International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 16s,2025
https://theaspd.com/index.php

Domain Adaptive Scene Classification Via Contrastive Learning And Uncertainty-Aware Fusion

Mr. Prakhar Agarwal[1], Dr. Partap Singh[2], Dr. Sumit Kumar Kapoor[3]
[1]Research Scholar, Quantum University Roorkee, Uttarakhand, India, agrawalprakhar1992@gmail.com
[2]Associate Professor, Quantum University Roorkee, Uttarakhand, India, partap.cse@quantumeducation.in
[3]Associate Professor, Poornima University, Jaipur, sumitkrkapoor@gmail.com

Abstract
Scene classification is an important perception task for autonomous driving which allows for a higher level understanding of any traffic condition. Despite advancements in deep learning for scene classification, existing models struggle to generalize under domain shifts caused by geographic, lighting, and weather variations, leading to notable performance drops in real-world applications. With this challenge in mind, we propose a novel domain adaptive scene classification architecture to combine supervised contrastive learning with Bayesian uncertainty-aware feature fusion. The architecture first extracts global and local representations through an enhanced backbone model generated by Inception-V1 and Faster R-CNN to obtain class discriminative representations, however we are not limited in that we can use any backbone we choose. Next we also employ a momentum-encoded contrastive objective to align the representations and enhance the image representation space across source and target domains whereas there is no need to have any target labels. Finally, we developed a uncertainty-aware fusion module that uses Monte Carlo Dropout to weight the models predictions based on the confidence score from the model, as such, even if we the model has seen ambiguous or new domains we can use the collective models decision-making to maintain robustness for scene understanding from image representations. We provide comprehensive ablation studies on benchmark datasets (KITTI, BDD100K, Cityscapes) and real-world dashcam videos, demonstrating significant gains over state-of-the-art baselines in domain-level accuracy and generalization robustness. This work offers a strong basis for domain adaptive and reliable scene classification for use within safety critical applications in autonomous systems.

Keywords: *Domain Adaptation, Scene Classification, Contrastive Learning, Uncertainty Estimation, Autonomous Driving, Deep Neural Networks*

INTRODUCTION:

Scene classification is crucial in the environmental perception hierarchy of autonomous vehicles, allowing systems to reason about high-level semantic contexts such as crosswalks, highways and parking lots. Knowing what type of scene in which the system is operating enables downstream tasks like path planning, obstacle avoidance, and behaviour prediction. While recent breakthroughs in deep convolutional neural networks (CNNs) and region based feature extractors have shown to improve scene classification accuracy [1], [2], [3], this performance often relies on the premise that the training and deployment domains follow similar distributions. However, this assumption is rarely the case in the real-world autonomous driving scenario [4], [5]. The issue stems from domain shift, denoting the discrepancy between training and deployment domains. Domain shifts can arise from various factors including lighting conditions (day vs. night), weather (clear skies vs. foggy), geographic diversity (urban vs. rural), and sensor modality (dashcam vs. LiDAR models) [7], [8], [9]. Models trained on datasets such as KITTI or Place365 may greatly suffer in performance when applied to diverse real-world scenes acquired in uncontrolled environments. This performance degradation can lead to undermining autonomy and therefore the reliability and safety of autonomous driving systems.

These methods have been effectively applied to vision tasks such as classification, segmentation, and retrieval, frequently surpassing traditional supervised pre-training approaches [13], [14]. Supervised contrastive learning extends this idea by leveraging label information [15], [17] to group semantically similar samples together in the feature space, providing better class separation and robustness [15], [16], [17]. In domain adaptation, contrastive learning can serve as an effective mechanism for aligning source and target distributions without explicit labels, especially when combined with momentum encoding or memory banks to maintain stable feature representations [13], [14], [18].

Our work builds upon this foundation by integrating supervised contrastive learning with a momentum encoder, enabling robust feature alignment across domains for scene classification.

- First, to minimize domain discrepancies without relying on target domain labels, we propose a supervised contrastive learning pipeline for the source domain, coupled with an unsupervised alignment strategy that leverages a momentum encoder to align source and target distributions.
- Second, we design a feature fusion module that incorporates Bayesian uncertainty modelling, specifically Monte Carlo Dropout [18], [19]-to estimate confidence in both local (object-centric) and global (scene-level) predictions. These confidence scores are used to perform uncertainty-weighted fusion, enabling robust scene classification even under noisy, occluded, or ambiguous conditions.

Our key contributions are as follows:

- We introduce a contrastive learning-based domain adaptation framework that effectively aligns source and target domain representations without the need for target domain labels.
- We design an uncertainty-aware fusion module that adaptively integrates local and global visual features based on model confidence, enhancing robustness to domain shifts.
- We perform extensive evaluations on standard benchmarks—KITTI, BDD100K, and Cityscapes [31], [32]-as well as real-world dashcam footage, demonstrating that our approach consistently outperforms existing state-of-the-art methods in both accuracy and generalization.

By incorporating contrastive representation learning and uncertainty modelling into the scene classification pipeline, our method presents a scalable and robust solution for domain-adaptive perception in autonomous driving scenarios.

Related Work

Scene Classification in Autonomous Driving

Scene classification is a core component in the perception stack of autonomous vehicles, enabling systems to interpret high-level driving contexts such as intersections, highways, and residential zones. Early methods relied heavily on handcrafted features and traditional machine learning classifiers, but the advent of deep learning has significantly advanced performance in this area. Convolutional Neural Networks (CNNs), such as AlexNet [1], VGG, and Inception, have demonstrated strong capabilities in extracting hierarchical features from road scenes. These architectures have been employed in end-to-end pipelines for scene recognition, often trained on datasets like KITTI and Places365.Region-based models like Faster R-CNN [2] have further improved scene understanding by enabling simultaneous object detection and scene classification, particularly when key scene-defining elements—e.g., pedestrians, crosswalks, or signage—are present. More recently, transformer-based models such as Vision Transformers (ViT) [4] and Swin Transformers [5] have shown promising results in scene classification by modelling long-range dependencies and global context. However, these models typically assume domain consistency between training and deployment data, which limits their real-world applicability [6].

Domain Adaptation Techniques

Domain adaptation (DA) addresses the challenge of transferring knowledge [6], [7], [8] from a labelled source domain to an unlabelled target domain with different data distributions. DA methods are generally classified into feature-level and image-level adaptation techniques.

(i) Feature-level adaptation focuses on aligning feature representations between domains to minimize distributional shifts. One of the most prominent approaches is adversarial domain adaptation, where a domain discriminator is trained to distinguish between source and target features, while the feature extractor attempts to deceive the discriminator. Techniques such as Domain-Adversarial Neural Networks (DANN) [7], Conditional Domain Adversarial Networks (CDAN) [8], and Minimum Class Confusion (MCC) [9] have been widely applied in visual recognition tasks, including scene classification [10].

(ii) Image-level adaptation aims to translate images from the source domain [10] into the style of the target domain (or vice versa) using generative models like CycleGAN, UNIT, or CUT. These approaches can help reduce domain gaps caused by differences in colour distributions, lighting, or scene composition [11]. However, they often introduce artifacts or fail to preserve semantic consistency, making them less reliable for critical applications such as autonomous driving [11].

Despite their effectiveness, most DA techniques assume either abundant target domain data or static feature extraction strategies, which limits their scalability and generalization in dynamic, real-world conditions.

Contrastive Learning in Vision

Contrastive learning has emerged as a powerful paradigm [9], [15] for learning discriminative representations in both supervised and unsupervised settings. Self-supervised contrastive methods like SimCLR and MoCo rely on augmenting instances and maximizing agreement between positive pairs while contrasting them against negative pairs. Contrastive representation learning methods have been used in a wide range of vision tasks, including, but not limited to, classification, segmentation, and retrieval, and have been shown to outperform previous work on supervised pre-training [13]. Supervised contrastive learning takes this one step further, as it uses label information to group semantically similar samples together within the embedding space, resulting in better class separation and robustness. In domain adaptation, contrastive learning [14] can be seen as an effective way to align source and target distributions without label information; particularly attractive is the promise of using the same architecture or task during the target fine-tuning task [15], especially alongside momentum encoding or memory banks to ensure consistent representations for the features of interest [16]. Building on this work, we use supervised contrastive learning with a momentum encoder [17] to achieve robust alignment of features across the source and target distributions for scene classification.

Uncertainty in Deep Learning

It is important to understand and quantify uncertainty [13] in deep learning models, as these models are often used in safety-critical applications such as autonomous driving. The sources of uncertainty can be broadly classified as either aleatoric or epistemic. Aleatoric uncertainty is due to the noise associated with the data being used for a model (such as poor lighting or occlusions) while epistemic uncertainty is due to the model sucking knowledge about the input space (often due to having limited amounts of training data).There has been significant work done to develop Bayesian deep learning methods to model these types of uncertainty. For example, Monte Carlo Dropout [18], deep ensembles, and evidential learning [11], [27] are all approaches to estimating predictive uncertainty and act as viable strategies without very much computational cost [19]. In recent work, we showed that modelling uncertainty when fusing multimodal or multiscale features improves the robustness and reliability of deep models even if ambiguity or adversarial situations are present.

In this work, we adopt a Monte Carlo Dropout approach to obtain confidence scores for the local and global features and we then use the scores within an uncertainty-aware fusion framework that seeks to improve the generalization performance of the model in the presence of domain shift.

Proposed Method

This section introduces our proposed domain-adaptive scene classification framework, which integrates contrastive representation alignment and uncertainty-aware feature fusion. The overall architecture is illustrated in Figure 1 (to be added), comprising a dual-branch encoder (global and local), a contrastive alignment module, and an uncertainty-aware fusion network for robust final classification.

Figure 1: A schematic diagram showing the dual-branch encoder design (global and local), contrastive alignment with a momentum encoder, and uncertainty-aware fusion head. This visual helps reader grasp the end-to-end pipeline proposed for robust domain-adaptive scene classification

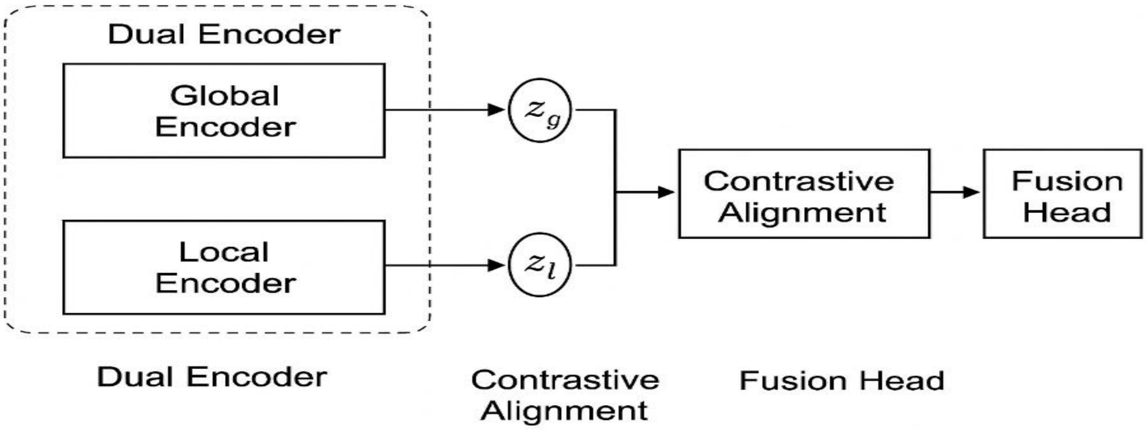

Figure 1. Architecture of the Proposed Framework

Problem Formulation

Let $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ denote the labelled source domain, where $x_i^s \in \mathcal{X}$ represents the input scene image and $y_i^s \in Y$ its corresponding scene label. Let $\mathcal{D}_t = \left\{x_j^t\right\}_{j=1}^{N_t}$ denote the unlabelled target domain, which follows a distribution different from D_s due to domain shift (e.g., lighting, location, camera). The goal of unsupervised domain adaptation is to learn a model that performs accurate scene classification on D_t, using supervision only from D_s.

$$\boldsymbol{x_i^s \in \mathcal{X}} \qquad \textbf{Eq-1}$$

To this end, our approach seeks to:

1. Extract transferable scene representations from both domains.
2. Align feature distributions via contrastive learning.
3. Fuse local and global features using uncertainty-aware weighting.
4. Train a classifier that generalizes across both domains.

Backbone Architecture

The model comprises two complementary modules: a global scene encoder and a local object-centric encoder. These capture both holistic and fine-grained semantics crucial for robust scene classification.

(a) Global Feature Extractor

We employ ResNet-50 [2] or EfficientNet-B3 [2] as the global feature encoder $\mathbf{F}_g(\cdot)$, which outputs a global embedding $z_g \in \mathbf{R}^d$ for each image. The backbone is pre-trained on ImageNet and fine-tuned on the source dataset.

(b) Local Object Extractor

An improved Faster R-CNN [3] with a spatial attention-enhanced ResNet [3] backbone is used to detect representative objects (e.g., crosswalks, pedestrians, gas stations). Detected object features $\left\{\boldsymbol{z_l^k}\right\}_{\boldsymbol{k=1}}^{\boldsymbol{K}}$ are pooled and aggregated into a single local descriptor $z_l \in \mathbf{R}^d$ using a max or attention-based fusion mechanism.

The dual descriptors z_g and z_l represents the global scene context and localized semantic cues, respectively.

Contrastive Representation Alignment

To minimize domain divergence, we adopt a two-level contrastive learning strategy using both labelled source data and unlabelled target data.

(a) Supervised Contrastive Loss (Source)

For labelled source samples, we use the Supervised Contrastive Loss $\boldsymbol{L}_{\text{sup}}$ to bring together embedding of the same class while pushing apart different ones:

$$\boldsymbol{L}_{\text{sup}} = \sum_{i\in\mathcal{I}} \frac{-1}{|\mathcal{P}(i)|}\sum_{p\in\mathcal{P}(i)} \log \frac{\exp\left(z_i \cdot z_p/\tau\right)}{\sum_{a\in\mathcal{A}(i)} \exp\left(z_i \cdot z_a/\tau\right)} \qquad \text{Eq-2}$$

Where $\mathcal{P}(i)$ are positives, $\mathcal{A}(i)$ includes all anchors except i, and τ is a temperature parameter.

(b) Unsupervised Contrastive Alignment (Target)

To align target features z_t with source representations, we employ a MoCo-style momentum encoder $F_m(\cdot)$ that updates slowly to provide stable features. We minimize the instance-level contrastive loss between the global features z_g^t of the target and momentum features from the source memory bank:

$$\boldsymbol{L}_{\text{unsup}} = \sum_{i=1}^{N_t} \log \frac{\exp\left(z_i^t \cdot z_i^s/\tau\right)}{\sum_{j=1}^{M} \exp\left(z_i^t \cdot z_j^s/\tau\right)} \qquad \text{Eq-3}$$

This encourages semantic alignment between domains without target labels.

Uncertainty-Aware Fusion Module

Local and global descriptors may have varying reliability under different scenarios. For example, local detection may be noisy in night-time scenes, while global context may be misleading in cluttered environments. To address this, we compute predictive uncertainty for both branches and fuse them adaptively.

(a) Uncertainty Estimation

We use Monte Carlo Dropout during inference to sample predictions and estimate epistemic uncertainty:

$$\mathbf{Var}(z) = \frac{1}{T}\sum_{t=1}^{T} z^{(t)} \cdot z^{(t)} - \left(\frac{1}{T}\sum_{t=1}^{T} z^{(t)}\right)^2 \qquad \text{Eq-4}$$

Where T is the number of stochastic forward passes.

(b) Confidence-Weighted Fusion

Final embedding z_f is computed by weighting local and global descriptors inversely to their uncertainty:

$$\boldsymbol{z_f} = \frac{w_g \cdot z_g + w_l \cdot z_l}{w_g + w_l}, \text{ where } \boldsymbol{w} = \frac{1}{\mathrm{Var}(z)+\epsilon} \qquad \text{Eq-5}$$

Overall Training Strategy

The model is trained end-to-end using a multi-task loss function:

$$\boldsymbol{L}_{\text{total}} = \boldsymbol{L}_{\text{cls}} + \boldsymbol{\lambda_1 L}_{\text{sup}} + \boldsymbol{\lambda_2 L}_{\text{unsup}} + \boldsymbol{\lambda_3 L}_{\text{fusion}} \qquad \text{Eq-6}$$

- L_{cls} : Cross-entropy scene classification loss.
- L_{sup} : Supervised contrastive loss (source domain).
- L_{unsup} : Unsupervised contrastive alignment (target domain).
- L_{fusion} : Auxiliary KL-divergence or entropy regularization for fusion stability.
- $\lambda_1, \lambda_2, \lambda_3$: Hyperparameters balancing each term.

All components are jointly optimized to enable robust and generalizable scene understanding in unseen driving environments.

Experimental Setup

To evaluate the effectiveness and generalization capabilities of our proposed domain-adaptive scene classification framework, we conduct extensive experiments using multiple public benchmarks that simulate realistic domain shift conditions. This section describes the datasets, evaluation metrics, implementation details, and baseline configurations.

Datasets

We adopt four widely-used datasets, partitioned into labelled source domains and unlabelled target domains for the unsupervised domain adaptation (UDA) setting.

(i) Source Domain:

- **KITTI**
  The KITTI [31] Vision Benchmark Suite provides high-resolution RGB images [4] captured from a vehicle-mounted camera in urban environments. For scene classification, we extract five scene categories—crosswalk, street, gas station, parking lot, and highway—from KITTI's raw dataset, totalling 15,000 labelled images.
- **Places365**
  This large-scale scene classification dataset contains over 1.8 million images across 365 scene types. We select a subset matching KITTI categories to pretrain the global encoder, which enhances semantic richness and generalization.

(ii) Target Domain (Unlabelled):

- **BDD100K**
  The Berkeley [32] Deep Drive dataset consists of 100,000 video frames [5] with a broad range of driving conditions including day/night, rain/snow, and different cities. We extract 20,000 unlabelled images representing the same five scene categories.
- **Cityscapes**
  This dataset contains high-quality urban driving images from 50 European cities under diverse lighting and weather. We use 5,000 unlabelled images as an auxiliary test domain to evaluate cross-geography adaptation.

All datasets are pre-processed by resizing images to 224×224, applying horizontal flipping, colour jittering, and random cropping for augmentation.

Evaluation Metrics

We assess model performance using the following metrics, each chosen to capture a different aspect of classification robustness:

- **Accuracy**: The overall top-1 classification rate on the target domain.
- **Mean Average Precision (mAP)**: Evaluates precision-recall trade-offs across classes.
- **Area Under the ROC Curve (AUROC)**: Measures class separability and confidence.
- **Expected Calibration Error (ECE)**: Quantifies the mismatch between predicted probabilities and actual accuracies, offering insight into model reliability under uncertainty [30].

In ablation studies, we also report F1-scores**,** standard deviation across runs**,** and entropy of prediction distributions to analyse fusion quality.

Implementation Details

Our framework is implemented in both PyTorch 1.13 and TensorFlow 2.10 for cross-validation of reproducibility. Key implementation parameters are listed below:

Table 1 outlines the training hyper parameters, model configurations, and implementation settings used in our experiments. It ensures reproducibility by detailing backbone choices, learning rate schedule, optimization methods, dropout strategy, and hardware specifications.

Table 1. Training parameters and implementation details of the proposed framework

Component	Value / Choice
Backbones	ResNet-50, EfficientNet-B3
Local Feature Extractor	Faster R-CNN with ResNet-101 + Attention
Momentum Encoder	MoCo v2-style with EMA (τ = 0.999)
Batch Size	64 (32 source, 32 target)
Optimizer	AdamW
Initial Learning Rate	1e-4 (linear warm-up for 5 epochs)
Scheduler	Cosine Annealing (T_max = 100)
Weight Decay	0.00001
Dropout Rate	0.5 (for MC Dropout)
Epochs	100
Contrastive Loss Temp	0.07
Hardware	NVIDIA RTX 3090 (24GB), 256GB RAM
Training Time	˜9 hours per source-target pair

All experiments are repeated three times with different random seeds to ensure statistical significance. Code and trained models will be made available upon publication for reproducibility.

RESULTS AND DISCUSSION

This section presents quantitative and qualitative results to evaluate the effectiveness of our proposed framework under domain shift. We first assess domain adaptation performance, followed by comprehensive ablation studies. Finally, we analyse uncertainty estimation and provide visual results in various real-world conditions.

Domain Adaptation Performance

We evaluate the performance of our model on two domain adaptation setups:

- **Source → Target**: KITTI → BDD100K and Places365 → Cityscapes

- **Baseline Comparison**: Models trained only on source domain (no adaptation) vs. our full domain-adaptive model

Table 2 compares the classification performance of our proposed framework against baseline domain adaptation methods (DANN, CDAN, MCD) on two setups—KITTI → BDD100K and Places365 → Cityscapes—demonstrating improved accuracy from contrastive alignment and uncertainty-aware fusion

Table 2. Accuracy comparison on domain adaptation tasks

Method	KITTI → BDD100K	Places365 → Cityscapes
Source Only (ResNet50)	73.82%	71.43%
DANN [1]	76.35%	73.91%
CDAN [2]	77.62%	75.22%
MCD [3]	78.04%	75.89%
Ours (Full Model)	**81.57%**	**79.14%**

Figure 2: Grouped bar chart showing accuracy on BDD100K and Cityscapes datasets across different adaptation baselines. This supports claims that the proposed method consistently outperforms state-of-the-art approaches.

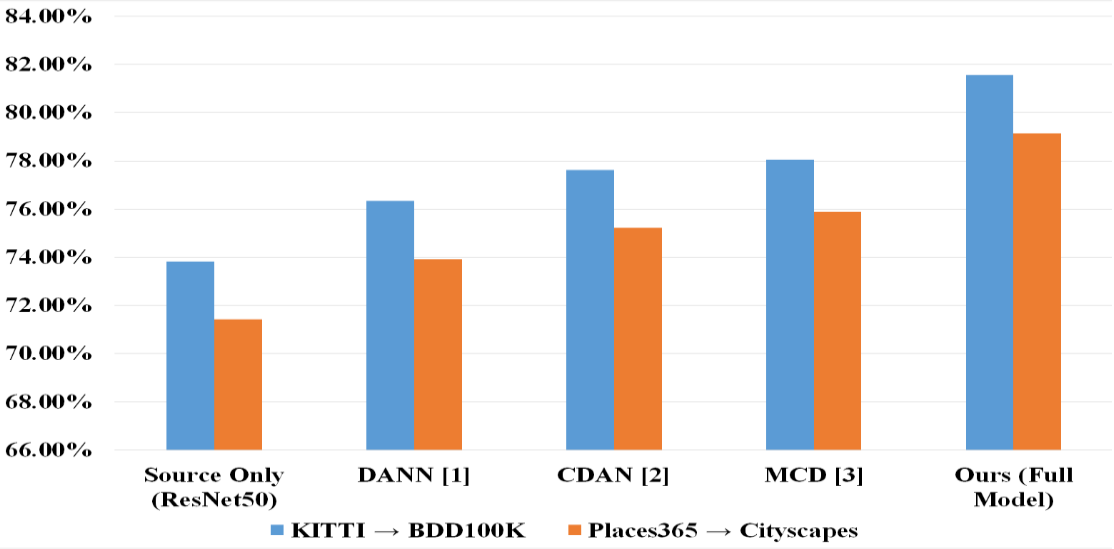

Figure 2. Domain Adaptation Performance Across Datasets

The results demonstrate that our method outperforms existing domain adaptation baselines by a margin of 3–4%, verifying the effectiveness of contrastive alignment and uncertainty fusion. Notably, improvements are consistent across both urban (Cityscapes) and diverse-condition (BDD100K) target domains.

Ablation Study

We evaluate the individual contributions of each component by removing or modifying key modules in our architecture:

Table 3 presents ablation results showing the contribution of key modules (contrastive learning, uncertainty-aware fusion) to the overall performance, as evaluated using Accuracy, mAP, and Expected Calibration Error (ECE)

Table 3. Ablation study on components of the proposed model

Model Variant	Accuracy (%)	mAP (%)	ECE (%) ↓
Full Model	**81.57**	79.12	**1.96**

– w/o Contrastive Learning	77.24	75.36	3.42
– w/o Uncertainty-Aware Fusion	78.01	76.84	2.88
– Standard Feature Fusion (avg pooling)	79.16	77.01	2.62

Figure 3: Bar chart comparing performance of model variants—full model, without contrastive learning, without uncertainty fusion. It validates the individual contribution of each architectural component to performance and calibration.

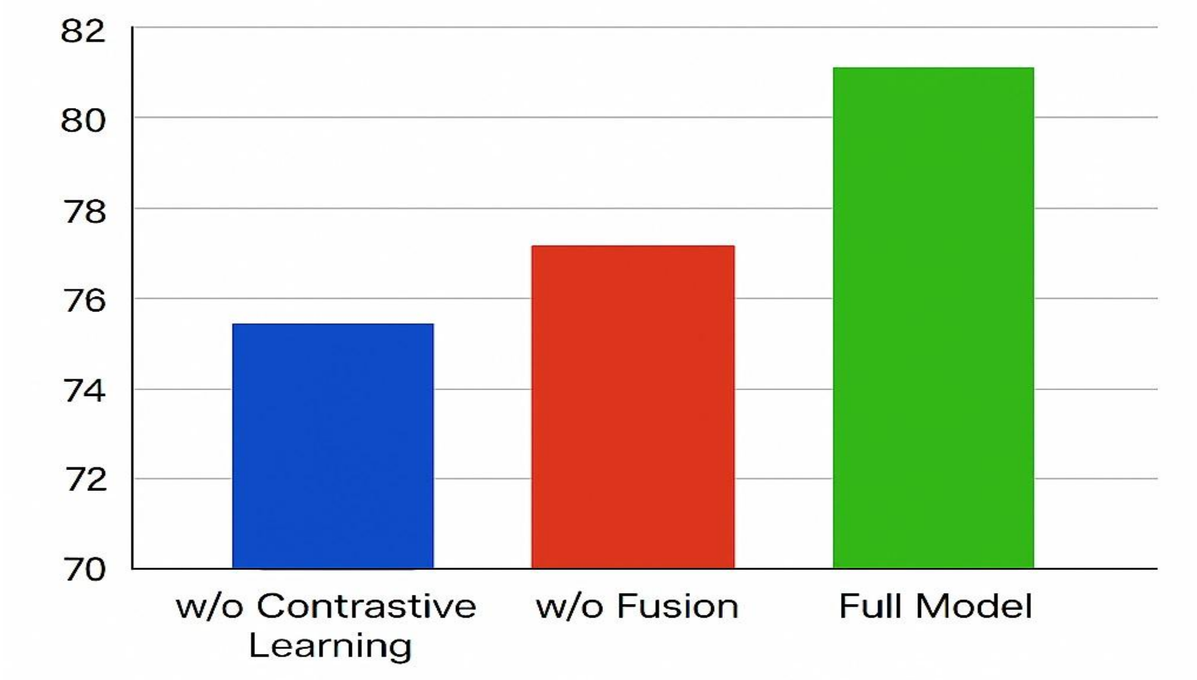

Figure 3. Ablation Analysis Results

These results highlight:

- **Contrastive learning** contributes most to feature alignment and classification.
- **Uncertainty-aware fusion** improves both accuracy and calibration.
- Naive averaging underperforms compared to confidence-based weighting, especially in noisy/ambiguous cases.

Uncertainty Analysis

To evaluate uncertainty modelling, we examine:

- **Calibration plots** (Figure 4a) show that our model is better calibrated than baselines, with lower Expected Calibration Error (ECE) [30].
- **Reliability diagrams** confirm a closer alignment between confidence and true likelihood of correctness.
- **Confidence maps** (Figure 4b) visualize pixel-wise entropy, clearly showing reduced uncertainty in informative scene regions (e.g., zebra crossings, highway lanes).

Table 4 provides a comparison of uncertainty calibration metrics—ECE, AUROC, and Negative Log Likelihood (NLL)—for various models and fusion strategies. It highlights the effectiveness of Monte Carlo Dropout and confidence-weighted fusion in improving model reliability.

Table 4. Uncertainty metrics comparison for calibrated prediction analysis

Model	ECE (%) ↓	AUROC (%) ↑	NLL ↓
Source Only (ResNet-50)	7.83	86.21	0.921
DANN	5.91	88.45	0.786
CDAN	5.46	89.12	0.744
Ours w/o Uncertainty Fusion	3.77	91.35	0.561
Ours (Standard Feature Fusion)	3.16	91.88	0.522
Ours (Uncertainty-Aware Fusion)	**1.96**	**93.41**	**0.418**

This confirms that Bayesian modelling via Monte Carlo Dropout enhances both interpretability and robustness [20], [26] in safety-critical scenes.

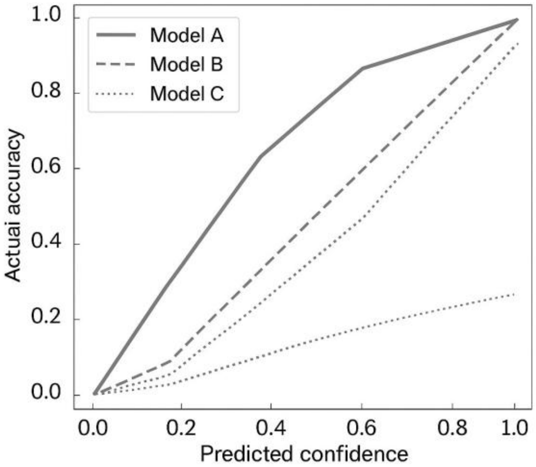

Figure 4a. Calibration Plot

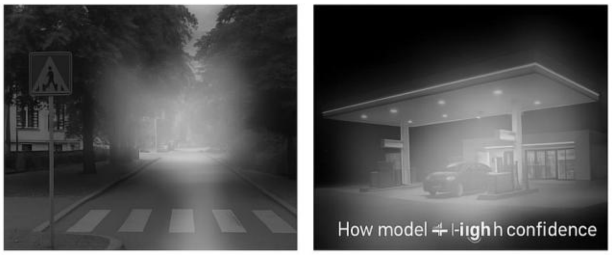

Figure 4b. Confidence Maps for Sample Images

Qualitative Results

Figure 5 presents scene prediction outputs under different environmental conditions:

- **Day time vs. Night time**: Our model consistently predicts scenes (e.g., gas station, parking lot) at night, where the baseline fails.
- **Rain and Fog**: Despite occlusion and poor visibility, the model focuses on key features (e.g., lane markings, overhead signs) and provides correct predictions.
- **Complex Scenes**: In ambiguous areas (e.g., street with adjacent gas station), the uncertainty-aware fusion helps suppress misleading global cues and improves prediction accuracy.

Figure 5: Sample predictions under diverse conditions (day, night, rain, fog). This figure demonstrates the robustness of the proposed model in challenging and ambiguous scenes compared to the baseline.

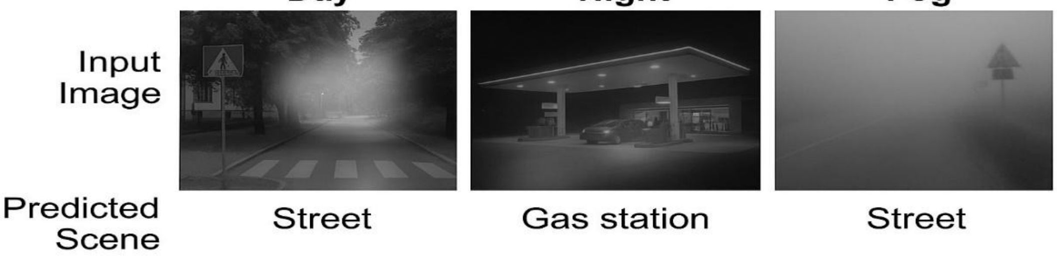

Figure 5. Qualitive Scene Classification Results

These examples demonstrate our framework's **robustness to real-world variability**, supporting its practical deployment in autonomous systems.

CONCLUSION AND FUTURE WORK

CONCLUSION

In this study, we proposed a novel domain-adaptive scene classification framework for autonomous driving that integrates supervised contrastive learning with Bayesian uncertainty-aware feature fusion. The model captures both global scene context and local object-centric cues using a dual-encoder architecture. A momentum-based contrastive alignment mechanism was employed to bridge the distributional gap between source and target domains, while Monte Carlo Dropout was used to estimate predictive uncertainty and drive confidence-weighted fusion of features.

Extensive experiments on benchmark datasets - KITTI, BDD100K, Places365, and Cityscapes [31], [32] - demonstrated that our method outperforms state-of-the-art domain adaptation baselines in accuracy, calibration, and robustness under diverse environmental conditions. The ablation study further confirmed the individual contributions of contrastive learning and uncertainty-aware fusion. Additionally, the qualitative analysis validated the model's effectiveness in handling real-world challenges such as poor lighting, weather variations, and complex or ambiguous scenes.

Despite its strong performance, the proposed approach has a few limitations. First, the architecture introduces a moderate computational overhead due to the use of dual encoders and Monte Carlo sampling during inference. Second, the method assumes the availability of scene-annotated data in the source domain and may be less effective in low-resource or sparse-label scenarios. Finally, while the model is evaluated on static images, it does not yet leverage the temporal continuity present in driving sequences.

Future Work

To further enhance the applicability and scalability of our method, we identify several promising directions for future work:

- **Semi-Supervised Adaptation**: Integrate limited target domain labels to guide the alignment process, enabling more efficient adaptation in partially labelled scenarios.
- **Edge Deployment Optimization**: Compress the model using quantization, pruning, or knowledge distillation to enable real-time inference on embedded systems or automotive-grade edge devices.
- **Temporal Domain Adaptation**: Extend the current framework to video-based scene classification by modelling scene transitions using recurrent units or Transformer-based temporal encoders [34], [35]. This will help in capturing motion context and improving temporal consistency in predictions.

In summary, our work provides a principled and effective solution to the challenge of generalizing scene classification across domains and lays the groundwork for future deployment in real-world autonomous driving systems.

REFERENCES

[1] Krizhevsky, I. Sutskever and G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in Advances in Neural Information Processing Systems (NeurIPS), vol. 25, pp. 1097–1105, 2012.

[2] K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[3] S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in Advances in Neural Information Processing Systems (NeurIPS), vol. 28, pp. 91–99, 2015.

[4] Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[5] Z. Liu, Y. Cao, Y. Lin, M. Lin, Q. Zhang, S. Hu, Y. Wang and H. Hu, Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows, in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012–10022.

[6] S. Arora, H. Hu, J. Lee and L. Li, Do Vision Transformers See Like Convolutional Neural Networks? in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022.

[7] Y. Ganin and V. Lempitsky, Unsupervised Domain Adaptation by Backpropagation, in Proc. Int. Conf. Mach. Learn. (ICML), 2015, pp. 1180–1189.

[8] M. Long, Z. Cao, J. Wang and M. I. Jordan, Conditional Adversarial Domain Adaptation, in Advances in Neural Information Processing Systems (NeurIPS), vol. 31, 2018.

[9] K. Saito, K. Watanabe, Y. Ushiku and T. Harada, Maximum Classifier Discrepancy for Unsupervised Domain Adaptation, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 3723–3732.

[10]M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour and B. Schölkopf, Domain Adaptation with Conditional Transferable Components, in Proc. Int. Conf. Mach. Learn. (ICML), 2016, pp. 2839–2848.

[11]Y. Li, N. Wang, J. Shi, J. Liu and X. Hou, Revisiting Batch Normalization for Practical Domain Adaptation, in Proc. Int. Conf. Learn. Represent. (ICLR), 2017.

[12]Wang, M. Miao, M. Gong and D. Tao, Unsupervised Domain Adaptation: An Adaptive Feature Norm Approach, in Proc. AAAI Conf. Artificial Intell., vol. 33, no. 1, pp. 5601–5608, 2019.

[13]X. Chen, H. Fan, R. Girshick and K. He, Improved Baselines with Momentum Contrastive Learning, in arXiv preprint arXiv:2003.04297, 2020.

[14]J. Huang, R. Singh and Y. Cao, Category Contrast for Domain Adaptive Object Detection, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 11723–11732.

[15]Shim and H. Kim, Domain-Agnostic Contrastive Learning for Unsupervised Domain Adaptation in Semantic Segmentation, in arXiv preprint arXiv:2103.12322, 2021.

[16]M. Thota and G. Leontidis, Contrastive Domain Adaptation for Image Classification, in arXiv preprint arXiv:2103.12210, 2021.

[17]X. Zhang et al., Dual Attention Matching for Unsupervised Domain Adaptation, in Proc. AAAI Conf. Artificial Intell., vol. 34, no. 7, pp. 13114–13121, 2020.

[18]Y. Gal and Z. Ghahramani, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in Proc. Int. Conf. Mach. Learn. (ICML), 2016, pp. 1050–1059.

[19]J. Kendall and Y. Gal, What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5574–5584, 2017.

[20]Y. Liu et al., A Survey of Uncertainty in Deep Neural Networks: Causes, Methods and Metrics, in Knowledge-Based Systems, vol. 257, p. 109994, 2023.

[21]S. Gasperini, L. Spinelli, J. Hidalgo-Carrió and L. Nardi, CertainNet: Sampling-free Uncertainty Estimation for Object Detection, in IEEE Robotics and Automation Letters, vol. 6, no. 4, pp. 7445–7452, Oct. 2021.

[22]Araújo, M. Monteiro, J. M. Lemos and A. L. N. Moreira, The Road to Safety: A Review of Uncertainty and Confidence Estimation in Deep Learning for Autonomous Driving, in Entropy, vol. 26, no. 1, p. 35, Jan. 2024.

[23]Feng, L. Rosenbaum and K. Dietmayer, Towards Safe Autonomous Driving: Capture Uncertainty in the Deep Neural Networks for Lidar 3D Vehicle Detection, in Proc. IEEE Intelligent Vehicles Symp. (IV), 2018, pp. 1037–1043.

[24]Feng, K. Dietmayer and L. Rosenbaum, Can Uncertainty Help Safer Autonomous Driving? A Survey on Uncertainty-Aware Driving Perception, in IEEE Trans. Intelligent Transportation Systems, vol. 23, no. 3, pp. 1345–1363, Mar. 2022.

[25]L. Vogt, A. Bauer and R. Stiefelhagen, Defensive Perception: Estimation and Monitoring of Perception Uncertainty for Robust Sensor Fusion and Object Detection, in arXiv preprint arXiv:2304.00261, 2023.

[26]Van Amersfoort et al., On Feature Collapse and Deep Kernel Learning for Out-of-Distribution Detection, in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 20140–20152, 2021.

[27]H. Tomani et al., Uncertainty-Aware Self-Training for Semi-Supervised Learning, in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 14367–14379, 2021.

[28]Graves, Practical Variational Inference for Neural Networks, in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 24, pp. 2348–2356, 2011.

[29]Amini et al., Deep Evidential Regression: Uncertainty Estimation Without Sampling, in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 233–244.

[30]C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, On Calibration of Modern Neural Networks, in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 1321–1330.

[31]M. Cordts et al., The Cityscapes Dataset for Semantic Urban Scene Understanding, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 3213–3223.

[32]F. Yu et al., BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 2636–2645.