Computer Vision based Large Scale Urban Mobility Audit and Parametric Road Scene Parsing

Thesis submitted in partial fulfillment
of the requirements for the degree of

*Master of Science in **Computer Science and Engineering** by Research*

by

Durga Nagendra Raghava Kumar Modhugu
2018900018
durga.nagendra@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
December 2022

Copyright © Durga Nagendra Raghava Kumar Modhugu, 2022
All Rights Reserved

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **" Computer Vision for Large Scale Urban mobility audit and Parametric Road Scene Parsing with Satellite Imagery"** by Durga Nagendra Raghava Kumar Modhugu, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Prof. C V Jawahar

To my beloved family

Acknowledgments

I would like to thank the people who have made my journey memorable and meaningful. First and foremost, I am deeply indebted to my advisors, Prof. C V Jawahar and Prof. Manmohan Chandraker for their constant guidance and support. Throughout my research journey, Prof. C V Jawahar has enabled me the freedom to grow and learn by making decisions and provided me with excellent opportunities to gain tremendous exposure, thereby making me a better researcher and professional. I sincerely thank Asokan Pichai for introducing me to the research opportunity at CVIT. I will also forever cherish enjoyable, thought-provoking discussions, with Prof. C V Jawahar, that have shaped my research outlook and personality.

I also wish to thank my friends, starting with my research project mates Harish Rithish and Ranjith Reddy. In retrospection, I think our rigorous and uncompromising conversations (most often debates :P) have helped us push each other's boundaries and resulted in better outcomes. I will certainly cherish all those memories and camaraderie throughout the thick and thin of the research. I also thank my lab mates Ashish, Deepak, Sangeeth, Rudrabha, Riya, Siddhanth and seniors Praveen Krishnan and Bhavani for all the help and guidance. This acknowledgment would be incomplete without thanking Arun KS for being with me throughout this journey.

Finally, it is impossible to thank the positive force behind each of my successful endeavors: my family for their selfless affection, support, and advice.

Abstract

The footprint of partial or fully autonomous vehicles is increasing gradually with time. The existence and availability of the necessary modern infrastructure are crucial for the widespread use of autonomous navigation. One of the most critical efforts in this direction is to build and maintain HD maps efficiently and accurately. The information in HD maps is organized in various levels 1) Geometric layer, 2) Semantic layer and 3) Map prior's layer. The conventional approaches to capturing and extracting information at different HD map levels rely heavily on huge sensor networks and manual annotation. This is not scalable to create HD maps for massive road networks. We propose two novel solutions to address the mentioned problems in this work. The first solution deals with the generation of the geometric layer with parametric information of the road scene and other one to update information on road infrastructure and traffic violations in the semantic layer.

Firstly, the creation of the geometric layer of the HD map requires understanding the road layout in terms of structure, number of lanes, lane width, curvature, etc. Prediction of these attributes as part of a generalizable parametric model with which road layout can be rendered would suite the creation of a geometric layer. Many previous works that tried to solve this problem rely only on ground imagery and are limited by the narrow field of view of the camera, occlusions, and perspective shortening. This work demonstrates the effectiveness of using aerial imagery as an additional modality to overcome the above challenges. We propose a novel architecture, Unified, that combines aerial and ground imagery features to infer scene attributes. We quantitatively evaluate on the KITTI dataset and show that our Unified model outperforms prior works. Since this dataset is limited to road scenes close to the vehicle, we supplement the publicly available Argoverse dataset with scene attribute annotations and evaluate far-away scenes. We quantitatively and qualitatively show the importance of aerial imagery in understanding road scenes, especially in regions farther away from the ego-vehicle.

Finally, we also propose a simple mobile imaging setup to address and audit several common problems in urban mobility and road safety, which can enrich the information in a semantic layer of HD maps. Recent computer vision techniques are used to identify street irregularities (including missing lane markings and potholes), absence of street lights, and defective traffic signs using videos obtained from a moving camera-mounted vehicle. Beyond the inspection of static road infrastructure, we also demonstrate the applicability of mobile imaging solutions to spot traffic violations. We validate our proposal on the long stretches of unconstrained road scenes covering over 2000Km and discuss practical challenges in applying computer vision techniques at such a scale. Exhaustive evaluation is carried

out on 257 long-stretches with unconstrained settings and 20 conditions-based hierarchical frame-level labels for different timings, weather conditions, road type, traffic density, and state of road damage. For the first time, we demonstrate that large-scale analytics of irregular road infrastructure is feasible with existing computer vision techniques.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Transportation plays a prominent role in serving the people's fundamental personal and economic needs. It helps link people and connect societies. Thus, enabling the ecosystem for development by promoting trade through markets. Road Transport has emerged as the dominant segment in the transportation landscape worldwide. For example, in India, road transportation contributes around 4.5% of its GDP and accounts for 87 % of the passenger traffic and 60% of the freight traffic movement [59]. Despite of its importance, road transport infrastructure around the world is poorly planned or managed, and the result is often inadequate and poorly maintained road infrastructure. The growing urban population, especially in Asia and Africa, compounds the challenges of road transport/mobility [61].

The United Nations' report on sustainable transportation defines it as the provision of services and infrastructure for the mobility of people and goods, advancing economic and social development to benefit today's and future generations in a manner that is safe, affordable, accessible, efficient, and resilient [61]. This recent definition of sustainable transport shifts the focus away from only the availability of the mobility infrastructure and emphasizes safety and inclusion in terms of affordability. This focus change sets ambitious goals and challenges to reshape the road transport/mobility sector towards sustainability. These challenges are enormous, and an opportunity to increase the use of innovative technologies such as computer vision and deep learning.

The commercial availability of ADAS and partially autonomous navigation systems is gradually increasing over time. Visual perception and understanding of the ego-vehicle surroundings, such as scene layout geometry, infrastructure, play a crucial role in downstream tasks such as path planning, motion forecasting etc. Modern infrastructure like high definition (HD) maps provides detailed information about the surrounding semantically and parametrically to become a critical component in the transportation infrastructure stack. HD maps released as part of the Argoverse dataset [13] provides information about lane level geometry as a vector map, driveable area and ground height in rasterized format. Nuscenes [11] HD maps provide lane geometry and status of infrastructure such as traffic lights. However, the HD maps are not ubiquitous; their production and maintenance costs are very high due to a wide range of expansive sensors used in data collection and the massive manual labour required to annotate this data. Automatic generation of HD maps and crowd-sourced maintenance methods are proposed

in [20], [63], [44]. These proposed methods only use either perspective view imagery or aerial imagery for the creation and maintenance of HD maps. As a result, they suffer from a lack of complementary understanding of both views.

Many techniques and control modules used in driver assistance systems rely on driver behavior, road infrastructure inspection, motion forecasting, and planning. Tasks such as regular road infrastructure audits and maintenance and traffic violation monitoring are crucial for safety. Road safety audit solutions range from manual to automatic. Manual techniques include field surveys and recorded video reviews. Automated solutions typically use computer vision-based detection techniques for road infrastructure audit [86], [105], [21], [7], [60], [66]. These methods work satisfactorily in well-conditioned and disciplined environments. However, they have difficulty producing good performance in unconstrained environments due to various challenges. On the other hand, these methods rely on large-scale sensor networks to assess road safety, which is expensive and difficult to maintain. Therefore, we need affordable and scalable solutions benchmarked extensively on unconstrained road scenes.

The thesis presents two solutions to provide an affordable alternative to HD maps. The first work, with the help of satellite and perspective view images, improves the understanding of complex road scenes parametrically to get the geometry of the road scene. The second work produces a framework for affordable large-scale road safety management in unconstrained settings to create a semantic layer detailing the quality of the infrastructure and road safety hazards. These works aim to provide effective and affordable alternatives to the expensive methods currently in use. The proposed methods are widely compared to related works to show their applicability and efficiency.

1.1 Scope of the Thesis

This thesis tackles the problem of visual perception on complex road scenes to improve safety and to understand the road scenes parametrically to help the downstream navigation tasks of the ego vehicle. To achieve the objectives mentioned above, we use perspective view imagery captured using front-facing camera mounted on the ego-vehicle, corresponding satellite imagery extracted using the GPS positioning.

1.1.1 Problem Definition

Looking Farther in Parametric Scene Parsing with Ground and Aerial Imagery:

Parametric models representing layout in terms of scene attributes are an attractive way to understand the road scene in autonomous navigation. Understanding a complex road scene parametrically in the Bird's eye view makes the downstream navigational tasks intuitive. However, the perspective view imagery recorded using the camera mounted on the ego-vehicle suffers from a narrow field of view of the camera, Occlusions and Foreshortening. Parametric understanding of BEV layout requires 3D understanding of the scene; expensive devices like LiDAR are used for this purpose. However, freely

available aerial images (satellite) overcome these problems with its complementary properties. We want to use aerial and ground images together to use their complementary properties to improve the parametric road scene understanding.

Evaluating Computer Vision Techniques for Urban Mobility safety on Large-Scale, Unconstrained Roads: Advanced Driver Assistance Systems (ADAS) rely on algorithms to characterize driver behavior, inspect road conditions, detect objects, predict motion, and identify abnormal events such as traffic violations. These tasks are becoming particularly difficult, and existing methods are unreliable in the unbridled road environment in developing countries. Infrastructure quality in unconstrained environments is uneven with insufficient lighting, signage, and damaged roads, making audit and maintenance critical to road safety. The current approaches rely on large camera networks. However, building such infrastructure on a large scale is costly and not scalable. Therefore, we need an affordable and scalable solution to detect road safety risks.

1.2 Contributions of this Work

This thesis addresses the above problems to propose novel solutions to parametric road scene understanding and scalable system to detect the dangers of road safety.

- We propose novel multi-modal multi-task architecture for parametric road scene understanding that leverages complementary properties of the ground and aerial imagery. This architecture provides new task-specific representations that give benefits in terms of field of view, occlusions, and evaluation further away from the ego vehicle.
- We also release a dataset with processed aerial imagery for the corresponding perspective view imagery for KITTI and Argoverse datasets and supplement them publicly with scene attribute annotations.
- We propose a novel setup and system to audit road infrastructure and traffic violations in a scalable and affordable manner. Also, we demonstrate the scalability and effectiveness of the vision algorithms on video covering 2000Km of the unconstrained road.
- We released those driving videos and related annotations providing the conditions in which data is captured in unconstrained road scenes.

1.3 Thesis Outline

In Chapter 2 provides an overview of the techniques used in this thesis. We explain and introduce various architectures of multi-modal deep networks, feature fusion, multi-task learning and object detection. The chapters (3, 4) provide a detailed information about our contributions to the two novel solutions.

Chapter 3 presents the problem of Parametric scene parsing with the ground and aerial imagery and extensively discusses relevant works in the literature on scene parsing using ground imagery, aerial imagery, LiDAR and Open Street Maps. We list both the advantages and disadvantages of the prior works to highlight the relevance of the proposed solution. Then, we provide information about the dataset, architecture used in this work and discuss the results.

Chapter 4 establishes the need for large-scale road safety audits and evaluation of computer vision techniques used for this purpose in a diverse set of conditions to understand their robustness and scalability. We also introduce a novel dataset of long videos in unconstrained settings. Then we discuss about the methods used to audit the infrastructure, traffic violations and extensively benchmark them on the data stretching across 2000KMs. Finally, we demonstrate a city scale semantic map of the road safety situation.

Chapter 5 concludes the discussion of this thesis by consolidating all the contributions made in our work.

Chapter 2

Background

2.1 Introduction

This chapter gives an overview of various components that form the backbone of all the experiments done throughout this work. First, we discuss deep multimodal networks and conventional feature fusion techniques. Second, we will discuss different deep neural network architectures for multitask learning. Then finally, details about Object detection and the Family of R-CNN and YOLO architecture.

2.2 Multi-Modal Deep Networks

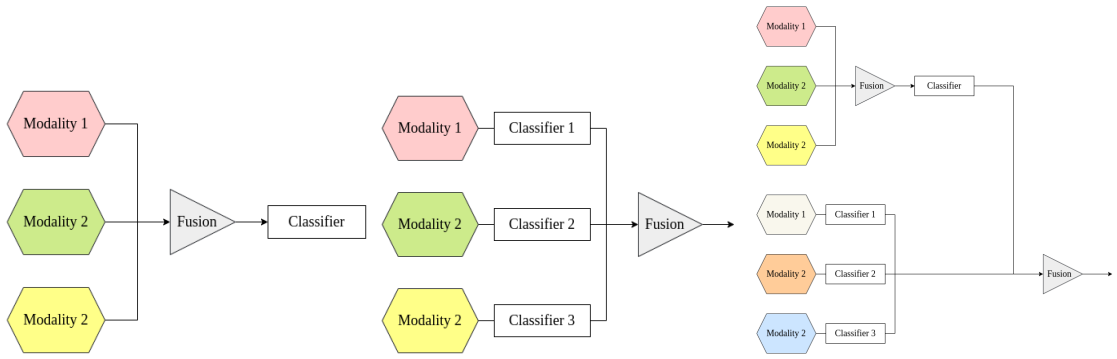
Modality in data refers to the form in which the information is stored. It is associated with the sensory perception of the world around vision, audio, and text. Even within a perceptual mode, different modalities of data can exist. For example, visual information about a scene stored in the image captured on the ground and satellite can be considered different modalities, considering the additional details captured in each setting. Human perception and understanding of the world are inherently multimodal. The use of multiple modalities of data can improve the performance of neural networks on the task at hand and enables a wide range of applications.

Learning an adequate representation of data to extract useful information for the prediction task is crucial in deep learning. Good representations are expressive, meaning that a reasonably sized learned model can capture many input configurations [5]. The representation of multiple modalities is even more complicated with challenges such as combining heterogeneous sources, eliminating noise, etc. Multi-modal learning is categorized as joint and coordinated [4]. The modality-specific embeddings are obtained from respective neural networks in a joint representation learning method. Then these modality-specific embeddings are fused to project into a common embedding space. Linear combination or concatenation of the embeddings are simple examples of feature fusion, more details in section 2.2.1. In coordinated representation learning, the embeddings of each modality are enforced to obtain desirable properties, like the higher similarity between them, etc.

2.2.1 Feature Fusion

The integration of features from multiple modalities to predict a discrete or continuous variable is called multimodal feature fusion. Feature fusion can increase the reliability of predictions by allowing complementary properties to be captured from multiple modalities. Conventionally, the fusion of features is classified as

- **Early fusion:** The features extracted from each modality are fused together before prediction.
- **Late fusion:** It is also known as decision fusion, as the predictions from each modality are fused together using maximum voting [57] and low rank fusion [50].
- **Hybrid fusion:** In this method, the features are fused with a combination of early and late fusion.



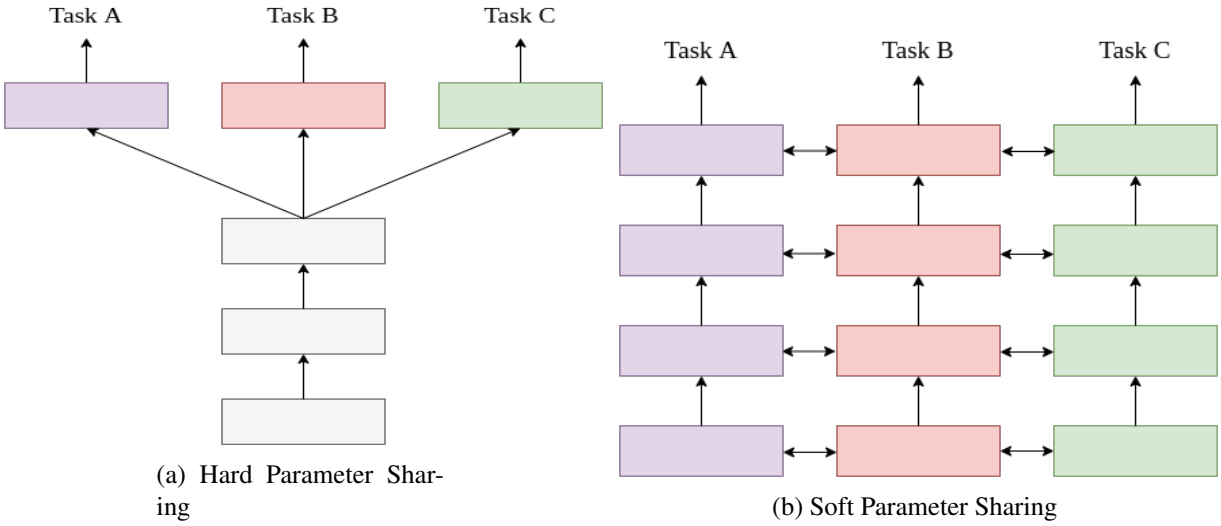
(a) Early Fusion (b) Late Fusion (c) Hybrid Fusion

2.3 Multi-Task learning with Deep Neural Networks

The learning paradigm in which models are trained on the same data, but simultaneously for multiple related tasks is popularly known as multi-task learning. Multi-task learning gives a faster learning speed without large-scale data for each task and helps reduce the computational demand for model training. However, training multi-task deep neural networks to produce desired results is not trivial. Especially negative transmission or destructive interference is a concern about improving task performance, resulting in reduced performance on other tasks due to conflicting needs. Mitigation of destructive interference and ablation of positive transfer in multi-task learning is an active research area.

Ruder [74] categorized multi-task learning models, namely hard parameter sharing and soft parameter sharing. The architectures that share parameters across all the tasks and are trained jointly to reduce loss due each task is hard parameter sharing networks. Whereas in soft-parameter sharing networks, the weights are specific to each task, but the output features at various levels combined or the loss function is regularized with the difference between the model parameters.

Different variations of hard and soft parameter sharing multi-task architectures are proposed for computer vision tasks. Liu et al. [49] proposed a multi-task attention network in which task-specific



(a) Hard Parameter Sharing

(b) Soft Parameter Sharing

Figure 2.2: Multi task networks can be broadly classified into two categories a) Hard Parameter Sharing b) Soft Parameter Sharing.

features are extracted using task-specific attention blocks on global shared features for various vision-based autonomous navigation tasks. The cross-stitch unit combines activations from multiple networks to learn the optimal combination of task-specific features and shared features is proposed by Misra et al [56]. Xu et al [100] proposed a network that first predicts a set of intermediate auxiliary tasks ranging from low level to high level. Then those predictions are distilled to predict the final tasks of depth estimation and scene parsing. Task routing layer, which uses a fixed task specific binary mask on convolutional layer outputs to effectively assign a subnetwork to a task, is proposed by Strezoski et. al. [82].

2.4 Object Detection

Object detection is a well known computer vision problem of detecting instance of semantic object of different classes in images. Object detection has a wide range of applications from autonomous navigation to healthcare. Many popular datasets like KITTI [28], PASCAL VOC [23] help immensely in improving the performance of object detection algorithms.

Object detection networks can be divided into single-stage and two-stage detectors. YOLO [69], R-CNN [30] are popular single and two stage detectors respectively.

2.4.1 R-CNN family

- **R-CNN [30]:** Region proposal module, feature extraction of proposed regions using convolutional neural network, linear SVM (Support Vector Machines) based object classifier, and bounding box regressor.
- **Fast R-CNN [31]:** Fast R-CNN extracts features of the entire image at once and passes the proposed regions to the pooling layer to obtain fixed-size representations. These fixed size representations are used for classification and bounding box regression using fully connected networks. As the feature representation of the entire image happens at once, unlike the multiple iterations in R-CNN. This improves the inference speed of Fast R-CNN highly.
- **Faster R-CNN [72]:** Faster R-CNN replaces the selective search with a fully convolutional region proposal network to predict objects of different scales and aspect ratios.

2.4.2 YOLO family

YOLO [69] (you look only once) runs at real-time speed for object detection at 45 FPS on Titan X GPU. The object detection problem is framed as a regression problem in YOLO, allowing it to directly class probabilities and bounding box values directly for each object.

YOLO divides the image into L $\times$ L grid. An object on image belongs to a grid when its center falls into that grid. Each grid predicts p bounding boxes and n probability scores for n classes. It uses a convolutional neural network pre-trained on ImageNet dataset for feature representation.

YOLOv2 [70] adopts a subset of design principles of its earlier version. It introduces Batch normalization in feature extraction network, introduces high-resolution input images for classification, predicts the size and aspect ratio of anchor boxes using K-means clustering and multi scale training. YOLOv3 [71] improves upon YOLOv2 by introducing multi-scale feature maps and objectness score at the output prediction along with bounding box, class probabilities.

Chapter 3

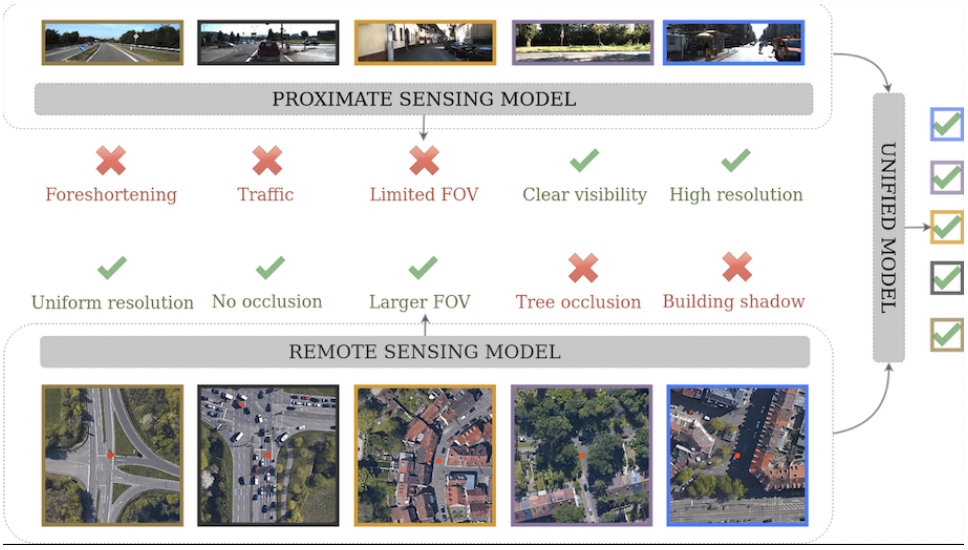
Looking Farther in Parametric Scene Parsing with Ground and Aerial Imagery

3.1 Introduction

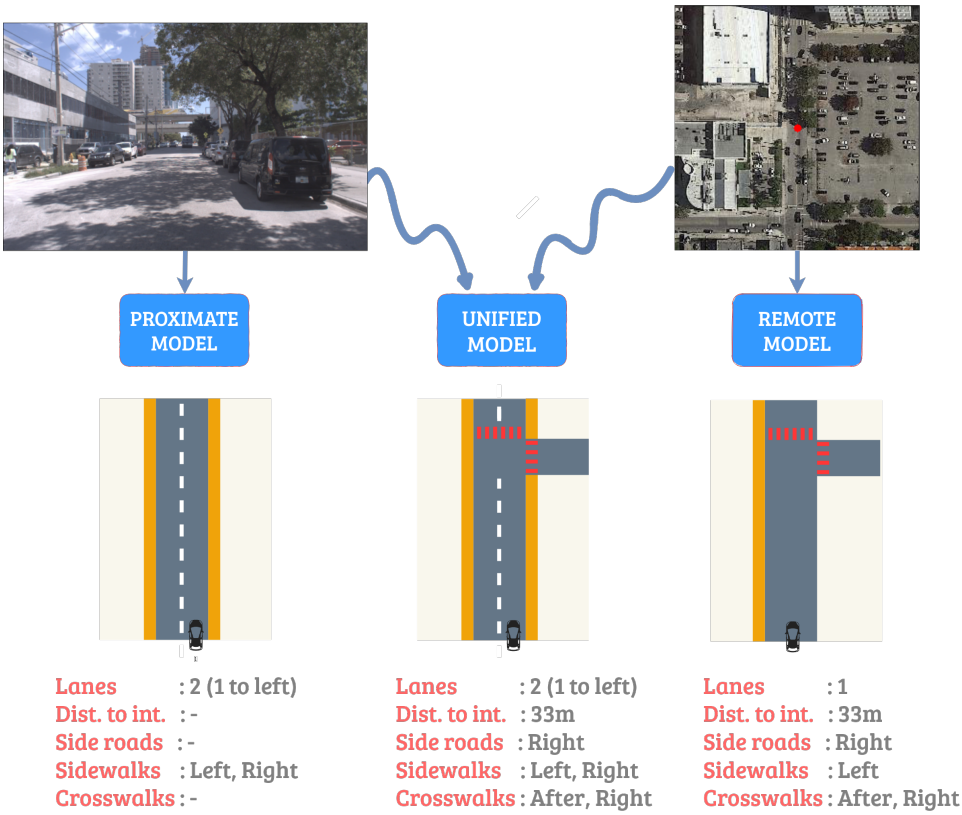
Understanding complex road scenes plays a crucial role in autonomous navigation and is an active area of research. Semantic understanding of the world can be obtained with non-parametric methods like semantic segmentation [14, 22, 79, 73] and depth estimation [32, 42, 101]. However, despite precise semantics, non-parametric outputs do not correspond to typical human interpretations associated with driving. Thus, it might not be intuitive to use for downstream navigational reasoning or decision making tasks.

In contrast to the above approaches, Wang et al [94] and Liu et al [48] propose a rich parametric model to represent the 3D scene layout from monocular ground imagery, which facilitates high-level reasoning. However, the model faces challenges in representing prominent aspects in the scene layout outside the field of view of the camera and in estimating distance, and semantics due to perspective foreshortening. Refinement, extraction, and understanding of global road topology [14, 3, 53, 34, 45] using aerial imagery is advantageous due to its larger field of view, uniform resolution for distance and semantic estimation in near and far-fields (due to a nearly orthographic projection). However, it cannot observe all local properties of the road topology due to occlusions in aerial imagery arising from vegetation and infrastructure. Given the complementary properties of the ground and aerial data, we propose that benefits may be available through their combination. However, this is a challenging problem for scene understanding, since both local details and global information are manifested very differently in perspective and aerial imagery.

We propose to use both, aerial and ground modalities, for the task of parametric representation of road scenes to exploit their complementary properties as illustrated in Figure 3.1. Several prior works exploit such complementary understanding from multiple data modalities such as ground and aerial imagery [54, 104, 96, 97, 52, 18, 91, 19] , ground imagery and LIDAR data from perspective view [92, 46], ground imagery, and Open Street Maps (OSM) [92], aerial imagery and OSM [92], as well



(a)



(b)

Figure 3.1: (a) While ground imagery provides strong cues for local properties of the road topology, they are limited by the narrow field of view of the camera, occlusions, and perspective foreshortening. In contrast, aerial imagery has the advantage of a larger field of view, presents a uniform resolution in both the near, and far fields and is free from severe occlusions due to traffic. (b) This paper derives a parametric representation of scene geometry and semantics in unconstrained traffic scenes. We leverage aerial imagery to look ahead and complement the local visual cues from ground imagery. Note that the bright red spot on the aerial image corresponds to the position of the ego vehicle.

as LIDAR and OSM [92, 83, 26]. We differ from all those works in using aerial imagery to obtain global context and ground imagery for strong visual cues of local measurements to derive a parametric representation of scene geometry and semantics in unconstrained traffic scenes. We propose a new scene model, network architecture, and multi-task training strategy to fully realize the complementary benefits of ground and aerial imagery.

The parameters used to define our scene layouts are existence and distance to the intersection, type of intersection, number of lanes on both sides, number of lanes on the opposite side etc. We use the annotations released by [93] on the KITTI dataset to compare the results with previous works. However, This dataset is limited to describing scenes that are only up to 30 meters from the ego-vehicle. Thus, we supplement the existing Argoverse dataset [13] with scene attributes annotations that can describe road scenes that far from the ego-vehicle. It is important to note that manual annotation for these parameters is very subjective, needs complex reasoning of the scene's geometry and expensive to collect at large scale. We thus leverage the publicly available Argoverse HD map [13] and automatically extract attributes from it. We then show the increasing importance of aerial imagery in predicting scene attributes that rely on distant visual cues.

To summarize, in this paper, we make the following contributions:

- A novel approach to parametric road scene understanding that leverages complementary properties of the ground and aerial imagery.
- Novel representations that yield advantages for the field of view, occlusions, and estimation farther from the ego-vehicle.
- A dataset with processed aerial imagery and scene attribute annotations supplement the publicly available Argoverse dataset.

3.2 Related Work

3.2.1 Parametric scene understanding

Parametric scene understanding is the task of approximating a road scene with a set of parameters, that are often human interpretable and thus intuitive to use for downstream navigational reasoning or decision-making tasks. Ess et. al. [22] propose a method to distinguish between different road layouts and also detect the presence of cars and pedestrians, while Mattyus et al. [53] estimate the width of OSM roads by utilizing aerial imagery. These methods only provide a crude understanding of the road scene. Geiger et al. [27] reason about the scene topology, geometry, and traffic activities from hand-crafted features and evaluate on a limited dataset of 113 images. Mattyus et al. [54] jointly infer the location and width of roads, cycling pavements, parking areas, and sidewalks using aerial and stereo ground imagery. They use hand-crafted features which model only straight roads and do not work on scenes with intersections. In contrast, our work can model complex road scenes, including intersections.

We find [94, 48, 78] to be the closest works to ours. Seff et al. [78] use CNN to automatically infer the scene attributes from a monocular RGB image. Wang et al. [94] estimate the semantic layout of ground imagery in Birds Eye View (BEV) and then extract scene attributes. Liu et al. [48] propose to use videos to benefit from cues of camera motion and long-term context. They make use of a Feature Transform Module to fuse features from nearby frames and a COLMAP [75, 76] based scene reconstruction of the whole video sequence to provide global information. However, all these models still suffer from issues of occlusion and perspective foreshortening inherently present in ground imagery, which we overcome by providing imagery from aerial modality. Additionally, we can provide surrounding scene context to the model by simply enlarging the field of view in aerial imagery.

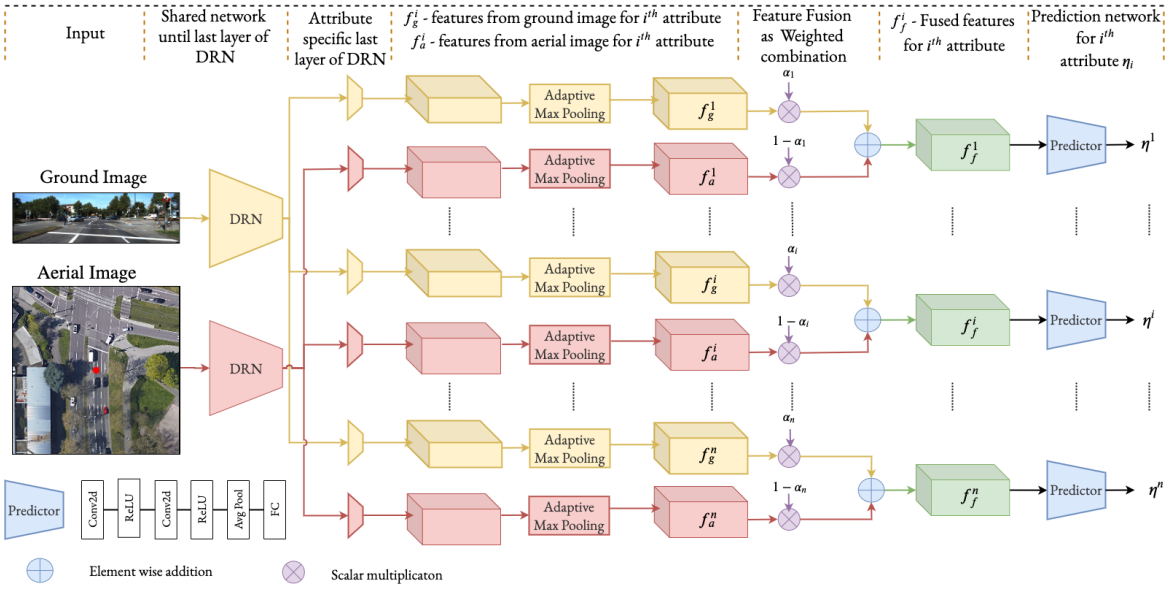


Figure 3.2: We employ two Dilated Residual Networks [103] (DRN), till the penultimate layer, to extract generic features from ground and aerial imagery. The ultimate layer of the DRN is then used to extract targeted features for each attribute, individually. The attribute-specific aerial and ground features are then passed through an adaptive max-pooling layer and fused using learned weights (α_i). The fused features are passed to a convolution neural network to predict binary, continuous, and multi-class attributes ($\eta_1 - \eta_n$), where n is the number of attributes. Note that the bright red spot in the aerial image corresponds to the position of the ego-vehicle.

3.2.2 Aerial-ground reasoning

Hu et al. [35] address the task of geo-localizing a ground-view image on an aerial image. Li et al. [43] propose a method to adapt to ground imagery in unseen regions with the help of aerial imagery. While these methods utilize both modalities of data, they focus on obtaining similar features from both views. In contrast, we leverage the complementary properties of the modalities to obtain more robust representations. Wegner et al. [95] detect trees by jointly reasoning from aerial and ground

imagery. They use per-view detectors to obtain detection proposals from each image separately. The proposals from each view are then combined to generate the final proposals. While they combine the two modalities at the output level, we combine the aerial and ground imagery at the feature-level. This provides the model with more vital cues to learn and predict from. Feng et al. [24] propose a method for urban zoning by using the semantic information from aerial imagery and the classification outputs from ground imagery. The extracted features are then passed to an MRF for inference. However, they use hand-crafted features, and their method is not trainable end-to-end. Workman et al. [97] propose an end-to-end trainable network for estimating geospatial functions such as population density, land cover, and land use. They use kernel regression and density estimation to convert features from ground-level images into a dense feature map and combine them with the aerial imagery features.

Manderson et. al [52] learn a navigation policy for off-road driving leveraging complementary inputs of ground and aerial imagery. In comparison, we learn novel representations that provide an accurate representation of scene geometry and semantics in unconstrained traffic scenes. Our architecture yields advantages with respect to the field of view, occlusions and estimation farther from the ego-vehicle.

3.3 Parametric scene parsing with ground and aerial imagery

Scene Model We can model a complex road scene using a set of binary, continuous and multi-class attributes. Binary attributes indicate the presence or absence of road components such as neighboring lanes, intersections, and side roads. Multi-class and continuous attributes provide a detailed understanding of these components by quantifying them, for example, the number of neighboring lanes and the distance to the intersection. In total, we have 14 binary (Θ_B), 2 multi-class (Θ_M) and 8 continuous (Θ_C) attributes for KITTI, and 8 binary, 2 multi-class and 2 continuous attributes for Argoverse. For more details, refer to [93] for KITTI, and Section 4 for Argoverse.

3.3.1 Architecture

As discussed earlier, each driving scene is described by a rich set of attributes. Since each attribute describes a different road component, naturally, the visual cues attended to in the image vary significantly. With the aerial and ground modalities providing complementary properties, it is important to fuse features efficiently to leverage this advantage. Additionally, the amount of context available from a particular modality differs for different attributes. Finally, due to the large number of attributes being learned, there is an inherent trade-off between increasing model capacity for learning discriminative features for each attribute and the resultant model size. These reasons make the architecture design for parametric scene understanding from complementary modalities challenging, which we seek to address below.

3.3.2 Attribute specific feature extraction

We use DRN as our backbone architecture for feature extraction. The choice of the feature extractor is not a focus of our work and can be replaced by other networks like ResNet [33]. Seff et al. [78] predict scene attributes by training a separate network for each attribute, thereby wasting computation by not learning shared features. On the other extreme, Wang et al. [94] learn and predict scene attributes from a single shared branch, thereby restricting the model from learning more discriminative features per attribute. We strike a balance by learning a shared DRN from each modality until the last layer, from where we branch. This allows the model to learn attribute-specific features from each modality with minimal impact on model size.

3.3.3 Multimodal fusion for leveraging complementary properties

Efficient fusion of features from the ground and aerial modalities is important to leverage their complementary properties since the importance of a modality varies for each attribute. For example, nearby lane information is more prominent in ground modality due to the high resolution of ground imagery, while information on side roads is clearer from aerial imagery due to the uniform resolution of the modality. Thus we fuse the features through a weighted sum given by the equation $f_f^i = \alpha_i f_g^i + (1 - \alpha_i) f_a^i$, where f_g^i, f_a^i, f_f^i are ground, aerial, and fused features of the i^{th} attribute respectively and α_i is a learnable parameter to fuse the ground and aerial features of the i^{th} attribute. Vielzeuf et al. [89] propose a multi-layer fusion approach and claim better results than the above mentioned fusion technique. However, while extending their technique to a feature extractor shared by multiple attributes, the influence of one modality on the other at the earlier layers leads to bias, resulting in the network performing well only on specific attributes. Thus, we extract features independently from each modality and then finally fuse features through a weighted sum.

3.3.4 Multi-attribute prediction

The fused features for each attribute are passed to a prediction network separately. The prediction network consists of two convolutional layers, a global average pooling layer, and a fully connected layer in sequence.

3.3.5 Loss function

We use weighted cross entropy and least squared error as our loss functions. As the losses of binary, continuous and multi-class are of different scales, we use multipliers to each of these losses, denoted by

γ_B, γ_C and γ_M respectively.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \gamma_B \,\mathrm{BCE}(\Theta_{B,i}, \eta_{B,i}) + \gamma_M \,\mathrm{CE}(\Theta_{M,i}, \eta_{M,i}) + \gamma_C \text{Ł2} \tag{3.1}$$

where BCE is Binary Cross Entropy, CE is Cross Entropy and $\{\Theta, \eta\}_{.,i}$ denotes the i^{th} sample in the dataset of length N and corresponding scene attributes.

$\eta_B = \{\eta^1, \eta^2, \ldots, \eta^b\}$, $\eta_M = \{\eta^{b+1}, \eta^{b+2}, \ldots, \eta^{b+m}\}$, $\eta_C = \{\eta^{b+m+1}, \eta^{b+m+2}, \ldots, \eta^{b+m+c}\}$, $n = b + m + c$

where b, m, c, n are the number of binary, multi-class, continuous and total attributes respectively in the scene model.

3.4 Experiments and Results

3.4.1 Datasets

We use two datasets, Argoverse and KITTI for ground imagery in perspective view. Wang et al. [93] provide scene attribute annotations for the KITTI dataset. Since the released annotations on the KITTI dataset is limited to road scenes up to 30m, we supplement the publicly available Argoverse dataset with scene attribute annotations for road scenes up to 60m. Since aerial imagery for KITTI and Argoverse are unavailable, we collect and process aerial imagery for both these datasets. For each of the above datasets, we acquire aerial imagery from Google Maps at zoom level 21. The Ground Sampling Distance (GSD) is 30cm for KITTI and 15cm for Argoverse. The aerial imagery is rotated such that the direction of the ego-vehicle always points towards the north of the aerial imagery. The scene attribute annotations on the Argoverse dataset, along with the processed aerial imagery for all datasets are released publicly. We further refer to these extended datasets on KITTI and Argoverse as KITTI-Air-PSU and Argo-Air-PSU, respectively.

3.4.1.1 KITTI-Air-PSU

We only use the left-front RGB images from the stereo camera in the KITTI dataset. Wang et al. [93] annotate each image with 14 binary, 2 multi-class and 8 continuous attributes. The attributes describe mainroad, intersections, sideroads, crosswalks, and sidewalks. There are in total, 16273 training and 2108 validation images. Please refer to [93] for detailed information.

3.4.1.2 Argo-Air-PSU

The KITTI-Air-PSU dataset is limited to describing scenes that are only up to 30 meters from the ego-vehicle. In an another work, Seff and Xiao [78] release annotations for 1 million Google Street

Table 3.1: Description of attributes of the Argo-Air-PSU scene model. B: Binary, M: Multi-class, C: Continuous.

ID	Description
B1	Are there lanes to the left of the ego-vehicle?
B2	Are there lanes to the right of the ego-vehicle?
B3	Is the main road one-way?
B4	Is there a side road to the left at the next intersection?
B5	Is there a side road to the right at the next intersection?
B6	Is the ego-vehicle at an intersection?
B7	Does the main road continue after the next intersection?
B8	Is the main road curved?
M1	Number of lanes to the left of the ego-vehicle?
M2	Number of lanes to the right of the ego-vehicle?
C1	Distance to the next intersection
C2	Radius of curvature

View (GSV) panoramas by automatically extracting attributes from the crowdsourced Open Street Maps (OSM). While this dataset is huge, there are several drawbacks associated with it. While there is a severe misalignment between GSV imagery and the roads in OSM, the annotations also are incomplete and error-prone due to the weak vetting process. In contrast, Argoverse provides HD maps from which we automatically query and extract accurate scene attribute annotations. We use the center-front images in the Argoverse tracking dataset as the ground imagery and create two different versions of the dataset. The first version covers scene attributes up to 30 meters in front of the camera, which we further call Argo-Air-PSU-30. The second version covers scene attributes up to 60 meters in front of the camera, which we further call Argo-Air-PSU-60. Since the Argoverse dataset only contains information pertaining to roads, we annotate the Argo-Air-PSU dataset on mainroads, intersections and sideroads. In total, we obtain annotations for 13122 training and 5017 validation images with 8 binary, 2 multi-class and 2 continuous attributes as described in Table 3.1.

Table 3.2: Comparison of our Unified model with online methods on the KITTI, Argo-Air-PSU-30 and Argo-Air-PSU-60 validation sets. Across datasets and attribute types, the Unified model shows better performance than prior works [94, 78] and baselines that use only a single modality.

Method	KITTI [27]			Argo-Air-PSU-30			Argo-Air-PSU-60		
	Accu-Bi. ↑	Accu-Mc. ↑	nMSE ↓	Accu-Bi. ↑	Accu-Mc. ↑	nMSE ↓	Accu-Bi. ↑	Accu-Mc. ↑	nMSE ↓
M-RGB [78]	.811	.778	.230	-	-	-	-	-	-
M-BEV [94, 77]	.820	.777	.141	-	-	-	-	-	-
Proximate (ours)	.833	.795	.168	.885	.838	.089	.883	.842	.088
Remote (ours)	.817	.796	**.116**	.93	.792	.058	.893	.811	.062
Unified (ours)	**.848**	**.819**	.118	**.939**	**.896**	**.047**	**.904**	**.852**	**.052**

3.4.1.3 Evaluation Metrics

For binary and multi-class attributes, we report on binary accuracy (Accu-Bi.) and multi-class accuracy (Accu-Mc.) respectively. Since few attributes in the dataset are highly biased, we experiment with F1 scores. However, since a single misprediction on the minor class results in heavy penalization, we observe that accuracy better reflects the performance of the model. For continuous attributes, we report on the normalized MSE (nMSE) scores. We report on IOU as the renderer required for computation is imperfect and the IOU scores are influenced by the implementation of the renderer.

3.4.2 Implementation Details

We set the batch size to 12 and used Adam for optimization with a learning rate of 10^{-4}. We set multipliers for binary, continuous, and multi-class losses to $\gamma_B = 30$, $\gamma_C = 0.01$ and $\gamma_M = 30$, so that their respective losses lie in the same scale. We train every model on 4 NVIDIA 1080 GPUs for 10 epochs. The code and the models are released publicly.

3.4.3 Baselines

To the best of our knowledge, [94], [48], and [78] are the only works that perform parametric road scene understanding. We note that recent works like [73, 77] generate semantic maps or occupancy grids in BEV, however, the outputs are non-parametric and thus, not directly comparable to us.

M-RGB [78] The features are extracted from ground imagery using a shared ResNet-101 [33]. The scene attributes are inferred directly by passing features through a fully connected (FC) network.

M-BEV [94, 77] The model constructs the BEV of ground imagery using semantic and depth labels. Features are extracted from BEV using CNN and followed by FC network for predictions.

Liu et al. [48] The model takes a video sequence from ground modality, converts them to BEV and fuses features from different frames using a Feature Transform Module. Further, an expensive COLMAP based reconstruction applied to the entire video sequence in an offline manner is provided as an additional input to the network.

Proximate (ours) We use the Unified model proposed by us but only use features from the ground imagery.

Remote (ours) We use the Unified model proposed by us but only use features from aerial imagery.

3.4.4 Dataset Details

Argoverse provides HD maps as a vector representation consisting of nodes and ways. A way is a lane segment containing a series of nodes (polyline). Multiple ways are connected to form lanes. It is important to note that each node has its‘ own GPS coordinate. For each image, we obtain the coordinates of the ego-vehicle and match it with the closest lane segment based on distance. We refer to the lane segment containing the ego-vehicle as the ego-lane.

3.4.4.1 Number of lanes

Each lane segment contains information on its immediate left and right neighboring lane segments. Using this, we can then recursively find all lane segments on either side of the ego-lane. We calculate the number of lanes on either side of the ego-vehicle by counting the neighbors in each side. Our annotation for lanes on the left of ego-vehicle also includes lanes in the opposite direction.

3.4.4.2 Oneway

Each lane segment has a direction associated with it, which indicates the direction of vehicular movement in that lane segment. While computing the neighbors of the ego-lane, we also compare the direction of the neighboring lane segments with the ego-lane. If there are no lane segments opposite the ego-lane, we annotate the main road as one way.

3.4.4.3 Ego-vehicle is at an intersection?

Each lane segment has a flag associated with it, indicating whether it is part of an intersection or not. Based on the flag's value for the ego-lane, we annotate if the ego-vehicle is at an intersection or not. It is important to note that we do not extract annotations for other attributes if the ego-vehicle is at an intersection.

3.4.4.4 Distance to intersection

Each lane segment has information on its‘ immediate successors. Lane segments have a single successor except for those just before intersections, where they might have more than one successor. We recursively find the next successor from the ego-lane, until we reach an intersection. We annotate the distance to the intersection as the sum of lengths of all lane segments up to that intersection. Since each lane segment is a polyline, and we calculate the distance as the length along with the nodes of

the polyline. For the ego-lane, we only calculate the distance from the ego-vehicle to the end of the ego-lane.

3.4.4.5 Side roads and intersection geometry

As mentioned earlier, each lane segment has a single successor except for those just before intersections. These lane segments before intersections may contain one or more successors, depending on the geometry of the intersection. Lane segments also contain information on turn direction. The turn direction indicates the relative direction of the lane segment with respect to its predecessor. Combining the above two pieces of information, we can extract attributes on side roads and intersection geometry. Firstly, we find the successor of the ego-lane that is just before the intersection. Then, we find all successors of this lane segment along with its turn direction. The turn directions are used to identify if there is a left or right side road and also if the main road continues after the intersection.

3.4.4.6 Is the main road curved?

We first find the closest three nodes to the ego-vehicle in the ego-lane. Using the coordinates of these three nodes, we find the radius of the circle that contains all three nodes. If this radius is below a threshold, we annotate the road to be curved. The threshold is empirically chosen to be 1000. Since this radius of curvature is an inaccurate approximation, we do not use it as a continuous attribute for training.

3.4.4.7 Explanation for using accuracy over F1 score

Most of the attributes that we predict to describe the road layout are very biased. The extreme imbalances in the dataset have resulted in F1-score being very sensitive just to one misprediction or correct prediction. Due to this, F1-score did not reflect the actual performance of the model. The closest work to ours is by Wang et al., [94] claimed the same observation as ours and used accuracy as a metric.

3.4.5 Results

3.4.5.1 Comparison on KITTI dataset (online methods)

In comparison to prior works, Table 3.2 shows our Unified model achieving a significant improvement in binary accuracy (.820 to .848), multi-class accuracy (.777 to .819), and on nMSE scores (.141 to .118) for continuous attributes. Comparing Proximate with M-RGB, where the only difference is the architecture being used, we can clearly observe the impact of our design choices. Despite M-BEV using additional semantic and depth information, our Remote model still performs better on multi-class and continuous attributes, while there is a minor performance drop on binary attributes. We now compare the performance of our Unified model against our Proximate and Remote models. While the Unified

Table 3.3: Normalizing constants used for evaluation of continuous attributes on all datasets are provided. Abbreviations used in the table are as follows:
mr: main road; sw: sidewalk; sr: side road; delim: delimiter; cw: crosswalk; dist.: distance; int.: intersection.

(a) KITTI-Air-PSU

Attributes	Constants
mr rotation	25.0
mr curvature	1000
left sr width	51.8
right sr width	51.8
mr delim width	15.0
sw delim width	8.0
dist. to right sr	31.85
dist. to left sr	31.85
dist. to cw on mr	12.0

(b) Argo-Air-PSU-30

Attributes	Constants
dist. to int.	30.0

(c) Argo-Air-PSU-60

Attributes	Constants
dist. to int.	60.0

Table 3.4: Comparison of our Unified model with offline methods on the KITTI validation set.

Method	KITTI [27]		
	Accu-Bi. ↑	Accu-Mc. ↑	nMSE ↓
Liu et al. [48]	.842	**.841**	.134
Unified (ours)	**.848**	.819	**.118**

model shows similar performance to Remote on nMSE scores, there is a marked increase in binary and multi-class accuracy over both Remote and Proximate models. Overall, the performance of the Unified model clearly shows the advantage of using both aerial and ground modalities for the parametric scene understanding.

3.4.5.2 Comparison on Argo-Air-PSU-30 and Argo-Air-PSU-60 datasets (online methods)

As shown in Table 3.2, the results indicate that the Unified model shows better performance than Remote and Proximate models on binary, multi-class, and continuous attributes. Importantly, they demonstrate the ability of aerial imagery in looking ahead at scenes that are farther away from the ego-vehicle. Looking at the continuous and binary attributes, the performance of the Remote model is superior to Proximate model in both the Argo-Air-PSU-30 and Argo-Air-PSU-60. Most of the binary attributes and

Table 3.5: The table shows the results of various ablation studies performed with the Unified model on the Argo-Air-PSU-30 dataset. Each row block corresponds to an experiment set. The results are to be compared within the block and with the final row, corresponding to our final Unified model. GAP: Global Average Pooling, AMP: Adaptive Max Pooling, Uni. sum: Uniform Sum, Wt. sum: Weighted sum, Pos: Position

Car Pos.	DRN Branch	Pooling	Fusion	Prediction N/W	Argo-Air-PSU-30		
					Acc-Bi ↑	Acc-Mc ↑	nMSE ↓
Middle	No	GAP	Concat.	Shared	.922	.786	.170
Middle	No	GAP	Concat.	Individual	.924	.816	.090
Middle	Yes	AMP	Concat.	Individual	.934	.805	.070
Middle	Yes	AMP	Uni. sum	Individual	.936	.816	.085
Middle	Yes	GAP	Wt. sum	Individual	.925	.863	.099
Middle	No	AMP	Wt. sum	Individual	.935	.852	.066
Bottom	Yes	AMP	Wt. sum	Individual	.923	.821	.080
Middle	Yes	AMP	Wt. sum	Individual	**.939**	**.896**	**.047**

all the continuous attributes correspond to global properties of the road topology. Thus, we can infer that addition of aerial modality improves the performance in predicting global properties, such as details of road intersection and side roads. Similarly, we can observe that the performance of the Proximate model is better than that of the Remote model on multi-class attributes i.e. lanes to the right of ego lane and lanes to the left of ego lane (local properties). The aerial imagery also aids in improving the performance of the Unified model on multi-class attributes in case of occlusions in ground imagery.

3.4.5.3 Comparison on KITTI dataset (offline methods)

From Table 3.4, we observe that though Liu et al. [48] utilize the complete video sequence for predicting scene attributes at a particular timestep, our Unified model still performs better on binary and continuous attributes, while observing imagery only from that current timestep. However, we perform worse on multi-class attributes constituting less than 10% of the total attributes. We note that [48] uses additional cues from scene reconstruction and vehicle localization, while we extract novel representations from aerial and ground modality without additional context and fuse them efficiently.

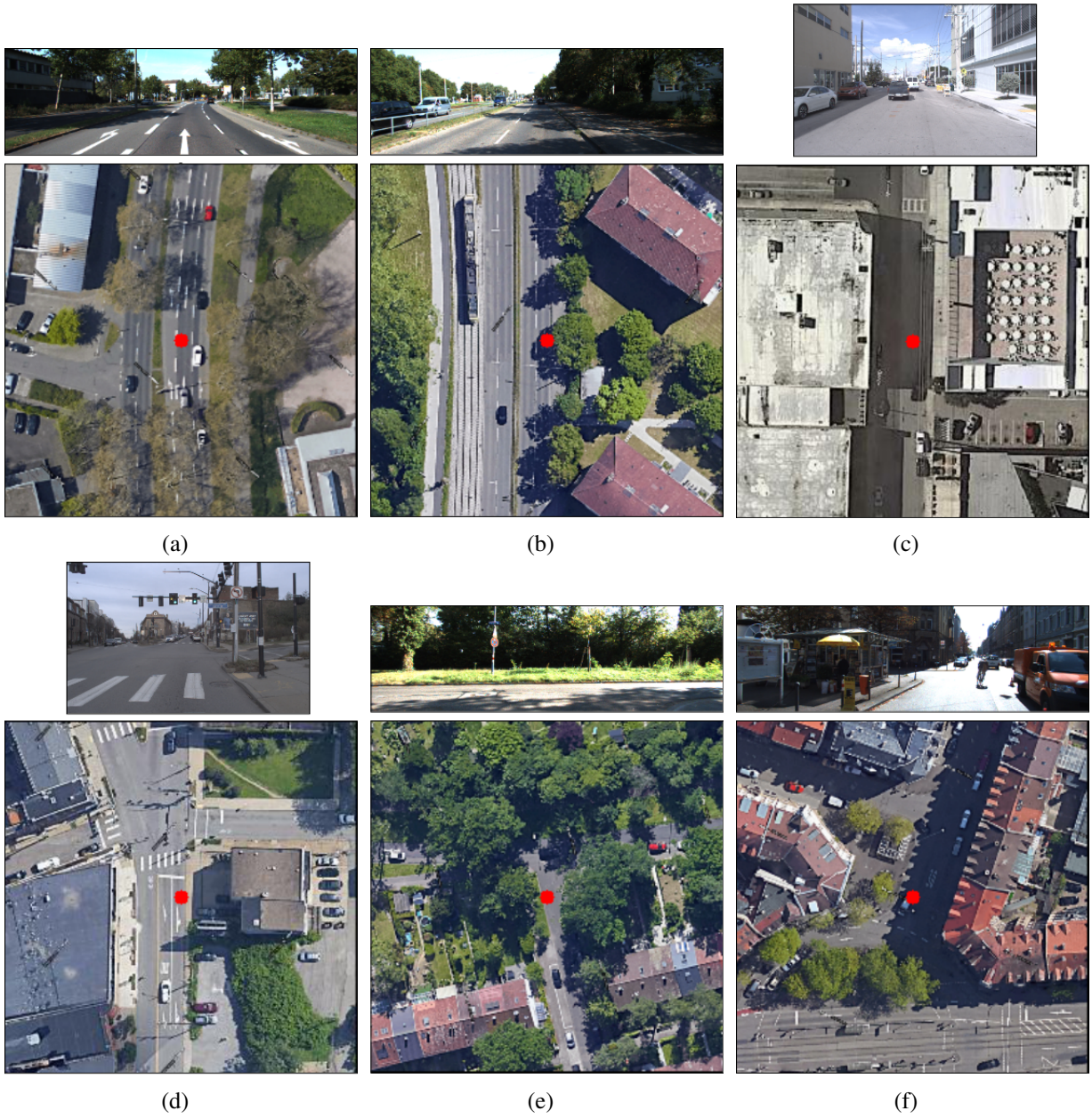
3.4.5.4 Ablation Experiments

To investigate the design choices of our Unified model, we conducted several ablation studies as shown in Table 3.5. Firstly, we observe that having an individual prediction network for each attribute is desirable, as the model can learn more discriminative features for prediction. Secondly, we look at different techniques for multimodal feature fusion. By learning optimal weightage for the two modalities

for each attribute individually, the model using weighted sum is able to best exploit the complementary properties of the two modalities. Thirdly, we look at the pooling techniques and observe that Adaptive Max Pooling performs significantly better than Global Average Pooling since it is able to retain the spatial context of features. Fourthly, we observe that even by branching only at the final layer of the DRN leads to significant improvement, validating the importance of having attribute-specific features before fusion. Finally, by placing the car at the middle of the aerial imagery, we are able to efficiently incorporate prior context behind the ego-vehicle, thereby resulting in significant performance gains.

3.4.5.5 Qualitative Results

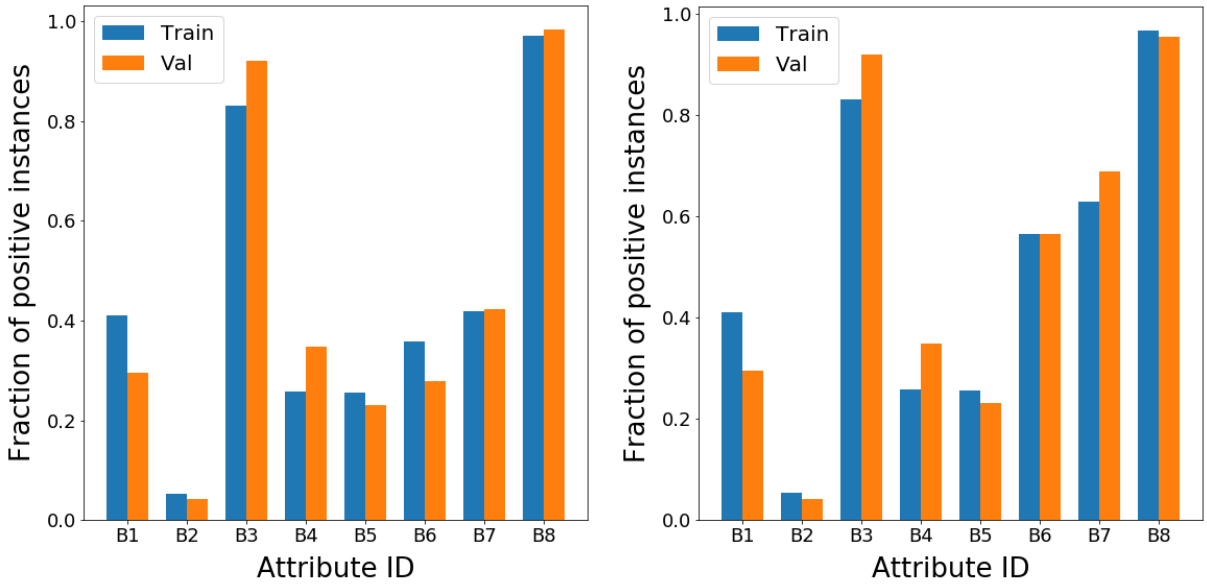
In Figure 3.3, we illustrate a few examples where our Unified model is able to overcome the individual shortcomings of aerial and ground imagery.



(a) (b) (c)

(d) (e) (f)

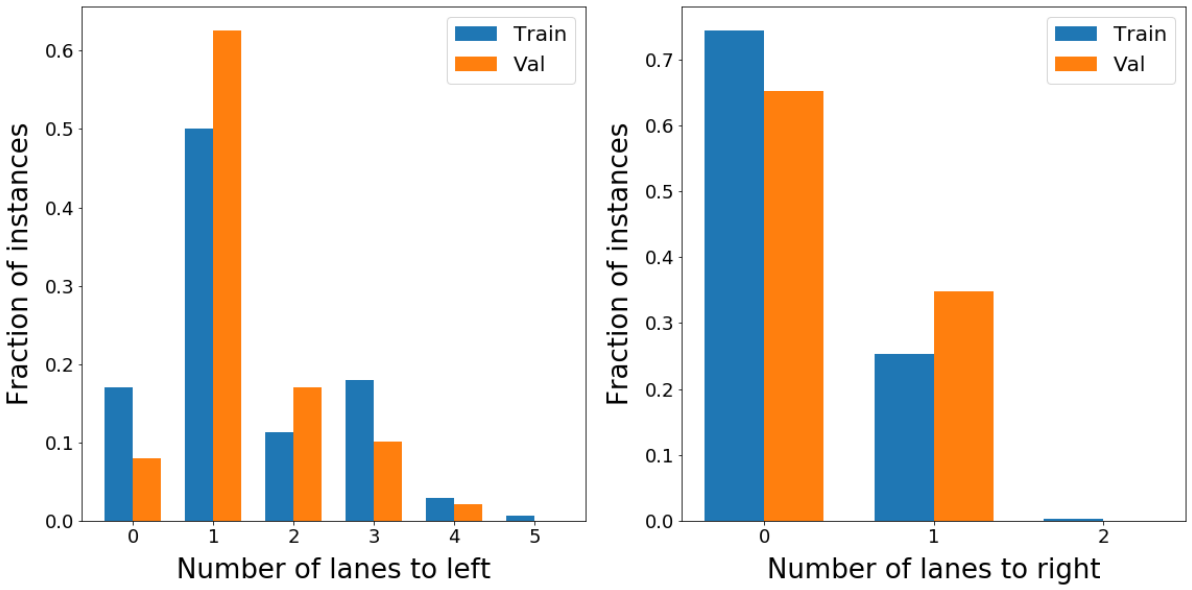
Figure 3.3: (a) Curvature of the road is visible in aerial imagery, (b) Right sidewalk is occluded by trees in aerial imagery, (c) Left sideroad is invisible in ground imagery due to occlusion by building, (d) Left sideroad is not visible in ground imagery due to limited field of view (e) Right sideroad is occluded by trees in aerial imagery (f) Aerial imagery incorrectly predicts as one-way due to building shadow. The above examples demonstrate the advantage of using both aerial and ground imagery. For all examples, the Unified model gives correct predictions. The (a) and (b) are examples where the Remote model predicts correctly while the Proximate model gives incorrect predictions. The (c) and (d) are examples where the proximate model predicts correctly, while the Remote model gives incorrect predictions. The reason for incorrect predictions are mentioned in the individual captions. Example (b) is taken from the Argo-Air-PSU-30 dataset while the rest are taken from KITTI-Air-PSU-30 dataset. Note that the bright red spot on the aerial image corresponds to the position of the ego-vehicle.



(a) Argo-Air-PSU-30

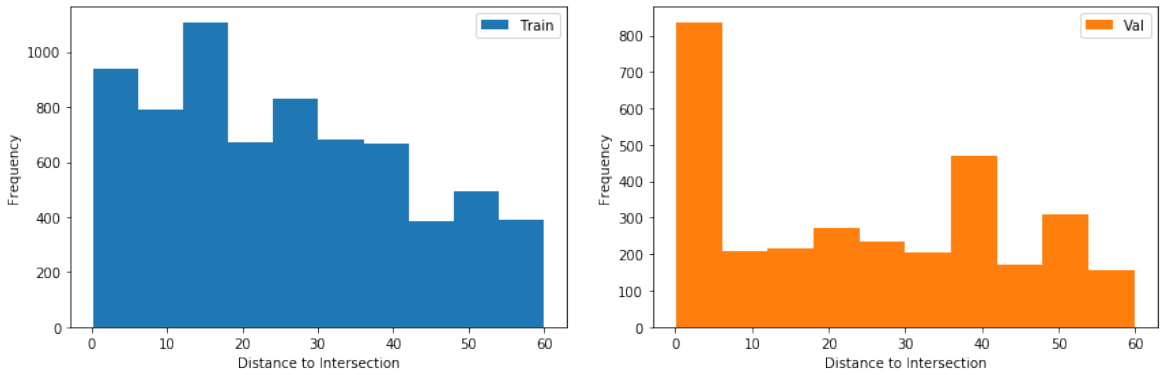
(b) Argo-Air-PSU-60

Figure 3.4: The distribution of binary attributes on either versions of the Argo-Air-PSU dataset is shown. We can see a change in distribution between versions only for attributes B6, B7 and B8. The distributions are similar for both validation and training sets. We can also observe there is an extremely high bias in the dataset for attributes B2, B3 and B8.



(a) (b)

Figure 3.5: The distribution of multi-class attributes on the Argo-Air-PSU dataset is shown. The distribution remains the same across both versions of the dataset. Since roads in Argoverse are right-side driving (USA), lanes to the left also include lanes in the opposite direction. The distributions are similar for both validation and training sets. We can observe that the number of instances of the last few classes is extremely low.



(a) Training (b) Validation

Figure 3.6: The distribution of distance to intersection on the Argo-Air-PSU-60 dataset is shown.

Chapter 4

Evaluating Computer Vision Techniques for Urban Mobility on Large-Scale, Unconstrained Roads

4.1 Introduction

The ever-increasing urban population on roads has resulted in an increase in the number of accidents and deaths [85], many of which could have been avoided with the effective use of technology. The escalating traffic also poses technical challenges in the urban mobility-related tasks to (i) driver assistance (ii) reliable audit and maintenance of the infrastructure, and (iii) detection and prevention of traffic violations. This is especially true for unstructured driving situations and unconstrained roads, which are common in developing countries.

Performance of driver assistance systems and automated control modules often rely heavily on the algorithms for (i) characterizing driven attention and behavior (ii) inspecting road conditions and infrastructures and (iii) detecting obstacles, predicting movement patterns, and evaluating motion trajectories. Aggregating the environmental information perceived through a camera is essential to the Advanced Driver Assistance Systems (ADAS). This becomes challenging in unconstrained road situations. Popular ADAS systems, even though they work satisfactorily in well-conditioned environments, struggle to identify the irregularities in roads and road objects in unconstrained road situations popularly seen in large parts of the world.

Unconstrained roads suffer from uneven quality of infrastructures, such as low lighting, inadequate signage, and damaged roads, making road safety a challenging problem [38]. Therefore, tasks such as regular audit and maintenance of road infrastructure and traffic violation monitoring become critical for safety. Current approaches rely heavily on large camera networks. This is not scalable or economical for massive lengths of roads in the ever-expanding urban settlements [51, 65].

However, current vision-based unstructured road detection algorithms are usually by continuously changing backgrounds, different road types (shape, color), variable lighting conditions, and weather conditions. Therefore, we need a highly automated, affordable, and scalable solution.

Figure 4.1: Irregularities in chaotic streets. Top to bottom: i) Sample street lights & traffic signs, ii) Road & Infrastructure defects, iii) Streets without street lights, markers (sometimes faded) & traffic signs, iv) Violations at varying orientations.

Figure 4.1 gives a glimpse of the potential problems seen in an unstructured situation. This includes (i) missing, rusted or nonfunctional signages, traffic lights, and street lights (ii) uneven and partly damaged roads (iii) illegal movement patterns and traffic violations. We detect such road situations with a simple camera mounted on a car and avoid expensive camera networks.

This paper demonstrates the effective use of computer vision algorithms in economically addressing the urban mobility challenges in unstructured driving situations. We inspect road infrastructure and conditions (such as no lane markings, potholes, absence of street lights), defective traffic signs, and also demonstrate the detection of traffic violations.

The contributions of this paper are as follows:

- Demonstrating the scalability and effectiveness of the vision algorithms on 2000Km with a systematic evaluation.

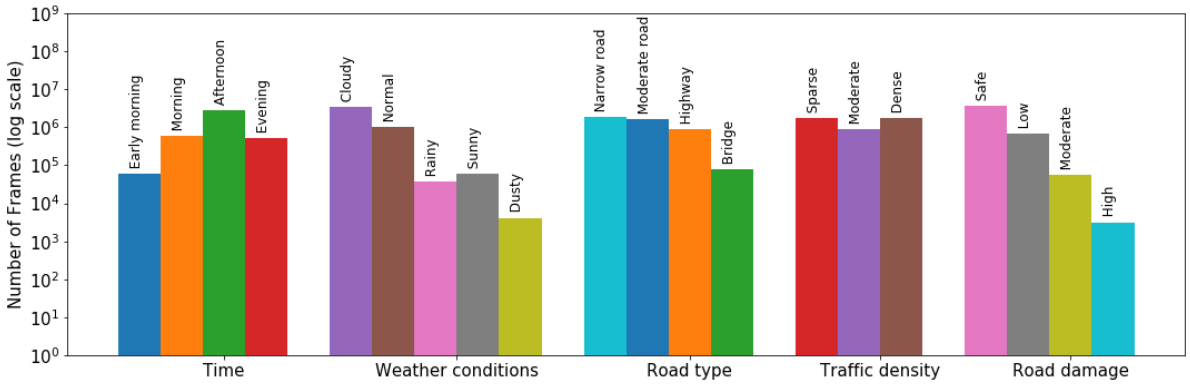


Figure 4.2: Distribution of 20 conditions-based hierarchical frame-level labels for different timings, weather conditions, road type, traffic density, and state of road damage on which we evaluate road surface, traffic infrastructure, and traffic violation models. Adverse weather conditions, congested lanes, heavy traffic, and damaged roads make our evaluation challenging and extensive

- We also release annotated resources that help further research on road safety on unstructured road situations.

4.2 Related Work

Many previous works have demonstrated the applications of the computer vision models for object detection, semantic segmentation, and vehicle tracking in urban mobility [62, 102, 9, 65]. The challenges like Nvidia AI city[58], CityFlow[84], and AutoNUE[1] have popularized city-scale mobility-related tasks.

4.2.1 Road Safety Systems

Early work on road safety [15, 64] demonstrated the effective collaboration between perception and control to increase the safety. Pedestrian detection [29], driver state and behaviour characterization etc., became an integral part of the ADAS system over the years. Zhang et al. [107] demonstrate a stereo-vision based training free approach to detect road regions accurately and robustly. Comprehensive surveys on vision-based traffic monitoring and vision applications in urban analytics are regularly conducted [17, 37, 36].

Dynamic objects around a vehicle also impact the safety and performance of the vehicle. Billones et al. [8] measure the speed of a vehicle by counting the frames it takes to pass a region of interest or reference lines. Singh et al. [81] present a framework for visual big data analytics for automatic detection of bike-riders without helmets in city traffic and discuss the challenges associated with city-scale surveillance for traffic control. Recent works have shown CNN to produce better results for helmet detection and classification [90]. Frossard et al. [25] detect driver's intent by detecting turn signals and

emergency flashers in video sequences. Datasets and robust models for pedestrian intent prediction are also present in the literature [40, 39, 41, 68, 47]. We use recent vision techniques to inspect helmet violations at scale on the unconstrained road scenes under different conditions in comparison to these works.

4.2.2 Road Inspection Systems

. Road infrastructure's soundness and suitable perceptibility form the essential elements of driver safety. Burschka et al. [10] introduce a vision-based system for traffic sign detection and ego-motion estimation. Another early work involves detecting cracked regions using a weakly supervised superpixel classifier trained on 220 images [87].

With the motivation to reduce asphalt pavement distresses, Kanza et al. [2] detect and localize potholes in 120 cropped road images using Histograms of oriented gradients (HOG) features and Naïve Bayes classifier. Zhang et al. [106] employ Convolutional Neural Network (CNN) to classify image patches of the road as defective or non-defective and evaluate it on 500 images. Yerram et al. [102] train a multi-step ERFNet based model with an attention mechanism on around 1000 unconstrained road scenes to segment road pixels into 9 different defects. However, these inspection methods apply road inspection at small scales compared to the approach in this paper.

The closest works to ours from this direction are i) Ma et al. [51], who utilize Fisher Vectors with Convolutional Neural Networks (CNN) to classify 700k images from 70k street segments into poor, fair, and good, and ii) Modhugu et al. [67], that demonstrate an automatic model for road infrastructure audit concerning traffic signs, street lights, and lane markings. In contrast to these works, we extend the range of functionalities, the volume of training data, and test the system's scalability extensively on a larger dataset spanning 2000 km of unconstrained roads.

4.2.3 Unconstrained Road Scenes and Datasets

Like many other areas, the data sets also drive research in understanding road scenes. Datasets like KITTI [**?**], Cityscapes [16], Argoverse [13], and NuScenes [12] facilitated the rapid development of this area. These results are often not directly applicable in unstructured road situations prevalent in large parts of the world. Developing economies' roads and traffic conditions in Latin America, Africa, and Asia are unstructured and unconstrained. The unstructured and unconstrained environments pose many challenging and complex problems creating scope for newer architectures and training methods. To meet the needs, Verma et al. [88] introduced an India Driving Dataset (IDD), a novel dataset for road scene understanding in unstructured environments. It consists of 10,000 images, with 34 classes collected over 182 drive sequences on roads in various parts of India. A rich label set with a four-level hierarchy represents the nuances of real driving behaviors in semantic segmentation.

Our attempt is to evaluate situations similar to that of IDD, but at a much large scale in terms of the length of the roads and the number of hours of driving. In contrast to the previous attempts in IDD, we

Table 4.1: Detection results for several tasks on their respective datasets.

Task / Dataset	Precision	Recall	F1	mAP@0.5
Street Lights	0.77	0.85	0.81	0.83
Traffic Signs	0.88	0.74	0.80	0.75
Traffic Participants	0.82	0.79	0.80	0.81
Helmet Violations	0.78	0.83	0.81	0.83
Potholes	0.66	0.53	0.59	0.54

do not focus on semantic segmentation or detection of common road objects. We focus on irregularities on the road surface and traffic infrastructure. Figure 4.2 demonstrates the diversity in the conditions we have used in the evaluation.

4.3 Road Safety Monitoring System

This section discusses the objectives and pipeline for the proposed vision-based road safety monitoring system.

4.3.1 Objectives

We perform road infrastructure audits and traffic violation monitoring on large-scale unconstrained road sequences. By computing on differing conditions of the road, traffic, and environment, we (i) evaluate existing computer vision techniques at scale for addressing road safety and (ii) summarize the state of road safety in the city and present our key findings. As part of the auditing road infrastructure, we identify possible irregularities in streets(including missing lane markings and potholes), absence of street lights, and defective traffic signs. Our selection of tasks is attributed to them providing essential visual cues for navigation and allowing for efficient mobility. Moreover, such infrastructure is widely prevalent, making them appropriate for evaluating the road infrastructure module at scale.

We focus on two-wheeled motor vehicle riders for monitoring traffic violations, as they account for nearly 30% of road accident deaths worldwide [98]. Around 74% of riders are involved in fatal accidents not found to be wearing protective helmets [55]. So, we find helmets to be critical for safety and specifically identify helmet violations. The widespread occurrence of helmet violations also allows for evaluating the traffic violations module under various conditions. We present the evaluation scores of existing methods on these tasks on large-scale unconstrained videos in Section 5 and the city-scale assessment of road safety in Section 6.

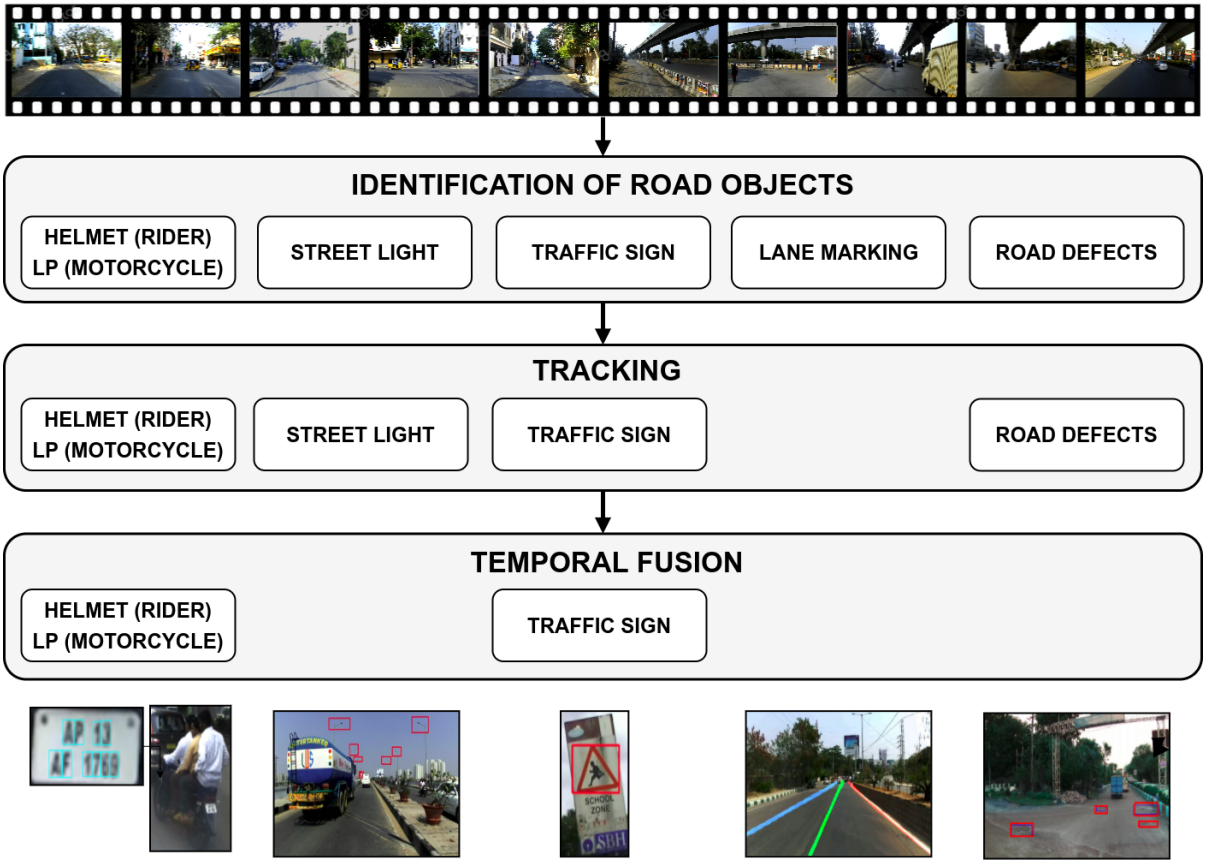


Figure 4.3: Pipeline of proposed road safety monitoring system. We identify road objects in each frame from a video feed, track them across the video, and temporally fuse predictions.

4.3.2 Pipeline

We simultaneously carry out several different tasks from a video feed, as visualized in Figure 4.3. We first detect (or segment) road objects of primary interest in each frame and then track the detected items across the video. We finally leverage temporal viewpoint variance to fuse and strengthen the predictions. For each of the task, the design choices are primarily motivated towards (i) using networks with high performance and widely adopted by the community and (ii) having a fast inference time to evaluate models on large-scale data. For initial training and quantitative analysis, we annotate 4.3k images from the Indian Driving dataset (IDD) by Verma et al. [88]. We use $80:20$ as a train:test split—the detection results for different tasks are shown in Table 4.1. We note that the network's choice is not a focus of this work and may be replaced with other architectures.

4.3.2.1 Identification of Road Objects

We now discuss the models this work employs for identifying the irregularities in lane markings, infrastructure, and traffic rules.

Lane markings are continuous and require strong structural priors for segmentation. We employ Spatial Convolutional Neural Network (SCNN) [99] to segment lane markings as it also extrapolates the lane markings occluded by other objects on the road surface.

Traffic signs, Street lights, Potholes and Traffic participants are directly detected from the incoming video feed. We employ YOLOv4 by Bochkovskiy et al. [9] for this purpose, as YOLOv4 provides a good trade-off between detection performance and speed. As part of traffic participants, we detect riders and motorcycles separately and associate them using simple heuristics.

Helmet violations are computed on the detected crops of riders and motorcycles. For detecting license plates, we use the WPOD-NET by Silva et al. [80], which detects and geometrically transforms the license plates to a planar front-view. For helmet detection on riders, we employ YOLOv4 for the same reasoning mentioned earlier.

4.3.2.2 Tracking Identified Objects

The detected objects, viz. traffic signs, street lights, potholes, and traffic participants, are passed to a tracker. By using a tracker, we (i) avoid redundant counting, (ii) temporally fuse predictions, and (iii) smoothen detections across the frames. We use Simple Online and Realtime Tracking (SORT) by Bewley et al. [6] as the tracker since it is extremely fast and provides relatively high accuracy. SORT tracks objects by employing the Kalman filter to handle motion prediction and the Hungarian method for frame-by-frame data association. They use bounding box overlap of detections as the association metric.

4.3.2.3 Temporal Fusion

Since subsequent frames provide a different viewpoint of the same object to the network, we combine individual frames' predictions and smoothen any aberration. We use majority voting, a standard late fusion technique, to combine frame-level predictions of traffic sign classification, rider-motorcycle association and helmet classification.

4.4 Data Capture

Constant wear and tear, heavy rains, and laying of underground cables necessitate routine maintenance of roads. Due to economic and regulatory hindrances, a lack of maintenance, weak accountability, and unreliable funding has resulted in unconstrained road conditions. As shown in Figure 4.5, rough terrains including muddy, bumpy roads and defects such as potholes, waterlogs and hazardous road objects

Figure 4.4: Capture under different lighting conditions. Clockwise from top-left: cloudy, sunny, rainy and shadows.

are a common occurrence. The problem is compounded due to varying level of occlusions, unstructured motion, and cluttered background, Other road infrastructure components that aid navigation is found in a similar state of poor planning and maintenance. Lane markings either do not exist or are covered in dust, traffic signs are worn out or are concealed amidst the background clutter, and street lights are obscured by vegetation.

As shown in Figure 4.6, there is also a significant variation in the traffic participants and their behavior. Two-wheelers (motorcycles) and three-wheelers (auto-rickshaws) are more prominent, while animals, including stray dogs and cows, are frequently spotted on the road. With the absence of sidewalks, pedestrians are often found walking alongside vehicles, and jaywalking is a common occurrence even during a traffic rush. Moreover, unsustainable population and vehicle density in cosmopolitan cities, combined with weak enforcement of traffic rules has brought about a disarray in driving behavior. Non-adherence to road lanes, traffic congestion, and chaotic unmanned junctions are widely prevalent. Additionally, violations of traffic rules such as wrong-side driving, swerving, vehicle overloading, and illegal parking cause further confusion.

In total, we record 257 video sequences amounting to 75 hours of capture and covering a distance of over 2000 Km. These road scenes are captured under diverse lighting conditions. Our setup consists of a car-mounted camera, recording 1920×1080 resolution videos at 15 FPS. The videos are synchronized to a GPS device, polling every second.

Figure 4.5: Dangerous road conditions. Clockwise from top: bumpy and muddy road, low-hanging wire with cows on road, narrow street with two-way traffic and under construction structure left unprotected.

4.5 Evaluation under different conditions

We discuss the performance of various models under differing conditions of time, weather, road damage, road type, and traffic density. We first annotate our data capture with a subcategory label for each category mentioned above. We then randomly sample 100 frames from each subcategory, totaling 2000, and annotate them with bounding boxes for each task. We provide our evaluation results in Table 4.2. Looking at the overall scores, we notice that illumination plays a crucial role in model performance, with low and moderate illumination as the optimum and high illumination as the adversary. This can be observed in the time category, where mAP scores are high during early morning and evening, while dropping during morning and afternoon. A similar trend was observed in the weather category, where low-illuminated cloudy weather performs similar to normal weather conditions, while struggling under sunny weather. Further, there is also performance drop when artifacts such as dust or raindrops are present in the field of view of the camera. Coming to differing conditions of road damage, the mAP scores decrease as the road quality worsens, mainly due to the model performance on potholes. An abnormality is seen for high road damage, whose score is influenced by the high performance on street lights, a task that does not get impacted by the road quality. Additionally, we notice that model performance is adversely affected by the traffic density due to difficulty in identifying traffic participants and their helmets in congested traffic scenes. However, our tasks do not exhibit any observable trend under differing conditions of road types. Overall, we find is a general trend of performance drop in unconstrained and unstructured road scenes.

Figure 4.6: Traffic participants and driving behavior. Clockwise from top-left: diversity in participants, pedestrian jaywalking during heavy traffic, chaotic unmanned junctions and wrong-side driving with illegal parking on the road.

Figure 4.7: Our challenging *LP-UC* dataset contains both single and double-line license plates with different fonts and varying character lengths. The dataset incorporates numerous variations, including license-plates that are broken, dented, blurred or rusted.

We now look at task-specific performance under challenging conditions. High traffic density and variations in traffic participants make it challenging for the model to correctly detect traffic participants, as illustrated in Fig. 4.8a and 4.8b. Often, nearby riders are detected as a single entity, while autorickshaw participants are falsely detected as motorbike riders, highlighting challenges posed by variations in data. Potholes, due to their inherent structure, have poor reflectivity, thereby appearing as dark patches to the camera. This property frequently causes the model to falsely detect circular shadows as potholes, as shown in Fig. 4.8c. Traffic signs under dusty, rainy, and sunny conditions is difficult. Contrasting to other tasks, the detection performance of street lights remains relatively uniform across varying conditions of weather, road and traffic, as observed in Fig. 4.8e. Further, we visualize in Fig. 4.8f, the erroneous detections majorly occur from falsely detecting tree branches and electric poles as street lights, whenever their structure appears exceedingly similar. Finally, we notice a marked difficulty in

Table 4.2: mAP scores of various models under differing conditions of road, traffic and environment

	Time				Weather Conditions				
Task/Dataset	Early Morning	Morning	Afternoon	Evening	Normal	Cloudy	Dusty	Rainy	Sunny
Street Lights	.91	.95	.80	.94	.86	.82	.93	.91	.80
Traffic Signs	.67	.32	.50	.50	.36	.49	.14	.19	.19
Traffic Participants	.94	.88	.87	.88	.92	.89	.87	.91	.95
Helmet Detection	.72	.60	.61	.68	.50	.56	.71	.53	.67
Potholes	.48	.66	.46	.54	.46	.34	.34	.43	.19
Overall	.74	.68	.65	.71	.62	.62	.60	.59	.56

	Road Damage				Road Type				Traffic density		
Task/Dataset	Safe	Low	Medium	High	Bridge	Narrow	Standard	Highway	Sparse	Moderate	Dense
Street Lights	.81	.73	.81	.91	.89	.81	.90	.82	.90	.90	.92
Traffic Signs	.56	.66	.49	.46	.33	.46	.66	.44	.54	.58	.37
Traffic Participants	.82	.88	.88	.86	.82	.98	.91	.83	.93	.85	.78
Helmet Detection	.66	.80	.69	.82	.72	.79	.63	.65	.72	.70	.65
Potholes	.71	.41	.32	.36	.55	.42	.71	.57	.55	.52	.35
Overall	.73	.67	.63	.73	.65	.67	.80	.67	.73	.71	.61

Table 4.3: The city-level assessment of various components of road safety are presented below.

Traffic signs		Street lights	Lanes	Potholes	Helmet
Visibility-Range	Defective	Avg. pair distance	No markings	Percentage of stretches	Percentage of Violating riders
9.7m	37.5 %	165m	60.3 %	4.0 %	45.9 %

detecting traffic signs under poor visibility conditions, as shown in Fig. 4.8g and Fig. 4.8h. This can be attributed to their position, where traffic signs are relatively far-away to the ego-vehicle and close to the ground-level.

4.6 City-scale assessment

Metrics for Road Safety Assessment: Traffic signs pre-inform riders of impending road scenes, and hence their range of visibility plays a crucial role. We measure the visibility range of a traffic sign as the distance (calculated using GPS location) traced by the ego-vehicle from the traffic sign's first frame of detection to its last frame of detection. Additionally, we also classify traffic signs as defective or normal. For lane markings and street lights, we expect them to occur all along the road periodically. Therefore, in these cases, the measure of interest is the road stretches without lane markings and street lights. We tag the detected street lights geographically and calculate the average distance between street lights along the route. For lane markings, we use the normalized lane regularity score in [67] that takes into account the percentage of pixels identified as lane markings along a video sequence stretch covering

50 meters of road. Based on this score, we classify the lane markings as fair, faded, or unfair.

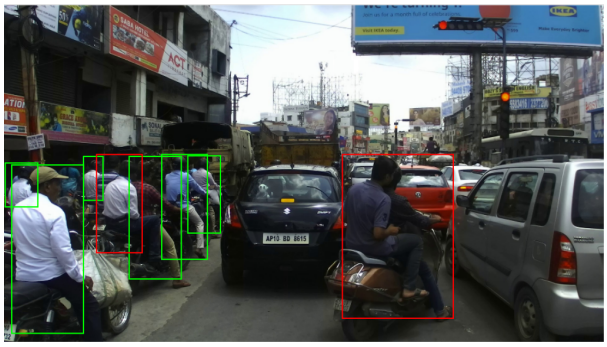
$$\text{Lane regularity score(r)} = \frac{\sum_{i \in f} \frac{l_i}{n_i}}{d \cdot |f|} \tag{4.1}$$

$$\text{Normalized lane regularity score} = \frac{r - r_{\min}}{r_{\max} - r_{\min}} \tag{4.2}$$

In equation 4.1, l_i is the total number of pixels identified as the lane marking in i^{th} frame, n_i is the number of pixels in the i^{th} frame, d is the length of the stretch ($= 50$), f is set of frames recorded to cover the stretch of distance d . In equation 4.2, r_{max} and r_{min} are maximum and minimum regularity scores determined experimentally over the complete data. To measure the quality of the road surface, we calculate the number of road defects for every 100 meters. We classify these stretches of road as poor, average or fair, based on their frequency of occurrence of road defects. For helmet violations, we calculate the percentage of riders not wearing results since it portrays the compliance level of road participants to traffic rules.

Quantitative Evaluation: We run the proposed system on videos collected from across the city and present a quantitative assessment on road safety in Table 4.3. The results shed light on the lack of maintenance of roads and road objects. Traffic signs are visible only at a close range of 10m, and a significant 37.5% percentage of them are rusted or faded out. Further, a street light is present only at every 165m on average, making night-driving a dangerous activity on such unconstrained roads. Lane markings are not a common sight, especially on side roads and streets. Seldom do road repairs result in 4.0% of road stretches with potholes. Finally, we observe that 45.9% of riders violate helmet rules, indicating low compliance with traffic rules.

Visualizations: By geotagging our predictions, we visualize areas in the city that are prone to helmet violations in Figure 4.9a. Similarly, Fig. 4.9b presents stretches in the city where lane markings are absent. Such visualizations allow authorities to zero down and attend to high-risk areas.



(a) Rider detection in dense traffic

(b) Rider detection in presence of diverse participants

(c) Pothole detection during shadow conditions

(d) Pothole detection on muddy roads

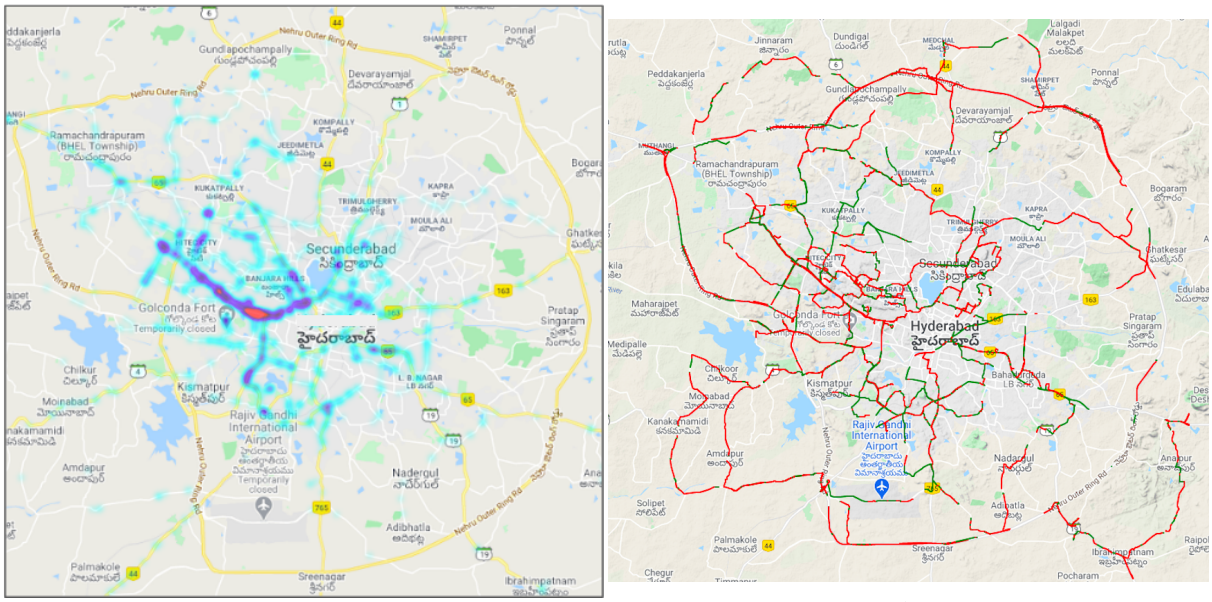
(e) Street lights detection in cloudy weather

(f) Street light detection with vegetation near roads

(g) Traffic sign detection on dusty roads

(h) Traffic sign detection under sunlight

Figure 4.8: Predictions under challenging conditions. For each image, only predictions of object of interest are shown. Green and red color boxes indicate correct and wrong predictions respectively.



(a) (b)

Figure 4.9: (a) Heatmap of traffic violations identified across the city. Red and neon blue indicate higher density and lower density regions respectively. (b) Red stretches indicate absence of streetlights whereas green stretches indicates presence streetlights.

Chapter 5

Conclusions and Future Directions

In this thesis, we investigate computer vision solutions for visual perception of road scenes to provide alternatives for HD maps in parametrically understanding the road scene and to inform the road safety hazards with a scalable setup. These solutions are a step towards an alternative to HD maps for autonomous vehicles.

The proposed architecture in chapter 3 exploits the complementary properties of aerial and ground imagery to derive a parametric representation of scene geometry. We start by creating a dataset with scene attribute annotations to supplement the publicly available Argoverse dataset. We propose a novel approach for parametric road scene understanding and show that our Unified model performs better than prior works. We also extensively show, both quantitatively and qualitatively, the advantage of jointly learning from both aerial and ground modalities.

Further in chapter 4, we propose simple mobile imaging setup to address several common problems in urban mobility and road safety. We work on long video stretches containing unconstrained road scenes captured from a car-mounted camera in a scalable and affordable manner. We discuss the performance of state-of-the-art computer vision models under 20 conditions of a different time, weather conditions, traffic density, and state of road damage. We finally demonstrate large-scale analytics of irregular road infrastructure and traffic violations is feasible with existing computer vision techniques for generating the semantic layer for HD maps. We release all the trained models, code, and annotations to encourage future work in this direction.

Our work exploiting complementary properties from aerial and ground imagery unlocks many further potential applications in autonomous navigation, such as landmark localization, motion forecasting, and planning. Mainly, the usage of aerial imagery as a scene layout prior has much scope to explore further. Another future direction for our work is identifying locations to recommend road infrastructure like traffic signs, speed breakers, etc., to enhance road safety.

Related Publications

1. **Looking Farther in Parametric Scene Parsing with Ground and Aerial Imagery** Raghava Modhugu*, Harish Rithish Sethuram*, Manmohan Chandraker, C.V. Jawahar, *IEEE International Conference on Robotics and Automation, 2021 (ICRA 2021)*

2. **Evaluating Computer Vision Techniques for Urban Mobility on Large-Scale, Unconstrained Roads** Raghava Modhugu*, Harish Rithish Sethuram*, Ranjith Reddy*, Rohit Saluja*, C.V. Jawahar, *arXiv 2021*

3. **Dear Commissioner, Please Fix These: A Scalable System for Inspecting Road Infrastructure** Raghava Modhugu*, Ranjith Reddy*, C.V. Jawahar, *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics, 2019*

(*equal contribution)

Bibliography

[1] *Autonomous navigation in unconstrained environments (AutoNUE) workshop and challenge at ECCV'18.*, 2018.

[2] K. Azhar, F. Murtaza, M. H. Yousaf, and H. A. Habib. Computer Vision Based Detection and Localization of Potholes in Asphalt Pavement Images. In *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–5. IEEE, 2016.

[3] S. M. Azimi, P. Fischer, M. Körner, and P. Reinartz. Aerial lanenet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 57(5):2920–2938, 2018.

[4] T. Baltrusaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. page 423–443, 2019.

[5] Y. Bengio, A. C. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2013.

[6] A. Bewley, G. Zongyuan, F. Ramos, and B. Upcroft. Simple online and realtime tracking. *ICIP*, 2016.

[7] R. Bhoraskar, N. Vankadhara, B. Raman, and P. Kulkarni. Wolverine: Traffic and road condition estimation using smartphone sensors. In *2012 Fourth International Conference on Communication Systems and Networks (COMSNETS 2012)*, pages 1–6, 2012.

[8] R. K. C. Billones, A. A. Bandala, E. Sybingco, L. A. G. Lim, and E. P. Dadios. Intelligent system architecture for a vision-based contactless apprehension of traffic violations. In *2016 IEEE Region 10 Conference (TENCON)*, pages 1871–1874. IEEE, 2016.

[9] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934*, 2020.

[10] D. Burschka and G. D. Hager. Vision-based 3d scene analysis for driver assistance. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 812–818. IEEE, 2005.

[11] 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.

[12] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[13] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019.

[14] G. Cheng, C. Wu, Q. Huang, Y. Meng, J. Shi, J. Chen, and D. Yan. Recognizing road from satellite images by structured neural network. *Neurocomputing*, 356:131–141, 2019.

[15] X. Cindy, F. Collange, F. Jurie, and P. Martinet. Object tracking with a pan-tilt-zoom camera: application to car driving assistance. In *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*, volume 2, pages 1653–1658. IEEE, 2001.

[16] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[17] S. R. E. Datondji, Y. Dupuis, P. Subirats, and P. Vasseur. A survey of vision-based traffic monitoring of road intersections. *IEEE transactions on intelligent transportation systems*, 17(10):2681–2698, 2016.

[18] J. Delmerico, A. Giusti, E. Mueggler, L. M. Gambardella, and D. Scaramuzza. "on-the-spot training" for terrain classification in autonomous air-ground collaborative teams. In D. Kulić, Y. Nakamura, O. Khatib, and G. Venture, editors, *2016 International Symposium on Experimental Robotics*. Springer International Publishing, 2017.

[19] J. Delmerico, E. Mueggler, J. Nitsch, and D. Scaramuzza. Active autonomous aerial exploration for ground robot path planning. *IEEE Robotics and Automation Letters*, 2(2):664–671, 2017.

[20] M. Elhousni, Y. Lyu, Z. Zhang, and X. Huang. Automatic building and labeling of hd maps with deep learning. In *AAAI*, 2020.

[21] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, and H. Balakrishnan. The pothole patrol: using a mobile sensor network for road surface monitoring. In *Proceedings of the 6th international conference on Mobile systems, applications, and services*, pages 29–39, 2008.

[22] A. Ess, T. Müller, H. Grabner, and L. J. Van Gool. Segmentation-based urban traffic scene understanding. In *BMVC*, volume 1, page 2, 2009.

[23] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. 2010.

[24] T. Feng, Q.-T. Truong, D. Thanh Nguyen, J. Yu Koh, L.-F. Yu, A. Binder, and S.-K. Yeung. Urban zoning using higher-order markov random fields on multi-view imagery data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 614–630, 2018.

[25] D. Frossard, E. Kee, and R. Urtasun. Deepsignals: Predicting Intent of Drivers Through Visual Signals. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9697–9703. IEEE, 2019.

[26] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3d traffic scene understanding from movable platforms. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):1012–1025, 2013.

[27] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[28] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[29] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf. Survey of Pedestrian Detection for Advanced Driver Assistance Systems. *IEEE transactions on pattern analysis and machine intelligence*, 32(7):1239–1258, 2009.

[30] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[31] R. B. Girshick. Fast r-cnn. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[32] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[33] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[34] N. Homayounfar, W.-C. Ma, S. Kowshika Lakshmikanth, and R. Urtasun. Hierarchical recurrent attention networks for structured online maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3417–3426, 2018.

[35] S. Hu, M. Feng, R. M. Nguyen, and G. Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018.

[36] M. R. Ibrahim, J. Haworth, and T. Cheng. Understanding cities with machine eyes: A review of deep computer vision in urban analytics. *Cities*, 96:102481, 2020.

[37] N. K. Jain, R. Saini, and P. Mittal. A review on traffic monitoring system techniques. In *Soft Computing: Theories and Applications*, pages 569–577. Springer, 2019.

[38] C. Jawahar and V. Padmanabhan. Technology interventions for road safety and beyond. *ACM*, 2019.

[39] C. G. Keller and D. M. Gavrila. Will the Pedestrian Cross? A Study on Pedestrian Path Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):494–506, 2013.

[40] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity Forecasting. In *European Conference on Computer Vision*, pages 201–214. Springer, 2012.

[41] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila. Context-Based Pedestrian Path Prediction. In *European Conference on Computer Vision*, pages 618–633. Springer, 2014.

[42] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[43] A. Li, H. Hu, P. Mirowski, and M. Farajtabar. Cross-view policy learning for street navigation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8100–8109, 2019.

[44] D. Li, J. Tang, and S. Liu. Brief industry paper: An edge-based high-definition map crowdsourcing task distribution framework for autonomous driving. In *2021 IEEE 27th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 453–456, 2021.

[45] W. Li, C. He, J. Fang, and H. Fu. Semantic segmentation based building extraction method using multi-source gis map datasets and satellite imagery. In *CVPR Workshops*, pages 238–241, 2018.

[46] J. Liang and R. Urtasun. End-to-end deep structured models for drawing crosswalks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 396–412, 2018.

[47] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles. Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction. *IEEE Robotics and Automation Letters*, 5(2):3485–3492, 2020.

[48] B. Liu, B. Zhuang, S. Schulter, P. Ji, and M. Chandraker. Understanding road layout from videos as a whole. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4414–4423, 2020.

[49] S. Liu, E. Johns, and A. J. Davison. End-to-end multi-task learning with attention. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1871–1880, 2019.

[50] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018.

[51] K. Ma, M. Hoai, and D. Samaras. Large-scale Continual Road Inspection: Visual Infrastructure Assessment in the Wild. In *BMVC*, 2017.

[52] T. Manderson, S. Wapnick, D. Meger, and G. Dudek. Learning to drive off road on smooth terrain in unstructured environments using an on-board camera and sparse aerial images. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1263–1269, 2020.

[53] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun. Enhancing road maps by parsing aerial images around the world. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1697, 2015.

[54] G. Máttyus, S. Wang, S. Fidler, and R. Urtasun. Hd maps: Fine-grained road segmentation by parsing ground and aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3611–3619, 2016.

[55] Ministry of Road Transport and Highways, India. *Road Accidents in India*, 2017.

[56] I. Misra, A. Shrivastava, A. K. Gupta, and M. Hebert. Cross-stitch networks for multi-task learning. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4003, 2016.

[57] E. Morvant, A. Habrard, and S. Ayache. Majority vote of diverse classifiers for late fusion. *ArXiv*, abs/1404.7796, 2014.

[58] M. Naphade, M.-C. Chang, A. Sharma, D. C. Anastasiu, V. Jagarlamudi, P. Chakraborty, T. Huang, S. Wang, M.-Y. Liu, R. Chellappa, J.-N. Hwang, and S. Lyu. The 2018 nvidia ai city challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[59] G. of India. Annual report. *Ministry of Road Transportation and Highways, India*, 2021.

[60] H. Oliveira and P. L. Correia. Automatic road crack detection and characterization. *IEEE Transactions on Intelligent Transportation Systems*, 14(1):155–168, 2013.

[61] S.-G. H.-L. A. G. on Sustainable Transport. Mobilizing sustainable transportation for development. *Analysis and Policy Recommendations from the United Nations*, 2014.

[62] A. Ošep, W. Mehner, P. Voigtlaender, and B. Leibe. Track, then Decide: Category-Agnostic Vision-based Multi-Object Tracking. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.

[63] D. Pannen, M. Liebner, W. Hempel, and W. Burgard. How to keep hd maps for automated driving up to date. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2288–2294, 2020.

[64] L. Petersson, N. Apostoloff, and A. Zelinsky. Driver assistance: An integration of vehicle monitoring and control. In *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, volume 2, pages 2097–2103. IEEE, 2003.

[65] Y. Qian, L. Yu, W. Liu, and A. G. Hauptmann. Electricity: An efficient multi-camera vehicle tracking system for intelligent city. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[66] M. Quintana, J. Torres, and J. M. Menéndez. A simplified computer vision system for road surface inspection and maintenance. *IEEE Transactions on Intelligent Transportation Systems*, 17(3):608–619, 2016.

[67] M. Raghava, R. Reddy, and C. Jawahar. Dear commissioner, please fix these: A scalable system for inspecting road infrastructure. *NCVPRIPG*, 2019.

[68] A. Rasouli, I. Kotseruba, and J. K. Tsotsos. Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 206–213, 2017.

[69] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[70] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[71] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767, 2018.

[72] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[73] T. Roddick and R. Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11138–11147, 2020.

[74] S. Ruder. An overview of multi-task learning in deep neural networks. *ArXiv*, abs/1706.05098, 2017.

[75] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[76] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[77] S. Schulter, M. Zhai, N. Jacobs, and M. Chandraker. Learning to look around objects for top-view representations of outdoor scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 787–802, 2018.

[78] A. Seff and J. Xiao. Learning from maps: Visual common sense for autonomous driving. *arXiv preprint arXiv:1611.08583*, 2016.

[79] S. Sengupta, P. Sturgess, L. Ladický, and P. H. S. Torr. Automatic dense visual semantic mapping from street-level imagery. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 857–862, 2012.

[80] S. M. Silva and C. R. Jung. License plate detection and recognition in unconstrained scenarios. In *European Conference on Computer Vision*. Springer, 2018.

[81] D. Singh, C. Vishnu, and C. K. Mohan. Visual Big Data Analytics for Traffic Monitoring in Smart City. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 886–891, 2016.

[82] G. Strezoski, N. van Noord, and M. Worring. Many task learning with task routing. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1375–1384, 2019.

[83] B. Suger and W. Burgard. Global outer-urban navigation with openstreetmap. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1417–1422. IEEE, 2017.

[84] Z. Tang, M. Naphade, M. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J. Hwang. CityFlow: A City-Scale Benchmark for Multi-Target Multi-Camera Vehicle Tracking and Re-Identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8798, 2019.

[85] D. Tran, E. Tadesse, D. Osipychev, J. Du, W. Sheng, Y. Sun, and H. Chen. A collaborative control framework for driver assistance systems. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6038–6043. IEEE, 2017.

[86] S. Varadharajan, S. Jose, K. Sharma, L. Wander, and C. Mertz. Vision for road inspection. In *IEEE Winter Conference on Applications of Computer Vision*, pages 115–122, 2014.

[87] S. Varadharajan, S. Jose, K. Sharma, L. Wander, and C. Mertz. Vision for Road Inspection. In *IEEE Winter Conference on Applications of Computer Vision*, pages 115–122. IEEE, 2014.

[88] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019.

[89] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[90] C. Vishnu, D. Singh, C. K. Mohan, and S. Babu. Detection of motorcyclists without helmet in videos using convolutional neural network. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017.

[91] L. Wang, D. Cheng, F. Gao, F. Cai, J. Guo, M. Lin, and S. Shen. A collaborative aerial-ground robotic system for fast exploration. In J. Xiao, T. Kröger, and O. Khatib, editors, *Proceedings of the 2018 International Symposium on Experimental Robotics*, pages 59–71, 2020.

[92] S. Wang, M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun. Torontocity: Seeing the world with a million eyes. *arXiv preprint arXiv:1612.00423*, 2016.

[93] Z. Wang, B. Liu, S. Schulter, and M. Chandraker. A dataset for high-level 3d scene understanding of complex road scenes in the top-view. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2019.

[94] Z. Wang, B. Liu, S. Schulter, and M. Chandraker. A parametric top-view representation of complex road scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10325–10333, 2019.

[95] J. D. Wegner, S. Branson, D. Hall, K. Schindler, and P. Perona. Cataloging public objects using aerial and street-level images-urban trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6014–6023, 2016.

[96] S. Workman, S. Richard, and N. Jacobs. Understanding and mapping natural beauty. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[97] S. Workman, M. Zhai, D. J. Crandall, and N. Jacobs. A unified model for near and remote sensing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2688–2697, 2017.

[98] World Health Organization. *Global status report on road safety*, 2018.

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

[100] D. Xu, W. Ouyang, X. Wang, and N. Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018.

[101] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3917–3925, 2018.

[102] Yarram Sudhir, Girish Varma, and C. V. Jawahar. City-scale road audit system using deep learning. *International Conference on Intelligent Robots and Systems*, 2018.

[103] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017.

[104] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875, 2017.

[105] L. Zhang, F. Yang, Y. Daniel Zhang, and Y. J. Zhu. Road crack detection using deep convolutional neural network. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3708–3712, 2016.

[106] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu. Road crack detection using deep convolutional neural network. In *2016 IEEE international conference on image processing (ICIP)*, pages 3708–3712. IEEE, 2016.

[107] Y. Zhang, J. Yang, J. Ponce, and H. Kong. Dijkstra Model for Stereo-Vision Based Road Detection: A Non-Parametric Method. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.