SAPIENZA
UNIVERSITÀ DI ROMA

DIAG
Dipartimento di Ingegneria
informatica, automatica e gestionale
Antonio Ruberti

PEDESTRIAN CROSSING PREDICTION FOR ENHANCED AUTONOMOUS DRIVING SAFETY

● PROJECT PRESENTATION

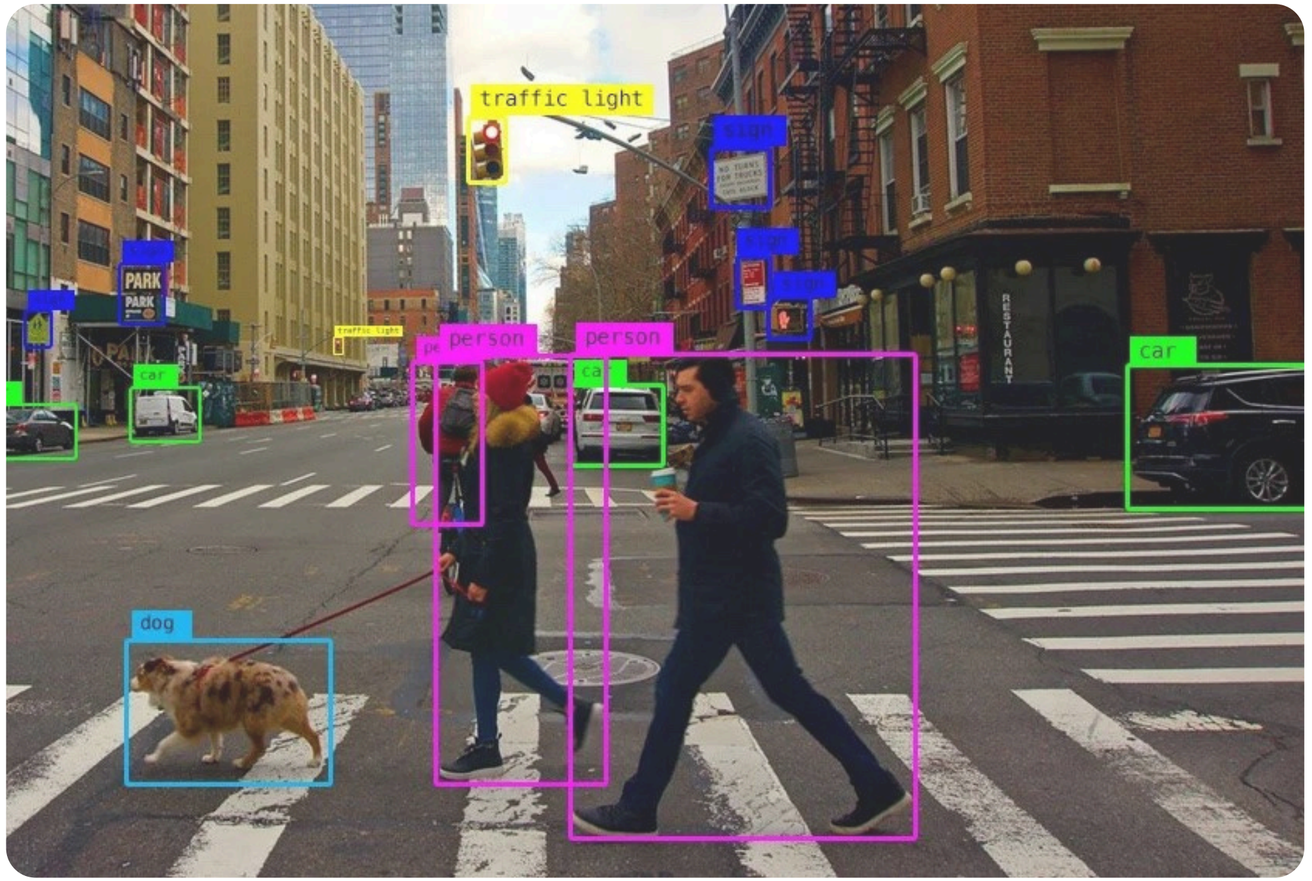
Gianmarco Donnesi
Matr. n. 2152311

Michael Corelli
Matr. n. 1938627

A.Y. 23/24

PROBLEM AND MOTIVATION

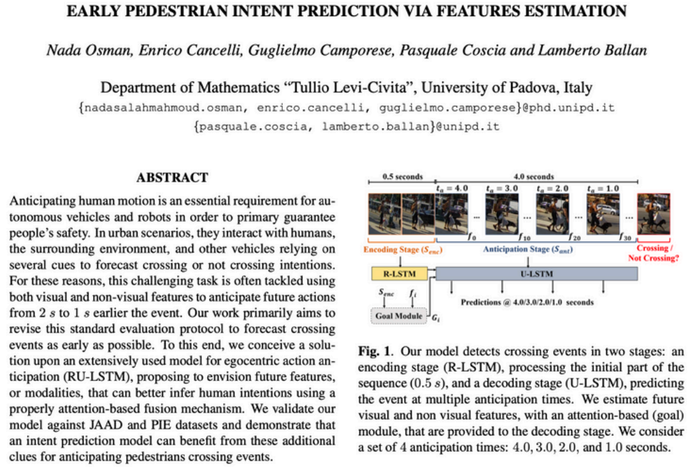
- **Pedestrian intention estimation** in urban environments is crucial for autonomous driving systems.
- **Predicting** whether a pedestrian will cross the street in real-time is complex due to the unpredictable nature of human actions.
- Enhancing **safety** and **reliability** in autonomous driving by accurately predicting pedestrian behavior.



RELATED WORKS

1) *Early Pedestrian Intent Prediction via Features Estimation (Nada Osman et al.)*

- Attention-based fusion mechanism to integrate visual and non-visual features.
- **Model**: modified version of RU-LSTM (Rolling-Unrolling LSTM) to anticipate future actions.
- **Datasets**: Validated on JAAD and PIE datasets.



EARLY PEDESTRIAN INTENT PREDICTION VIA FEATURES ESTIMATION

Nada Osman, Enrico Cancelli, Guglielmo Camporese, Pasquale Coscia and Lamberto Ballan

Department of Mathematics "Tullio Levi-Civita", University of Padova, Italy
{nadasalahmahmoud.osman, enrico.cancelli, guglielmo.camporese}@phd.unipd.it
{pasquale.coscia, lamberto.ballan}@unipd.it

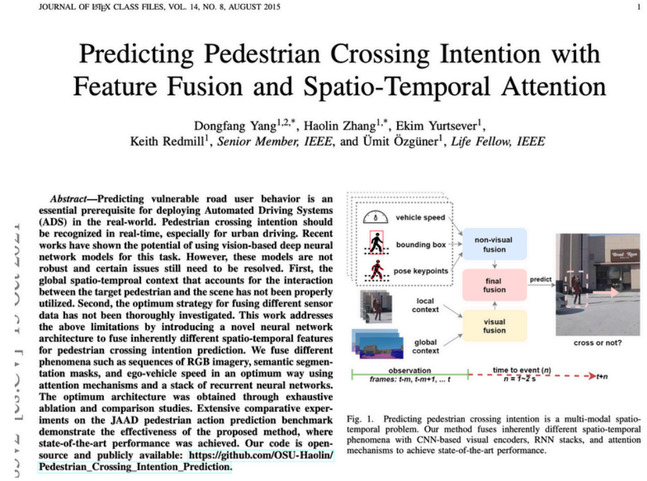
ABSTRACT

Anticipating human motion is an essential requirement for autonomous vehicles and robots in order to primary guarantee people's safety. In urban scenarios, they interact with humans, the surrounding environment, and other vehicles relying on several cues to forecast crossing or not crossing intentions. For these reasons, this challenging task is often tackled using both visual and non-visual features to anticipate future actions from 2 s to 1 s earlier the event. Our work primarily aims to revise this standard evaluation protocol to forecast crossing events as early as possible. To this end, we conceive a solution upon an extensively used model for egocentric action anticipation (RU-LSTM), proposing to envision future features, or modalities, that can better infer human intentions using a properly attention-based fusion mechanism. We validate our model against JAAD and PIE datasets and demonstrate that an intent prediction model can benefit from these additional clues for anticipating pedestrians crossing events.

Fig. 1. Our model detects crossing events in two stages: an encoding stage (R-LSTM), processing the initial part of the sequence (0.5 s), and a decoding stage (U-LSTM), predicting the event at multiple anticipation times. We estimate future visual and non visual features, with an attention-based (goal) module, that are provided to the decoding stage. We consider a set of 4 anticipation times: 4.0, 3.0, 2.0, and 1.0 seconds.

2) *Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention (Dongfang Yang et al.)*

- Fusion of different spatio-temporal features
- **Model**: CNNs (for spatial feature extraction) + RNNs (specifically GRUs) with **attention mechanisms** for temporal analysis.
- **Datasets**: state-of-the-art performance on the JAAD



Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention

Dongfang Yang[1,2,*], Haolin Zhang[1,*], Ekim Yurtsever[1], Keith Redmill[1], *Senior Member, IEEE*, and Ümit Özgüner[1], *Life Fellow, IEEE*

***Abstract*—Predicting vulnerable road user behavior is an essential prerequisite for deploying Automated Driving Systems (ADS) in the real-world. Pedestrian crossing intention should be recognized in real-time, especially for urban driving. Recent works have shown the potential of using vision-based deep neural network models for this task. However, these models are not robust and certain issues still need to be resolved. First, the global spatio-temproal context that accounts for the interaction between the target pedestrian and the scene has not been properly utilized. Second, the optimum strategy for fusing different sensor data has not been thoroughly investigated. This work addresses the above limitations by introducing a novel neural network architecture to fuse inherently different spatio-temporal features for pedestrian crossing intention prediction. We fuse different phenomena such as sequences of RGB imagery, semantic segmentation masks, and ego-vehicle speed in an optimum way using attention mechanisms and a stack of recurrent neural networks. The optimum architecture was obtained through exhaustive ablation and comparison studies. Extensive comparative experiments on the JAAD pedestrian action prediction benchmark demonstrate the effectiveness of the proposed method, where state-of-the-art performance was achieved. Our code is open-source and publicly available: https://github.com/OSU-Haolin/Pedestrian_Crossing_Intention_Prediction.**

Fig. 1. Predicting pedestrian crossing intention is a multi-modal spatio-temporal problem. Our method fuses inherently different spatio-temporal phenomena with CNN-based visual encoders, RNN stacks, and attention mechanisms to achieve state-of-the-art performance.

[1] Osman, Nada, E. Cancelli, G. Camporese, P. Coscia, and L. Ballan. "Early Pedestrian Intent Prediction via Features Estimation." In Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 206-213. 2022.
[2] Yang, Dongfang, H. Zhang, E. Yurtsever, K. Redmill, and Ü. Özgüner. "Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention." Journal of LaTeX Class Files, vol. 14, no. 8, pp. 1-10. 2021.

OVERVIEW OF OUR APPROACH

JAAD Dataset

Our model was trained on JAAD dataset, which includes annotated video sequences of pedestrian behaviors, providing a comprehensive source for training and evaluating our model.

Bboxes & Pose Keypoints

Used bounding box annotations to track pedestrian movements across video frames and extracted pose keypoints to enhance the model's understanding of pedestrian behavior.

VGG19 + LSTM

Implemented a modified version of the pre-trained VGG19 Convolutional Neural Network (CNN) followed by a Long Short-Term Memory (LSTM) network to analyze spatial and temporal patterns in pedestrian movements.

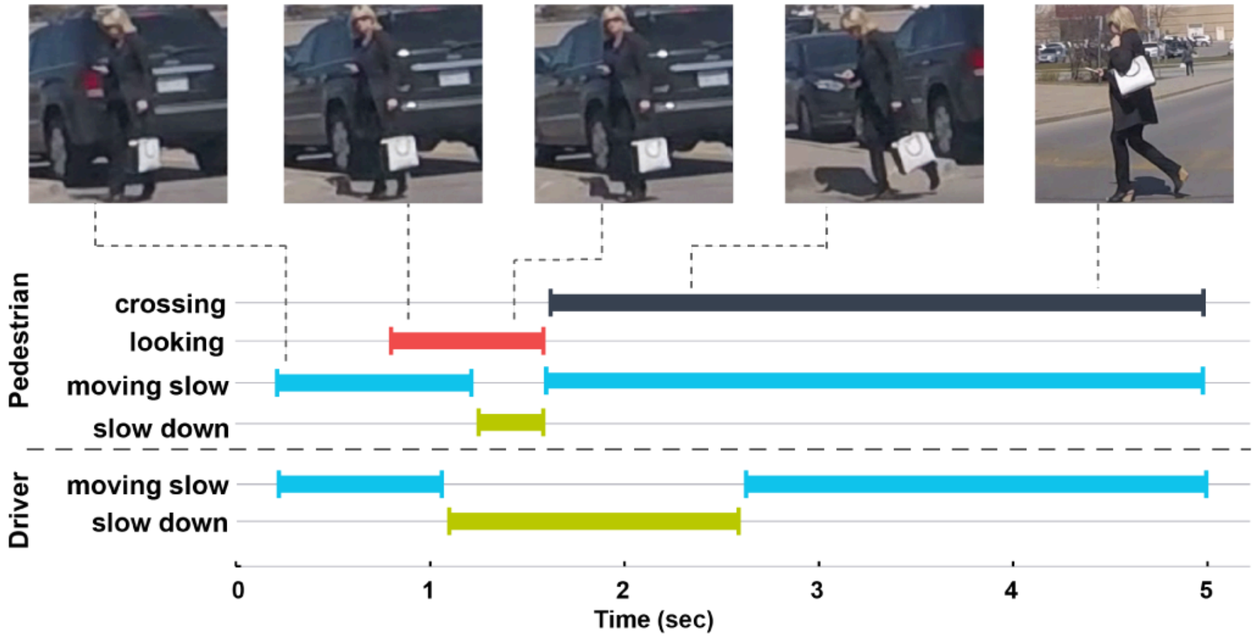
Performance Evaluation

Evaluated model performance using standard metrics such as accuracy, recall, and F1-score, ensuring robust and reliable predictions of pedestrian crossing behavior.

JAAD 2.0 DATASET

- The **JAAD** (*Joint Attention for Autonomous Driving*) 2.0 dataset is designed for research in autonomous driving, focusing on **pedestrian behavior** and **intention prediction**.

- It includes annotated video sequences of pedestrian behaviors captured in various urban environments.



- Provides precise localization of pedestrians in video frames using **Bounding Boxes** coordinates and indicates the level of visibility of each pedestrian including **occlusion informations**.

[3] Rasouli, Amir, Iuliia Kotseruba, and John 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, pp. 206-213. 2017.
[4] Rasouli, Amir, Iuliia Kotseruba, and John K. Tsotsos. "*Agreeing to cross: How drivers and pedestrians communicate.*" In IEEE Intelligent Vehicles Symposium (IV), pp. 264-269. 2017.

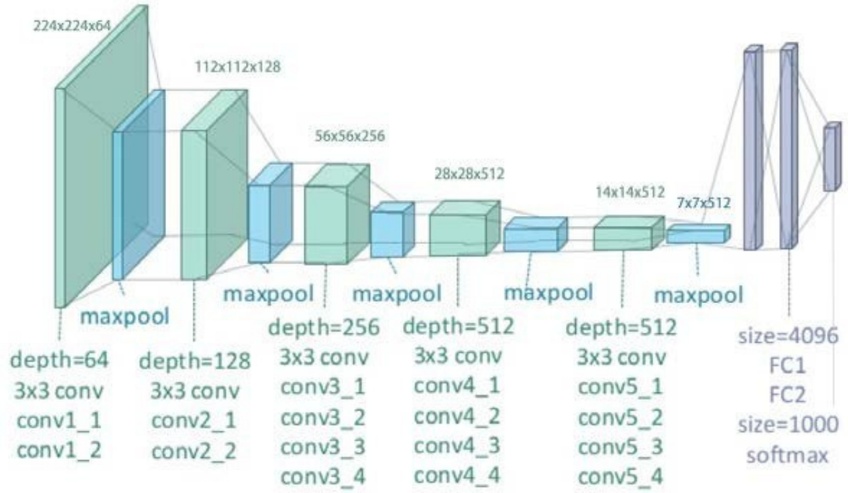
JAAD 2.0 DATASET - ANNOTATIONS

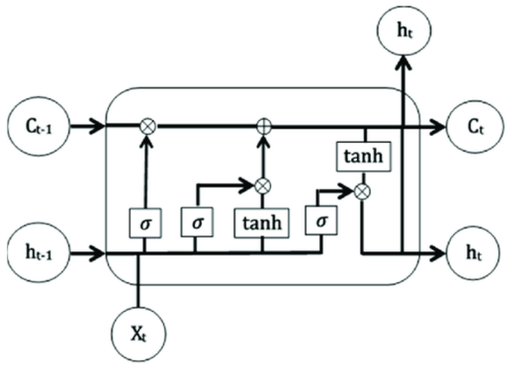
Annotation Types:

- **Generic Annotations**: Video attributes (time of day, weather, location), pedestrian bounding box coordinates, occlusion information, and activities (e.g., walking, looking).
- **Attributes**: Information regarding pedestrian demographics, crossing points, crossing characteristics.
- **Appearance**: Pedestrian appearance details such as pose, clothing, objects carried (high visibility videos).
- **Traffic**: Information about traffic signs and traffic lights for each frame.
- **Vehicle**: Vehicle actions per frame (e.g., moving fast, speeding up).

MODEL ARCHITECTURE: VGG19 + LSTM

Our model combines the **VGG19 Convolutional Neural Network** (CNN) for spatial feature extraction and **Long Short-Term Memory (LSTM)** networks for temporal pattern analysis.





VGG19 Feature Extraction:

- Pre-trained VGG19 model is used to extract features from the input images.
- Layers up to the 36th are frozen to utilize pre-trained weights and reduce training time.
- The extracted features are pooled and passed through the VGG19 classifier (excluding the final layer).

LSTM for Temporal Analysis:

- The output from the VGG19 is reshaped and fed into an LSTM network.
- LSTM captures the temporal dependencies in the sequence of frames, crucial for understanding pedestrian behavior over time.

MODEL ARCHITECTURE: SOFT ATTENTION MECHANISM

➔ **Attention Weights:**

- The LSTM outputs a sequence of hidden states.
- Each hidden state is passed through a series of linear layers and ReLU activation to compute **attention scores**.
- The attention scores are normalized using a softmax function to produce **attention weights.**

➔ **Context Vector:**

- The attention weights are applied to the LSTM hidden states.
- A weighted sum of the hidden states is computed, resulting in a **context vector**. This vector captures the most relevant temporal features from the sequence.

```python
# Attention mechanism
class SoftAttention(nn.Module):
    def __init__(self, hidden_dim):
        super(SoftAttention, self).__init__()
        self.hidden_dim = hidden_dim
        # Attention network: 2 linear layers + ReLU activation
        self.attention = nn.Sequential(
            nn.Linear(hidden_dim, hidden_dim),
            nn.ReLU(inplace=True),
            nn.Linear(hidden_dim, 1)
        )

    def forward(self, lstm_output):
        # lstm_output: [batch_size, seq_len, hidden_dim]
        # Calculate attention weights
        attn_weights = self.attention(lstm_output)
        # Normalize the attention weights using softmax
        attn_weights = torch.softmax(attn_weights, dim=1)
        # Compute the context vector as a weighted sum of LSTM outputs
        context = torch.sum(attn_weights * lstm_output, dim=1)
        return context, attn_weights
```

BOUNDING BOXES & KEYPOINT EXTRACTION WITH MEDIAPIPE

- **Pose Estimation:** the frame is resized and converted to RGB before being processed by MediaPipe to detect pose keypoints.
- **Bounding Boxes**: Annotated on each frame to capture pedestrian locations.
- **Drawing and Saving Keypoints:** if pose landmarks are detected, the keypoints are extracted, drawn on the frame, and saved as '*.npy*' file for each image.

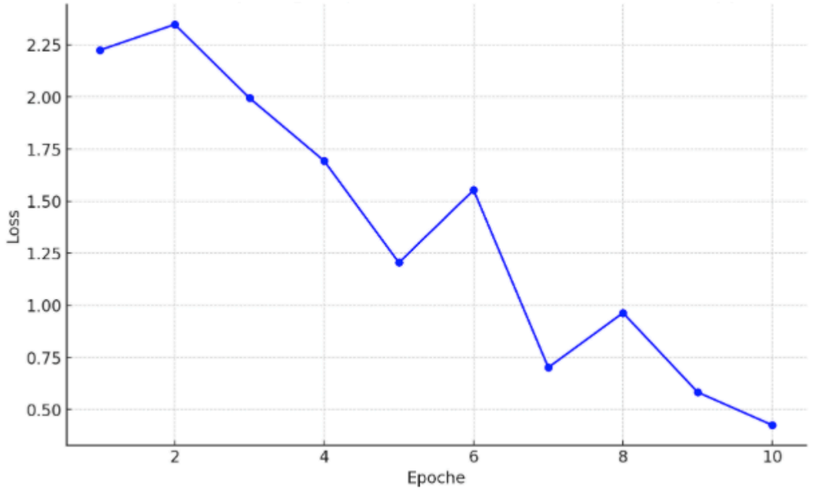
- **Media Pipe** is an open-source framework from Google that provides tools for working with media data or for media processing.
  - Provides high-accuracy pose estimation.
  - Efficient and works in real-time, suitable for processing video frames.

Google MediaPipe

MODEL TRAINING AND TESTING

The JAAD dataset was split into **training** and **testing** sets with:

- *.pkl* files contain preprocessed video data
- *.pt* files contain the processed frames, keypoints, and additional information: traffic, vehicle, appearance and attributes.



Train:

- **Mixed Precision Training**: GradScaler and autocast are used for mixed precision training to speed up computation and reduce memory usage.
- **Optimizer and Scheduler**: the Adam optimizer and StepLR learning rate scheduler are set up.
- **Loss Criterion**: the loss function is defined as BCEWithLogitsLoss.

Test:

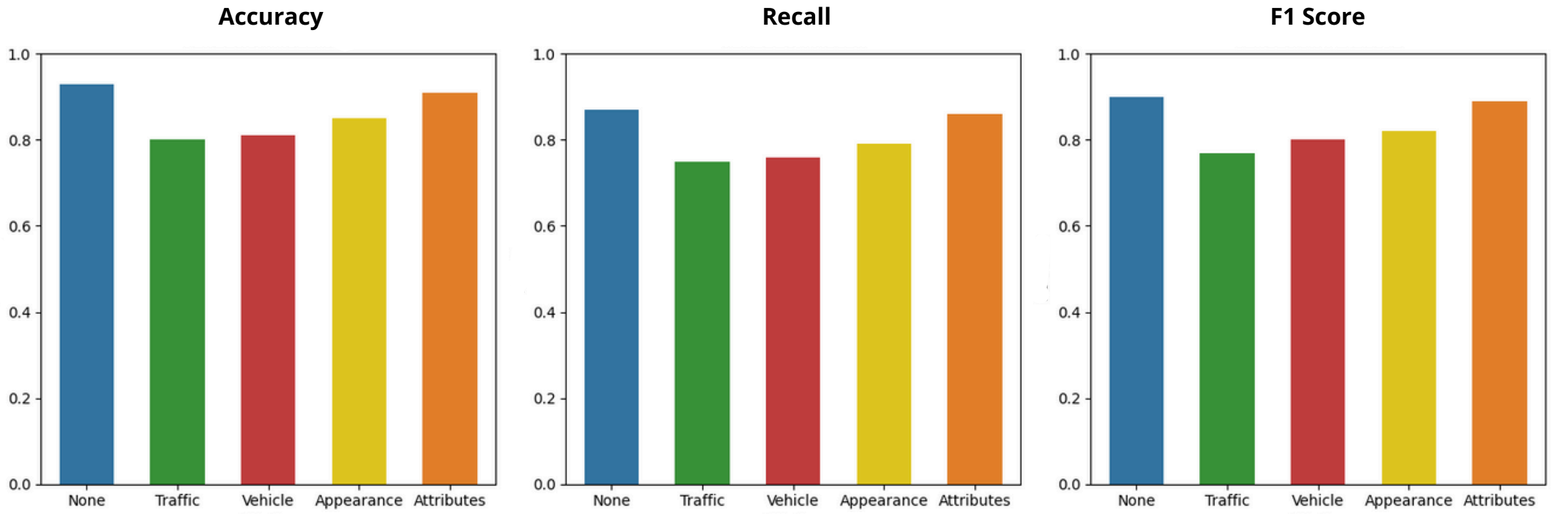
- average loss
- accuracy
- recall
- F1 score

Depending on the specified ablation type, the corresponding feature is zeroed out to study its impact on model performance.

PERFORMANCE EVALUATION: ABLATION STUDIES

Ablation Studies

- Evaluates the model with and without specific features to understand their impact on performance.
- The function iterates over different ablation types (or no ablation)



Results

- **Ablation "traffic" and “vehicle“:**
  - The removal of traffic information drastically reduces model performance, underscoring its importance for accurate predictions.
  - In contrast, removing vehicle information has a less severe impact, suggesting the model can still perform relatively well without it.

SAPIENZA
UNIVERSITÀ DI ROMA

DIAG
Dipartimento di Ingegneria
informatica, automatica e gestionale
Antonio Ruberti

THANK YOU!

- FOR YOUR ATTENTION

Gianmarco Donnesi
Matr. n. 2152311

Michael Corelli
Matr. n. 1938627

A.Y. 23/24