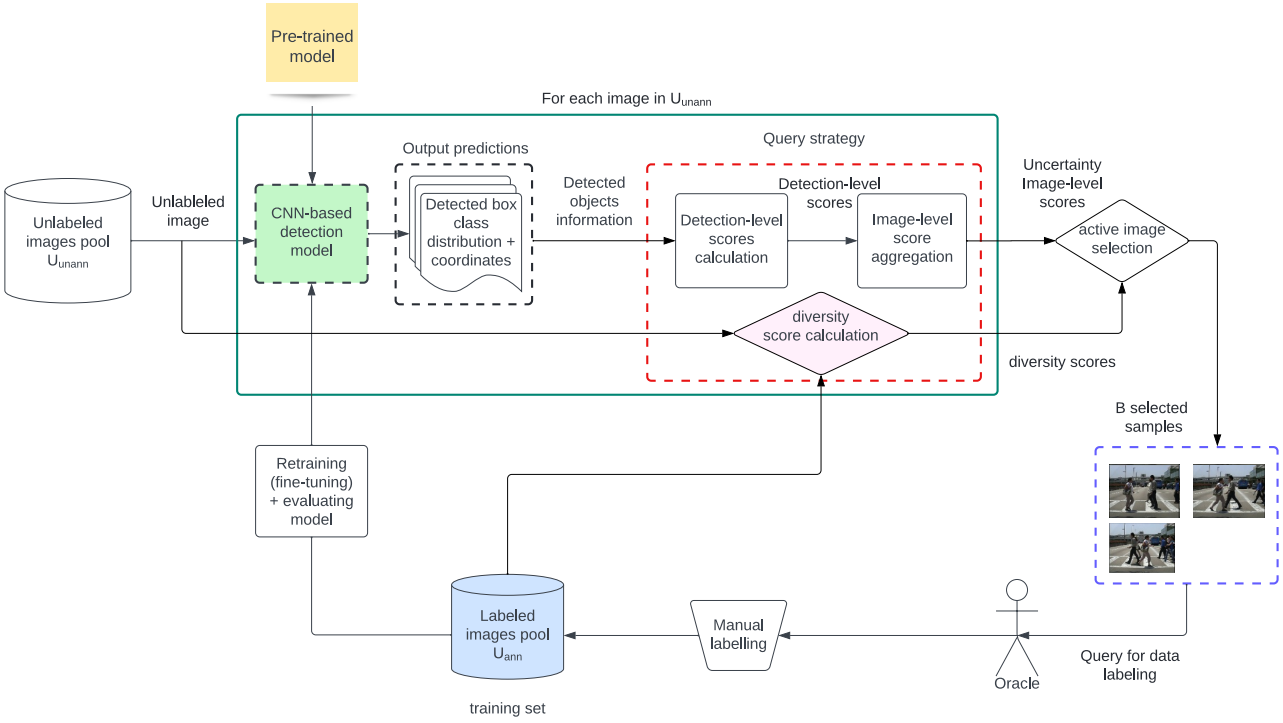

Figure 1. Overview of the proposed deep active learning framework

model relies on baseline CNN model for performing feature learning and extra head layers for performing classification and bounding box regression. The overall deep architecture is trained in end-to-end manner while a post-processing method is performed to obtain the final detection outputs. As detector prediction, the 2D map of probabilities per class and bounding boxes coordinates are used to rank examples in recent uncertainty-based deep active learning works [44]. For the reasons of reducing the training cost and exploring the influence of domain-shift on the overall performance, we prefer to fine-tune a pre-trained object detector using transfer learning paradigm, instead of training from scratch.

C. Deep Active Learning for CNN-based Object Detector

To train the underling detection model, the active learning method should carefully employ a properly designed query strategy for querying the labels while identifying the cost of preforming the selection/annotation process. Independently from the underlying detection model architecture, our query strategies are performed as explained below.

1) Uncertainty-based Deep Active Learning

Despite the effectiveness of uncertainty based Deep Active Learning for classification task, it needs to be revised for object detection. Basically, these selection strategies suffer from querying outliers and they are less efficient in evaluating images data in autonomous driving datasets when using only the label uncertainty of CNN-based model. Thus, the pick of more valuable unlabeled images may fail, with a negative impact on the detection performance. In order to address these challenges, we suggest to incorporate regression together with classification on an uncertainty sampling strategy as explained bellow.

Classification uncertainty sampling: Given an example x , CNN-based object detector estimates the probability distribution of the label $p(c|x)$ over C classes per detected Bounding Boxes. Such predictions are evaluated by a scoring function to measure the uncertainty metric and form a detection-level scores for each detected object, while using the uncertainty sampling for such purpose. For a given bounding box Bb, its classification uncertainty $U_C(Bb)$ is defined as $U_C(Bb) = 1 - P_{max}(Bb)$ where $P_{max}(Bb)$ is the highest probability distribution among all classes.

Regression uncertainty sampling: Since CNN-based detector predicts bounding box coordinates, the regression uncertainty could be measured by estimating distribution probability density [9]. By adopting Gaussian Mixture Model (GMM), each bounding boxes' distribution probability density (denoted L) is estimated in terms of calculated log probability. Then the obtained L is clipped as Lb=min(-99, L). Finally, the regression uncertainty (U_r) is calculated using the following uncertainty formulation

$$U_r = \begin{cases} 0.05 * (Lb + 10) + 0.5, Lb \geq -10 \\ 0.5 * \frac{Lb+100}{90}, Lb < -10 \end{cases} \quad (1)$$

WCR Deep Active Learning: Inspired by [9], our proposed weighted classification-regression (WCR)

uncertainty-based deep active learning algorithm uses both classification uncertainty U_c and regression uncertainty U_r to perform the query strategy. However, an unlabeled image could not selected for querying unless the WCR image-level uncertainty, denoted as U_s, was aggregated from detection-level scores for each detected box (object) in it, as:

$$U_s = agg(U_c(Bb) \times U_r(Bb))$$

, where Bb $\subseteq$ detected Bboxes. In our work, the aggregating methods are performed as inspired by [44].

- **Sum**: given an unlabeled image x, the aggregate score, from the detected bounding boxes D, can be obtained as follows

$$v_{Sum}(x) = \sum_{Bb \subseteq D} U_C(Bb) \quad (2)$$

- **Average**: with less sensitivity to the number of detections, the main idea is averaging all detection-level scores.

$$v_{Avg}(x) = \frac{1}{|D|} \sum_{Bb \subseteq D} U_C(Bb) \quad (3)$$

- **Maximum**: the maximum of detection-level scores is kept. Despite the robustness to zero valued detections (as noise), a substantial information can be lost.

$$v_{Max}(x) = \max_{Bb \subseteq D} U_C(Bb) \quad (4)$$

According to one-by-one query method, the query function can select a group of B unlabeled images with higher WCR uncertainty while ignoring outliers. Our contribution is highlighted in algorithm 1.

Algorithm 1 WCR deep Active Learning Implementation Details

Require: annotated images pool U_{ann}, unannotated images pool U_{unann}, object detector OD, testset U_{test}, objects' categories C

$U_{ann} \leftarrow \emptyset$
$OD \leftarrow$ pre-trained object detector OD_0
repeat
  for each image x in U_{unann} **do**
    Fed x into the object detector OD
    Get bounding boxes Dx with corresponding posterior probability p(c—Bb) and coordinates after post-processing operation (NMS)
    for each object Bb in Dx **do**
      Use objects' information to calculate U_c and U_r
    end for
    Calculate WCR uncertainty U_s using each x object's U_c and U_r
  end for
  Sort U_{unann} (in descending order) using the assigned WCR uncertainty U_s scores
  Select B high ranking images as queries for annotation by an oracle
  $U_{ann} \leftarrow U_{ann} + B$, $U_{unann} \leftarrow U_{unann} - B$
  $OD_t \leftarrow OD_{t-1}$ fine-tuned on U_{ann}
  Test OD_t using U_{test}
  Evaluate the detection performance (detection loss)
until The required performance is reached or query budge

Ensure: detector model parameters W_F and the final detector model OD_F

Despite its robustness, this solution could select redundant images which are less effective in training process. Thus, the visual patterns diversity in typical urban road scenario cannot be captured. Also, the repetitive selection procedure of one instance at time can lead to produce an inefficient and time consuming training process and therefore an expensive burden is taken by the annotator expert. Other issues are related to the impractical setting of B (ranging from increasing time-to-completion to some sort of uniform sampling of images) and the fixed assumption about the cost of labeling. Meanwhile, to address the above noted issues, one can attempt to incorporate our proposed uncertainty measure in batch sample query strategy aiming at ensuring a cost-effective training and labeling tasks, where a true diversity within a group of instances is guaranteed.

2) Cost-effective Deep Batch-mode Active Learning

Two critical design points, namely batch query and batch size selection, are carried out as explained below.

Deep Batch-mode Active Learning: Compared to one-by-one query strategy, several Deep Batch Mode Active Learning researches have shown the efficiency of hybrid batch-based query strategy in training of a CNN-based object detector [7]. In such setting, the final score used for ranking the unlabeled images and picking diverse samples with high uncertainty is calculated as [45]

$$finalScore = \alpha \times (1.0 - similarityScore) + (1.0 - \alpha) \times uncertaintyScore$$

, where the parameter α aims to weight the impact of each factor as

$$\alpha = \frac{|U_{unann}|}{|U_{unann}| + |U_{ann}|}$$

In our work, we investigate the previous presented WCR uncertainty as informativeness criterion, while choosing Euclidean distance as similarity measurement. This method favors the selection of the furthest unlabeled sample x_i from its closest labeled neighbor where the distance between them is computed as follows [46]

$$div_i = \min_{j=1,2,...,n} \|x_i - x_j\|^2, x_i \subseteq U_{unann}, x_j \subseteq U_{ann}$$

However, one of the fundamental issues in using a batch-based query strategy lies in the batch size that might produce worse results and make the labeling effort inefficient [45]. To relax this limitation, our contribution consists of selection of diverse batch with optimal size at each iteration under the constraint of a given budget and desired

performance.

Cost-effective Deep Batch-mode Active Learning: In autonomous driving context, the selection of optimal batch of instances with positive impact on detection performance is driven by determining the batch size, which ensures the adaptive response to varied labeling time. Independently from a particular batch size and inspired by [47], the optimal batch size selection is reduces to the 0-1 Knapsack problem, which maximizes the uncertainty, maintains the annotation costs and can be solved with dynamic programming. First, we pick a set of batches with size Q_i from unlabeled images pool, where $Q_i \subseteq 100 \ldots |$ unlabeled images pool $|$. Given such batches set, where each item have a weight T_i and value V_i, one can formulate a 0-1 Knapsack problem. It's worth noting that the batch uncertainty V_i is defined by summing the uncertainty of top-Q_i images within the batch i.e.

$$V_i = \sum_{j=1}^{Q_i} V_{ij}$$

while its annotation time is estimated as annotation cost T_i.

Estimating annotation costs: As cited in [11], the annotation time for baseline methods, given a batch of queried images, is calculated using the following formula

$$T_i = 7.8 \times Q_i + 34.5 \times bQ_i$$

where Q_i is the batch size, bQ_i is the total objects in it and $T_i <$ T. In our work, bQ_i is the total number of the predicted BBox within the batch.

As result, most useful instances with low redundant information could be selected to label while improving the performance at every iteration and saving immense amounts of labeling loads given fixed budget. The overall operations in our CEDBMAL is depicted in algorithm 2.

Algorithm 2 CEDBM Active Learning Implementation Details

Require: annotated images pool U_{ann}, unannotated images pool U_{unann}, object detector OD, testset U_{test}, objects' categories C

$U_{ann} \leftarrow \emptyset$
$OD \leftarrow$ pre-trained object detector OD_0
repeat
  for each batch size Q_i in 100 ... —U_{unann}— **do**
    for each image x in U_{unann} **do**
      Fed x into the object detector OD
      Get bounding boxes Dx with corresponding posterior probability p(c—Bb) and coordinates
      for each object Bb in Dx **do**
        Use objects' information to calculate U_c and U_r
      end for
      Calculate WCR uncertainty U_s, as $UncertaintyScore_x$, using each x object's U_c and U_r
      Calculate $similarityScore_x$ using Euclidean distance
      calculate $score_x = \alpha\times$ (1.0 { $similarityScore_x$) + (1.0 { α) $\times UncertaintyScore_x$
    end for
    Sort U_{unann} (in descending order) using the assigned $scores_x$
    Select a batch of instances B_{Qi} with largest $score_x$
    $bQ_i \leftarrow 0$
    $V_i \leftarrow 0$
    for each image x in B_{Qi} **do**
      $bQ_i \leftarrow bQ_i + D_x$
      $V_i \leftarrow V_i + UncertaintyScore_x$
    end for
    Estimate the annotation time T_i (as cost) using Q_i and bQ_i
  end for
  Estimate the optimal batch size by solving a 0-1 Knapsack problem using T_i and Uncertainty V_i for each batch size
  Select the batch B_{best}, with the best batch size, as queries for annotation by an oracle
  $U_{ann} \leftarrow U_{ann} + B_{best}$, $U_{unann} \leftarrow U_{unann}$–$B_{best}$
  Fine-tune OD_{t-1} using U_{ann} to get OD_t
  Test OD_t using U_{test}
  Evaluate the detection performance (detection loss)
until The required performance is reached or query budge

Ensure: detector model parameters W_F and the final detector model OD_F

5. Experiments and Results

A. Experimental Setup

To study how our DAL framework could ensure a cost-effective annotation and training processes while reducing manual annotation effort and guaranteeing expected detection performance along over a dataset for autonomous driving, we use it to fine-tune a pre-trained Tiny-YOLOv3 for detecting pedestrian (as uses case) on the Caltech Pedestrian Detection Benchmark [48] while evaluating various setting of B. In our experiments, we retain only training frames labeled as "person" with a height taller or equal 20 pixels which are used for simulating the oracle annotation and approaching the safety risk assess of the trained model by using partial specifications[3]. In addition, the test set is used to evaluate the performance (detection loss) of the detector using the Piotr's Matlab Toolbox while providing a fair and comprehensive comparison against two other alternatives: transfer learning and random sampling. In such case, the "Reasonable" scenario is preferred. For validation purpose, we split the training set by 10% as validation set. In the following, we will discuss the details of target dataset, the tiny version of the detection model and the alternative sampling methods.

1) Dataset

The Caltech Pedestrian dataset [48] consists of ~ 10 hours of 640x480 30Hz urban driving video with 350K labeled bounding boxes whereas 2,300 unique pedestrians were annotated. Over the 11 sessions, it results in 42,782 training images (set00-set05) and 4,024 test images (set06-set10) sampled every 30th video frame. The log-average miss rate is used to evaluate the detection performance and

is calculated by averaging miss rate on false positive per-image (FPPI) points where the relevant point is defined as FPPI = 10^{-1}. 4 testing scenarios which are "All", "Reasonable", "Scale=near", and "Scale=medium" are defined. The statistics of the frames with bounding boxes are labeled as "person" with a height of 20 pixels is summarized in TABLE I.

TABLE I. "person" labeled frames statistics in train and test sets

# unlabeled frames		"Person" label	
Train	**Test**	**# labeled images**	**# Bounding Boxes**
4250	4024	2006	4987

2) Tiny-YOLOv3

In this work, we focus on Tiny-YOLOv3, a simplified version of YOLOv3 [49]. Emphasis on the Darknet-53 backbone, its low-complexity architecture motivates its suitability for constrained environments with a significant detection speed, but at the cost of some loss of the detection accuracy. In our experiments, we use a pre-trained version of the Tiny-YOLOv3 on COCO benchmark [50] which contains 82 object categories. Also, we perform two training scenarios. Firstly, we freeze the Darknet backbone layers and fine-tune the other layers while setting the training parameters as: learning rate: $1e^{-3}$, number of iterations: 60, mini-batch size: 16. After that, we unfreeze all layers wile fine tuning all of the layers using the following training parameters: initial learning rate: $1e^{-4}$, initial number of epochs: 60, number of epochs: 120, batch size: 16. For the both scenarios, we use Adam as optimization algorithm while the learning rate is decayed by a factor of 0.1.

3) Random Sampling

Random sampling, as passive learning, is a naive sampling technique that aims to choose the frame to be labelled uniformly at random from the unlabelled pool. Thus, the selected frames are independent and not known beforehand[51].

4) Transfer Learning

Generally applied in deep learning, the transfer learning focuses on the transfer of knowledge from source domains to target domains while fine-tuning a pre-trained deep model. Thus, the performance of deep model could be improved by exploiting parameter sharing with low dependence on a large number of data and a tedious training process [52]. Incorporating this paradigm into our empirical experiments can allow us to provide a comprehensive comparison in terms of the number of training examples and iteration.

B. Results

1) Pre-trainted Tiny-YOLOv3 vs Transfer Learning

For further comparison with our method, we explore the benefit of transfer learning in improving Tiny-YOLOv3 detection performance. Figure 2 shows the quantitative results of the COCO pre-trained Tiny-YOLOYv3 versus the fine-tuned model on the Caltech Pedestrian dataset, in terms of miss rate and false-positive per image (FPPI).

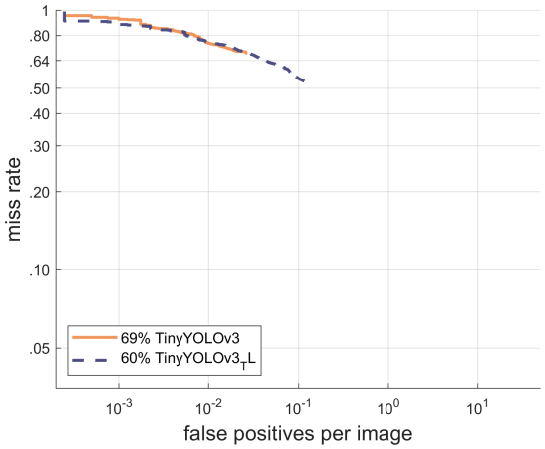

Figure 2. Performance curves of Caltech Pedestrian fine-tuned Tiny-YOLOv3 vs COCO pre-trained Tiny-YOLOv3

As shown in Figure 2, the transfer learning technique reduced the detection loss of pre-trained Tiny-YOLOv3 by 9%. This observation is explained by the fact that domain adaptation is achieved by training the pre-trained model on a target fully labeled dataset. Using the visual pattern knowledge learned from the COCO dataset as the source domain, the output of Tiny-YOLOv3 model was guided from the detection of various object classes to the precise location of pedestrian objects in Caltech Pedestrian dataset, as target domain, with high objectness score (from multiclass to binary object detection). Some qualitative examples of detection results on the Caltech test set, using the two Tiny-YOLOv3 models, are shown in Figure 3 and Figure 4.

2) Random Sampling

In this scenario, we investigate the random selection technique to randomly sample B instances for query manual labeling at each cycle, while setting B = 500 as indicated in [4]. TABLE II illustrates that, starting from the 2^{nd} cycle, the updated model performs close to the pre-trained and fine-tuned Tiny-YOLOv3 models. However, starting from the 6^{th} cycle, the updated model outperforms clearly the both previous models (57% against the 69% and 60% respectively) with only 3000 labeled frames. This is due to more knowledge being gained from the Caltech Pedestrian dataset by the trained model as the labeled frames in U_{ann} are increased.

In addition, one can observe a varied number of detected bounding boxes from cycle to another. This is due to the exploitation, by the model, of random knowledges from the training set consisting of randomly selected informative

Figure 3. Qualitative results of the pre-trained Tiny-YOLOv3 model on the Caltech Pedestrian dataset (detect 80 object categories)

Figure 4. Qualitative results of the fine-tuned Tiny-YOLOv3 model on the Caltech Pedestrian dataset(detect only pedestrian object)

samples.

3) Experiment on WCR

To further analyze the effectiveness of our DAL algorithm, we carefully evaluated the classification and regression uncertainty based sampling strategy for selecting a fixed B value, defined as in the previous experiment, of informative samples while considering an equal annotation cost for the overall unlabeled images.

TABLE II. Evaluation performance results for random sampling experiment

	cyc	#SI	#IBx	#BxC	#Bx	FPPI
RS B=500	1	500	218	486	486	85%
	2	1000	442	601	1087	69%
	3	1500	683	655	1742	62%
	4	2000	930	637	2379	62%
	5	2500	1163	547	2926	61%
	6	3000	1411	605	3531	57%
	7	3500	1637	579	4110	57%
	8	4000	1875	583	4693	55%
	9	4250	2006	294	4987	53%
PTY3						69%
TLTY3		4250	2006		4987	60%

cyc:cycles ,**#SI**:Number-selected-images ,**#IBx**:Number-images-with-Bboxes ,**#BxC**:Number-detected-Bboxes-per-cyc ,**#Bx**:Number-detected-Bboxes#

- *Experiment using classification uncertainty*:

To provide short analysis of our classification uncertainty selection strategy (U_c), we compare its overall performance, in terms of miss rate and false-positive per image (FPPI), to those of pre-trained and fine-tuned models, while evaluating three aggregation methods, namely sum, avg and max, for the earliest DAL cycles. Figure 5 and Figure 6 together provide the quantitative results.

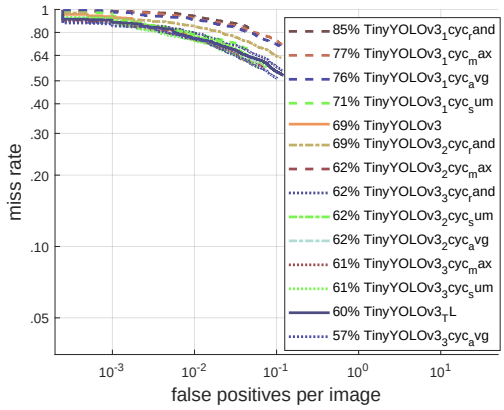

Figure 5. Performance curves of Pre-trained Tiny-YOLOv3 vs. TL-fine-tuned Tiny-YOLOv3 vs. random selection vs. Actively-fine-tuned Tiny-YOLOv3 (score function: Uc, aggregation method: sum, max and avg) at different training cycles on the Caltech Pedestrian dataset.

At the 1^{st} and 2^{nd} Uc-DAL cycles, the results demonstrate the potential of the DAL strategy to yield the same or lower detection loss with only a few labeled frames. Besides, in contrast to labeled frames randomly sampled or selected by sum and max aggregation methods at the 3^{thd} cycle, the miss rate is decreased to 57% when the 1500 labeled frames, selected by the avg method, are involved in

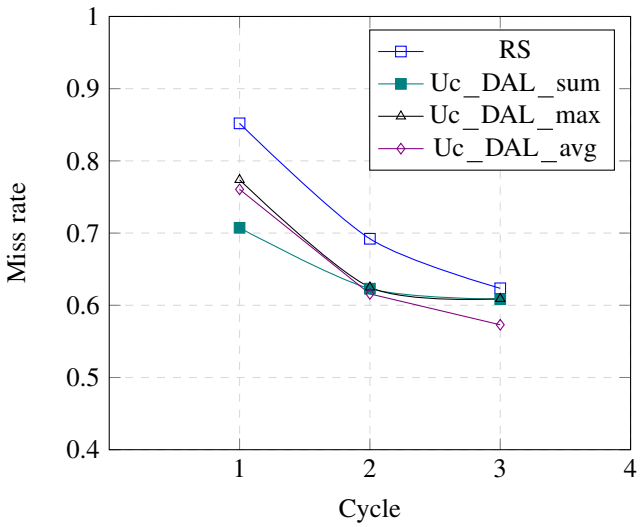

Figure 6. Miss rate of random selection vs. variants of our UC_DAL method based on sum, max and avg aggregation method.

the detector's training. This is mainly due to the ignorance of outliers, as noisy samples, during samples selection.

Using a subset of more informative samples selected during DAL cycles, the avg aggregation method is able to build detector model with lower detection loss compared to the transfer learning on the fully labeled Caltech Pedestrian dataset and the other methods. This can be also seen in Figure 6, showing the miss rate per cycle. However, sum and max aggregation methods could achieve the same performance but at the cost of more burden due to the outliers' influence and the visual similarity between the selected frames.

Overall, we can claim that the DAL algorithm based on classification uncertainty, is effective in training a detection model that guarantees the expected performance with less training effort and manual labeling. Yet, this comes at the cost of the negative influence of both outliers and visual similarity.

- *Experiment using regression uncertainty incorporated with classification uncertainty (WCR)*:

In this experiment, we analyze the exploration of model awareness about the class and localization prediction in addressing the aforementioned issues. The high scoring frames are selected according to weighted selection "WCR" (U_s) criterion based on U_c and U_r. Keeping the same fixed value of B, TABLE III to TABLE V show the results of miss rate obtained with respect to the number of pedestrian instances (instance-level labels) selected by every aggregation method after completion of a WCR-DAL cycle.

Compared to "sum" and "max" aggregation methods, the fine-tuning of the underling detector model using the top ranked labeled pedestrian instances according to the "avg" aggregation method is more accurate. ≅3195 labeled boxes, collectively contained in 1500 frames selected by "avg"

TABLE III. Evaluation performance results for WCR deep active learning experiment (with "sum" aggregation method)

	cyc	#SI	#IBx	#BxC	#Bx	FPPI
WCR-DAL sum	1	500	440	1764	1764	71%
	2	1000	802	862	2626	61%
	3	1500	1024	462	3088	60%
	4	**2000**	1231	415	3503	**57%**
	5	2500	1486	437	3940	58%
	6	3000	1709	518	4458	56%
	7	3500	1817	183	4641	56%
	8	4000	1872	98	4739	54%
	9	**4250**	2006	248	4987	**55%**

cyc:cycles ,**#SI**:Number-selected-images ,**#IBx**:Number-images-with-Bboxes ,**#BxC**:Number-detected-Bboxes-per-cyc ,**#Bx**:Number-detected-Bboxes#, **WCR-DAL**: WCR-DAL training, agg func= "Sum" , B=500

TABLE IV. Evaluation performance results for WCR deep active learning experiment (with "avg" aggregation method)

	cyc	#SI	#IBx	#BxC	#Bx	FPPI
WCR-DAL avg	1	500	418	1405	1405	72%
	2	1000	849	1337	2742	61%
	3	**1500**	1062	453	3195	**57%**
	4	2000	1263	428	3623	55%
	5	2500	1479	347	3970	57%
	6	3000	1652	265	4235	54%
	7	3500	1746	152	4387	57%
	8	**4000**	1863	245	4632	**51%**
	9	4250	2006	355	4987	52%

cyc:cycles ,**#SI**:Number-selected-images ,**#IBx**:Number-images-with-Bboxes ,**#BxC**:Number-detected-Bboxes-per-cyc ,**#Bx**:Number-detected-Bboxes#, **WCR-DAL**: WCR-DAL training, agg func= "Avg" , B=500

TABLE V. Evaluation performance results for WCR deep active learning experiment (with "max" aggregation method)

	cyc	#SI	#IBx	#BxC	#Bx	FPPI
WCR-DAL max	1	500	426	1632	1632	74%
	2	1000	819	1004	2636	63%
	3	1500	1049	483	3119	58%
	4	2000	1282	506	3625	58%
	5	2500	1502	369	3994	56%
	6	3000	1631	200	4194	55%
	7	3500	1733	168	4362	55%
	8	**4000**	1874	306	4668	**54%**
	9	4250	2006	319	4987	55%

cyc:cycles ,**#SI**:Number-selected-images ,**#IBx**:Number-images-with-Bboxes ,**#BxC**:Number-detected-Bboxes-per-cyc ,**#Bx**:Number-detected-Bboxes#, **WCR-DAL**: WCR-DAL training, agg func= "Max" , B=500

method are more accurate (with 57% of detection loss) than ≅3119 selected by "max" method (with 58% of detection loss) and ≅3088 selected by "sum" method (with 60% of detection loss). This can be explained by the effectiveness of the "avg" method in avoiding outlier selection, which is the main issue of the uncertainty-based sampling strategy. As a result, the selection of frames with sparse object density can be avoided and more informative pedestrian instances can be highly ranked in hopes of rapidly reducing detection loss.

Moreover, one can note that the U_c scoring function performs slightly close to WCR counterpart. Figure 5 and Figure 6 illustrate this observation by comparing the miss rate of the three aggregation methods. This is mainly due to the failure of the sampling strategy in capturing the visual patterns similarity in subsequent frames.

4) Experiment on CEDBMAL

In this experiment, we analyze the importance of involving a dynamic batch selection to address the variable annotation cost issue and improve the performance. To this end, a group of frames, with best batch size B, is sampled according to the labeling time cost of frames and the distribution of objects over them.

TABLE VI shows the results of miss rate according to "avg" aggregation function and the best B value selected at each CEDBMAL cycle. It is observed that the miss rate, in the 2^{nd} cycle, is decreased to 57% with only 783 labeled frames that contain ≅ 2226 pedestrians. The same miss rate is obtained using random sampling method after 6 cycles (3000 selected frames for labeling containing ≅ 3531 pedestrians as reported in TABLE II), and using WCR-DAL method in 3^{thd} cycle, but at the cost of more labeled frames (1062 frames which contain ≅ 3195 pedestrians) and a fixed group size (see TABLE IV). Such observation is explained by two reasons: (1) the picking up, in cost-aware manner, of the best group with few important frames that contain relatively diverse and fewer (but more informative) detected pedestrians. (2) The integration of WCR uncertainty to estimate pedestrian objects amount during batch sampling and optimal batch size selection.

5) Comparisons with State of the Art Approaches

In the following parts, we compare the detection performance of our proposed method with baseline pedestrian detector and after that compare it with existing DAL technique in the related literature for training pedestrian detector.

- *Comparisons with baseline pedestrian detector*:

In this part, we compare the results of using our DAL strategy versus standard training strategies for building pedestrian detector. The experiment is conducted using representative shallow learning (handcrafted feature)-based and deep (feature) learning-based pedestrian detectors whose results are published on Caltech Pedestrian detection benchmark [53], [54]. All methods considered here were trained on fully labeled Caltech-USA and INRIA datasets without referring to DAL algorithms.

TABLE VI. Evaluation performance results for CEDBMAL experiment (with "avg" as aggregation method)

	cyc	#B	#IBx	#BxC	#Bx	FPPI
CEDBMAL	0	500	218	486	486	85%
	1	900	707	1507	1993	61%
	2	100	**783**	233	2226	**57%**
	3	500	1025	670	2896	55%
	4	500	1276	632	3528	56%
	5	100	1399	159	3687	57%
	6	500	1501	364	4051	57%
	7	100	1600	147	4198	55%
	8	500	1773	385	4583	55%
	9	100	1828	108	4691	55%
	10	100	1909	104	4795	55%
	11	350	2006	192	4987	**53%**

cyc:cycles ,**#B**:Best-batch-size ,**#IBx**:Number-images-with-Bboxes ,**#BxC**:Number-detected-Bboxes-per-cyc ,**#Bx**:Number-detected-Bboxes#, **CEDBMAL**: CEDBMAL training, agg func= "Avg"

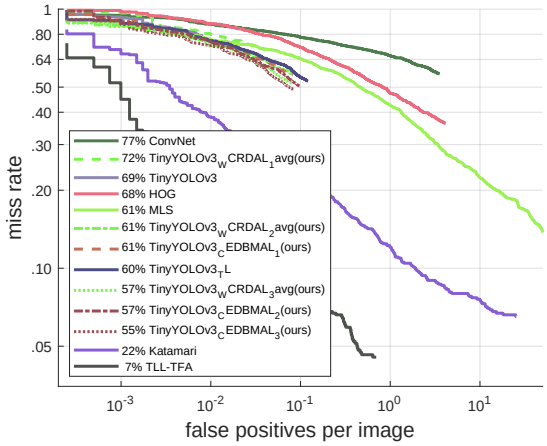

Figure 7. Pedestrian detection on the Caltech Pedestrian dataset.

Figure 7 provides quantitative results in terms of miss rate and false-positive per image (FPPI). The result depict that the deep learned features train a more accurate pedestrian detector than handcrafted features. This is due to the fact of the model's sensitivity towards the training strategy and the amount of data used for knowledge learning.

Moreover, the results report that the subset of labeled training data, actively selected and accumulated by our proposed methods, is enough to yield the best performance and outperforms some pedestrian detectors with more than 14% reduction in miss rate in the early DAL cycles (57% MR of CEDBMAL with 2226 labeled pedestrian objects against 57% MR of WCR-DAL with 3503 labeled

TABLE VII. Listing of methods for pedestrian detection considered in comparison on Caltech-USA dataset

method	Td	Ts	Fe	Cl	#LtD	#Dp	MR
ConvNet [55]	INRIA	Sot	learning(Pixels)	DeepNet	21845	-	0.77
TinyYOLOv3_WCRDAL_1avg(ours)	Caltech	AL + TL	learning(Pixels)	DeepNet	418	1405	0.72
TinyYOLOv3	COCO	Pre	learning(Pixels)	DeepNet	165482	-	0.69
HOG [56]	INRIA	Scratch	handcrafted	LinearSVM	≅ 14658	-	0.68
MLS [57]	INRIA	boosting	handcrafted	Adaboost	≅ 14658	-	0.61
TinyYOLOv3_WCRDAL_2avg(ours)	Caltech	AL + TL	learning(Pixels)	DeepNet	849	2742	0.61
TinyYOLOv3_CEDBMAL_1(ours)	Caltech	AL + TL	learning(Pixels)	DeepNet	707	1993	0.61
TinyYOLOv3_TL	Caltech	TL	learning(Pixels)	DeepNet	2006	-	0.59
TinyYOLOv3_WCRDAL_3avg(ours)	Caltech	AL + TL	learning(Pixels)	DeepNet	1062	3195	0.57
TinyYOLOv3_CEDBMAL_2(ours)	Caltech	AL + TL	learning(Pixels)	DeepNet	783	2226	0.57
TinyYOLOv3_CEDBMAL_3(ours)	Caltech	AL + TL	learning(Pixels)	DeepNet	1025	2896	0.55
Katamari [53]	Caltech	scratch	handcrafted	-	-	-	0.22
TLL-TFA [58]	Caltech	scratch	learning(Pixels)	DeepNet	≅ 42782	-	0.07

Td:Training-dataset ,**Ts**:Training-strategy ,**Fe**:Features ,**Cl**:Classifier ,**#Ltd**:Number-labeled-training-data ,**#Dp**:Number-detected-pedestrian ,**MR**:Miss-Rate ,**Sot**:Stochastic online training ,**TL**:Transfer learning ,**AL**:Active learning ,**Pre**:Pre-trained ,**CEDBMAL_x**: CEDBMAL training, agg func= "Avg", cycle number= "x" ,**WCRDAL_xavg**: WCRDAL training, agg func= "Avg", cycle number= "x"

pedestrian objects against 77% MR of ConvNet with fully labeled dataset).

Beyond labeling cost awareness, we can claim that using batch mode DAL together with TL could lead to efficiently train a deep learning based approaches with less amount of training data, less architecture complexity and less negative effect of outlier, redundancy data and domain shift problem.

TABLE VII reports additional details on the training data and the miss rate versus the labeled training data as well as number of detected pedestrians.

- *Comparisons with DAL-based pedestrian detector*:

In this part, we evaluate our method compared to published results of the related DAL technique [4] for training pedestrian detector, while performing per-cycle comparisons. In this comparison, we analyze the miss rate by examining the importance of the number of detected pedestrian and the batch size. Overall, the comparison settings are summarized in TABLE VIII. Quantitative results are reported in Figure 8 together with Figure 9.

As shown, the results indicate that the three DAL-based methods outperform the random sampling strategy. Since the first DAL cycle can mine hard instances, it contributes the most in terms of reducing the miss rate compared to randomly sampled instances (see Figure 8).

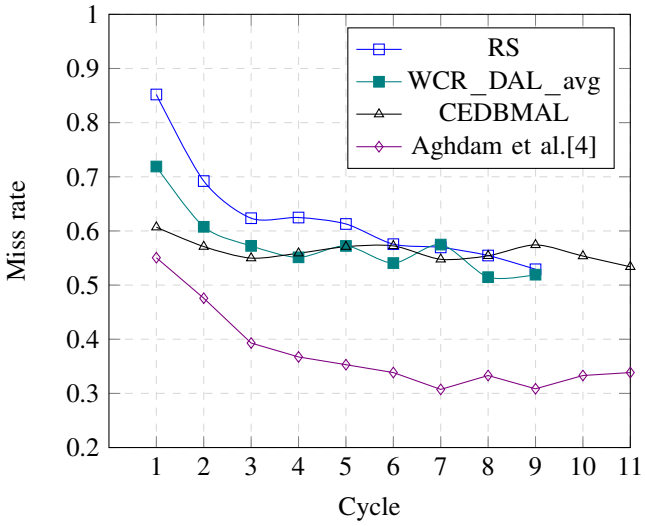

Figure 8. Miss rate of random selection vs. WCR_DAL_avg (ours) vs. CEDBMAL (ours) vs. Aghdam et al.[4] .

Nevertheless, at the end of 3^{thd} cycle, the network trained on the 1500 frames selected by the method [4] is more accurate than the networks trained on same number of selected frames by our DAL methods. From Figure 8 and Figure 9, we can see a reduction of miss rate about 15% under the cost of labeling ≅ 1900 predicted pedestrian instances against 15% with the cost of labeling ≅ 3195 and about 20% for a labeling cost of 2K predicted pedestrian instances. This is due to the ability of the method [4] to query the labeling of the most useful detected pedestrian instances, which provide more knowledge about the target object to the network.

Compared to per-instance sampling strategies, our pro-

TABLE VIII. Comparison settings of our method to Aghdam et al.[4] .

method	Td	OD	Qs	#Cyc	Bs	B	Sf	Af
Aghdam et al.[4]	CP, C, BD	dDNa	oBo	14	Fi	500	pSf, MC-D, En	dAf
WCRDAL(ours)	C	TYv3	oBo	9	Fi	500	Un	avg, max, sum
CEDBMAL(ours)	C	TYv3	Bat	11	Dy	Dy	Un and Di	avg

Td:Training-dataset ,**OD**:Object-detector ,**QS**:Query-strategy ,**#Cyc**:DAL-cycles-number ,**Bs**:Batch-size-selection ,**B**:Batch-size ,**Sf**:Scoring-functuin ,**Af**:Aggregation-function ,**dDNa**:defined-Deep-Network-architecture ,**dAf**:defined-Aggregation-function ,**MC-D**:Monte Carlo-Dropout ,**Fi**:fixed (static) ,**pSf**:pixel-level Sf ,**En**:entropy ,**oBo**:one-by-one query method ,**CP**:CityPerson Pedestrian ,**BD**:BDD100K ,**Un**:Uncertainty ,**Di**:Diversity ,**Bat**:batch query method ,**CP**:TinyYOLOv3 ,**TYv3**:TinyYOLOv3 ,**Dy**: dynamic ,**C**: Caltech ,**CEDBMAL**: CEDBMAL training

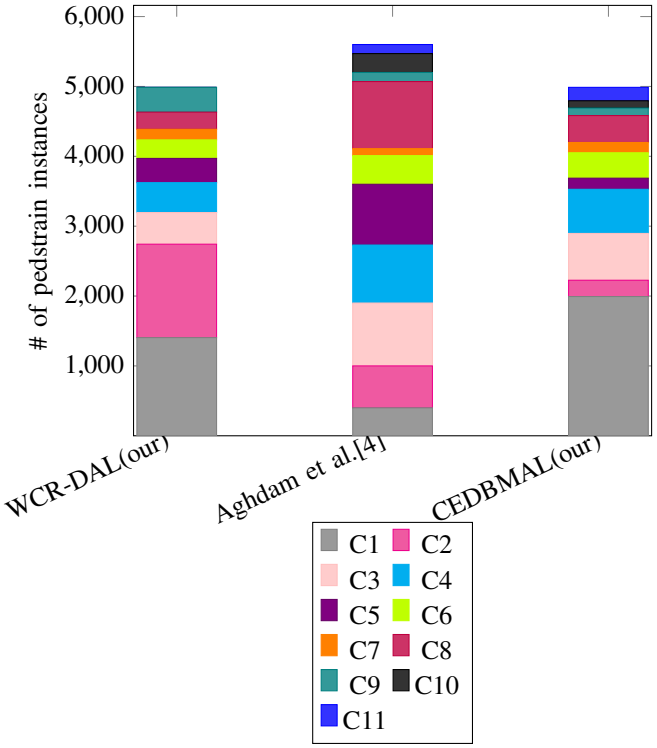

Figure 9. Number of pedestrian instances in training set at each DAL cycle

posed adaptive batch query strategy performs better than our WCR-DAL method and exhibits a performance close to the method reported in [4]. This observation is emphasized by the gains in handling the bounding box distribution, across DAL cycles, regardless of the underlying detector architecture. Consequently, CEDBMAL's dynamic selection of batch size based on object amount not only helps to effectively maintain the cost of data labeling, but also reduces DAL selection cycles and naturally supports the commonly used mini-batch training concept.

C. Discussion

As could be seen, the reported results on Caltech Pedestrian dataset are very promising. However, the transfer learning presented a worse performance. Such observation can be explained by the fact that the taking into account outliers and redundant data, during the training process, degraded the detector model performance. Because it does not suffer from these issues, the random sampling selection technique, coupled with transfer learning, surpasses the transfer learning method. Compared to previous methods, our proposed DAL method effectively decreases the detection loss while minimizing the annotation and training costs and dealing with the negative influence of noisy training data. Regardless of the aggregation method being used, both U_c and WCR query strategies can discover gradually more knowledge from few frames, leading to min more informative boxes (hard examples) that provide good signal for fast convergence and annotation cost reduction, while the overall performance of both strategies remains close to each other. Even thought WCR-DAL could select high uncertainty frames with more pedestrian objects in early cycles, similar object distribution in consecutive frames does not always yield an improvement and yet decreases the performance of the detector. Throughout the adaptive selection of best batch according to its size, target object distribution and annotation cost, it is clear that CEDBMAL cost-effectively fine-tunes more robust CNN-based detection model and conserves detection loss close to existing performance results. This is due to maintain outliers' selection and diversity between selected frames, which is highly expected to decrease the detection loss while saving annotation time within a given budget. However, the success of our method is a matter of critical factors, namely the underlying detector architecture complexity, scoring functions and the query strategy. Although existing object detection algorithms have achieved good results, it is still a challenge to effectively handle the correlation between sample selection criteria, dynamic batch selection, and noisy data identification to minimize the cost of data labeling.

6. Conclusions and Future Work

In this paper, a novel, cost-effective deep active learning framework for object detection was proposed. An adaptive batch sampling strategy was designed. At the frame-level, a set of batches, consisting of the top-ranked samples within the batch, is actively selected based on both WCR uncertainty and diversity. Afterward, an annotation cost-based batch selection is performed considering detected objects that provide training benefits. The results demon-

strated that CEDBMAL framework contributes to a smaller decrease in detection loss, but a greater reduction in manual annotation effort, with less than 50% of labeled data, as well as an effective avoidance of the impractical batch size determination, the equal cost assumption and the budget exceedance. In addition, providing a deeper insight regarding cost-effective batch size selection by applying optimization techniques, such as dynamic programming, in DAL. However, its major shortcomings are the negative influence of class imbalance and similar instances between batches, as well as the increased learning time trigged by the active learning cycle. As a future work, we will address the consistency metric in minimizing labeling cost of batch DAL. Furthermore, integrating other learning paradigms to reduce human supervision.

References

[1] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE access.*, vol. 8, pp. 58 443–58 469, 2020.

[2] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3d object detection methods for autonomous driving applications," *IEEE Trans on Intell Trans Sys.*, vol. 20, no. 10, pp. 3782–3795, 2019.

[3] Q. Rao and J. Frtunikj, "Deep learning for self-driving cars: Chances and challenges," in *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems*. IEEE, 2018, pp. 35–38.

[4] H. H. Aghdam, A. Gonzalez-Garcia, J. v. d. Weijer, and A. M. López, "Active learning for deep detection neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 2019, pp. 3672–3680.

[5] B. Settles, "Active learning literature survey," *Computer Sciences Technical Report 1648, University of Wisconsin-Madison*, 2009.

[6] ——, "Active learning," ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012, vol. 6, pp. 1—-114, https://doi.org/10.2200/s00429ed1v01y201207aim018.

[7] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, "A survey of deep active learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 9, pp. 1–40, 2021.

[8] M. Wu, C. Li, and Z. Yao, "Deep active learning for computer vision tasks: Methodologies, applications, and challenges," *Applied Sciences*, vol. 12, no. 16, p. 8103, 2022.

[9] Z. Qu, J. Du, Y. Cao, Q. Guan, and P. Zhao, "Deep active learning for remote sensing object detection," *arXiv: Computer Vision and Pattern Recognition*, 2020.

[10] T. Yuan, F. Wan, M. Fu, J. Liu, S. Xu, X. Ji, and Q. Ye, "Multiple instance active learning for object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5326–5335.

[11] A. L. Chandra, S. V. Desai, V. N. Balasubramanian, S. Ninomiya, and W. Guo, "Active learning with point supervision for cost-effective panicle detection in cereal crops," *Plant Methods.*, vol. 16, no. 1, pp. 1–16, 2020.

[12] D. Feng, X. Wei, L. Rosenbaum, A. Maki, and K. Dietmayer, "Deep active learning for efficient training of a lidar 3d object detector," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 667–674.

[13] S. Roy, A. Unmesh, and V. P. Namboodiri, "Deep active learning for object detection," in *BMVC*, vol. 362. BMVA, 2018, p. 91.

[14] A. Kirsch, J. v. Amersfoort, and Y. Gal, *BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[15] S. H. Park and S. B. Kim, "Robust expected model change for active learning in regression," *Applied Intelligence*, vol. 50, no. 2, pp. 296–313, 2020.

[16] Y. Zhao, Z. Shi, J. Zhang, D. Chen, and L. Gu, "A novel active learning framework for classification: using weighted rank aggregation to achieve multiple query criteria," *Pattern Recognition*, vol. 93, pp. 581–602, 2019.

[17] Y. Li, B. Fan, W. Zhang, W. Ding, and J. Yin, "Deep active learning for object detection," *Information Sciences*, vol. 579, pp. 418–433, 2021.

[18] X. Gui, X. Lu, and G. Yu, "Cost-effective batch-mode multi-label active learning," *Neurocomputing*, vol. 463, pp. 355–367, 2021.

[19] J. Wu, V. S. Sheng, J. Zhang, H. Li, T. Dadakova, C. L. Swisher, Z. Cui, and P. Zhao, "Multi-label active learning algorithms for image classification: Overview and future promise," *ACM Computing Surveys (CSUR)*, vol. 53, no. 2, pp. 1–35, 2020.

[20] T. Yao, W. Wang, and Y. Gu, "A deep multiview active learning for large-scale image classification," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[21] C. Su, Z. Yan, and G. Yu, "Cost-effective multi-instance multilabel active learning," *International Journal of Intelligent Systems*, vol. 36, no. 12, pp. 7177–7203, 2021.

[22] G. Yu, Y. Xing, J. Wang, C. Domeniconi, and X. Zhang, "Multiview multi-instance multilabel active learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2021.

[23] J. Guo, Z. Pang, M. Bai, P. Xie, and Y. Chen, "Dual generative adversarial active learning," *Applied Intelligence*, pp. 1–12, 2021.

[24] J. Wang, Y. Yan, Y. Zhang, G. Cao, M. Yang, and M. K. Ng, "Deep reinforcement active learning for medical image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 33–42.

[25] N. Nadagouda, A. Xu, and M. A. Davenport, "Active metric learning and classification using similarity queries," *arXiv preprint arXiv:2202.01953*, 2022.

[26] N. Zemmal, N. Azizi, M. Sellami, S. Cheriguene, A. Ziani, M. AlDwairi, and N. Dendani, "Particle swarm optimization based swarm intelligence for active learning improvement: Application on medical data classification," *Cognitive Computation*, vol. 12, no. 5, pp. 991–1010, 2020.

[27] Z. Deng, Y. Yang, K. Suzuki, and Z. Jin, "Fedal: An federated active learning framework for efficient labeling in skin lesion analysis," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2022, pp. 1554–1559.

[28] Q. Jin, M. Yuan, Q. Qiao, and Z. Song, "One-shot active learning for image segmentation via contrastive learning and diversity-based sampling," *Knowledge-Based Systems*, vol. 241, p. 108278, 2022.

[29] N. Bougie and R. Ichise, "Goal-driven active learning," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS '22. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2022, p. 1923–1925.

[30] J. Shim and S. Kang, "Domain-adaptive active learning for cost-effective virtual metrology modeling," *Computers in Industry*, vol. 135, p. 103572, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166361521001792

[31] K. Wang, L. Lin, X. Yan, Z. Chen, D. Zhang, and L. Zhang, "Cost-effective object detection: Active sample mining with switchable selection criteria," *IEEE transactions on neural networks and learning systems.*, vol. 30, no. 3, pp. 834–850, 2018.

[32] P. Mi, J. Lin, Y. Zhou, Y. Shen, G. Luo, X. Sun, L. Cao, R. Fu, Q. Xu, and R. Ji, "Active teacher for semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 482–14 491.

[33] H. V. Vo, O. Siméoni, S. Gidaris, A. Bursuc, P. Pérez, and J. Ponce, "Active learning strategies for weakly-supervised object detection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*. Springer, 2022, pp. 211–230.

[34] J. K. Mandivarapu, B. Camp, and R. Estrada, "Deep active learning via open-set recognition," *Frontiers in Artificial Intelligence*, vol. 5, p. 2, 2022.

[35] K.-P. Ning, X. Zhao, Y. Li, and S.-J. Huang, "Active learning for open-set annotation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 41–49.

[36] P. Kumar and A. Gupta, "Active learning query strategies for classification, regression, and clustering: a survey," *Journal of Computer Science and Technology*, vol. 35, no. 4, pp. 913–945, 2020.

[37] W. Yu, S. Zhu, T. Yang, and C. Chen, "Consistency-based active learning for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3951–3960.

[38] J. Wu, J. Chen, and D. Huang, "Entropy-based active learning for object detection with progressive diversity constraint," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9397–9406.

[39] S. Sivaraman and M. M. Trivedi, "Active learning for on-road vehicle detection: A comparative study," *Machine vision and applications.*, vol. 25, no. 3, pp. 599–611, 2014.

[40] T. Yang, J. Li, Q. Pan, C. Zhao, and Y. Zhu, "Active learning based pedestrian detection in real scenes," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4, 2006, pp. 904–907.

[41] A. Moses, S. Jakkampudi, C. Danner, and D. Biega, "Localization-based active learning (local) for object detection in 3d point clouds," in *Geospatial Informatics XII*, vol. 12099. SPIE, 2022, pp. 44–58.

[42] Z. Liang, X. Xu, S. Deng, L. Cai, T. Jiang, and K. Jia, "Exploring diversity-based active learning for 3d object detection in autonomous driving," *arXiv preprint arXiv:2205.07708*, 2022.

[43] D. L. Li, M. Prasad, C.-L. Liu, and C.-T. Lin, "Multi-view vehicle detection based on fusion part model with active learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 3146–3157, 2021.

[44] C.-A. Brust, C. Käding, and J. Denzler, "Active and incremental learning with weak supervision," *KI-Künstliche Intelligenz.*, pp. 1–16, 2020.

[45] T. N. Cardoso, R. M. Silva, S. Canuto, M. M. Moro, and M. A. Gonçalves, "Ranked batch-mode active learning," *Information Sciences.*, vol. 379, pp. 313–337, 2017.

[46] Y. Gu, D. Zydek, and Z. Jin, "Active learning based on random forest and its application to terrain classification," in *Progress in Systems Engineering*, ser. Advances in Intelligent Systems and Computing, H. Selvaraj, D. Zydek, and G. Chmaj, Eds. Springer, 2015, vol. 366, pp. 273–278.

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

[48] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence.*, vol. 34, no. 4, pp. 743–761, 2011.

[49] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Computer Vision and Pattern Recognition*. Springer Berlin/Heidelberg, Germany, 2018, pp. 1804–02 767.

[50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[51] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *International Conference on Learning Representations*, 2018.

[52] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE.*, vol. 109, no. 1, pp. 43–76, 2020.

[53] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II 13*. Springer, 2015, pp. 613–627.

[54] N. Ragesh and R. Rajesh, "Pedestrian detection in automotive safety: Understanding state-of-the-art," *IEEE Access*, vol. 7, pp. 47 864–47 890, 2019.

[55] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3626–3633.

[56] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[57] W. Nam, B. Han, and J. H. Han, "Improving object localization using macrofeature layout selection," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 1801–1808.

[58] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation," *arXiv preprint arXiv:1807.01438*, 2018.

Mohib Eddine Khebbache (mohibeddine-khabache@univ-eloued.dz) is serving as an assistant professor of computer science at the University of El Oued, Algeria. He received the engineer degree in computer science from the University of Biskra, Algeria, and the MS degree in computer science from the University of Batna, Algeria, in 2007 and 2014, respectively. He is currently working toward the PhD degree at the University of Biskra. His research interests include wireless networks and machine learning methods for autonomous driving.

Salim Bitam (s.bitam@univ-biskra.dz) is a full professor in the Computer Science Department at the University of Biskra, Algeria, as well as a senior member of the LESIA Laboratory at the University of Biskra, and an associate member of the LiSSi Laboratory at the University of ParisEst Créteil VdM (UPEC), France. He received an Engineer degree in computer science from the University of Constantine, Algeria, his Master's and Ph.D. in computer science from the University of Biskra, and a Doctorate of Sciences (Habilitation) diploma from the Higher School of Computer Science - ESI, Algiers, Algeria. His main research interests are vehicular ad hoc networks, cloud computing, and bio-inspired methods for routing and optimization. He has to his credit more than 30 publications in journals, books, and conferences, for which he has received two best paper awards. He has served as an editorial board member and a reviewer of several journals for IEEE, Elsevier, Wiley, and Springer, and on the technical program committees of several international conferences (IEEE GLOBECOM, IEEE ICC, IEEE/RSJ IROS, and others).

Abdelhamid Mellouk (mellouk@u-pec.fr) is a Full Professor at University of Paris-Est (UPEC), Networks & Telecommunications (N&T) Department and LiSSi/TincNet Laboratory France. He graduated in computer network engineering from the Computer Science High Eng. School, University Oran-EsSenia, Algeria, and the University of Paris Sud XI Orsay, received the Ph.D. in computer science from the same university, and a Doctorate of Sciences (Habilitation) diploma from UPEC. Founder of the Network Control Research activity in UPEC with extensive international academic and industrial collaborations, his general area of research is in adaptive realtime bio-inspired control for high-speed new generation dynamic wired/wireless networking in order to maintain acceptable quality of service/experience for added value services. He is an active member of the IEEE Communications Society and held several offices including leadership positions in IEEE Communications Society Technical Committees. He has published/coordinated 11 books, 3 lecture notes, and several refereed international publications in journals, conferences, and books, in addition to numerous keynotes and plenary talks in flagship venues. He serves on the Editorial Boards or as Associate Editor for several journals, and he is chairing or has chaired (or co-chaired) some of the top international conferences and symposia, including the TPS Chair of ICC 2017.