Acquiring and Accruing Knowledge from Diverse Datasets: a New Approach to Multi-label Driving Scene Classification

Ke Li, Chenyu Zhang, Ruwen Qin
Department of Civil Engineering, Stony Brook University
Stony Brook University

Introduction

Driving scenes are complex

- Characterized by multiple attributes
- Unbalanced data distribution on the high-dimensional attribute space

Existing open-source driving scene datasets

- Each provides labels of one or a few attributes
- Lack comprehensive multi-label annotation

Domain shift cross datasets

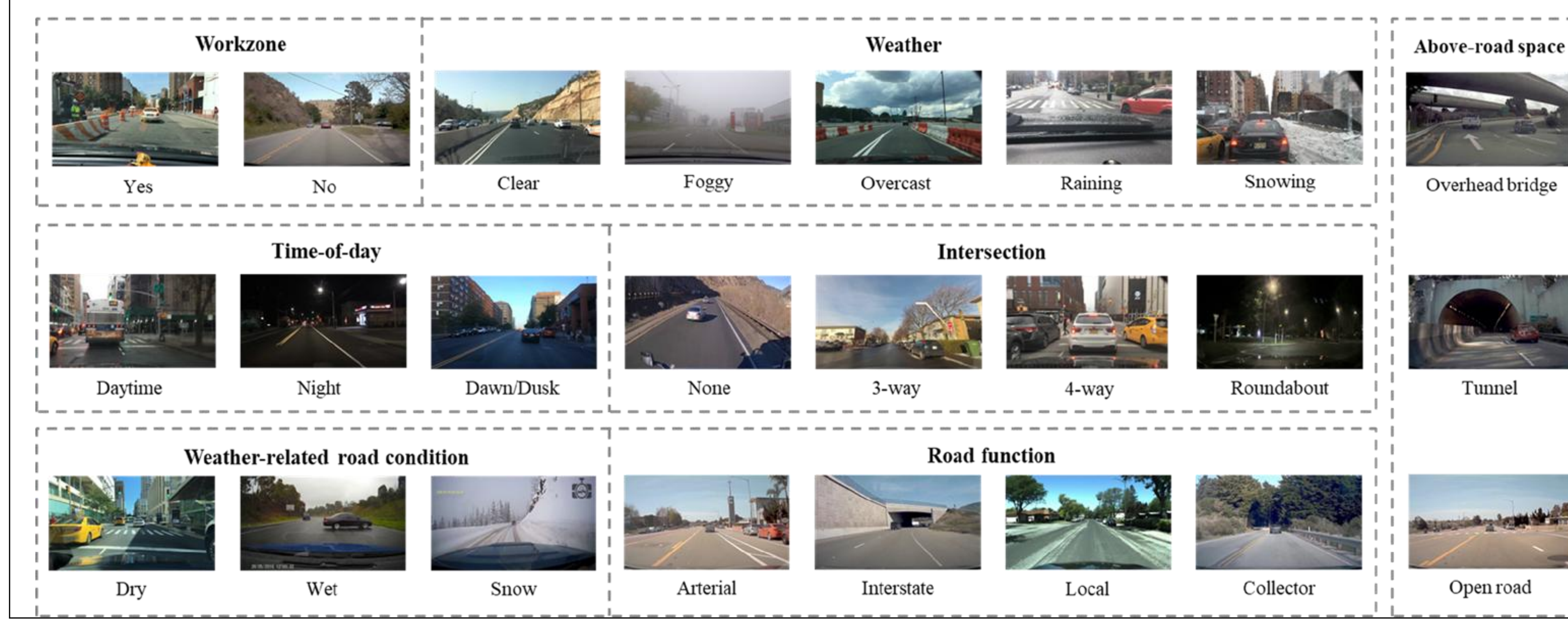
- Domain discrepancy across datasets cause challenges in adapting mono-task models across the datasets
- Leads to the misalignment between extracted feature and target feature for each scene attribute

Research questions

- How to extract knowledge from diverse sources of datasets and accumulate the knowledge into one foundation model?
- How to address the cross-dataset domain shift issue for complex driving scenes?

Driving Scene Identification(DSI) Dataset

- 7 datasets of 32k images collectively contribute 24 possible scene labels
- Each label is associate with one and only one of the 7 attributes



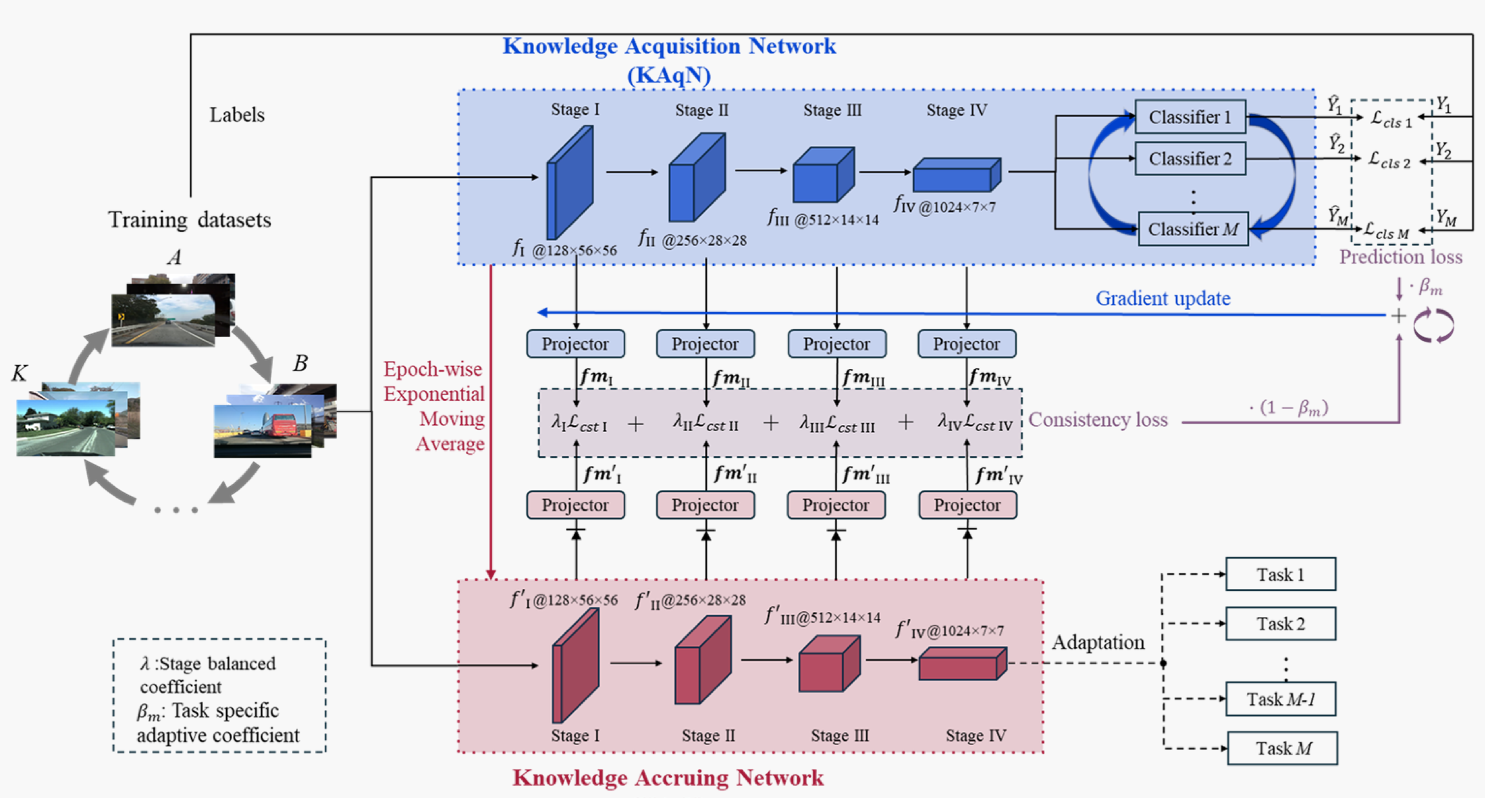
Datasets and classes	Sample Size Trn	Vld	Tst
Road function	*3,891*	*1210*	*639*
arterial	1,105	260	100
collector	798	260	100
local	1,038	432	339
interstate	950	258	100
Intersection related	*1,981*	*332*	*367*
four-way	673	116	115
three-way	358	50	91
no	801	147	111
roundabout	149	19	50
Above-road space	*4,874*	*1,866*	*1,025*
overhead bridge	3,000	1,000	500
open	1,136	332	216
tunnel	738	534	309
Workzone	*2,121*	*1,498*	*662*
no	703	534	309
yes	1418	964	353

Datasets and classes	Sample Size Trn	Vld	Tst
Time-of-day	*1,656*	*1,022*	*300*
daytime	734	400	100
dawn/dusk	216	164	100
night	706	485	100
Weather	*2,798*	*1,400*	*500*
clear	653	300	100
frog	572	300	100
overcast	654	300	100
raining	358	250	100
snowing	561	250	100
Road condition	*2,295*	*957*	*441*
dry	811	353	145
snow	936	325	150
wet	548	279	146

Methodology

Knowledge Acquisition & Accruement Network (KA2N)

- It utilizes teacher-student network architecture
- The feature extractors adopt Swin Transformer base's architecture
- KAqN sequentially and cyclically learns individual tasks from diverse datasets that each provides labels of one attribute
- Learned knowledges accrues in KArN via epoch-wise exponential moving average
- The loss function both guides learning and mitigates forgetting
- Leading to one foundation model with the knowledge to recognize multiple driving scene attributers



Consistency Active Learning (CAL)

- For every scene attribute i, it searches samples from the training datasets without the ground truth label for that attribute, which are in high similarity with attribute i's test data in terms of the feature consistency measure
- A small portion of the identified samples are recommended to domain experts to let them provide the multi-label annotation
- The foundation model is refined using the expert-annotated multi-label samples
- Cross-dataset adaptation is achieved iteratively.

Algorithm 1 Framework of Consistency Active Learning (CAL)

Input:
training dataset D^T, test dataset D^U, budget B for 1 iteration, initial weight ω, network Φ, maximum iteration I, CAL training dataset L
Initialization: Φ with ω, $L = \emptyset$
Output:
for $i = 1, ..., I$ **do**
  for $t \in D^T, u \in D^U$
    Compute consistency score $con(t, u)$ using Eq. (1)
  S_i ←Select top B samples from D^U based on $con(t, u)$
  L_i ← The pair of S_i from D^T
  Update $L = L \cup L_i$
  Acquire labels Y_L for samples in L
  Training Φ with (L, Y_L) using Eq. (2)
  Update $D^T = D^T - L_i$
return Φ

Results

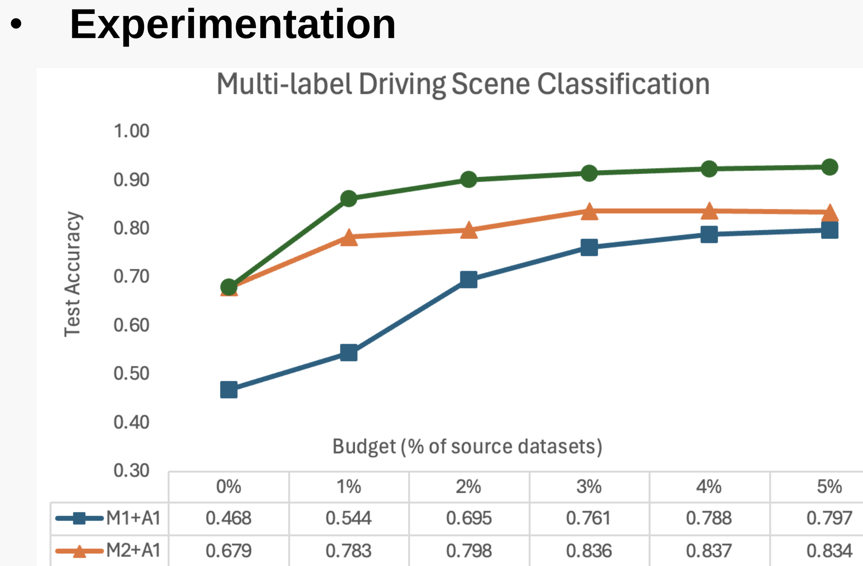
Effectiveness of Knowledge Accruement

- Compare the multi-task foundation model with mono-task models on each of the individual datasets
- The foundation model achieves classification accuracy comparable to mono-task models (±3%)
- The foundation model provides the pretrained weights for cross-dataset adaptation

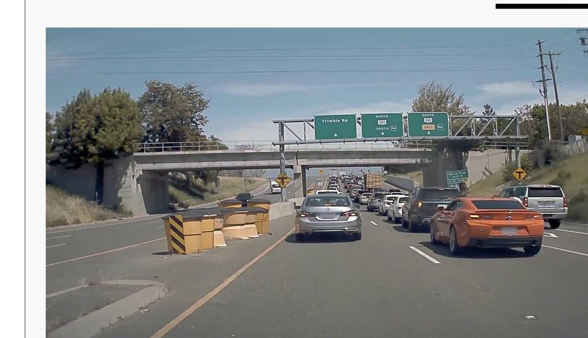
Attributes	Mono-task	Foundation Model (KArN)
Time-of-day	**0.993**	0.990 ↓0.003
Weather	0.900	**0.912** ↑0.012
Road function	0.998	0.998
Road condition	0.980	0.980
Intersection	0.867	**0.894** ↑0.027
Above-road space	0.948	**0.978** ↑0.030
Workzone	**0.943**	0.913 ↓0.030

Effectiveness of Cross-Dataset Adaptation

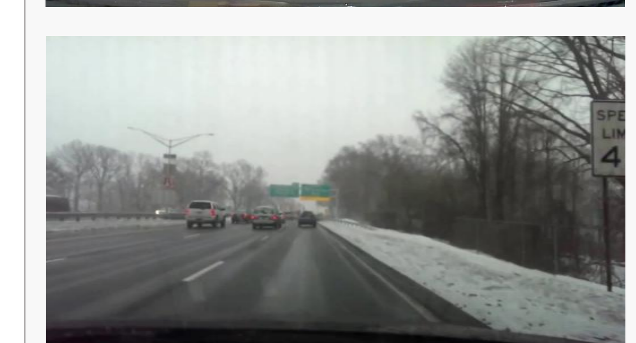
- **Models**: pretrained on ImageNet (M1) vs. trained using KA2N (M2)
- **Adaptation methods**: using randomly-selected expert-annotated samples (A1) vs. CAL (A2)
- **Budget**: up to 5% of source datasets
- **Experimentation**



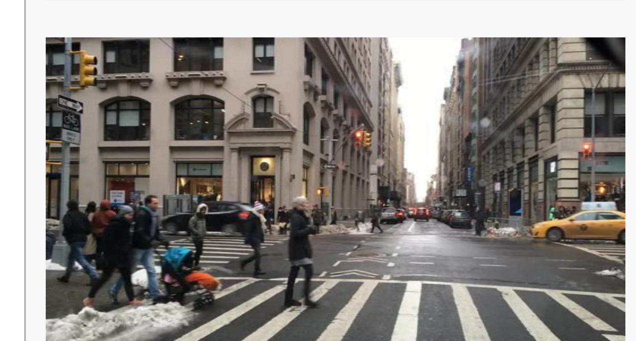
Examples



Time-of-day: daytime 1.00 → 1.00
Weather:: overcast 0.78 → 0.74
Road Function::interstate 0.96 → 0.90
Road Condition::snow 0.95 → dry 0.97
Intersection: no 0.78 → 0.90
Workzone: yes 0.99 → 0.78
Above-road space: bridge 1.00 → 1.00



Time-of-day: dawn 0.98 → daytime 0.90
Weather: snowy 0.98 → 0.88
Road Function: local 0.99 → interstate 0.93
Road Condition: snow 0.99 → 0.97
Intersection: no 0.97 →0.90
Workzone: no 0.99 → 0.99
Space: open road 1.00 → 0.99



Time-of-day: dawn 0.99 → daytime 0.87
Weather: Snowy 0.92→ overcast 0.72
Road function: local 1.00 → 0.99
Road condition: snow 0.99 → 0.91
Intersection: 4-way 0.97 → 0.99
Workzone: no 0.99 → 0.99
Space: open road 1.00 → 0.99

Findings

- The foundation model (w/o cross-dataset adaptation) achieves low test accuracy (67.9%) in multi-label driving scene classification
- But it still outperforms Swin Transformer by 21.1%.
- The advantage of the foundation model over Swin Transformer is diminishing as the budget for adaptation increases
- The cross-dataset adaptation of the foundation model using CAL increases test accuracy by 18.3% with a 1% budget and by 24.8% with a 5% budget
- CAL outperforms the adaptation using randomly selected training samples by 7.8%~10.3%
- The overall gain over the baseline model (M1+A1) is 13% ~ 32%, depending on the budget for adaptation.

Acknowledgement

Contact Information

Ke Li
(ke.li.1@stonybrook.edu)

Conclusions

- The KA2N framework can learn multi-label driving scene classification from diverse sources of datasets that each teaches KA2N on one task. KA2N accrues the learned knowledge to produce a foundation model.
- The challenge of domain shift facing the foundation model can be addressed by CAL through cross-dataset adaptation
- Our proposed approach to multi-label driving scene classification can achieve 86.2% test accuracy with only 1% annotation budget and 92.7% accuracy with 5% annotation budget