Multi-label Scene Classification for Autonomous Vehicles: Acquiring and Accumulating Knowledge from Diverse Datasets

Ke Li, *Graduate Student Member, IEEE*, Chenyu Zhang, *Graduate Student Member, IEEE*, Yuxin Ding, Xianbiao Hu, Ruwen Qin*, *Member, IEEE*

***Abstract*—Driving scenes are inherently heterogeneous and dynamic. Multi-attribute scene identification, as a high-level visual perception capability, provides autonomous vehicles (AVs) with essential contextual awareness to understand, reason through, and interact with complex driving environments. Although scene identification is best modeled as a multi-label classification problem via multitask learning, it faces two major challenges: the difficulty of acquiring balanced, comprehensively annotated datasets and the need to re-annotate all training data when new attributes emerge. To address these challenges, this paper introduces a novel deep learning method that integrates Knowledge Acquisition and Accumulation (KAA) with Consistency-based Active Learning (CAL). KAA leverages monotask learning on heterogeneous single-label datasets to build a knowledge foundation, while CAL bridges the gap between single- and multi-label data, adapting the foundation model for multi-label scene classification. An ablation study on the newly developed Driving Scene Identification (DSI) dataset demonstrates a 56.1% improvement over an ImageNet-pretrained baseline. Moreover, KAA-CAL outperforms state-of-the-art multi-label classification methods on the BDD100K and HSD datasets, achieving this with 85% less data and even recognizing attributes unseen during foundation model training. The DSI dataset and KAA-CAL implementation code are publicly available at https://github.com/KELISBU/KAA-CAL.**

***Index Terms*—Autonomous Vehicles, Driving Scene Understanding, Foundation Model, Deep Learning**

I. Introduction

DRIVING scene identification involves assigning non-exclusive labels to each scene captured by an onboard camera. As illustrated in Fig. 1, a complex scene is characterized using various attributes, such as time of day, weather, roadway function, intersection type, among others. Scene identification is a high-level perception crucial for autonomous vehicles (AVs), as it provides the contextual awareness for reasoning and decision-making [1], [2]. For example, the friction coefficient of the road surface varies depending on whether the road is dry, wet, snowy, or icy. Identifying driving scenes based on the road surface condition allows vehicles to adjust their estimations of braking distance and turning speed accordingly, thus improving safety. Driving scene identification also supports data generation, as it provides scene attribute labels that serve as semantic priors to guide image and video synthesis. For example, after aligning visual and semantic representations, the intersection type class serves as an explicit conditioning information for fine-grained generation (e.g., three-way, four-way, or roundabout).

Driving scenes are inherently heterogeneous and dynamic, requiring the simultaneous recognition of many attributes. This makes scene identification a multi-label classification problem that goes far beyond a simple perception task. The complexity nature of this important problem highlights the critical need for a foundation model possessing the comprehensive knowledge to accurately characterize these intricate scenes.

While seemingly straightforward for driving scene identification, standard multitask learning methods face critical limitations. A primary challenge is the lack of a comprehensively annotated multi-label training dataset. In a high-dimensional attribute space, driving scene distribution is highly imbalanced [3], [4]. Some scenarios are rare, making it difficult to collect balanced multi-label training samples. This likely explains why image- and video-level annotations in popular driving scene datasets are either single-labeled (e.g., [5], [6]) or multi-labeled with just limited scene attributes (e.g., [7], [8]), and why existing models are predominantly single-label classification models (e.g., [9]–[11]). Meanwhile, as the transportation system evolves, the need to identify additional attributes constantly emerges. However, the standard multitask learning framework proves inefficient for integrating desired new knowledge. This is because it requires not only re-labeling entire training datasets with the required new labels but also a further data collection, a process that is neither scalable nor cost-effective.

The aforementioned challenges, combined with the abundance of single-label datasets and the convenience of monotask learning, motivate the exploration of a learning approach leading to a driving scene foundation model. As illustrated in Fig. 1, this strategy constructs the model by learning from heterogeneous datasets and consolidating transferable representations. The resulting model can not only recognize complex driving scenes using a variety of attributes but also quickly learn to recognize new ones.

The learning process to be introduced in this paper is inspired by how students learn. A student model concentrates on learning one classification task from a corresponding single-label dataset at a time in a sequential manner and consolidates the acquired knowledge in a teacher model, which serves as

Ke Li, Chenyu Zhang, and Ruwen Qin are with the Department of Civil Engineering, Stony Brook Univerisity, Stony Brook, NY 11794, USA.

Yuxin Ding and Xianbiao Hu are with the Department of Civil and Environmental Engineering, Pennylvania State University, University Park, PA 16802, USA.

* Corresponding author: Ruwen Qin, email: ruwen.qin@stonybrook.edu

Manuscript received Month dd, yyyy; revised Month dd, yyyy.

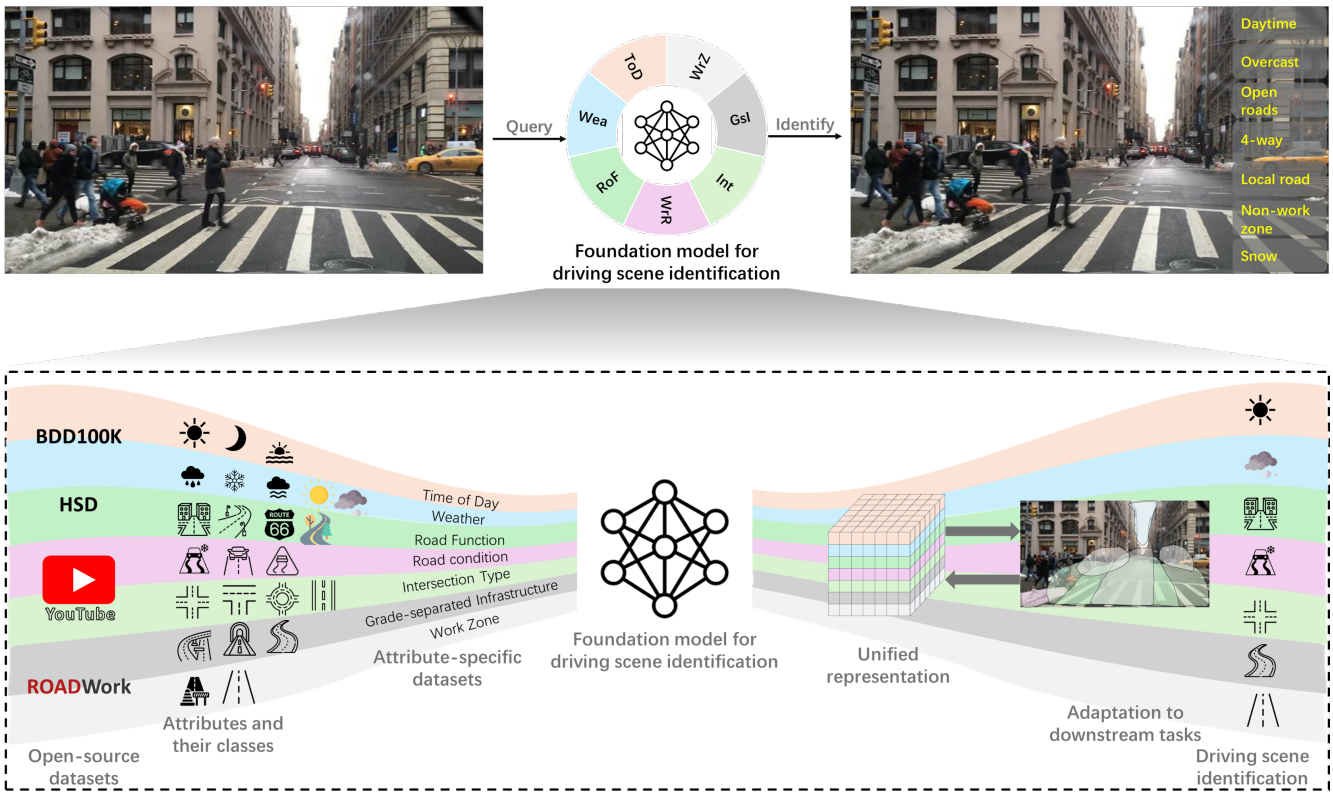


Fig. 1. A foundation model for multi-attribute driving scene identification, with its comprehensive knowledge obtained from diverse single-label datasets

a reference for the student model's continual learning and improvement. This sequential process goes cyclically, progressively transforming knowledge from heterogeneous single-label datasets into a unified representation, forming a foundation model for multi-attribute driving scene identification. Yet, this approach confronts an inevitable domain shift problem, which arises from the discrepancy between the marginal distributions of individual attributes and their joint distribution. To mitigate this, a strategy is developed to guide the model's adaptation, transforming the knowledge it acquires from diverse single-label datasets into a comprehensive understanding suitable for complex scene recognition.

This paper reports an effort in designing and validating the described new deep learning method for complex driving scene identification. In addressing challenges confronted in this study, the paper makes the following contributions:

- A new deep learning system, **K**nowledge **A**cquisition and **A**ccumulation (KAA), is proposed to learn from heterogeneous single-label datasets and consolidate learned knowledge into a unified representation, laying the groundwork for multi-label driving scene identification.
- In demonstrating KAA, a comprehensive and scalable, **D**riving **S**cene **I**dentification (DSI) dataset is constructed, comprising seven single-label subsets, each annotated according to one specific scene attribute, representing heterogeneous sources of information about driving scenes.
- An adaptation algorithm, **C**onsistency-based **A**ctive **L**earning (CAL), is developed to couple with KAA, addressing the domain shift problem it confronts.

Remainder of the paper is organized as follows. Section II summarizes recent work related to this study. Then, Section III details the components of the new DSI dataset. After that, the proposed KAA-CAL learning method is introduced in Section IV, with the implementation delineated in Section V. Section VI further reports the experimentation and results. At the end, Section VII concludes the study by summarizing important findings and future research directions.

II. Related Work

This paper is built upon a spectrum of research areas. The most relevant literature from these fields is summarized below.

A. Public Driving Scene Datasets

Various driving video datasets have been developed to support different perception tasks for AVs, including scene classification. BDD100K [7] is a large-scale driving video dataset featuring image-level annotations for six weather conditions, six scene types, and three distinct times of day. Honda Scene Dataset (HSD) [8] contains 80 hours of driving video data clips collected in the San Francisco Bay Area. HSD provides video-level identification of eleven road places, four categories of road environment, and four weather conditions, thereby broadening the diversity of both attributes and classes for urban scene identification. ROADWork [6] is a public dataset focused on work zones identification with data collected from eighteen cities in the U.S. This dataset also provides fine-grained annotations for instance and semantic segmentation. DENSE++ [12] provides multi-label annotations

for environmental conditions including daytime, precipitation, fog, road condition, roadside condition, and scene-setting for 12,997 images collected from Northern Europe. While the annotation method is provided, the image labels are not publicly available. WZ-Traffic [5] is a scene dataset with 6,035 single-label images in 20 categories. Other open-source driving datasets, including Cityscapes [13], KITTI [14], and nuScenes [15], are primarily annotated for object detection, semantic segmentation, and instance segmentation tasks.

Few datasets provide comprehensive annotations for multi-label scene classification. Some provide class labels either at the frame level or the video level, yet they do so based on one or a few attributes of their interest. Integrating the information in heterogeneous datasets as a unified representation of driving scenes is greatly desired.

B. Multi-label Driving Scene Classification

Multi-attribute scene identification is commonly modeled as a multi-label image classification proble,. A few studies have tackled this problem in various approaches. Duong et al. [3] developed CF-Net, which uses single-label classifiers to enhance the main multi-label classifier by utilizing feature fusion and stacking. The model was trained on a modified BDD100K dataset with three attributes: location, weather, and time of day. Chen et al. [4] proposed to solve the multi-label scene classification problem by incorporating the single-label training procedure into the multi-label architecture. Additionally, a deep data integration strategy was utilized to improve the classification ability. RECNet [12] proposed a hierarchical strategy for annotating six environmental conditions, including daytime, fog, precipitation, road condition, roadside condition, and infrastructure, leading to a fully annotated dataset DENSE++. This dataset was used to train a multi-label classification model, which uses the EfficientNet-B2 as the shared backbone for the six downstream classifiers. Prykhodchenko and Skruch [16] also trained a multi-label scene classification model on BDD100K via multitask learning. Data augmentation and balancing strategies were utilized to address the class imbalance issue in the BDD100K dataset.

Existing methods generally adhere to the standard multi-task learning framework, thus relying on multi-label training datasets. New methodologies that can transcend this framework will offer an opportunity to utilize the abundant driving scene images or video data.

C. Knowledge Distillation

Knowledge distillation (KD) typically utilizes a teacher-student framework to distill the knowledge accumulated in a deep or large model (the teacher) into a smaller or shallow one (the student). It supports the student network's continual acquisition and consolidation of scene identification knowledge.

KD methods are generally categorized into three types: response-based KD, where the student directly mimics the teacher's final predictions; feature-based KD, which transfers knowledge by leveraging features extracted by the teacher; and relation-based KD, which explores relationships between different layers or samples [17]. When both the teacher and student networks are deep, regulating multiple hidden layers of the student network leads to improved performance compared to response-based KD [18]. Ma et al. [19] introduced ARK to accrue and reuse feature-based knowledge, addressing the challenge of inconsistent annotations in various medical image datasets. Using ARK, a foundation model is constructed for medical image analysis. When KD occurs progressively, there is a risk of forgetting previously acquired knowledge. To mitigate this issue, Van de Ven et al. [20] proposed a brain-inspired method that replays internal or hidden representations in continual learning. While sharing some similarities with proposed KAA, ARK primarily focuses on addressing annotation heterogeneity across different medical datasets.

D. Deep Active Learning

Deep active learning can efficiently adapt models to new domains or tasks, achieved by strategically selecting the most informative and representative samples for adaptation. Some uncertainty-based methods such as maximum entropy, margin sampling, least-confidence sampling, and Bayesian Active Learning by Disagreement rely on uncertainty measures based on a task [21]. In contrast, a task-agnostic method introduced in [22] integrates a loss prediction module to select top samples of the highest prediction loss from the unlabeled pool. Uncertainty-based methods may cause sample redundancy.

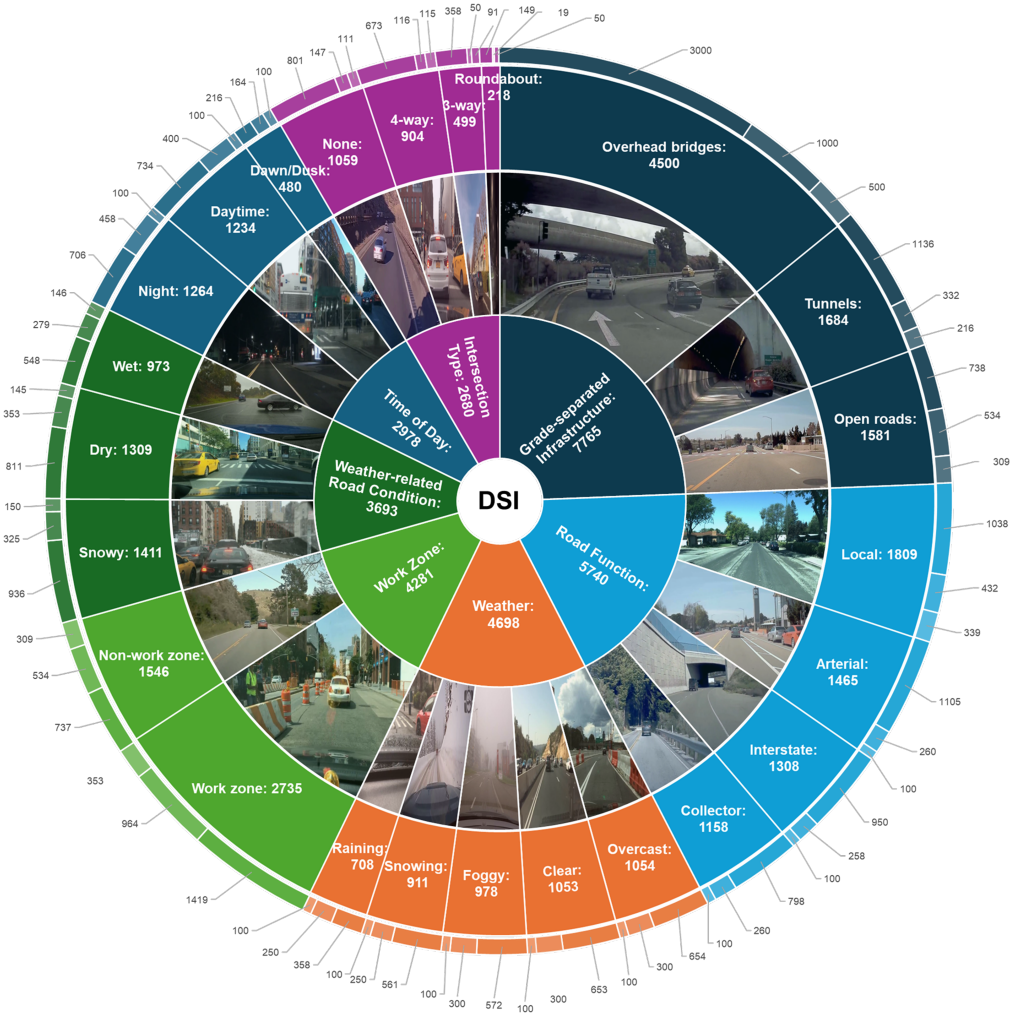
A diversity-based method selects instances from the unlabeled pool to represent a broad data distribution. Beyond traditional clustering methods like K-means, Core-Set Section [23] computes the distance of features from a designated hint layer within a deep learning model. A greedy search algorithm is employed to iteratively select samples that exhibit the greatest distance from their nearest neighbors until the selection budget is exhausted. Diversity-based methods may overlook the most informative instances.

Hybrid methods, including Wasserstein Adversarial Active Learning and Batch Active learning by Diverse Gradient Embeddings, effectively address limitations of both uncertainty- and diversity-based methods [21]. For example, Hekimoglu et al. [24] quantified uncertainty using an inconsistency score computed as the maximum loss between the initial-refined pairwise task predictions. Simultaneously, a reconstructed feature embedding that condenses information across all tasks measures the diversity score. The combination of these two components forms the basis for deep active learning. Yet this approach cannot handle the domain shift caused by monotask learning from diverse single-label datasets.

III. Driving Scene Identification (DSI) Dataset

The data distribution in the high-dimensional scene attribute space is highly imbalanced. Some combinations of scene classes are corner cases difficult to collect data for. This limits the creation of comprehensively annotated datasets to directly train multi-label scene classification models via multitask learning. However, collecting data from respective marginal distributions of scene attributes is relatively straightforward.

This paper introduces DSI, a heterogeneous single-label dataset for demonstrating KAA-CAL, the proposed learning



Attributes and classes	Trn	Vld	Tst
Grade-separated Infrastructure	4,874	1,866	1,025
Overhead bridges	3,000	1,000	500
Tunnels	1,136	332	216
Open roads	738	534	309
Road Function	3,891	1210	639
Local	1,038	432	339
Arterial	1,105	260	100
Interstate	950	258	100
Collector	798	260	100
Weather	2,798	1,400	500
Overcast	654	300	100
Clear	653	300	100
Foggy	572	300	100
Snowing	561	250	100
Raining	358	250	100
Work Zone	2,121	1,498	662
Work zone	1418	964	353
Non-work zone	703	534	309
Weather-related Road Condition	2,295	957	441
Snowy	936	325	150
Dry	811	353	145
Wet	548	279	146
Time of Day	1,656	1,022	300
Night	706	485	100
Daytime	734	400	100
Dawn/Dusk	216	164	100
Intersection Type	1,981	332	367
None	801	147	111
4-way	673	116	115
3-way	358	50	91
Roundabout	149	19	50
Total	19,616	8,285	3,934

Fig. 2. Overview of the DSI dataset

method. DSI consists of 31,835 scene images sampled from public driving video datasets, including BDD100K [7], HSD [8], and ROADWork Data [6], and is supplemented with additional images sampled from YouTube videos. An overview of the DSI dataset and its statistics are presented in Fig. 2. The dataset is available to the public at https://github.com/KELISBU/KAA-CAL.

The DSI dataset consists of M (7) subsets, denoted as $\{\mathcal{D}_m\}_{m=1}^{M}$, where m is the index of attributes for scene identification. Each subset is annotated according to one unique scene attribute and further partitioned into training (Trn), validation (Vld), and test (Tst) sets:

$$\mathcal{D}_m = \mathcal{D}_m^{\text{Trn}} \cup \mathcal{D}_m^{\text{Vld}} \cup \mathcal{D}_m^{\text{Tst}}. \tag{1}$$

Fig. 2 illustrates the scene attributes and respective classes. In total, DSI provides image-level labels for 24 classes, with each falling exclusively under one of the seven scene attributes. Classifying driving scenes according to these scene attributes offers valuable insights for AVs. Notably, DSI can be conveniently expanded by adding new single-label subsets or new classes within existing subsets (e.g., a new design of intersection type) as needs for new knowledge about scene identification emerge.

Grade-separated Infrastructure. Road segments featuring grade-separated infrastructure, such as bridges and tunnels, often present specific challenges for AVs. Bridges above the road may limit the sensor field of view, cause sudden lighting changes, and introduce gusts of wind. Tunnels, on the other hand, have different lighting conditions, limited visibility, and potential congestion. The Grade-separated Infrastructure subset in DSI comprises three classes: Overhead bridges, Tunnels, and Open roads, totaling 7,765 images. This subset was curated from the HSD dataset using query keywords like "overhead bridge" and "tunnel". To ensure these structures are visually perceivable, additional query criteria such as "Approaching" and "Entering" were used for data sampling.

Road Function. Each road type has specific characteristics, such as a speed limit range, traffic density, vehicle type distribution, terrain, and travel purposes. Thus, understanding the road function assists adaptive driving behavior. The US

road functional system has not been systematically annotated in public datasets. To address this, the Road Function subset was created by querying with keywords like "road function" on YouTube and sampling every fifth frame from the collected videos. This dataset consists of 5,740 images across 4 classes: Local, Arterial, Interstate, and Collector.

Weather. Identification of the weather condition is critical because an adverse weather condition risks driving safety and the performance of visual tasks. The Weather subset includes 4,698 frames sampled from the BDD100K dataset, encompassing Overcast, Clear, Foggy, Snowing, and Raining conditions. To mitigate the issue of data imbalance caused by the rarity of foggy scenes, synthetic images were generated to ensure a more balanced distribution across weather conditions.

Work Zone. Driving safety is a concern in work zones, where construction equipment, temporary barriers, road workers, and other related facilities are present. Work zones often involve lane closures. Confusion arises when old, not fully removed lane markings mix with new ones, especially for AVs. Identifying work zones allows AVs to adjust their lane position and speed dynamically, minimizing crash risks. The training and validation data of the Work Zone subset mainly come from BDD100K and HSD, and its testing data are primarily from the ROADWork dataset.

Weather-related Road Condition. The friction coefficient of the road surface changes depending on weather conditions like snow or rain. Identifying these weather-related road conditions allows AVs to adjust their control for better adaptation to the specific surface characteristics. The Weather-related Road Condition subset includes 3,693 images collected from both YouTube and BDD100K, intended for identifying Snowy, Dry, and Wet conditions.

Time of Day. The lighting condition directly impacts the visibility of objects, roads, and the surrounding environment because illumination influences key visual functions like contrast sensitivity, visual acuity, depth perception, and peripheral vision. Consequently, the lighting condition affects not only human drivers' visual perception and reaction times but also machine vision due to the affected quality of camera-captured data. For example, nighttime and dawn/dusk images often exhibit increased noise, reduced contrast, and color distortion. The Time of Day subset comprises 2,978 images labeled as Night, Daytime, or Dawn/Dusk, sampled from the BDD100K dataset.

Intersection Type. More than half of the combined total of fatal and injury crashes occur at or near intersections due to increased conflict points and complex interactions among vehicles, pedestrians, and other road users. Recognizing the type of intersection allows AVs to anticipate potential conflict points and make informed decisions, such as yielding, stopping, or merging in complex environments. The Intersection Type subset consists of 2,680 images, categorized into four classes: None, 3-way, 4-way, and Roundabout. The data were collected by integrating YouTube videos, which provide roundabout images, along with the HSD and BDD100K datasets, which contribute images of none, 3-way, and 4-way intersections. 3-way and 4-way intersections are typically more noticeable when ego-vehicles are entering the intersection. Therefore, the query keyword "entering" was used when searching for such images in the HSD dataset.

IV. Methodology

Multi-label driving scene classification involves assigning multiple non-exclusive class labels to each input scene image $\boldsymbol{x}$ in the domain $\mathcal{I} = \{\boldsymbol{x} | \boldsymbol{x} \in \mathbb{R}^{3\times H\times W}\}$, comprising color images with three color channels, a height of H pixels, and a width of W pixels. $\boldsymbol{x}$ can be classified based on M different scene attributes, Y_m, for $m = 1, \ldots, M$. Each attribute Y_m is a categorical variable with its support $\mathcal{U}_m$ comprising the exclusive classes for this attribute. A multi-label classification model $\mathcal{C}$ approximates the mapping, $\mathcal{I} \rightarrow \mathcal{U}_1 \times \cdots \times \mathcal{U}_M$, which estimates the class labels for each scene image $\boldsymbol{x}$:

$$[\widehat{\boldsymbol{y}}_1, \ldots, \widehat{\boldsymbol{y}}_M] = \mathcal{C}(\boldsymbol{x}; \boldsymbol{\Theta}), \tag{2}$$

where $\widehat{\boldsymbol{y}}_m \in \mathbb{R}^{|\mathcal{U}_m|}$ is the prediction in probability for the one-hot encoded true label $\boldsymbol{y}_m \in \mathbb{B}^{|\mathcal{U}_m|}$ of $\boldsymbol{x}$, and $\boldsymbol{\Theta}$ represents the learnable parameters of the classification model $\mathcal{C}$. To build such a model, this section proposes a new deep learning method named KAA-CAL. First, KAA learns a model $\widetilde{\mathcal{C}} = \mathcal{C}(\cdot, \widetilde{\boldsymbol{\Theta}})$ that maximizes the classification accuracy in isolated per-attribute identification. Following this, CAL further adapts $\widetilde{\mathcal{C}}$ to become $\mathcal{C}^* = \mathcal{C}(\cdot, \boldsymbol{\Theta}^*)$, one that is proficient in simultaneous all-attribute identification.

A. Knowledge Acquisition and Accumulation (KAA)

1) Overview of the Learning Process: KAA, depicted in Fig. 3, is a deep learning system designed to acquire and accumulate knowledge from various single-label datasets for multi-label driving scene classification. KAA utilizes a teacher-student network architecture. The knowledge acquisition network (KAqN) serves as the student network, responsible for learning new knowledge and consolidating it into the existing knowledge in the learning system. The knowledge accumulation network (KAcN), acting as the teacher network, stores the consolidated knowledge and applies it to guide the student network in continual learning and improvement.

KAqN's deep encoder is trained to attain the knowledge for performing M downstream tasks, indexed by m, each focused on classifying the input driving scene image according to one attribute. The overall approach of KAqN is a sequential and cyclical learning process, as outlined in Algorithm 1. A training epoch, indexed by t, is one learning cycle. During each learning cycle, KAqN goes through M learning iterations to sequentially learn on these classification tasks, utilizing a dedicated single-label dataset for each task. Therefore, the learning iteration i corresponds to a specific learning cycle t and learning task m:

$$\begin{aligned} t &= \lceil i/M \rceil, \\ m &= i - (t-1)M. \end{aligned} \tag{3}$$

Upon completion of one learning cycle, the newly acquired knowledge is consolidated into the existing knowledge stored in KAcN, forming an updated foundation for continual learning in the subsequent cycle. This cyclical and sequential learning process continues until the stopping criterion is met. KAA's architecture and design are further delineated below.

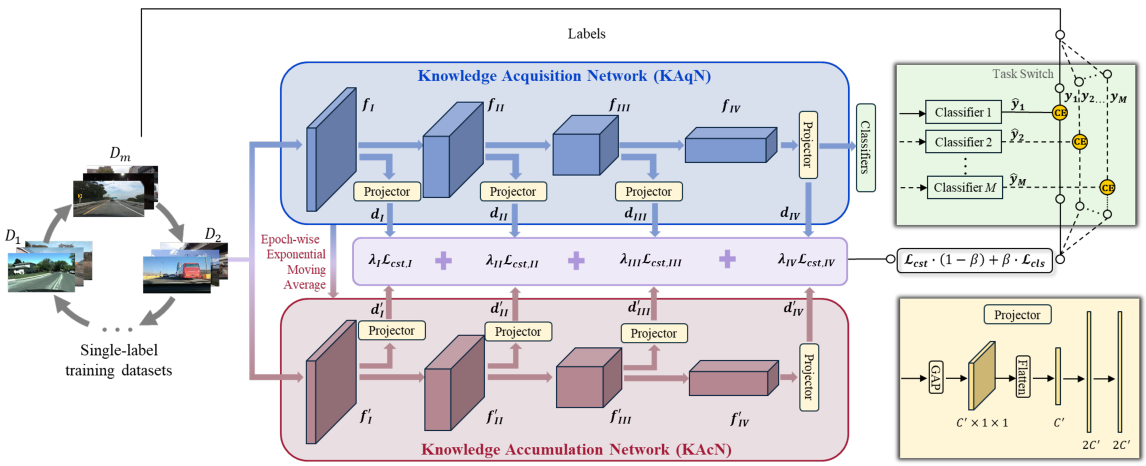


Fig. 3. Overview of the proposed Knowledge Acquisition and Accumulation (KAA) learning system

Algorithm 1: KAA Algorithm

INPUT:

$\{\mathcal{D}_m \mid m = 1, \ldots, M\}$:the DSI dataset comprising M single-label subsets

INITIALIZATION:

$i \leftarrow 0$: index of learning iteration
$t \leftarrow 0$: index of learning cycle
$\boldsymbol{\theta}(0)$: initial values for KAqN
$\boldsymbol{\phi}_m(0)$: initial values for classifier m, for $m = 1, \ldots, M$
$\boldsymbol{\theta}'(0)$: initial values for KAcN

CYCLICAL & SEQUENTIAL LEARNING PROCESS:

$i = 0$
for *cycle* $t = 1, \ldots, T$ **do**
  KAqN acquires new knowledge sequentially:
  for *task* $m = 1, \ldots, M$ **do**
    $i = i + 1$
    learning classification task m on dataset $\mathcal{D}_m$:
      $\boldsymbol{\theta}(i-1)$ is updated to become $\boldsymbol{\theta}(i)$
      $\boldsymbol{\phi}_m(t-1)$ is updated to become $\boldsymbol{\phi}_m(t)$
  end
  $\boldsymbol{\theta}(Mt)$ is consolidated with $\boldsymbol{\theta}'(t-1)$:
    $\boldsymbol{\theta}'(t) \xleftarrow{\boldsymbol{\theta}(Mt)} \boldsymbol{\theta}'(t-1)$
  Termination of the learning process:
  if *stopping criterion is satisfied* **then**
    $\tilde{t} = t$;
    break
  end
end
OUTPUT: $\widetilde{\boldsymbol{\Theta}} \leftarrow \{\boldsymbol{\theta}'(\tilde{t}), \boldsymbol{\phi}_1(\tilde{t}), \ldots, \boldsymbol{\phi}_M(\tilde{t})\}$

2) Knowledge Acquisition: KAqN utilizes the backbone of Swin-B [25], a vision transformer pretrained on ImageNet, as the deep feature encoder. The encoder embeds each input image as a feature map for the downstream classification tasks. Specifically, Swin-B processes an input scene image with a resolution of 244×244 by first partitioning it into non-overlapping 4×4 patches. It then sequentially encodes these patches into a feature map through four stages, progressively reducing the spatial resolution of the input scene image while increasing the depth (number of channels) of the feature maps. Each stage applies either a linear embedding (LE) or a patch merging (PM) layer, followed by multiple Swin Transformer (ST) blocks:

$$\begin{aligned}
\boldsymbol{f}_{\mathrm{I}} &= \mathrm{ST}(\mathrm{ST}(\mathrm{LN}(\boldsymbol{x}; \boldsymbol{\theta}^{\mathcal{E}}_{\mathrm{I},0}); \boldsymbol{\theta}^{\mathcal{E}}_{\mathrm{I},1}); \boldsymbol{\theta}^{\mathcal{E}}_{\mathrm{I},2}), \\
\boldsymbol{f}_{\mathrm{II}} &= \mathrm{ST}(\mathrm{ST}(\mathrm{PM}(\boldsymbol{f}_{\mathrm{I}}; \boldsymbol{\theta}^{\mathcal{E}}_{\mathrm{II},0}); \boldsymbol{\theta}^{\mathcal{E}}_{\mathrm{II},1}); \boldsymbol{\theta}^{\mathcal{E}}_{\mathrm{II},2}), \\
\boldsymbol{f}_{\mathrm{III}} &= \mathrm{ST}(\ldots \mathrm{ST}(\mathrm{PM}(\boldsymbol{f}_{\mathrm{II}}; \boldsymbol{\theta}^{\mathcal{E}}_{\mathrm{III},0}); \boldsymbol{\theta}^{\mathcal{E}}_{\mathrm{III},1}); \boldsymbol{\theta}^{\mathcal{E}}_{\mathrm{III},6}), \\
\boldsymbol{f}_{\mathrm{IV}} &= \mathrm{ST}(\mathrm{ST}(\mathrm{PM}(\boldsymbol{f}_{\mathrm{III}}; \boldsymbol{\theta}^{\mathcal{E}}_{\mathrm{IV},0}); \boldsymbol{\theta}^{\mathcal{E}}_{\mathrm{IV},1}); \boldsymbol{\theta}^{\mathcal{E}}_{\mathrm{IV},2}).
\end{aligned} \tag{4}$$

Here, $\boldsymbol{\theta}^{\mathcal{E}}$ designates all the learnable parameters of KAqN's deep encoder. The extracted feature maps have the following dimensions: $\boldsymbol{f}_{\mathrm{I}} \in \mathbb{R}^{128 \times 56 \times 56}$, $\boldsymbol{f}_{\mathrm{II}} \in \mathbb{R}^{256 \times 28 \times 28}$, $\boldsymbol{f}_{\mathrm{III}} \in \mathbb{R}^{512 \times 14 \times 14}$, and $\boldsymbol{f}_{\mathrm{IV}} \in \mathbb{R}^{1024 \times 7 \times 7}$.

Then, four projectors respectively embed the individual feature maps derived from each stage as feature vectors,

$$\boldsymbol{d}_s = \mathrm{MLP}(\mathrm{Flatten}(\mathrm{GAP}(\boldsymbol{f}_s)); \boldsymbol{\theta}^{\mathcal{P}}_s), \tag{5}$$

for $s = \mathrm{I}, \ldots, \mathrm{IV}$. Here, $\boldsymbol{\theta}^{\mathcal{P}}$ designates the learnable parameters for the four projectors. Eq. (5) states that each projector first downsamples the input feature map $\boldsymbol{f}_s$ through global average pooling (GAP), reducing its spatial dimensions to 1 while keeping the number of channels unchanged. Then, the downsampled feature map is flattened (Flatten) as a vector, which then goes through a multi-layer perceptron (MLP) with two hidden layers. The resulting feature embeddings are $\boldsymbol{d}_{\mathrm{I}} \in \mathbb{R}^{256}$, $\boldsymbol{d}_{\mathrm{II}} \in \mathbb{R}^{512}$, $\boldsymbol{d}_{\mathrm{III}} \in \mathbb{R}^{1024}$, and $\boldsymbol{d}_{\mathrm{IV}} \in \mathbb{R}^{2048}$.

Finally, the feature embedding from the last stage, $\boldsymbol{d}_{\mathrm{IV}}$, flows into classifier m, which is active in the current learning iteration, to predict the probability distribution for the classes of attribute m:

$$\widehat{\boldsymbol{y}}_m = \mathrm{SM}(\mathrm{FC}(\boldsymbol{d}_{\mathrm{IV}}; \boldsymbol{\phi}_{\mathrm{FC},m}); \boldsymbol{\phi}_{\mathrm{SM},m}). \tag{6}$$

The fully-connected (FC) layer of the classifier reduces the dimension of feature embedding $\boldsymbol{d}_{\mathrm{IV}}$, and the softmax operator (SM) normalizes the output as the probability distribution over classes. $\boldsymbol{\phi}_m$ in Eq. (6) represents the learnable parameters of classifier m.

KAqN's performance on the classification task m in any learning iteration is evaluated by calculating the corresponding cross-entropy loss:

$$\mathcal{L}_{cls,m} = -\boldsymbol{y}_m \cdot \log \widehat{\boldsymbol{y}}_m, \tag{7}$$

where $\cdot$ denotes the inner product of the one-hot encoded ground truth class label, $\boldsymbol{y}_m$, and the predicted probability distribution over classes, $\widehat{\boldsymbol{y}}_m$. The prediction loss in Eq. (7) is part of the total loss for guiding the modeling training.

Within a learning cycle t, KAqN learns classification tasks one at a time, using a dedicated single-label dataset for each task. In completing a learning iteration described in Eqs. (4, 5, 6, 14) , KAqN's parameter values are updated:

$$\boldsymbol{\theta}(i) \xleftarrow{\text{monotask learning}} \boldsymbol{\theta}(i-1), \tag{8}$$

and so the classifier for that task:

$$\boldsymbol{\phi}_m(t) \xleftarrow{\text{monotask learning}} \boldsymbol{\phi}_m(t-1). \tag{9}$$

$\boldsymbol{\theta}$ in Eq. (8) is the collection of the encoder's parameters $\boldsymbol{\theta}^{\mathcal{E}}$ and the projectors' parameters $\boldsymbol{\theta}^{\mathcal{P}}$. At the end of learning cycle t, KAqN completes its training on all the M tasks. The resulting parameter values , $\boldsymbol{\theta}(Mt)$, embed the acquired new knowledge about the driving scene classification.

3) Knowledge Accumulation: Knowledge acquired by KAqN in the cyclical training process accumulates in KAcN. Fig. 3 shows that KAcN's network architecture is identical to KAqN. The feature map extracted in each stage of the KAcN's encoder is $\boldsymbol{f}_s^{'}$, which is further embedded as a feature vector $\boldsymbol{d}_s^{'}$, for $s = \mathrm{I}, \ldots, \mathrm{IV}$. For simplicity, $\boldsymbol{\theta}^{'}$ designates all the learnable parameters of KAcN. The identical network architecture of KAcN and KAqN greatly simplifies the process of knowledge accumulation delineated as below.

Upon the completion of the learning cycle t, KAA needs to consolidate the new knowledge acquired by KAqN, $\boldsymbol{\theta}(Mt)$, with the previously learned knowledge stored in KAcN, $\boldsymbol{\theta}^{'}(t-1)$. For this consolidation, KAA adopts exponential moving average (EMA), a temporal mechanism for accumulating knowledge from cyclical learning [26]:

$$\boldsymbol{\theta}'(t) = \alpha(t)\boldsymbol{\theta}'(t-1) + (1-\alpha(t))\boldsymbol{\theta}(Mt), \tag{10}$$

where $\alpha(t)$ is the stability coefficient, a real-valued parameter within the range $[0.9, 1]$. $\alpha(t)$ determines the proportion of previously learned knowledge to retain in cycle t.

KAA, as a learning system, needs to keep a balance between retaining already learned knowledge and acquiring new knowledge, a phenomenon called stability-plasticity tradeoff in neuroscience [27], incremental learning [28], and continual learning [29]. If KAA is too stable, it will struggle to learn new knowledge effectively. Conversely, if it is too plastic, it risks forgetting important knowledge that has already been acquired. This paper designs a cosine scheduler for progressively updating the stability coefficient $\alpha(t)$ in Eq. (10) from 0.9 to 1 in the cyclical learning process:

$$\alpha(t) = 0.9 + 0.05\left[1 - \cos\left(t\pi/T\right)\right], \tag{11}$$

where T is the maximum learning cycles for KAA. Eq. (11) indicates that $\alpha(t)$ is a monotonically increasing function. That is, in the early phases of learning, KAA leans toward plasticity, allowing itself to rapidly absorb new knowledge. In the later phases, it prioritizes stability to prevent the disruption of the established cognitive capability.

4) Knowledge Retention: The cyclical nature of training can cause catastrophic forgetting, thus impeding continual learning and improvement. This issue can be effectively addressed by setting the accumulated knowledge as the reference for the subsequent knowledge acquisition. KAA achieves this through feature-based KD, which penalizes large stage-wise discrepancies between the feature embedding extracted by KAqN, $\boldsymbol{d}_s(i)$, in each learning iteration and the corresponding feature embedding derived by KAcN, $\boldsymbol{d}_s'(t-1)$. The penalty is calculated as the mean squared error (MSE):

$$\mathcal{L}_{cst,s}(i) = \frac{1}{|\boldsymbol{d}_s|}\|\boldsymbol{d}_s(i) - \boldsymbol{d}_s^{'}(t-1)\|_2^2, \tag{12}$$

for stages $s = \mathrm{I}, \ldots, \mathrm{IV}$. This particular KD design fully leveraging the advantages of deep networks, which has been shown to be effective [18].

Then, the overall consistency loss is the weighted sum of stage-wise consistency losses:

$$\mathcal{L}_{cst}(i) = \sum_{s=\mathrm{I}}^{\mathrm{IV}} \lambda_s \mathcal{L}_{cst,s}(i), \tag{13}$$

where λ_s is the coefficient for stage s, which is set to be one for all stages in this study.

5) Loss Function as the Learning Guidance: The consistency loss in Eq. (13) is a regularization term for training KAqN, ensuring that the new knowledge acquisition is based on already attained knowledge. Therefore, the per-sample loss function in learning iteration i can be formulated as the weighted sum of the prediction loss and the consistency loss:

$$\mathcal{L}_{tot}(i) = \beta(i)\mathcal{L}_{cls}(i) + (1-\beta(i))\mathcal{L}_{cst}(i), \tag{14}$$

where $\beta(i)$, a real-valued parameter in the range $[0, 1]$, is the performance-based acquisition-retention indicator (PARI) for balancing the acquisition of new knowledge and the retention of already attained knowledge. $\beta(i)$ is adjusted dynamically with respect to learning needs. Lower performance anticipated for KAqN on the task in learning corresponds to a greater need for continual learning and improvement on this task. Therefore, $\beta(i)$ is designed as a monotonically decreasing function of the estimated task performance, $\widehat{p}(i)$:

$$\beta(i) = (1 - \widehat{p}(i)^{\psi})^{1/\psi}, \tag{15}$$

where $\psi > 1$ is a real-valued parameter defining the shape of the curve, which is set to be 4 in this study.

$\widehat{p}(i)$ in Eq. (15) estimates classification accuracy of KAqN in learning iteration i, obtained based on an EMA process:

$$\widehat{p}(i) = \omega\widehat{p}(i-M) + (1-\omega)p(i-M), \tag{16}$$

where $\widehat{p}(i-M)$ is the estimated accuracy of KAqN for the last learning cycle and $p(i-M)$ is the accuracy that KAqN actually achieved. The EMA process ensures that the performance estimation is based on the entire record from past training cycles. The parameter ω is set to be 0.9 in this study, meaning that the performance estimation puts more weight on the long-term trend than the most recent observation.

Summing the loss function in Eq. (14) over all the training data yields the total training loss. By minimizing this loss with respect to KAqN's learnable parameters in each learning iteration and across learning cycles, the classification model defined in Eq. (2) is optimized. The resulting classification model $\widetilde{C}$, described in Algorithm 1, is a foundation model for driving scene identification.

B. Consistency-based Active Learning (CAL)

KAA learns the knowledge for multi-label scene classification through monotask learning across different single-label datasets. While simplifying data collection and reducing learning complexity, this approach confronts a domain shift issue. That is, the class label distributions of single-label datasets represent the marginal probability distributions of individual attributes, different from their joint distribution. The same issue occurs in the extracted features. Therefore, the foundation model can make mistakes when it attempts to simultaneously identify a scene using multiple attributes.

This paper proposes CAL (Algorithm 2) to tackle the aforementioned issue, which seeks to cost-effectively adapt the attained foundation model to real-world multi-attribute scene identification. Fig. 4 illustrates the overall approach of the proposed CAL. A multi-label test sample, $\mathcal{D}^T$, representing the driving scene distribution in the joint space of scene attributes, is prepared using a stratified sampling strategy (Stf). It randomly samples κ test images per class from DSI's test sets:

$$\mathcal{D}^T = \cup_{m=1}^{M} \mathrm{Stf}(\mathcal{D}_m^{\mathrm{Tst}}; \mathcal{U}_m, \kappa), \tag{17}$$

which is then fully annotated with respect to all M attributes. The adaptation test set may also be drawn from alternative domains or tasks to which the foundation model has been adapted. The adaptation test set may also be drawn from other domains or tasks to which the foundation model has been adapted.

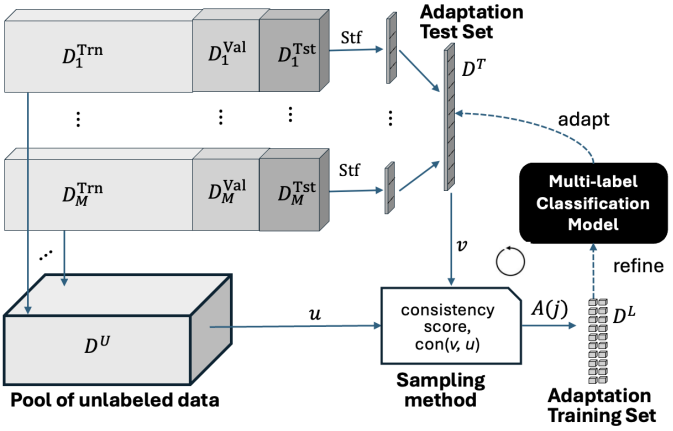


Fig. 4. Schematic diagram for Consistency-based Active Learning (CAL)

Algorithm 2: CAL Algorithm

INPUT:
$\mathcal{D}^U$: Unlabeled pool
$\mathcal{D}^T$: Test dataset
b_j: Budget for adaptation iteration j
N_{CAL}: Maximum iterations

INITIALIZATION:
$\mathcal{D}^L = \emptyset$: initial labeled training set
$\boldsymbol{\Theta}^{'}(0) = \widetilde{\boldsymbol{\Theta}}$: initial learnable parameters

CONSISTENCY ACTIVE LEARNING:
for $j = 1, \ldots, N_{CAL}$ **do**
  compute consistency score $\mathrm{con}(v, \overline{u})$, $\forall\, v \in \mathcal{D}^T$;
  $\mathcal{A}(j)$ comprises b_j samples from $\mathcal{D}^U$ with top consistency scores, labeled by the Oracle;
  add $\mathcal{A}(j)$ to the labeled dataset:
    $\mathcal{D}^L \leftarrow \mathcal{D}^L \cup \mathcal{A}(j)$;
  remove $\mathcal{A}(j)$ from unlabeled dataset:
    $\mathcal{D}^U \leftarrow \mathcal{D}^U \setminus \mathcal{A}(j)$;
  train the classification model for multitasking:
    $\boldsymbol{\Theta}^{'}(j) \xleftarrow{\text{multitask learning}} \boldsymbol{\Theta}^{'}(j-1)$.
  if *stopping criterion is satisfied* **then**
    $j^* = j$.
    break
  end
end

Output: $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}^{'}(j^*)$

Then, a training set $\mathcal{D}^L$ needs to be labeled for adapting the classification model to multitask learning over iterations. Although the ground truth labels for the training set $\mathcal{D}_m^{\mathrm{Trn}}$ were seen by task m, other tasks have never seen these labels. Therefore, the union of DSI's training sets forms an "unlabeled" pool,

$$\mathcal{D}^U = \cup_{m=1}^{M} \mathcal{D}_m^{\mathrm{Trn}}, \tag{18}$$

from which training samples can be selected and prepared according to Algorithm 2.

CAL searches the unlabeled pool D^U for data points that are similar to those in the test set D^T. A consistency function, $\mathrm{con}(v, u)$, is defined to quantify the similarity between any image in the test set, $v \in \mathcal{D}^T$, and any one in the unlabeled pool, $u \in \mathcal{D}^U$:

$$\mathrm{con}(v, u) = -\|v, u\|_{\boldsymbol{d}'_{2,\mathrm{IV}}}, \tag{19}$$

where $\|\cdot,\cdot\|_{\boldsymbol{d}'_{2,\mathrm{IV}}}$ measures the Euclidean distance between two data points in terms of their final-stage feature embeddings. Thereby, for any image in the test set $v \in \mathcal{D}^T$, the corresponding image in the unlabeled pool which is most similar to it is identified:

$$\overline{u} = \underset{u \in \mathcal{D}^U}{\arg\max}\, \mathrm{con}(v, u). \tag{20}$$

Then, these unlabeled images are ranked in decreasing order of their consistency score:

$$\text{con}(v^{(1)}, \overline{u}^{(1)}) > (v^{(2)}, \overline{u}^{(2)}) > \dots \tag{21}$$

to prioritize them for annotation. Following the ranking, the top b_j unlabeled images with the highest consistency scores are selected for annotation:

$$\mathcal{A}(j) = \{\overline{u}^{(1)}, \dots, \overline{u}^{(b_j)}\}, \tag{22}$$

where b_j is the annotation budget in iteration j. After that, $\mathcal{A}(j)$ is moved from the unlabeled pool to the labeled training set, $\mathcal{D}^L$:

$$\mathcal{D}^L \leftarrow \mathcal{D}^L \cup \mathcal{A}(j), \tag{23}$$

$$\mathcal{D}^U \leftarrow \mathcal{D}^U \setminus \mathcal{A}(j). \tag{24}$$

The dataset $\mathcal{D}^L$ is used to refine the classification model into a multitask model.

In refining the classification model, the focal loss is evaluated for each data point in the training set $\mathcal{D}^L$ to focus on hard samples for each task,

$$\mathcal{L}_{MT} = -\sum_{m=1}^{M} \boldsymbol{y}_m \cdot ((\mathbf{1} - \widehat{\boldsymbol{y}}_m)^{\gamma_m} \odot \log \widehat{\boldsymbol{y}}_m), \tag{25}$$

where $\odot$ stands for the element-wise product, $\cdot$ is the dot product, and γ_m is the focusing parameter for task m, which is set to 1 here. The total loss of the multitask learning process is obtained by summing up this loss across in Eq. (25) across all data points in $\mathcal{D}^L$, which guides the model refinement iteratively:

$$\boldsymbol{\Theta}'(j) \xleftarrow{\text{multitask learning}} \boldsymbol{\Theta}'(j-1), \tag{26}$$

where $\boldsymbol{\Theta}'(0) = \widetilde{\boldsymbol{\Theta}}$. This process can go for several iterations until satisfaction, resulting in the optimized classification model $\mathcal{C}^*$ parameterized with $\boldsymbol{\Theta}^*$.

V. Implementation Details

The proposed KAA-CAL learning method was implemented using PyTorch 1.10.0 on a server equipped with an Nvidia Tesla V100 featuring 32 GB of memory. Model training utilized the AdamW optimizer, a batch size of 32, and a maximum of 100 training epochs. Input images were resized to 224×244 and augmented with random rotations, crops, and pixel normalization.

The deep encoder for KAA in Eq. (4) was initialized with the weights pretrained on ImageNet-22k [30]. The projectors in Eq. (5) and the classifiers in Eq. (6) were initialized using the default Xavier initialization method. Then, learnable parameters were progressively optimized using the method delineated in Algorithm 1. The learning rate for KAA implementation was adjusted in two phases. The first phase is a 10-epoch warm-up, where the learning rate linearly increased from an initial value of 1e-6 to 5e-4. In the second phase, the learning rate decayed to a final value of 1e-5 using a cosine schedule. Upon convergence at the end of all training epochs, the model from the final epoch is the foundation model.

In implementing CAL, a per-iteration budget (b_j) of 200 images, approximately 1% of the single-label training data in DSI, was allocated for annotating the data recommended by Algorithm 2. This process ran for up to five iterations to evaluate the efficiency of CAL. The data annotation process confronted a challenge - not all images can be reliably annotated with all the seven driving scene attributes. For example, annotators found the weather condition in some nighttime images was hard to identify. When a reliable label was unavailable, a sentinel value "-1" was assigned. Such labels were excluded from the loss function evaluation to ensure training integrity. Aiming to retain knowledge learned from KAA, the first three stages of the deep encoder were frozen for finetuning multitask classification. The learning rate scheduler for the CAL implementation was similarly configured.

VI. Experiments and Results

Experimental studies were conducted, aiming to verify the merits of the proposed KAA-CAL learning method.

A. Deep Encoder Selection

Various deep network architectures are available to choose from for the deep encoder. Vision transformers (ViT) demonstrate unique advantages over convolutional neural networks in learning deep features for image classification. This study compared SOTA transformer-based models, including Swin-S, Swin-B, ViT-B, and ViT-L. The comparison spanned the seven independent image classification tasks, evaluating models from the perspectives of both classification accuracy and model complexity. The results are summarized in TABLE I.

TABLE I
Classification accuracy (%) comparison of vision transformers as the deep encoder in monotask learning

	Swin-B	Swin-S	ViT-B	ViT-L
MODEL COMPLEXITY				
FLOPs (G)	15.4	8.7	17.6	76.9
The number of parameters (M)	88	50	86	307
ACCURACY (%)				
Time of Day	**99.6**	91.3	89.3	96.0
Weather	**92.2**	92.0	90.9	92.1
Weather-related Road Condition	**98.4**	97.7	96.7	96.6
Road Function	**99.8**	99.2	95.4	95.8
Intersection	88.6	86.9	93.4	**95.8**
Work Zone	**95.2**	87.9	88.5	90.2
Grade-separated Infrastructure	94.6	93.7	97.2	**98.4**
Average	**95.5**	92.7	93.4	95.0

The top performance for each task is bolded.

As expected, Swin-B outperforms Swin-S on all seven tasks, and the same pattern holds for ViT-L compared to ViT-B. This confirms that the more complex transformers (Swin-B and ViT-L) achieve better performance than their less complex counterparts. While both are powerful, Swin transformers address limitations of the original ViT, exhibiting improved computational efficiency and the capability for multi-scale feature learning. TABLE I shows that Swin-B outperformed ViT-L on five out of seven tasks with a margin ranging from 0.1% to 4.9%. Importantly, the FLOPs for Swin-B are only 20% of ViT-L. Given the highest average accuracy and more affordable model complexity, Swin-B is chosen as the deep encoder architecture for the classification model.

B. Effectiveness of Stage-wise Knowledge Distillation

The ability to retain previously acquired knowledge while learning new information is essential for the multi-attribute scene identification model to be scalable. Sections IV-A4 and IV-A5 explained that, when KAqN continues acquiring new knowledge, the already acquired knowledge accumulated in KAcN is retained through stage-wise feature-based KD, and the intensity of this regularization is moderated by PARI. To demonstrate the effectiveness of this knowledge retention design for KAA, an ablation study about KD was performed with results summarized in TABLE II.

The Baseline model in TABLE II is learned solely by a student network because both stage-wise KD and PARI are removed. It results in an average accuracy of 92.7%, representing a 2.5% decrease compared to our model (Model IV). This gap will increase when the KAA attempts to acquire more diverse knowledge from many datasets.

Model I adds the last-stage KD to the baseline model, analogous to a teacher who provides less guidance to the student's learning. This model achieves a negligible performance gain of merely 0.1%. Conversely, Model III incorporates stage-wise KD into the baseline model, leading to an average accuracy increase of 1.8%. This comparison highlights the importance of stage-wise KD, since features of different scales offer comprehensive information about driving scenes.

Model II augments the last-stage KD with PARI, which improves average accuracy by 1.2% over Model I. Similarly, Model IV pairs PARI with stage-wise KD, increasing Model III's well-performed average accuracy by an additional 0.7%. These observations verify the importance of dynamically adjusting the weight for new knowledge acquisition according to specific learning needs, regardless of the KD method.

C. Performance of the Foundation Model

After verifying the strengths of KAA's learning architecture design, the experimental study continued to assess the resulting foundation model's performance. Fig. 5 compares the foundation model to the seven individual monotask models. As illustrated, the foundation model is comparable to the individual monotask models in task accuracy, with variations ranging from -3.9% to 3.2% across the seven tasks. Specifically, the foundation model is 3.2% more accurate than the monotask model for identifying Grade-separated Infrastructure. However, the foundation model's accuracy in identifying Work Zone is 3.9% lower. This drop can be attributed to the challenge of image-level classification in capturing the fine-grained features critical for identifying work zones, such as safety cones and work zone signs. Approaches based on object detection or semantic segmentation (e.g., [31]) are generally more capable of detecting such localized features. For the remaining five tasks, their performance differences are within $\pm 1\%$. The foundation model can substitute for individual models dedicated to different classification tasks, as its comparable performance demonstrates.

TABLE II
EFFECTIVENESS OF KNOWLEDGE RETENTION THROUGH STAGE-WISE FEATURE-BASED KNOWLEDGE DISTILLATION

Model	Last-stage KD	Stage-wise KD	PARI	Avg. (%)
Baseline				92.7
Model I	✓			92.8
Model II	✓		✓	94.0
Model III		✓		94.5
Model IV (Ours)		✓	✓	**95.2**

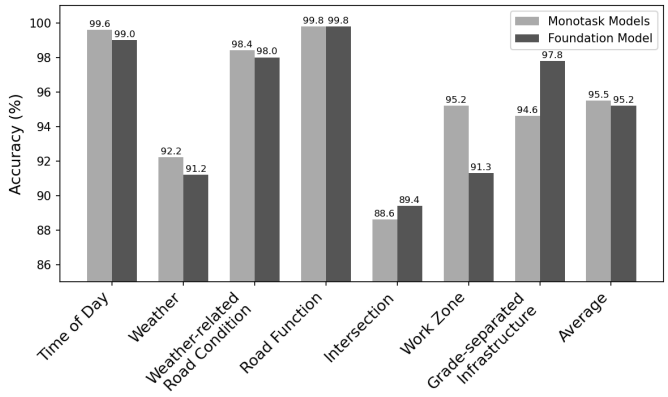


Fig. 5. The foundation model vs. corresponding monotask models

D. Effectiveness of Consistency-based Active Learning

The foundation model, obtained from diverse single-label datasets via monotask learning, must quickly adapt to downstream tasks. CAL proposed in Section IV-B is designed for this purpose, addressing the limitations inherent in KAA's monotask learning approach. To demonstrate its effectiveness, CAL was compared to its variants and SOTA deep active learning methods, detailed as follows.

- **Ran-I**'s deep encoder is initialized with the weights pretrained on ImageNet-22k. Samples for each class are randomly selected from the unlabeled pool and then annotated. Compared to CAL, Ran-I does not leverage the driving scene knowledge offered by the foundation model, nor does it employ the feature consistency-based criterion for sampling.
- **Ran-D** uses the same sampling method as Ran-I, but it adopts the foundation model as its deep encoder.
- **Core-Set-I** is a diversity-based active learning approach proposed in [23]. Its deep encoder is initialized with the weights pretrained on ImageNet-22k. In the first iteration, the model is refined with an initial training set, helping expose the encoder to some multi-label training data. In each subsequent iteration, a k-Center-Greedy strategy selects additional samples to represent various clusters in the unlabeled pool.
- **LLAL-I** follows the same initialization as Core-Set-I, but it introduces an additional loss prediction component [22], which learns to estimate a pseudo-loss value for each sample. In selecting samples from the unlabeled pool, those with the highest pseudo-loss values are pri-

oritized, as they represents the most informative samples to learn from.

Fig. 6 compares the proposed CAL method with those introduced above, presenting the task-level accuracies in plots a-g and the average accuracy in plot h. Each plot illustrates the test accuracy values before adaptation (the training set $\mathcal{D}^L$ has 0% labeled data from $\mathcal{D}^U$) and over five iterations of adaptation. Each iteration selects 1% of data from the unlabeled pool. Fig. 6(h) shows that, prior to adaptation, models that are initialized with the foundation model's weights (i.e., CAL and Ran-D) demonstrate a higher average accuracy than those initialized with weights pretrained on ImageNet (i.e., Ran-I, LLAL-I, and Core-set-I), with a margin of 31.3%. The initial performance gap between these two groups highlights the benefit of utilizing the foundation model for driving scene identification, as it possesses the knowledge about the seven classification tasks.

After the first iteration, CAL is still the top one although all methods improve their task performance. It surpasses the best competing method's accuracy by a significant margin of 3.3% to 10.0% across all tasks, except for the task of identifying Time of Day (Fig. 6(a)). Effectively, CAL brings the classification model's average accuracy up to 86.2% after just one iteration of adaptation. With another iteration, CAL boosts the average accuracy to 90.1%, with task accuracy ranging from 83.6% to 96.3%. Through additional iterations, CAL further enhances the accuracy for identifying Weather and Road Function, thus increasing the average accuracy to 92.7%. Figure 6(h) reveals CAL's superior performance and cost-effectiveness. It consistently outperforms the best competing method in every iteration, with margins ranging from 5.0% to 12.9%. Importantly, CAL has achieved its major improvements in just two iterations, while all other methods require significantly more iterations and so more training data for adaptation.

The importance of CAL's sample recommendation method is demonstrated by a comparison between Ran-D and CAL, which differ in their sampling strategies. Ran-D uses a simple stratified random sampling strategy, whereas CAL samples data based on features consistency measurement between unlabeled data and the test set for adaptation. CAL consistently improves task accuracy over iterations, whereas Ran-D lowers the accuracy in identifying Road Function and Grade-separated Infrastructure. Moreover, CAL achieves a higher saturated accuracy than Ran-D on all tasks. Consequently, CAL dominates Ran-D in Fig. 6(h).

The groundwork that the driving scene foundation model has laid for multi-attribute scene identification is highlighted by reviewing Ran-I, LLAL-I, and Core-Set-I that do not utilize the foundation model. Those methods need significantly more iterations to improve task performance. Among those, Core-Set-I is the most competitive. After five iterations, it has attained the task accuracy comparable to CAL's in identifying Weather-related Road Condition and Grade-separated Infrastructure. But CAL still beat Core-Set-I on the remaining five tasks by a margin ranging from 1.3% to 13.9%. Consequently, Core-Set-I's average accuracy is still 11.8% lower than that of CAL after three iterations and 5.0% lower after five iterations.

The comparison in Fig. 6 indicates that CAL benefits from the foundation model, which not only sets up a higher starting point for adaptation but also gives CAL good knowledge to identify the most useful training data for ongoing learning and refinement. This pairing of a capable active learning method with a knowledgeable foundation model is the underlying reason for the efficiency and effectiveness of model adaptation.

E. Ablation Study

KAA and CAL have now been verified as advantageous designs: KAA builds a knowledge foundation for driving scene identification by acquiring and accumulating information from heterogeneous single-label source data via monotask learning. CAL adapts this foundation model to the multi-label target data. An ablation study was conducted to gauge their respective contributions, with results summarized in TABLE III.

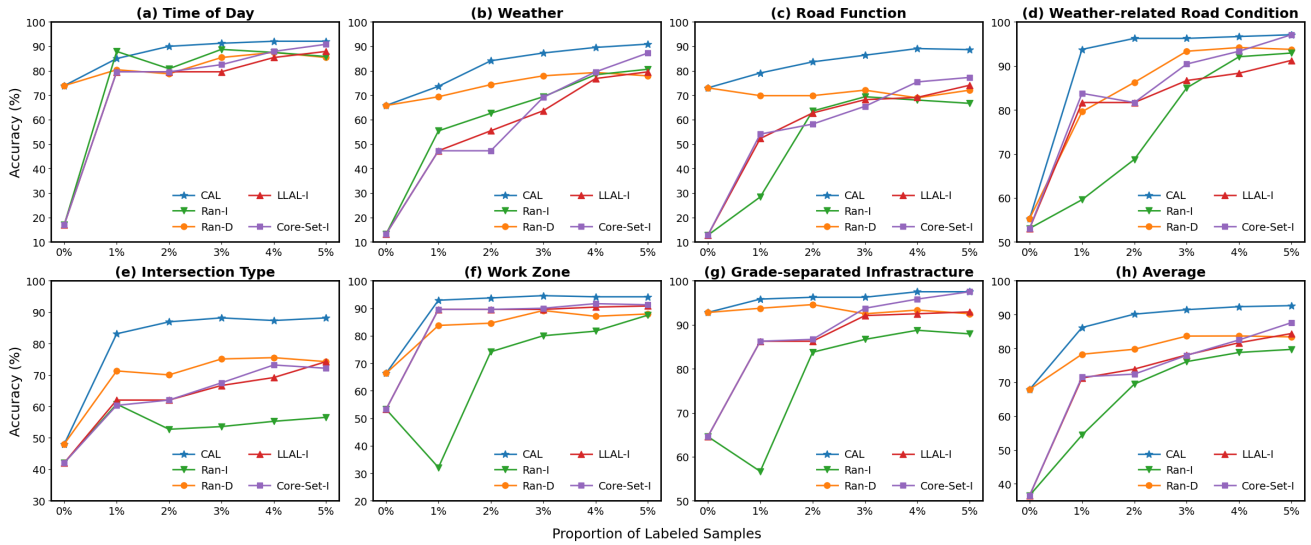


Fig. 6. Comparison of active learning methods in terms of task and average accuracies (%)

TABLE III
PROGRESSIVE IMPROVEMENT IN TASK AND AVERAGE ACCURACIES (%) IN THE ABLATION STUDY

Model	KAA	CAL	ToD	Wea	RoF	WrR	Int	WrZ	GsI	Average
Baseline			17.1	13.2	12.7	53.1	42.1	53.3	64.6	36.6
Foundation	✓		73.9 56.8↑	65.8 52.6↑	73.0 60.3↑	55.3 2.2↑	48.1 6.0↑	66.4 13.1↑	92.8 28.2↑	67.9 31.3↑
with Adaptation	✓	✓(2 iters)	90.0 16.1↑	84.1 18.3↑	83.6 10.6↑	96.3 41.0↑	86.9 38.8↑	93.8 27.4↑	96.3 3.5↑	90.1 22.2↑
with Adaptation	✓	✓(5 iters)	**92.1** 2.1↑	**90.9** 6.8↑	**88.6** 5.0↑	**97.1** 0.8↑	**88.2** 1.3↑	**94.2** 0.4↑	**97.5** 1.2↑	**92.7** 2.6↑

ToD: Time of Day; Wea: Weather; RoF: Road Function; WrR: Weather-related Road Condition; Int: Intersection type; WrZ: Work Zone, GsI: Grade-separated Infrastructure.

The baseline model in this ablation study is a multitask classification model whose deep encoder is Swin-B pretrained on ImageNet. This baseline model achieves task accuracies ranging from 17.1% to 64.4%, resulting in an average accuracy of 36.6%. Its low accuracy indicates that ImageNet possesses insufficient knowledge regarding driving scene identification.

Adoption of the foundation model as the deep encoder boosts the average accuracy to 67.9%, representing a substantial increase of 31.3%. This gain is attributed to the comprehensive knowledge regarding driving scenes acquired by KAA from diverse single-label datasets. For instance, the accuracy in identifying Road Function, a unique attribute specific to driving scenes, had an increase of 60.3%. Furthermore, the task accuracies for identifying the environmental attributes, Time of Day and Weather, each rise by over 50%. Consequently, the task accuracies now span from 48.1% to 92.8%. However, room for improvement are clearly present. The gains in identifying Weather-related Road Condition and Intersection are limited, primarily due to a large shift in driving scene data distribution compared to the marginal distributions in the single-labeled datasets.

CAL, designed to address the limitation of KAA's monotask learning approach, effectively improves the average accuracy to 90.1% with only two iterations, corresponding to another increase of 22.2%. At the task level, accuracies have reached the range from 83.6% to 96.3%. With three additional iterations, the task accuracies in identifying Weather and Road Function are improved by at least 5%, further boosting the average accuracy to 92.7%.

F. Driving Scene Identification Examples

To supplement the ablation study in the preceding section, seven examples (a-g) are further provided, with their corresponding results detailed in TABLE IV. The baseline model demonstrates insufficient knowledge about driving scene identification. It fails in multi-label scene classification, only correctly identifying each scene with at most three attributes.

The foundation model, which assimilates knowledge regarding driving scene identification from diverse single-label datasets, significantly outperforms the baseline model that lacks such knowledge, correctly identifying scenes with respect to four to six attributes. However, an observable limitation of the foundation model is its frequent misidentification of dry road surface conditions as snowy, daytime as dawn, non-work zones as work zones, other road types as local roads, and non-intersection or three-way intersections as four-way intersections. For instance, the foggy weather condition in scenario (d) may have led the foundation model to erroneously classify daytime as dawn and the dry road surface as snowy. Furthermore, the vehicle's close proximity to the intersection makes it difficult to differentiate three-leg and four-leg features. Scenario (f) depicts an interstate highway that is slightly more complex than typical interstate driving scenes. The advertisement banner on the left and the open construction space on the right likely lead the model to misidentify it as a four-way intersection on a local road. Additionally, the atypical work zone features and the relatively farther distance of the safety drums from the ego-vehicle fail the model in recognizing the work zone scenario. Overall, certain combinations of scene attribute labels are less prevalent in single-labeled datasets, thereby limiting the foundation model's exposure and subsequent ability to accurately identify these specific scenes.

Utilizing feature proximity to recommend new training samples that are similar to the test data, CAL adapts the foundation model quickly to multi-attribute sence identification, demonstrating superior ability in classifying driving scenes comprehensively, as illustrated in the two rightmost columns in TABLE IV. Notably, the model adapted via CAL with only two iterations corrects most misclassifications by the foundation model, demonstrating the effectiveness of CAL. Meanwhile, with iterations up to five, the classification model learns the more fine-grained feature distributions, thus improving its ability to identify driving scenes. For instance, in scenes a, b, d, and e, the foundation model misclassified the weather-related road condition as snowy. Assisted by CAL, now the adapted model can correctly recognize the category and reassign the class label. Furthermore, CAL mitigates dominant biases in feature distributions across single-label classification datasets. Specifically, in recognizing a four-way intersection, the model relies more on the complex behavior of traffic participants rather than on the actual geometric layout of the intersection (as seen in scenes c and g). CAL addresses this by providing additional class labels to supervise the learning of features that were previously either unseen or unaddressed in monotask learning.

G. Comparison to SOTA Multitask Classification Models

In the end, the proposed KAA-CAL learning method was compared to SOTA models reviewed in Section II-B. To demonstrate the generalization of the KAA-CAL learning method, the comparative study was conducted on both BDD100K [7] and HSD [8]. BDD100K has labels for three driving scene attributes, including Time of Day, Weather, and

TABLE IV
GROUNDTRUTH AND PREDICTED LABELS FOR DRIVING SCENE IDENTIFICATION EXAMPLES

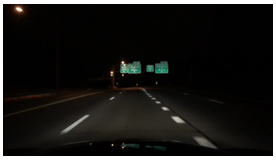
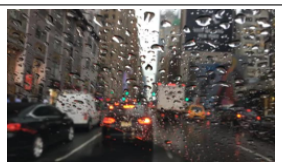
Example	Driving scene	Groundtruth	Predicted Labels			
			Baseline	Foundation	KAA-CAL (2 iters)	KAA-CAL (5 iters)
a		Night x Interstate Dry Non-intersection Non-work zone Open roads	Dawn x Collector Wet Non-intersection Work zone Open roads	Night x Local road Snowy Non-intersection Non-work zone Open roads	Night x Interstate Dry Non-intersection Non-work zone Open roads	Night x Interstate Dry Non-intersection Non-work zone Open roads
b		Daytime Overcast Arterial Dry 4-way Non-work zone Open roads	Dawn Snowing Collector Wet 3-way Work zone Open roads	Daytime Overcast Local road Snowy 4-way Work zone Open roads	Daytime Overcast Local road Dry 4-way Non-work zone Open roads	Daytime Overcast Arterial Dry 4-way Non-work zone Open roads
c		Daytime Overcast Local road Snowy 4-way Non-work zone Open roads	Dawn Snowing Arterial Snowy Non-intersection Work zone Open roads	Dawn Snowing Local road Snowy 4-way Non-work zone Open roads	Daytime Overcast Local road Snowy 4-way Non-work zone Open roads	Daytime Overcast Local road Snowy 4-way Non-work zone Open roads
d		Daytime Foggy Local road Dry 3-way Non-work zone Open roads	Dawn Snowing Collector Snowy Non-intersection Work zone Open roads	Dawn Foggy Local road Snowy 4-way Non-work zone Open roads	Daytime Foggy Local road Dry 3-way Non-work zone Open roads	Daytime Foggy Local road Dry 3-way Non-work zone Open roads
e		Daytime Clear Interstate Dry Non-intersection Non-work zone Overhead bridges	Dawn Snowing Collector Wet Non-intersection Work zone Open roads	Daytime Overcast Interstate Snowy Non-intersection Work zone Overhead bridges	Daytime Overcast Interstate Dry Non-intersection Non-work zone Overhead bridges	Daytime Clear Interstate Dry Non-intersection Non-work zone Overhead bridges
f		Daytime Clear Interstate Dry Non-intersection Work zone Open roads	Dawn Snowing Collector Snowy Non-intersection Work zone Open roads	Dawn Clear Local road Dry 4-way Non-work zone Open roads	Daytime Clear Arterial Dry Non-intersection Work zone Open roads	Daytime Clear Arterial Dry Non-intersection Work zone Open roads
g		Daytime Rainy Local road Wet Non-intersection Non-work zone Open roads	Dawn Clear Collector Snowy Non-intersection Work zone Open roads	Daytime Rainy Local road Wet 4-way Non-work zone Open roads	Dawn Rainy Local road Wet Non-intersection Non-work zone Open roads	Daytime Rainy Local road Wet Non-intersection Non-work zone Open roads

TABLE V
BDD100K AND HSD DATASETS

Dataset	Training	Validation	Testing
BDD100K	70,000	7,000	3,000
HSD*	53,115	5,500	2,275

*Scene images were sampled every 100 frames from 80 hours of driving videos.

Scene Type. The HSD dataset includes labels for four attributes: Intersection Type, Surface Condition, Driving Types, and Weather. Despite the limited number of attributes provided, these two datasets offer an opportunity for objectively evaluating KAA-CAL against SOTA methods. TABLE V summarizes the sizes of these datasets in terms of training, validation, and testing partitions.

Our multi-attribute scene identification model utilizes the foundation model as its deep encoder. It was subsequently refined via CAL to identify driving scenes in BDD100K and HSD, respectively, using a total of 15% of their training data over three iterations. In contrast, the four SOTA models, whose encoders were pretrained on ImageNet, were trained on the full BDD100K and HSD training datasets, respectively. Results of the comparative study are summarized in TABLE VI.

KAA-CAL, which uses only 15% of the training data, achieves task accuracy comparable to SOTA methods that utilize all the training data. Specifically, on BDD100K, KAA-CAL achieves the same highest accuracy as ResNet-18 and CF-NET in identifying Time of Day. KAA-CAL's accuracy in identifying Weather (81.8%) exceeds the second-best method,

TABLE VI
CLASSIFICATION ACCURACY (%) COMPARISON WITH SOTA METHODS ON BDD100K AND HONDA HSD DATASET

	BDD100K			HSD			
	Time of Day	Weather	Scene Type	Intersection	Surface Condition	Driving Type	Weather
ResNet18 [16]	<u>92.7</u>	80.6	77.7	<u>88.9</u>	<u>91.0</u>	80.5	67.1
ViT-B-16 [16]	92.2	80.1	76.9	88.0	<u>91.0</u>	<u>81.1</u>	<u>68.3</u>
CF-NET [3]	<u>92.7</u>	<u>81.0</u>	<u>78.5</u>	**89.0**	90.8	78.2	66.7
RECNet [12]	92.6	80.2	76.7	87.0	90.4	80.7	68.0
KAA-CAL (ours)*	**92.7**	**81.8**	**78.6**	87.8	**91.1**	**81.4**	**68.6**

The best performance is bolded, and the second best is underlined. $*$ trained using only 15% of the training data.

CF-NET (81.0%), by 0.8%, and it slightly surpasses CF-NET by 0.1% in identifying Scene Type.

On HSD, KAA-CAL achieves the highest accuracy in identifying Surface Condition (91.1%), Driving Type (81.4%), and Weather (68.6%), exceeding the second best method by a small margin ranging from 0.1% to 0.3%. However, its accuracy in identifying Intersection Type (87.8%) is 1.2% below the best method, CF-NET (89.0%).

Notably, Scene Type in BDD 100K and Driving Type in HSD are not attributes that the foundation learned to recognize. With the foundation model's comprehensive knowledge about driving scenes, and facilitated by CAL, the model efficiently gains the knowledge to identify driving scenes using these new attributes. While SOTA models listed in TABLE VI attain the performance comparable to KAA-CAL, they are not ready to characterize driving scenes using other attributes.

VII. CONCLUSION

This paper presents KAA-CAL, a deep learning method for multi-attribute driving scene identification, a fundamental yet challenging visual perception capability for AVs. KAA improved the average classification accuracy by 31.3% compared to the baseline model pretrained on ImageNet. CAL further boosts the performance to 92.7%. Additionally, KAA-CAL outperforms SOTA methods on BDD100K and HSD, achieving this while using only 15% of the training data and even recognizing attributes for which the foundation models was not previously train.

KAA-CAL lays a methodological foundation for not only multi-attribute scene identification - a high level perception - but other vision capabilities that AV needs. A universal vision foundation model can be built by synergizing the knowledge of those constituents. Due to a deliberate focus on the methodological development, KAA-CAL was not evaluated on a broader range of datasets. Yet, the number of scene attributes can be scaled up in this foundation model by following the data preparation and learning approach outlined in this paper. Furthermore, despite its efficiency, CAL has not yet achieved the ideal few-shot model refinement for downstream tasks. Future studies can also leverage the domain-specific prior and semantic guidance embedded in the driving scene identification model with vision-language models to advance video understanding and reasoning for AVs. Nevertheless, the work presented in this paper lays the groundwork for exploring these opportunities for generalization, scaling, improvement, and extension.

ACKNOWLEDGMENT

Qin receives funding from the Rural Safe Efficient Advanced Transportation Center, a Tier-1 University Transportation Center funded by the United States Department of Transportation (USDOT), through agreement number 69A3552348321. The contents of this paper reflect the views of the authors. USDOT assumes no liability for the contents or use thereof.

REFERENCES

[1] S. Luo, W. Chen, W. Tian, R. Liu, L. Hou, X. Zhang, H. Shen, R. Wu, S. Geng, Y. Zhou, L. Shao, Y. Yang, B. Gao, Q. Li, and G. Wu, "Delving into multi-modal multi-task foundation models for road scene understanding: From learning paradigm perspectives," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 12, pp. 8040–8063, 2024.

[2] C. Ryu, H. Seong, D. Lee, S. Moon, S. Min, and D. H. Shim, "Words to wheels: Vision-based autonomous driving understanding human language instructions using foundation models," in *2025 IEEE Intelligent Vehicles Symposium (IV)*, 2025, pp. 2200–2207.

[3] T. T. Duong, T. P. Nguyen, and J. W. Jeon, "Combined classifier for multi-label network in road scene classification," in *2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, 2020, pp. 1–4.

[4] L. Chen, W. Zhan, W. Tian, Y. He, and Q. Zou, "Deep integration: A multi-label architecture for road scene recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4883–4898, 2019.

[5] F. Wu, "Wz-traffic dataset," https://github.com/Fangyu0505/traffic-scene-recognition, 2019.

[6] I. Admin, A. Ghosh, R. Tamburo, S. Narasimhan, S. Zheng, J. A. Padilla, M. Cardei, N. Dunn, and H. Zhu, "ROADWork Data," 7 2024. [Online]. Available: https://kilthub.cmu.edu/articles/dataset/ROADWork_Data/26093197

[7] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[8] A. Narayanan, I. Dwivedi, and B. Dariush, "Dynamic traffic scene classification with space-time coherence," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5629–5635.

[9] F. Wu, S. Yan, J. S. Smith, and B. Zhang, "Deep multiple classifier fusion for traffic scene recognition," *Granular Computing*, vol. 6, no. 1, pp. 217–228, 2021.

[10] R. Prykhodchenko and P. Skruch, "Road scene classification based on street-level images and spatial data," *Array*, vol. 15, p. 100195, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005622000467

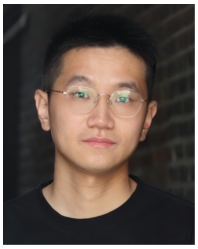
[11] J. Ni, K. Shen, Y. Chen, W. Cao, and S. X. Yang, "An improved deep network-based scene classification method for self-driving cars," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[12] M. Introvigne, A. Ramazzina, S. Walz, D. Scheuble, and M. Bijelic, "Real-time environment condition classification for autonomous vehicles," 2024. [Online]. Available: https://arxiv.org/abs/2405.19305

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

[14] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.
[15] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *arXiv preprint arXiv:1903.11027*, 2019.
[16] R. Prykhodchenko and P. Skruch, "Efficient multi-task learning for road scene classification: Scene, time, and weather predictions," in *2024 IEEE 20th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2024, pp. 1–7.
[17] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.
[18] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.
[19] D. Ma, J. Pang, M. B. Gotway, and J. Liang, "Foundation ark: Accruing and reusing knowledge for superior and robust performance," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, Eds. Cham: Springer Nature Switzerland, 2023, pp. 651–662.
[20] G. M. Van de Ven, H. T. Siegelmann, and A. S. Tolias, "Brain-inspired replay for continual learning with artificial neural networks," *Nature communications*, vol. 11, no. 1, p. 4069, 2020.
[21] X. Zhan, Q. Wang, K. hao Huang, H. Xiong, D. Dou, and A. B. Chan, "A comparative survey of deep active learning," 2022. [Online]. Available: https://arxiv.org/abs/2203.13450
[22] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
[23] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," 2018. [Online]. Available: https://arxiv.org/abs/1708.00489
[24] A. Hekimoglu, M. Schmidt, and A. Marcos-Ramiro, "Active learning with task consistency and diversity in multi-task networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 2503–2512.
[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
[26] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.
[27] M. Mermillod, A. Bugaiska, and P. Bonin, "The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects," p. 504, 2013.
[28] G. Wu, S. Gong, and P. Li, "Striking a balance between stability and plasticity for class-incremental learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 1124–1133.
[29] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3366–3385, 2022.
[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
[31] V. Sundharam, A. Sarkar, A. Svetovidov, J. S. Hickman, and A. L. Abbott, "Characterization, detection, and segmentation of workzone scenes from naturalistic driving data," *Transportation Research Record*, vol. 2677, no. 3, pp. 490–504, 2023. [Online]. Available: https://doi.org/10.1177/03611981221115724

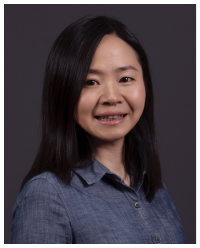
Ke Li (Graduate Student Member, IEEE) received the B.S. degree in transportation from Southeast University, China, in 2021, and the M.S. degree in Civil Engineering from University of Illinois Urbana-Champaign, IL, USA, in 2023. He is currently pursuing the Ph.D. degree in the Department of Civil Engineering at Stony Brook University, NY, USA. He has been a research assistant since 2023, working on perception and recognition for autonomous driving systems and intelligent systems using deep learning methods.



Chenyu Zhang (Graduate Student Member, IEEE) received the B.S. degree in Bridge Engineering in 2018 and the M.S. degree in Civil Engineering in 2021, both from the School of Highway, Chang'an University, Xi'an, China. He is currently pursuing the Ph.D. degree in Civil Engineering at Stony Brook University, Stony Brook, NY, USA. His research interests include intelligent transportation systems, transportation asset management, and resilient infrastructure systems.

Yuxin Ding received his Bachelor degree from the Civil Aviation University Of China, and Master degree from Georgia Institute Of Technology. He is currently pursuing the Ph.D. with the Department of Civil and Environmental Engineering, Penn State University, Pennsylvania, USA. His research is focused on autonomous vehicles and dynamic system modeling.

Xianbiao Hu received the B.Eng. and M.Eng. degrees in Transportation Engineering from Tongji University, China, and the Ph.D. degree in Transportation Engineering from the University of Arizona, USA. His research focuses in the area of Smart Mobility System, Dynamic System Modeling, Vehicle Technology, Active Demand Management, Automated Vehicles, and Transportation Electrification.



Ruwen Qin (Member, IEEE) received the B.E. and M.S. degrees in spacecraft design from Beijing University of Aeronautics and Astronautics and the Ph.D. degree in industrial engineering & operations research from Pennsylvania State University. She is an Associate Professor of civil engineering at Stony Brook University. Her research is focused on sensing and deep learning methods to build perception and cognition abilities for intelligent systems like autonomous vehicles. She is a member of IEEE Intelligent Transportation Systems Society.