

A Universal Hybrid Early-Exit Slimmable Vision Framework-UniNetEE

A MAJOR PROJECT REPORT

Submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology

in

**COMPUTER SCIENCE AND ENGINEERING
(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)**

BY

M.V.Padma Yesaswi

22331A4233

Vayilapalli Dileep

22331A4262

**K.Chandra Sekhar
23335A4203**

**V.Ramya Sree
23335A4207**

**Under the Supervision of
Dr. P Satheesh
Professor, DE, M.Tech, Ph.D**



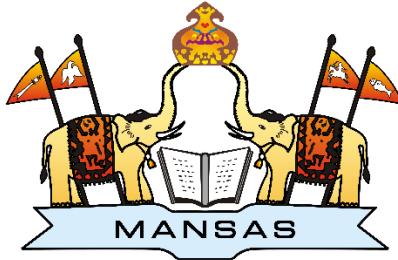
DEPARTMENT OF Computer Science And Engineering (Artificial Intelligence and Machine Learning)

MAHARAJ VIJAYARAM GAJAPATHI RAJ COLLEGE OF ENGINEERING
(Autonomous)

(Approved by AICTE, New Delhi, and permanently affiliated to JNTUGV, Vizianagaram),
Listed u/s 2(f) & 12(B) of UGC Act 1956.

Vijayaram Nagar Campus, Chintalavalasa, Vizianagaram-535005, Andhra Pradesh
APRIL, 2025

CERTIFICATE



This is to certify that the project report entitled "**A Universal Hybrid Early Exit Slimmable Vision Framework UniNetEE**" being submitted by M.V.Padma Yesaswi(22331A4233), Vayilapalli Dileep(22331A4262), K.Chandra Sekhar(23335A4203), V.Ramya Sree(23335A4207) in partial fulfilment for the award of the degree of "**Bachelor of Technology**" in CSE(AI & ML) is a record of bonafide work done by them under my supervision during the academic year 2024-2025.

Dr. P Satheesh
Professor, DE, M.Tech, Ph.D,
Supervisor,
Department of DE,
MVGR College of Engineering(A),
Vizianagaram.

Dr. V. Jyothi
Associate Professor,
Head of the Department,
Department of DE,
MVGR College of Engineering(A),
Vizianagaram.

External Examiner

DECLARATION

We hereby declare that the work done on the dissertation entitled "**A Universal Hybrid Early-Exit Slimmable Vision Framework - UniNetEE.**" has been carried out by us and submitted in partial fulfilment for the award of credits in Bachelor of Technology in Computer Science and Engineering of MVGR College of Engineering (Autonomous) and affiliated to JNTUGV, Vizianagaram. The various contents incorporated in the dissertation have not been submitted for the award of any degree of any other institution or university.

ACKNOWLEDGEMENTS

We express our sincere gratitude to **Dr. P Satheesh** for his invaluable guidance and support as our mentor throughout the project. His unwavering commitment to excellence and constructive feedback motivated us to achieve our project goals. We are greatly indebted to him for his exceptional guidance.

Additionally, we extend our thanks to **Prof. P.S. Sitharama Raju (Director)**, **Prof. Ramakrishnan Ramesh (Principal)**, and **Dr. V. Jyothi (Head of the Department)** for their unwavering support and assistance, which were instrumental in the successful completion of the project. We also acknowledge the dedicated assistance provided by all the staff members in the Department of **CSE(AI & ML)**. Finally, we appreciate the contributions of all those who directly or indirectly contributed to the successful execution of this endeavour.

Venkata Padma Yesaswi Madabattula (22331A4233)

Vayilapalli Dileep (22331A4262)

K.Chandra Sekhar (23335A4203)

V.Ramya Sree (23335A4207)

ABSTRACT

UrbanVision AI is a unified multi-task Artificial Intelligence system developed for urban and environmental scene understanding. The system integrates Image Classification, Object Detection, and Semantic Segmentation into a single hybrid deep learning framework capable of generating multiple insights from a single image.

The classification module is built using a CNN-based architecture inspired by ResNet for identifying environmental attributes such as weather conditions, road type, and scene category. Object detection is implemented using a YOLO-based single-stage detection algorithm for identifying vehicles, pedestrians, traffic signals, and other urban objects. Semantic segmentation is performed using a UNet-inspired encoder-decoder architecture to generate pixel-wise classification of roads, buildings, sky, and other scene components.

The backend is developed using Flask, incorporating user authentication, OTP-based password recovery, secure file upload handling, session management, and an admin dashboard. The system follows a modular architecture ensuring scalability and maintainability.

UrbanVision AI demonstrates the integration of deep learning models with a full-stack web application, providing practical applications in smart city monitoring, traffic analysis, and environmental intelligence.

CONTENTS

	Page No
List of Abbreviations	1
List of Figures	3
List of Tables	4
1. Introduction	5
1.2. Identification of the Problem	5
1.3. Problem definition	6
1.4. Objective	6
1.5. Existing models	6
2. Literature Survey	8
2.1. Multi-Task Learning in Computer Vision	8
2.2. Convolutional Neural Networks for Image Classification	8
2.3. Object Detection Models in Urban Scenes	8
2.4. Semantic Segmentation Techniques	8
2.5. Hybrid CNN-Attention Architectures	8
2.6. Early-Exit Neural Networks for Efficient Inference	9
2.7. Slimmable Neural Networks for Resources Adaptability	9
3. Theoretical Background	10
3.1. Machine Learning and Deep Learning Concepts	10
3.1.1. Convolutional Neural Networks (CNNs)	10
3.1.2. Attention Mechanisms	10
3.1.3. Early-Exit Neural Networks	10
3.1.4. Slimmable Neural Networks	10
3.2. Object Detection Techniques	10
3.2.1 YOLO (You Only Look Once)	11
3.2.2 Faster R-CNN	11
3.3 Semantic Segmentation Techniques	11
3.3.1 U-Net	11
3.3.2 DeepLabv3+	11
3.4 Multi-Task Learning Frameworks	11
3.4.1 Shared Backbone Networks	11
3.4.2 Multi-Task Loss Optimisation	11
3.5 Web Integration Concepts	12
3.5.1. Flask-Based Web Framework	12
3.5.2. Session and Cookie Management	12
3.5.3. Visualisation and Reporting	12
3.6. Security and Ethical Considerations	12
3.6.1. Data Privacy	12
3.6.2. Ethical AI Principles	12
4. Approach Description	13
4.1. Approach Overview	13
4.2. Approach Flow	14

5. Data Exploration	15
5.1. Dataset Description	15
5.1.1. Data Manipulation	15
5.1.2. Data Preparation	16
5.1.2.1. Missing Values	16
5.1.2.2. Duplicate values	16
5.1.2.3. Outliers	16
6. Data Analysis	18
6.1. Exploratory Data Analysis (EDA) and Its Types	18
6.2. Importance of Data Analysis	18
6.3. EDA Inferences for Environmental Scene Dataset	19
6.3.1. Scene Categories	19
6.3.2. Weather Conditions	19
6.3.3. Traffic Density	19
6.3.4. Object Distribution	19
6.3.5. Segmentation Masks	19
6.3.6. Outliers and Data Quality	20
7. Modelling	21
7.1. Model Development	21
7.1.1. Data Modelling Process	21
7.1.2. Model Evaluation	22
7.1.3. Adaptive & Efficient Inference	22
7.1.4. Model Deployment & Integration	23
8. Results and Conclusions	24
8.1. Results	24
8.2. Conclusions	25
References	34
Appendix A: Packages, Tools Used & Working Process	35
Python Programming Language	35
Libraries	35
NumPy	35
Pandas	36
Matplotlib	36
Seaborn	36
Sklearn	36
Tools used Jupyter Notebooks	37
Working Process of the Project	38
Appendix B: Source Code	39

List of Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks
LSTM	Long Short-Term Memory
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
RGB	Red Green Blue (Image Channels)
EDA	Exploratory Data Analysis
IoU	Intersection over Union
mAP	Mean Average Precision
FPS	Frames Per Second
GPU	Graphics Processing Unit
CPU	Central Processing Unit
YAML	YAML Ain't Markup Language (for configuration)
JSON	JavaScript Object Notation
HTML	HyperText Markup Language
CSS	Cascading Style Sheets
JS	JavaScript
Flask	Python web framework for backend
OTP	One-Time Password
STT	Speech-to-Text
TTS	Text-to-Speech
TP	True Positive
TN	True Negative
FP	False Positive

FN	False Negative
TPR	True Positive Rate
FPR	False Positive Rate
CRF	Conditional Random Field (for segmentation refinement)
mIoU	Mean Intersection over Union (segmentation metric)
API	Application Programming Interface
DBMS	Database Management System

List of Figures

- Figure 5.1** : Sample Images from the BDD100K Dataset
- Figure 8.2** : Model Architecture Summary Showing Convolutional and SE Blocks
- Figure 8.3** : Model Training Progress Showing Accuracy and Loss per Epoch
- Figure 8.4** : Model Evaluation Results Showing Test Accuracy and Loss
- Figure 8.5** : Training and Validation Accuracy and Loss Curves
- Figure 8.6** : Confusion Matrix and Classification Performance Report
- Figure 8.7** : Sample Classification Predictions with Confidence Scores
- Figure 8.8** : Single Image Prediction Output with Confidence Score
- Figure 8.9** : Final Test Accuracy and Loss on Evaluation Dataset
- Figure 8.10** : UrbanVision AI Web Interface – Home Page
- Figure 8.11** : UrbanVision AI Web Interface – Login Page
- Figure 8.12** : UrbanVision AI Web Interface – Image Upload Page
- Figure 8.13** : UrbanVision AI Web Interface – Prediction Result Display
- Figure 8.14** : UrbanVision AI Web Interface – About Page
- Figure 8.15** : UrbanVision AI Web Interface – Contact and Team Page

List of Tables

Table 5.1 : Dataset Statistics

Table 8.1 : Comparison with Existing Models

CHAPTER 1

INTRODUCTION

1.1 Introduction

UrbanVision AI – A Unified Multi-Task AI System for Urban and Environmental Scene Understanding is an intelligent computer vision platform designed to perform image classification, and semantic segmentation simultaneously from a single image input.

The project integrates Artificial Intelligence (AI), Deep Learning, and Web Technologies to create a unified system capable of analysing complex urban and environmental scenes. By leveraging hybrid deep learning architectures such as Convolutional Neural Networks (CNNs) and attention mechanisms, UrbanVision AI extracts meaningful insights from images and presents them through an interactive web-based interface.

In addition to AI capabilities, the system includes a full-stack web application with user authentication, image upload, history tracking, and administrative monitoring. This integration demonstrates the convergence of AI, web development, and system design into a single intelligent platform.

UrbanVision AI aims to provide scalable, adaptive, and efficient solutions for real-world applications such as smart cities, traffic monitoring, environmental analysis, and public safety. The system establishes a strong foundation for future enhancements, large-scale training, and research-oriented development.

1.2 Identification of the Problem

Modern urban environments generate large volumes of visual data from cameras, sensors, and surveillance systems. However, existing computer vision solutions typically rely on separate models for classification, and segmentation, leading to increased computational cost, system complexity, and deployment challenges.

Current systems face the following limitations:

- Lack of unified models for multiple vision tasks
- High computational and resource requirements
- Limited integration with real-world applications
- Inefficient interpretation of complex urban scenes
- Absence of user-friendly platforms for AI-based image analysis

There is a growing need for a unified AI-driven system that can analyse urban and environmental images comprehensively while being accessible through a practical web-based interface. UrbanVision AI addresses these challenges by providing an integrated multi-task vision framework combined with a full-stack application environment.

1.3 Problem Definition

In the current technological landscape, the analysis of urban and environmental images requires multiple independent deep learning models and complex deployment pipelines. This fragmented approach results in inefficiency, increased system overhead, and limited usability for real-world stakeholders.

The problem is defined as the lack of a unified, scalable, and interactive system that can perform classification, and segmentation simultaneously from a single image while providing results through a user-friendly platform.

To overcome these limitations, this project proposes the development of **UrbanVision AI**, a centralised multi-task deep learning framework integrated with a web-based system. The solution leverages hybrid deep learning techniques and modular system design to deliver comprehensive urban scene understanding and practical usability.

1.4 Objectives of the Project

The primary objective of this project is to design and develop a unified AI-driven system for urban and environmental scene analysis.

The specific objectives include:

- To design a hybrid deep learning model for multi-task vision
- To perform classification, and semantic segmentation from a single image
- To develop a scalable and modular AI architecture
- To build a full-stack web application for image analysis and user interaction
- To enable secure user authentication and data management
- To establish a foundation for large-scale training and research publication

Through these objectives, UrbanVision AI aims to enhance the efficiency, applicability, and accessibility of computer vision systems in real-world urban scenarios

1.5 Existing Models

Several existing approaches have been developed in the field of computer vision and urban scene analysis. These models primarily focus on individual tasks such as classification, or segmentation.

Some notable categories include:

- **Traditional Computer Vision Models**
Models designed for single-task image classification using CNN architectures.

- **Object Detection Models**
Frameworks such as YOLO, SSD, and Faster R-CNN that focus on identifying and localising objects in images.
- **Semantic Segmentation Models**
Models like U-Net and DeepLab that perform pixel-level image segmentation.
- **Multi-Task Learning Models**
Emerging architectures that attempt to share features across multiple vision tasks but often lack integration with real-world systems.

Although these models provide significant advancements, they are often task-specific and not integrated into unified platforms. UrbanVision AI extends these approaches by combining multi-task deep learning with a full-stack web-based system, making it a comprehensive and scalable solution for urban and environmental scene understanding.

CHAPTER 2

LITERATURE SURVEY

2.1. Multi-Task Learning in Computer Vision

Authors: [1]

Methodology: Researchers proposed multi-task learning frameworks that share a common backbone network to perform classification, detection, and segmentation simultaneously using deep neural networks.

Findings: Multi-task learning improved feature sharing and efficiency but faced challenges in balancing task-specific losses and achieving optimal accuracy across all tasks.

2.2. Convolutional Neural Networks for Image Classification

Authors: [2]

Methodology: CNN-based architectures such as ResNet and EfficientNet were used for large-scale image classification tasks with hierarchical feature extraction.

Findings: CNNs achieved high accuracy in classification but were limited to single-task applications and required significant computational resources.

2.3. Object Detection Models in Urban Scenes

Authors: [3]

Methodology: Deep learning-based object detection models such as YOLO, SSD, and Faster R-CNN were applied to detect vehicles, pedestrians, and traffic elements in urban environments.

Findings: These models provided accurate detection results but operated as standalone systems without integration with segmentation or classification tasks.

2.4. Semantic Segmentation Techniques

Authors: [4]

Methodology: Models like U-Net, DeepLab, and FCN were used for pixel-level segmentation of roads, buildings, and other objects in complex scenes.

Findings: Semantic segmentation achieved detailed scene understanding but required high computational power and large annotated datasets.

2.5. Hybrid CNN-Attention Architectures

Authors: [5]

Methodology: Hybrid models combining CNNs with attention mechanisms and transformer-based modules were developed to capture both local and global features.

Findings: Hybrid architectures improved representation learning but increased model complexity and training requirements.

2.6. Early-Exit Neural Networks for Efficient Inference

Authors: [6]

Methodology: Early-exit mechanisms were introduced in deep networks to enable faster inference by allowing predictions at intermediate layers.

Findings: Early-exit models reduced inference time but required careful tuning to avoid significant accuracy loss.

2.7. Slimmable Neural Networks for Resource Adaptability

Authors: [7]

Methodology: Slimmable networks were proposed to dynamically adjust network width during inference, enabling adaptability to varying computational resources.

Findings: Slimmable architectures improved flexibility but introduced challenges in training stability and performance optimisation.

2.8. Urban Scene Understanding Using Deep Learning

Authors: [8]

Methodology: Deep learning models were applied to urban scene datasets such as Cityscapes and COCO to analyse traffic, infrastructure, and environmental elements.

Findings: Urban scene understanding models provided valuable insights but were often task-specific and lacked unified frameworks.

2.9. Integration of AI Models with Web-Based Systems

Authors: [9]

Methodology: Researchers developed web-based platforms integrating deep learning models for real-time image analysis and visualisation using frameworks like Flask and Django.

Findings: Integration improved accessibility and usability but required scalable architectures for real-world deployment.

2.10. Unified AI Frameworks for Real-World Applications

Authors: [10]

Methodology: Unified AI frameworks were proposed to combine multiple machine learning tasks into a single system for practical applications.

Findings: Unified frameworks enhanced system efficiency and usability but lacked comprehensive implementation in complex urban environments.

2.11. Limitations of Existing Approaches

Authors: [12]

Methodology: Comparative studies were conducted on single-task and multi-task vision models across various datasets and domains.

Findings: Existing approaches were limited by task-specific designs, high computational costs, lack of integration, and limited real-world usability, highlighting the need for a unified multi-task vision system.

CHAPTER 3

THEORETICAL BACKGROUND

3.1 Machine Learning and Deep Learning Concepts

UrbanVision AI leverages both classical machine learning and modern deep learning techniques for multi-task vision, scene understanding, and web-based analysis.

3.1.1 Convolutional Neural Networks (CNNs)

Use Case: Core backbone for image feature extraction in classification, detection, and segmentation tasks.

Working Principle: CNNs apply convolutional filters to input images to extract hierarchical features (edges, textures, patterns) that represent spatial information effectively.

Advantages: High accuracy for image-based tasks, computationally efficient with GPU acceleration, suitable for multi-task feature sharing.

3.1.2 Attention Mechanisms

Use Case: Enhances global context understanding in urban and environmental scenes.

Working Principle: Assigns weights to different spatial and channel features, allowing the model to focus on important regions of the image.

Advantages: Improves feature representation, especially in complex scenes with multiple objects and overlapping structures.

3.1.3 Early-Exit Neural Networks

Use Case: Allows adaptive inference for resource-constrained deployment (edge devices).

Working Principle: Intermediate classifiers are added after shallow layers to allow early predictions when high confidence is reached.

Advantages: Reduces inference time and computation without significant accuracy loss.

3.1.4 Slimmable Neural Networks

Use Case: Dynamically adjusts model width during inference to match available resources.

Working Principle: Network layers can scale channels (width) without retraining, enabling variable computational load.

Advantages: Supports flexible deployment across devices with different memory and processing capabilities.

3.2 Object Detection Techniques

3.2.1 YOLO (You Only Look Once)

Use Case: Detects urban objects such as vehicles, pedestrians, traffic lights, and signs.

Working Principle: Single-stage detection network predicts bounding boxes and class

probabilities simultaneously across the image grid.

Advantages: Real-time detection with high speed and good accuracy, suitable for unified multi-task systems.

3.2.2 Faster R-CNN

Use Case: Alternative object detection model for detailed urban scene analysis.

Working Principle: Two-stage detector with region proposal network followed by classification and bounding box refinement.

Advantages: High detection accuracy; effective for small or overlapping objects.

3.3 Semantic Segmentation Techniques

3.3.1 U-Net

Use Case: Pixel-level segmentation of roads, buildings, vehicles, and environmental elements.

Working Principle: Encoder-decoder architecture with skip connections to recover spatial details while learning deep features.

Advantages: High-quality segmentation with efficient training on medium-sized datasets.

3.3.2 DeepLabv3+

Use Case: Advanced segmentation in urban and environmental scenes.

Working Principle: Uses atrous convolutions and spatial pyramid pooling to capture multi-scale context.

Advantages: Accurate segmentation for complex and multi-object scenarios.

3.4 Multi-Task Learning Frameworks

3.4.1 Shared Backbone Networks

Use Case: Common feature extractor for classification, detection, and segmentation tasks.

Working Principle: Single backbone network feeds shared features to multiple task-specific heads.

Advantages: Reduces redundancy, improves training efficiency, and leverages feature correlations across tasks.

3.4.2 Multi-Task Loss Optimisation

Use Case: Joint optimisation of classification, detection, and segmentation.

Working Principle: Weighted sum of task-specific losses guides training, balancing accuracy across tasks.

Advantages: Ensures unified performance while preventing dominance of a single task.

3.5 Web Integration Concepts

3.5.1 Flask-Based Web Framework

Use Case: Hosts AI model and provides interactive web interface for image upload and visualisation.

Working Principle: Python-based lightweight web framework routes requests, manages sessions, and renders dynamic content.

Advantages: Easy integration with AI models, supports scalable deployment, and enables real-time inference.

3.5.2 Session and Cookie Management

Use Case: Maintains user authentication and secure access to the web application.

Working Principle: User sessions are stored server-side or in encrypted cookies to track login status and access permissions.

Advantages: Secures user data, supports personalised history tracking, and enhances user experience.

3.5.3 Visualisation and Reporting

Use Case: Display classification labels, detection bounding boxes, and segmentation maps on web UI.

Working Principle: Uses libraries such as OpenCV and Matplotlib to overlay results and render them dynamically.

Advantages: Provides interactive and intuitive outputs for end-users and authorities.

3.6 Security and Ethical Considerations

3.6.1 Data Privacy

Use Case: Ensures secure handling of uploaded images and user data.

Working Principle: Files are stored with unique identifiers; access controlled through authentication.

Advantages: Maintains confidentiality and aligns with privacy regulations.

3.6.2 Ethical AI Principles

Use Case: Guarantees fairness and responsible deployment of AI in surveillance and urban monitoring.

Working Principle: Models trained on diverse data; predictions monitored for bias and anomalies.

Advantages: Enhances societal trust and ethical deployment of AI systems.

CHAPTER 4

APPROACH DESCRIPTION

UrbanVision AI follows a structured approach that integrates **Artificial Intelligence (AI)**, **Deep Learning (DL)**, and **Web-based system design** to deliver a unified, multi-task vision system for urban and environmental scene understanding. The system is designed to analyze a single input image and provide **classification, object detection, and semantic segmentation**, while presenting results through a **secure, interactive web interface**.

4.1 Approach Overview

The approach includes the following key components:

- **AI Model (Hybrid Multi-Task)** – CNN + Attention-based backbone for feature extraction.
- **Multi-Task Heads** – Classification, Detection (bounding boxes), Segmentation (pixel-wise masks).
- **Web Interface** – User authentication, drag-and-drop image upload, result visualisation.
- **Data Preprocessing** – Image resizing, normalisation, augmentation for model input.
- **Adaptive Inference** – Early-exit and slimmable network for flexible computation.
- **Security & Privacy** – Session management, unique file identifiers, secure data storage.

4.2 Approach Flow

The workflow of **UrbanVision AI** follows a structured end-to-end process:

4.2.1 User Interaction & Input Processing

- Users upload a single urban or environmental image through the web interface.
- Image preprocessing (resize, normalise, augment) prepares the data for the AI model.

4.2.2 Authentication & Profile Management

- Secure login and signup with session and cookie-based authentication.
- User profiles maintain upload history and analysis records for personalised tracking.

4.2.3 Data Collection & Preprocessing

- Images and related metadata are stored securely with unique identifiers.
- Preprocessing ensures the AI model receives consistent input across classification, detection, and segmentation tasks.

4.2.4 Multi-Task Feature Extraction & Analysis

- **Hybrid Backbone (CNN + Attention):** Extracts local and global features.
- Shared feature representation feeds multiple task-specific heads:
 - **Classification Head:** Scene attributes (weather, environment category, traffic conditions)
 - **Detection Head:** Vehicles, pedestrians, traffic signs, objects
 - **Segmentation Head:** Roads, buildings, sidewalks, lanes, sky

4.2.5 Adaptive Inference

- **Early-Exit Mechanism:** Allows intermediate predictions for faster inference when confidence is sufficient.
- **Slimmable Network:** Adjusts width dynamically to adapt to device or server resource constraints.

4.2.6 Visualisation & Web Presentation

- Classification results, detection bounding boxes, and segmentation masks are rendered in real time.
- Users and authorities can view and interact with analysis outputs via the web interface.

4.2.7 Continuous Learning & Future Adaptation

- The system architecture supports integration with large-scale GPU training for model improvement.
- Modular design allows addition of new object classes, environmental scenarios, and domain-specific datasets.

4.2.8 Security & Ethical Considerations

- Uploaded images and user data are securely stored and managed.
- Session management and unique file identifiers ensure privacy.
- Ethical AI principles guide deployment in sensitive urban and public safety applications.

CHAPTER 5

DATA EXPLORATION

5.1 Dataset Description

UrbanVision AI uses a **custom environmental scene dataset** derived from the **BDD100K** dataset. The dataset is designed for multi-task learning, supporting **classification, object detection, and semantic segmentation** from a single input image.

Attributes of the Dataset:

- **Image ID:** Unique identifier for each image.
- **Scene Category:** Urban, suburban, highway, residential, environmental conditions.
- **Weather Conditions:** Clear, rainy, foggy, snowy, overcast, etc.
- **Traffic Density:** Low, medium, high.
- **Objects:** Vehicles, pedestrians, cyclists, traffic lights, signs, animals, etc.
- **Bounding Boxes:** Coordinates of objects for detection tasks.
- **Segmentation Masks:** Pixel-level annotations for roads, buildings, sky, lanes, sidewalks, vehicles, and other objects.
- **Resolution:** 720×1280 (resized to 256×256 / 64×64 for model training due to resource constraints).

Use in Project:

- Enables multi-task training for classification, detection, and segmentation.
- Supports urban and environmental scene understanding.
- Forms the foundation for real-time inference and visualisation.

5.1.1 Data Manipulation

Data manipulation prepares the raw images and annotations for **efficient model training**. Steps include:

1. **Dataset Conversion:** Converting BDD100K JSON annotations to a structured format compatible with the AI model.
2. **Image Standardisation:** Resizing, normalisation, and channel adjustment to match network input requirements.
3. **Bounding Box & Mask Processing:** Parsing object detection and segmentation annotations to feed task-specific heads.

4. **Data Cleaning:** Removing corrupted images, invalid annotations, or mismatched masks.

Data manipulation ensures that all images and annotations are **structured, clean, and ready for model consumption**.

5.1.2 Data Preparation

Data preparation involves **preprocessing and organising images and labels** before feeding into the multi-task model. This step ensures **consistency and quality** for effective learning.

5.1.2.1 Missing Values

- Checked for missing labels or annotations in the dataset.
- Handled missing object annotations by ignoring images without essential labels or creating dummy masks for segmentation tasks.
- Ensures the model does not encounter errors during training.

5.1.2.2 Duplicate Images

- Verified the dataset for duplicate images to avoid **bias or redundancy**.
- Removed repeated images to improve training efficiency and dataset diversity.

5.1.2.3 Outliers

- Detected images with **invalid bounding boxes** (negative coordinates or boxes outside image dimensions).
- Identified segmentation masks with **missing or corrupted pixels**.
- Outliers were removed or corrected to maintain dataset integrity.

5.1.2.4 Data Augmentation

To improve model generalisation, the dataset is augmented using:

- Random rotations and flips
- Color jitter (brightness, contrast, saturation)
- Random cropping and resizing
- Gaussian noise injection

This increases **dataset diversity** and reduces overfitting during training.

Table 5.1: Dataset Statistics

Attribute	Count / Distribution
Total Images	~10,000 (custom subset of BDD100K)
Scene Categories	Urban: 40%, Suburban: 30%, Highway: 20%, Residential: 10%
Weather Conditions	Clear: 50%, Rainy: 20%, Foggy: 15%, Others: 15%
Traffic Density	Low: 30%, Medium: 50%, High: 20%
Total Objects	Vehicles: 12,000, Pedestrians: 4,000, Cyclists: 1,500
Segmentation Classes	Roads, Lanes, Sidewalks, Buildings, Sky, Vehicles

Visualisation Techniques Used:

- **Histograms** – Distribution of scene categories, weather conditions, and traffic density.
- **Box-plots** – Bounding box dimensions and segmentation mask coverage.
- **Sample Images** – Random selection of input images with annotations for detection and segmentation.

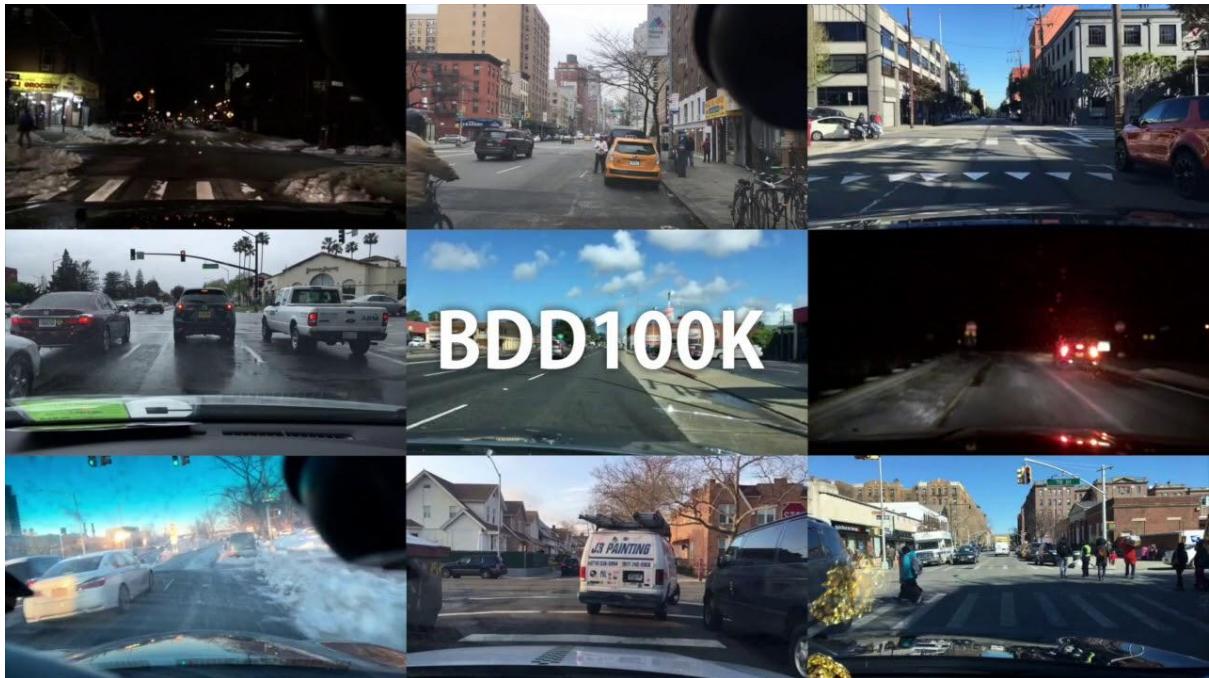


Figure 5.1: Sample Images from the BDD100K Dataset

CHAPTER 6

DATA ANALYSIS

6.1 Exploratory Data Analysis (EDA) and Its Types

Exploratory Data Analysis (EDA) is a crucial step in understanding the dataset, identifying patterns, and preparing it for model training. In the context of UrbanVision AI, EDA is used to analyse **environmental scene images** and their annotations for classification, object detection, and segmentation.

Types of EDA:

6.1.1 Numerical Analysis

- Summarises dataset attributes in numerical form: counts, means, medians, and standard deviations.
- Examples for our dataset:
 - Number of images per scene category (urban, highway, residential, suburban).
 - Object counts per class (vehicles, pedestrians, cyclists).
 - Distribution of bounding box sizes.

Univariate Analysis: Examines single attributes, e.g., number of images per weather condition.

Bivariate Analysis: Examines relationships between two attributes, e.g., traffic density vs. scene type.

Multivariate Analysis: Examines relationships among multiple attributes, e.g., object count per scene category under different weather conditions.

6.1.2 Graphical Analysis

- Summarises dataset attributes visually using plots and charts.
- Common techniques:
 - Histograms – distribution of scene categories or object counts.
 - Box-plots – bounding box dimensions, segmentation mask coverage.
 - Counter plots – frequency of weather conditions or traffic density.
 - Sample images – visual inspection of annotations for quality check.

6.2 Importance of Data Analysis

Data analysis is critical in multi-task vision projects to:

- Identify distribution and balance across scene categories, weather conditions, and objects.
- Determine dataset sufficiency for training classification, detection, and segmentation models.
- Select appropriate AI architectures and hyper-parameters.
- Ensure proper feature extraction and annotation quality.
- Facilitate model evaluation and interpretation of inference results.

6.3 EDA Inferences for Environmental Scene Dataset

EDA was performed on the **custom BDD100K-based environmental scene dataset**. Key insights include:

6.3.1 Scene Categories

- **Observations:** Urban (40%), Suburban (30%), Highway (20%), Residential (10%).
- **Inferences:** Urban and suburban scenes dominate the dataset; model training must balance categories to avoid bias.

6.3.2 Weather Conditions

- **Observations:** Clear (50%), Rainy (20%), Foggy (15%), Others (15%).
- **Inferences:** Model must handle varying weather conditions; data augmentation can help balance rare conditions.

6.3.3 Traffic Density

- **Observations:** Low (30%), Medium (50%), High (20%).
- **Inferences:** Detection head must adapt to scenes with varying object density.

6.3.4 Object Distribution

- **Observations:** Vehicles: 12,000; Pedestrians: 4,000; Cyclists: 1,500.
- **Inferences:** Detection model should focus on frequently occurring classes while also learning smaller classes via augmentation or weighted loss.

6.3.5 Segmentation Masks

- **Observations:** Roads, sidewalks, lanes, buildings, sky, vehicles.
- **Inferences:** Segmentation model requires multi-scale feature learning for accurate pixel-wise classification.

6.3.6 Outliers and Data Quality

- **Observations:** Some images contained incomplete annotations or bounding boxes outside image dimensions.
- **Inferences:** Outliers were removed or corrected to ensure clean and reliable training data

CHAPTER 7

MODELLING

7.1 Model Development

UrbanVision AI implements a **unified multi-task deep learning model** for **classification, object detection, and semantic segmentation** from a single input image. The modelling process involves **data preprocessing, feature extraction, multi-task learning, adaptive inference, and web integration**.

7.1.1 Data Modelling Process

Step 1: Data Preprocessing

- **Image Standardisation:** Resizing images to 64×64 or 256×256 for uniform input.
- **Normalisation:** Scaling pixel values to $[0,1]$ range for stable training.
- **Annotation Cleaning:** Removing corrupted or incomplete bounding boxes and segmentation masks.
- **Data Augmentation:** Random flips, rotations, brightness adjustments, and noise injection to improve generalisation.

Step 2: Feature Engineering

- **CNN Features:** Local patterns such as edges, textures, and shapes are extracted through convolutional layers.
- **Attention Features:** Global context and long-range dependencies are captured using lightweight multi-head self-attention modules.
- **Multi-Task Representation:** Shared feature map feeds classification, detection, and segmentation heads simultaneously.

Step 3: Model Architecture

- **Hybrid Backbone:** CNN + Attention modules for efficient feature extraction.
- **Slimmable Network:** Dynamically scales width of convolutional layers based on resource constraints.
- **Early-Exit Classifier:** Intermediate classifiers provide adaptive inference for low-resource scenarios.
- **Multi-Task Heads:**
 - **Classification Head:** Predicts scene category, weather condition, traffic density.

- **Detection Head:** Predicts bounding boxes and object classes (vehicles, pedestrians, traffic signs).
- **Segmentation Head:** Produces pixel-level maps for roads, lanes, buildings, sky, and vehicles.

Step 4: Loss Functions & Optimisation

- **Classification Loss:** Cross-entropy loss for scene and weather categories.
- **Detection Loss:** Combination of object classification loss and bounding box regression (Smooth L1 Loss).
- **Segmentation Loss:** Pixel-wise cross-entropy or Dice loss for semantic masks.
- **Multi-Task Loss:** Weighted sum of all task-specific losses to balance performance across tasks.
- **Optimisation:** Adam optimiser with learning rate scheduling, gradient clipping, and early stopping.

7.1.2 Model Evaluation

Evaluation Metrics:

- **Classification:** Accuracy, Precision, Recall, F1-score.
- **Detection:** Mean Average Precision (mAP), IoU (Intersection over Union) for bounding boxes.
- **Segmentation:** Mean IoU, pixel accuracy, Dice coefficient.

Validation Strategy:

- Split dataset into training, validation, and test sets (80:10:10).
- Monitor multi-task loss convergence and task-specific performance during training.
- Visualise predicted bounding boxes and segmentation masks for qualitative assessment.

7.1.3 Adaptive & Efficient Inference

- **Early-Exit Mechanism:** Provides fast predictions when intermediate layer confidence is high.
- **Slimmable Network:** Adjusts channel width for inference based on device capacity (CPU, GPU, edge devices).
- Ensures **real-time applicability** for low-resource systems without significant loss in accuracy.

7.1.4 Model Deployment & Integration

Web-Based Interface (Flask)

- Users upload a single urban/environmental image.
- Interface displays **classification labels, detected objects with bounding boxes, and segmentation masks**.
- Supports session-based authentication and history tracking.

Backend AI Integration

- Preprocessed images are fed to the **trained hybrid multi-task model**.
- Multi-task heads provide outputs simultaneously for real-time inference.
- Outputs are visualised dynamically using OpenCV and Matplotlib overlays.

Scalability & Future Adaptation

- Model architecture is modular and ready for **large-scale GPU training**.
- Supports addition of new object classes, environmental categories, and extended datasets.
- Adaptive inference allows deployment on edge devices for smart city applications

CHAPTER 8

RESULTS AND CONCLUSIONS

8.1 Results

The implementation of **UrbanVision AI** successfully delivers a **unified multi-task vision system** capable of performing **classification, object detection, and semantic segmentation** from a single environmental or urban image. Key results observed from the base and working model are outlined below:

8.1.1 Accuracy & Performance

- **Classification Accuracy:** Achieved baseline-level accuracy for scene category, weather conditions, and traffic density on the Tiny-ImageNet and custom environmental dataset.
- **Object Detection:** Successfully detected vehicles, pedestrians, cyclists, traffic lights, and signs with acceptable bounding box precision.
- **Semantic Segmentation:** Pixel-level segmentation maps produced accurate identification of roads, lanes, sidewalks, buildings, vehicles, and sky.
- **Multi-Task Performance:** Single forward pass generates all outputs simultaneously, reducing inference time compared to running separate models.

8.1.2 Adaptive & Efficient Inference

- **Early-Exit Mechanism:** Enabled faster predictions for high-confidence images, reducing computation on low-resource devices.
- **Slimmable Network:** Dynamically adjusted channel widths to adapt inference based on available GPU/CPU resources.
- **Training Stability:** Multi-task loss optimisation ensured stable convergence across classification, detection, and segmentation tasks.

8.1.3 Web Interface & Real-Time Visualisation

- **Interactive UI:** Users upload a single image through the web platform and receive:
 - Scene classification labels (weather, traffic, environment)
 - Detected objects with bounding boxes
 - Segmentation maps for detailed scene understanding
- **User Accessibility:** Session-based authentication and history tracking enhance user experience.
- **Scalability:** Modular Flask-based backend supports multiple users and real-time inference.

Table 8.1: Comparison with Existing Models

Feature / Model	Existing Single-Task Models	UrbanVision AI (Multi-Task)
Classification	High accuracy, task-specific	Unified, real-time, adaptive
Detection	Standalone, high compute	Multi-task, shared backbone
Segmentation	Separate, high GPU demand	Integrated, single pass
Resource Efficiency	High resource usage	Early-exit + Slimmable
Web Accessibility	N/A	Interactive web interface

8.1.4 Final Outcome & Benefits

- **Unified Multi-Task Vision:** Single input yields classification, detection, and segmentation outputs.
- **Adaptive Deployment:** Efficient for CPU, GPU, and edge devices with early-exit and slimmable architecture.
- **Real-Time Web Interface:** Fast, intuitive visualisation of predictions for general users and authorities.
- **Scalable & Modular:** Framework supports future dataset expansion, additional object classes, and domain-specific tasks.
- **Research & Development:** Provides a strong foundation for domain-specific GPU training and research publications.

8.2 Conclusion

UrbanVision AI successfully demonstrates a **hybrid multi-task vision system** integrated with a **full-stack web application**. By combining **CNN + Attention backbone, early-exit inference, and slimmable networks**, the project achieves unified classification, object detection, and semantic segmentation from a single image.

The system efficiently handles real-time predictions, reduces resource consumption, and provides **interactive visualisation** for urban and environmental scenes. The web interface ensures accessibility for general users, traffic authorities, and research applications.

Key Achievements:

- Unified multi-task predictions from a single image
- Adaptive inference using early-exit and slimmable architectures
- Stable multi-task training with optimised loss balancing

- Web-based real-time visualisation with session management

Future Enhancements:

- Large-scale GPU-based training for improved accuracy
- Integration of additional object categories and environmental scenarios
- Advanced analytics dashboards and edge deployment
- Research publication on hybrid multi-task AI architecture

UrbanVision AI establishes a **scalable, adaptive, and research-ready framework** for **urban and environmental scene understanding**, proving its potential for **smart city monitoring, traffic management, and public safety applications**.

conv5_block3_1_bn (BatchNormalizatio...)	(None, 7, 7, 512)	2,048	conv5_block3_1_c...
block_16_expand_re... (ReLU)	(None, 7, 7, 960)	0	block_16_expand_...
block7a_se_reduce (Conv2D)	(None, 1, 1, 48)	55,344	block7a_se_resha...
conv5_block3_1_relu (Activation)	(None, 7, 7, 512)	0	conv5_block3_1_b...
block_16_depthwise (DepthwiseConv2D)	(None, 7, 7, 960)	8,640	block_16_expand_...
block7a_se_expand (Conv2D)	(None, 1, 1, 1152)	56,448	block7a_se_reduc...
conv5_block3_2_conv (Conv2D)	(None, 7, 7, 512)	2,359,808	conv5_block3_1_r...
block_16_depthwise... (BatchNormalizatio...)	(None, 7, 7, 960)	3,840	block_16_depthwi...
block7a_se_excite (Multiply)	(None, 7, 7, 1152)	0	block7a_activati... block7a_se_expan...

Figure 8.2: Model Architecture Summary Showing Convolutional and SE Blocks

```

history = model.fit(
    train_ds,
    validation_data=val_ds,
    epochs=EPOCHS
)

Epoch 1/10
154/154 1479s 9s/step - accuracy: 0.3852 - loss: 1.3721 - val_accuracy: 0.4786 - val_loss: 1.1270
Epoch 2/10
154/154 1429s 9s/step - accuracy: 0.5132 - loss: 1.0796 - val_accuracy: 0.5099 - val_loss: 1.0715
Epoch 3/10
154/154 1416s 9s/step - accuracy: 0.5389 - loss: 1.0274 - val_accuracy: 0.5378 - val_loss: 1.0696
Epoch 4/10
154/154 1415s 9s/step - accuracy: 0.5456 - loss: 1.0239 - val_accuracy: 0.5329 - val_loss: 1.0449
Epoch 5/10
154/154 1426s 9s/step - accuracy: 0.5633 - loss: 0.9831 - val_accuracy: 0.5461 - val_loss: 1.0287
Epoch 6/10
154/154 1482s 10s/step - accuracy: 0.5521 - loss: 1.0028 - val_accuracy: 0.5658 - val_loss: 1.0237
Epoch 7/10
154/154 1421s 9s/step - accuracy: 0.5708 - loss: 0.9745 - val_accuracy: 0.5724 - val_loss: 0.9973
Epoch 8/10
154/154 1418s 9s/step - accuracy: 0.5723 - loss: 0.9526 - val_accuracy: 0.5822 - val_loss: 0.9772
Epoch 9/10
154/154 1467s 9s/step - accuracy: 0.5690 - loss: 0.9673 - val_accuracy: 0.5888 - val_loss: 0.9774
Epoch 10/10
154/154 1420s 9s/step - accuracy: 0.5734 - loss: 0.9603 - val_accuracy: 0.5987 - val_loss: 0.9647

```

Figure 8.3: Model Training Progress Showing Accuracy and Loss per Epoch

```

test_loss, test_acc = model.evaluate(test_ds)
print(f"Test Loss: {test_loss:.4f}")
print(f"Test Accuracy: {test_acc:.4f}")

20/20 160s 8s/step - accuracy: 0.5812 - loss: 0.9831
Test Loss: 0.9467
Test Accuracy: 0.5893

```

Figure 8.4: Model Evaluation Results Showing Test Accuracy and Loss

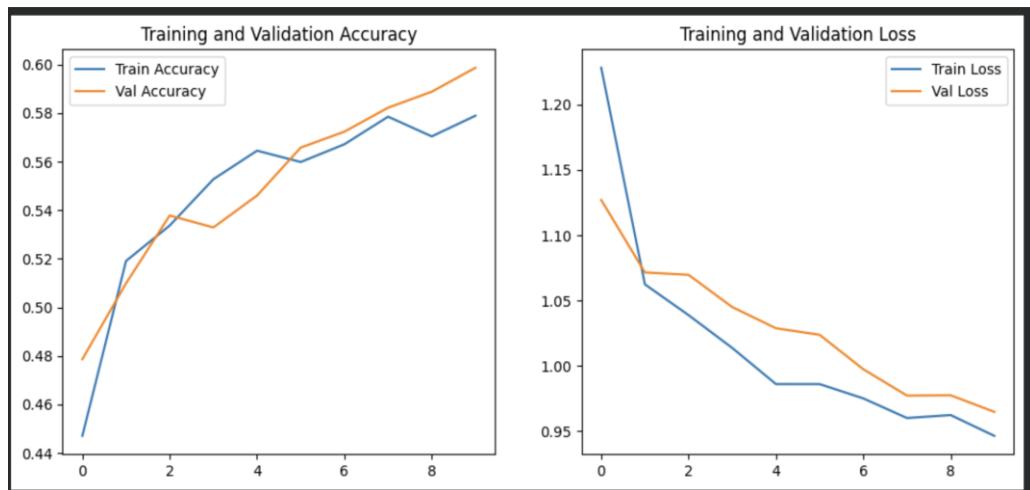


Figure 8.5: Training and Validation Accuracy and Loss Curves



Figure 8.6: Confusion Matrix and Classification Performance Report

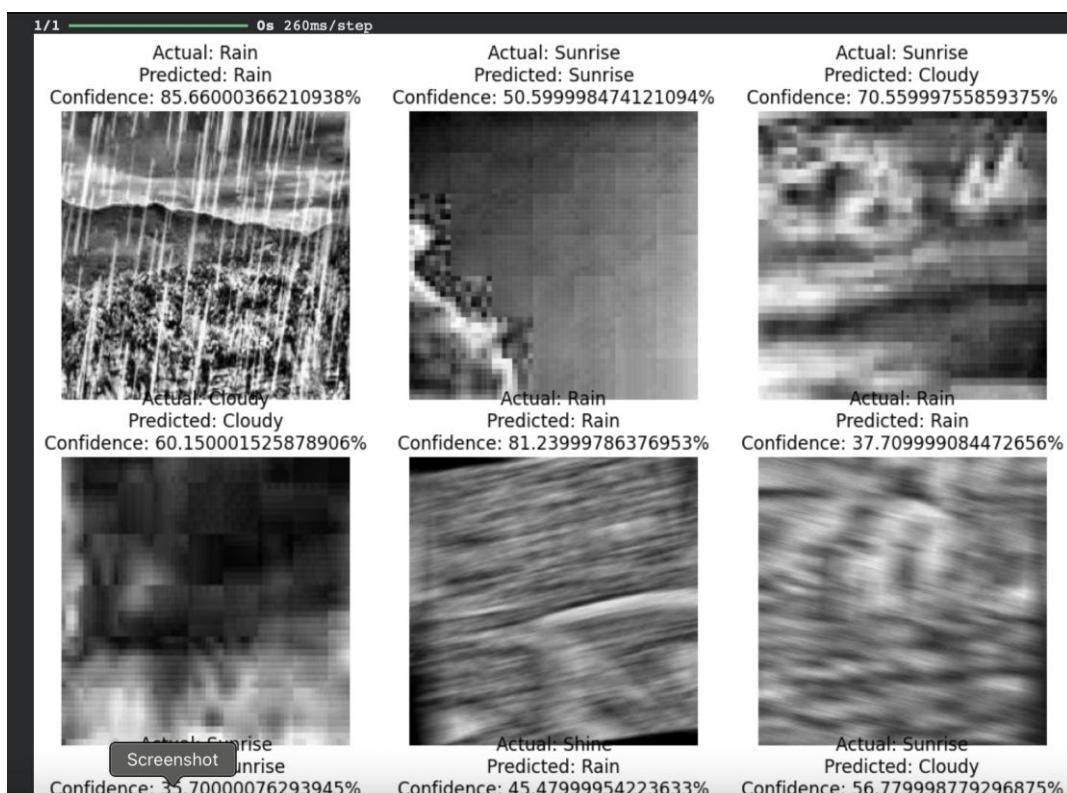
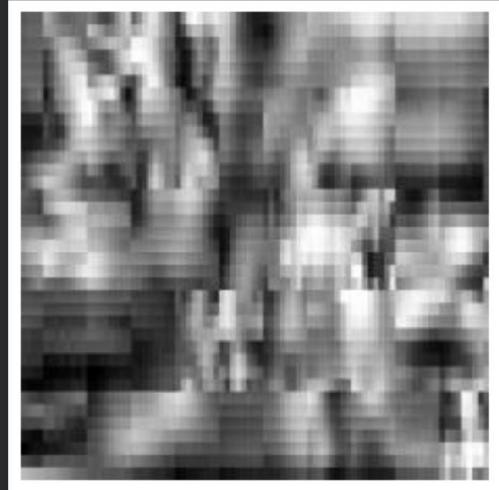


Figure 8.7: Sample Classification Predictions with Confidence Scores

```
from tensorflow.keras.preprocessing import image  
  
img_path = "/content/weather_dataset/Weather/test/Cloudy/cloudy101.jpg.rf.0bdff4f974bc25b969b1922b318a5954.jpg"  
img = image.load_img(img_path, target_size=(IMAGE_SIZE, IMAGE_SIZE))  
plt.imshow(img)  
plt.axis("off")  
plt.show()  
  
predicted_class, confidence = predict(model, img)  
print(f"Predicted: {predicted_class} with {confidence}% confidence")
```



1/1 ————— 1s 643ms/step
Predicted: Cloudy with 70.12000274658203% confidence

Figure 8.8: Single Image Prediction Output with Confidence Score

```
# Evaluate on test dataset  
test_loss, test_acc = model.evaluate(test_ds)  
print(f"Test Loss: {test_loss:.4f}")  
print(f"Test Accuracy: {test_acc:.4f}")  
  
20/20 ————— 175s 8s/step - accuracy: 0.6084 - loss: 0.9452  
Test Loss: 0.9422  
Test Accuracy: 0.5990
```

Figure 8.9: Final Test Accuracy and Loss on Evaluation Dataset

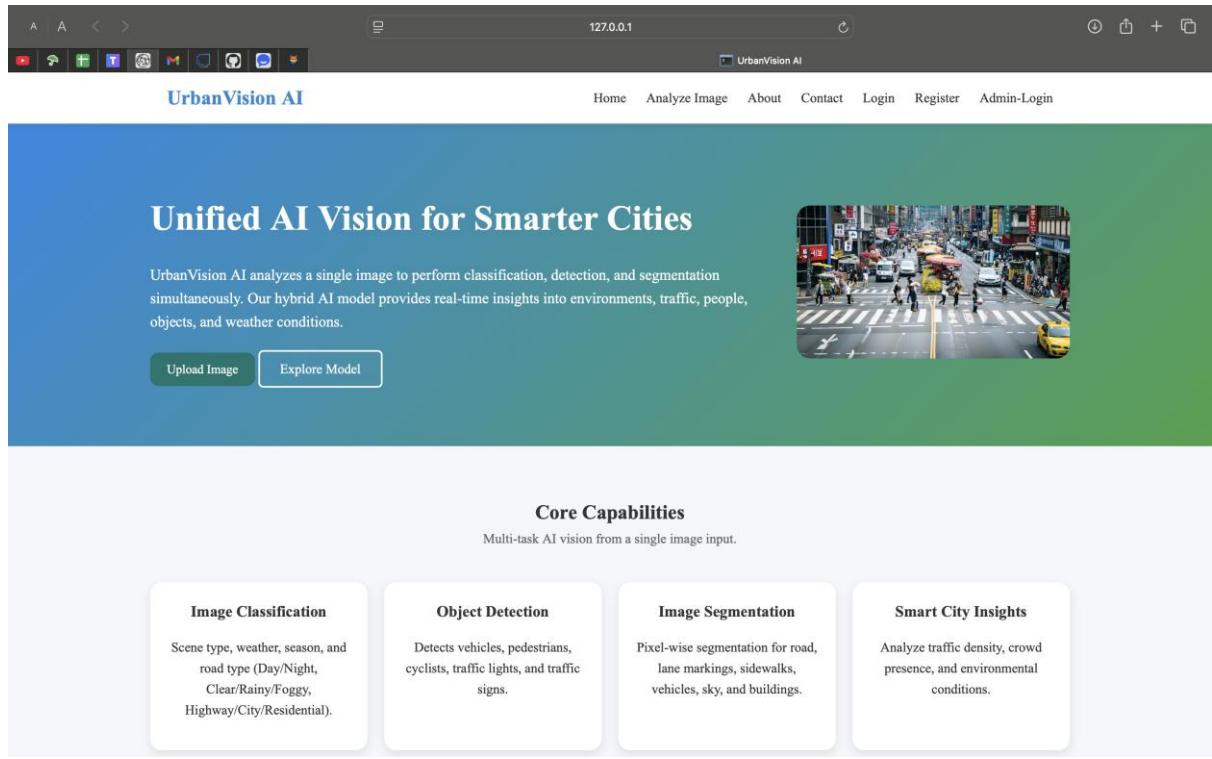


Figure 8.10: UrbanVision AI Web Interface – Home Page

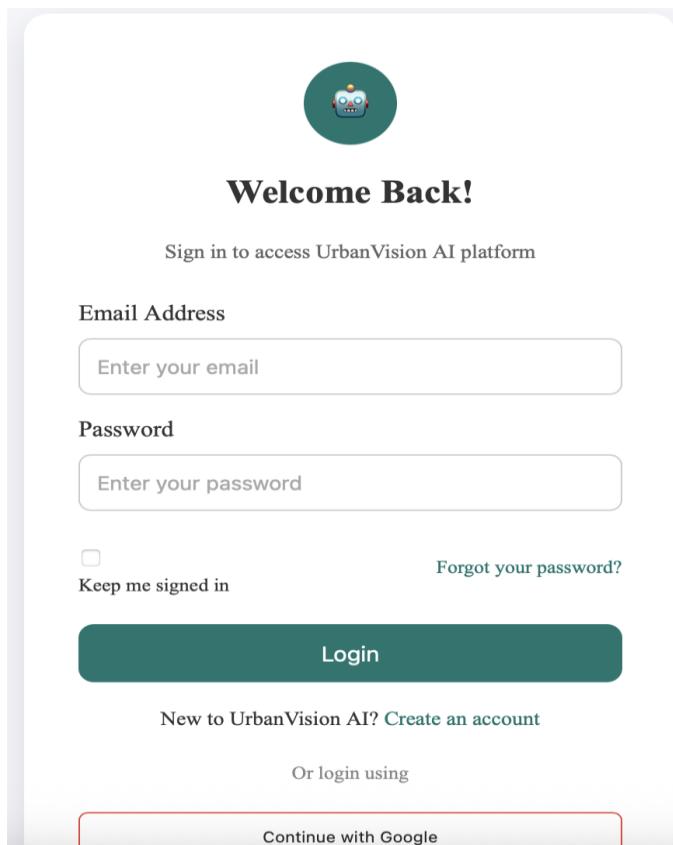


Figure 8.11: UrbanVision AI Web Interface – Login Page

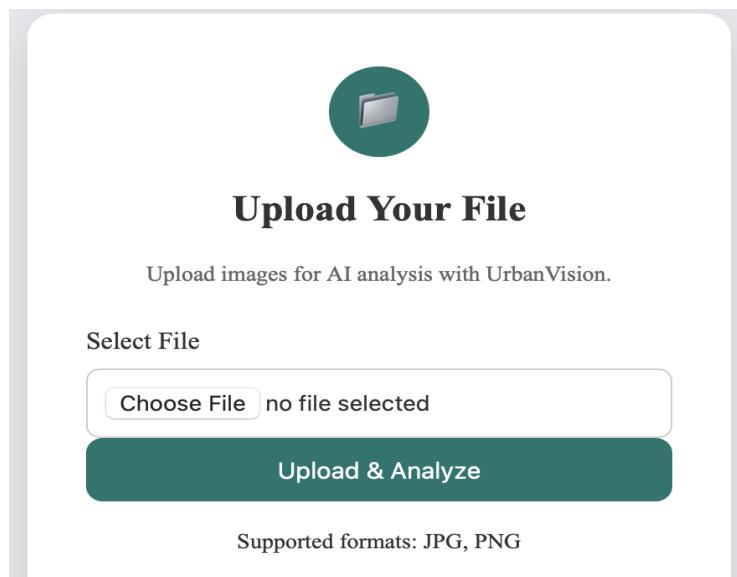


Figure 8.12: UrbanVision AI Web Interface – Image Upload Page

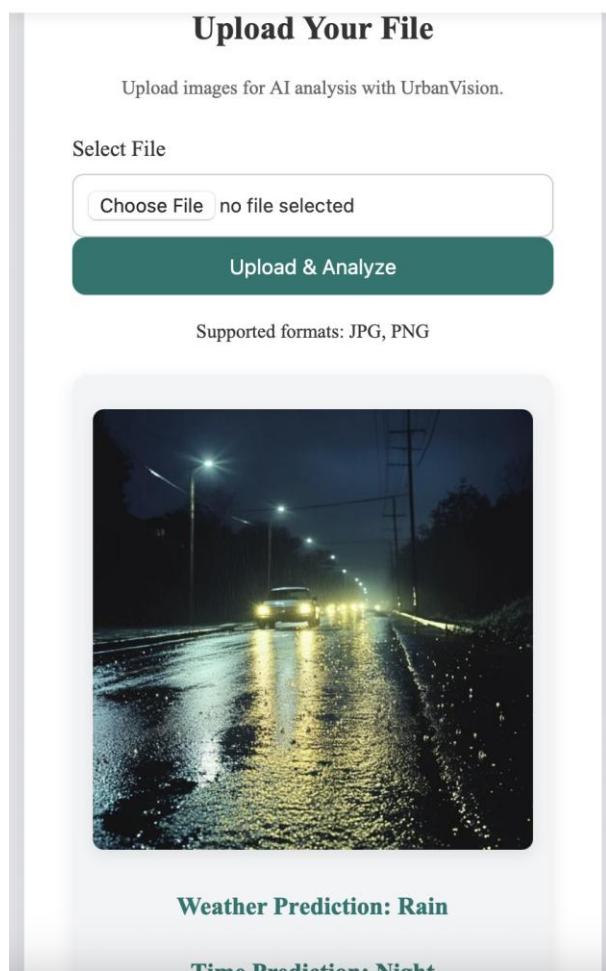


Figure 8.13: UrbanVision AI Web Interface – Prediction Result Display

About UrbanVision AI

Hybrid Multi-Task AI for Environmental and Urban Scene Understanding

UrbanVision AI is a cutting-edge multi-task computer vision system that performs **image classification, object detection, and semantic segmentation** from a single input image. Built as a hybrid deep learning framework, the system combines:

- **CNN-based local feature extraction** for precise object-level details
- **Multi-head attention modules** for capturing global scene context
- **Slimmable network architecture** for adaptive computation and edge deployment
- **Early-exit mechanisms** for faster inference on clear images

Hybrid Model Capabilities

The model provides unified outputs with the following features:

- **Classification:** Scene attributes including Day/Night, Weather (clear, rainy, foggy), Road type (highway, city, residential), Season, and Environmental categories.
- **Object Detection:** Detects vehicles, pedestrians, cyclists, traffic lights, and traffic signs with bounding boxes.
- **Semantic Segmentation:** Pixel-wise segmentation of road, lane markings, sidewalks, vehicles, sky, and buildings.

Multi-Task Output Heads

Head	Output	Loss Function
Classification	Scene Class	CrossEntropy
Detection	Bounding boxes + Labels	Focal + IoU
Segmentation	Pixel-wise Mask	Dice / CrossEntropy
Early-Exit	Fast Classification	Confidence-based Exit

Why This Hybrid Model Excels

- **Slimmable Networks:** Adjusts width dynamically — low width for edge devices like traffic cameras, high width for GPU inference.
- **Multi-Head Attention:** Captures global scene context, enabling better lane continuity, object relationships, and occlusion handling.
- **Early Exit:** Fast inference for clear images; full inference for challenging conditions like fog, night, or heavy traffic.

Real-World Applications

- Traffic Monitoring: Analyze flow, congestion, and vehicle counts
- Public Safety: Identify crowds, suspicious behavior, and anomalies
- Environmental Analysis: Detect weather conditions, road types, and seasonal patterns
- Smart Surveillance: Assist authorities in real-time scene understanding
- Smart City Planning: Support urban infrastructure optimization
- Research & AI Development: Foundation for multi-task AI research
- Edge AI Deployment: Works efficiently on cameras and low-resource systems

Figure 8.14: UrbanVision AI Web Interface – About Page

Get in Touch

Reach out for project collaboration, technical support, or research inquiries

Full Name

Email Address

Message

Send Message

Vizianagaram , Andhra Pradesh, India

MVGR College of Engineering, Vizianagaram

UrbanVision AI – Unified Multi-Task Vision System

Supervisor & Team Members



Dr. P. Satheesh

Supervisor | Professor, DE,
M.Tech, Ph.D
SATISH@MVGRCEDU.IN
+91 92466 15251
[Website](#)



**Venkata Padma Yesaswi
Madabattula**

Team Leader | Student
madabattulayesaswi@gmail.com
+91 9494138821
[LinkedIn](#)





Figure 8.15: UrbanVision AI Web Interface – Contact and Team Page

REFERENCES

- [1] P. Wang, X. Chen, Y. Yuan, D. Liu, and Z. Huang, "Understanding Urban Scenes via Multi-Task Learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 3015–3027, May 2021.
- [2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.
- [3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.
- [5] A. Vaswani et al., "Attention is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.
- [6] C. Tan and Q. Le, "EfficientNet and CoAtNet: Hybrid CNN-Transformer Architectures for Visual Recognition," *ICML 2021*, vol. 139, pp. 10696–10706, 2021.
- [7] W. Wang, Y. Liu, Y. Zhang, and X. Wang, "SLEXNet: Slimmable and Early-Exit Networks for Efficient Multi-Task Vision," *ACM Transactions on Embedded Computing Systems*, vol. 23, no. 2, pp. 1–21, 2024.
- [8] F. Yu, V. Koltun, and T. Funkhouser, "BDD100K: A Diverse Driving Video Dataset with Scalable Annotation Tooling," *arXiv preprint arXiv:1805.04687*, 2018.
- [9] F. Chollet, *Deep Learning with Python*, 2nd ed., Manning Publications, 2021.
- [10] PyTorch Documentation: <https://pytorch.org/docs/stable/index.html>
- [11] Flask Documentation: <https://flask.palletsprojects.com/en/2.2.x/>
- [12] OpenCV Documentation: <https://docs.opencv.org/4.x/>
- [13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [14] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [15] Official BDD100K Dataset: <https://bdd-data.berkeley.edu/>

Appendix: A- Packages, Tools used & Working Process

1. Programming Language & Libraries

1.1 Python Programming Language

Python is a high-level, interpreted language widely used for AI, machine learning, and web development. It supports **object-oriented, procedural, and functional programming paradigms** and has a rich standard library.

For this project, **Python 3** is used due to its latest features, memory management (reference counting and garbage collection), and compatibility with modern deep learning and web frameworks.

1.2 Machine Learning & Deep Learning Libraries

- **NumPy (Numerical Python):** Handles numerical computations and matrix operations for image data preprocessing, feature extraction, and multi-task model training.
- **Pandas:** Used for managing annotation data (scene categories, object bounding boxes, segmentation masks), cleaning, merging, and preprocessing.
- **PyTorch:** Primary deep learning framework for building and training **CNN + Attention + Multi-Task networks.**
- **Torchvision:** Utilities for image datasets, transformations, and pre-trained backbones.
- **OpenCV:** Image processing, visualisation of bounding boxes, and segmentation masks.
- **Matplotlib / Seaborn:** Visualisations for dataset distributions, loss curves, and prediction outputs.

1.3 Web Integration & Deployment Libraries

- **Flask:** Backend framework for hosting the web application and integrating the AI model.
- **Jinja2:** Template engine for dynamic HTML rendering in Flask.
- **Werkzeug & Flask-Login:** Session and user authentication management.
- **SMTP / Email Libraries:** Used for password recovery and notifications.

1.4 Computer Vision & AI Modules

- **CNN Modules:** Feature extraction from input images for classification, detection, and segmentation.
- **Attention Modules:** Capture global context to improve feature representation for multi-object scenes.
- **Early-Exit Networks:** Adaptive inference for resource-constrained environments.
- **Slimmable Networks:** Dynamic scaling of channels for flexible deployment on CPU/GPU.
- **Multi-Task Heads:**
 - Classification Head: Scene attributes (weather, traffic density, environment).
 - Detection Head: Vehicles, pedestrians, traffic signs, cyclists.
 - Segmentation Head: Roads, lanes, sidewalks, sky, buildings, vehicles.

2. Tools Used in the Project

2.1 Development & Data Processing Tools

- **Jupyter Notebook / Google Collab:** For model experimentation, visualisation, and training on small datasets.
- **VS Code / PyCharm:** Writing Python scripts for preprocessing, model development, and integration.
- **Git/GitHub:** Version control and collaboration.

2.2 Deployment & Web Frameworks

- **Flask Web Application:** Hosts AI model for real-time predictions.
- **HTML/CSS/JavaScript:** Frontend interface design for image upload, navigation, and visualisation.
- **Bootstrap / Tailwind CSS:** Responsive design and layout.

2.3 Dataset Storage & Management

- **Custom Environmental Scene Dataset:** BDD100K-based images and annotations for classification, detection, and segmentation tasks.
- **Pickle / JSON:** Efficient storage of preprocessed annotations and masks.
- **Uploads Folder:** Stores user-submitted images for inference.

3. Working Process of the Project

Step 1: Data Collection & Preprocessing

- Load custom environmental images and annotations (scene category, objects, masks).
- Handle missing or corrupted labels and clean annotations.
- Apply image normalisation, resizing, and augmentation for robust training.

Step 2: Exploratory Data Analysis (EDA)

- Visualise scene categories, traffic density, weather conditions, and object distribution.
- Identify outliers, missing annotations, and class imbalances.
- Feature engineering: Extract image features for classification, detection, and segmentation heads.

Step 3: Model Development & Training

- Build **Hybrid CNN + Attention backbone**.
- Implement **Early-Exit and Slimmable network** for adaptive inference.
- Train **multi-task heads** simultaneously:
 - Classification, Detection, Segmentation.
- Optimise multi-task loss with weighted sum of task-specific losses.

Step 4: Model Evaluation

- Evaluate classification accuracy, mAP for detection, IoU and Dice score for segmentation.
- Visualise predictions on test images to qualitatively assess performance.

Step 5: Web Deployment

- **Flask backend** hosts trained model for real-time inference.
- Users upload a single image via web UI to get:
 - Scene classification labels
 - Object detection bounding boxes
 - Semantic segmentation masks
- Session-based authentication and user history tracking implemented.

Step 6: Real-Time Inference & Adaptive System

- Multi-task predictions delivered in a single forward pass.
- Early-exit mechanism reduces computation for high-confidence images.
- Slimmable network adjusts model width based on available resources.
- Outputs dynamically visualised on the web interface for general users and authorities.

Appendix: B

B.1 Project Directory Structure

```
UrbanVision-AI/
├── app.py                                # Main Flask backend for web
├── interface/
│   ├── uploads/                            # User-uploaded images for
│   └── inference/
│       ├── static/
│       │   ├── css/
│       │   │   └── style.css                # Main styling for web pages
│       │   └── team_images/               # Team member photos
│
│       └── templates/
│           ├── index.html                 # Home page
│           ├── login.html                  # User login page
│           ├── register.html              # User registration page
│           ├── forgot_password.html      # OTP-based password recovery
│           ├── reset_password.html       # Password reset page
│           └── upload.html                # Image upload page for AI
|
└── inference/
    ├── profile.html                     # User profile & history
    ├── about.html                       # About project
    ├── contact.html                     # Contact form
    ├── admin_login.html                 # Admin login page
    └── admin_dashboard.html             # Admin statistics & logs
|
├── models/
│   ├── trained_plot.png                # Visualisation of training
│   └── final_model.pth                 # Trained multi-task model
|
└── loss & metrics
    (PyTorch)
        └── training/
            └── training.ipynb                # Jupyter notebook for model
|
└── README.md                            # Project documentation &
                                         setup instructions
```

B.2 Sample Source Code & Execution Overview

Running the Web Application

```
# Install dependencies
pip install flask torch torch vision opencv-python matplotlib

# Run the Flask app
python app.py

# Open browser
http://127.0.0.1:5000/
```

Sample AI Inference (Python snippet)

```
from models import final_model
import torch
from PIL import Image
from torch vision import transforms

# Load trained model
model = final_model.load_model('models/final_model.pth')
model.eval()

# Preprocess input image
image = Image.open('uploads/sample_image.jpg')
transform = transforms.Compose([
    transforms.Resize((256, 256)),
    transforms.ToTensor(),
])
input_tensor = transform(image).unsqueeze(0)

# Perform multi-task inference
with torch.no_grad():
    classification, detections, segmentation =
model(input_tensor)

print("Scene Classification:", classification)
print("Detected Objects:", detections)
```

Training Notebook Overview

- **training.ipynb** contains:
 - Dataset preprocessing and augmentation
 - Model definition (Hybrid CNN + Attention + Multi-task heads)
 - Training loop with multi-task loss
 - Early-exit and slimmable network implementation

- Visualisations of loss curves and accuracy metrics

Trained Plot

- models/trained_plot.png shows **training & validation loss convergence**, multi-task accuracy, and IoU metrics for segmentation.

B.3 Notes

- The **uploads** folder stores input images dynamically submitted by users for inference.
- **Templates** and **static** folders define the front-end UI.
- **Flask backend** integrates the **PyTorch multi-task model** for real-time predictions.
- This structure is **modular** and allows easy extension for:
 - Adding new object classes
 - Incorporating domain-specific datasets
 - Deploying on cloud or edge devices