Pedestrian Intention Estimation

Computer Vision Project

Master thesis – Scaccia Simone

SAPIENZA UNIVERSITÀ DI ROMA

ALCOR Lab

Overview

- Introduction
- State of the art and challenges
- Proposed method
- Conclusion

Pedestrian crossing behavior

- Problem: In level 4 autonomous driving, **pedestrian crossing behavior** is one of the most important behaviors that needs to be studied urgently. In urban scenarios, vehicles frequently interact with crossing pedestrians. If the autonomous system failed to handle vehicle-pedestrian interaction, casualties will most likely occur. [1]

- Solutions [2]:
  - 1:Pedestrian action prediction through **trajectory prediction**
    - Pros: effective when the pedestrians are already crossing or are about to do so.
    - Cons: these algorithms react to an **action already in progress** instead of anticipating it.

Pedestrian crossing behavior

- Problem: In level 4 autonomous driving, **pedestrian crossing behavior** is one of the most important behaviors that needs to be studied urgently. In urban scenarios, vehicles frequently interact with crossing pedestrians. If the autonomous system failed to handle vehicle-pedestrian interaction, casualties will most likely occur. [1]

- Solutions [2]:
  - 2: **Pedestrian Intention Estimation**
    - Pros: a remedy for the common drawbacks of trajectory-based algorithms is to **anticipate the action** by estimating its underlying cause or intention (crossing/non-crossing).
    - Cons:
      - **Labeling videos is expensive and time-consuming**. Even though there are a lot of videos captured by street or commodity cameras, they cannot be used to improve the prediction performance without annotation. [5]
      - Given the same driving scenes, two studies reported that **human drivers** not only **disagreed on pedestrian crossing intentions** at the same pre-determined critical frames but also tended to **estimate crossing/non-crossing at different timings**. These phenomena reflect the uncertainty of understanding complex pedestrian-crossing driving scenes, which are highly dynamic, non-deterministic, and context dependent. [4]
      - We need a **representative and heterogeneous sample** of persons to obtain label data on intentions.
      - Proposed solution: unsupervised approach

Pedestrian Intention

- Ad hoc dataset (PIE [2]):
  - The dataset contains several hours of naturalistic video footage of pedestrians in urban environments. In addition to bounding box and behavior annotations, we augment our dataset with **human reference data for pedestrian intention estimation** established via a large-scale experiment.
  - **Video dataset**: The PIE dataset consists of over 6 hours of driving footage captured with calibrated monocular dashboard camera Waylens Horizon equipped with 157° wide angle lens. All videos are recorded in HD format (1920 × 1080 px) at 30 fps. The camera was placed inside the vehicle below the rear-view mirror. For convenience, videos are split into approx. 10 minute long chunks and grouped into 6 sets. The entire dataset was recorded in downtown Toronto, Canada during daytime under sunny/overcast weather conditions
  - Annotations. For each pedestrian close to the road that can potentially interact with the driver we provide the following annotations: bounding boxes with occlusion flags, as well as **crossing intention confidence** and text labels for pedestrians' actions ("walking", "standing", "looking", "not looking", "crossing", "not crossing"). Each pedestrian has a unique id and can be tracked from the moment of appearance in the scene until going out of the frame.
  - We collected 27,630 responses from over 700 subjects (ages 19 − 88).

Pedestrian Intention

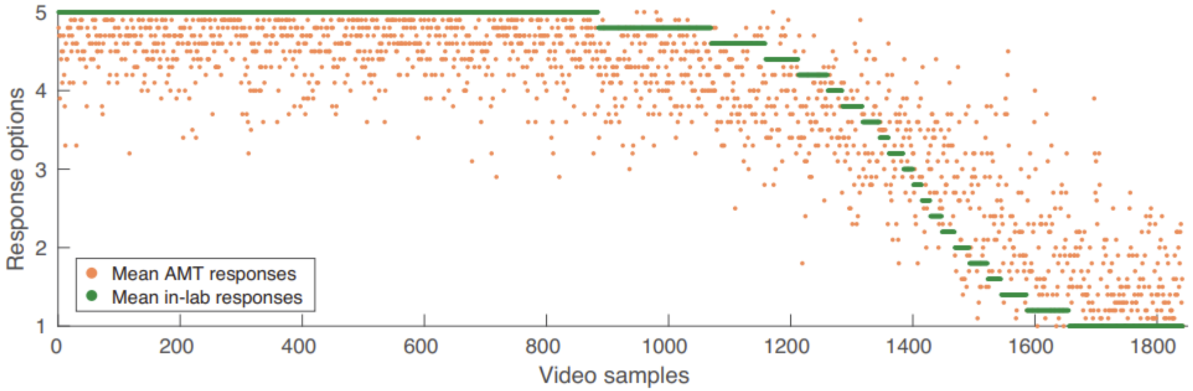
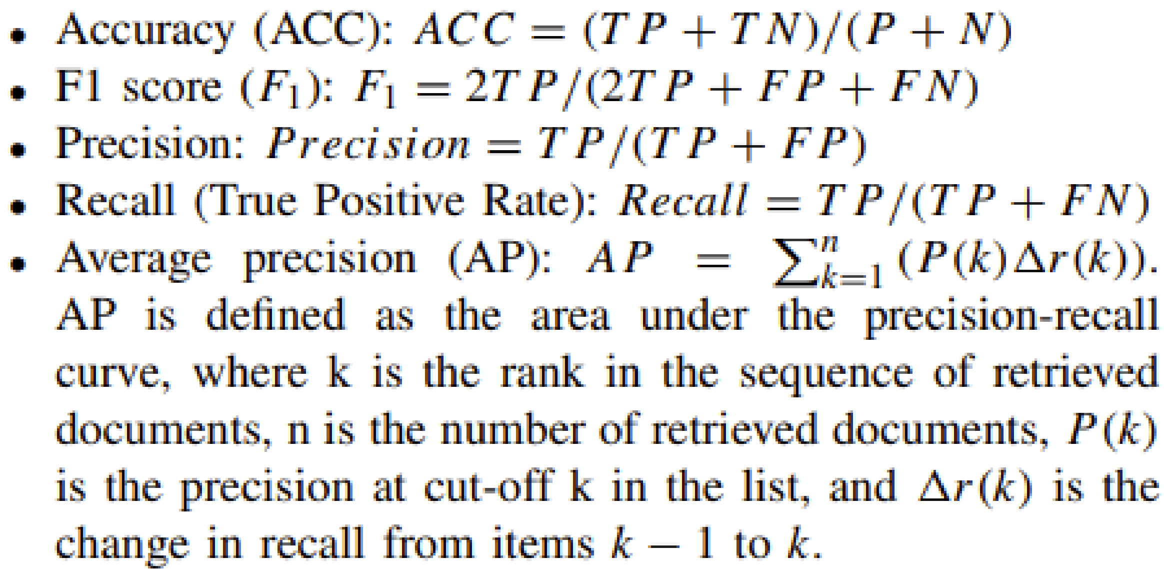


Figure 2. A plot of average responses to the question "Does this pedestrian want to cross?" for each of the 1842 video samples containing a single pedestrian of interest. Answer option 5 is selected for the presence and option 1 for the absence of crossing intention respectively. Answer options in between represent various levels of uncertainty. In-lab and AMT responses are shown as green and red dots respectively. Average responses are sorted in descending order for clarity.

Intention Evaluation Metrics [6]

- The evaluation metrics for intention prediction are listed below, with the number of positives P, negatives N, true positives TP, true negatives TN, false positives FP, and false negatives FN.



- Accuracy (ACC): $ACC = (TP + TN)/(P + N)$
- F1 score (F_1): $F_1 = 2TP/(2TP + FP + FN)$
- Precision: $Precision = TP/(TP + FP)$
- Recall (True Positive Rate): $Recall = TP/(TP + FN)$
- Average precision (AP): $AP = \sum_{k=1}^{n} (P(k)\Delta r(k))$. AP is defined as the area under the precision-recall curve, where k is the rank in the sequence of retrieved documents, n is the number of retrieved documents, $P(k)$ is the precision at cut-off k in the list, and $\Delta r(k)$ is the change in recall from items $k - 1$ to k.

Intention Evaluation Metrics

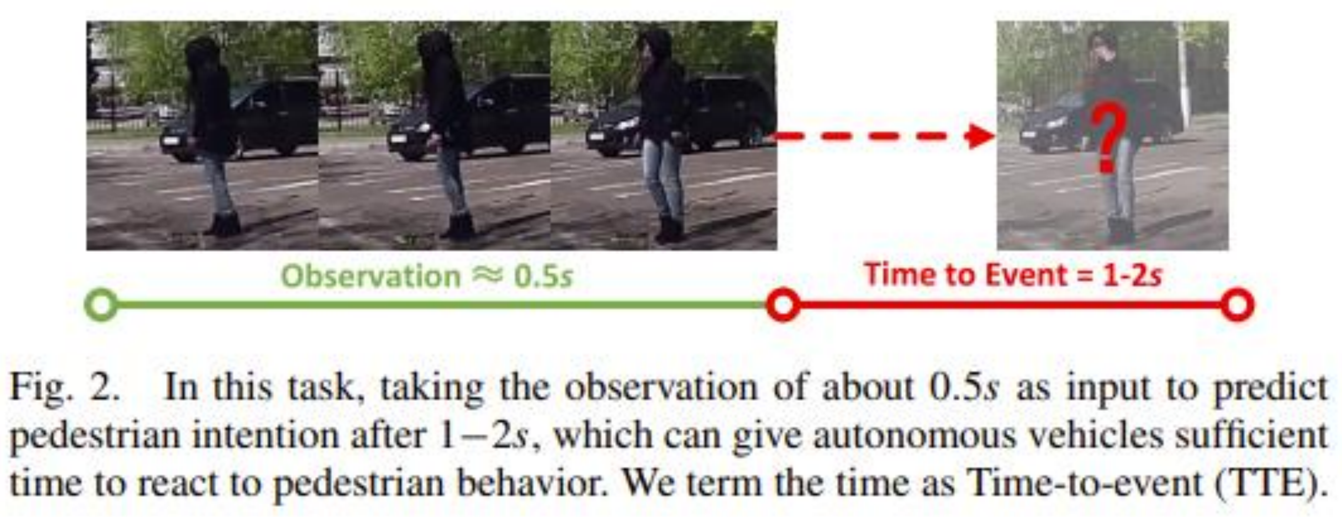
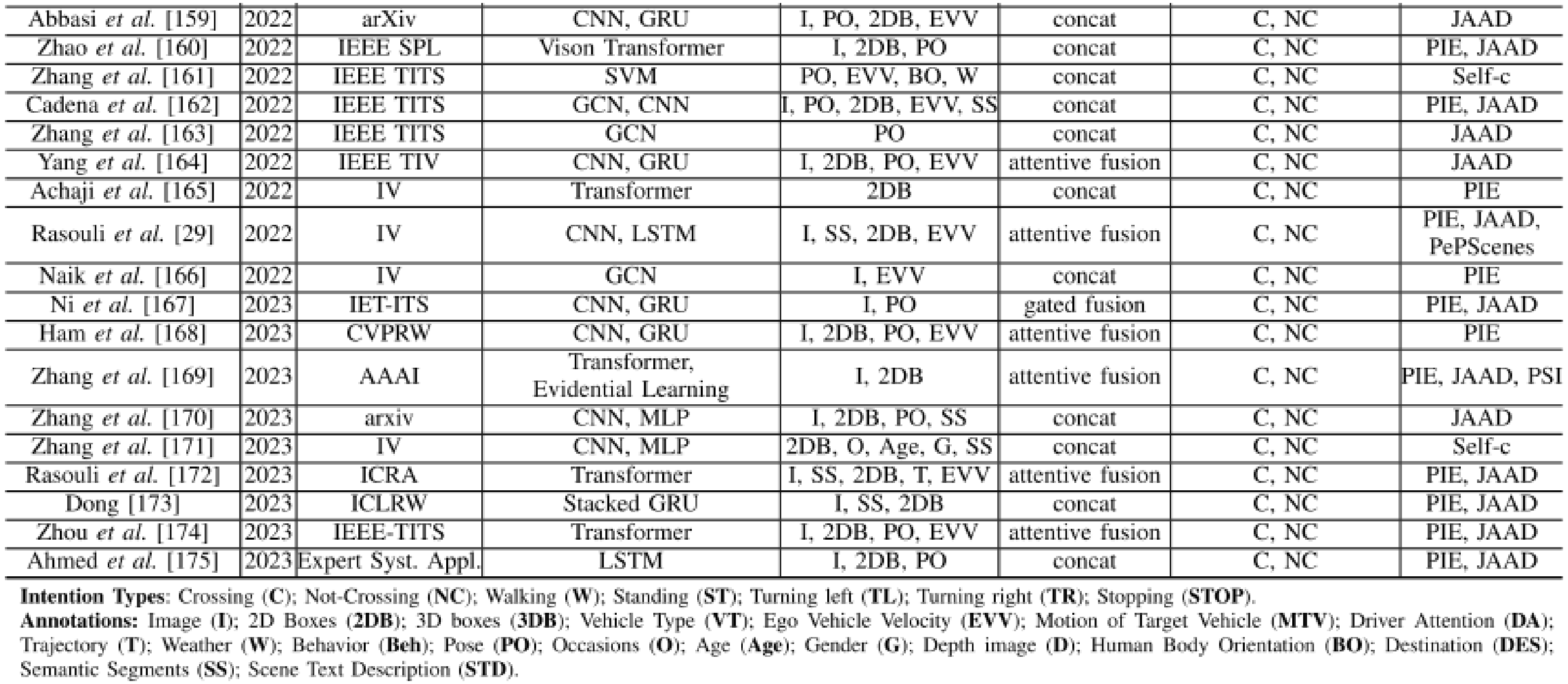


Fig. 2. In this task, taking the observation of about 0.5s as input to predict pedestrian intention after $1-2s$, which can give autonomous vehicles sufficient time to react to pedestrian behavior. We term the time as Time-to-event (TTE). [10]

Recent SOTA on PIE

With the successful application of deep learning, pedestrian centric BIP methods based on Convolutional Neural Networks (**CNNs**), Recurrent Neural Networks (**RNNs**), Long Short Term Memory (**LSTM**) networks, Gated Recurrent Units (**GRUs**), Graph Neural Networks (**GCNs**), and Transformer networks have become popular in this field. The JAAD and PIE datasets have the dominant position for performance evaluation. [7]



Abbasi *et al.* [159]	2022	arXiv	CNN, GRU	I, PO, 2DB, EVV	concat	C, NC	JAAD
Zhao *et al.* [160]	2022	IEEE SPL	Vison Transformer	I, 2DB, PO	concat	C, NC	PIE, JAAD
Zhang *et al.* [161]	2022	IEEE TITS	SVM	PO, EVV, BO, W	concat	C, NC	Self-c
Cadena *et al.* [162]	2022	IEEE TITS	GCN, CNN	I, PO, 2DB, EVV, SS	concat	C, NC	PIE, JAAD
Zhang *et al.* [163]	2022	IEEE TITS	GCN	PO	concat	C, NC	JAAD
Yang *et al.* [164]	2022	IEEE TIV	CNN, GRU	I, 2DB, PO, EVV	attentive fusion	C, NC	JAAD
Achaji *et al.* [165]	2022	IV	Transformer	2DB	concat	C, NC	PIE
Rasouli *et al.* [29]	2022	IV	CNN, LSTM	I, SS, 2DB, EVV	attentive fusion	C, NC	PIE, JAAD, PePScenes
Naik *et al.* [166]	2022	IV	GCN	I, EVV	concat	C, NC	PIE
Ni *et al.* [167]	2023	IET-ITS	CNN, GRU	I, PO	gated fusion	C, NC	PIE, JAAD
Ham *et al.* [168]	2023	CVPRW	CNN, GRU	I, 2DB, PO, EVV	attentive fusion	C, NC	PIE
Zhang *et al.* [169]	2023	AAAI	Transformer, Evidential Learning	I, 2DB	attentive fusion	C, NC	PIE, JAAD, PSI
Zhang *et al.* [170]	2023	arxiv	CNN, MLP	I, 2DB, PO, SS	concat	C, NC	JAAD
Zhang *et al.* [171]	2023	IV	CNN, MLP	2DB, O, Age, G, SS	concat	C, NC	Self-c
Rasouli *et al.* [172]	2023	ICRA	Transformer	I, SS, 2DB, T, EVV	attentive fusion	C, NC	PIE, JAAD
Dong [173]	2023	ICLRW	Stacked GRU	I, SS, 2DB	concat	C, NC	PIE, JAAD
Zhou *et al.* [174]	2023	IEEE-TITS	Transformer	I, 2DB, PO, EVV	attentive fusion	C, NC	PIE, JAAD
Ahmed *et al.* [175]	2023	Expert Syst. Appl.	LSTM	I, 2DB, PO	concat	C, NC	PIE, JAAD

Intention Types: Crossing (**C**); Not-Crossing (**NC**); Walking (**W**); Standing (**ST**); Turning left (**TL**); Turning right (**TR**); Stopping (**STOP**).
Annotations: Image (**I**); 2D Boxes (**2DB**); 3D boxes (**3DB**); Vehicle Type (**VT**); Ego Vehicle Velocity (**EVV**); Motion of Target Vehicle (**MTV**); Driver Attention (**DA**); Trajectory (**T**); Weather (**W**); Behavior (**Beh**); Pose (**PO**); Occasions (**O**); Age (**Age**); Gender (**G**); Depth image (**D**); Human Body Orientation (**BO**); Destination (**DES**); Semantic Segments (**SS**); Scene Text Description (**STD**).

Challenges [3]

- **Camera and Lidar**:
  in case of **poor or adverse lighting conditions**, adequate feature extraction of the pedestrian and the road scene becomes tough leading to misleading intention predictions. Another preferred sensor offering strong spatial coverage of the scene is **LiDAR** which also provides robustness to all lighting conditions, unlike RGB cameras.

- **Inadequate annotated ground-truth** data availability:
  a large proportion of techniques present for pedestrian intention prediction rely on a supervised learning approach that demands a large-scale labelled dataset for sufficient training. However, **labelling the pedestrian dataset is a highly complex and painstaking task**.

- **Real-time scenario**:
  too complex an algorithm or a huge number of modalities may produce **delayed inference** rendering it useless for real-time operations.

Self-supervised learning for videos [14]

- **Clustering**: clustering is an approach to self-supervised learning that focuses on the optimal grouping of videos in a cluster. We want that **similar videos aggregate** (pedestrian who wants to cross) **while dissimilar videos separate**.

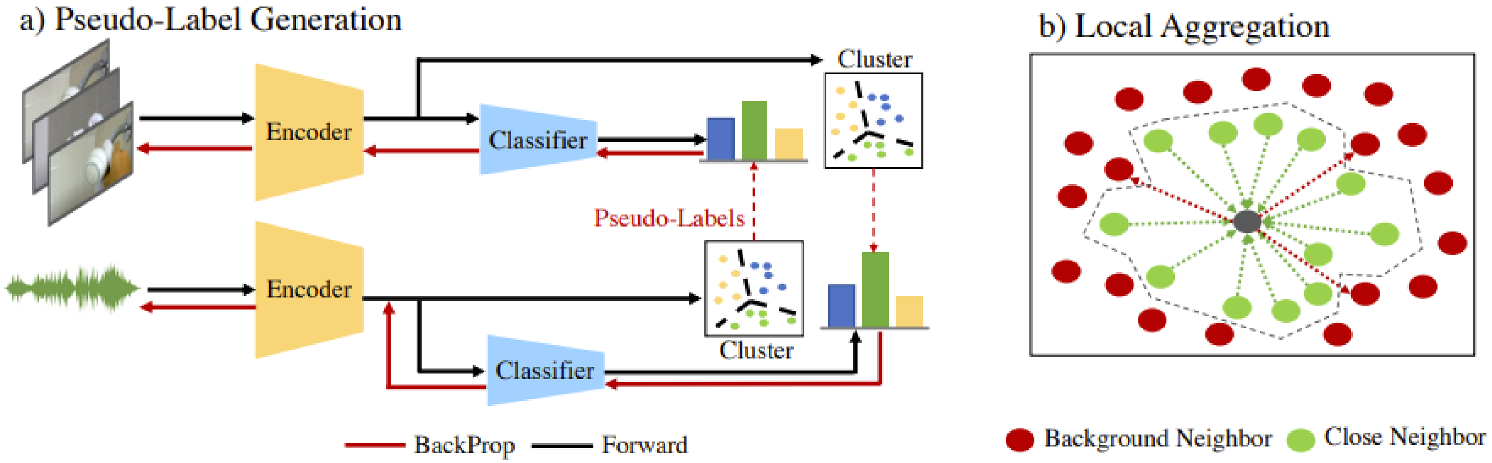


Figure 14: Toy examples of how clustering is used in self-supervised learning for video. a) Pseudo-Label generation uses cluster assignments as pseudo labels with multiple modalities as in Alwassel et al. [2020]. The cluster assignments from one signal are used as pseudo-labels for the other. b) Local Aggregation focuses on the latent space where for each sample "close" neighbors are pulled closer and "background" neighbors are pushed further away like in Zhuang et al. [2019], He et al. [2017], Wang and Schmid [2013], Tokmakov et al. [2020].

VIE: UL Deep Neural Embeddings [15]

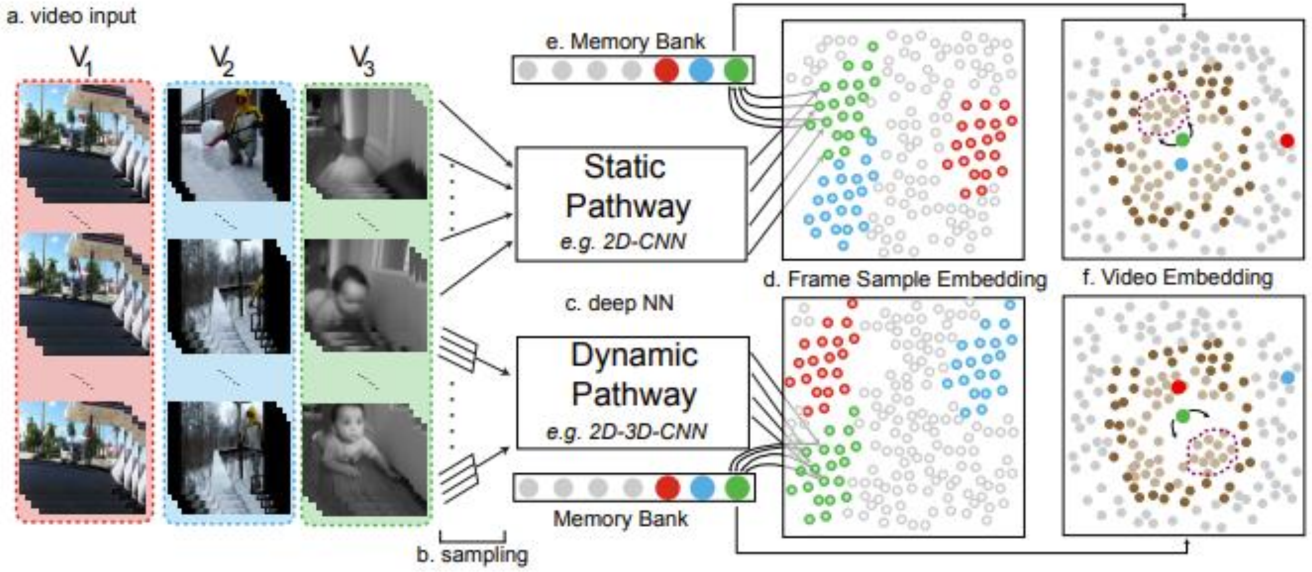
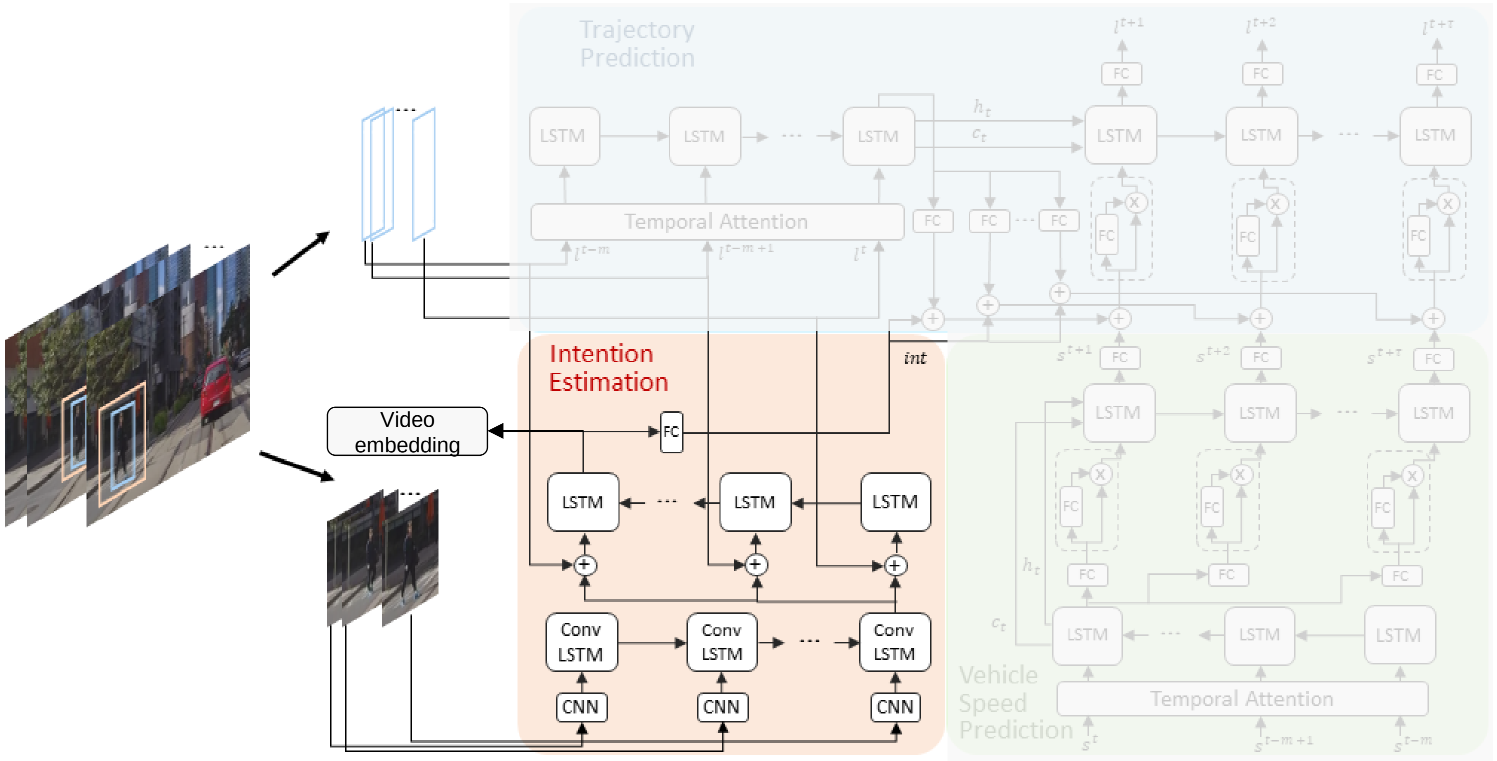


Figure 1: **Schematic of the Video Instance Embedding (VIE) Framework. a.** Frames from individual videos ($\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$) are **b.** sampled into sequences of varying lengths and temporal densities, and input into **c.** deep neural network pathways that are either static (single image) or dynamic (multi-image). **d.** Outputs of frame samples from either pathway are vectors in the D-dimensional unit sphere $S^D \subset \mathbf{R}^{D+1}$. The running mean value of embedding vectors are calculated over online samples for each video, **e.** stored in a memory bank, and **f.** at each time step compared via unsupervised loss functions in the video embedding space. The loss functions require the computation of distribution properties of embedding vectors. For example, the Local Aggregation (LA) loss function involves the identification of Close Neighbors $\mathbf{C}_i$ (light brown points) and Background Neighbors $\mathbf{B}_i$ (dark brown points), which are used to determine how to move target point (green) relative to other points (red/blue).

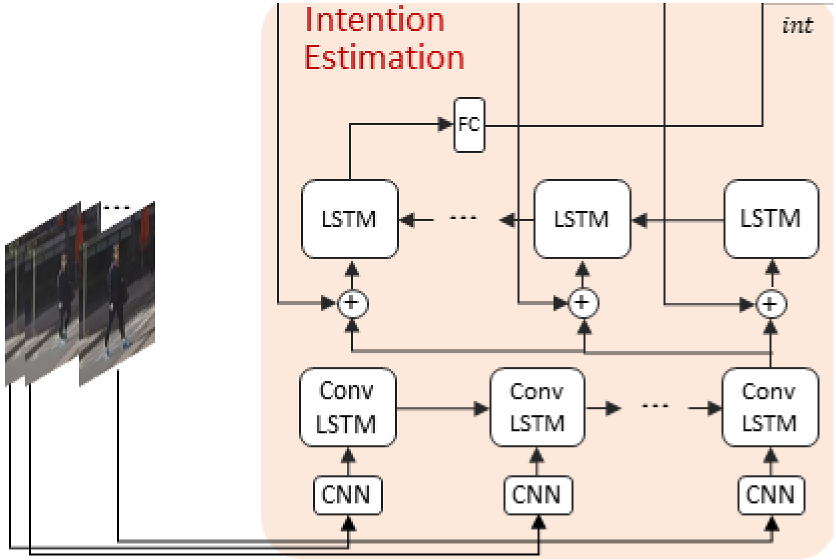
Unsupervised learning for PIE

- Computing embeddings:
  - Frames can have multiple pedestrian, so we need an ad hoc model to predict his intention.
  - First and most simple SOTA using PIE: PIEPredict [2]



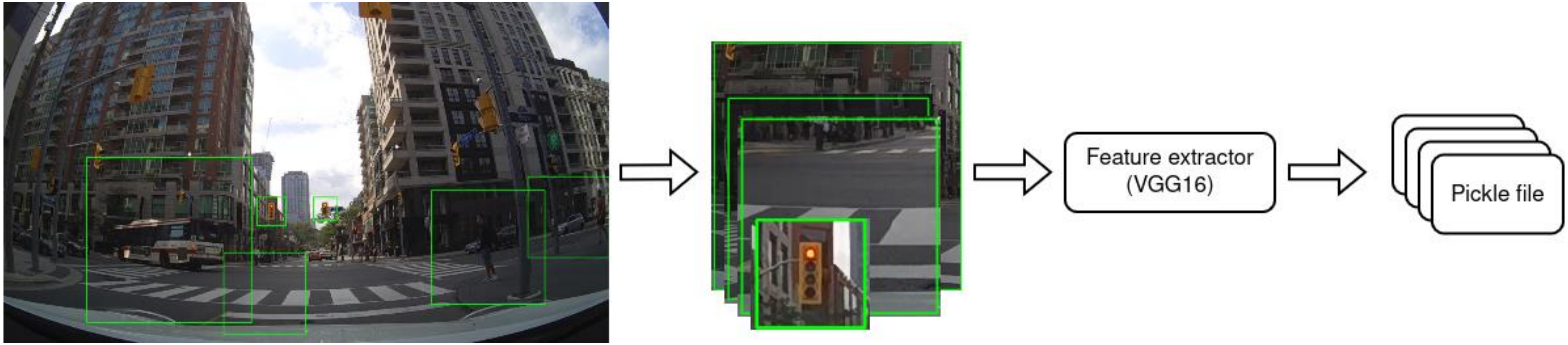
PIE and PIEPredict constraints

- PIE dataset size:
  - **69GB of videos.**
  - Extracting all **annotated frames** requires **1.1TB**.
- PIEPredict pipeline:
  - For each annotated image **extract pedestrian features** using a **pretrained model** VGG16.
  - **Save the feature** as a pickle file.
  - During the next epoch the image's feature is already computed and we can **reuse them** instead of the image.
- Considerations:
  - Storing all the annotated images requires a lot of resources 69GB + 1.1TB + features.
  - **Images are only required the first epoch**, then only its extracted features are used.
  - Image size: 2.5MB, picke file size: 16.5KB using VGG16 fc1.

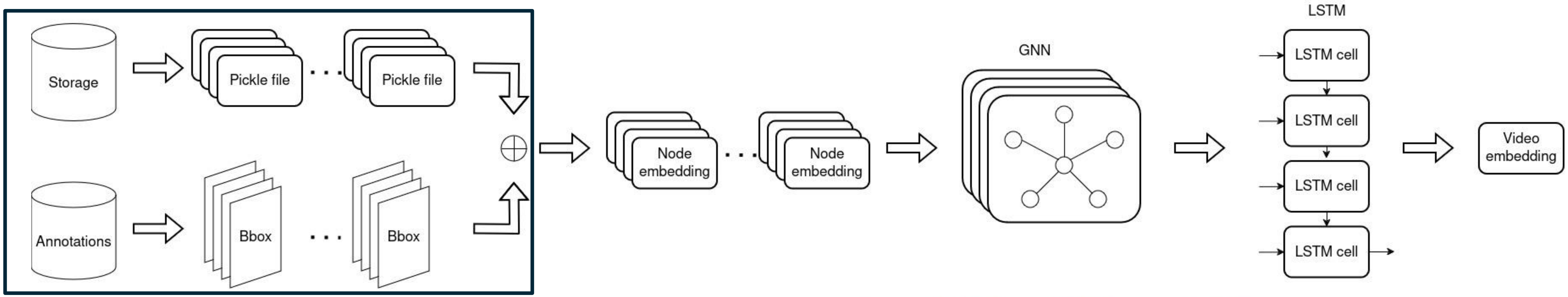


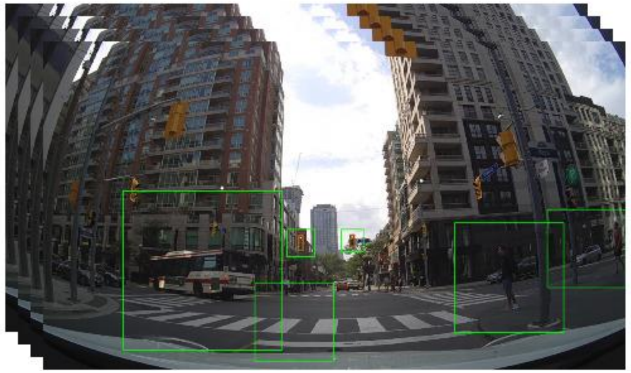
PIE and PIEPredict constraints

- Proposed solution: **preprocessing**
  - PIE suggests extracting and storing all annotated frames, then extracting the image's features during training.
  - To reduce the dataset size, we can **preprocess the image features** during the extraction of the annotated frames. **Instead of storing all the annotated images**, we can store only the extracted features, thus reducing the dataset size.
  - The **final dataset size** is only **10GB**.



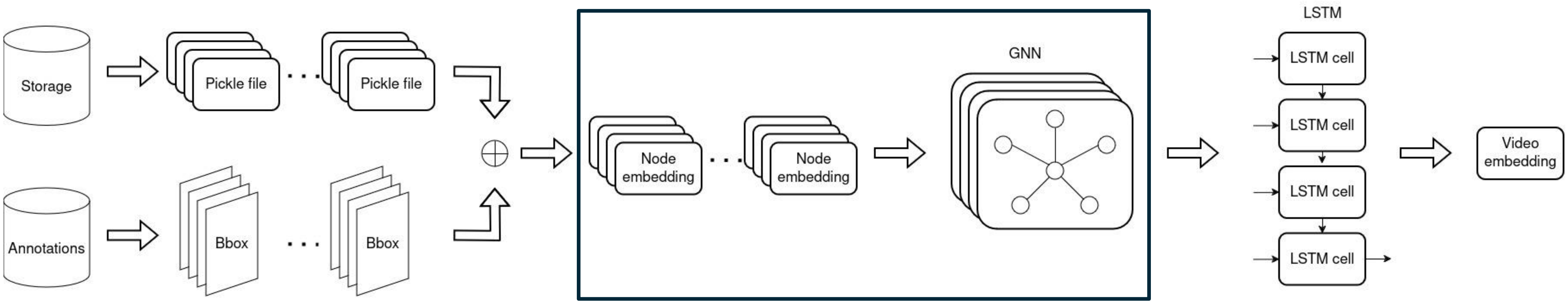
UnPIE network: loading phase

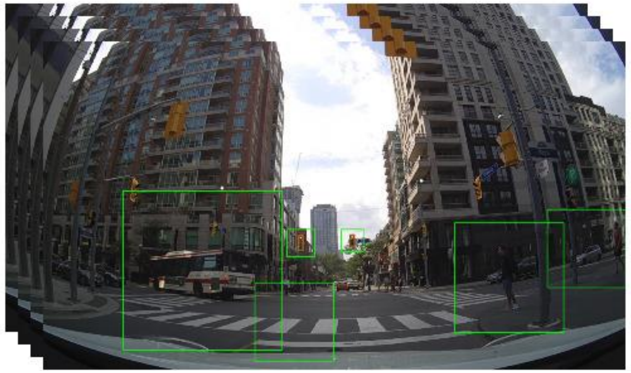


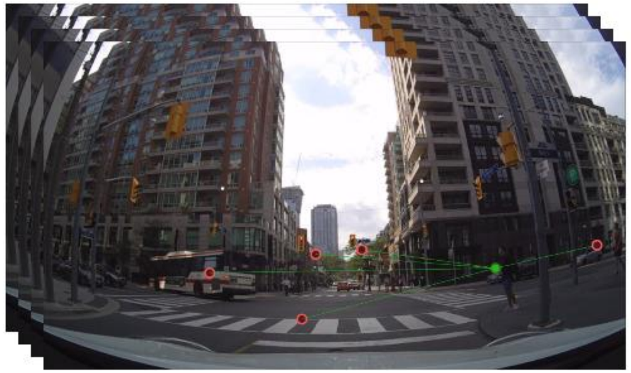


1. **Loading feature phase**: Image feature are loaded and then concatenated with the bbox to encode the relative position between features

UnPIE network: GCN phase

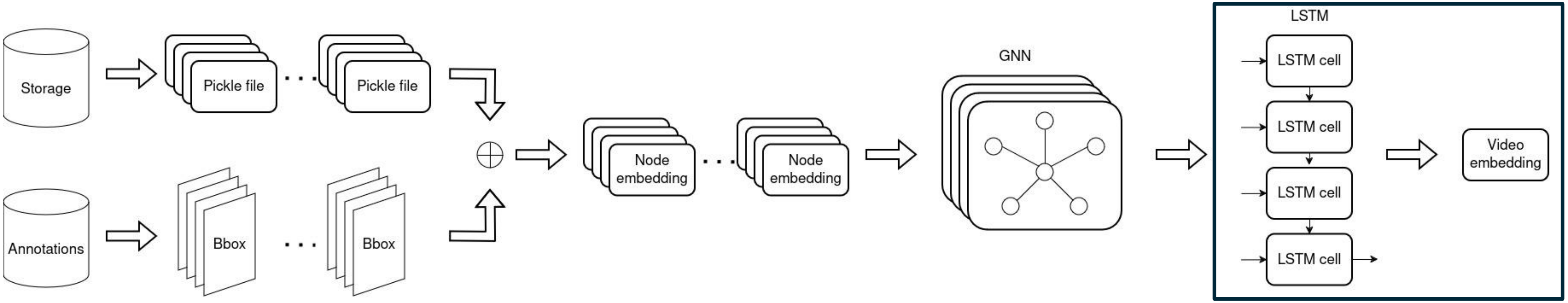


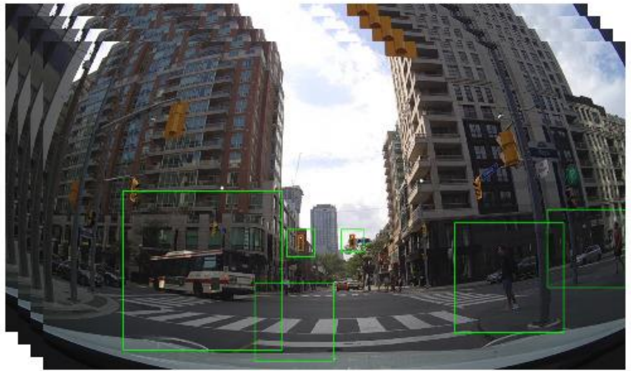




2. **Graph computation phase**: Features are combined in a star graph to be computed by the GCN. Its output is a pedestrian feature embedding for each frame of the clip. The GCN aim is to encode all the features in a single frame, and give in output the representation of a pedestrian intent.

UnPIE network: LSTM phase

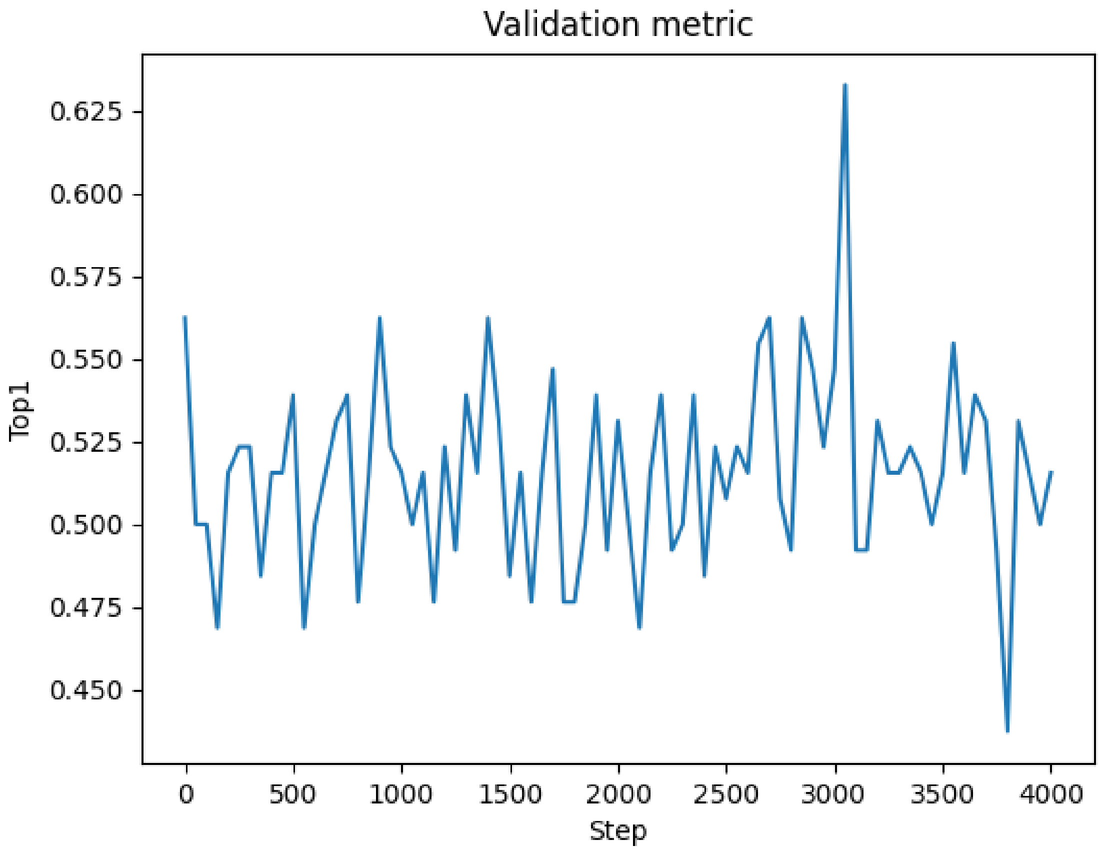




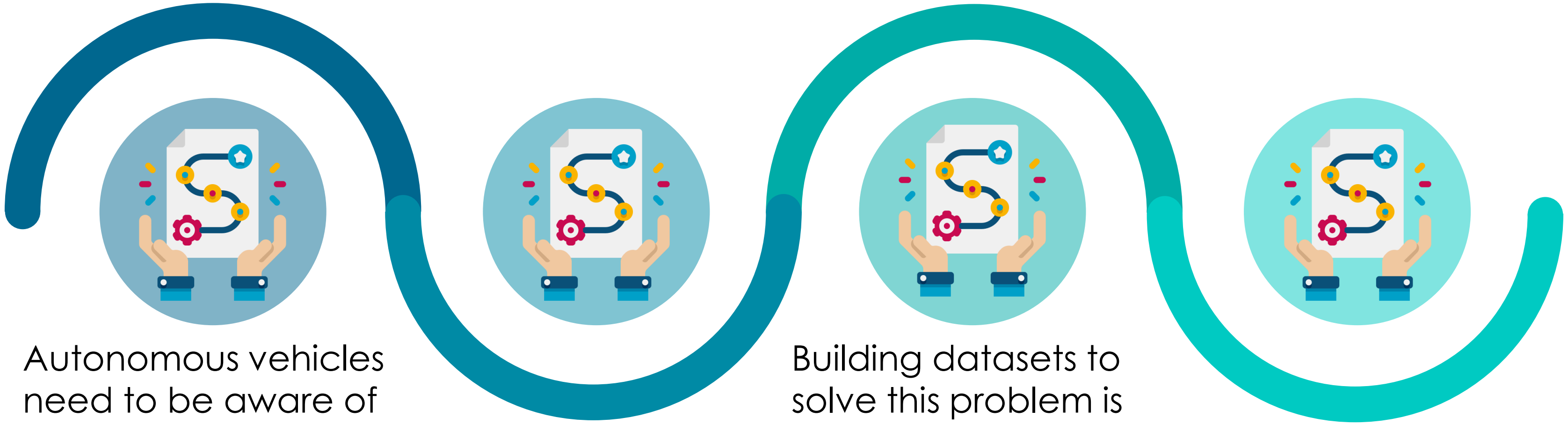
3. **LSTM computation phase**: the LSTM takes in input the representation of the pedestrian intent frame per frame and outputs the aggregate intent representation. The embedding produced should better represent the pedestrian over time. Finally. the embedding will be computed by the VIE framework to train the network parameters.

UnPIE network: accuracy

- After a small training of 4000 epochs the accuracy of the model is slightly better than a random function.
- The causes can be various.
- One of them could be a nonoptimal combination of hyperparameters.
- A different implementation of the Graph Neural Network could drastically improve the performances.



Conclusion



Autonomous vehicles need to be aware of pedestrian crossing behavior.

To avoid casualties, we can predict its trajectory and its intentions.

Building datasets to solve this problem is very expensive and time-consuming

An unsupervised model like UnPIE can be useful to avoid this task

References

1. Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention
   2022, cit. 96
2. PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction
   2019, cit. 320
3. Pedestrian Intention Prediction for Autonomous Vehicles: A Comprehensive Survey
   2022, cit. 30
4. TrEP: Transformer-Based Evidential Prediction for Pedestrian Intention with Uncertainty
   2023, cit. 12
5. AutoTrajectory: Label-Free Trajectory Extraction and Prediction from Videos Using Dynamic Points
   2020, cit. 15
6. Pedestrian Behavior Prediction Using Deep Learning Methods for Urban Scenarios: A Review
   2023, cit. 10

References

7. Multi-scale pedestrian intent prediction using 3D joint information as spatio-temporal representation
   2023, cit. 7
8. Pedestrian Intention Prediction: A Multi-task Perspective
   2020, cit. 37
9. PIT: Progressive Interaction Transformer for Pedestrian Crossing Intention Prediction
   2023, cit. 14
10. PedFormer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning
    2023, cit. 12
11. CIPF: Crossing Intention Prediction Network based on Feature Fusion Modules for Improving Pedestrian Safety
    2023, cit. 7

References

13. Scene Spatio-Temporal Graph Convolutional Network for Pedestrian Intention Estimation
    2022, cit. 5
14. Self-supervised learning for videos: A survey
    2023, cit. 93
15. Unsupervised Learning from Video with Deep Neural Embeddings
    2020, cit. 63

Contacts

Alcor Lab

WEBSITE
https://alcorlab.diag.uniroma1.it/

EMAIL
alcor@diag.uniroma1.it

Personal contacts

Scaccia Simone
scaccia.2045976@studenti.uniroma1.it