

Spatiotemporal Event Graphs for Dynamic Scene Understanding

Salman Khan

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Computer Science and Mathematics

Visual Artificial Intelligence Laboratory (VAIL)
School of Engineering, Computing and Mathematics
Oxford Brookes University

Supervised by :
Prof. Fabio Cuzzolin and Dr. Tjeerd Olde Scheper

January 2023

*The author holds the **copyright** of this thesis. Any person(s) intending to use a part or whole of the materials in the thesis in a proposed publication must seek copyright release from the author.*

Thesis/Assessment Committee

Prof. Dima Damen (External Examiner)
Professor of Computer Vision
Lead of Machine Learning and Computer Vision Group
Department of Computer Science
University of Bristol

Dr. Matthias Rolf (Internal Examiner)
Reader
School of Engineering, Computing and Mathematics
Oxford Brookes University

Preface

This dissertation is submitted for the degree of Doctor of Philosophy at Oxford Brookes University, United Kingdom. The research presented herein was undertaken under the kind supervision of Prof. Fabio Cuzzolin and Dr. Tjeerd Olde Scheper between February 2020 and February 2023. To the best of my knowledge, this work is original, excluding those previous works for which acknowledgment and reference have been made. Neither this nor any substantially similar dissertation has been or is being submitted for any other degree, diploma, or qualification at any other university. Part of this work has been published as follows:

- **Salman Khan**, Izzeddin Teeti, Andrew Bradley, Mohamed Elhoseiny, and Fabio Cuzzolin, “A Hybrid Graph Network for Complex Activity Detection in Video”, Pre-print arXiv, (2023).
- Gurkirt Singh, ..., **Salman Khan**, ..., and Fabio Cuzzolin, “ROAD: The ROAd event Awareness Dataset for autonomous Driving”, in IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI) 45, pp. 1036-1054, January 2023.
- E. Giunchiglia, **Salman Khan**, ..., F. Cuzzolin and T. Lukasiewicz, “ROAD-R: the Autonomous Driving Dataset for Learning with Requirements”, Machine Learning, Springer, September 2022 (accepted with minor revision)
- A. Shahbaz, **Salman Khan**, ... and F. Cuzzolin, “International Workshop on Continual Semi-Supervised Learning: Introduction, Benchmarks and Baselines”, in CSSL 2021, Lecture Notes in Artificial Intelligence Vol. 13418, pages 1-14, Springer, 2022
- **Salman Khan** and Fabio Cuzzolin, “Spatiotemporal Deformable Scene Graphs for Complex Activity Detection”, in Proceedings of The 32nd British Machine Vision Conference (BMVC), 2021.
- Eleonora Giunchiglia, Mihaela Catalina Stoian, **Salman Khan**, Fabio Cuzzolin and Thomas Lukasiewicz, “ROAD-R: The Autonomous Driving Dataset with Logical Requirements”, in Proceedings of IJCAI 2022 Workshop on Artificial Intelligence for Autonomous Driving (AI4AD 2022) and IJCLR 2022.
- Ajmal Shahbaz, **Salman Khan**, Mohammad Asiful Hossain, Vincenzo Lomonaco, Kevin Cannons, and Fabio Cuzzolin, International Workshop on Continual Semi-Supervised Learning: Introduction, Benchmarks and Baselines, in Proceedings of

IJCAI Workshop, International Workshop on Continual Semi-Supervised Learning (2021).

- Fabio Cuzzolin, **Salman Khan**, Ajmal Shahbaz “Continual Semi-Supervised Learning for Classification via a Time-Varying Hidden Markov Model”, Pre-print arXiv 2023.

Dedication

In memory of my father
Late *Mushtaq Ahmad*
With love and eternal appreciation

&

I would like to dedicate this work to my parents, siblings, and wife. Without whom none of my success would be possible

“The best things that a man can leave behind are three: A righteous son who will pray for him, ongoing charity whose reward will reach him, and knowledge which is acted upon after his death.”

[The Messenger of Allah: Prophet Muhammad SAW
- Sunan Ibn Majah Vol. 1, Book 1, Hadith 241]

Acknowledgements

First and foremost, I would like to praise Allah the Almighty, the Most Gracious, and the Most Merciful for His blessing given to me during my study and in completing this thesis. May Allah's blessing goes to His final Prophet Muhammad (peace be upon him), his family, and his companions.

I would like to express my heartfelt thanks to my PhD supervisor Prof. Fabio Cuzzolin, and co-supervisor Dr. Tjeerd Olde Scheper for providing such an exceptional opportunity for me to pursue a doctorate in such an interesting topic as Computer Vision! I also thank my supervisors for their invaluable time, guidance and support. In particular, Fabio's strive for excellence, openness, and positive criticism always inspired me to push myself beyond my limits and allowed me to gradually improve my scientific, technical and academic skills. Moreover, human values like elegance and generosity that I found in Prof. Cuzzolin have inspired me to work in the Visual Artificial Intelligence Lab (VAIL), School of Engineering, Computing and Mathematics, Oxford Brookes University.

I would also like to thank my co-authors and collaborators Dr. Andrew Bradley, Dr. Gurkirt Singh, Dr. Eleonora Giunchiglia, and Izzeddin Teeti for their time and effort spent during our long technical discussions in our various projects.

I want to thank: Tjeerd Olde Scheper and Matthias Rolf for their invaluable guidance and support as research tutors; Jill Organ, Catherine Joyejob, and Claire Ferone for helping me with all administrative tasks; all the members of VAIL with whom I had a chance to work, especially Inna Skarga-Bandurova, Vivek Singh, Neha Bhargava, Dinesh Jackson Samuel, Mohamed Mohamed, Bogdan Cirstea, Ajmal Shahbaz, Shireen Kudukkil, Aytac Kanaci, and Amirul Islam.

I am grateful to all external collaborators I worked with: Reza Javanmard, Kevin Cannons, Naeemullah Khan, Mihaela Catalina Stoian, and Valentina Musat.

My special thanks go to Prof. Mohamed H. Elhoseiny for offering me an internship at his lab Vision-CAIR, King Abdullah University of Science and Technology (KAUST), Saudi Arabia. I feel fortunate to have had the chance to work in a highly stimulating environment and thankful to the Vision-CAIR members for their invaluable time and guidance during my stay at KAUST.

Due acknowledgement goes to my MS supervisors Prof. Sung Wook Baik and Prof. Khan Muhammad for their time, advice and support during my MS dissertation. I would also like to thank my BS supervisor Prof. Muhammad Sajjad for his guidance and support.

I am thankful to my friends, who have greatly supported me for many years, Dr. Amin Ullah, Dr. Ijaz Ul Haq, Dr. Tanveer Husaain, Dr. Fath U Min Ullah,

Muhammad Irfan, Imran Ullah, Hayat Ullah, Abid Ali, Hamza Ikram, Yahya Shah, Malik Yasir, and Uzair Shah.

Last but not least, I am thankful for my *parents* and my *family* in Pakistan for their endless patience, encouragement, and generosity. Especially, my late Father *Mushtaq Ahmad* who did not only raise and mature me but also struggled over the years for my education to make me what I am today, and my elder brother *Dr. Ishtiaq Ahmad* to be my mentor since my childhood. Without their perseverance, I would not be writing this today. My heartfelt thanks go to my wife *Malaika* for her unbounded patience and pushing me to be my best self.

Abstract

Dynamic scene understanding is the ability of a computer system to interpret and make sense of the visual information present in a video of a real-world scene. In this thesis, we present a series of frameworks for dynamic scene understanding starting from road event detection from an autonomous driving perspective to complex video activity detection, followed by continual learning approaches for the life-long learning of the models.

Firstly, we introduce the ROad event Awareness Dataset (ROAD) for Autonomous Driving, to our knowledge the first of its kind. ROAD is designed to test an autonomous vehicle’s ability to detect road events, defined as triplets composed by an active agent, the action(s) it performs and the corresponding scene locations. Due to the lack of datasets equipped with formally specified logical requirements, we also introduce the ROad event Awareness Dataset with logical Requirements (ROAD-R), the first publicly available dataset for autonomous driving with requirements expressed as logical constraints, as a tool for driving neurosymbolic research in the area.

Next, we extend event detection to holistic scene understanding by proposing two complex activity detection methods. In the first method, we present a deformable, spatiotemporal scene graph approach, consisting of three main building blocks: action tube detection, a 3D deformable RoI pooling layer designed for learning the flexible, deformable geometry of the constituent action tubes, and a scene graph constructed by considering all parts as nodes and connecting them based on different semantics. In a second approach evolving from the first, we propose a hybrid graph neural network that combines attention applied to a graph encoding of the local (short-term) dynamic scene with a temporal graph modelling the overall long-duration activity. Our contribution is threefold: i) a feature extraction technique; ii) a method for constructing a local scene graph followed by graph attention, and iii) a graph for temporally connecting all the local dynamic scene graphs.

Finally, the last part of the thesis is about presenting a new continual semi-supervised learning (CSSL) paradigm, proposed to the attention of the machine learning community. We also propose to formulate the continual semi-supervised learning problem as a latent-variable.

Contents

Preface	vii
Dedication	ix
Acknowledgements	xi
Abstract	xiii
List of Figures	xxi
List of Tables	xxvii
Nomenclature	xxxi
1 Introduction	1
1.1 Video Action Recognition and Detection	1
1.2 Event Detection	2
1.2.1 Road Event Detection	2
1.3 Temporal Action Localisation	2
1.4 Complex Activity Detection	3
1.5 Scene Graphs	3
1.6 Continual Learning	5
1.7 Contributions of the Thesis	5
1.8 Organisation of the Thesis	6
1.9 Publications	7
1.10 List of Packages and Demos	9
2 Related Work	11
2.1 Autonomous Driving Datasets	11
2.2 Temporal Action Localisation	12
2.3 Complex Activity Detection	12
2.3.1 One-Stage vs Two-Stage Approaches	13
2.4 Graph Networks	13

2.4.1	GCN-Based Temporal Activity Localisation	14
2.5	Continual Learning	15
2.5.1	Continual Unsupervised Learning	15
2.5.2	Time-Varying Hidden Markov Models	15
2.6	Benchmark Datasets	16
2.6.1	ROAD	16
2.6.2	SARAS-ESAD	17
2.6.3	Thumos-14	17
2.6.4	ActivityNet-1.3	17
2.6.5	50-Salad	17
2.7	Evaluation Metrics	18
2.7.1	Classification Accuracy	18
2.7.2	Precision	18
2.7.3	Recall	18
2.7.4	F1-score	19
2.7.5	Mean Average Precision	19
2.7.6	Mean Absolute Error	19

Part I : ROAD (The ROad event Awareness Dataset for Autonomous Driving) **21**

3 The ROad event Awareness Dataset for Autonomous Driving **23**

3.1	Introduction	23
3.1.1	Motivation	24
3.1.2	ROAD: a multi-label, multi-task dataset	26
3.1.3	ROAD-R	28
3.1.4	Main Contributions	29
3.2	The Dataset	31
3.2.1	A Multi-Label Benchmark	31
3.2.2	Data Collection	32
3.2.3	Annotation Process	33
3.2.4	Tasks	34
3.2.5	Quantitative Summary	35
3.3	Baseline and Challenge	36
3.3.1	3D-RetinaNet Architecture	36
3.3.2	Online Tube Generation via Agentness Score	37
3.3.3	The ROAD Challenge	38
3.4	ROAD-R	39
3.5	Experimental Validation	41

3.5.1	Implementation Details	41
3.5.2	Experimental Results on ROAD	42
3.5.3	Qualitative Results	46
3.5.4	Challenge Results	46
3.5.5	Experimental Results of ROAD-R	49
3.6	Further Extensions	51
3.7	Summary of the Chapter	54
Part II : Complex Activity Detection		55
4	Spatiotemporal Deformable Scene Graphs for Complex Activity Detection	57
4.1	Introduction	57
4.1.1	Motivation	58
4.1.2	Objective	59
4.1.3	Main Contributions	59
4.2	Proposed Method	61
4.2.1	Action Tube Detection	62
4.2.2	3D Part-Based Deformable Layer	63
4.2.3	Graph Convolutional Network	63
4.2.4	Complete Framework	65
4.3	Experiments	67
4.3.1	Datasets and Evaluation Metrics	67
4.3.2	Action Tube Detection	69
4.3.3	Complex Activity Recognition	69
4.3.4	Temporal Activity Detection - Comparison with State-of-the-Art	70
4.3.5	Ablation Studies	72
4.4	Summary of the chapter	73
4.4.1	Limitations and Future Work	73
4.4.2	Conclusion	74
5	A Hybrid Graph Network for Complex Activity Detection in Video	75
5.1	Introduction	75
5.1.1	Motivation	76
5.1.2	Objectives	78
5.1.3	Contributions	79
5.2	Proposed Method	80
5.2.1	Tube Detection and Feature Extraction	81
5.2.2	Scene Graph Attention	82

5.2.3	Temporal Graph Localisation	83
5.2.4	Loss Functions	84
5.3	Experiments	85
5.3.1	Datasets	85
5.3.2	Implementation Details	86
5.3.3	Comparison with State-of-the-Art	87
5.3.4	Ablation Studies	90
5.4	Summary of the chapter	92
5.4.1	Future Work	92
5.4.2	Conclusion	93

Part III : Continual Learning 95

6 Continual Semi-Supervised Learning 97

6.1	Introduction	97
6.1.1	Motivation	98
6.1.2	Overview of the CSSL Challenge	98
6.1.3	Overview of Hidden Markov Model as Latent Variable	99
6.1.4	Main Contributions	100
6.2	Continual Semi-Supervised Learning Workshop and Challenge	102
6.2.1	Learning	102
6.2.2	Testing	102
6.2.3	Benchmarks	103
6.2.4	Challenges	105
6.2.5	Tasks	106
6.2.6	Baselines	107
6.3	HMM Latent Variable Model for Unsupervised Data-Streams	109
6.3.1	Principles of HMM Latent Variable Model	109
6.3.2	A Two-Stage Learning Process	110
6.3.3	Joint Distribution of the Data Stream	110
6.3.4	Discrete-State T-HMM for Continuous Classification	111
6.3.5	Random Sampling based Memory Replay	115
6.3.6	Alternative to HMM as Latent Variable Model	116
6.4	Experimental Analysis of Challenge Baselines	116
6.4.1	Results on Activity Recognition	117
6.4.2	Results on Crowd Counting	118
6.5	Experimental Analysis of HMM latent Variable	118
6.5.1	Datasets and Setups	118
6.5.2	Quantitative Analysis	120

6.6	Summary of the Chapter	121
7	Conclusions and Future Research Directions	123
7.1	Summary of the Thesis	123
7.1.1	ROad event Awareness Dataset for Autonomous Driving . .	123
7.1.2	Complex Activity Detection	124
7.1.3	Continual Semi-Supervised Learning	125
7.2	Future Research Directions	126
7.2.1	Extension of ROAD	126
7.2.2	Holistic Graph Approach for Complex Activity Detection and Prediction	126
7.2.3	Extension to Spatiotemporal Sentence Grounding	126
7.2.4	Continual Learning Extension to Regression	127
	References	129

List of Figures

1.1	Our complex activity recognition framework workflow for two videos from the ROAD (autonomous driving) and SARAS (surgical robotics) datasets. The overall concept of detecting complex activities via a part-based deformation mechanism is the same in both cases. However, in ROAD background frames exist which do not contain any complex road activity. In contrast, the SARAS complex activities are actually the phases of a surgical procedure, which are contiguous without the need for a background label.	4
3.1	Use of labels in ROAD to describe typical road scenarios. (a) A green car is in front of the AV while changing lanes, as depicted by the arrow symbol. The associated event will then carry the following labels: <i>in vehicle lane</i> (location), <i>moving left</i> (action). Once the event is completed, the location label will change to: <i>in outgoing lane</i> . (b) Autonomous vehicle turning left from lane 6 into lane 4: lane 4 will be the <i>outgoing lane</i> as the traffic is moving in the same direction as the AV. However, if the AV turns right from lane 6 into lane 4 (a wrong turn), then lane 4 will become the <i>incoming lane</i> as the vehicle will be moving into the incoming traffic. The overall philosophy of ROAD is to use suitable combinations of multiple label types to fully describe a road situation, and allow a machine learning algorithm to learn from this information.	25
3.2	Number of instances of each class of individual label-types, in logarithmic scale.	35
3.3	Proposed 3D-RetinaNet architecture for online video processing. . .	36
3.4	Success cases in which our model detects the actions and locations correctly, and only for those agents which are active. Ground truth bounding boxes and labels are shown in green, while the predictions of our model are shown in red.	47

3.5	Some of the failure modes of our model, shown detecting inactive agents and/or assigning incorrect action and location labels. Ground truth bounding boxes and labels are shown in green, while the predictions of our model are shown in red.	48
3.6	Examples of violations of $\{\neg\text{RedTL}, \neg\text{GreenTL}\}$	51
3.7	ROAD-R and SOTA models. In the x -axis, there is the threshold $\theta \in [0.1, 0.9]$, step 0.1.	52
4.1	Overall pipeline of our long-term complex activity detection framework. (i) The input video is first divided into snippets. (ii) Snippets are passed to an action tube detector module one by one. (iii) A part-based deformation module receives 3D VGG features and action tube locations and returns features for the salient (non-background) parts of each tube instance. (iv) A GCN module represents activity parts (action tubes) as nodes with the features generated by ROI pooling and builds edges with different semantics to construct a spatiotemporal graph representation. Finally (v), the graph representation features produced by GCN inference on the consolidated graph are used for temporal activity detection.	60
4.2	(a) AMTNet [1] generates micro-action tubes by predicting bounding boxes for the start and the end frame of a snippet. Predictions for a fixed number of intermediate frames are constructed by bilinear interpolation. (b) Complete action tubes are created by temporally linking micro action tubes by dynamic programming.	62
4.3	Our 3D deformable RoI pooling layer takes feature maps and action tube locations as input and arranges them into a fixed-size grid of components (here illustrated for size 3×3). For each grid component an offset is generated and multiplied by the original tube feature to produce the final component features.	64
4.4	Our graph module takes as input the features generated for each tube by the ROI pooling layer $T_{1,2,3...K}$ and builds edges between them according to different semantics. (order of appearance, spatial similarity, label). The overall graph is processed by a GCN to deliver a fixed-size graph representation.	65
4.5	Complex activity classification performance over ROAD dataset.	70
4.6	Complex activity classification performance on SARAS-ESAD dataset.	71

4.7	The performance of different components of our proposed method over ROAD and SARAS-ESAD dataset. The mAP at a threshold of 0.5 is reported for the test set of both datasets.	72
4.8	The performance comparison of various edge types in our GCN modules. The mAP at a threshold of 0.5 is reported for the test set of both datasets.	73
5.1	An overview of our Complex Activity Detection (CompAD) framework. The input video is divided into fixed-size snippets $S_1, \dots, S_N$ (top); each snippet is then processed in three major steps (bottom). A) Firstly, scene objects (agents) are detected and tracked throughout the snippet to form agent tubes. 3D features are then extracted from all the cropped agent tubes (f_{a^i}) as well as the local scene (f_{s_i}). B) Next, a fully-connected scene graph is constructed where agent nodes (in gray) are connected to each other and to the snippet node (in light blue). The local scene graph is processed using a graph attention network (GAT), resulting in intermediate scene features (f''_{s_i}). C) Finally, all local scene features associated with individual snippets are temporally connected and processed using a global temporal graph to identify the boundaries of the activity using anchor proposals (shown in different colors). Each of the anchors represents a binary graph mask where snippets which belong to the activity receive label $y_{br_i} = 1$, 0 otherwise.	76
5.2	Comparing TAL datasets (ActivityNet-1.3 and Thumos-14) with a CompAD dataset (ROAD) in terms of average activity and video durations and mean number of agents per snippet.	78
5.3	Visualisation of our agent detection and tracking stage using a bird's-eye view of the spatiotemporal volume corresponding to a video segment of <i>ActivityNet-1.3</i> dataset. The first half of the graph shows the plotting of agent tubes while the second half illustrates the scene graph of the detection agents. In this figure, we show examples of activities performed by multiple actors (agent tubes).	80
5.4	Visualisation of our agent detection and tracking stage using a bird's-eye view of the spatiotemporal volume corresponding to a video segment of <i>Thumos-14</i> dataset. The first half of the graph shows the plotting of agent tubes while the second half illustrates the scene graph of the detection agents. Both of the examples show the activities performed by multiple agents (recognising sports activities).	82

5.5	Visualisation of our agent detection and tracking stage using a bird’s-eye view of the spatiotemporal volume corresponding to a video segment of the <i>ROAD</i> dataset. The upper section shows an example of activity performed by only three agents. Below is the example of a night video where the activity is performed by multiple agents. The first half of the graph shows the plotting of agent tubes while the second half illustrates the scene graph of the detection agents. . . .	83
5.6	Our scene graph attention layer (Stage B of our approach, Fig. 5.1) takes the node features generated in the feature extraction Stage A as an input, and updates each node’s features using node attention with respect to its all connected neighbours. Multi-head attention with H heads is applied to each node to further robustify the representation – averaging yields the final node features. To obtain a fixed-size overall representation for a specific local scene (snippet), the features of all its nodes are aggregated using a learnable weight matrix W_2	84
5.7	Class-wise average precision (%) with IoU threshold of 0.5 for <i>ROAD</i> dataset.	87
5.8	Class-wise average precision (%) with IoU threshold of 0.5 for <i>Thumos-14</i> dataset.	89
5.9	Performance comparison with the state-of-the-art over all the three datasets considered, as a function of average activity duration and average no. of agents per snippet.	90
5.10	Average mAP of variants of our method with: attended scene graph (‘Agents with GAT’), scene graph (‘Agents without GAT’) and ‘Scene features only’, over the all three datasets.	91
5.11	The qualitative results of our proposed method for all three datasets. The green rectangles covering the snippets (local scenes) are the ground truth while the yellow boxes show the prediction of our model.	93
5.12	The qualitative results of our proposed method for all three datasets. The green rectangles covering the snippets (local scenes) are the ground truth while the yellow boxes show the prediction of our model.	94
6.1	Overview of the proposed latent variable pipeline.	100

6.2	Overall pipeline of our CAR baseline. 1) Firstly, a pre-trained EfficientNet model is fine-tuned over the supervised fold of the dataset (jointly over the 15 sequences). 2) The unlabeled validation and test folds are divided into subfolds. In each session video frames from each sub-fold are used in a self-training cycle in which a pseudo label P_L with prediction probability score P_s is generated for each frame, but only the pseudolabels with P_s above a fixed threshold (namely, 0.4) are considered as ground truth for retraining the model. Frames with lower P_s are ignored. The model updated by self-training in session n is used as initial model in session $n + 1$	107
6.3	Time-varying HMM (dashed bars) vs. traditional HMM (colored bars).	120

List of Tables

2.1	Comparison of ROAD with similar datasets for perception in autonomous driving in terms of diversity of labels. The comparison is based on the number of classes portrayed and the availability of action annotations and tube tracks for both pedestrians and vehicles, as well as location information. Most competitor datasets do not provide action annotation for either pedestrians or vehicles.	12
3.1	ROAD labels.	28
3.2	ROAD tasks and attributes.	34
3.3	Constraint statistics. Π_n is the set of constraints r in Π with $ r = n$, i.e., with n positive and negative labels. $\mathcal{C} = \{\neg A : A \in \mathcal{C}\}$. Each row shows the number of rules r with $ r = n$, and the average number of negative and positive labels in such rules.	40
3.4	Splits of training, validation and test sets for the ROAD dataset with respect to weather conditions. The table shows the number of videos in each set or split. For splits, the first figure is the number of training videos, the second number that of validation videos.	41
3.5	Frame-level results (mAP %) averaged across the three splits of ROAD. The considered models differ in terms of backbone network (2D, I3D, and Slowfast) and clip length (08 vs 32). The performance of YOLOv5 on agent detection is also reported. Detection threshold $\delta = 0.5$. Both validation and test performance are reported for each entry.	43
3.6	Video-level results (mAP %) averaged across the three ROAD splits. The models differ in terms of backbone network (2D, I3D and Slowfast) and test time clip length (08 vs 32). The performance of YOLOv5 on agent detection is also reported. Both validation and test performance are reported for each entry.	43

3.7	Number of video- and frame-level instances for each label (individual or composite), left. Corresponding frame-/video-level results ($mAP\%$) for each of the three ROAD splits (right). Val- n denotes the validation set for Split n . Results produced by an I3D backbone.	44
3.8	AV temporal action detection results (frame $mAP\%$) averaged across all three splits.	49
3.9	Results (in video- mAP) of the winning entries to the ICCV 2021 ROAD challenge compared with the Slowfast and YOLOv5 baselines, at a detection threshold of 0.2.	49
3.10	f- $mAP@0.5$ and f- $mAP@0.75$ for the current SOTA models.	52
4.1	List of ROAD complex activities with number of train and test snippets splits and brief description.	67
4.2	SARAS video phases with number of train and test snippets and a brief description.	68
4.3	Action tube detection performance on both the ROAD and SARAS-ESAD datasets. Both Frame- mAP and Video- mAP at different IoU thresholds are reported for evaluation.	69
4.4	Comparative analysis of temporal activity localisation performance on ROAD and SARAS-ESAD, reporting mAP (%) at five different IoU thresholds.	70
4.5	ROAD activity localisation performance (mAP , %) for each complex road activity, at a standard IoU threshold of 0.5.	71
4.6	SARAS-ESAD activity localisation performance (mAP , %) for each activity at a fixed IoU threshold of 0.5.	71
4.7	The performance comparison of different sequence lengths over both ROAD and SARAS-ESAD dataset, reporting mAP (%) at two different IoU thresholds.	73
5.1	Comparing our approach with the state-of-the-art methods for CompAD available on the ROAD dataset. The mAP at the various standard thresholds is reported. Best results are in bold and second best <u>underlined</u> .	87
5.2	Activity detection performance comparison with state-of-the-art methods on Thumos-14. mAP values at different IoU thresholds are reported for the test Thumos-14. Best results are in bold and second best <u>underlined</u> .	88

5.3	Activity detection performance comparison with state-of-the-art methods on ActivityNet-1.3. mAP values at different IoU thresholds are reported for validation sets of ActivityNet-1.3. Best results are in bold and second best <u>underlined</u>	89
5.4	Effect of local scene graph topologies and feature aggregation strategies on the performance of our proposal. The average mAP for the relevant IoU thresholds is reported for each dataset.	91
5.5	Performance of our method as a function of different snippet lengths over the three datasets. The average mAP for the relevant IoU thresholds is reported for each dataset.	92
5.6	Performance of our method as a function of different temporal graph lengths over the three datasets. The average mAP for the relevant IoU thresholds is reported for each dataset.	92
6.1	Performance of the initial supervised model versus that of the incrementally updated models (separately on V and T or in combination), using three standard evaluation metrics. For each metric we report both class average (C) and weighted average (W).	117
6.2	Evaluation on the CCC validation (V) and test (T) folds for different experimental setups. Here $sup-ft$, $upd-V$, and $upd-V+T$ refer to the supervised model fine-tuned on S , the model incrementally updated via self-training in sessions on the validation fold V , and the model incrementally updated on the combined validation V and test T folds, respectively. The average test MAE is reported for each model.	117
6.3	Time-varying HMM (TV HMM) vs. traditional HMM (HMM) based on one-off training, where P, R, and F1 are precision, recall, and F1-score, respectively.	119
6.4	Time-varying HMM (TV HMM) vs. traditional HMM (HMM) based on continual retraining, where P, R, and F1 are precision, recall, and F1-score, respectively.	120
6.5	Time-varying HMM (TV HMM) vs. traditional HMM (HMM) based on incremental update, where P, R, and F1 are precision, recall, and F1-score, respectively.	120

Nomenclature

Acronyms

C3D: Convolutional 3D Network

I3D: Inflated Convolutional 3D Network

mAP: mean Average Precision

RGB: Red Green Blue, 3-Channel Camera output

TAL: Temporal Action Localisation

AUC: Area Under the Curve

AV: Autonomous Vehicle

BB: Bounding Box

CompAD: Complex Activity Detection

CL: Constrained Loss

CO: Constrained Output

CAR: Continual Activity Recognition

CCC: Continual Crowd Counting

CSSL: Continual Semi-Supervised Learning

CNN: Convolutional Neural Network

EM: Expectation Maximisation

FPN: Feature Pyramid Network

fps: frames per second

GAT: Graph Attention Network

GCN: Graph Convolutional Network

HMM: Hidden Markov Model

HMC: Hierarchical Multi-label Classification

IoU: Intersection over Union

LSTM: Long Short-Term Memory recurrent neural network

MAE: Mean Absolute Error

MC: Multi-label Classification

1D: One Dimensional

RNN: Recurrent Neural Network

RE: ROAD Event

SSL: Semi-Supervised Learning

ESAD: the Endoscopic Surgeon Action Detection Dataset

ROAD: The ROad Awareness Dataset for Autonomous Driving

ROAD-R: The ROad Awareness Dataset with logical Requirements

3D: Three Dimensional (spatiotemporal for video)

T-HMM: Time-varying Hidden Markov Model

2D: Two Dimensional (spatially for image)

Chapter 1

Introduction

Scene understanding from videos has a significant role in the computer vision community due to its vast real-world applications such as autonomous driving [2, 3], surveillance [4], medical robotics [5] or team sports analysis [6]. The scene understanding problem can be addressed by a variety of techniques starting from simple video action recognition to complex activity detection. In this thesis, we address the problem in a comprehensive fashion, starting from individual event detection to holistic complex scene analysis (complex activity detection).

1.1 Video Action Recognition and Detection

Video action recognition is the task of identifying and classifying human actions in video sequences. *Video action detection*, on the other hand, is the process of identifying and locating the actions of an individual within a video, in both space and time. In the current literature, a number of convolutional neural networks (CNNs) frameworks are used to learn spatial features from individual video frames, which are then fed to recurrent neural networks (RNNs) to model the temporal dynamics across frames [7]. Another approach is to use 3D CNNs, which can learn spatiotemporal features directly from the video data [8].

One of the key challenges in video action recognition and detection is the need to deal with the large variability in the appearance and motion of the actions across different videos. To address this, researchers often use techniques such as data augmentation and transfer learning to improve the robustness of the models [9, 10]. Additionally, some researchers have proposed using attention mechanisms to allow the model to focus on the most relevant parts of the video when detecting and recognising actions [11]. Video action recognition and detection have applications to a wide range of areas such as surveillance, human-computer interaction and sports analytics, to name a few.

1.2 Event Detection

Video event detection refers to the task of identifying and localising specific events or activities in a video. Such activities may be performed by a single agent (object) or by multiple agents at the same time. Examples include activities such as sports events [12], traffic incidents [13], disaster management [14, 15] or criminal activities. The most popular deep learning-based method for video event detection is the two-stream CNN, which uses two streams of CNNs, one for processing the RGB frames and the other for processing the optical flow frames [1]. Another popular method is the 3D CNN, which uses 3D convolutional kernels that can learn spatiotemporal features from the video. These features can then be used to detect events over time, making 3D CNNs useful for tasks such as road events for decision-making in autonomous driving.

1.2.1 Road Event Detection

In this thesis, we define *road events* to be the collection of three labels (agent, action, and location) associated with a single “tube” (the spatiotemporal representation of a series of bounding boxes). The agent is an active road user (object); the event triple also contains the action(s) the latter performs (possibly more than one at the same time), and the location(s) in which this takes place (which may vary from the start to the end of the event itself), as seen from the point of the view of an autonomous vehicle.

1.3 Temporal Action Localisation

The ability to recognise and localise activities temporally from untrimmed videos is a research problem which has been attracting significant attention, due in part to the rapid growth of online video generation. The current state-of-the-art in action detection is *Temporal Action Localisation* (TAL) - whose approaches not only recognise the action label(s) present in a video, but also identify the start and end time of each action instance occurring.

TAL can be instrumental in generating sports highlights [16], understanding road scenes in autonomous driving [17], video summarisation in surveillance videos [18] and video captioning [19]. A number of TAL methods [20–29] have recently been proposed. Whereas various new datasets have been recently proposed, the two most common relevant benchmarks remain ActivityNet 1.3 [30], and Thumos-14 [31]. State-of-the-art performance on Thumos-14 has improved in four years by some 19% [32].

Almost all TAL approaches contemplate two major aspects: *features/scene representation* and *temporal localisation*. In the scene representation stage, snippets (continuous sequences of frames) are processed to understand the local scene in the video. Methods [27, 28] typically employ pre-extracted features obtained using a sequential learning model (e.g., I3D [33]), often pre-trained on the Kinetics [34] dataset. Such features are then processed, e.g., via a temporal or semantic graph neural network, by applying appropriate feature encoding techniques or by generating temporal proposals, in an object detection style [35]. In the second step of TAL, actions need to be temporally localised. Existing methods vary in the way this is done, e.g. through temporal graphs [21, 25], boundary regression and proposals generation [20, 22, 28] or encoder-decoder approaches [32]. All such methods, however, are only applicable to short- or mid-duration actions that last for a few seconds (e.g., a person jumping or pitching a baseball).

1.4 Complex Activity Detection

Complex activities, in opposition, can be defined as longer-term events, typically comprising a series of shorter actions. For example, an autonomous vehicle (AV) “negotiating a pedestrian crossing” presents a complex activity - first the AV drives along the road, then the traffic lights switches to red, the vehicle stops and several pedestrians cross the road. Eventually, the lights change back to green, and the AV drives off again. Whereas a number of studies have attempted to address the localisation of longer-term activities [36–38], these approaches mostly focus on activities performed by single actor. To the author’s knowledge, [36] is the only existing study which attempts to model truly complex activities, performed by multiple actors over an extended period of time. The work, however, relies upon the availability of heavily-annotated datasets which include granular labels for the individual actions which make up a complex activity, and the corresponding frame-level bounding boxes. A general overview of a complex activity framework is visualised in Figure 1.1.

1.5 Scene Graphs

A *scene graph* is a graph-based representation of a scene, where the nodes of the graph represent entities in the scene (such as objects, people, and locations) and the edges represent relationships between those entities (such as proximity, or interaction) [39]. Scene graphs are used in computer vision [40] and natural language processing [41] to represent the semantic structure of a scene and to facilitate tasks

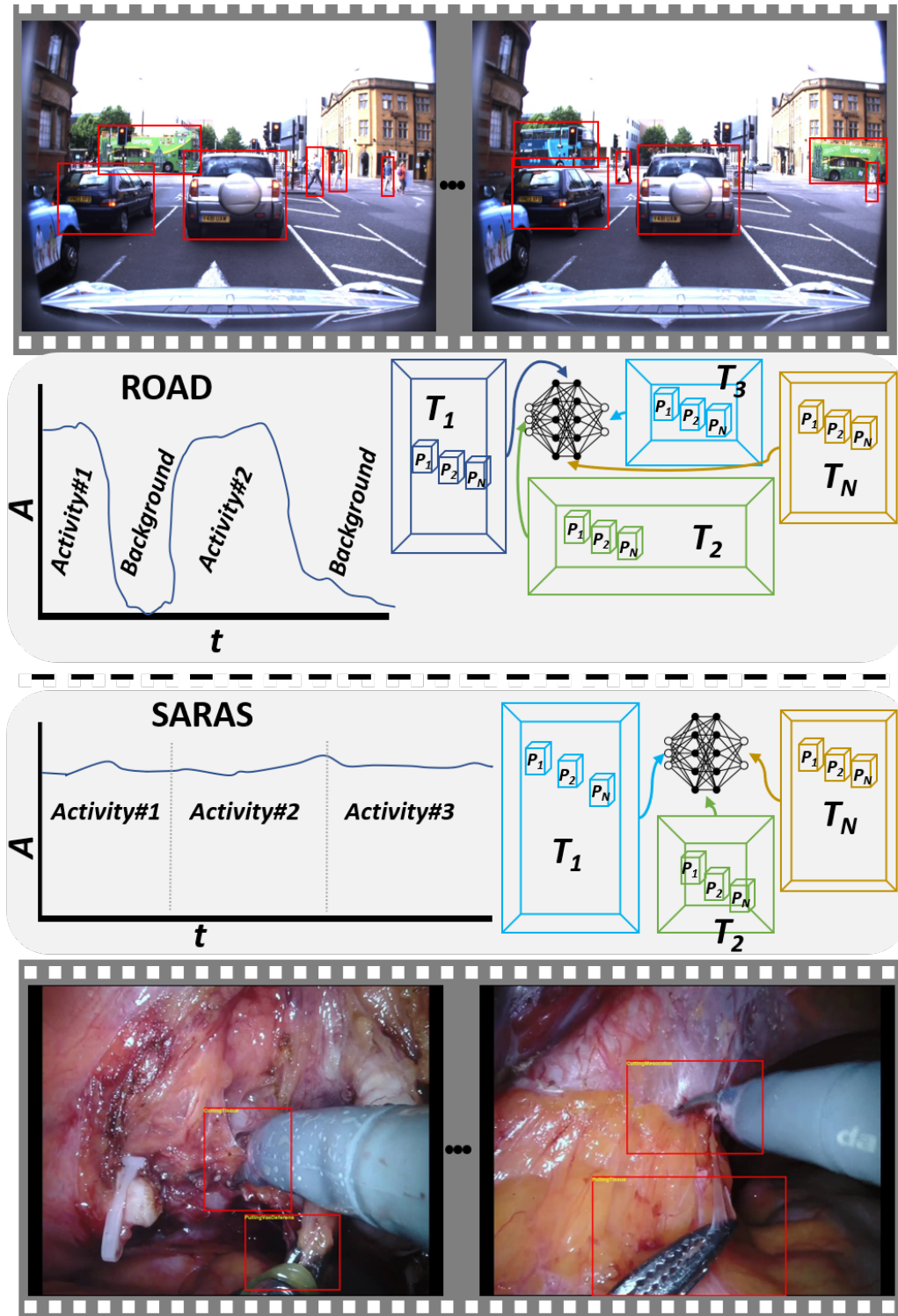


Figure 1.1: Our complex activity recognition framework workflow for two videos from the ROAD (autonomous driving) and SARAS (surgical robotics) datasets. The overall concept of detecting complex activities via a part-based deformation mechanism is the same in both cases. However, in ROAD background frames exist which do not contain any complex road activity. In contrast, the SARAS complex activities are actually the phases of a surgical procedure, which are contiguous without the need for a background label.

such as object detection, scene understanding, and natural language generation.

Scene graphs are typically constructed by analysing images or videos, using techniques such as object detection, instance segmentation, and scene understanding [42]. The nodes of the graph are typically labeled with object categories, and the edges are labeled with the type of relationship between the entities.

Due to the vast importance of scene graphs in scene understanding, in this thesis, we also adopt a scene graph representation for modelling complex activities.

1.6 Continual Learning

Continual learning refers to the ability of a machine learning model to learn and adapt to new information over time, without forgetting previously acquired knowledge [43]. In other words, it allows a model to continuously improve its performance by learning from new data, even after it has been deployed in the real world. This is an important aspect of developing AI systems that can function in dynamic, real-world environments and adapt to changing conditions [44]. Continual learning is an active area of research in machine learning, with the goal of developing models that can learn and adapt throughout their lifetime. A typical example is that of an object detector which needs to be extended to include classes not originally in its list (e.g., ‘donkey’) while retaining its ability to correctly detect, say, a ‘horse’. The hidden assumption there is that we are quite satisfied with the model we have, and we merely wish to extend its capabilities to new settings and classes.

In the last part of the thesis, we also proposed two continual learning paradigms in Chapter 6 for updating the machine learning model continually in a semi-supervised manner.

1.7 Contributions of the Thesis

The main contributions of this thesis are the following:

1. A ROad event Awareness Dataset for Autonomous Driving (ROAD), the first of its kind, designed to support a paradigm shift toward the detection of semantically meaningful road events, and allow the testing of a range of tasks related to situation awareness for autonomous driving: agent and/or action detection, event detection, ego-action classification.
2. The ROad event Awareness Dataset with logical Requirements (ROAD-R), the first publicly available dataset whose commonsense requirements are expressed in full propositional logic, equipped with baseline models.

3. A spatiotemporal scene graph generation and processing mechanism for complex activity detection able to cope with a variable number of parts while learning the overall semantics of an activity class.
4. An extension of 2D deformable RoI pooling to 3D for flexibly pooling features from the various components of the detected tubes, to create an overall representation for activity parts.
5. A hybrid graph network approach for general complex activity detection, comprising a local scene graph as well as a global temporal graph, capable of localising both long and short-term activities.
6. A feature extraction strategy for fine-grained temporal activity localisation without the need for additional annotation, consisting in detecting and tracking the relevant scene agents to construct tubes, followed by feature extraction from both the whole scene and each agent tube individually, as opposed to existing methods which only rely on snippet-level features.
7. A formal definition of the continual semi-supervised learning (CSSL) problem, and the associated training and testing protocols, with a suitable baseline. The first two benchmark datasets for the validation of semi-supervised continual learning approaches, one for classification (continual activity recognition) and one for regression (continual crowd counting), which we propose as the foundations for the first challenge in this new domain area.
8. A principled two-stage approach to the CSSL problem, based on a latent model formulation which reduces the task to continual supervised learning.

1.8 Organisation of the Thesis

Firstly, we review all the relevant related work, including the benchmark datasets and evaluation metrics, in Chapter 2. Next, the main body of the thesis (Chapters 3 to 7) is divided into three parts; Part I, Part II, and Part III. Finally, the thesis is concluded in Chapter 7 by a summary and a discussion of possible future extensions.

The theme of Part I 2.7.6 is ROAD (The ROad event Awareness Dataset for Autonomous Driving) which includes two chapters 3. In Chapter 3, we present the baseline ROAD dataset, which is designed to test an autonomous vehicle’s ability to detect road events, defined as triplets composed by an active agent, the action(s) it performs, and the corresponding scene locations. We also introduce the ROad event Awareness Dataset with logical Requirements (ROAD-R) in Part I, the first publicly

available dataset for autonomous driving with requirements expressed as logical constraints, as a neurosymbolic extension to ROAD.

Next, in Part II 3.7 we present the concept of Complex Activity Detection (CompAD), over two Chapters (Chapter 4 and Chapter 5). Chapter 4 presents a novel deformable, spatiotemporal scene graph approach for CompAD, consisting of three main building blocks: (i) action tube detection, (ii) the modelling of the deformable geometry of parts, and (iii) a graph convolutional network. The latter is further improved in Chapter 5 by addressing the CompAD problem using a hybrid graph neural network that combines attention applied to a graph encoding the local (short-term) dynamic scene with a temporal graph modelling the overall long-duration activity.

Finally, Part III 5.4.2 is about continual learning and is composed of a single Chapter 6. There, we present a new continual semi-supervised learning (CSSL) paradigm, which we proposed to the attention of the machine learning community via the IJCAI 2021 International Workshop on Continual Semi-Supervised Learning. In addition to the baseline made available for the challenge, we also propose to formulate the continual semi-supervised learning problem as a latent-variable one, in which the labels of a series of unlabelled, streaming data points play the role of the hidden states, while the instances play that of the observations.

1.9 Publications

Part of thesis: the following is the list of publications that form the core of this PhD thesis.

1. **Salman Khan**, Izzeddin Teeti, Andrew Bradley, Mohamed Elhoseiny, and Fabio Cuzzolin, “A Hybrid Graph Network for Complex Activity Detection in Video”, Pre-print arXiv, (2023).
2. Gurkirt Singh, ..., **Salman Khan**, ..., and Fabio Cuzzolin, “ROAD: The ROad event Awareness Dataset for autonomous Driving”, in IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI) 45, pp. 1036-1054, January 2023.
3. E. Giunchiglia, **Salman Khan**, ..., F. Cuzzolin and T. Lukasiewicz, “ROAD-R: the Autonomous Driving Dataset for Learning with Requirements”, Machine Learning, Springer, September 2022 (accepted with minor revision)
4. A. Shahbaz, **Salman Khan**, ... and F. Cuzzolin, “International Workshop on Continual Semi-Supervised Learning: Introduction, Benchmarks and Baselines”, in CSSL 2021, Lecture Notes in Artificial Intelligence Vol. 13418, pages 1-14, Springer, 2022

5. **Salman Khan** and Fabio Cuzzolin, “Spatiotemporal Deformable Scene Graphs for Complex Activity Detection”, in Proceedings of The 32nd British Machine Vision Conference (BMVC), 2021.
6. Eleonora Giunchiglia, Mihaela Catalina Stoian, **Salman Khan**, Fabio Cuzzolin and Thomas Lukasiewicz, “ROAD-R: The Autonomous Driving Dataset with Logical Requirements”, in Proceedings of IJCAI 2022 Workshop on Artificial Intelligence for Autonomous Driving (AI4AD 2022) and IJCLR 2022.
7. Ajmal Shahbaz, **Salman Khan**, Mohammad Asiful Hossain, Vincenzo Lomonaco, Kevin Cannons, and Fabio Cuzzolin, International Workshop on Continual Semi-Supervised Learning: Introduction, Benchmarks and Baselines, in Proceedings of IJCAI Workshop, International Workshop on Continual Semi-Supervised Learning (2021).
8. Fabio Cuzzolin, **Salman Khan**, Ajmal Shahbaz “Continual Semi-Supervised Learning for Classification via a Time-Varying Hidden Markov Model”, Preprint arXiv 2023.

Other works: Below is the list of other works which I published in the course of my PhD term, but are not directly connected to the topic of this thesis.

1. Muhammad, Khan, Hayat Ullah, **Salman Khan**, Mohammad Hijji, and Jaime Lloret. “Efficient Fire Segmentation for Internet-of-Things-Assisted Intelligent Transportation Systems.” IEEE Transactions on Intelligent Transportation Systems, September 2022.
2. **Salman Khan**, Khan Muhammad, Tanveer Hussain, Javier Del Ser, Fabio Cuzzolin, Siddhartha Bhattacharyya, Zahid Akhtar, and Victor Hugo C. de Albuquerque. “DeepsMOKE: Deep learning model for smoke detection and segmentation in outdoor environments.” Expert Systems with Applications 182 (2021): 115125.
3. Teeti, Izzeddin, **Salman Khan**, Ajmal Shahbaz, Andrew Bradley, and Fabio Cuzzolin. “Vision-based Intention and Trajectory Prediction in Autonomous Vehicles: A Survey.” Proceedings of IJCAI 2022 (Survey track), pages 5630-5637.
4. Teeti, Izzeddin, Valentina Musat, **Salman Khan**, Alexander Rast, Fabio Cuzzolin, and Andrew Bradley. “Vision in adverse weather: Augmentation using CycleGANs with various object detectors for robust perception in autonomous racing.” ICML 2022 - Workshop on Safe Learning for Autonomous Driving (SL4AD 2022).

5. Onyeogulu, Tochukwu, Amirul Islam, **Salman Khan**, Izzeddin Teeti, and Fabio Cuzzolin. "Situation Awareness for Automated Surgical Check-listing in AI-Assisted Operating Room." arXiv preprint arXiv:2209.05056 (2022).
6. Jin, Kaizhe, Adrian Rubio-Solis, Ravi Naik, Tochukwu Onyeogulu, Amirul Islam, **Salman Khan**, Izzeddin Teeti et al. "Identification of Cognitive Workload during Surgical Tasks with Multimodal Deep Learning." arXiv preprint arXiv:2209.06208 (2022).

1.10 List of Packages and Demos

- All resources of our ROAD project (IEEE TPAMI 2022 [45] and The ROAD Workshop and Challenge: Event Detection for Situation Awareness in Autonomous Driving, ICCV 2021) are publicly available at:

Workshop and Challenge link: <https://sites.google.com/view/roadchallengeiccv2021/>

ROAD dataset: <https://github.com/gurkirt/road-dataset>

3D RetinaNet baseline: <https://github.com/gurkirt/3D-RetinaNet>

Inference demo: <https://www.youtube.com/watch?v=CmxPjHhiarA>

Challenge evalAI <https://eval.ai/web/challenges/challenge-page/1059/overview>.

- The resources of our ROAD-R project [46] are publicly available at:

code: <https://github.com/EGiunchiglia/ROAD-R>

video: https://www.youtube.com/watch?v=_-Ll5d1VQXY.

- The source code of our BMVC 2021 [17] is publicly available at: https://github.com/salmank255/SDSG_Complex_Activity.

- The source code of our CompAD paper Pre-print is publicly available at: https://anonymous.4open.science/r/HG_CompAD-7697/README.md.

- All resources of the continual learning project (Continual Semi-Supervised Learning Workshop and Challenge at IJCAI 2021 [47]) are publicly available at:

Workshop link: <https://sites.google.com/view/sscl-workshop-ijcai-2021/>

Code: <https://github.com/salmank255/cssl>

CAR challenge baseline: <https://github.com/salmank255/IJCAI-2021-Continual>

CAR evalAI: <https://eval.ai/web/challenges/challenge-page/984/overview>

CCC challenge baseline: https://github.com/Ajmal70/IJCAI_2021_Continual_Crowd_Counting_Challenge

CCC evalAI: <https://eval.ai/web/challenges/challenge-page/986/overview>.

Chapter 2

Related Work

This chapter presents a literature review of the most relevant methods in the areas (discussed in Chapter 1) which most contribute to scene understanding. We start from a comparison of the existing major autonomous driving datasets (Section 2.1) with our ROAD dataset, which is presented in Chapter 3. Next, we provide a literature review of temporal action localisation, complex activity detection, graph convolutional networks, and continual learning in Sections 2.2, 2.3, 2.4, and 2.5, respectively. We also discuss the benchmark datasets used for the evaluation of our proposed methods in Section 2.6. Lastly, we introduce all the evaluation metrics (Section 2.7) used for the evaluation of the different tasks targeted in this thesis.

2.1 Autonomous Driving Datasets

In recent years a multitude of AV datasets have been released, mostly focusing on object detection, scene segmentation, trajectory prediction, and action detection. We can categorise them into two main bins: (1) RGB without range data (single modality) which is relatively less time-consuming and expensive than building multimodal datasets including range data from LiDAR or radar and (2) RGB with range data (multimodal) that contain multiple types of data, such as lidar, camera, and radar data.

To compare ROAD (proposed in Chapter 3) with autonomous driving perception datasets, we provide a comparative analysis of the recent state-of-the-art datasets in Table 2.1. As it can be noted in the table, the unique feature of ROAD is its diversity in terms of the types of actions and events portrayed, by all types of road agents in the scene. With 12 agent classes, 30 action classes and 15 location classes ROAD provides (through a combination of these three elements) a much more refined description of road scenes.

Table 2.1: Comparison of ROAD with similar datasets for perception in autonomous driving in terms of diversity of labels. The comparison is based on the number of classes portrayed and the availability of action annotations and tube tracks for both pedestrians and vehicles, as well as location information. Most competitor datasets do not provide action annotation for either pedestrians or vehicles.

Dataset	Class Num.	Ann. frames	Countries	Loc. lab.	Avg. obj./f	Action Ann		Tube Ann	
						Ped.	Veh.	Ped.	Veh.
SYNTHIA [48]	13	13400	1	pixelwise ann.	-	-	-	-	-
Cityscapes [49]	30	5000	2	pixel level sem.	-	-	-	-	-
Waymo [50]	4	1,200,000	1	-	20	-	-	✓	✓
ApolloScape [51]	25	140,000	1	pixel level sem.	-	-	-	✓	✓
PIE [52]	6	293,437	1	-	2.5	✓	-	✓	-
TITAN [53]	50	75,262	1	-	5.2	✓	✓	✓	✓
KITTI360 [54]	19	83,000	1	sem. ann.	-	-	-	-	-
A*3D [55]	7	39,000	1	-	-	4.8	-	-	-
H3D [56]	8	27,721	1	-	16	-	-	✓	✓
Argoverse [57]	15	70,557	1	-	18	-	-	✓	✓
NuSense [2]	23	40,000	2	3D sem. seg.	20	-	-	✓	✓
DriveSeg [58]	12	5,000	1	sem. ann.	-	-	-	-	-
Spatiotemporal action detection datasets									
UCF24 [59]	24	80175	-	-	1.5	✓	-	✓	-
AVA [60]	80	1,600,000	-	-	3	✓	-	✓	-
Multisports [61]	66	37701	-	-	2.5	✓	-	✓	-
ROAD [45]	43	122,000	1	✓	5	✓	✓	✓	✓

Ped. Pedestrian, Veh. Vehicle, ann. annotation, sem. seg. semantic segmentation

2.2 Temporal Action Localisation

Significant progress has been made in recent years toward temporal action localisation (TAL) in untrimmed videos. The main goal of TAL is to remove the irrelevant background from the input videos and identify the start and end instants of the action of interest. Temporal localisation methods can be divided into supervised [62, 63] versus weakly/unsupervised ones [64–66]. While the former requires temporal annotations to learn a model able to identify the start and end of an action, in weakly-supervised settings the model is learned from partial annotation.

In Part II of the thesis, our goal is the same as temporal activity localisation, with the difference that we wish to target *complex, long duration events articulated into a possibly large number of parts* (complex activity detection), whose internal structure cannot be leveraged by naive temporal segmentation methods.

2.3 Complex Activity Detection

Most recent work on complex activity recognition concerns scalar sensors [67–69] or a combination of both scalar and vision sensors [70, 71]. Recently, though, several vision-based complex activity recognition methods have been proposed [40, 63, 72–75] with the goal of understanding an overall scene by recognising and segmenting atomic actions. These methods can be further divided into (i) sliding windows ap-

proaches [38, 76], in which an activity classifier is applied to each snippet, and (ii) boundaries analyses [77, 78], in which a model is trained to identify the start and end time of each action. Overall, current activity recognition methods are geared to recognise short-term activities via a combination of small atomic actions.

Unlike existing approaches, the objective of our complex activity model (proposed in Chapter 4) is to understand *long-term* activities in dynamic scenes, such as the *phases* a surgical procedure is broken into, whose detection is crucial to inform the decision making of automated robotic assistants.

2.3.1 One-Stage vs Two-Stage Approaches

Complex activity methods can be broadly divided into *one-stage* and *two-stage* approaches. The former [23, 79–81] detect actions/activities in a single shot, and can be easily trained in an end-to-end manner. For example, Wang et al. [82] detect actions using transformers which, unlike RNNs, do not suffer from nonparallelism and vanishing gradients. Most one-stage methods, however, only perform action classification, rather than spatiotemporal localisation. In contrast, Lin et al. [81] have recently proposed an anchor-free, one-stage light model which generates proposals, locates actions within them, and classifies them end-to-end.

The latter group of methods [20, 22, 62, 83, 84], instead, consists of two stages, similarly to region-proposal object detectors. The first stage generates suitable proposals for predicting the start and end time of an activity, while the second stage extracts features and processes the proposals before passing them to both a classification head and a regression head (for temporal localisation). Some work, including [22], focus on the first stage to improve the quality of the proposals, while others focus on the processing or refining of the proposals. [84], for instance, uses an off-the-shelf method for proposal generation. The second stage consists of two networks, a ‘disentanglement’ network to separate the classification and regression representations, and a ‘context aggregation’ network to add them together. Such methods are not trainable end-to-end and limited to short- or mid-duration actions. In contrast, we propose a *complex activity detection* method in Chapter 5, which exhibits the advantages of both classes of methods, thanks to our hybrid graph approach capable of recognising and localising both short- and long-duration activities.

2.4 Graph Networks

Recently, Graph Convolutional Networks (GCNs) have been widely used for action and activity detection and recognition, building on their success in different areas of computer vision such as point cloud segmentation [85, 86] and 3D object detec-

tion [87]. Relevant GCN approaches have been broadly focussing on either action recognition [88–90] or TAL [21, 25]. In the former, videos are represented in different spatiotemporal formats such as 3D point clouds and time-space region graphs, and methods focus on recognising atomic actions only. In contrast, Zeng et al. [21] use GCN for temporal activity localisation by considering action proposals as nodes and a relation between two proposals as an edge. In opposition, in our model nodes are action tubes and their connections are based on an array of semantics. In another recent study, Xu et al. [25] generate graphs by considering temporal snippets as nodes and drawing connections between them based on temporal appearance and semantic similarity.

Most graph-based activity detection methods [21, 25, 91] construct a graph for a whole video by taking snippets as nodes and their temporal linkage as edges, not paying much attention to the constituent atomic actions within each snippet, and are typically limited to shorter videos or memory dependent in case of larger videos due to adjacency matrices. For a large number of snippets (nodes) in the video required memory can be substantial. The adjacency matrix of the graph with n nodes (snippets) requires $O(n^2)$ memory, which can be impractical for large videos.

In contrast, our proposed framework presented in Chapter 4 is designed to construct a scene graph for each snippet rather than the whole video which reflects the structure of a dynamic scene in terms of atomic action tubes (nodes) and the different types of relationships between them.

2.4.1 GCN-Based Temporal Activity Localisation

GCNs have been extensively investigated for TAL [21, 25, 62, 92, 93]. GCN-based TAL methods can also be further divided into two-stage and one-stage methods. The former, once again, perform localisation after generating suitable proposals. For instance, in [21] two different types of boundary proposals are generated and then individually passed to the same graph, resulting in both an action label and a temporal boundary.

One-stage GCN-based TAL methods, instead, solve the detection problem without proposals in one go by learning spatiotemporal features in an end-to-end manner. For examples, in [25] a graph is first generated by connecting the snippets both temporally and by virtue of meaningful semantics. This graph is then divided into sub-graphs (anchors), where each anchor represents the activity in an untrimmed video. In contrast, Khan et al. [17] proposed a spatiotemporal scene graph-based long-term TAL method where each of the snippets is considered as a separate graph, which is heavily dependent on the particular actions present in the scene, and is only applicable to datasets providing (label and bounding box) annotations for each individual

action.

In Chapter 5, we propose a new framework leveraging GCNs, by incorporating them in an overall hybrid graph capable of modelling both the local scenes, via a Graph Attention Network (GAT) [94], and the overall global activity via a temporal graph. GATs build on the transformer concept by applying attention to graphs, and were originally proposed in [94] for node classification. The idea is to update the representation of the current node with respect to its neighbours by applying attention to learn the importance of the various connections.

In Chapter 5, GAT is used at the local scene level to learn the features of each node (active agent tube), to generate a more robust local scene representation.

2.5 Continual Learning

Current continual learning methods can be categorised into three major families based on how the information of previous data is stored and used. *Prior-focused* methods [95] use a penalty term to regularise the parameters rather than a hard constraint. *Parameter isolation* methods [96] dedicate different parameters for different tasks to prevent interference. Finally, *replay-based* approaches store the examples' information in a 'replay buffer' or a generative model, which is used for rehearsal/retraining or to provide constraints on the current learning step [97].

2.5.1 Continual Unsupervised Learning

Continual unsupervised learning is a relatively unexplored territory due to the heavy dependency of existing works on the availability of ground truth information. Another challenge lies in the clear distinction of class boundaries for controlling model updates [98]. Clustering approaches that provide structure to the unlabelled stream are a popular theme of existing continual unsupervised works [98–100]. A self-adaptive clustering approach [98] is proposed which does not rely on labelled data for model updates. The labelled data is initially leveraged to learn the cluster-to-class associations. Similarly, in [99] the pseudo labels for new data are obtained by a K-means global clustering approach, whereas the updated model without fully connected layers is employed as a feature extractor.

2.5.2 Time-Varying Hidden Markov Models

A number of competing formulations exist for HMMs whose parameters depend on time. They have been employed recently for useful life prediction in engineering [101], analysing time-varying networks in EEG data [102], and incorporating

periodic variability in animal movement [103]. Hidden *semi-Markov* models [104], in particular, have been proposed, which augment HMMs with time-varying parameters, so that the transition probability to a new state varies in accordance with the time spent in the current state. In opposition, the TVHMM model formulated in [105] encodes non-stationarity into a finite sequence of time varying transition densities and can simulate state duration of infinite length. A common approach [103, 106] is to model the transition probabilities as functions of a covariate process $\{z_t\}$, which is observed in parallel to the main observation process [107]. As the covariates z_t are simultaneously correlated with the hidden variables x_t , it may be preferable to use lagged informative variables or artificial variables. In fact, in [108] a time-varying hidden Markov model with latent information was proposed in which an additional hidden variable (‘signal’) with AR(1) dynamics drives the transition probabilities in the main hidden variable, x_t .

In [109], the authors observed that the states respond more to their own past values than any other variable. They thus proposed a model in which the state depends on the sign of an autoregressive latent variable, which induces time-varying probabilities in a non-Markovian process, as the probability of a certain state at time t depends on the lagged value of the latent variable and not the state at the previous time.

2.6 Benchmark Datasets

In the literature, several datasets have been proposed to evaluate the performance of scene understanding from simple action detection to event detection and complex activity recognition. Here, in this chapter, we describe all the datasets used in this thesis.

2.6.1 ROAD

ROAD (The ROad event Awareness Dataset for Autonomous Driving) [45] is a multi-labeled dataset proposed for road agent, action, and location detection. The combination of these three labels is referred to in [45] as a ‘road event’. The ROAD dataset consists of a total of 22 videos with an average duration of 8 minutes, captured by the Oxford RobotCar [110] under diverse lighting and weather conditions. The dataset was further extended as a testbed for complex activity detection in [17]. Complex activities in the ROAD dataset belong to six different classes, including: negotiating an intersection, negotiating a pedestrian crossing, waiting in a queue, merging into the (ego) vehicle lane, sudden appearance (of other vehicles/agents), and (people) walking in the middle of the road. Activities can span up to two minutes and in-

volve a large number of road agents. We introduce ROAD in Chapter 3 and evaluate our 3D-RetinaNet baseline in the same chapter. We also use ROAD extension for complex activity detection in Chapter 4 and Chapter 5

2.6.2 SARAS-ESAD

ESAD (the Endoscopic Surgeon Action Detection Dataset) [111] is a benchmark devised for surgeon action detection in real-world endoscopic surgery videos. The goal of ESAD is contributing to increasing the potential and dependability of surgical assistant robots by realistically testing their awareness of the actions performed by a surgeon. In ESAD, surgeon actions are classed into 21 different categories and annotated with the help of professional surgeons. This dataset is used in Chapter 4 for complex activity detection.

2.6.3 Thumos-14

Thumos-14 [31] is one of the benchmark datasets for temporal action localisation. Thumos-14 contains 413 untrimmed temporally annotated videos categorised into 20 actions. Videos are characterised by a large variance in duration, from one second to 26 minutes. On average, each video contains 16 action instances. To compare our performance with the state-of-the-art, we also adopt the standard practice of using the validation set (200 videos) for training while evaluating our model on the test set (213 videos). We use this dataset for the evaluation of our complex activity framework in Chapter 5.

2.6.4 ActivityNet-1.3

ActivityNet-1.3 [30] is one of the largest action localisation datasets with around 20K untrimmed videos comprising 200 action categories. The videos are divided into training, validation, and testing folds according to a ratio of 2:1:1, respectively. The number of action instances per video is 1.65, which is quite low compared to Thumos-14. Following the previous art, we train our model on the training set and test it on the validation set. We use this dataset in Chapter 4 for the evaluation of our complex activity framework.

2.6.5 50-Salad

50-Salad [112] is an activity segmentation dataset with the aim of stimulate research on recognizing manipulative gestures. It consists of 50 videos of 25 cooks (actors)

preparing salads, where each cook recorded two videos of themselves preparing salads by following different sequence of steps. The dataset contains two levels of annotations: “high-level” labels only span three classes, whereas “low-level” annotation involves a total of 51 classes. 50-Salad is a multimodal dataset contains other modality besides RGB videos including RGB-D (depth information), 3-axis accelerometer devices attached to a knife, a mixing spoon, a small spoon, a peeler, a glass, an oil bottle, and a pepper dispenser. This dataset is used in Chapter 6 for the evaluation of our latent model.

2.7 Evaluation Metrics

We also provide the standard evaluation metrics used for the evaluation of our methods proposed in this thesis.

2.7.1 Classification Accuracy

Classification accuracy is a measure of how well a classification model is able to correctly predict the class labels of a given dataset. It is defined as the proportion of correctly classified instances to the total number of instances in the dataset. Classification accuracy is used in Chapter 4 for recognition of activities.

2.7.2 Precision

Precision is a measure of the accuracy of a classification model’s positive predictions. It is defined as the proportion of true positive predictions (correctly predicted positive instances) to the total number of positive predictions (true positives plus false positives).

A high precision value means that the model has a low number of false positives, or that it is good at not classifying negative instances as positive. In other words, it means that the model is good at identifying only the relevant instances among the positive predictions. Precision is used in Chapter 4 and Chapter 6 for recognition of activities.

2.7.3 Recall

Recall, also known as Sensitivity or the True Positive Rate (TPR), is a measure of a classification model’s ability to correctly identify all relevant instances. It is defined as the proportion of true positive predictions (correctly predicted positive instances) to the total number of actual positive instances (true positives plus false negatives).

A high recall value means that the model has a low number of false negatives, or that it is good at identifying all relevant instances. In other words, it means that the model is good at detecting all the positive instances among all the actual positive instances. Recall is used in Chapter 4 and Chapter 6 for recognition of activities.

2.7.4 F1-score

The F1 Score is a measure of a classification model's accuracy that combines precision and recall. It is defined as the harmonic mean of precision and recall.

The F1 Score is a single value that represents the balance between precision and recall. A higher F1 Score indicates that the model has a better balance between precision and recall.

It is particularly useful when the classes are imbalanced, or when the cost of false positives and false negatives is different. In these cases, it is possible to have a high accuracy score but a low F1 score if precision and recall are not in balance. F1 score is used in Chapter 4 and Chapter 6 for recognition of activities.

2.7.5 Mean Average Precision

Mean Average Precision (mAP) is a common evaluation metric used in action detection, which is a subfield of computer vision that deals with recognizing and locating instances of specific actions within video sequences. MAP calculates the average precision across all classes of actions in a video dataset, by taking into account both the precision and recall of the action detection algorithm.

Precision measures the proportion of true positive detections (correctly detected actions) to the total number of positive detections (both true positives and false positives). Recall measures the proportion of true positive detections to the total number of instances of the action that should have been detected.

mAP is calculated by first computing the average precision for each action class, and then taking the mean of those values. A higher MAP value indicates a better performance of the action detection algorithm. This evaluation metric is used for two purposes; firstly for the spatiotemporal detection of events in Chapter 3. Secondly, used for the temporal evaluation of complex activities in Chapter 4 and Chapter 5.

2.7.6 Mean Absolute Error

Mean Absolute Error (MAE) is a measure of the average magnitude of the errors in a set of predictions, without considering their direction. It measures the difference between the predicted values and the true values.

MAE is a commonly used regression evaluation metric and it is relatively easy to understand and interpret. It is expressed in the same unit as the target variable and it gives an idea of the magnitude of the error, regardless of the direction of the error. MAE is used in Chapter 6 for the evaluation of continual crowd counting (regression problem).

**Part I : ROAD (The ROad event Awareness Dataset for
Autonomous Driving)**

Chapter 3

The ROad event Awareness Dataset for Autonomous Driving

3.1 Introduction

As anticipated, the first part of this thesis concerns individual event detection, specifically the detection of road events from an autonomous driving perspective. Injecting capabilities such as the understanding of dynamic road events and their evolution into autonomous vehicles has the potential to take situational awareness and decision-making closer to human-level performance. To this purpose, in this chapter, we introduce the ROad event Awareness Dataset (ROAD) for Autonomous Driving. ROAD is designed to test an autonomous vehicle’s ability to detect road events. To the best of our knowledge, ROAD is the first multi-label multi-action autonomous driving dataset where each of the bounding box is associated with three labels (agent, action, and location). This unique labeling schema makes ROAD more demanding and challenging in comparison to its competitors. ROAD comprises videos originally from the Oxford RobotCar Dataset, annotated with bounding boxes showing the location in the image plane of each road event. We benchmark various detection tasks, proposing as a baseline a new incremental algorithm for online road event awareness termed 3D-RetinaNet. We also report the performance on the ROAD tasks of Slowfast and YOLOv5 detectors, as well as that of the winners of the ICCV2021 ROAD challenge, which highlights the challenges faced by situation awareness in autonomous driving. ROAD is designed to allow scholars to investigate exciting tasks such as complex (road) activity detection, future event anticipation and continual learning.

Further down the line, we also introduce logical constraints to ROAD. The main intuition behind these logical constraints is that neural networks have proven to be very powerful at computer vision tasks. However, they often exhibit unexpected behaviours, violating known requirements and expressing background knowledge.

This calls for models (i) able to learn from the requirements, and (ii) be guaranteed to be compliant with the requirements themselves. Unfortunately, the development of such models is hampered by the lack of datasets equipped with formally specified requirements. Therefore, we introduce the ROad event Awareness Dataset with logical Requirements (ROAD-R), the first publicly available dataset for autonomous driving with requirements expressed as logical constraints.

3.1.1 Motivation

The history of *autonomous driving* has witnessed significant milestones and advancements since the Defense Advanced Research Projects Agency (DARPA¹) Grand Challenge 2004, which marked a pivotal moment in the field. The DARPA Grand Challenge was a groundbreaking competition organized by the U.S. Department of Defense’s DARPA agency. Its aim was to accelerate the development of autonomous vehicle technology. The challenge required participating vehicles to navigate a designated route in a desert environment, relying solely on their autonomous capabilities. The main purpose of the Challenge was to push the boundaries of autonomous vehicle technology. The goal was to foster innovation and encourage advancements that could potentially benefit military applications, such as unmanned ground vehicles for reconnaissance or supply missions [113]. The DARPA Grand Challenge helped establish autonomous driving as a viable research area and laid the foundation for subsequent advancements in the field.

In recent years, *autonomous driving* has emerged as a fast-growing research area. The race towards fully autonomous vehicles pushed many large companies, such as Google, Toyota, and Ford, to develop their own concept of *robot-car* [114–116]. While self-driving cars are widely considered to be a major development and testing ground for the real-world application of artificial intelligence, major reasons for concern remain in terms of safety, ethics, cost, and reliability [117]. From a safety standpoint, in particular, smart cars need to robustly interpret the behaviour of the humans (drivers, pedestrians or cyclists) they share the environment with, in order to cope with their decisions. *Situation awareness* and the ability to understand the behaviour of other road users are thus crucial for the safe deployment of autonomous vehicles (AVs). Therefore, in this chapter, we only consider the context of *vision-based* autonomous driving [118] from video sequences captured by cameras mounted on the vehicle in a streaming, online fashion.

While detector networks [119] are routinely trained to facilitate object and actor recognition in road scenes, this simply allows the vehicle to ‘see’ what is around it. The philosophy of this work is that robust self-driving capabilities require a deeper,

¹<https://www.darpa.mil/>

more human-like understanding of dynamic road environments (and of the evolving behaviour of other road users over time) in the form of semantically meaningful concepts, as a stepping stone for intention prediction and automated decision making. One advantage of this approach is that it allows the autonomous vehicle to focus on a much smaller amount of relevant information when learning how to make its decisions, in a way arguably closer to how decision-making takes place in humans.

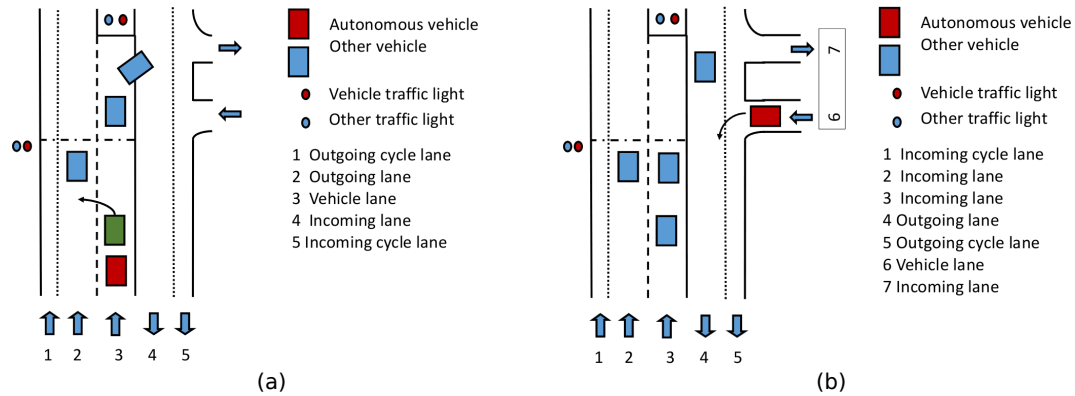


Figure 3.1: Use of labels in ROAD to describe typical road scenarios. (a) A green car is in front of the AV while changing lanes, as depicted by the arrow symbol. The associated event will then carry the following labels: in vehicle lane (location), moving left (action). Once the event is completed, the location label will change to: in outgoing lane. (b) Autonomous vehicle turning left from lane 6 into lane 4: lane 4 will be the outgoing lane as the traffic is moving in the same direction as the AV. However, if the AV turns right from lane 6 into lane 4 (a wrong turn), then lane 4 will become the incoming lane as the vehicle will be moving into the incoming traffic. The overall philosophy of ROAD is to use suitable combinations of multiple label types to fully describe a road situation, and allow a machine learning algorithm to learn from this information.

On the opposite side of the spectrum lies end-to-end reinforcement learning. There, the behaviour of a human driver in response to road situations is used to train, in an imitation learning setting [120], an autonomous car to respond in a more ‘human-like’ manner to road scenarios. This, however, requires an astonishing amount of data from a myriad of road situations. For highway driving only, a relatively simple task when compared to city driving, Fridman et al. in [121] had to use a whole fleet of vehicles to collect 45 million frames. Perhaps more importantly, in this approach, the network learns a mapping from the scene to control inputs, without attempting to model the significant facts taking place in the scene or the reasoning of the agents therein. As discussed in [122], many authors [123, 124] have recently highlighted the insufficiency of models which directly map observations to actions [125], specifically in the scenario of the self-driving car.

3.1.2 ROAD: a multi-label, multi-task dataset

Concept. This work aims to propose a new framework for situation awareness and perception, departing from the disorganised collection of object detection, semantic segmentation or pedestrian intention tasks which is the focus of much current work. We propose to do so in a “holistic”, multi-label approach in which agents, actions and their locations are all ingredients in the fundamental concept of *road event* (RE). Road events are defined as triplets $E = (Ag, Ac, Loc)$ composed by an active road agent Ag , the action(s) Ac it performs (possibly more than one at the same time), and the location(s) Loc ² in which this takes place (which may vary from the start to the end of the event itself), as seen from the point of the view of an autonomous vehicle. The complete list of all three label types is provided in Table 3.1. This takes the problem to a higher conceptual level, in which AVs are tested on their *understanding of what is going on* in a dynamic scene rather than their ability to describe what the scene *looks like*, putting them in a position to use that information to make decisions and a plot course of action. Modelling dynamic road scenes in terms of road events can also allow us to model the causal relationships between what happens; these causality links can then be exploited to predict further future consequences.

To transfer this conceptual paradigm into practice, this chapter introduces ROAD, the first *ROad event Awareness in Autonomous Driving Dataset*, as an entirely new type of dataset designed to allow researchers in autonomous vehicles to test the situation awareness capabilities of their stacks in a manner impossible until now. Unlike all existing benchmarks, ROAD provides ground truth for the action performed by all road agents, not just humans. In this sense ROAD is unique in the richness and sophistication of its annotation, designed to support the proposed conceptual shift. We are confident this contribution will be very useful moving forward for both the autonomous driving and the computer vision community. However, at the same time, it is also worth mentioning that the quantisation of the action and activities in the context of autonomous driving with several limitation and challenges including computational complexity, robustness to uncertainty, and control and decision-making.

Features. ROAD is built upon (a fraction of) the Oxford RobotCar Dataset [126], by carefully annotating 22 carefully selected, relatively long-duration videos. Road events are represented as ‘tubes’, i.e., time series of frame-wise bounding box detections. ROAD is a dataset of significant size, most notably in terms of the richness and complexity of its annotation rather than the raw number of video frames. A total of 122K video frames are labelled for a total of 560K detection bounding boxes in turn associated with 1.7M unique individual labels, broken down into 560K agent

²Note: The label location Loc in our case does not refer to the spatiotemporal coordinates (x, y) of the bounding box.

labels, 640K action labels and 499K location labels.

The dataset was designed according to the following principles.

- A *multi-label* benchmark: each road event is composed by the label of the (moving) agent responsible, the label(s) of the type of action(s) being performed, and labels describing where the action is located.
- Each event can be assigned *multiple instances* of the same label type whenever relevant (e.g., an RE can be an instance of both *moving away* and *turning left*).
- The labelling is done *from the point of view of the AV*: the final goal is for the autonomous vehicle to use this information to make the appropriate decisions.
- The meta-data is intended to contain all the information required to fully describe a road scenario: an illustration of this concept is given in Figure 3.1. After closing one's eyes, the set of labels associated with the current video frame should be sufficient to recreate the road situation in one's head (or, equivalently, sufficient for the AV to be able to make a decision).

In an effort to take action detection into the real world, ROAD moves away from human body actions almost entirely, to consider (besides pedestrian behaviour) actions performed by humans as drivers of various types of vehicles, shifting the paradigm from *actions performed by human bodies* to *events caused by agents*. As shown in our experiments, ROAD is more challenging than current action detection benchmarks due to the complexity of road events happening in real, non-choreographed driving conditions, the number of active agents present and the variety of weather conditions encompassed.

Tasks. ROAD allows one to validate manifold tasks associated with situation awareness for self-driving, each associated with a label type (agent, action, location) or a combination thereof:

1. *Agent detection*, the output is in the form of 'agent tubes' collecting the BBs associated with an active road agent in consecutive frames.
2. *Action detection*, the output is in the form of 'action tubes' (series of bounding box (BB) detections with an attached action label).
3. *Location detection*, the output is in the form of 'agent-location' tubes collecting the BBs associated with an agent and its location in consecutive frames.
4. *Agent-action detection*, the detector is trained on pairs of agent-action labels. The output is again in the form of a tube.

5. *Road event detection*, as well as the AV’s decisions make use of all the three types of information provided by ROAD, detecting RE triplets is crucial for autonomous driving applications.
6. *Temporal AV actions detection*, the output of this task is the class of action the AV performs in each video frame, from a list suitably compiled.

For each task, one can assess both *frame-level* detection, which outputs independently for each video frame the bounding box(es) of the instances there present and the relevant class labels, and *video-level* detection, which consists in regressing the whole series of temporally-linked bounding boxes (i.e., in current terminology, a ‘tube’) associated with an instance, together with the relevant class label. In this chapter, we conduct tests on both. All tasks come with both the necessary annotation and a shared baseline, which is described in Section 3.3.

Table 3.1: ROAD labels.

Agents	Actions	Locations
Pedestrian	Move away	AV lane
Car	Move towards	Outgoing lane
Cyclist	Move	Outgoing cycle lane
Motorbike	Brake	Incoming lane
Medium vehicle	Stop	Incoming cycle lane
Large vehicle	Indicating left	Pavement
Bus	Indicating right	Left pavement
Emergency vehicle	Hazards lights on	Right pavement
AV traffic light	Turn left	Junction
Other traffic light	Turn right	Crossing location
	Overtake	Bus stop
	Wait to cross	Parking
	Cross from left	
	Cross from right	
	Crossing	
	Push object	
	Red traffic light	
	Amber traffic light	
	Green traffic light	

3.1.3 ROAD-R

In this chapter, we also generalize hierarchical multi-label classification (HMC) problems by introducing *multi-label classification problems with (full) propositional logic requirements*. Thus, given a multi-label classification problem with labels A , B , and

C , we can, for example, write the requirement:

$$(\neg A \wedge B) \vee C,$$

stating that for each data point in the dataset either the label C is predicted, or B but not A are predicted.

Then, we present the ROad event Awareness Dataset with logical Requirements (ROAD-R), the first publicly available dataset for autonomous driving with requirements expressed as logical constraints. ROAD-R extends the ROAD dataset by manually annotated ROAD-R with 243 constraints, each verified to hold for each bounding box. A typical constraint is thus “a traffic light cannot be red and green at the same time”, while there are no constraints like “pedestrians should cross at crossings”, which should always be satisfied in theory, but which might not be in real-world scenarios.

Given ROAD-R, we considered 6 current state-of-the-art (SOTA) models, and we showed that they are not able to learn the requirements just from the data points, as more than 90% of the times, they produce predictions that violate the constraints.

3.1.4 Main Contributions

The major contributions of this work are thus the following.

- A conceptual shift in situation awareness centred on a formal definition of the notion of road event, as a triplet composed by a road agent, the action(s) it performs and the location(s) of the event, seen from the point of view of the AV.
- A new ROad event Awareness Dataset for Autonomous Driving (ROAD), the first of its kind, designed to support this paradigm shift and allow the testing of a range of tasks related to situation awareness for autonomous driving: agent and/or action detection, event detection, ego-action classification.
- A multi-label classification problem with propositional logic requirements.
- Introduction of ROAD-R, which is the first publicly available dataset whose requirements are expressed in full propositional logic.
- Considered 6 SOTA models and show that on ROAD-R, they produce predictions violating the requirements more than $\sim 90\%$ of the times,

Instrumental to the introduction of ROAD as the benchmark of choice for semantic situation awareness, we propose a robust baseline for online action/agent/event detection (termed *3D-RetinaNet*) which combines state-of-the-art single-stage object

detector technology with an online tube construction method [127], with the aim of linking detections over time to create *event tubes* [128, 129]. Results for two additional baselines based on a Slowfast detector architecture [130] and YOLOv5³ (for agent detection only) are also reported and critically assessed.

Related publications: The work of this chapter appeared in IEEE transactions on pattern analysis and machine intelligence (IEEE TPAMI 2022) [45] and The ROAD Workshop and Challenge: Event Detection for Situation Awareness in Autonomous Driving at ICCV 2021⁴. The dissertation’s author was the co-author of the TPAMI paper and co-organiser of the workshop and challenge. The main contributions of the dissertation’s author to this project are: adding a new backbone (SlowFast) to the 3D RetinaNet architecture (baseline for ROAD), implementing a YOLOv5 for agents detection, and setting up the evaluation site for the challenge. All the resources of this project are publically available and are accessed from *ROAD dataset*⁵, *3D RetinaNet baseline*⁶, *Inference demo*⁷, and *Challenge evalAI*⁸.

The work of ROAD-R appeared in IJCAI 2022 Workshop on Artificial Intelligence for Autonomous Driving (AI4AD 2022)⁹ and IJCLR 2022 the International Joint Conference on Learning and Reasoning [46]. This work received **the best paper award** at IJCAI 2022 AI4AD and won **the best student paper award** at IJCLR 2022. The dissertation’s author was the co-author in both papers. The main contribution of the dissertation’s author to this project is providing the vision baselines for on ROAD dataset for logical constraint. All the resources of this project are publically available and are accessed from *code*¹⁰, *video*¹¹.

Outline: The remainder of the chapter is organised as follows. Section 3.2 presents our ROAD dataset in full detail, including: its multi-label nature (Sec. 3.2.1), data collection (Sec. 3.2.2), annotation (Sec. 3.2.3), the tasks it is designed to validate (Sec. 3.2.4), and a quantitative summary (Sec. 3.2.5). Section 3.3 presents an overview of the proposed 3D-RetinaNet baseline, and recalls the ROAD challenge organised by some of us at ICCV 2021 to disseminate this new approach to situation awareness within the autonomous driving and computer vision communities, using ROAD as the benchmark. Section 3.4 describes ROAD-R. Experiments are described

³<https://github.com/ultralytics/yolov5>.

⁴<https://sites.google.com/view/roadchallengeiccv2021/>

⁵<https://github.com/gurkirt/road-dataset>

⁶<https://github.com/gurkirt/3D-RetinaNet>

⁷<https://www.youtube.com/watch?v=CmxPjHhiarA>

⁸<https://eval.ai/web/challenges/challenge-page/1059/overview>

⁹<https://learn-to-race.org/workshop-ai4ad-ijcai2022/>

¹⁰<https://github.com/EGiunchiglia/ROAD-R>

¹¹https://www.youtube.com/watch?v=__-Ll5d1VQXY

in Section 3.5, where a number of ablation studies are reported and critically analysed in detail, together with the results of the ROAD challenge’s top participants. Section 3.6 outlines additional exciting tasks the dataset can be used as a benchmark for in the near future, such as future event anticipation, decision making and machine theory of mind [122]. Conclusions and future work are outlined in Section 3.7.

3.2 The Dataset

3.2.1 A Multi-Label Benchmark

The ROAD dataset is specially designed from the perspective of self-driving cars, and thus includes actions performed not just by humans but by all road agents in specific locations, to form *road events* (REs). REs are annotated by drawing a bounding box around each active road agent present in the scene, and linking these bounding boxes over time to form ‘tubes’. As explained, to this purpose three different types of labels are introduced, namely: (i) the category of *road agent* involved (e.g. *Pedestrian*, *Car*, *Bus*, *Cyclist*); (ii) the *type of action* being performed by the agent (e.g. *Moving away*, *Moving towards*, *Crossing* and so on), and (iii) the *location* of the road user relative the autonomous vehicle perceiving the scene (e.g. *In vehicle lane*, *On right pavement*, *In incoming lane*). In addition, ROAD labels the actions performed by the vehicle itself. Multiple agents might be present at any given time, and each of them may perform multiple actions simultaneously (e.g. a *Car* may be *Indicating right* while *Turning right*). Each agent is always associated with at least one action label. The complete list of all three labels types are given in Table 3.1.

Agent labels. Within a road scene, the objects or people able to perform actions which can influence the decision made by the autonomous vehicle are termed *agents*. We only annotate *active agents* (i.e., a parked vehicle or a bike or a person visible to the AV but located away from the road are not considered to be ‘active’ agents). Three types of agent are considered to be of interest, in the sense defined above, to the autonomous vehicle: people, vehicles and traffic lights. For simplicity, the AV itself is considered just like another agent: this is done by labelling the vehicle’s bonnet. People are further subdivided into two sub-classes: pedestrians and cyclists. The vehicle category is subdivided into six sub-classes: car, small-size motorised vehicle, medium-size motorised vehicle, large-size motorised vehicle, bus, motorbike, emergency vehicle. Finally, the ‘traffic lights’ category is divided into two sub-classes: *Vehicle traffic light* (if they apply to the AV) and *Other traffic light* (if they apply to other road users). Only one agent label can be assigned to each active agent present in the scene at any given time.

Action labels. Each agent can perform one or more *actions* at any given time

instant. For example, a traffic light can only carry out a single action: it can be either red, amber, green or ‘black’. A car, instead, can be associated with two action labels simultaneously, e.g., *Turning right* and *Indicating right*. Although some road agents are inherently multitasking, some action combinations can be suitably described by a single label: for example, pushing an object (e.g. a pushchair or a trolley-bag) while walking can be simply labelled as *Pushing object*. The latter was our choice.

AV own actions. Each video frame is also labelled with the action label associated with what the AV is doing. To this end, a bounding box is drawn on the bonnet of the AV. The AV can be assigned one of the following seven action labels: *AV-move*, *AV-stop*, *AV-turn-left*, *AV-turn-right*, *AV-overtake*, *AV-move-left* and *AV-move-right*.

Location labels. Agent *location* is crucial for deciding what action the AV should take next. As the final, long-term objective of this project is to assist autonomous decision making, we propose to label the location of each agent from the perspective of the autonomous vehicle. For example, a pedestrian can be found on the right or the left pavement, in the vehicle’s own lane, while crossing or at a bus stop. The same applies to other agents and vehicles as well. There is no location label for the traffic lights as they are not movable objects, but agents of a static nature and well-defined location. To understand this concept, Fig. 3.1 illustrates two scenarios in which the location of the other vehicles sharing the road is depicted from the point of view of the AV. *Traffic light* is the only agent type missing location labels, all the other agent classes are associated with at least one location label.

3.2.2 Data Collection

ROAD is composed of 22 videos from the publicly available Oxford Robot Car Dataset [126] (OxRD) released in 2017 by the Oxford Robotics Institute¹², covering diverse road scenes under various weather conditions. The OxRD dataset, collected from the narrow streets of the historic city of Oxford, was selected because it presents challenging scenarios for an autonomous vehicle due to the diversity and density of various road users and road events.

Note, however, that our labelling process (described below) is not limited to OxRD. In principle, other autonomous vehicle datasets (e.g. [131, 132]) may be labelled in the same manner to further enrich the ROAD benchmark, we plan to do exactly so in the near future.

Video selection. Within OxRD, videos were selected with the objective of ensuring diversity in terms of weather conditions, times of the day and types of scenes recorded. Specifically, the 22 videos have been recorded both during the day (in strong sunshine, rain or overcast conditions, sometimes with snow present on the

¹²<http://robotcar-dataset.robots.ox.ac.uk/>

surface) and at night.

Preprocessing. First, the original sets of video frames were downloaded and demosaiced, in order to convert them to red, green, and blue (RGB) image sequences. Then, they were encoded into proper video sequences using `ffmpeg`¹³ at the rate of 12 frames per second (fps). Although the original frame rate in the considered frame sequences varies from 11 fps to 16 fps, we uniformised it to keep the annotation process consistent.

3.2.3 Annotation Process

Annotation tool. Annotating tens of thousands of frames rich in content is a very intensive process; therefore, a tool is required which can make this process both fast and intuitive. For this work, we adopted Microsoft’s VoTT¹⁴. The most useful feature of this annotation tool is that it can copy annotations (bounding boxes and their labels) from one frame to the next, while maintaining a unique identification for each box, so that boxes across frames are automatically linked together. Moreover, VoTT also allows for multiple labels, thus lending itself well to ROAD’s multi-label annotation concept.

Annotation protocol. All salient objects and actors within the frame were labelled, with the exception of inactive participants (mostly parked cars) and objects/actors at large distances from the ego vehicle, as the latter were judged to be irrelevant to the AV’s decision-making. This can be seen in the attached 30-minute video¹⁵ portraying ground truth and predictions. As a result, pedestrians, cyclists, and traffic lights were always labelled. Vehicles, on the other hand, were only labelled when active (i.e., moving, indicating, being stopped at lights or stopping with hazard lights on the side of the road). As mentioned, only parked vehicles were not considered active (as they do not arguably influence the AV’s decision-making) and were thus not labelled.

Event label generation. Using the annotations manually generated for actions and agents in the multi-label scenario as discussed above it is possible to generate *event-level* labels about agents, e.g. *Pedestrian / Moving towards the AV On right pavement* or *Cyclist / Overtaking / In vehicle lane*. Any combinations of location, action and agent labels are admissible. If location labels are ignored, the resulting event labels become location-invariant.

Namely, given an agent tube and the continuous temporal sequence of action labels attached to its constituent bounding box detections, we can generate action tubes by looking for changes in the action label series associated with each agent tube. For instance, a *Car* appearing in a video might be first *Moving away* before *Turning left*.

¹³<https://www.ffmpeg.org/>

¹⁴<https://github.com/Microsoft/VoTT/>

¹⁵<https://www.youtube.com/watch?v=CmXPjHhiarA>.

The agent tube for the car will then be formed by two contiguous agent-action tubes: a first tube with label pair *Car / Moving away* and a second one with pair *Car / Turning left*.

Table 3.2: ROAD tasks and attributes.

Task type	Problem type	Output	Multiple labels
Active agent	Detection	Box&Tube	No
Action	Detection	Box&Tube	Yes
Location	Detection	Box&Tube	Yes
Event	Detection	Box&Tube	Yes
AV-action	Temporal detection	Start/End	No

3.2.4 Tasks

ROAD is designed as a sandbox for validating the six tasks relevant to situation awareness in autonomous driving outlined in Sec. 3.1.2. Five of these tasks are detection tasks, while the last one is a frame-level action recognition task sometimes referred to as 'temporal action segmentation' [133], Table 3.2 shows the main attributes of these tasks.

All detection tasks are evaluated both at frame-level and at video- (tube-)level. *Frame-level detection* refers to the problem of identifying in each video frame the bounding box(es) of the instances there present, together with the relevant class labels. *Video-level detection* consists in regressing a whole series of temporally-linked bounding boxes (i.e., in current terminology, a 'tube') together with the relevant class label. In our case, the bounding boxes will mark a specific active agent in the road scene. The labels may issue (depending on the specific task) either from one of the individual label types described above (i.e., agent, action or location) or from one of the meaningful combinations described in 3.2.3 (i.e., event).

Below we list all the tasks for which we currently provide a baseline, with a short description.

1. *Active agent detection* (or *agent detection*) aims at localising an active agent using a bounding box (frame-level) or a tube (video-level) and assigning a class label to it.
2. *Action detection* seeks to localise an active agent occupied in performing a specific action from the list of action classes.
3. In *agent location detection* (or *location detection*) a label from the relevant list of locations (as seen from the AV) is sought and attached to the relevant bounding box or tube.

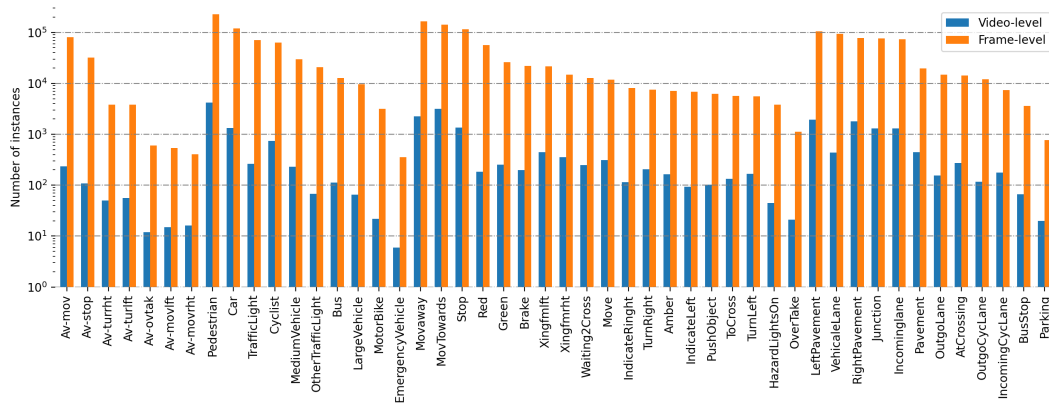


Figure 3.2: Number of instances of each class of individual label-types, in logarithmic scale.

4. *Road event detection* (or *event detection*) consists in assigning to each box or tube a triplet of class labels.
5. *Autonomous vehicle temporal action segmentation* is a frame-level action classification task in which each video frame is assigned a label from the list of possible AV own actions. We refer to this task as 'AV-action segmentation', similarly to [133].

3.2.5 Quantitative Summary

Overall, 122K frames extracted from 22 videos were labelled, in terms of both AV own actions (attached to the entire frame) and bounding boxes with attached one or more labels of each of the three types: agent, action, location. In total, ROAD includes 560K bounding boxes with 1.7M instances of individual labels. The latter figure can be broken down into 560K instances of agent labels, 640K instances of action labels, and 499K instances of location labels. Based on the manually assigned individual labels, we could identify 454K instances of triplets (event labels).

The number of instances for each individual class from the three lists is shown in Fig. 3.2 (frame-level, in orange). The 560K bounding boxes make up 7,029,981,58,040, and 8,394 tubes for the label types agent, action, location, and event, respectively. Figure 3.2 also shows the number of tube instances for each class of individual label types as number of video-level instances (in blue). The figure also highlights the presence of an imbalance in the dataset distribution among the classes. Specifically, certain classes exhibit a significant disparity in sample sizes, for example, "car" has a considerably larger number of samples compared to "emergency vehicle."

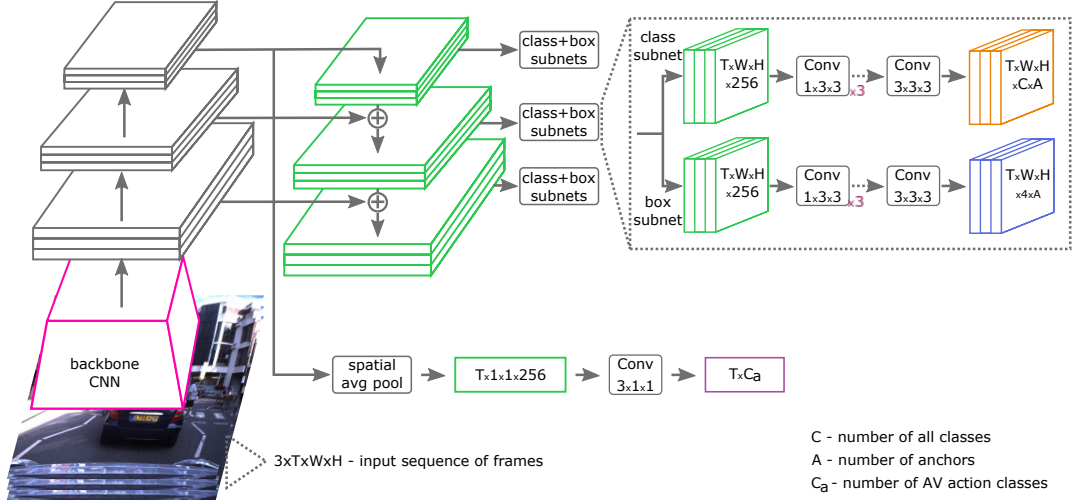


Figure 3.3: Proposed 3D-RetinaNet architecture for online video processing.

3.3 Baseline and Challenge

Inspired by the success of recent 3D CNN architectures [33] for video recognition and of feature-pyramid networks (FPN) [134] with focal loss [135], we propose a simple yet effective 3D feature pyramid network (3D-FPN) with focal loss as a baseline method for ROAD’s detection tasks. We call this architecture *3D-RetinaNet*.

3.3.1 3D-RetinaNet Architecture

The data flow of 3D-RetinaNet is shown in Figure 3.3. The input is a sequence of T video frames. As in classical FPNs [134], the initial block of 3D-RetinaNet consists of a backbone network outputting a series of forward feature pyramid maps, and of lateral layers producing the final feature pyramid composed by T feature maps. The second block is composed by two sub-networks which process these features maps to produce both bounding boxes (4 coordinates) and C classification scores for each anchor location (over A possible locations). In the case of ROAD, the integer C is the sum of the numbers of agent, action, location, and agent-action-location (event) classes, plus one reserved for an *agentness* score. The extra class agentness is used to describe the presence or absence of an active agent. As in FPN [134], we adopt ResNet50 [136] as the backbone network.

2D versus 3D backbones. In our experiments we show results obtained using three different backbones: frame-based ResNet50 (2D), inflated 3D (I3D) [33] and Slowfast [130], in the manner also explained in [130, 137]. Choosing a 2D backbone makes the detector completely online [127], with a delay of a single frame. Choosing an I3D or a Slowfast backbone, instead, causes a 4-frame delay at detection time. Note that, as Slowfast and I3D networks makes use of a max-pool layer with stride

2, the initial feature pyramid in the second case contains $T/2$ feature maps. Nevertheless, in this case we can simply linearly upscale the output to T feature maps.

AV action prediction heads. In order for the method to also address the prediction of the AV's own actions (e.g. whether the AV is stopping, moving, turning left etc.), we branch out the last feature map of the pyramid (see Fig. 3.3, bottom) and apply spatial average pooling, followed by a temporal convolution layer. The output is a score for each of the C_a classes of AV actions, for each of the T input frames.

Loss function. As for the choice of the loss function, we adopt a binary cross-entropy-based focal loss [135]. We chose a binary cross entropy because our dataset is multi-label in nature for which we only train a single binary cross entropy function. The rationale behind selecting a focal-type loss function stems from the anticipation that it could potentially help the neural network in addressing the challenge of long tail distribution of data (see Figure 3.2), where some classes have significantly more samples than others. The focal-type loss accomplishes this by giving more weight to samples from minority classes during training, enabling the model to better learn their distinctive characteristics.

3.3.2 Online Tube Generation via Agentness Score

The autonomous driving scenario requires any suitable method for agent, action or event tube generation to work in an *online* fashion, by incrementally updating the existing tubes as soon as a new video frame is captured. For this reason, this work adopts a recent algorithm proposed by Singh *et al.* [127], which incrementally builds action tubes in an online fashion and at real-time speed. To the best of our knowledge, [127] was the first online multiple action detection approach to appear in the literature, and was later adopted by almost all subsequent works [138–140] on action tube detection.

Linking of detections. We now briefly review the tube-linking method of Singh *et al.* [127], and show how it can be adapted to build agent tubes based on an 'agentness' score, rather than build a tube separately for each class as proposed in the original paper. This makes the whole detection process faster, since the total number of classes is much larger than in the original work [127]. The proposed 3D-RetinaNet is used to regress and classify detection boxes in each video frame potentially containing an active agent of interest. Subsequently, detections whose score is lower than 0.025 are removed and non-maximal suppression is applied based on the agentness score.

At video start, each detection initialises an agentness tube. From that moment on, at any time instance t the highest scoring tubes in terms of mean agentness score up to $t - 1$ are linked to the detections with the highest agentness score in frame t which display an Intersection-over-Union (IoU) overlap with the latest detection in

the tube above a minimum threshold λ . The chosen detection is then removed from the pool of frame- t detections. This continues until the tubes are either assigned or not assigned a detection from current frame. Remaining detections at time t are used to initiate new tubes. A tube is terminated after no suitable detection is found for n consecutive frames. As the linking process takes place, each tube carries scores for all the classes of interest for the task at hand (e.g., action detection rather than event detection), as produced by the classification subnet of 3D-RetinaNet. We can then label each agentness tube using the k classes that show the highest mean score over the duration of the tube.

3.3.3 The ROAD Challenge

To introduce the concept of road event, our new approach to situation awareness and the ROAD dataset to the computer vision and AV communities, some of us have organised in October 2021 the workshop "The ROAD challenge: Event Detection for Situation Awareness in Autonomous Driving"¹⁶. For the challenge, we selected (among the tasks described in Sec. 3.2.4) only three tasks: agent detection, action detection and event detection, which we identified as the most relevant to autonomous driving.

As standard in action detection, evaluation was done in terms of video mean average precision (video-mAP). 3D-RetinaNet was proposed as the baseline for all three tasks. Challenge participants had 18 videos available for training and validation. The remaining 4 videos were to be used to test the final performance of their model. This split was applied to all the three challenges (split 3 of the ROAD evaluation protocol, see Section 3.5.2).

The challenge opened for registration on April 1 2021, with the training and validation folds released on April 30, the test fold released on July 20 and the deadline for submission of results set to September 25. For each stage and each Task the maximum number of submissions was capped at 50, with an additional constraint of 5 submissions per day. The workshop, co-located with ICCV 2021, took place on October 16 2021.

In the validation phase we had between three and five teams submit between 15 and 17 entries to each of three challenges. In the test phase, which took place after the summer, we noticed a much higher participation with 138 submissions from 9 teams to the agent challenge, 98 submissions from 8 teams to the action challenge, and 93 submission from 6 teams to the event detection challenge.

The methods proposed by the winners of each challenge are briefly recalled in Section 3.5.4.

¹⁶<https://sites.google.com/view/roadchallengeiccv2021/>.

Benchmark maintenance. After the conclusion of the ROAD @ ICCV 2021 workshop, the challenge has been re-activated to allow for submissions indefinitely. The ROAD benchmark will be maintained by withholding the test set from the public on the `eval.ai` platform¹⁷, where teams can submit their predictions for evaluation. Training and validation sets can be downloaded from <https://github.com/gurkirt/road-dataset>.

3.4 ROAD-R

ROAD-R extends the ROAD dataset by introducing a set Π of 243 constraints that specify the space of admissible outputs. In order to improve the usability of our dataset, we write the constraints in a way that allows us to easily express Π as a single formula in conjunctive normal form (CNF). The above can be done without any loss in generality, as any propositional formula can be expressed in CNF, and is important because many solvers expects formulas in CNF as input. Thus, each requirement in Π has form:

$$l_1 \vee l_2 \vee \dots \vee l_n, \quad (3.1)$$

where $n \geq 1$, and each l_i is either a negative label $\neg A$ or a positive label A . The requirements have been manually specified following three steps:

1. an initial set of constraints Π_1 was manually created,
2. a subset $\Pi_2 \subset \Pi_1$ was retained by eliminating all those constraints that were entailed by the others,
3. the final subset $\Pi \subset \Pi_2$ was retained by keeping only those requirements that were always satisfied by the ground-truth labels of the entire ROAD-R dataset.

Finally, redundancy in the constraints has been automatically checked with REL-SAT¹⁸. Note that our process of gathering and further selecting the logical requirements follows more closely the software engineering paradigm rather than the machine learning view. To this end, we ensured that the constraints were consistent with the provided labels from the ROAD dataset in the sense that they were acting as strict conditions to be absolutely satisfied by the ground-truth labels, as emphasized in the third step of the annotation pipeline above. Table 3.3 gives a high-level description of the properties of the set Π of constraints. Notice that, with a slight abuse of notation, in the tables we use a set based notation for the requirements. Each requirement of form (3.1) thus becomes

$$\{l_1, l_2, \dots, l_n\}.$$

¹⁷<https://eval.ai/web/challenges/challenge-page/1059/overview>

¹⁸<https://github.com/roberto-bayardo/relsat/>

Table 3.3: Constraint statistics. Π_n is the set of constraints r in Π with $|r| = n$, i.e., with n positive and negative labels. $\mathcal{C} = \{\neg A : A \in \mathcal{C}\}$. Each row shows the number of rules r with $|r| = n$, and the average number of negative and positive labels in such rules.

n	$ \Pi_n $	$\text{avg}_{r \in \Pi_n}(r \cap \mathcal{C})$	$\text{avg}_{r \in \Pi_n}(r \cap \mathcal{C})$
2	215	1.995	0.005
3	5	1	2
7	1	1	6
8	6	1	7
9	6	1	8
10	1	0	10
12	1	1	11
14	1	0	14
15	7	1	14
Total	243	1.87	0.96

Such notation allows us to express the properties of the requirements in a more succinct way. In addition to the information in the tables, we report that of the 243 constraints, there are two in which all the labels are positive (expressing that there must be at least one agent and that every agent but traffic lights has at least one location), and 214 in which all the labels are negative (expressing mutual exclusion between two labels). All the constraints with more than two labels have at most one negative label, as they express a one-to-many relation between actions and agents (like in our dataset “if something is crossing, then it is a pedestrian or a cyclist”). Constraints like “pedestrians should cross at crossings”, which might not be satisfied in practice, are not included. Additionally embedding such logical constraints would require, e.g., using modal operators and, while it would be an interesting study to see the impact on the model’s predictions when adding more complex layers to the expressivity of our logic, we opted for using a simpler logic in this first instance. This also provides more transparency to the wider research community, as the full propositional logic covers a vast range of applications that do not require extra logical operators. Notice that the 243 requirements restrict the number of admissible prediction to $4985868 \sim 5 \times 10^6$, thus ruling out $(2^{41} - 4985868) \sim 10^{12}$ non-admissible predictions.¹⁹ In principle, the set of admissible predictions can be further reduced by adding other constraints. Indeed, the 243 requirements are not guaranteed to be complete from every possible point of view: as standard in the software development cycle, the requirement specification process deeply involves the stakeholders of the system (see, e.g., [141]). For example, we decided not to include constraints like “it is not possible to both move towards and move away”, which were not satisfied by all

¹⁹The number of admissible predictions has been computed with relsat: <https://github.com/roberto-bayardo/relsat/>.

Table 3.4: Splits of training, validation and test sets for the ROAD dataset with respect to weather conditions. The table shows the number of videos in each set or split. For splits, the first figure is the number of training videos, the second number that of validation videos.

Condition	sunny	overcast	snow	night
Training and validation	7	7	1	3
<i>Split-1</i>	7/0	4/3	1/0	3/0
<i>Split-2</i>	7/0	7/0	1/0	0/3
<i>Split-3</i>	4/3	7/0	1/0	3/0
Testing	1	1	1	1

the data points because of errors in the ground truth labels. In these cases, we decided to dismiss the constraint in order to maintain (i) consistency between the knowledge provided by the constraints and by the data points, and (ii) backward compatibility.

As an additional point, we underline that, even though the annotation of the requirements introduces some overhead in the annotation process, it is also the case that the effort of manually writing 243 constraints (i) is negligible when compared to the effort of manually annotating the 22 videos, and (ii) can improve such last process, e.g., allowing to prevent errors in the annotation of the data points.

3.5 Experimental Validation

In this section we present results on the various tasks the ROAD dataset is designed to benchmark (see Sec. 3.2.4).

3.5.1 Implementation Details

The results are evaluated in terms of both frame-level bounding box detection and tube detection. In the first case, the evaluation measure of choice is *frame mean average precision* (f-mAP). We set the Intersection over Union (IoU) detection threshold to 0.5 (signifying a 50% overlap between the predicted and true bounding box). For the second set of results, we use *video mean average precision* (video-mAP), as information on how the ground-truth BBs are temporally connected is available. These evaluation metrics are standard in action detection [127, 138, 142–144].

We also evaluate actions performed by AV, as described in 3.2.1. Since this is a temporal segmentation problem, we adopt the mean average precision metric computed at frame-level, as standard on the Charades [133] dataset.

We use sequences of $T = 8$ frames as input to 3D-RetinaNet. Input image size is set to 512×682 . This choice of T is the result of GPU memory constraints; however, at test time, we unroll our convolutional 3D-RetinaNet for sequences of 32 frames, showing that it can be deployed in a streaming fashion. We initialise the

backbone network with weights pretrained on Kinetics [34]. For training we use an SGD optimiser with step learning rate. The initial learning rate is set to 0.01 and drops by a factor of 10 after 18 and 25 epochs, up to an overall 30 epochs.

The parameters of the tube-building algorithm (Sec. 3.3.2) are set by cross validation. For ROAD we obtain $\lambda = 0.5$ and $k = 4$.

3.5.2 Experimental Results on ROAD

Three splits: modelling weather variability

For the benchmarking of the ROAD tasks, we divided the dataset into two sets. The first set contains 18 videos for training and validation purposes, while the second set contains 4 videos for testing, equally representing the four types of weather conditions encountered.

The group of training and validation videos is further subdivided into three different ways ('splits'). In each split, 15 videos are selected for training and 3 for validation. Details on the number of videos for each set and split are shown in Table 3.4. All 3 validation videos for Split-1 are overcast; 4 overcast videos are also present in the training set. As such, Split-1 is designed to assess the effect of different overcast conditions. Split-2 has all 3 night videos in the validation subset, and none in the training set. It is thus designed to test model robustness to day/night variations. Finally, Split-3 contains 4 training and 3 validation videos for sunny weather: it is thus designed to evaluate the effect of different sunny conditions, as camera glare can be an issue when the vehicle is turning or facing the sun directly.

Note that there is no split to simulate a bias towards snowy conditions, as the dataset only contains one video of that kind. The test set (bottom row) is more uniform, as it contains one video from each environmental condition.

Results on the various tasks

Results are reported for the tasks discussed in Section 3.2.4.

Frame-level results across the five detection tasks are summarised in Table 3.5 using the frame-mAP (f-mAp) metric, for a detection threshold of $\delta = 0.5$. The reported figures are averaged across the three splits described above, in order to assess the overall robustness of the detectors to domain variations. Performance within each split is evaluated on both the corresponding validation subset and test set. Each row in the Table shows the result of a particular combination of backbone network (2D, I3D, or Slowfast) and test-time sequence length (in a number of frames, 8 and 32). Frame-level results vary between 16.8% (events) and 65.4% (agentness) for I3D, and between 23.9% and 69.2% for Slowfast.

Table 3.5: Frame-level results (mAP %) averaged across the three splits of ROAD. The considered models differ in terms of backbone network (2D, I3D, and Slowfast) and clip length (08 vs 32). The performance of YOLOv5 on agent detection is also reported. Detection threshold $\delta = 0.5$. Both validation and test performance are reported for each entry.

Model	Agentness	Agents	Actions	Locations	Events
2D-08	51.8/63.4	30.9/39.5	15.9/22.0	23.2/30.8	10.6/12.8
2D-32	52.4/64.2	31.5/39.8	16.3/22.6	23.6/31.4	10.8/13.0
I3D-08	52.3/65.1	32.2/39.5	19.3/25.4	24.5/34.9	12.3/16.5
I3D-32	52.7/65.4	32.3/39.2	19.7/25.9	24.7/35.3	12.6/16.8
Slowfast-08	68.8/ 69.2	41.9/47.5	26.9/31.1	34.6/ 37.3	18.1 /23.7
Slowfast-32	69.3 /68.7	42.6/43.7	27.3 / 31.7	34.8 /36.4	18.0/ 23.9
YOLOv5	-	57.9 / 56.9	-	-	-

Table 3.6: Video-level results (mAP %) averaged across the three ROAD splits. The models differ in terms of backbone network (2D, I3D and Slowfast) and test time clip length (08 vs 32). The performance of YOLOv5 on agent detection is also reported. Both validation and test performance are reported for each entry.

Model	Agents	Actions	Locations	Events
Detection threshold $\delta = 0.2$				
2D-08	22.2/25.1	10.3/13.9	18.2/24.8	12.8/14.7
2D-32	22.6/25.0	11.2/14.5	18.5/25.9	13.0/15.3
I3D-08	23.2/26.5	14.1/15.8	20.8/25.8	14.9/17.4
I3D-32	24.4/26.9	14.3/17.5	21.3/27.1	15.9/17.9
Slowfast-08	24.1/29.0	16.0 / 20.5	28.3/ 33.0	18.9/22.4
Slowfast-32	24.2/28.6	16.0 /19.55	29.0 /29.7	19.1 / 22.5
YOLOv5	38.8 / 43.3	-	-	-
Detection threshold $\delta = 0.5$				
2D-08	8.9/7.5	2.3/3.0	5.2/6.1	5.1/5.3
2D-32	8.3/8.0	2.7/3.3	5.6/7.1	5.0/5.7
I3D-08	9.2/9.6	4.0 /4.3	5.8/6.9	4.6/5.4
I3D-32	9.7/10.2	4.0 /4.6	6.4/7.7	4.8/6.1
Slowfast-08	7.1/8.9	3.9/ 4.7	7.1/ 11.0	6.5 /6.6
Slowfast-32	8.3/9.8	3.7/4.4	8.4 /10.0	5.3/ 7.3
YOLOv5	18.7 / 13.9	-	-	-

Video-level results are reported in terms of video-mAP in Table 3.6. As for the frame-level results, tube detection performance (see Sec. 3.3.2) is averaged across the three splits. One can appreciate the similarities between frame- and video-level results, which follow a similar trend albeit at a much lower absolute level. Again, results are reported for different backbone networks and sequence lengths.

Streaming deployment. Increasing test sequence length from 8 to 32 does not much impact performance. This indicates that, even though the network is trained on 8-frame clips, being fully convolutional (including the heads in the temporal direction), it can be easily unrolled to process longer sequences at test time, making it easy to deploy in a streaming fashion. Being deployable in an incremental fashion is a must for autonomous driving applications; this is a quality that other tubelet-

Table 3.7: Number of video- and frame-level instances for each label (individual or composite), left. Corresponding frame-/video-level results (mAP@%) for each of the three ROAD splits (right). Val- n denotes the validation set for Split n . Results produced by an I3D backbone.

		Number of instance					Frame-mAP@0.5/Video-mAP@0.2					
Train subset		#Boxes/#Tubes					Train-1		Train-2		Train-3	
Eval subset	All	Val-1	Val-2	Val-3	Test		Val-1	Test	Val-2	Test	Val-3	Test
Agent	559142/7029	60103/781	79119/761	83750/809	82465/1138		44.5/30.1	34.0/25.7	17.2/16.0	40.9/27.4	35.3/27.1	42.6/27.5
Action	639740/9815	69523/1054	89142/1065	95760/1111	94669/1548		26.2/17.0	26.6/17.4	11.7/11.4	25.3/17.3	21.2/14.6	25.7/17.9
Location	498566/8040	56594/851	67116/864	77084/914	70473/1295		34.9/28.6	35.2/26.4	13.7/12.1	33.9/26.3	25.4/23.2	36.7/28.6
Event	453626/8394	43569/883	65965/963	72152/967	64545/1301		17.7/18.6	15.9/15.8	6.4/11.8	16.4/18.9	13.7/17.2	18.1/18.9
		Number of instances					Frame-AP					
AV-action	122154/490	17929/67	18001/56	16700/85	20374/82		57.9	45.7	33.5	43.6	43.7	48.2

based online action detection methods [138–140] fail to exhibit, as they can only be deployed in a sliding window fashion. Interestingly, the latest work on streaming object detection [145] proposes an approach that integrates latency and accuracy into a single metric for real-time online perception, termed ‘streaming accuracy’. We will consider adopting this metric in the future evolution of ROAD.

Impact of the backbone. Broadly speaking, the Slowfast [130] and I3D [33] versions of the backbone perform as expected, much better than the 2D version. A Slowfast backbone can particularly help with tasks that require the system to ‘understand’ movement, e.g. when detecting actions and road events, at least at 0.2 IoU. Under more stringent localisation requirements ($\delta = 0.5$), it is interesting to notice how Slowfast’s advantage is quite limited, with the I3D version often outperforming it. This is due to the fact that the 3D-RetinaNet baseline model’s choice of fixed-length input that may not fully leverage the strengths of the Slowfast backbone, which excels in scenarios requiring motion information and temporal context, such as video analysis. Therefore, any critique of the model should consider this limitation and its potential impact on capturing temporal dependencies and motion-related features. Finally, even stronger backbones using transformers [146, 147] can be plugged in our 3D-RetinaNet baseline as backbone.

Level of task challenge. The overall results on event detection (last column in both Table 3.5 and Table 3.6) are encouraging, but they remain in the low 20s at best, showing how challenging situation awareness is in road scenarios.

Comparison across tasks. The headline figures are not really comparable since, as we know, the number of classes per task varies. More importantly, within-class variability is often lower for composite labels. For example, the score for *Indicating right* is really low, whereas *Car / Indicating-right* has much better performance. This is because the within-class variability of the pair *Car / Indicating-right* is much lower than that of *Indicating right*, which puts together instances of differently-looking types of vehicles (e.g. buses, cars and vans) all indicating right. Interestingly, results on agents are comparable among the four baseline models (especially for f-mAP and

v-mAP at 0.2, see Tables 3.5 and 3.6).

YOLOv5 for Agent detection. For completeness, we also trained YOLOv5²⁰ for the detection of active agents. The results are shown in the last row of both Table 3.5 and Table 3.6. Keeping in mind that YOLOv5 is trained only on single input frames, it shows a remarkable improvement over the other baseline methods for active agent detection. We believe that is because YOLOv5 is better at the regression part of the detection problem – namely, Slowfast has a recall of 71% compared to the 94% of YOLOv5, so that Slowfast has a 10% lower mAP for active agent detection. We leave the combination of YOLOv5 for bounding box proposal generation and Slowfast for proposal classification as a promising future extension, which could lead to a general improvement across all tasks.

Validation vs test results. Results on the test set are, on average, superior to those on the validation set. This is because the test set includes data from all weather/visibility conditions (see Table 3.4), whereas for each split the validation set only contains videos from a single weather condition. E.g., in Split 2 all validation videos are nighttime ones.

Results under different weather conditions

Table 3.7 shows, instead, the results obtained under the three different splits we created on the basis of the weather/environmental conditions of the ROAD videos, discussed in Section 3.5.2 and summarised in Table 3.4. Note that the total number of instances (boxes for frame-level results or tubes for video-level ones) of the five detection tasks is comparable for all the three splits.

We can see how Split-2 (for which all three validation videos are taken at night and no nighttime videos are used for training, see Table 3.4) has the lowest validation results, as seen in Table 3.7 (Train-2, Val-2). When the network trained on Split-2's training data is evaluated on the (common) test set, instead, its performance is similar to that of the networks trained on the other splits (see Test columns). Split-1 has three overcast videos in the validation set, but also four overcast videos in the training set. The resulting network has the best performance across the three validation splits. Also, under overcast conditions one does not have the typical problems with nighttime vision, nor glares issues as in sunny days, as it has sunny videos in both train and validation sets.

These results seem to attest a certain robustness of the baseline to weather variations, for no matter the choice of the validation set used to train the network parameters (represented by the three splits), the performance on test data (as long as the latter fairly represents a spectrum of weather conditions) is rather stable.

²⁰<https://github.com/ultralytics/yolov5>

Results of AV temporal action detection

Table 3.8 shows the results of using 3D-RetinaNet to temporally detect AV-action classes, averaged across all three splits on both validation and test set. As we can see, the results for classes *AV-move* and *AV-stop* are very good, we think because these two classes are predominately present in the dataset. The performance of the ‘turning’ classes is reasonable, but the results for the bottom three classes are really disappointing. We believe this is mainly due the fact that the dataset is very heavily biased (in terms of number of instances) towards the other classes. As we do intend to further expand this dataset in the future by including more and more videos, we hope the class imbalance issue can be mitigated over time. A measure of performance weighing mAP using the number of instances per class could be considered, but this is not quite standard in the action detection literature. At the same time, ROAD provides an opportunity for testing methods designed to address class imbalance.

3.5.3 Qualitative Results

Finally, we provide some qualitative results of our baseline model in terms of success and failure modes. Cases in which the baseline work accurately are illustrated in Figure 3.4, where the model is shown to detect only those agents which are active (i.e., are performing some actions) and ignore all the inactive agents (namely, parked vehicles). Agent prediction is very stable across all the examples, whereas action and location prediction show some weakness in some case: for instance, the night-time example in the second row of the second column, where both the cars in front are moving away in the outgoing lane but our method fails to label their location correctly.

In contrast, the failure modes illustrated in Figure 3.5 are cases in which the model fails to assign to agents the correct label, and also detects agents which are not active (e.g. often parked cars, see the white vehicle in the top row, first column or the red vehicle in the third row, first column).

3.5.4 Challenge Results

Table 3.9 compares the results of the top teams participating in our ROAD @ ICCV 2021 challenge with those of the Slowfast and YOLOv5 baselines, at a tube detection threshold of 0.2. The challenge server remains open at <https://eval.ai/web/challenges/challenge-page/1059/overview>, where one can consult the latest entries.

Agent detection. The agent detection challenge was won by a team formed by Chenghui Li, Yi Cheng, Shuhan Wang, Zhongjian Huang, Fang Liu of Xidian Uni-



Figure 3.4: Success cases in which our model detects the actions and locations correctly, and only for those agents which are active. Ground truth bounding boxes and labels are shown in green, while the predictions of our model are shown in red.

versity, with an entry using YOLOv5 with post-processing. In their approach, agents are linked by evaluating their similarity between frames and grouping them into a tube. Discontinuous tubes are completed through frame filling, using motion information. Also, the authors note that YOLOv5 generates some incorrect bounding boxes, scattered in different frames, and take advantage of this by filtering out the shorter tubes. As shown in Table 3.9, the postprocessing applied by the winning entry significantly outperforms our off-the-shelf implementation of YOLOv5 on agent detection.

Action detection. The action detection challenge was won by Lijun Yu, Yijun



Figure 3.5: Some of the failure modes of our model, shown detecting inactive agents and/or assigning incorrect action and location labels. Ground truth bounding boxes and labels are shown in green, while the predictions of our model are shown in red.

Qian, Xiwen Chen, Wenhe Liu and Alexander G. Hauptmann of team CMU-INF, with an entry called “ArgusRoad: Road Activity Detection with Connectionist Spatiotemporal Proposals”, based on their Argus++ framework for real-time activity recognition in extended videos in the NIST ActEV (Activities in Extended Video ActEV) challenge²¹. They had to adapt their system to be run on ROAD, e.g. to construct tube proposals rather than frame-level proposals. The approach is a rather complex cascade of object tracking, proposal generation, activity recognition and temporal localisation stages [148]. Results show a significant (5%) improvement

²¹<https://actev.nist.gov/>.

Table 3.8: AV temporal action detection results (frame mAP%) averaged across all three splits.

Model	No instances	Frame-mAP@0.5			
		I3D		2D	
Eval subset	All	Val	Test	Val	Test
Av-move	81196/233	92.0	96.6	83.0	87.8
Av-stop	31801/108	92.2	98.5	65.3	68.4
Av-turn-right	3826/50	46.1	63.0	35.0	57.7
Av-turn-left	3787/56	69.0	59.8	55.1	42.9
Av-overtake	599/12	4.9	1.1	2.7	2.5
Av-move-left	537/15	0.5	0.8	0.5	0.5
Av-move-right	408/16	10.5	0.6	4.0	2.0
Total/Mean	122154/490	45.0	45.8	35.1	37.4

Table 3.9: Results (in video-mAP) of the winning entries to the ICCV 2021 ROAD challenge compared with the Slowfast and YOLOv5 baselines, at a detection threshold of 0.2.

Task	Top team	Slowfast	YOLOv5	Winners
Agent detection	Xidian	29.0	43.3	52.4
Action detection	CMU-INF	20.5	-	25.6
Event detection	IFLY	22.4	-	24.7

over the Slowfast baseline, which is close to state-of-the-art in action detection, but still at a relatively low level (25.6%)

Event detection. The event detection challenge was won by team IFLY (Yujie Hou and Fengyan Wang, from the University of Science and Technology of China and IFLYTEK). The entry consisted in a number of amendments to the 3D-RetinaNet baseline, namely: bounding box interpolation, tuning of the optimiser, ensemble feature extraction with RCN, GRU and LSTM units, together with some data augmentation. Results show an improvement of above 2% over Slowfast, which suggests event better performance could be achieved by applying the ensemble technique to the latter.

3.5.5 Experimental Results of ROAD-R

As a first step, we ran 6 SOTA temporal feature learning architectures as part of a 3D-RetinaNet model [149] (with a 2D-ConvNet backbone made of Resnet50 [136]) for event detection and evaluated to which extent constraints are violated. We considered:

1. *2D-ConvNet (C2D)* [150]: a Resnet50-based architecture with an additional temporal dimension for learning features from videos. The extension from 2D to 3D is done by adding a pooling layer over time to combine the spatial

features.

2. *Inflated 3D-ConvNet (I3D)* [33]: a sequential learning architecture extendable to any SOTA image classification model (2D-ConvNet based), able to learn continuous spatio-temporal features from the sequence of frames.
3. *Recurrent Convolutional Network (RCN)* [8]: a 3D-ConvNet model that relies on recurrence for learning the spatio-temporal features at each network level. During the feature extraction phase, RCNs exploit both 2D convolutions across the spatial domain and 1D convolutions across the temporal domain.
4. *Random Connectivity Long Short-Term Memory (RCLSTM)* [151]: an updated version of LSTM in which the neurons are connected in a stochastic manner, rather than fully connected. In our case, the LSTM cell is used as a bottleneck in Resnet50 for learning the features sequentially.
5. *Random Connectivity Gated Recurrent Unit (RCGRU)* [151]: an alternative version of RCLSTM where the GRU cell is used instead of the LSTM one. GRU makes the process more efficient with fewer parameters than the LSTM.
6. *SlowFast* [130]: a 3D-CNN architecture that contains both slow and fast pathways for extracting the sequential features. A Slow pathway computes the spatial semantics at low frame rate while a Fast pathway processes high frame rate for capturing the motion features. Both of the pathways are fused in a single architecture by lateral connections.

We trained 3D-RetinaNet²² using the same hyperparameter settings for all the models: (i) batch size equal to 4, (ii) sequence length equal to 8, and (iii) image input size equal to 512×682 . All the models were initialized with the Kinetics pre-trained weights. An SGD optimizer [152] with step learning rate was used. The initial learning rate was set to 0.0041 for all the models except SlowFast, for which it was set to 0.0021 due to the diverse nature of slow and fast pathways. All the models were trained for 30 epochs and the learning rate was made to drop by a factor of 10 after 18 and 25 epochs. The machine used for the experiments has 64 CPUs (2.2 GHz each) and 4 Titan RTX GPUs having 24 GB of RAM each.

To measure the models' performance, we used the *frame mean average precision* (f-mAP), which is the standard metric used for action detection (see, e.g., [153, 154]), with IoU threshold equal to 0.5 and 0.75, indicated as f-mAP@0.5 and f-mAP@0.75, respectively. The results for the SOTA models at IoU threshold 0.5 and 0.75 are reported in Table 3.10, column "SOTA".

²²<https://github.com/gurkirt/3D-RetinaNets>.



Figure 3.6: Examples of violations of $\{\neg\text{RedTL}, \neg\text{GreenTL}\}$.

To measure the extent to which each system violates the constraints, we used the following metrics:

- the percentage of non-admissible predictions,
- the average number of violations committed per prediction, and
- the percentage of constraints violated at least once,

while varying the threshold θ from 0.1 to 0.9 with step 0.1. The results are in Fig. 5.9, where (to improve readability) we do not plot the values corresponding to $\theta = 0.0$ and $\theta = 1.0$. For $\theta = 0.0$ (resp., $\theta = 1.0$), all the predictions are positive (resp., negative), and thus the corresponding values are (in order) 100%, 214, and 214/243 (resp., 100%, 2, and 2/243).

3.6 Further Extensions

By design, ROAD is an open project which we expect to evolve and grow over time.

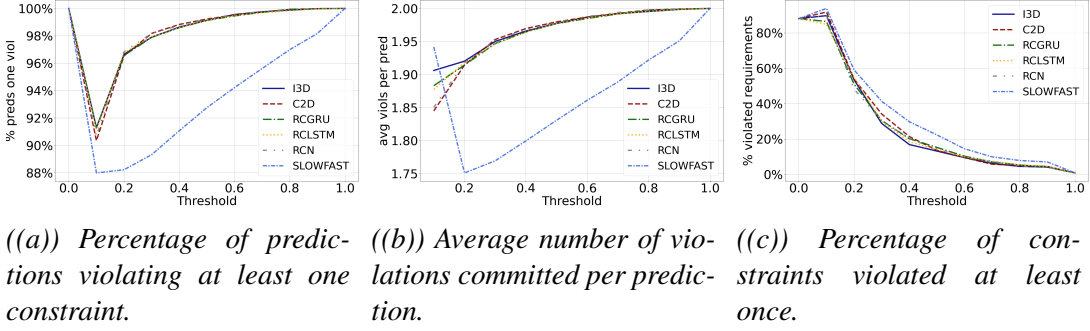


Figure 3.7: ROAD-R and SOTA models. In the x-axis, there is the threshold $\theta \in [0.1, 0.9]$, step 0.1.

Table 3.10: $f\text{-mAP}@0.5$ and $f\text{-mAP}@0.75$ for the current SOTA models.

Model	IoU threshold	
	0.5	0.75
C2D	27.57	12.66
I3D	30.12	16.29
RCGRU	30.78	15.98
RCLSTM	30.49	16.64
RCN	29.64	15.82
SlowFast	28.79	14.40

Extension to other datasets and environments. In the near future, we will work towards completing the multi-label annotation process for a larger number of frames coming from videos spanning an even wider range of road conditions. Further down the line, we plan to extend the benchmark to other cities, countries and sensor configurations, to slowly grow towards an even more robust, 'in the wild' setting. In particular, we will initially target the Pedestrian Intention Dataset (PIE, [52]) and Waymo [155]. The latter one comes with spatiotemporal tube annotation for pedestrian and vehicles, much facilitating the extension of ROAD-like event annotation there.

Event anticipation/intent prediction. ROAD is an oven-ready playground for action and event anticipation algorithms, a topic of growing interest in the vision community [156, 157], as it already provides the kind of annotation that allows researchers to test predictions of both future event labels and future event locations, both spatial and temporal. Anticipating the future behaviour of other road agents is crucial to empower the AV to react timely and appropriately. The output of this Task should be in the form of one or more future tubes, with the scores of the associated class labels and the future bounding box locations in the image plane [158]. We will shortly propose a baseline method for this Task, but we encourage researchers in the area to start engaging with the dataset from now.

Autonomous decision making. In accordance with our overall philosophy, we will design and share a baseline for AV decision making from intermediate semantic

representations. The output of this Task should be the decision made by the AV in response to a road situation [159], represented as a collection of events as defined in this chapter. As the action performed by the AV at any given time is part of the annotation, the necessary meta-data is already there. Although we did provide a simple temporal segmentation baseline for this task seen as a classification problem, we intend in the near future to propose a baseline from a decision making point of view, making use of the intermediate semantic representations produced by the detectors.

Machine theory of mind [160] refers to the attempt to provide machines with (limited) ability to guess the reasoning process of other intelligent agents they share the environment with. Building on our efforts in this area [122], we will work with teams of psychologists and neuroscientists to provide annotations in terms of mental states and reasoning processes for the road agents present in ROAD. Note that theory of mind models can also be validated in terms of how close the predictions of agent behaviour they are capable of generating are to their actual observed behaviour. Assuming that the output of a theory of mind model is intention (which is observable and annotated) the same baseline as for event anticipation can be employed.

Continual event detection. ROAD’s conceptual setting is intrinsically incremental, one in which the autonomous vehicle keeps learning from the data it observes, in particular by updating the models used to estimate the intermediate semantic representations. The videos forming the dataset are particularly suitable, as they last 8 minutes each, providing a long string of events and data to learn from. To this end, we plan to set a protocol for the continual learning of event classifiers and detectors and propose ROAD as the first continual learning benchmark in this area [161].

Alternatives to ROAD-R. Throughout our ROAD-R, we used logical constraints to satisfy the requirements and avoid the prediction of impossible scenarios. The main intuition behind using logical constraints instead of probabilistic priors is that probabilistic priors do not give any guarantee that the constraints will be satisfied. In ROAD-R, the constraints we employed predominantly pertained to impossible scenarios. This means that the majority of our constraints are rooted in expressing physical impossibilities. This distinction between impossible and improbable scenarios is pivotal when considering alternative probabilistic methods. In the future, we are aiming to consider alternative probabilistic methods that can effectively handle both impossible and improbable scenarios. This reflection can enhance the robustness and applicability of our research findings, ensuring that they are not limited solely to situations of physical impossibility but also account for highly improbable yet impactful events. By doing so, we aim to provide a more comprehensive and versatile framework for decision-makers and stakeholders who rely on our research to navigate uncertain real-world situations.

3.7 Summary of the Chapter

This chapter proposed a strategy for situation awareness in autonomous driving based on the notion of road events, and contributed a new ROad event Awareness Dataset for Autonomous Driving (ROAD) as a benchmark for this area of research. The dataset, built on top of videos captured as part of the Oxford RobotCar dataset [126], has unique features in the field. Its rich annotation follows a multi-label philosophy in which road agents (including the AV), their locations and the action(s) they perform are all labelled, and road events can be obtained by simply composing labels of the three types. The dataset contains 22 videos with 122K annotated video frames, for a total of 560K detection bounding boxes associated with 1.7M individual labels.

Baseline tests were conducted on ROAD using a new 3D-RetinaNet architecture, as well as a Slowfast backbone and a YOLOv5 model (for agent detection). Both frame-mAP and video-mAP were evaluated. Our preliminary results highlight the challenging nature of ROAD, with the Slowfast baseline achieving a video-mAP on the three main tasks comprised between 20% and 30%, at low localisation precision (20% overlap). YOLOv5, however, was able to achieve significantly better performance. These findings were reinforced by the results of the ROAD @ ICCV 2021 challenge, and support the need for an even broader analysis, while highlighting the significant challenges specific to situation awareness in road scenarios.

Our dataset is extensible to a number of challenging tasks associated with situation awareness in autonomous driving, such as event prediction, trajectory prediction, continual learning and machine theory of mind, and we pledge to further enrich it in the near future by extending ROAD-like annotation to major datasets such as PIE and Waymo.

We also proposed a new dataset for the task of learning with requirement, called ROAD-R. We showed that SOTA models most of the time violate the requirements, and how it is possible to exploit the requirements to create models that are compliant with (i.e., strictly satisfy) the requirements while improving their performance.

Part II : Complex Activity Detection

Chapter 4

Spatiotemporal Deformable Scene Graphs for Complex Activity Detection

4.1 Introduction

In the second part of the thesis, we are taking one step ahead of ROAD event detection presented in the previous chapter (Chapter 3) and address a novel problem of Complex Activity Detection. In this chapter, long-term complex activity recognition and localisation framework is presented and can be crucial for decision-making in autonomous systems such as smart cars and surgical robots. Here we address the problem via a novel deformable, spatiotemporal scene graph approach, consisting of three main building blocks: (i) action tube detection, (ii) the modelling of the deformable geometry of parts, and (iii) a graph convolutional network. Firstly, action tubes are detected in a series of snippets. Next, a new 3D deformable RoI pooling layer is designed for learning the flexible, deformable geometry of the constituent action tubes. Finally, a scene graph is constructed by considering all parts as nodes and connecting them based on different semantics such as order of appearance, sharing the same action label and feature similarity. We also contribute fresh temporal complex activity annotation for the recently released ROAD autonomous driving and SARAS-ESAD surgical action datasets and show the adaptability of our framework to different domains. Our method is shown to significantly outperform graph-based competitors on both augmented datasets.

4.1.1 Motivation

Complex activity recognition is attracting much attention in the computer vision research community due to its significance for a variety of real-world applications, such as autonomous driving [2, 3], surveillance [4], medical robotics [5] or team sports analysis [6]. In autonomous driving, for instance, it is extremely important that the vehicle understands dynamic road scenes, in order, e.g., to accurately predict the intention of pedestrians and forecast their trajectories to inform appropriate decisions. In surveillance, group activities rather than actions performed by individuals need to be monitored. Robotic assistant surgeons need to understand what the main surgeon is doing throughout a complex surgical procedure composed of many short-term actions and events [162], in order to suitably assist them.

Prior to the advent of deep learning techniques, the study of *atomic* and *complex* activities was a vibrant research area that encompassed a range of methodologies [163, 164]. The fundamental works from the 1990s and the early 2000s focused on understanding and modeling these activities through various approaches such as grammars [165], Bayesian networks [166], and graph-based recognition [167]. Researchers explored the decomposition of complex actions into fundamental *atomic actions*, which were basic and indivisible components of more intricate behaviors. These foundational works laid the groundwork for understanding how activities are composed and structured, providing insights into the underlying patterns and dependencies. By analyzing the relationships between atomic actions and their compositions, researchers were able to formulate models that captured the essence of human actions and interactions. This chapter revisits the concepts of *atomic* and *complex* activities within the framework of the present era, building upon the legacy of past research while also addressing the novel challenges and opportunities introduced by deep learning techniques.

Recent methods for action or activity recognition and localisation can be broadly divided into two categories; single atomic action [25, 168–170] and multiple atomic action recognition/localisation [40, 63, 72–75]. The former methods only focus on identifying the start and end of an action performed in a short video portraying a single instance, leveraging datasets such as UCF-101 [171] or Charades [172]. The latter set of approaches consider videos which contain a number of atomic actions or multiple repetitions of the same action. Methods in this category do address complex activity recognition, as their aim is to understand an overall, dynamic scene by detecting and identifying its constituent components. Datasets used for complex activity detection are Epic-kitchens [173], THUMOS14 [31] or ActivityNet v1.3 [30]. Both classes of methods are geared towards merely recognising and localising short term action or activities that lasts for only a few frames or seconds.

4.1.2 Objective

Unlike all existing methods, in this chapter, we present a framework capable of recognising *complex, long-term activities*, validated in the fields of autonomous driving and surgical robotics but of general applicability and extendable to other domains. More precisely, by ‘complex activity’ we mean *an ensemble of ‘atomic’ actions, each performed by an individual agent present in the environment, which extends over a period of time and which collectively has a meaning*. These activities can be contiguous, occurring sequentially without any gap or background (in the case of surgical robotics). Alternatively, these activities may occur at different positions within the video, with background segments interspersed between them (in the case of autonomous driving). For instance, in the autonomous driving context, one can define the complex activity *Negotiating intersection* as composed of the following atomic actions: an autonomous vehicle (AV) moves along its lane; the traffic light regulating the vehicle’s lane turns red, while the light turns green for a traversing road; a number of other vehicles pass through the intersection from both sides; the AV’s light turns green again; the AV resumes moving and crosses the intersection. The proposed pipeline (Fig. 4.1) is divided into three parts: (i) action tube detection, (ii) part-based feature extraction and learning via 3D deformable RoI pooling, and (iii) a graph generation strategy to process a variable number of parts and their connections, aimed at learning the overall semantics of a dynamic scene representing a complex activity. Action tube detection is a necessary pre-processing step, aimed at spatially and temporally locating the atomic actions present [127, 128, 158, 174–177]. Note that the tube detector needs to ensure a fixed-size representation for each activity part (atomic action). Here, in particular, we adopt AMTNet [1], as the latter describes action tubes of any duration using a fixed number of bounding box detections.

4.1.3 Main Contributions

Our contribution is twofold. Firstly, our novel 3D deformable RoI pooling layer, inspired by standard deformable and modulated RoI pooling [178, 179], is not only designed to work with 3D data but is also capable of learning feature representations for tubes of variable spatiotemporal shape. Secondly, an original Graph Convolutional Network (GCN) module constructs a graph by considering individual tubes as nodes and connecting them via edges encoding diverse semantics, namely: appearance order, sharing of the same action label, and spatial feature similarity. The spatiotemporal scene graph so constructed is then processed by a stack of graph convolutional layers resulting in graph representation features, which are used to train a classifier for recognising complex activities, followed by a localisation stage which

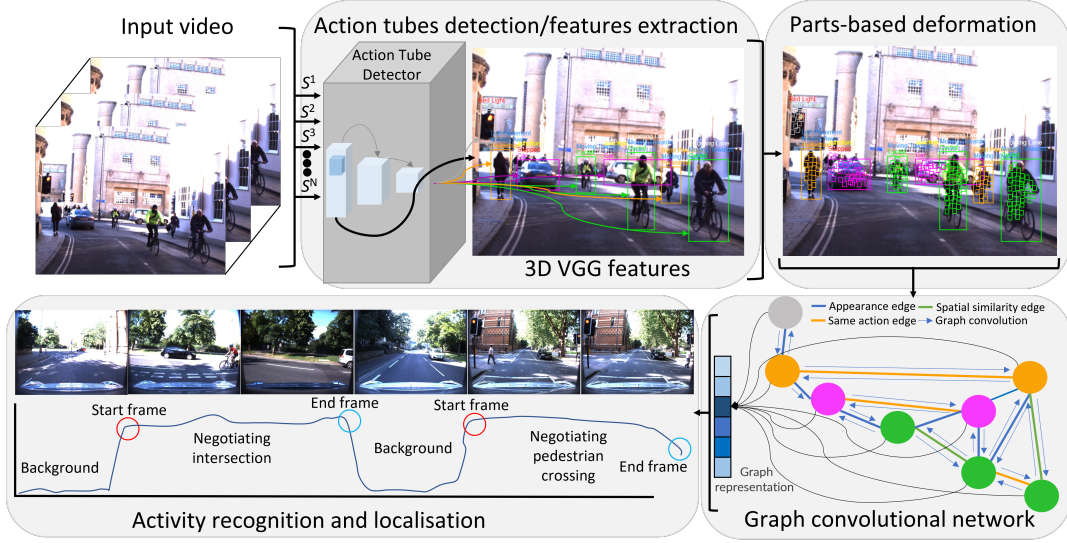


Figure 4.1: Overall pipeline of our long-term complex activity detection framework. (i) The input video is first divided into snippets. (ii) Snippets are passed to an action tube detector module one by one. (iii) A part-based deformation module receives 3D VGG features and action tube locations and returns features for the salient (non-background) parts of each tube instance. (iv) A GCN module represents activity parts (action tubes) as nodes with the features generated by ROI pooling and builds edges with different semantics to construct a spatiotemporal graph representation. Finally (v), the graph representation features produced by GCN inference on the consolidated graph are used for temporal activity detection.

uses a sliding window approach.

The framework is evaluated using two real-world datasets springing from completely different domains: ROAD [180] for situation awareness in autonomous driving and SARAS-ESAD [181] for surgical action detection, both providing video-level annotation in the form of (atomic) action tubes. In this work we augment these datasets with suitable annotation on the start and end time of each instance of complex activity (road activities in ROAD vs surgical phases in SARAS-ESAD). The main contributions of this work are therefore:

- A novel framework for long-term complex activity recognition and localisation.
- An original deformable 3D RoI pooling approach for flexibly pooling features from the various components of the detected tubes, to create an overall representation for activity parts.
- A spatiotemporal scene graph generation and processing mechanism able to cope with a variable number of parts while learning the overall semantics of an activity class.
- Augmented annotation for two newly-released datasets aimed at making them suitable benchmarks for future work on complex activity detection.

Our results clearly indicate that the detection task (both at atomic and complex activity level) is extremely challenging on the real-world data which forms these newly annotated benchmarks, when compared to existing academic datasets. We hope this will stimulate further original thinking to address these challenges. Our method is shown to clearly outperform two recent state of the art graph-based competitors [21, 25] on both augmented datasets.

Related publications: The work presented in this chapter is published in BMVC 2021 [17]. The dissertation author was the primary investigator in [17]. Our code is also available online¹.

Outline: This chapter is organised as follows. Firstly, we start by presenting the proposed methodology of our spatiotemporal deformable scene graphs framework in Section 4.2. The proposed methodology is divided into building blocks including Action Tube Detection, 3d Part-Based Deformable Layer, and Graph Convolutional Network described in Section 4.2.1, Section 4.2.2, and Section 4.2.3, respectively. Next, the detailed experimental evaluation of our proposed method is reported in Section 4.3. Finally, a summary with limitations of the proposed method is given in Section 4.4.

4.2 Proposed Method

Whereas modern complex activity recognition and localisation methods focus on single action or collections of short-term atomic actions, the issue of detecting complex activities articulated into a number atomic actions and spanning longer intervals of time is still rather unexplored.

Crucial to the identification of complex video activities is the modelling of the relations among the constituent actions. Deformable models that represented activities as graphs of atomic parts provide such a framework. The existing part-based methods are investigated for action detection and prediction tasks, but their potential for the recognition of *long-term* activities has not been explored so far.

In this chapter, we propose to achieve this via a combination of deformable pooling of features and a spatiotemporal graph representation that employs multiple semantics. The workflow of our approach is illustrated in Figure 4.1. The input video is divided into fixed-sized snippets which are fed to an action tube detector, followed by a novel 3D deformable RoI pooling stage which computes parts features, which are later processed by GCN for activity detection.

¹https://github.com/salmank255/SDSG_Complex_Activity

4.2.1 Action Tube Detection

To provide a fixed-size representation for the instances of atomic actions composing a complex activity, we adopt AMTnet [1]. AMTnet is a two-stream online action tube detector that uses both RGB and optical flow information (although here we only use the RGB stream). The main rationale for using AMTnet is that it generates tubes in an incremental manner while preserving a fixed-size representation.

Architectural Details. AMTNet uses VGG-16 [182] as baseline CNN feature extractor. The last two fully-connected layers of VGG-16 are replaced by two convolutional layers, and add four extra convolutional layers at the end. AMTNet takes sequence of RGB frames as an input with a fixed temporal interval Δ between consecutive frames, i.e., $\{f_t, f_{t+\Delta}\}$. The input to AMTNet is in the format $[BS \times Sq \times D \times H \times W]$, where BS is the training batch size, Sq is the sequence length (in this case a pair), D is the dimensionality (equal to 3 as we are dealing with RGB frames), while H and W are the height and width of each frame (300×300 in our case). As typical in action detection, AMTNet uses both a classification and a regression layer for recognition and detection, respectively, with the goal of predicting action ‘micro-tubes’ defined by pairs of consecutive detections. The method predicts bounding boxes for a pair of frames separated by fixed gap Δ , while the bounding boxes for intermediate frames are generated by interpolation. In this work, atomic action instances are represented as 3 micro-tubes with $\Delta = 3$ for an overall tube length of $L = 12$ frames, aligned with our snippet length. Complete action tubes are incrementally generated by AMTNet by temporally linking the micro-tubes predicted by the network. (see Figure 4.2).

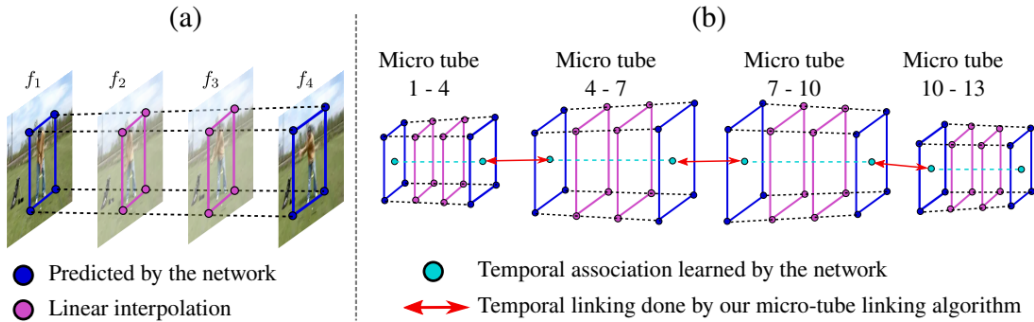


Figure 4.2: (a) AMTNet [1] generates micro-action tubes by predicting bounding boxes for the start and the end frame of a snippet. Predictions for a fixed number of intermediate frames are constructed by bilinear interpolation. (b) Complete action tubes are created by temporally linking micro action tubes by dynamic programming.

4.2.2 3D Part-Based Deformable Layer

The feature extractor in our framework is a novel 3D deformable RoI pooling layer which encodes the spatiotemporal geometry of the action tubes which correspond to the activity parts. This is an extension of the existing standard deformable RoI pooling layer [178] that has the ability to extract and learn features from an action tube rather than a 2D bounding box. The rationale behind using the 3D deformable RoI pooling layer is that it allow us to learn at training time how the geometric shape of atomic action tubes (as regions of the video considered as a spatiotemporal volume) varies across instances of the same class. Intuitively, the shape of the bounding box detections forming a tube (and therefore the shape of the tube itself) will vary with e.g. the viewpoint, as well as the particular style with which the action is performed by a certain agent (for instance, a cyclist can turn right making a narrower as opposed to a wider turn).

The principle of our 3D deformable RoI pooling operation is shown in Figure 4.3. Like the classical deformable RoI pooling layer, our module also includes standard RoI pooling (used in all region proposal-based object detection methods), a fully connected layer, and *offsets* which encode the amount of geometric deformation. Firstly, standard RoI pooling is applied to the provided feature map X and bounding boxes locations forming an action tube ($L \times [x, y, w, h]$), by subdividing the tube into a pooled feature map grid of fixed-size in both the spatial and the temporal dimensions: $L \times k \times k$. Here $L = 12$ is the fixed action tube length, while k is a free parameter which determines the ‘bin size’, i.e., the number of grid locations detections are divided into in each spatial dimension (see Figure 4.3 again). Next, for each bin in the grid, normalised offsets (representing the degree of deformation of the grid components of each action tube) are generated for these feature maps using a fully-connected layer, which are then transformed using an element-wise product with the original RoI’s width and height. Offsets are also multiplied by a scalar value to modulate their magnitude (empirically set to 0.1), making them invariant to the different possible sizes of the RoI. In our framework, this layer takes the VGG features extracted by AMTNet and each detected action tube separately as an input and returns an overall feature map which encodes both the appearance and the shape (through the above offsets) of each atomic action.

4.2.3 Graph Convolutional Network

As our purpose is to achieve a comprehensive understanding of the dynamic scene which comprises a complex activity, we propose to use a graph convolutional network to model and exploit the relations between the constituent action tubes. Unlike the tree structure of classical part-based models (which requires to fix the number of

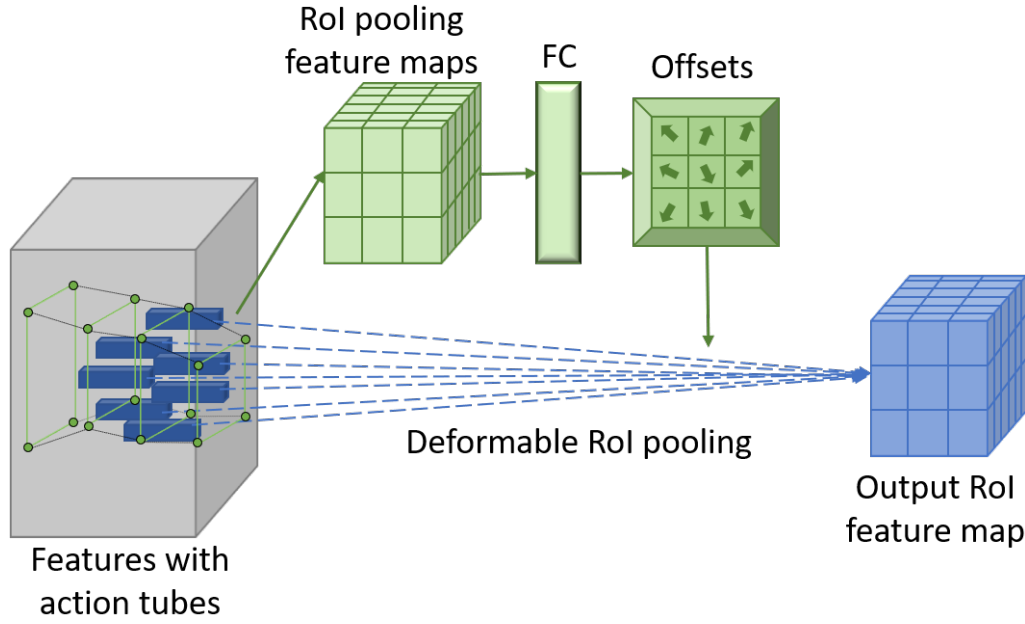


Figure 4.3: Our 3D deformable RoI pooling layer takes feature maps and action tube locations as input and arranges them into a fixed-size grid of components (here illustrated for size 3×3). For each grid component an offset is generated and multiplied by the original tube feature to produce the final component features.

parts [183]), (spatiotemporal) graphs allow us to flexibly describe a complex activity composed by a variable number of actions (nodes) of different type, and to encode the different semantic relationships between them. The functioning of our GCN module is illustrated in Figure 4.4.

Graph Construction. When constructing the activity graphs, the input video is subdivided into consecutive, non-overlapping, fixed-length snippets. For each snippet, a separate graph is built with a variable number of nodes corresponding to the number of detected activity parts (action tubes) within the snippet. The initial representation of the nodes is provided by their RoI features. We consider three different types of connections: (i) the *order of appearance* (from left to right) of each action tube (bearing in mind that in autonomous driving, for instance, road activities tend to follow a specific order, e.g., pedestrian crossing the road followed by vehicles engaging an intersection); (ii) the *spatial similarity* of the node features inspired from the edge convolution proposed in [85], which dynamically constructs edges between the nodes according to their feature distances. The distance threshold in our case is fixed, but can be a learnable parameter by updating via an edge convolution [85]; (iii) *node type*, meant as the sharing the same action label, as this provides very relevant information for the determination of the activity class. As a result three spatiotemporal scene graphs are constructed having the same nodes but with different edges. While the second and third graph are undirected, the appearance order graph is a directed one. However, when merging the three graphs an undirected version of the

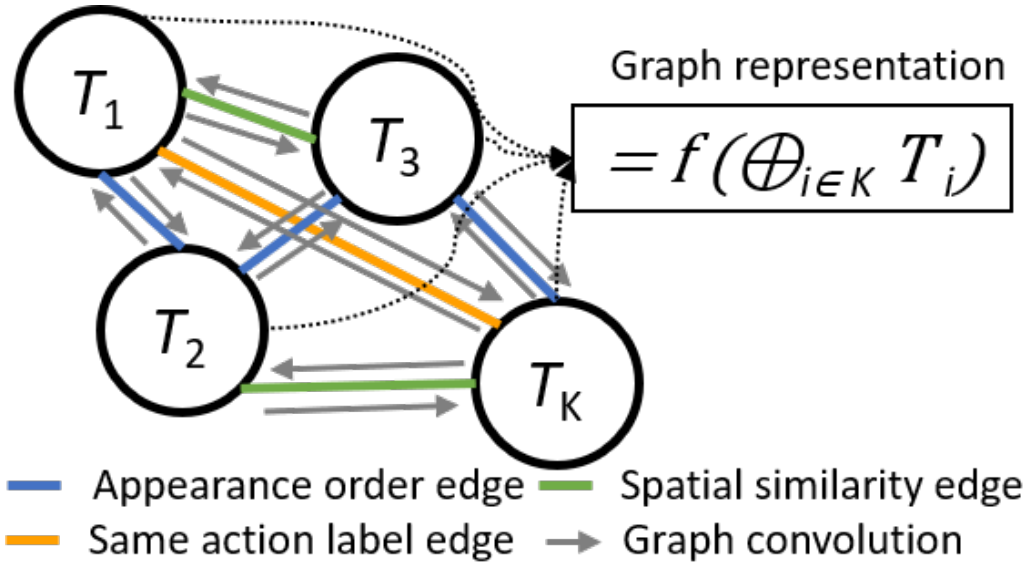


Figure 4.4: Our graph module takes as input the features generated for each tube by the ROI pooling layer $T_{1,2,3...K}$ and builds edges between them according to different semantics. (order of appearance, spatial similarity, label). The overall graph is processed by a GCN to deliver a fixed-size graph representation.

order graph is used. These graphs are then combined by taking a union of all edges to create a single homogeneous graph representing the overall scene. Namely, two nodes are connected in the merged graph iff they are connected in at least one of the three graphs.

Graph Convolution and Representation. Given the final graph, global graph embedding is applied to extract the context of each snippet portraying a complex activity. In our GCN approach we apply a stack of three graph convolutional layers followed by a graph readout layer. The latter encapsulates the final graph representation by taking the mean of the hidden convolutional representation, resulting in a fixed-sized feature representation which is invariant to the number of nodes and edges.

4.2.4 Complete Framework

The complete framework is the concatenation of the aforementioned three modules. Firstly, we divide the video V into N Snippets $S_{1,2,3,...,N}$, with each Snippet S_i consisting of a fixed (M) number of frames: $S_i = F^{1,2,3,...,M}$. Each snippet is passed to the action tube detection module AT which returns K action tubes each composed by $L = 12$ bounding boxes with labels B_A and intermediate VGG features X_A represented as $B_A^{M \times K \times 5}$, $X_A^{M \times 64 \times 300 \times 300} \in AT$. Action tube locations and features are then passed to our 3D deformable RoI pooling layer $DRoI$ which returns a fixed-sized (i.e., 7×7) grid of components whose dimensionality is equal to the number (64) of convolutional layers: $X_{DRoI}^{K \times 64 \times M \times 7 \times 7}$. These features are then fed to our

Algorithm 4.1 Complex Activity Detection

```

1 Input: Input Video  $V$  divided into  $N$  number of Snippets  $S$  consists of fixed  $M$ 
   number of frames  $F^{1,2,3,...,M}$ 
2 Initialization: Load pretrained action tube detector  $ATD$  on desire dataset, import
   our 3D deformable RoI pooling  $DRoI$  and sparsity  $Sp$  modules for  $i \in 1,2,3,...,$ 
    $N$  do
4   Extract features and detect action tubes  $B_A^{M \times K \times 5}, X_B^{M \times 64 \times 300 \times 300} \leftarrow$ 
    $ATD(S_i)$ 
5    $X_{DRoI}^{K \times 64 \times M \times 7 \times 7} \leftarrow DRoI(B_A^{M \times K \times 5}, X_B^{M \times 64 \times 300 \times 300})$ 
6    $X_{Sp}^{K \times 32 \times M \times 7 \times 7} \leftarrow Sp(X_{DRoI}^{K \times 64 \times M \times 7 \times 7})$ 
7    $X_{Fc1}^{K \times M \times 512} \leftarrow Fc1(X_{Sp}^{K \times 32 \times M \times 7 \times 7})$ 
8    $X_{Fc2}^{2048} \leftarrow Fc2(X_{Fc1}^{K \times M \times 512})$ 
9    $CA_i \leftarrow Softmax(X_{Fc2}^{2048})$ 
10  if  $CA_{i-1} == CA_i$  then
11    Same activity
12  else
13    if  $CA_{i-1} == CA_{i+1}$  then
14      same activity
15    else
16      end of activity
17      start new activity
18    end
19  end
20 end
21 Output: Start and end frames with activity labels

```

GCN module G , where a graph with K nodes and E edges is processed to yield a fixed-sized feature representation $X_G^{2048} \in G$. Finally, the latter features are fed to a Softmax classifier to classify the snippet into their respective activity category.

For localisation we use a sliding window approach in which snippets are sequentially classified, using a dual verification mechanism. Namely, as our purpose is to recognise and detect *long-term* activities, if there appears to be a random false positive or false negative between two snippets belonging to the same class we simply ignore it and consider it as having the same activity label. The detection algorithm is described in detail in Algorithm 4.1.

Implementation. Before training our overall architecture, we separately train AMTNet for action tube detection over both datasets. Note that we had to design from scratch suitable data loaders for the two datasets, as the format of the annotation there is completely different from that of the original datasets AMTNet was validated upon. As mentioned, our 3D RoI pooling layer includes a temporal dimension to learn and extract tube parts of a tube rather than of a 2D object. In our experiments we also convert a more recent version of deformable RoI pooling called 'modulated' deformable RoI pooling to the 3D case. In the GCN module we construct graphs

Table 4.1: List of ROAD complex activities with number of train and test snippets splits and brief description.

No	Complex activity	Train snippets	Test snippets	Scenario/Description
1	Negotiating intersection	523	414	Autonomous vehicle (AV) stops at intersection/junction, waits for other vehicles to pass, then resumes.
2	Negotiating pedestrian crossing	381	25	AV Stops at pedestrian crossing signal, waits for pedestrian(s) and cycles to cross, then resumes.
3	Waiting in a queue	570	447	AV stops and waits in a queue for a signal (either pedestrian crossing or junction traffic light).
4	Merging into vehicle lane	68	3	AV stops and waits for bus/car to merge into the vehicle lane from the opposite side of the road.
5	Sudden appearance	3	3	A pedestrian/cycle/vehicle suddenly appears in front of the vehicle, coming from a side street.
6	Walking in the middle of road	13	19	A pedestrian walking in the middle of the road in front of the AV, causing the AV to slow down or stop.

online for each snippet during training using a PyTorch data loader [184]. For the design of the GCN architecture we use the Deep Graph Library (DGL) [185] with a PyTorch back-end, which supports the processing of graphs of various length in a single mini-batch. Overall, architecture is implemented using the PyTorch [184] deep learning library with OpenCV and Scikit-learn. For training we used a machine equipped with 4 Nvidia GTX 1080 GPUs with 12GB VRAM each.

4.3 Experiments

4.3.1 Datasets and Evaluation Metrics

In this chapter, we used two datasets for evaluating our approach, both already annotated at video level for action tubes detection.

ROAD [180]: ROAD (the ROad event Awareness Dataset for autonomous driving) is annotated for road action and event detection. The complete detail of ROAD dataset is given in Section 2.6.

In this work, we augmented the annotation of the ROAD dataset for complex road activity detection. We used a total of 19 videos with an average duration of 8 minutes each, 12 of which were selected for training and the remaining 7 for testing. We temporally annotated the ROAD videos by specifying the start and end frame for six different classes of complex road activities we inferred from video inspection. For example, a ‘Negotiating intersection’ activity class can be defined which is made up of the following ‘atomic’ events: Autonomous Vehicle (AV)-move + Vehicle traffic

Table 4.2: SARAS video phases with number of train and test snippets and a brief description.

No	Complex Phase #	activity/	Train snip- pets	Test snip- pets	Scenario/Description
1	Phase#1	(Bladder mobilization)	1808	999	Tools only used for mobilization
2	Phase#2	(AFP Dis-section)	6263	4455	Anytime only Anterior Prostatic Fat (AFP) is removed
3	Phase#3	(Bladder neck transection)	8353	2419	Whenever the scissor cut the neck, until all the neck is transected
4	Phase#4	(NVP dis-section)	361	46	Located just under the seminal vesicles.
5	Phase#5	(Exposure of seminal vesicles)	2858	800	White bags surrounding the bladder neck. Comprises fat removal.
6	Phase#6	(Urethral division)	2553	1685	Start even when removing fat on the urethra.
7	Phase#7	(Prostate liberation)	1092	2903	After urethral division. It is not AFP. Stops only when completely free.
8	Phase#8	(Vesicourethral anastomosis)	4430	4690	Every time there is string and needle in the scene. Stops when urethra is linked to bladder

light / Green + AV-stop + Vehicle(s) / Stopped / At junction+ AV-move. Activity class statistics are listed and described in Table 4.1.

SARAS-ESAD [181]: ESAD (the Endoscopic Surgeon Action Detection Dataset) is a benchmark devised for surgeon action detection in real-world endoscopic surgery videos described in Section 2.6.

In this work, we took a step forward and annotated ESAD in terms of complex activities corresponding to the different *phases* of the surgical procedure portrayed by the videos (namely, radical prostatectomy). For example, Phase # 3 corresponds to ‘Bladder neck transection’, in which a scissor cuts the neck of the bladder until it is transected. Phases and their statistics are again reported in Table 4.2. The complex activities (surgical phases) in SARAS were defined by professional surgeons experts in radical prostatectomy. As standard in the surgical context, such phases are consecutive without the need for any background activity class. For more details please see [181].

Evaluation Metrics: For the evaluation of action tube detection performance we used the standard frame/video mean Average Precision (mAP) at different IoU thresholds δ (namely, 0.2, 0.3, 0.5, 0.75) on both datasets. Complex activity recognition was evaluated using classification accuracy, precision, recall and F-score. For complex activity localisation we used the standard protocol mAP over the temporal dimension used by all relevant methods.

Table 4.3: Action tube detection performance on both the ROAD and SARAS-ESAD datasets. Both Frame-*mAP* and Video-*mAP* at different IoU thresholds are reported for evaluation

Methods / IoU threshold δ	ROAD				SARAS-ESAD			
	0.2	0.3	0.5	0.75	0.2	0.3	0.5	0.75
Singh et al. [180] (frame- <i>mAP</i>)	-	-	25.9	-	-	-	-	-
Singh et al. [180] (video- <i>mAP</i>)	17.5	-	4.6	-	-	-	-	-
Bawa et al. [181] (frame- <i>mAP</i>)	-	-	-	-	-	24.3	12.2	-
AMTNet (frame- <i>mAP</i>)	22.3	18.1	15.4	11.0	30.4	24.6	18.7	7.9
AMTNet (video- <i>mAP</i>)	11.6	7.9	3.8	-	13.7	10.1	8.8	5.4

4.3.2 Action Tube Detection

A detailed comparative analysis of AMTNet over different action detection datasets can be found in the original paper [1]. Here we briefly report the performance of AMTNet on our two datasets of choice, as AMTNet was never tested there. Table 4.3 reports both frame-*mAP* and video-*mAP* results, and compares AMTNet with the proposed baselines for the two datasets: the ROAD baseline (termed *3D RetinaNet* [180]), and the ESAD baseline [181], a vanilla implementation of RetinaNet (only providing frame-level results). AMTNet performed better than [181] on SARAS-ESAD, while being inferior to [180] on ROAD. Remember that the main rationale for using AMTNet is that it can provide a fixed-size representation for the tubes (as required by our framework), motivating us to compromise on accuracy.

4.3.3 Complex Activity Recognition

Next, we provide a detailed analysis of complex activity *classification* using our approach on both the ROAD and SARAS-ESAD datasets. The performance for each class in both datasets is illustrated in Fig. 4.5 and Fig. 4.6 using all metrics. It is apparent that the ROAD dataset is characterised by significant fluctuations in class-wise performance, with higher recognition accuracy for activities that appear more often, e.g. ‘waiting in a queue’, as opposed to infrequent ones (e.g. ‘sudden appearance’). In SARAS-ESAD each activity class does contain enough samples for good training, while the diversity in phases still poses a challenge. In fact, performance on this dataset is a function of the complexity of the phases and the visual similarity of the constituent atomic actions, not just the amount training data. For example, Phase 5 and 6 both include ‘fat removal’ actions, complicating their differentiation.

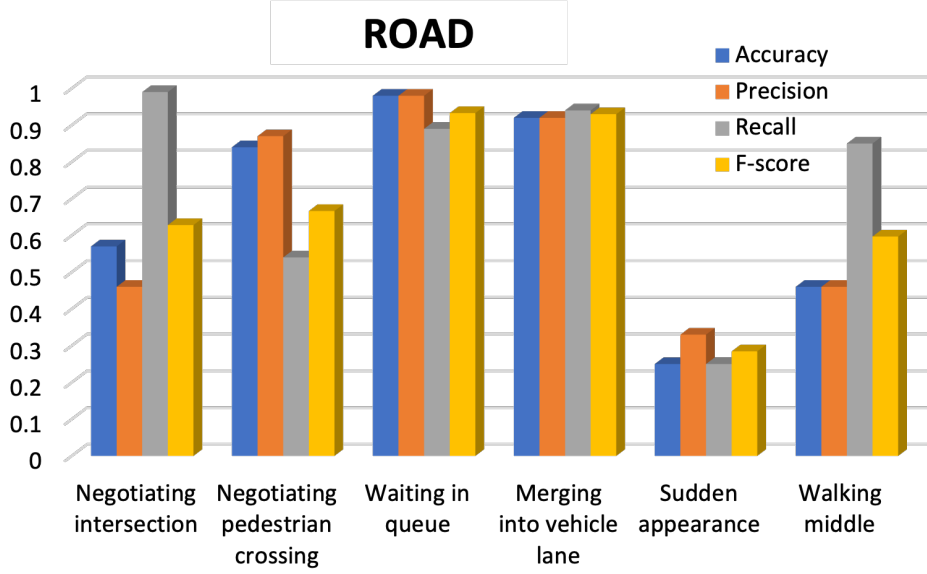


Figure 4.5: Complex activity classification performance over ROAD dataset.

Table 4.4: Comparative analysis of temporal activity localisation performance on ROAD and SARAS-ESAD, reporting mAP (%) at five different IoU thresholds.

Methods / IoU threshold δ	ROAD					SARAS-ESAD				
	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
P-GCN [21]	60.0	56.7	53.9	50.5	46.4	57.9	55.6	53.4	49.0	45.1
G-TAD [25]	62.1	59.8	55.6	52.2	48.7	59.1	56.7	54.5	49.8	46.9
Ours	77.3	74.6	71.2	66.7	59.4	62.9	59.6	58.2	55.3	51.5

4.3.4 Temporal Activity Detection - Comparison with State-of-the-Art

To evaluate the performance of our complex activity detection approach we reimplemented two state-of-the-art activity localisation methods – P-GCN [21] and G-TAD [25], as neither ROAD nor ESAD were ever used for complex activity detection. The major changes we made during re-implementation are: (i) data loading (as both P-GCN and G-TAD were designed to be trained and tested on pre-extracted features), and (ii) replacing the regression part with a sliding window approach for the SARAS-ESAD dataset, as the latter lacks a background class.

ROAD. For activity detection in ROAD we used an additional ‘background’ class, which indicates either no action or presence of action(s) without any solid indication. Whenever no action tube was detected we would use the entire frame as RoI for our parts deformation module to understand the overall scene. Temporal activity detection performance on ROAD, measured via mAP at five different IoU thresholds, is reported in Table 4.4 for both our approach and the two competitors.

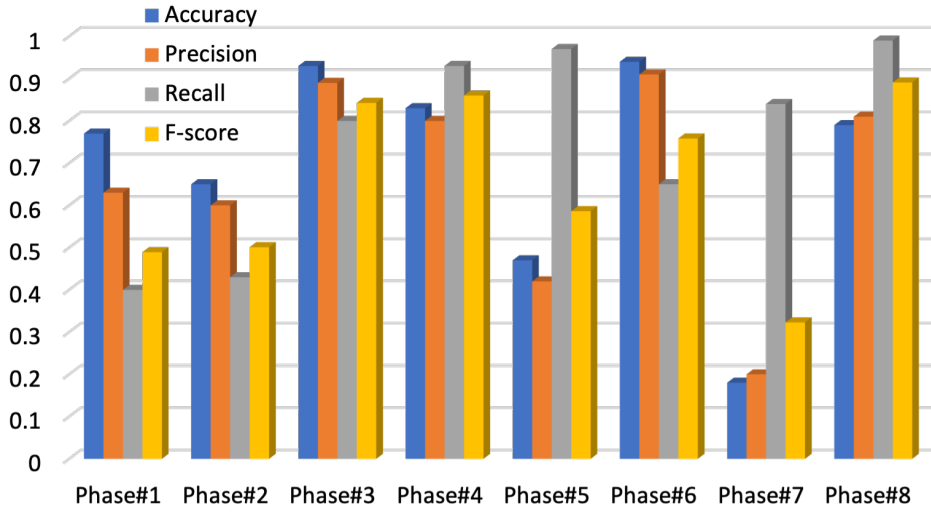


Figure 4.6: Complex activity classification performance on SARAS-ESAD dataset.

Table 4.5: ROAD activity localisation performance (mAP, %) for each complex road activity, at a standard IoU threshold of 0.5.

Method / Activities	Negotiating intersection	Negotiating pedestrian crossing	Waiting in queue	Merging into vehicle lane	Sudden appearance	Walking middle of road
P-GCN [21]	44.3	53.8	74.4	50.1	21.7	34.1
G-TAD [25]	47.8	57.3	70.6	55.2	24.3	37.1
Ours	51.2	72.3	89.8	84.1	17.8	41.3

Class-wise results for each complex activity at a standard IoU threshold of 0.5 are reported in Table 4.5.

SARAS-ESAD. Temporal activity detection on this dataset much relates to activity recognition, as surgical phases are contiguous. Both the average *mAP* of the methods at five IoU thresholds and the class-wise performance for each activity (phase) at a standard IoU threshold of 0.5 are reported in Table 4.4 and 4.6, respectively. From the results it is clear how our method outperforms the chosen state-of-the-art methods by a reasonable margin.

Table 4.6: SARAS-ESAD activity localisation performance (mAP, %) for each activity at a fixed IoU threshold of 0.5.

Method / Activities	Phase#1	Phase#2	Phase#3	Phase#4	Phase#5	Phase#6	Phase#7	Phase#8
P-GCN [21]	56.7	43.2	52.3	59.1	33.8	59.4	14.8	41.2
G-TAD [25]	51.1	46.6	57.2	63.8	29.4	62.2	19.3	45.7
Ours	57.5	54.1	69.3	60.2	31.1	71.3	16.5	52.4

4.3.5 Ablation Studies

We ran a number of ablation studies on both datasets to understand the influence of the various components of our model.

Effect of GCN. First of all, we demonstrated the significance of our GCN by conducting three different types of experiments; i) Snippet features from the entire scene without engaging in action tube detection or utilising GCN.. ii) Action features without GCN where we first detect the action tubes and then extract the deformable features from the tubes. It is worth noting that, in this particular instance, we constrained the number of action tubes to a fixed limit of 10. iii) Action features with GCN, our proposed framework, incorporating action features and integrating the GCN. The clear difference among these three types of setups can be observed in Fig. 4.7.

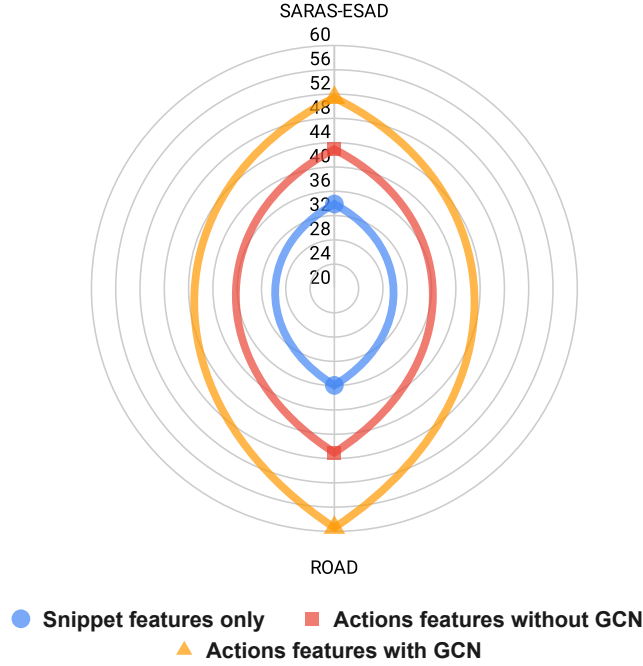


Figure 4.7: The performance of different components of our proposed method over ROAD and SARAS-ESAD dataset. The mAP at a threshold of 0.5 is reported for the test set of both datasets.

Effect of Edges. Next, we ablated the effect of various types of edges employed in our GCN for graph construction. This ablation analysis includes all three types of edges used for graph construction (explained in Section 4.2.3) individually and combined. The comprehensive performance for each edge type over both ROAD and SARAS-ESAD datasets is presented in Fig. 4.8.

Effect of Sequence Length. Lastly, we showed how snippet duration, which represents the time span of the local scene impacts our model. We conducted exper-

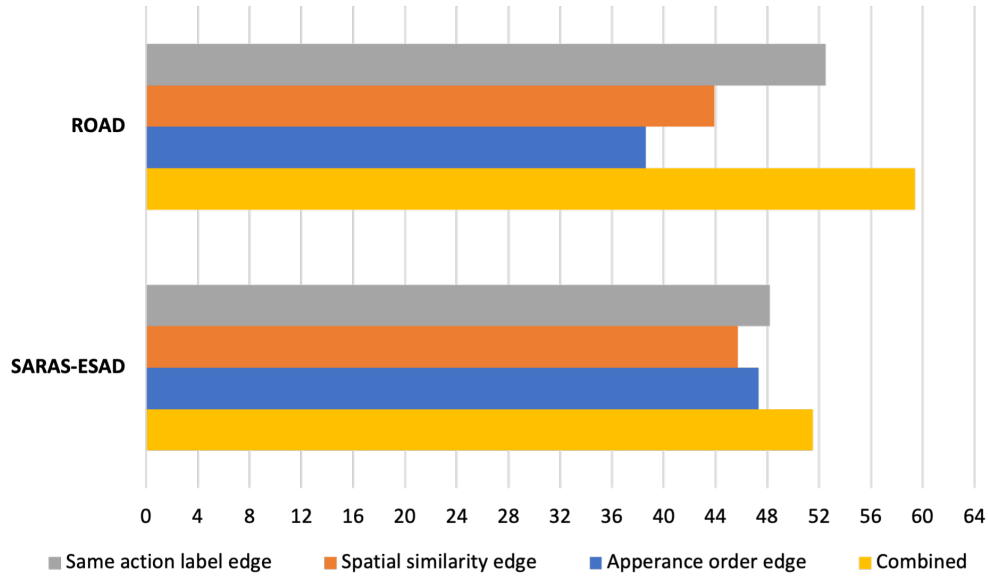


Figure 4.8: The performance comparison of various edge types in our GCN modules. The mAP at a threshold of 0.5 is reported for the test set of both datasets.

iments with four different sizes (12, 18, 24, and 30), as given in Table 4.7. Notably, when evaluating our model on ROAD, the most favorable results were achieved with a sequence length of 18. This choice is attributed to the datasets’ characteristics, as the activities within them tend to be of average duration. Conversely, when evaluating on SARAS-ESAD, we observed superior performance with a sequence length of 24, aligning with the predominantly the much longer duration of activities in this dataset.

Table 4.7: The performance comparison of different sequence lengths over both ROAD and SARAS-ESAD dataset, reporting mAP (%) at two different IoU thresholds.

Sequence length / IoU threshold δ	ROAD		SARAS-ESAD	
	0.1	0.5	0.1	0.5
12	77.3	59.4	62.9	51.5
18	79.7	60.8	64.6	52.8
24	74.1	55.3	67.2	54.5
30	70.6	52.2	66.1	53.7

4.4 Summary of the chapter

4.4.1 Limitations and Future Work

The main limitation of this work is that it relies on action tube detection. From our results, the existing tube detectors are not reliable enough to perform well over

challenging real-world datasets such as those we adopted here. Clearly, if the tube detector misses an important atomic action this will affect the overall activity detection performance. Nevertheless our results show that, even when using a suboptimal detector, our approach is capable of significantly outperforming state of the art methods on our new benchmarks.

Detection is challenging on SARAS and ROAD because of their real-world nature: surgical images are indistinct, road scenes come with incredible variations. These benchmarks show how even the best detectors suffer a huge drop in performance when moving from ‘academic’ benchmarks to real ones. We hope the realism of these two extremely challenging datasets will stimulate real progress and new original thinking in the field.

In the future our primary target will be the design of a more accurate action tube detector with the ability to perform better in challenging scenarios such as those portrayed in ROAD or SARAS-ESAD. We will also explore the end-to-end training of the entire model in all its three components. Further down the line, we will update our S/T scene graph approach to properly model the heterogenous nature of the graph [186], and extend it to a more complete representation of complex dynamic events in which nodes (rather than correspond all to action tubes) may be associated with any relevant elements of a dynamic scene, such as objects, agents, actions, locations and their attributes (e.g. red, fast, drivable, etc).

4.4.2 Conclusion

In this chapter, we presented a spatiotemporal complex activity detection framework which leverages both part deformation and a heterogenous graph representation. Our approach is based on three building blocks; action tube detection, part-based deformable 3D RoI pooling for feature extraction and a GCN module which processes the variable number of detected action tubes to model the overall semantics of a complex activity. In an additional contribution, we temporally annotated two recently released benchmark datasets (ROAD and ESAD) in terms of long-term complex activities. Both datasets come with video-level action tube annotation, making them suitable benchmarks for future work in this area. We thoroughly evaluated our method, showing the effectiveness of our 3D part-based deformable model approach for the detection of complex activities.

Chapter 5

A Hybrid Graph Network for Complex Activity Detection in Video

5.1 Introduction

Following the limitations of our previous work (relies on action tube detection) explained in Chapter 4. In this Chapter, we propose a new **Complex Activity Detection** (CompAD) paradigm that can be applicable to a variety of computer vision domains, including autonomous driving and sports analytics without additional action annotation. CompAD can be seen as the advanced version of Temporal Action Localisation (TAL), which aims to recognise what actions take place in a video and identify their start and end time. While TAL mostly focuses on short-term actions, CompAD extends the analysis to long-term activities, and does so by modelling a complex video activity’s internal structure. In this chapter, we address the CompAD problem using a hybrid graph neural network that combines attention applied to a graph encoding the local (short term) dynamic scene with a temporal graph modelling the overall long-duration activity. Our contribution is threefold. i) Firstly, we propose a novel feature extraction technique which, for each video snippet, generates spatiotemporal ‘tubes’ for the active elements (‘agents’) in the (local) scene by detecting individual objects, tracking them and then extracting 3D features from all the agent tubes as well as the overall scene. ii) Next, we construct a local scene graph where each node (representing either an agent tube or the scene) is connected to all other nodes. Attention is then applied to this graph to obtain an overall representation of the local dynamic scene. iii) Finally, all local scene graph representations are temporally connected to each other via a temporal graph, with the aim of estimating the complex activity class together with its start and end time. The proposed framework outperforms all previous state-of-the-art methods on three benchmark datasets: Thumos-14, ActivityNet-1.3, and the more recent ROAD.

5.1.1 Motivation

The ability to detect and recognise activities from untrimmed videos is a research problem which is attracting significant attention, due in part to the rapid growth of online video generation in domains including sports [6], autonomous driving [2], medical robotics [187] and surveillance [188]. The current state-of-the-art in action detection is *Temporal Action Localisation* (TAL) - whose approaches not only recognise the action label(s) present in a video, but also identify the start and end time of each action instance occurring. TAL can be instrumental in generating sports highlights [16, 61], understanding road scenes in autonomous driving [17], video summarisation in surveillance videos [18] and video captioning [19, 189]. A number of TAL methods [20–29] have recently been proposed, competing to achieve state-of-the-art performance on selected benchmark datasets. Whereas various new datasets have been recently proposed, the two most common relevant benchmarks remain ActivityNet 1.3 [30], and Thumos-14 [31]. State-of-the-art performance on Thumos-14 has improved in four years by some 19% [32].

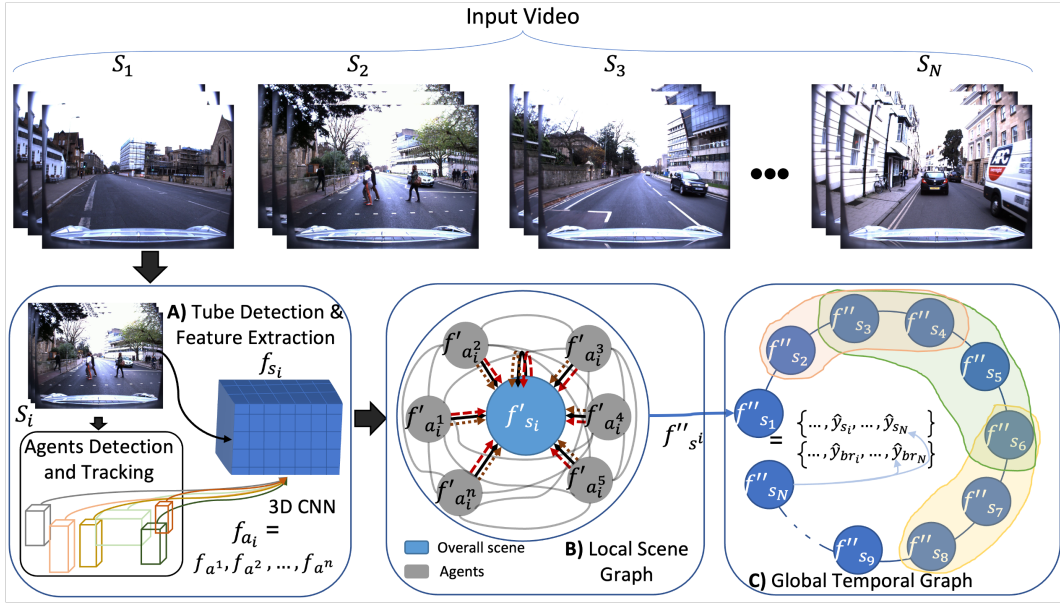


Figure 5.1: An overview of our **Complex Activity Detection (CompAD)** framework. The input video is divided into fixed-size snippets $S_1, \dots, S_N$ (top); each snippet is then processed in three major steps (bottom). A) Firstly, scene objects (agents) are detected and tracked throughout the snippet to form agent tubes. 3D features are then extracted from all the cropped agent tubes (f_{a_i}) as well as the local scene (f_{s_i}). B) Next, a fully-connected scene graph is constructed where agent nodes (in gray) are connected to each other and to the snippet node (in light blue). The local scene graph is processed using a graph attention network (GAT), resulting in intermediate scene features (f'_{s_i}). C) Finally, all local scene features associated with individual snippets are temporally connected and processed using a global temporal graph to identify the boundaries of the activity using anchor proposals (shown in different colors). Each of the anchors represents a binary graph mask where snippets which belong to the activity receive label $y_{br_i} = 1$, 0 otherwise.

Almost all TAL approaches contemplate two major aspects: *features/scene representation* and *temporal localisation*. In the scene representation stage, snippets (continuous sequences of frames) are processed to understand the local scene in the video. Methods [26–29] typically employ pre-extracted features obtained using a sequential learning model (e.g., I3D [33]), often pre-trained on the Kinetics [34] dataset. Such features are then processed, e.g., via a temporal or semantic graph neural network, by applying appropriate feature encoding techniques or by generating temporal proposals, in an object detection style [35]. In the second step of TAL, actions need to be temporally localised. Existing methods vary in the way this is done, e.g. through temporal graphs [21, 25], boundary regression and proposals generation [20, 22, 28] or encoder-decoder approaches [32]. All such methods, however, are only applicable to short- or mid-duration actions that last for a few seconds (e.g., a person jumping or pitching a baseball).

As recently pointed out in e.g. [36, 37], in real-world applications a challenge is posed by *complex activities*, longer-term events comprising a series of elementary actions, often performed by multiple agents. For example, an Autonomous Vehicle (AV) negotiating a pedestrian crossing is engaged in a complex activity: First it drives along the road, then the traffic lights change to red, the vehicle stops and several pedestrians cross the road. Eventually, the lights turn green again and the AV drives off.

Theoretically, TAL methods can be employed to temporally segment complex activities, in practice such approaches are only employed to detect short- or mid-duration actions lasting a few seconds at most (e.g., a person jumping or pitching a baseball). The activities contemplated by the most common datasets are of this nature. Fig. 5.2 compares two standard TAL benchmarks with the recently released ROAD dataset, explicitly designed for complex activity detection. Activities in the ROAD dataset last longer than those in ActivityNet or Thumos, with twice as many agents per snippet, making them more complex in nature.

In this paper, we argue that standard TAL approaches are ill equipped to detect complex activities, as they fail to model both the global temporal structure of the activity and its fine-grained structure, in terms of the agents and elementary actions involved and their relations.

We may thus define (strong) **Complex Activity Detection** (CompAD) as the task of recovering, given an input video, the temporal extent of the activities there present, *as well as* their inner structure in terms of the agents or elementary actions involved. A weaker CompAD is one in which the only expected output is the temporal segmentation, with the internal structure of the activity estimated as a means for achieving segmentation - with no annotation available.

While a small number of studies have attempted to detect complex, long-duration

activities [36–38], to our knowledge [17] is the only existing study which attempts to tackle CompAD as defined above. The work, however, relies upon the availability of heavily-annotated datasets which include granular labels for the individual actions which make up a complex activity, and the corresponding frame-level bounding boxes. This is a serious limitation, for it prevents [17] from being usable for pure temporal segmentation and compared with prior art there. Further, [17] focuses only on the graph representation of snippets, neglecting the long-term modelling of complex activities.

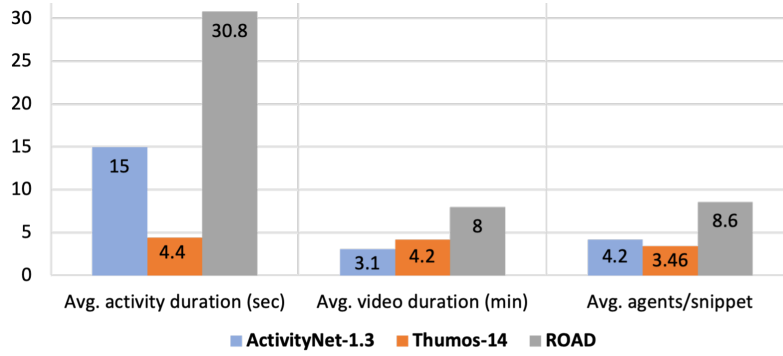


Figure 5.2: Comparing TAL datasets (ActivityNet-1.3 and Thumos-14) with a CompAD dataset (ROAD) in terms of average activity and video durations and mean number of agents per snippet.

5.1.2 Objectives

This work aims to push the boundaries of temporal action localisation towards full-blown *Complex Activity Detection* (CompAD), intended as the ability to detect long-term, complex activities, leveraging datasets providing temporal segmentation annotation *only*. We do so by modelling and leveraging a complex video activity’s internal structure, but without resorting to any additional fine-grained annotation concerning individual actions.

Our **proposed framework** (Fig. 5.1) is composed of three stages: A) feature extraction; B) a scene graph attention network designed to learn the importance of each active object (‘agent’) within the local dynamic scene; and C) a temporal graph of attended scene graphs for the localisation of complex activities of arbitrary duration.

Our feature extraction scheme (A) differs from what typically done in prior art - where spatiotemporal features are extracted from whole snippets only. In contrast, we first detect the relevant active objects in the scene and track them within each snippet to build for each of them an *agent tube* (a series of related bounding boxes). A pre-trained tracker is used to allow the method to be deployable to any datasets with only temporal annotation (no bounding box annotation for the scene agents is required), while enabling a finer-grained description of the internal structure of a

complex activity. Agent tubes are resized to a standard spatial size to create ‘mini videos’ from which features can be extracted. A pre-trained 3D feature extraction model is then used to encode both the ‘cropped’ tubes and the whole snippet.

To suitably represent the local dynamic scene within each snippet, a *local scene graph* (B) is constructed using three different topologies including a star graph, star graph with same agent label, and a fully-connected (the effects of all three topologies are explained in Section 5.3.4 - Ablation study). The local scene graph is then processed using a *scene graph attention (SGAT) network* for extracting the overall scene representation. The rationale behind using the SGAT is its ability to compute the importance of each node in the context of its neighbors, thus modelling the structure of complex scenes.

Finally (C), the learned local scene graphs are connected to each other by constructing a *temporal graph*, with the aim of recognising the activity label and identifying its temporal boundaries (start and end time) in a class-agnostic manner. Each node in the temporal graph learns two things; differentiating between the background and the activity region and recognising the activity label. Such a strategy enables us to model and detect both short and long-duration activities.

5.1.3 Contributions

Our main contributions are, therefore:

- A hybrid graph network approach for general complex activity detection, comprising a local scene graph as well as a global temporal graph, capable of localising both long and short-term activities.
- A feature extraction strategy for fine-grained temporal activity localisation without the need for additional annotation, consisting in detecting and tracking the relevant scene agents to construct tubes, followed by feature extraction from both the whole scene and each agent tube individually, as opposed to existing methods which only rely on snippet-level features.
- Comprehensive experiments showing how our approach leveraging weak CompAD *outperforms the most recent TAL state-of-the-art across the board* on both Thumos-14 and the recent ROad event Awareness Dataset (ROAD) [17,45], which portrays long-duration road activities involving multiple road agents over sometimes several minutes, while being extremely competitive on the standard ActivityNet 1.3 benchmark, showing a clear dominance on classical TAL approaches as the duration and the complexity of the activities increase.

Related publications: The work presented in this chapter is a preprint ArXiv. The dissertation author is the primary investigator of this research work. Our code is also available online¹.

Outline: The remainder of the Chapter is organised as follows. The proposed methodology of our framework is presented in Section 5.2 while the main building blocks including tube detection and features extraction in Section 5.2.1, scene graph attention in 5.2.2, temporal graph localisation in 5.2.3, and loss functions in 5.2.4, respectively. Further, the detailed experimental evaluation of the proposed method is reported in Section 5.3. Finally, a summary with limitations of the proposed method is given in Section 5.4.

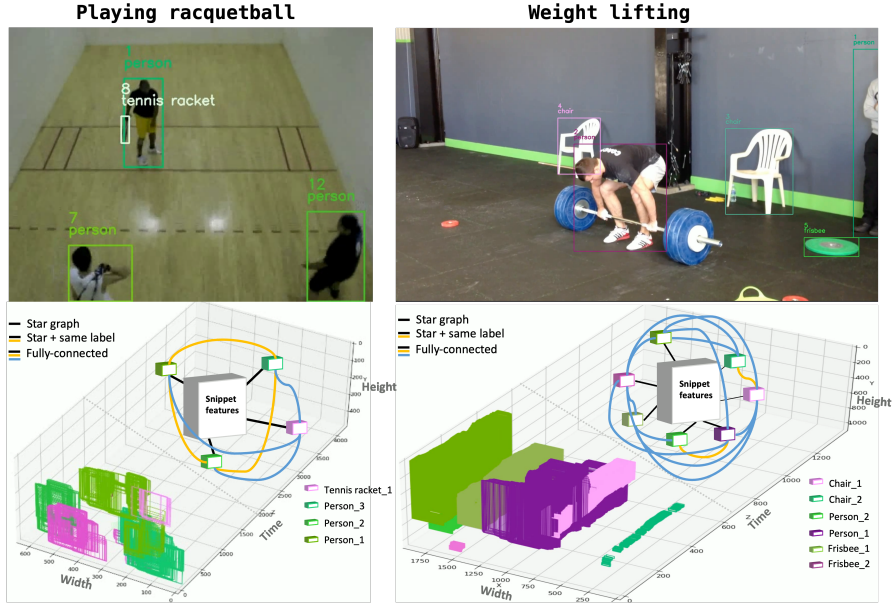


Figure 5.3: Visualisation of our agent detection and tracking stage using a bird's-eye view of the spatiotemporal volume corresponding to a video segment of ActivityNet-1.3 dataset. The first half of the graph shows the plotting of agent tubes while the second half illustrates the scene graph of the detection agents. In this figure, we show examples of activities performed by multiple actors (agent tubes).

5.2 Proposed Method

The proposed framework, outlined in the Introduction, is illustrated in Fig. 5.1. An input untrimmed video V is divided into N snippets $S = S_1, \dots, S_i, S_{i+1}, \dots, S_N$ (each snippet is a pre-defined constant length of consecutive frames). Each snippet is then passed to the tube detection and feature extraction module which returns a feature vector for both the snippet f_{s_i} and the individual agent tubes $f_{a_i^1}, f_{a_i^2}, \dots, f_{a_i^n}$, where n

¹https://anonymous.4open.science/r/HG_CompAD-7697/README.md

is the number of agents present in snippet i . These features are then forwarded to the local scene graph attention layer for learning the attention of each agent in the context of its neighbours. This returns an aggregate feature representation for the whole scene (f''_{s_i}). These aggregate local scene features for all the snippets, $f''_{s_1}, f''_{s_2}, \dots, f''_{s_N}$, are then connected using a global temporal graph for the generation of the activity class label $\hat{y}_{s_i}$ and activity boundary labels $\hat{y}_{br_i}$ using anchor proposals in a class-agnostic manner.

5.2.1 Tube Detection and Feature Extraction

As mentioned, one of the contributions of this chapter is a new strategy for feature extraction and representation which consists in analysing the finer structure of the local dynamic scene, rather extracting features from whole snippets only.

Objects Detection and Tracking. We first detect scene objects in each frame of the snippet using an object detector pre-trained on the COCO dataset [190], which comprises of 80 different types of objects. However, we select a relevant list of object types (termed *agents*) which depends on the dataset (the lists of classes selected for each dataset is given in Section 5.3.2 – Implementation Details). The detected agents are then tracked using a pre-trained tracker throughout the snippet in order to construct agent tubes. The latter are of variable length, depending on the agent itself and its role in each snippet. A pictorial illustration of agent tubes in an example video segment from all three datasets, showing both a sample frame, with overlaid the detection bounding boxes, and a bird’s-eye view of the agent tubes and a local scene graph for the snippet it belongs to.

ActivityNet-1.3 is one of the largest temporal action localisation datasets which includes activities performed by both individual and multiple objects (agents). Therefore, we illustrate the example of multi-agent activities in Fig. 5.3.

Thumos-14 covers different types of sports actions performed by single or mostly multiple agents. Two sample actions; baseball pitch and soccer penalty are visualised in Fig. 5.4. In the figure, it can be seen that our model construct the seen graph of almost all agents despite the low-resolution videos.

ROAD. The activities in the road dataset are mostly performed by a large number of agents, however, there are a few cases where the number of detected agents is less in number as shown in Fig. 5.5 (above). On the other hand, in Fig. 5.5 (below) we also show an example of a night video where the activity is performed by a large number of agents moving very promptly.

Feature Extraction. Next, all the detected agent tubes are brought to a standard size and passed to the pre-trained 3D CNN model along with the whole snippet for spatiotemporal feature extraction. The adopted 3D CNN model allows variable

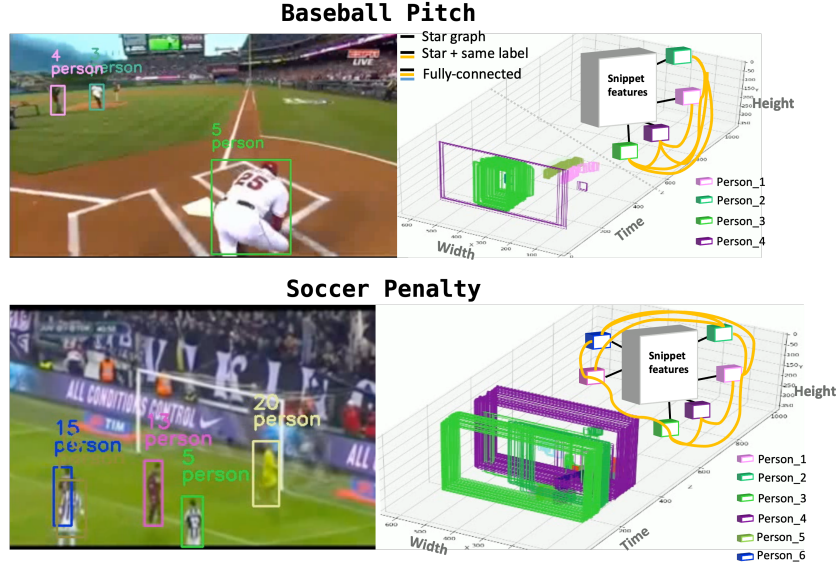


Figure 5.4: Visualisation of our agent detection and tracking stage using a bird’s-eye view of the spatiotemporal volume corresponding to a video segment of Thumos-14 dataset. The first half of the graph shows the plotting of agent tubes while the second half illustrates the scene graph of the detection agents. Both of the examples show the activities performed by multiple agents (recognising sports activities).

length inputs, and outputs a fixed-sized feature vector for each of the agent tubes and the snippet.

5.2.2 Scene Graph Attention

In a second stage, a scene graph representation is used to describe the scene in terms of features extracted from both the overall snippet and the agent tubes.

Graph Generation. The scene graph is constructed using three different topologies: a star graph connecting each agent node to the scene node, a star topology with also links between agents sharing a label, and a fully-connected one. The influence of topology is shown in Sec. 5.3.4.

Graph Attention. As mentioned, our scene graph attention layer is inspired by the GAT concept [94], originally designed for node classification. While we follow a similar attention mechanism, here we amend the attention layer in order to extract aggregate features from all the nodes, to be passed in turn to our localisation layer in the third stage.

The workflow of our scene graph attention layer is shown in Fig. 5.6. All input node features are linearly transformed using a weight matrix W_1 , followed by *self-attention* to find the importance of each node with respect to its connected neighbours. An activation function LeakyReLU is applied to the features for nonlinearity, resulting in final output features for all of the nodes. These are then normalised using

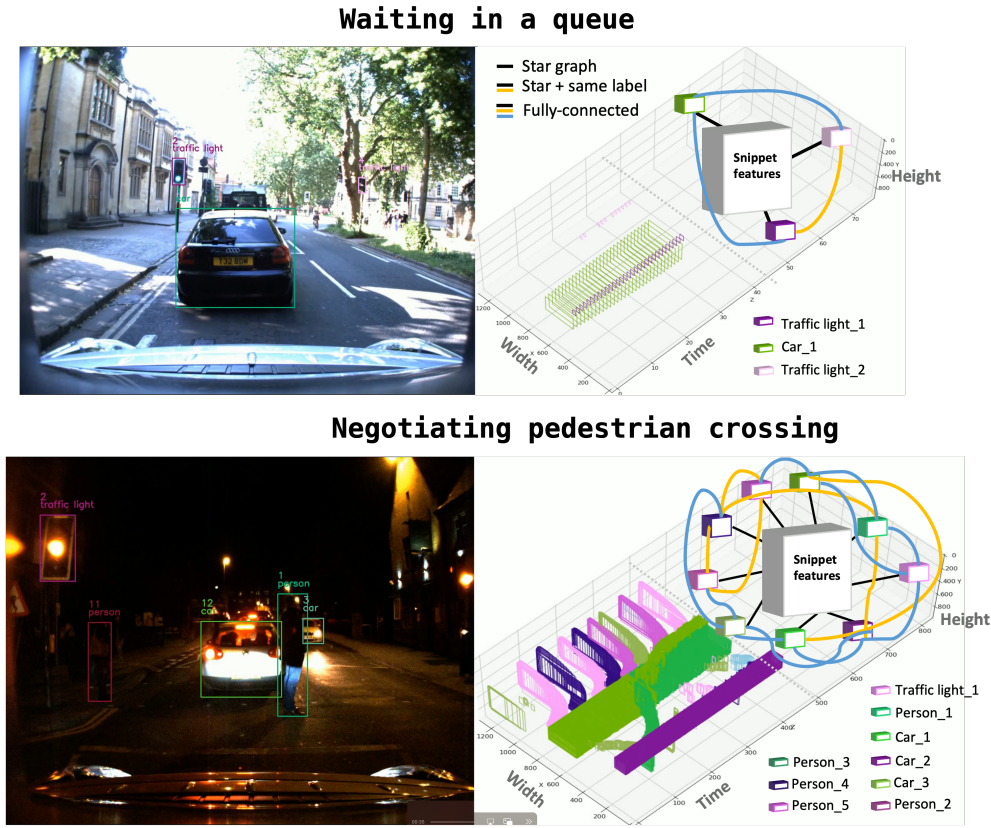


Figure 5.5: Visualisation of our agent detection and tracking stage using a bird’s-eye view of the spatiotemporal volume corresponding to a video segment of the ROAD dataset. The upper section shows an example of activity performed by only three agents. Below is the example of a night video where the activity is performed by multiple agents. The first half of the graph shows the plotting of agent tubes while the second half illustrates the scene graph of the detection agents.

a softmax function to make each node representation ($f'_{s_i a_i^j}$) comparable with that of all the nodes connected to it (f'_{s_i}).

The self-attention process is further improved by applying a *multi-head attention* strategy inspired by transformers [94]. Attention is applied to the node features individually. For each node, the average over the H heads is computed, resulting in an attended feature vector for each node.

Finally, to get a fixed-size representation for the whole scene, the output features of all the nodes are aggregated using another learnable weight matrix W_2 , which outputs the final feature representation f''_{s_i} for the whole (local) scene.

5.2.3 Temporal Graph Localisation

In Stage C of our framework, activity recognition and localisation are performed using a GCN. The final features from all the local scenes ($f''_{s_1}, \dots, f''_{s_i}, f''_{s_{i+1}}, \dots, f''_{s_N}$), as outputted by the scene graph attention layer, are temporally connected to build a

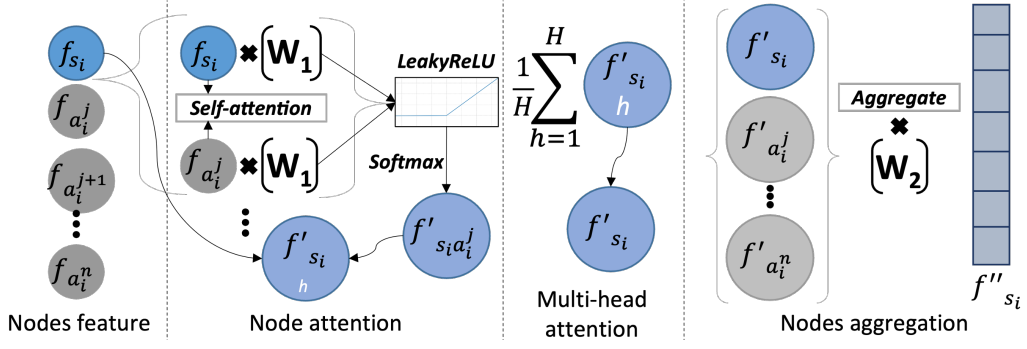


Figure 5.6: Our scene graph attention layer (Stage B of our approach, Fig. 5.1) takes the node features generated in the feature extraction Stage A as an input, and updates each node’s features using node attention with respect to its all connected neighbours. Multi-head attention with H heads is applied to each node to further robustify the representation – averaging yields the final node features. To obtain a fixed-size overall representation for a specific local scene (snippet), the features of all its nodes are aggregated using a learnable weight matrix W_2 .

global temporal graph (see Fig. 5.1, C).

Our GCN network for processing the global temporal graph is divided into two parts. The first part consists of three 1D convolutional layers designed to learn the temporal appearance of all the local scenes with boundaries. Each convolutional layer is followed by a sigmoid activation for non-linearity. The second part generates the anchor proposals of the temporally learned features via pre-defined anchors, where each of the anchors acts as a binary mask over the whole graph.

Overall, the GCN module provides two different outputs. i) *Activity classification*: the list of predicted activity labels for all the snippets in the video ($\hat{y}_{s_1}, \dots, \hat{y}_{s_i}, \dots, \hat{y}_{s_N}$) is produced, where the dimensionality of the output vector $\hat{y}$ is equal to the number of classes. ii) *Activity localisation*: activities are localised using binary masked class-agnostic anchor proposals. The Intersection over Union (IoU) measure between each anchor and the ground truth (true temporal extension of the activity) is computed, and the anchors with maximum IoU are selected to train the model for localising the boundaries of any activity, regardless of its activity label. The final output of our temporal graph is a one-hot binary vector ($\hat{y}_{br_1}, \dots, \hat{y}_{br_i}, \dots, \hat{y}_{br_N}$) for each series of snippets (video), where $\hat{y}_{br_i} = 1$ iff snippet S_i belongs to the activity, $= 0$ when the snippet does not belong to it.

5.2.4 Loss Functions

Our problem is multi-objective, as we aim at not only recognising the label of the activity taking place but also finding its boundary (start and end time). Therefore, we combined two losses one for learning the class label (classification) and the other for finding the start and end time (regression). The details of both losses are as follows:

Given the ground truth $y_s = \{y_{s_1}, \dots, y_{s_i}, \dots, y_{s_N}\}$ and predicted $\hat{y}_s = \{\hat{y}_{s_1}, \dots, \hat{y}_{s_i}, \dots, \hat{y}_{s_N}\}$ activity labels, and the ground truth $y_{br} = \{y_{br_1}, \dots, y_{br_i}, \dots, y_{br_N}\}$ and predicted $\hat{y}_{br} = \{\hat{y}_{br_1}, \dots, \hat{y}_{br_i}, \dots, \hat{y}_{br_N}\}$ temporal extent labels, our overall loss function is formulated as:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{Act} + \mathcal{L}_{Br}. \quad (5.1)$$

The second component denotes the binary cross entropy (BCE) loss

$\mathcal{L}_{Br} = \{l_1, \dots, l_i, \dots, l_N\}^T$, where

$$l_i = -w_i \left[y_{br_i} \cdot \log \hat{y}_{br_i} + (1 - y_{br_i}) \cdot \log(1 - \hat{y}_{br_i}) \right],$$

driving the recognition of the activity's temporal extent in class-agnostic manner.

$\mathcal{L}_{Act}$ denotes, instead, the *BCEWithLogitsLoss* [191], defined as

$\mathcal{L}_{Act} = \{l_{1,c}, \dots, l_{i,c}, \dots, l_{N,c}\}^T$, where

$$l_{i,c} = -w_{i,c} \left[p_c \cdot y_{s_i,c} \cdot \log \sigma(\hat{y}_{s_i,c}) + (1 - y_{s_i,c}) \cdot \log(1 - \sigma(\hat{y}_{s_i,c})) \right]. \quad (5.2)$$

Here i is the index of the sample in the batch, c is the class number, p_c is the weight of a positive sample for class c , and σ is the sigmoid function. This loss component is used to predict the activity labels $\{y_{s_i}\}$, and combines classical binary cross entropy with a sigmoid layer [191]. This combination is proven to be numerically stronger (as it leverages the log-sum-exp trick) than using a plain sigmoid separately followed by a BCE loss.

Finally, λ is a weight term, which we set to the number of selected anchor proposals (128 in our case).

5.3 Experiments

5.3.1 Datasets

To evaluate the performance of our proposed method, we used three benchmark datasets including ROAD, Thumoas-14, and ActivityNet-1.3. Firstly, ROAD is used to compare our method with state-of-the-art CompAD methods. Next, Thumos-14 and ActivityNet-1.3 are used to show the effectiveness of our method over temporal action localisation methods. The complete detail of all these three datasets can be found in Section 2.6.

5.3.2 Implementation Details

Evaluation metrics. In our experiments, mean Average Precision (mAP) was used as an evaluation metric (explained in Section 2.7), using different IoU thresholds for the different datasets. According to the official protocols for the various benchmarks, the following lists of temporal IoU thresholds were selected: $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ for ROAD, $\{0.3, 0.4, 0.5, 0.6, 0.7\}$ for Thumos-14 and $\{0.5, 0.7, 0.95\}$ for ActivityNet-1.3.

Feature extraction. Firstly, the agent tubes are constructed by detecting scene objects using a YOLOv5 detector [192] pre-trained on the COCO dataset. Detections are then tracked throughout a snippet using DeepSort [193]. Then, features are extracted using an I3D network pre-trained on the Kinetics dataset [34], from both the entire snippet and each cropped agent tube individually. The object categories were reduced to six for the ROAD dataset to only cover the agents actually present in the road scenes portrayed there. As the other two datasets (ActivityNet-1.3 and Thumos-14) are general purpose, in their case we retained all the 80 classes present in the COCO dataset.

Scene Graph Attention. The local scene graph was generated by producing a list of tuples $[(0, 1), (0, 2), \dots]$, where the first index relates to the source node and the second number indexes the target node. The reason for preferring this structure over an adjacency matrix was to limit memory usage. For node attention learning, we initialised our architecture using the weights of the GAT model [94] pre-trained on the PPI dataset [194] and applied 4 attention layers with $\{4, 4, \text{num of classes}, \text{num of classes}\}$ heads, respectively. The number of classes *num of classes* is equal to 201, 21, and 7 in ActivityNet-1.3, Thumos-14 and ROAD, respectively.

The **Temporal Graph** is a stack of three 1D convolution layers on the final representation of the temporally connected local scenes. The size of the input to the first convolutional layer is the number N of local scenes (snippets), multiplied by the number of heads in the last attention layer.

The length of our temporal graph is fixed to N . Videos with number of snippets less than or equal to N are passed directly to the temporal graph; longer videos are split into multiple chunks containing N snippets each. The output is a one-hot vector of activity labels of size N and a collection of 128 proposals (binary graph masks) also of length N .

Loss Functions. Our problem is multi-objective, as we aim at not only recognising the label of the activity taking place but also finding its boundary (start and end time). Our overall loss function is thus the weighted sum of *BCEWithLogitsLoss* [191] (for activity classification) and standard binary cross entropy (for tem-

poral localisation). Full details can be found in the Section 5.2.4.

5.3.3 Comparison with State-of-the-Art

ROAD. To validate our CompAD approach, we first compared ourselves with the prior art available for the ROAD dataset *and* re-implemented the current state-of-the-art TAL methods; TallFormer [37] and ActionFormer [195] there (see Table 5.1). The standard IoU thresholds for mAP computation on ROAD range from 0.1 to 0.5. The detailed class-wise average precision of our methods is visualised in Fig. 5.7.

Table 5.1: Comparing our approach with the state-of-the-art methods for CompAD available on the ROAD dataset. The mAP at the various standard thresholds is reported. Best results are in **bold** and second best underlined.

Methods	ROAD					
	0.1	0.2	0.3	0.4	0.5	Avg
P-GCN [21]	60.0	56.7	53.9	50.5	46.4	53.5
G-TAD [25]	62.1	59.8	55.6	52.2	48.7	55.6
STDSG [17]	77.3	74.6	71.2	<u>66.7</u>	<u>59.4</u>	<u>69.8</u>
TallFormer [37]	<u>78.4</u>	<u>74.9</u>	70.3	63.8	57.1	68.9
ActionFormer [195]	76.5	73.7	72.6	64.4	58.2	69.0
Ours	82.1	77.4	73.3	69.5	62.9	73.0

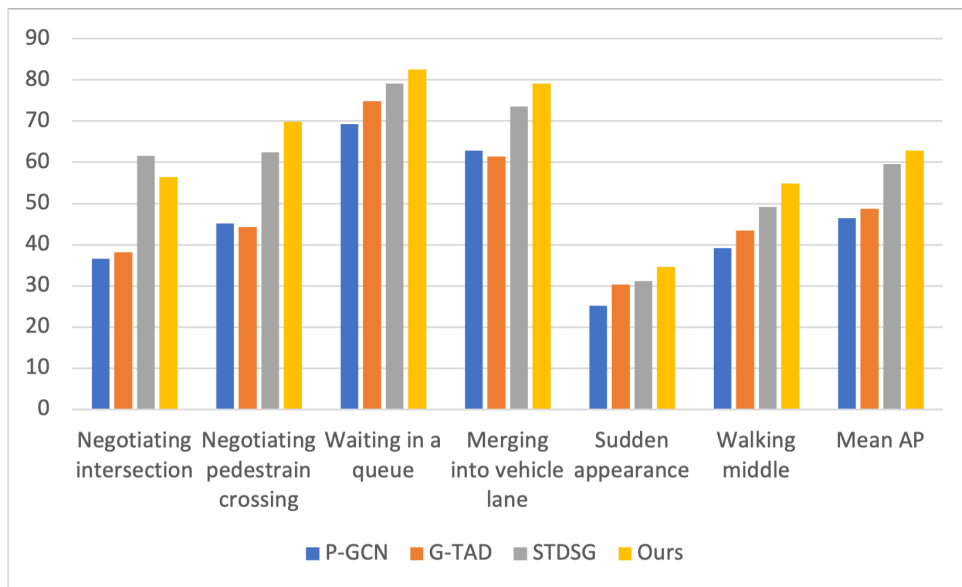


Figure 5.7: Class-wise average precision (%) with IoU threshold of 0.5 for ROAD dataset.

Thumos-14. Our hybrid graph approach is compared with the state of the art on Thumos-14 in Table 5.2. We followed the standard evaluation metric there: mAP with IoU thresholds ranging from 0.3 to 0.7 and their average. Competitor

methods were published in top vision conferences in 2018-2022. The recent ActionFormer [195] approach achieves the highest average mAP 66.8. The detailed class-wise average precision of our method is visualised in Fig. 5.7.

*Table 5.2: Activity detection performance comparison with state-of-the-art methods on Thumos-14. mAP values at different IoU thresholds are reported for the test Thumos-14. Best results are in **bold** and second best underlined.*

Methods	Conference	Thumos-14					
		0.3	0.4	0.5	0.6	0.7	Average
BSN [20]	ECCV'18	53.5	45.0	36.9	28.4	20.0	36.8
P-GCN [21]	ICCV'19	63.6	57.8	49.1	—	—	—
BMN [22]	ICCV'19	56.0	47.4	38.8	29.7	20.5	38.5
GTAN [23]	CVPR'19	57.8	47.2	38.8	—	—	—
MGG [24]	CVPR'19	53.9	46.8	37.4	29.5	21.3	37.8
G-TAD [25]	CVPR'20	54.5	47.6	40.2	30.8	23.4	39.3
BU-MR [196]	ECCV'20	53.9	50.7	45.4	38.0	28.5	43.3
BC-GNN [62]	ECCV'20	57.1	49.1	40.4	31.2	23.1	40.2
BSN++ [197]	AAAI'21	59.9	49.5	41.3	31.9	22.8	41.1
TCANet [198]	CVPR'21	60.6	53.2	44.6	36.8	26.7	44.4
AFSD [81]	CVPR'21	67.3	62.4	55.5	43.7	31.1	52.0
MUSES [199]	CVPR'21	68.9	64.0	56.9	46.3	31.0	53.4
VSGN [200]	ICCV'21	66.7	60.4	52.4	41.0	30.4	50.2
ContextLoc [201]	ICCV'21	68.3	63.8	54.3	41.8	26.2	50.9
RTD-Net [202]	ICCV'21	68.3	62.3	51.9	38.8	23.7	49.0
CPN [28]	WACV'22	68.2	62.1	54.1	41.5	28.0	50.7
E2E-TAD [27]	CVPR'22	69.4	64.3	56.0	46.4	34.9	54.2
RefactorNet [26]	CVPR'22	70.7	65.4	58.6	47.0	32.1	54.8
RCL [203]	CVPR'22	70.1	62.3	52.9	42.7	30.7	51.7
LDCLR [32]	AAAI'22	72.1	<u>65.9</u>	57.0	44.2	28.5	53.5
TallFormer [37]	ECCV'22	76.1	—	<u>63.2</u>	—	34.5	59.2
STALE [204]	ECCV'22	68.9	64.1	57.1	46.7	31.2	52.9
ActionFormer [195]	ECCV'22	82.1	77.8	71.0	59.4	43.9	66.8
Ours	—	<u>78.2</u>	<u>69.5</u>	62.7	<u>50.1</u>	<u>36.9</u>	<u>59.8</u>

ActivityNet-1.3. Table 5.3 also compares our proposal with the state-of-the-art methods on ActivityNet-1.3. The mAP IoU thresholds used for comparison are, according to protocol, 0.5, 0.75 and 0.95. Our approach performs the best on average and for IoU = 0.5 and 0.95, while coming a very close second for IoU = 0.75 with RefactorNet [26], proving our hypothesis that modelling the fine-grain structure of activity is conducive to better performance even for short-term actions.

To sum up, our proposal clearly outperforms all previous approaches over the ROAD dataset, showing the potential of this approach for modelling and detecting long, complex activities. Further, our approach outperforms all competitors over ActivityNet and achieves a reasonable performance (second best) on Thumos-14,

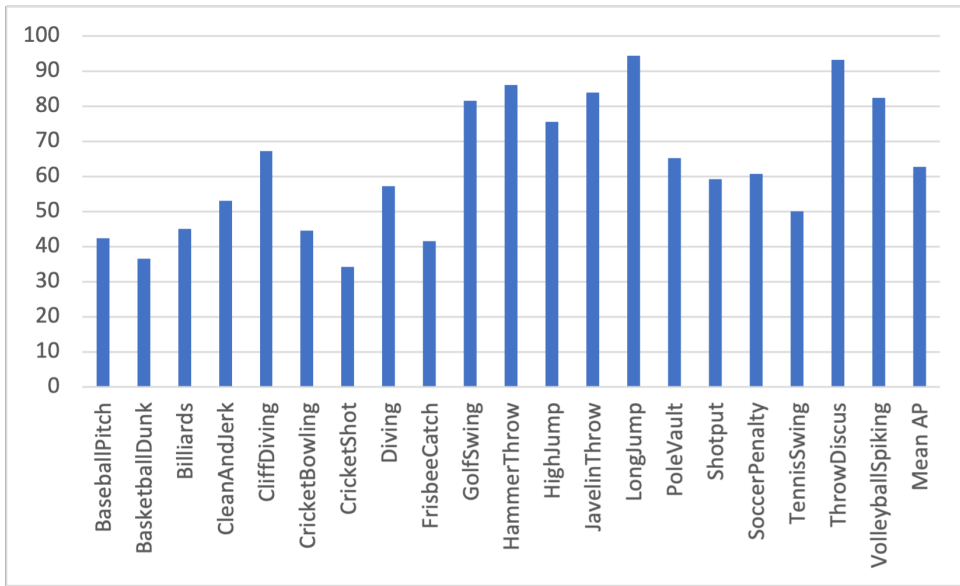


Figure 5.8: Class-wise average precision (%) with IoU threshold of 0.5 for Thumos-14 dataset.

Table 5.3: Activity detection performance comparison with state-of-the-art methods on ActivityNet-1.3. mAP values at different IoU thresholds are reported for validation sets of ActivityNet-1.3. Best results are in **bold** and second best underlined.

Methods	Conference	ActivityNet-1.3			
		0.5	0.75	0.95	Average
BSN [20]	ECCV'18	46.45	29.96	8.02	30.03
P-GCN [21]	ICCV'19	48.26	33.16	3.27	31.11
GTAN [23]	CVPR'19	52.61	34.14	8.91	34.31
G-TAD [25]	CVPR'20	50.36	34.60	9.02	34.09
BU-MR [196]	ECCV'20	43.47	33.91	9.21	30.12
BC-GNN [62]	ECCV'20	50.6	34.8	9.4	34.3
BSN++ [197]	AAAI'21	51.27	35.70	8.33	34.88
TCANet [198]	CVPR'21	51.91	34.92	7.46	34.43
AFSD [81]	CVPR'21	52.4	35.3	6.5	34.4
MUSES [199]	CVPR'21	50.02	34.97	6.57	33.99
VSGN [200]	ICCV'21	52.4	36.0	8.4	35.1
ContextLoc [201]	ICCV'21	56.01	35.19	3.55	34.23
RTD-Net [202]	ICCV'21	47.2	30.7	8.6	30.8
E2E-TAD [27]	CVPR'22	50.47	35.99	<u>10.83</u>	35.10
RefactorNet [26]	CVPR'22	56.6	40.7	7.4	<u>38.6</u>
RCL [203]	CVPR'22	55.15	39.02	8.27	37.65
LDCLR [32]	AAAI'22	<u>58.14</u>	36.30	6.16	35.24
TallFormer [37]	ECCV'22	54.1	36.2	7.9	35.6
STALE [204]	ECCV'22	56.5	36.7	9.5	36.4
ActionFormer [195]	ECCV'22	53.5	36.2	8.2	35.6
Ours	—	60.6	<u>40.3</u>	11.1	39.3

which is characterised by shorter, simpler activities. This proves that modelling the fine-grain structure of activities is conducive to better performance, even for simpler events. Fig. 5.9 illustrates how our method increasingly outperforms the prior art as the duration and complexity of the activities increases.

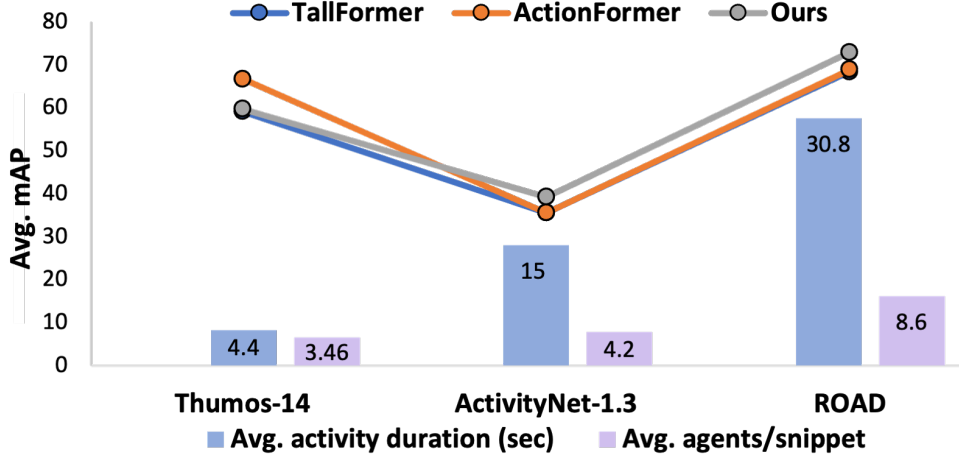


Figure 5.9: Performance comparison with the state-of-the-art over all the three datasets considered, as a function of average activity duration and average no. of agents per snippet.

5.3.4 Ablation Studies

We ran a number of ablation studies to understand the influence of the various components.

Effect of Agent Nodes. Firstly, we showed the advantage of using a scene graph modelling and connecting single agents compared to only using whole scene features. Namely, we removed the local scene graph from our pipeline (Fig. 5.1, stage B) and passed the 3D features of the whole scene as a node to the global temporal graph. The significant performance drop can be clearly observed in Fig. 5.10 over all three datasets.

Effect of Edges and Aggregation. Next, we constructed our local scene graph using three different types of edge connections. Firstly, a fully-connected scene graph where each node is connected to every other node. Secondly, a star structure where all of the agent nodes are connected to the scene node only. Lastly, a star structure with additional connections between agent nodes sharing the same label.

We also validated two different techniques for extracting the final representation from the local scene graph, which we named ‘Aggregated’ and ‘Scene’. In the former, the feature representation is extracted by aggregating those of all the attended nodes. In the latter, only the feature vector related to the scene node (after attention) is retained. The effect of the possible combinations of graph topologies and aggregation strategies is shown in Table 5.4 – a fully-connected scene graph with Aggregated

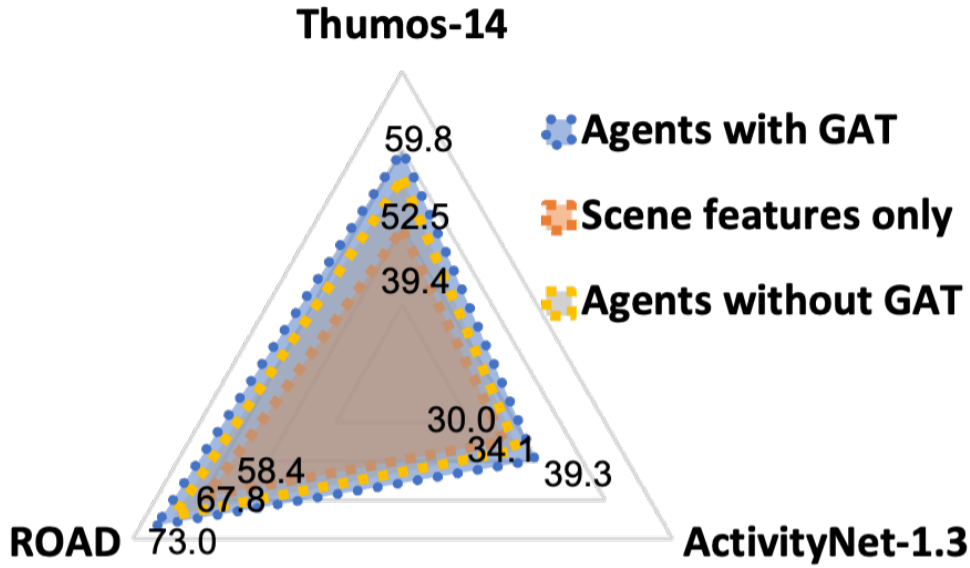


Figure 5.10: Average mAP of variants of our method with: attended scene graph ('Agents with GAT'), scene graph ('Agents without GAT') and 'Scene features only', over the all three datasets.

Table 5.4: Effect of local scene graph topologies and feature aggregation strategies on the performance of our proposal. The average mAP for the relevant IoU thresholds is reported for each dataset.

Methods	Thumos-14	ActivityNet-1.3	ROAD
Fully connected (Aggregated)	59.8	37.4	73.0
Fully connected (Scene)	49.2	31.4	62.7
Star structure (Aggregated)	51.2	39.3	62.3
Star structure (Scene)	44.8	36.6	57.9
Star + same label (Aggregated)	52.2	35.3	64.9
Star + same label (Scene)	41.9	35.3	59.2

features performs the best in two cases over three, while the star topology performed best for ActivityNet.

Effect of Sequence Length. To explore the effect on our model of snippet duration (the temporal extent of the local dynamic scene), we performed experiments with four different sizes (12, 18, 24, and 30), reported in Table 5.5. On Thumos-14 and ROAD the top scores were obtained by selecting a sequence length of 24, due to the nature of the activities present in these datasets, which last longer. On the other hand, on ActivityNet-1.3 we achieved the best performance using a sequence length of 18, as most activities there are shorter in duration.

Effect of Temporal Graph Length. In addition to the temporal span of the local scene graph, we also ablated the effect of the length of the temporal graph. By this, we mean the number of local scene nodes used in the temporal graph for recognition and localisation. Four different temporal lengths (128, 256, 512, 1024) are com-

Table 5.5: Performance of our method as a function of different snippet lengths over the three datasets. The average mAP for the relevant IoU thresholds is reported for each dataset.

Snippet length	Thumos-14	ActivityNet-1.3	ROAD
12	52.7	31.3	56.9
18	59.8	37.6	62.6
24	54.3	39.3	73.0
30	49.7	34.5	70.0

Table 5.6: Performance of our method as a function of different temporal graph lengths over the three datasets. The average mAP for the relevant IoU thresholds is reported for each dataset.

Temporal length	Thumos-14	ActivityNet-1.3	ROAD
128	59.8	29.6	48.3
256	52.9	32.7	58.5
512	50.3	39.3	69.2
1,024	46.8	34.1	73.0

pared in Table 5.6. The best performance is achieved on ActivityNet-1.3 using a shorter temporal length, due to the average duration of the videos there and its portraying shorter-term activities. In contrast, on Thumos-14 and ROAD our approach performed the best using longer temporal graphs, as activities there are typically of longer duration.

Qualitative Results. To help the reader visualise the output of our proposed method, we show some qualitative detection results on all the three datasets in Fig. 5.11 and Fig. 5.12. The figure shows one sample per dataset, and portrays a series of local scenes (snippets), skipping some for visualisation purposes, with superimposed the ground truth (in green) and the prediction of our model (in red). For example, for ActivityNet-1.3 an instance of the ‘Layup drill in basketball’ class is shown in which the activity starts with snippet 7 and ends with snippet 42. Our model predicts the activity to start from snippet 6 and end with snippet 43.

5.4 Summary of the chapter

5.4.1 Future Work

In the future we intend to progress further from incremental inference to incremental training, by learning to constructing activity graphs in an incremental manner, thus opening the way to applications such as future activity anticipation [205] and pedestrian intent prediction [206], while overcoming the need for a fixed-size snippet rep-

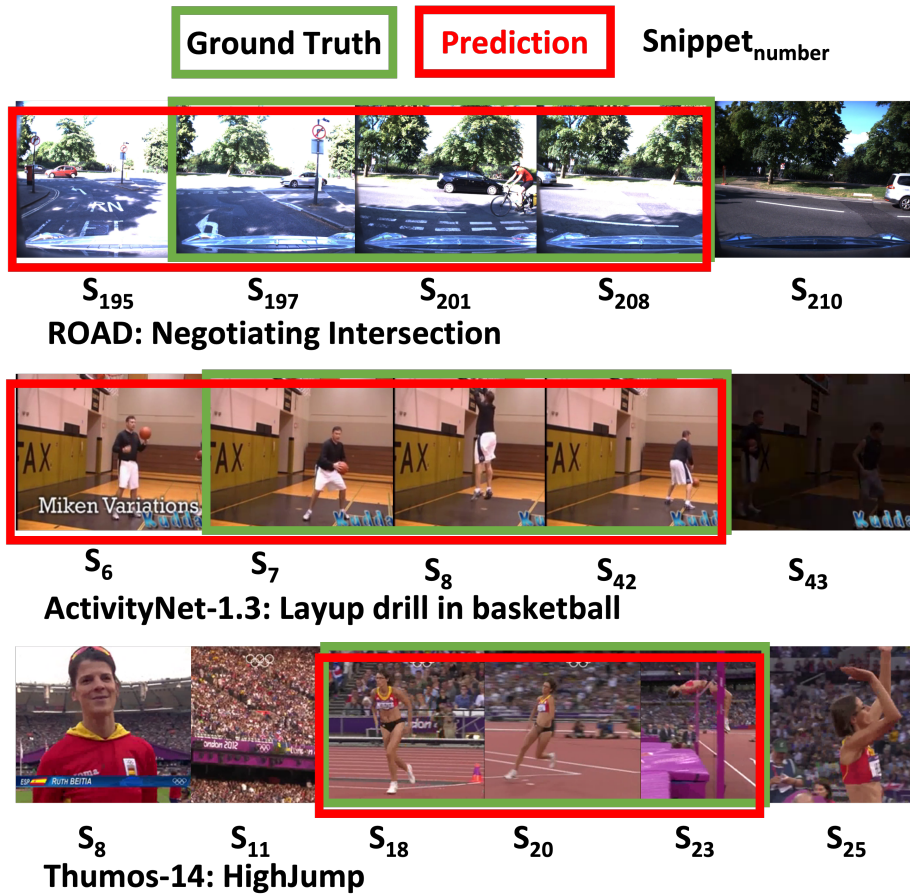


Figure 5.11: The qualitative results of our proposed method for all three datasets. The green rectangles covering the snippets (local scenes) are the ground truth while the yellow boxes show the prediction of our model.

resentation. A further exciting line of research is the modelling of the uncertainty associated with complex scenes, in either the Bayesian [207] or the full epistemic setting [208, 209].

5.4.2 Conclusion

Whereas most existing work focuses on short-term actions, this chapter explicitly addressed the problem of detecting longer-term, complex activities using a novel hybrid graph neural network-based framework which combines both scene graph attention and a temporal graph to model activities of arbitrary duration. Our proposed framework is divided into three main building blocks: agent tube detection and feature extraction; a local scene graph construction with attention; and a temporal graph for recognising the class label and localising each activity instance. We tested our method on three benchmark datasets, showing the effectiveness of our method in detecting both short-term and long-term activities, thanks to its ability to model their finer-grained structure without the need for extra annotation.



Figure 5.12: The qualitative results of our proposed method for all three datasets. The green rectangles covering the snippets (local scenes) are the ground truth while the yellow boxes show the prediction of our model.

Part III : Continual Learning

Chapter 6

Continual Semi-Supervised Learning

6.1 Introduction

The last part of the thesis is about continual learning, while interesting works have been recently directed at continual learning from streaming data in a fully supervised setting, especially focusing on avoiding catastrophic forgetting, continual learning in a semi-supervised setting remains a wide-open research question. Thus in this chapter, we present a new continual semi-supervised learning (CSSL) paradigm, proposed to the attention of the machine learning community via the IJCAI 2021 International Workshop on Continual Semi-Supervised Learning (CSSL@IJCAI)¹, with the aim of raising the field’s awareness about this problem and mobilising its effort in this direction. After a formal definition of continual semi-supervised learning and the appropriate training and testing protocols, the chapter introduces two new benchmarks specifically designed to assess CSSL on two important computer vision tasks: activity recognition and crowd counting. We describe the Continual Activity Recognition (CAR) and Continual Crowd Counting (CCC) challenges built upon those benchmarks, the baseline models proposed for the challenges, and describe a simple CSSL baseline which consists in applying batch self-training in temporal sessions, for a limited number of rounds. The results show that learning from unlabelled data streams is extremely challenging, and stimulate the search for methods that can encode the dynamics of the data stream.

In addition to the challenge baseline, we also propose to formulate the continual semi-supervised learning problem as a latent-variable one, in which the labels of a series of unlabelled, streaming datapoints play the role of the hidden states, while the instances play that of the observations. In any realistic setting the process generating instances and labels, modelled by a probability distribution, will vary in time. The aim is for the model to be updated in such a way as to adapt to a (slowly) shifting

¹<https://sites.google.com/view/sscl-workshop-ijcai-2021/>

target domain, without artificial subdivisions into tasks. To accommodate for this, latent variable temporal graphical models whose parameters also vary in time need to be considered. In particular, we thoroughly analyse the proposed framework in a classification setting, for the case in which the data stream is modelled by a time-varying hidden Markov model.

6.1.1 Motivation

The continual learning problem has been recently the object of much attention in the machine learning community, with a focus on preventing the model updated in the light of new data from ‘catastrophic forgetting’ its initial abilities. A typical example is that of an object detector which needs to be extended to include classes not originally in its list (e.g., ‘donkey’) while retaining its ability to correctly detect, say, a ‘horse’. The unspoken assumption there is that we are quite satisfied with the model we have, and merely wish to extend its capabilities to new settings and classes.

This way of posing the continual learning problem, however, is in rather stark contrast with widespread real-world situations in which an initial model is trained using limited data, only for it to then be deployed without any additional supervision. Think of a person detector used for surveillance purposes on a busy street. Even after having been trained extensively on public datasets, experience shows that its performance in its target setting will suboptimal. In this scenario, the goal is for the model to be incrementally updated using the new (unlabelled) data, in order to adapt to its target domain.

In such settings, the goal is quite the opposite of that of the classical scenario described above: the initial model is usually poor and not a good match for the target domain. To complicate things further, the target domain itself may change over time, both periodically (e.g., night/day cycles) and in asynchronous, discrete steps (e.g., when a new bank opens within the camera’s field of view or a new building is erected with a new entrance).

6.1.2 Overview of the CSSL Challenge

In the first part of this chapter, we formalise this problem as one of *continual semi-supervised learning*, in which an initial training batch of labelled data points is available and can be used to train an initial model, but then the model is incrementally updated exploiting the information provided by a time series of unlabelled data points, each of which is generated by a data generating process (modelled by a probability distribution) which *varies with time*. Our formulation of CSSL is similar to *online domain adaptation* (ODA). Both CSSL and ODA involve adaptation to changing

data distributions, they have distinct focuses and characteristics. The main difference between ODA and our CSSL is that ODA primarily deals with adapting a model to changes in the data distribution across different domains (tasks). This is particularly relevant in applications where the target domain is dynamic and constantly evolving [210–212]. CSSL, on the other hand, focuses on the scenario where a model incrementally learns from both labeled and unlabeled data over time. The key challenge is to continually adapt to new data without catastrophic forgetting of previously learned knowledge. While this may involve adapting to some degree of changing data distributions, the primary emphasis is on incremental learning and avoiding model degradation over time. In CSSL, we do not assume any artificial subdivision into ‘tasks’, but allow the data-generating distribution to be an arbitrary function of time.

While batch semi-supervised learning (SSL) has seen renewed interest in recent times, thanks to relevant work in the rising field of unsupervised domain adaptation [213] and various approaches based on classical SSL self-training [214, 215], where the dominant trend is based on adversarial approaches to reduce the discrepancy between source and target domain features [213]. A panoply of deep self-training approaches have been proposed, inspired by classical SSL self-training [214], which alternate between generating pseudo-labels for the unlabelled data-points and re-training a deep classifier [215]. In opposition, continual SSL is still a rather unexplored field. The reason is that, while in the supervised case it is clear what information the streaming data points carry, in the semi-supervised case it is far from obvious what relevant information carried by the streaming instance should drive model update.

6.1.3 Overview of Hidden Markov Model as Latent Variable

In response to the CSSL challenge, we propose to model the stream of instances and ground truth values as a (temporal) *latent variable model*, in which the instances play the role of the observations, and the targets that of the hidden variables. To accommodate for data distributions that shift in time, this latent variable model needs to be time-variant, i.e., its parameters must be a function of time. In this chapter we investigate, in particular, the case in which the latent variable model belongs to the class of *time-varying hidden Markov models* (T-HMMs) [105, 108], although deep generative models can also be envisaged. We focus on *classification* tasks, applied to the *continual activity recognition* problem setting defined in 6.1.2.

Our approach is two-step. For each time instant, t , in the first stage, the parameters of the latent model of the data stream are updated incrementally, together with the sequence of state (label) estimates up to the present time instant. In the case of T-HMMs, the parameters are: i) the time-varying matrix of transition probabilities

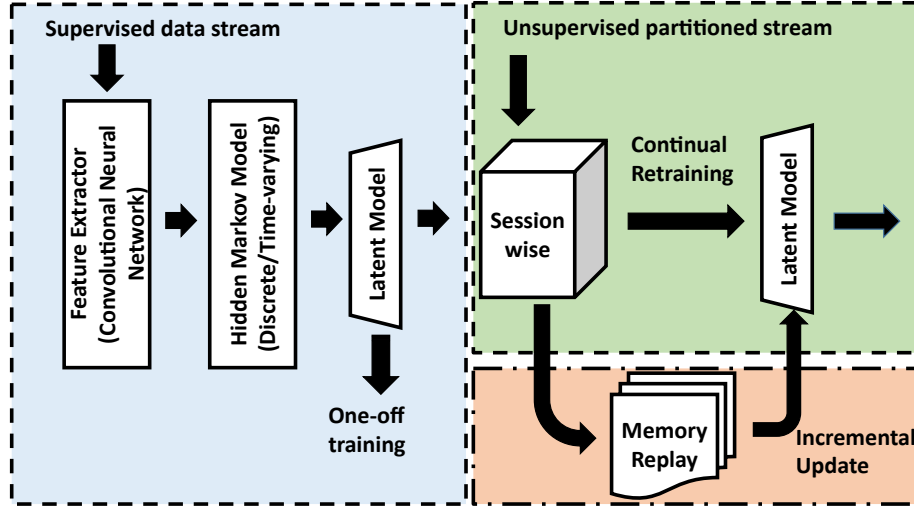


Figure 6.1: Overview of the proposed latent variable pipeline.

between labels and ii) the time-varying parameters of the emission probability densities (typically, Gaussian) for generating observations given the current state (label). Such a T-HMM latent variable model can be suitably initialised using the information provided by the initial, supervised training set. Based on the obtained sequence of most likely (pseudo-)label estimates², a suitable continual supervised learning approach (in our case, memory replay with random sampling [97]) can be applied to incrementally update the model. Figure 6.1 shows the overall pipeline of the proposed framework.

6.1.4 Main Contributions

Summarising, the main contributions of this chapter are:

1. A formal definition of the continual semi-supervised learning problem, and the associated training and testing protocols.
2. The first two benchmark datasets for the validation of semi-supervised continual learning approaches, one for classification (continual activity recognition, CAR) and one for regression (continual crowd counting, CCC), which we propose as the foundations of the first challenges in this domain.
3. Results produced by a simple strategy in which self-training is applied in sessions to competitive baseline models on both benchmarks, as the foundations for the above challenges.
4. A principled two-stage approach to the problem based on a latent model formulation reducing the task to continual supervised learning.

²In fact, an HMM outputs a series of probability vectors $P(y_t | x_1, \dots, x_t)$, where y_t is the current label and x_τ denotes the observation at time τ . See Conclusions.

5. An efficient solution based on a time-varying hidden Markov models, able to model streaming data issued from a time-shifting target data distribution.
6. State-of-the-art results produced by our approach on the recently released Continual Activity Recognition (CAR) dataset, as well as the 50Salads dataset adapted to continual learning thanks to a new, original protocol.

Related publications: The work of this chapter appeared in International Workshop on Continual Semi-Supervised Learning [47], Continual Semi-Supervised Learning Workshop and Challenge at IJCAI 2021³, and under preparation IEEE transactions on pattern analysis and machine intelligence (IEEE TPAMI). The dissertation's author was co-organiser of the workshop and challenge and co-author of the IEEE TPAMI paper. The main contributions of the dissertation's author to this project are: leading the effort of continual activity recognition (continual semi-supervised classification), HMM latent baseline implementation, and CAR baseline for challenge. All the resources of this project are publically available and are accessed from *Code*⁴, *CAR challenge baseline*⁵, *CAR evalAI*⁶, *CCC challenge baseline*⁷, and *CCC evalAI*⁸.

Outline: The structure of the chapter is as follows. Firstly, we explained Continual Semi-Supervised Learning Workshop and Challenge in Section 6.2 including learning (Section 6.2.1) and testing (Section 6.2.2) protocols. The CSSL workshop and challenge also include benchmarks (Section 6.2.3), challenges (Section 6.2.4), and baselines (Sections 6.2.6. Secondly, we then explained HMM latent variable model in Section 6.3. The latent variable model is then divided into principles (Section 6.3.1), learning process (Section 6.3.2), distribution of data stream (Section 6.3.3, discrete-state T-HMM for continuous classification (Section 6.3.4), and memory replay (Section 6.3.5). Next, the experimental results of CSSL workshop and challenge and latent variable are given in Section 6.4 and Section 6.5, respectively. Finally, in Section 6.6 we provide the summary of the chapter.

³<https://sites.google.com/view/sssl-workshop-ijcai-2021/>

⁴<https://github.com/salmank255/cssl>

⁵<https://github.com/salmank255/IJCAI-2021-Continual-Activity-Recognition-Challenge>

⁶<https://eval.ai/web/challenges/challenge-page/984/overview>

⁷https://github.com/Ajmal70/IJCAI_2021_Continual_Crowd_Counting_Challenge

⁸<https://eval.ai/web/challenges/challenge-page/986/overview>

6.2 Continual Semi-Supervised Learning Workshop and Challenge

We begin by formalising the continual semi-supervised learning problem, and by defining the appropriate training and testing protocols for this new learning setting.

6.2.1 Learning

The *continual semi-supervised learning* problem can be formulated as follows.

Given an initial batch of supervised training examples,

$$\mathcal{T}_0 = \{(x^j, y^j), j = 1, \dots, N\},$$

and a stream of unsupervised examples,

$$\mathcal{T}_t = \{x_1, \dots, x_t, \dots\},$$

we wish to learn a model $f(x|\theta)$ mapping instances $x \in \mathcal{X}$ to target values $y \in \mathcal{Y}$, depending on a vector $\theta \in \Theta$ of parameters.

When $\mathcal{Y}$ is discrete (i.e., a list of labels) the problem is one of continual semi-supervised *classification*; when $\mathcal{Y}$ is continuous we talk about continual semi-supervised *regression*.

In both cases we wish the model to be updated incrementally as the string of unsupervised samples come in: namely, based on the current instance x_t at time t , model θ_{t-1} is mapped to a new model θ_t .

6.2.2 Testing

How should such a series of models be evaluated? In continual supervised learning (see e.g. [97]), a series of ground-truth target values is available and can be exploited to update the model. To avoid overfitting, the series of models outputted by a continual supervised learner, $f(x|\theta_1), \dots, f(x|\theta_t), \dots$ cannot be tested on the same data it was trained upon. Thus, basically all continual (supervised) learning papers set aside from the start a set of instances upon which the incremental models are tested (e.g., in the Core50 dataset the last three sessions are designated as test fold [216]).

In principle, however, as each model $f(x|\theta_t)$ is estimated at time t it should also be tested on data available at time t , possibly coming from a parallel (test) data stream. An idealised such scenario is one in which a person detector is continually trained on data coming from one surveillance camera, say in a shopping mall, and the resulting model is deployed (tested) in real time on all other cameras in the same mall.

In our continual semi-supervised setting, model update cannot make use of any ground truth target values. If the latter are somehow available, but not shown to the learner for training purposes, the performance of the learner can in fact be evaluated on the stream of true target values by testing at each time t model $f(x|\theta_t)$ on the data pair (x_t, y_t) . A reasonable choice for a performance measure is the average loss of the series of models on the *contemporary* data pair:

$$\sum_{t=1, \dots, T} l(f(x_t|\theta_t), y_t). \quad (6.1)$$

In our experiments, a standard 0/1 loss, which translates into classical accuracy, is used for classification tasks.

6.2.3 Benchmarks

To empirically validate CSSL approaches in a computer vision setting we created two benchmark datasets, one designed to test continual classification (CAR), and one for continual regression (CCC).

Continual activity recognition (CAR) dataset

The MEVA dataset

To allow the validation of continual semi-supervised learning approaches in a realistic classification task we created a new *continual activity recognition* (CAR) dataset derived from the very recently released MEVA (Multiview Extended Video with Activities) dataset [217].

MEVA is part of the EctEV (Activities in Extended Video) challenge⁹. As of December 2019, 328 hours of ground-camera data and 4.2 hours of Unmanned Aerial Vehicle video had been released, broken down into 4304 video clips, each 5 minutes long. These videos were captured at 19 sites (e.g. School, Bus station, Hospital) of the Muscatatuck Urban Training Center (MUTC), using a team of over 100 actors performing in various scenarios. There are annotations for 22.1 hours (266 five-minute-long video clips) of data. The original annotations are available on GitLab¹⁰. Each video frame is annotated in terms of 37 different activity classes relevant to video surveillance (e.g. *person_opens_facility_door*, *person_reads_document*, *vehicle_picks_up_person*). Each activity is annotated in terms of class labels and bounding boxes around the activity of interest. Whenever activities relate to objects or other persons (e.g., in *person_loads_vehicle* the person usually puts an object into

⁹<https://actev.nist.gov/>

¹⁰<https://gitlab.kitware.com/meva/meva-data-repo/tree/master/annotation/DIVA-phase-2/MEVA/>

the vehicle’s trunk; in *person_talks_to_person* a number of people listen to the person speaking), these object(s) or people are also identified by a bounding box, to allow human-object interaction analyses.

The CAR dataset

The original MEVA dataset comes with a number of issues, from our standpoint: (i) multiple activity classes can take place simultaneously, whereas in our formulation (at least for now) only one label can be emitted at any given time instant; (ii) the quality of the original annotation is uneven, with entire instances of activities missing. As our tests are about classification, we can neglect the bounding box information.

For these reasons we set about creating our own continual activity recognition (CAR) dataset by selecting 45 video clips from MEVA and generating from scratch a modified set of annotations spanning a reduced set of 8 activity classes (e.g. *person_enters_scene_through_structure*, *person_exits_vehicle*, *vehicle_starts*). Those 8 classes have been suitably selected from the original 37 to ensure that activity instances from different classes do not temporally overlap, so that we can assign a single label to each frame. Each instance of activity is annotated with the related start and end frame. Frames that do not contain any relevant activity label are assigned to a ‘background’ class. The goal is to classify the activity label of each input video frame.

For some MEVA sites, contiguous annotated videos exist with no gap between the end of the first video and the start of the second video. For some sites, ‘almost’ contiguous videos separated by short (5 or 15 min) gaps are available. Finally, videos from a same site separated by hours or days exist. Accordingly, our CAR benchmark is composed of 15 sequences, broken down into three groups:

1. Five 15-minute-long sequences from sites G326, G331, G341, G420, and G638 formed by three original videos which are contiguous.
2. Five 15-minute-long sequences from sites G329, G341, G420, G421, G638 formed by three videos separated by a short gap (5-20 minutes).
3. Five 15-minute-long sequences from sites G420, G421, G424, G506, and G638 formed by three original videos separated by a long gap (hours or days).

Each of these three evaluation settings is intended to simulate a different mix of continuous and discrete domain dynamics.

The CAR dataset including annotation and scripts is available on GitHub¹¹.

¹¹<https://github.com/salmank255/IJCAI-2021-Continual-Activity-Recognition-Challenge>

The continual crowd counting (CCC) dataset

Crowd counting is the problem of, given a video frame, counting the number of people present in the frame. While intrinsically a classification problem, crowd counting can be posed as a regression problem by manually providing for each training frame a density map [218]. To date, crowd counting is mostly considered an image-based task, performed on single video frames. Few attempts have been made to extend the problem to the video domain [219], [220], [221].

To the best of our knowledge, continual crowd counting has never been posed as a problem, not even in the fully supervised context – thus, there are no standard benchmarks one can adopt in this domain. For this reason we set about assembling the first benchmark dataset for *continual crowd counting* (CCC). Our CCC dataset is composed by 3 sequences, taken from existing crowd counting datasets:

1. A single 2,000 frame sequence originally from the Mall dataset¹² [222].
2. A single 2,000-frame sequence originally from the UCSD dataset¹³ [223].
3. A 750-frame sequence from the Fudan-ShanghaiTech (FDST) dataset¹⁴, composed by 5 clips, 150 frames long, portraying a same scene [224].

The ground truth for the CCC sequences (in the form of a density map for each frame) was generated by us for all three datasets following a standard annotation protocol¹⁵.

The CCC dataset complete with scripts for download and generating the baseline results is available on GitHub¹⁶.

6.2.4 Challenges

CAR challenge

From our problem definition (Sec. 6.2), once a model is fine-tuned on the supervised portion of a data stream it is then both incrementally updated using the unlabelled portion of the same data stream and tested there, using the provided ground truth. Incremental training and testing are supposed to happen independently for each sequence, as the aim is to simulate real-world scenarios in which a smart device with continual learning capability can only learn from its own data stream. However, as CAR sequences do not contain instances of all 9 activities, the initial supervised training is run there on the union of the supervised folds for each sequence.

¹²<https://www.kaggle.com/c/counting-people-in-a-mall>

¹³<http://www.svcl.ucsd.edu/projects/peoplecnt/>

¹⁴<https://drive.google.com/drive/folders/19c2X529VTNj13YL1EYweBg60G70G2D-w>

¹⁵<https://github.com/svishwa/crowdcount-mcnn>

¹⁶https://github.com/Ajmal70/IJCAI_2021_Continual_Crowd_Counting_Challenge

Split. Each data stream (sequence) in our benchmark(s) is divided into a supervised fold (S), a validation fold (V) and a test fold (T), the last two unsupervised. Specifically, in CAR given a sequence the video frames from the first five minutes ($5 \times 60 \times 25 = 7,500$ samples) are selected to form the initial supervised training set S . The second clip (another 5 minutes) is provided as validation fold to tune the CSSL strategy. The remaining clip (5 minutes) is used as test fold for testing the performance of the (incrementally updated) classifier.

Evaluation. Performance is evaluated as the average performance of the incrementally updated classifier over the test fold for all the 15 sequences. More specifically, we evaluate the average accuracy (percentage of correctly classified frames) over the test folds of the 15 sequences. Remember that, however, in our setting each test frame is classified by the *current* model available at time t *for that specific sequence*. This distinguishes our evaluation setting from classical ones in which all samples at test time are processed by the same model, whose accuracy is then assessed.

CCC challenge

Unlike the CAR challenge, in the crowd counting case (as this is a regression problem) each sequence is treated completely independently.

Split. For the CCC challenge we distinguish two cases. For the 2,000-frame sequences from either the UCSD or the Mall dataset, S is formed by the first 400 images, V by the following 800 images, and T by the remaining 800 images. For the 750-frame sequence from the FDST dataset, S is the set of the first 150 images, V the set of the following 300 images, and T the set of remaining 300 images.

Evaluation. For continual crowd counting, MAE (Mean Absolute Error) is adopted (as standard in the field) to measure performance. MAE is calculated using predicted and ground truth density maps in a regression setting.

6.2.5 Tasks

For our IJCAI challenge we therefore set four different validation experiments (*tasks*).

1. In the first task (*CAR-absolute*) the goal is to achieve the best average performance across all the (test folds of the) 15 sequences in the CAR dataset. The choice of the baseline action recognition model is left to the participants.
2. In the second task (*CAR-incremental*) the goal is to achieve the best performance differential between the model as updated through the chosen CSSL strategy and the original model fine-tuned over the supervised fold. We thus evaluate the difference between the average performance of the incrementally

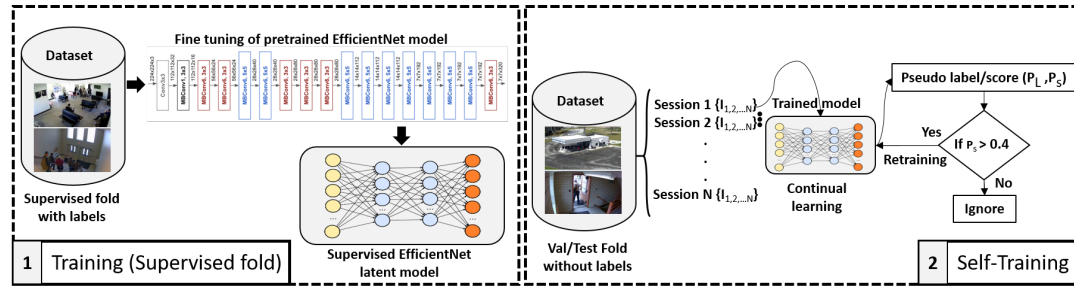


Figure 6.2: Overall pipeline of our CAR baseline. 1) Firstly, a pre-trained EfficientNet model is fine-tuned over the supervised fold of the dataset (jointly over the 15 sequences). 2) The unlabeled validation and test folds are divided into subfolds. In each session video frames from each sub-fold are used in a self-training cycle in which a pseudo label P_L with prediction probability score P_s is generated for each frame, but only the pseudolabels with P_s above a fixed threshold (namely, 0.4) are considered as ground truth for retraining the model. Frames with lower P_s are ignored. The model updated by self-training in session n is used as initial model in session $n + 1$.

updated model on the test fold and the average performance of initial model, also on the test fold. The baseline recognition model is set by us (see Baselines).

3. Task *CCC-absolute* seeks the best average performance over the test fold of the 3 sequences of the CCC dataset. The choice of the baseline crowd counting model is left to the participants to encourage them to push its performance to the limit.
4. Finally, task *CCC-incremental* seeks the best performance differential between the initial and the updated model over the test fold, averaged across the three sequences. The baseline crowd counting model is set.

6.2.6 Baselines

To provide a baseline for the above tasks, and assess the degree of difficulty of the challenges, we decided to adopt a simple strategy which consists of classical (batch) semi-supervised self-training *performed in a series of sessions*, for both the activity and the crowd counting experiments. Baselines are provided regarding the initial action recognition model to be used in the CAR tests, the base crowd counter to be used in the CCC tests, and the semi-supervised incremental learning process itself.

Baseline activity recognition model

The baseline model is the recent EfficientNet network (model EfficientNet-B5) [225], pre-trained on the large-scale ImageNet dataset over 1000 classes. For our tests we initialised the model using those weights and changed the number of classes to 9

activities (see Section 6.2.3). Detailed information about its implementation, along with pre-trained models, can be found on Github¹⁷. The model is easily downloadable using the Python command “pip” (*pip install efficientnet-pytorch*).

Baseline crowd counter

For the baseline crowd counting model, we selected the Multi-Column Convolutional Neural Network (MCNN) [226], whose (unofficial) implementation is publicly available¹⁸ and uses PyTorch. MCNN was considered state-of-the-art when released in 2016, but is still commonly used as a standard baseline in more recent works, due to its competitive performance on public datasets (e.g., [227, 228]). MCNN made significant contributions to crowd counting by proposing a network architecture better-equipped to deal with differing head sizes due to image resolution, perspective effects, distances between the camera and the people within the scene, etc. In a nutshell, this robustness to scale was achieved via a network composed of three parallel CNN columns, each of which used filters with different receptive field sizes, allowing each column to specialize for a particular scale of human head.

Pre-trained MCNN models are available for both the ShanghaiTech A and the ShanghaiTech B datasets. For our tests, as well as the Challenge, we chose to adopt the ShanghaiTechB pre-trained model.

Baseline incremental learning strategy

As mentioned, our baseline for incremental learning from unlabelled data stream is instead based on a *vanilla self-training* approach [229]. For each sequence, the unlabelled data stream (without distinction between validation and test folds) is partitioned into a number of sub-folds. Each sub-fold spans 1 minute in the CAR challenges, so that each unlabelled sequence is split into 10 sub-folds. Sub-folds span 100 frames in the CCC challenges, so that the UCSD and MALL sequences comprise 16 sub-folds whereas the FDST sequence contains only 6 sub-folds.

Starting with the model initially fine-tuned on the supervised portion of the data stream, self-training is successively applied in a batch fashion in sessions, one session for each sub-fold, for a single epoch (as we noticed that using multiple epochs would lead to results degradation). Self-training requires to select a confidence threshold above which predictions are selected as pseudolabels. In our tests we set a confidence threshold of 0.4. The predictions generated by the model obtained after one round of self-training upon a sub-fold are stored as baseline predictions for the current sub-fold. The model updated after each self-training session is used as initial model for

¹⁷<https://github.com/lukemelas/EfficientNet-PyTorch>

¹⁸<https://github.com/svishwa/crowdcount-mcnn>

the following session.

6.3 HMM Latent Variable Model for Unsupervised Data-Streams

In the second part of the chapter, the problem is defined as the same as in Section 6.2 including *Learning* and *Testing*. However, in this part, we only consider the continual semi-supervised classification.

6.3.1 Principles of HMM Latent Variable Model

In this chapter, we wish to devise an approach to the continual semi-supervised learning problem adhering to the following principles.

1. Continual semi-supervised learning, as defined in Section 6.2, should reduce to continual supervised learning once one or more hypotheses are made about the target values of the unsupervised stream.
2. At each time step t , a constant number, M , of such hypotheses should be examined.
3. The admissible target assignments for the unlabelled data stream up to time t should be sampled according to their likelihood to be correct, assuming this information is available.
4. For each target assignment hypothesis under consideration, the model should be updated incrementally.
5. To realistically model real-world applications, a budget should be enforced on the total number of (past) example instances one has access to at any given time step, t .
6. Assuming ground truth target values $y_1, \dots, y_t, \dots$ are available, the validation of the learner should take place simultaneously with the training process, by measuring at each time instant, t , the loss of the current model, $f(x|\theta_t)$, on the current data pair, (x_t, y_t) , and taking the average over the entire data stream.

Principle 3) assumes that a likelihood model for the data stream of targets is maintained at each time step, t , so that we can sample from it. In the classification case (when $\mathcal{Y}$ is finite), a likelihood model for the admissible label assignments over

the (current) sequence of instances, $x_1, \dots, x_t$, is a finite probability distribution,¹⁹ $P_Y(y_1, \dots, y_t | \Theta_Y^t)$ over $\mathcal{Y}^t \doteq \mathcal{Y} \times \dots \times \mathcal{Y}$ (t times), conditional on a vector of parameters $\Theta_Y^t = \{\theta_Y^1, \dots, \theta_Y^t\}$. A similar likelihood model can be defined for continual regression problems.

6.3.2 A Two-Stage Learning Process

In this chapter, we propose a general approach to continual semi-supervised learning (Section 6.2) which adheres to the principles stated in Section 6.3.1. The approach is incremental, i.e., the ‘best’ model, $f(x|\theta)$, mapping instances, $x \in \mathcal{X}$, to target values, $y \in \mathcal{Y}$, is updated at each time step, t , based on the new evidence, x_t . Namely, at each time step t :

1. The most likely sequence(s) of target values $y_1, \dots, y_t$ are estimated based on the new piece of evidence (unlabelled instance) x_t .
2. Based on the current estimate of the most likely labellings, a continual supervised learning task is posed and a memory replay-based approach is used to update the model.

More specifically, in this work we propose to estimate the most likely target assignments for the current data-stream (Step 1) via a *latent variable model* of the data stream of pairs (x_t, y_t) . This is discussed in Section 6.3. Although in principle any such model can be employed, including deep generative models, we elaborate in particular on the case in which *time-varying hidden Markov models* are used to model the data and the way the data-generating distribution changes over time for a continual classification task.

While the pseudo-labels estimated by the latent variable model are used to incrementally update the model at training time, at testing time the performance of the series of models (and, as a consequence, of the overall learner) can be measured versus the series of ground-truth labels, assuming those are available, as dictated by Principle 6.

6.3.3 Joint Distribution of the Data Stream

We assume that the unsupervised data stream, up to the current time instant t , can be modelled by the joint distribution

$$p_D^{1:t}((x_1, y_1), \dots, (x_t, y_t) | \theta_D^1, \dots, \theta_D^t), \quad (6.2)$$

¹⁹In this chapter we use P (uppercase) to denote finite probability distributions, and p (lowercase) to denote probability densities.

under the additional principle that:

7. The data-generating probability distribution $p_D^t = p((x_t, y_t)|\theta_D^t)$, from which the pair (x_t, y_t) is assumed to be sampled, can be a function of time.

Any likelihood model for the target assignments (Principle 3) should thus be derived from the joint, time-varying probability distribution (6.2), possibly under sensible simplifying assumptions.

Principle 7) allows the target distribution to vary over time, reflecting material, realistic real-world conditions. This principle makes our approach more general than those which assume the existence, next to the source domain, of a single target domain, as in classical domain adaptation [230]. It is also more general than ‘task-based’ approaches to continual learning [231], in which the data are assumed to be generated in a series of discrete tasks, each associated with its own data generating process, whose boundaries typically are given.

6.3.4 Discrete-State T-HMM for Continuous Classification

The data stream of pairs is composed by two parallel series $x_1, \dots, x_t, \dots$ and $y_1, \dots, y_t, \dots$, of which only the sequence of instances is observed, while the (unknown) target values form a series of hidden variables. This amounts to representing the unsupervised data stream using a *latent variable model* in which the targets $y_1, \dots, y_t, \dots$ play the role of the hidden variables.

In this chapter, we consider *hidden Markov models* (HMMs) [232] as one specific instance of a latent variable model that is applicable to the posed problem setup. In the continual classification case, states are discrete and correspond to the labels. To cater to Principle 7), we seek to learn a *time-varying* HMM from the data stream which can model a data distribution that shifts with time.

Model formulation

In a discrete-state, continuous observation time-varying hidden Markov model [105, 108, 233] the joint distribution (6.2) assumes the form:

$$p_D^{1:t}((x_1, y_1), \dots, (x_t, y_t)|\theta_D^1, \dots, \theta_D^t) = P(y_1|\theta_D^1)p(x_1|y_1, \theta_D^1) \cdot \prod_{\tau=2}^t P(y_\tau|y_{\tau-1}, \theta_D^{\tau-1})p(x_\tau|y_\tau, \theta_D^\tau). \quad (6.3)$$

The model parameters $\theta_D^\tau = \{\theta^0, \theta_T^\tau, \theta_F^\tau\}$ are *time-varying*, to model the fact that, according to Principle 7), the (target) data distribution varies in general with time. These parameters describe²⁰: (i) the (discrete) probability of the initial state, $P(y_1|\theta^0)$,

²⁰The notation used for HMM parameters follows that in [232].

(ii) the discrete, time-dependent transition probability for the hidden variables (in this case, the labels), $A^\tau \doteq P^\tau(y_{\tau+1}|y_\tau) = P(y_{\tau+1}|y_\tau, \theta_T^\tau)$, where $\tau \in \{1, \dots, t\}$ denotes any time instant in the current data-stream, and (iii) the conditional density of the (continuous) observations (instances), given the state (label), $\Gamma^\tau(x_\tau|y_\tau) = p(x_\tau|y_\tau, \theta_\Gamma^\tau)$.

Under this model, the structure of the time-dependent data probability distribution is

$$p_D^t((x_t, y_t)|\theta_D^t) = p(x_t|y_t, \theta_\Gamma^t) \cdot P(y_t), \quad (6.4)$$

where

$$P(y_t) = P(y_t|y_{t-1}, \theta_T^{t-1}) \cdot \dots \cdot P(y_2|y_1, \theta_T^1) \cdot P(y_1|\theta^0). \quad (6.5)$$

In other words, the probability distribution generating the data at time t is parameterised by: the parameters, θ_Γ^t , that determine the distribution of the instances given the label at time t , those that determine the series of transitions $1 \rightarrow 2 \rightarrow \dots \rightarrow t$ of the Markov chain of labels, and the probability of the initial label. This formulation constrains the way the target domain distribution $p((x, y)|\theta_D)$ can vary in time, while still allowing for a rather general process model.

Learning the Parameters of a Discrete-State T-HMM

Modelling time dependency with covariates. While the above formulation is completely general, the mainstream approach to T-HMM is to model the dependency of the parameters on time (and the transition probabilities, in particular) using a multinomial logistic model [234] of the kind

$$A^\tau(i, j) = P(y_\tau = i|y_{\tau-1} = j) \doteq \frac{\exp(g(z_\tau, i, j))}{1 + \sum_{h=1}^{N-1} \exp(g(z_\tau, i, h))}, \quad (6.6)$$

where N is the number of states, $i, j \in \{1, \dots, N\}$, and $g(z_\tau, i, j)$ is a function of a *covariate* process $\{z_\tau, \tau = 1, \dots\}$, which is also *observable*. As each entry of the transition matrix is a function of the covariate variables, their values drive the dynamics of the transition probabilities.

The transition matrix can then be expressed as

$$A^\tau = P(y_{\tau+1}|y_\tau, \theta_T^\tau) = P(y_{\tau+1}|y_\tau, \theta_L, z_\tau),$$

where θ_L is the vector of parameters of the logistic model (6.6) (in particular, of the function g), and z_τ is the covariate vector at time τ . Covariates on the initial probability can also be accounted for, in the form of a single vector, z_0 .

Covariates as functions of time. The shape of the function, g , is a design choice. A common option is $g(z_\tau, i, j) = \phi_{ij} + z_\tau \vartheta_{ij}$, where ϕ_{ij} and ϑ_{ij} are unknown coef-

ficients, which are unique to the transition between state i and state j . The resulting transition matrix has $|\theta_L| = N \cdot (N - 1)$ parameters, twice those of a traditional, time-invariant transition matrix.

In many practical applications, time and date stamps can act as effective covariates. For instance, in surveillance settings, the time stamp can induce a suitable periodic behaviour in a system, e.g. to detect people entering a building. The date stamp can induce annual cycles, or facilitate the gradual shift of the parameter values over time. As shown in [103], a range of transition models can be designed by plugging in a different function of time (intended as the main covariate quantity) as the argument $g(z_\tau, i, j)$ of the exponential in (6.6).

Parameter estimation via Expectation Maximisation. Just as in traditional HMMs, the Expectation-Maximisation (EM) algorithm [235] can be used to recursively estimate the parameters of a T-HMM in the logistic dependency from covariates formulation, as well as to estimate the most likely series of states [236], given one or more training sequences of observations: $\{(x_1^m, \dots, x_{T_m}^m), m = 1, \dots, M\}$.

In the T-HMM case, the parameter estimation algorithm takes as input, in addition to each training sequence of observations $x_1, \dots, x_T$, a parallel sequence of covariate vectors, $z_1, \dots, z_T$, and outputs both the parameters θ_L of the logistic model (6.6) and those of the emission probabilities. EM is well-known to be prone to converge to local, as opposed to global, minima of the log-likelihood of the data. In our experiments, EM is run multiple times on the same sequence in order to extract ‘consensus’ parameters.

Initialisation via the supervised fold. While the initial probabilities of the hidden variables (labels) can be estimated via EM with covariates in the same framework explained above for the transition matrix, for our problem it makes more sense to estimate them using the information provided by the initial, labelled training set $\mathcal{T}_0$, sampled from a ‘source’ data distribution $p((x, y)|\theta_D^0)$. In fact, from $\mathcal{T}_0$ we can extract the empirical marginal of the labels, $\hat{P}(y|\theta_D^0)$, by counting and normalising the occurrences of the labels in the initial training data, and set:

$$P(y_1|\theta^0) = \hat{P}(y|\theta_D^0). \quad (6.7)$$

Covariates in the emission probabilities. To the best of our knowledge, the extension of the dependence on covariates to the emission probabilities $p(x_\tau|y_\tau, \theta_\Gamma^\tau)$ has not been discussed in the literature. Nevertheless, this appears to be conceptually straightforward, for instance by plugging a logistic dependency of the mean vector μ^τ and covariance matrix Σ^τ in the classical multivariate Gaussian approach from a sequence of covariates (e.g., time or functions thereof). We will address this extension in future work.

Discrete state HMM vs. continuous state space The choice of using labels as the state space in HMMs is often a strong and meaningful solution for specific tasks. For example, in our case, the primary objective is to train the model on the sequences that are already available and are inherently discrete in nature (train set). We are not expecting a change in the number of states as each state represents a specific class and these classes will remain the same during the training and prediction. In contrast, using a continuous state space related to physics or perception may introduce complexity that is unnecessary for tasks focused on discrete events. This complexity can hinder model interpretability and complicate the mapping between states and meaningful labels. The model with continuous state space would be useful in other continual learning settings such as class-incremental learning, where the objective is to train the model on new classes.

Protocols for training the latent model

In the incremental learning setting described in Section 6.3, the desired model (in this case, the classifier) is incrementally updated at each time instant t using the unlabelled observation x_t , and tested on the ground-truth output y_t . In our latent-model approach to the problem, this incremental updating exploits the series of pseudo-outputs (labels) generated by the latent model.

Classical (T-)HMM training, however, assumes that model parameters are learned in a batch way from one or more training sequences of observations. The question thus arises of what training sequences the latent model should be learned from, whether or not in fact the latent model should be learned from the single unsupervised data stream at disposal, and at what time instants in the data-stream T-HMM learning should take place.

We can identify a number of options.

1. **Separate (batch) training.** The option which is closest to the traditional setting is to identify the latent variable model once and for all from one or more training sequences of instances gathered from the site of interest, and set aside for this purpose. This obviously assumes that such sequences are available, which might not be the case. Alternatively, if only one sequence is available for both training and testing, one can learn the latent model from the labelled section $\mathcal{T}_0$ of the data stream and never update it, but simply use it to generate pseudolabels for the unlabelled section $\mathcal{T}_t$.
2. **Continuous (batch) training.** The second option is to learn the latent model from the test unsupervised data stream itself (as the only one available), but to learn a new version of the T-HMM as each new instance becomes available. Given a new instance x_t :

- (a) We use the current sequence $x_1, \dots, x_t$ to re-estimate the parameters $\{\theta_L, \theta_\Gamma\}$ of the T-HMM.
- (b) We estimate the probability $P(y_1, \dots, y_t)$ of all possible labellings for the current sequence under the newly learnt model.

This protocol fits well our philosophy of continually learning and revising our view of the world as new observations are gathered. However, it is very computationally expensive and possibly unfeasible for deployment on edge devices such as surveillance cameras.

3. **Continual (batch) training.** An intermediate, less expensive option is to update the latent model at regular intervals, rather than for each new unlabelled sample (hence, *continually*, rather than *continuously*). In this case:

- (a) The latent model is first initialised using the string of instances (which are temporally ordered) in the supervised training set $\mathcal{T}_0$ (without making use of the label information).
- (b) The latent model is then re-trained in *sessions*, each corresponding to a section of the unlabelled data stream.

This approach requires us to suitably partition the data stream $\mathcal{T}_t$ into sessions. The limit case is that in which a single learning session is used, after which the latent model is not updated further. We will elaborate on this option in our Experiments.

Sampling labellings as a state estimation problem

Once the parameters θ_L, θ_Γ of the latent T-HMM model have been learned from the training set of the unsupervised data streams, given a new unsupervised data stream $x_1, \dots, x_T$ computing its most likely labelling $\hat{y}_1, \dots, \hat{y}_T$ reduces to estimating the most likely sequence of hidden states. In fact, a sequence of discrete probability distributions $P(y_t | x_1, \dots, x_t; z_1, \dots, z_t; \theta^0, \theta_L, \theta_\Gamma)$ over the hidden variables can be obtained by simply ‘running’ the model in a forward pass.

6.3.5 Random Sampling based Memory Replay

After obtaining the estimated pseudo-labels from session-wise continual retraining, the continual semi-supervised learning problem is simplified to continual supervised learning. A random sampling [97] based memory replay method is employed to incrementally update the model. An initial memory buffer of size M is comprised

of 1000 samples randomly selected from the supervised training set, $\mathcal{T}_0$, and the first session, $\mathcal{T}_1$, of the suitably partitioned data stream.

The model is then incrementally updated at the end of each session via a randomly selected memory buffer. At the end of every session, the memory buffer is filled with new samples. Then, the memory buffer is updated to contain 1000 samples that are obtained through random selection. This incremental updating process continues until the last session of the partitioned unlabelled stream is reached. Details on how the data stream is partitioned can be found in Section 6.5.1. The effectiveness of random sampling-based memory replay along with one-off training and continual retraining is detailed in Section 6.5.2.

6.3.6 Alternative to HMM as Latent Variable Model

In our method, we used HMMs as latent variable modeling for pseudo-label generation, however, there are alternative models that you can consider. These alternatives may offer different strengths and may be more suitable for specific tasks or data characteristics. Some of the alternatives are as follows:

- **Variational Autoencoders (VAEs):** In CSSL, you can train a VAE on the labeled data and then generate pseudo-labels for unlabeled data by mapping them to the learned latent space.
- **Gaussian Mixture Models (GMMs):** GMMs are probabilistic models that assume data points are generated from a mixture of Gaussian distributions. They can be used for clustering and semi-supervised learning. In this context, GMMs can help cluster data points and assign pseudo-labels based on cluster assignments.
- **Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs):** RNNs and LSTMs are neural network architectures capable of modeling sequential data. They can be used to capture temporal dependencies in data and generate pseudo-labels based on the learned representations.

6.4 Experimental Analysis of Challenge Baselines

We ran experiments on both continual classification and continual regression using two new benchmarks of our own creation, which formed the basis of an IJCAI 2021 workshop and challenge²¹. The following description of the experimental setups mimics very closely the guidelines provided for the IJCAI challenge.

²¹<https://sites.google.com/view/sscl-workshop-ijcai-2021/>

Table 6.1: Performance of the initial supervised model versus that of the incrementally updated models (separately on V and T or in combination), using three standard evaluation metrics. For each metric we report both class average (C) and weighted average (W).

Methods / Evaluation metrics	Validation Fold						Test Fold					
	Precision		Recall		F1-Score		Precision		Recall		F1-Score	
	C	W	C	W	C	W	C	W	C	W	C	W
$sup - ft - union$	0.20	0.74	0.25	0.68	0.16	0.70	0.16	0.82	0.15	0.77	0.14	0.79
$upd - V / upd - T$	0.18	0.73	0.19	0.73	0.14	0.72	0.14	0.81	0.15	0.81	0.14	0.80
$upd - V + T$	0.17	0.73	0.19	0.73	0.14	0.72	0.14	0.80	0.14	0.79	0.13	0.79

Table 6.2: Evaluation on the CCC validation (V) and test (T) folds for different experimental setups. Here $sup - ft$, $upd - V$, and $upd - V + T$ refer to the supervised model fine-tuned on S , the model incrementally updated via self-training in sessions on the validation fold V , and the model incrementally updated on the combined validation V and test T folds, respectively. The average test MAE is reported for each model.

	Validation Fold V			Test Fold T		
	$FDST$	$UCSD$	$MALL$	$FDST$	$UCSD$	$MALL$
$sup - ft$	5.17	6.36	6.65	9.59	8.34	13.56
$upd - V$	6.05	7.58	7.40	8.00	7.45	16.38
$upd - V + T$	8.32	8.28	9.36	9.16	8.93	9.46

6.4.1 Results on Activity Recognition

Three types of experiments were performed to measure: (i) the performance of the baseline model after fine-tuning on the union of the supervised folds for the 15 CAR sequences ($sup - ft - union$), without any continual learning; (ii) the performance obtained by incrementally updating the model (in sessions) using the unlabelled validation ($upd - V$) or test ($upd - T$) data streams, considered separately; (iii) the performance of self-training over validation and training folds considered as a single contiguous data stream, again in a session-wise manner ($upd - V + T$).

The results are shown in Table 6.1. The standard classification evaluation metrics precision, recall, and F1-score were used for evaluation. For each evaluation metric we computed both the *class average* (obtained by computing the score for each class and taking the average without considering the number of training samples in each class, i.e., $F1_{class1} + F1_{class2} \cdots + F1_{class9}$) and the *weighted average* (which uses the number of samples W_i in each class i , i.e., $F1_{class1} * W_1 + F1_{class2} * W_2 \cdots + F1_{class9} * W_9$).

It can be noted that on the validation fold continual updating does improve performance to some extent, especially under Recall and F1 score. Improvements in Recall are visible on the test fold as well. All in all the baseline appears able to extract information from the unlabelled data stream.

6.4.2 Results on Crowd Counting

Table 6.2 shows a quantitative analysis of the performance of the fine-tuned supervised model (*sup*) and two incrementally updated models (*upd*) on the validation and the test split, respectively, using the mean absolute error (MAE) metric.

The experiments were performed under three different experimental settings: (i) the initial model fine tuned on the supervised folds (*sup-ft*), (ii) a model incrementally updated on the validation fold (*upd-V*), and (iii) a model incrementally updated on both the validation and test folds considered as a single data stream (*upd-V+T*). In Table 6.2 all three models are tested on both the *V* and *T* portions of the data streams. The initial model was trained for 100 epochs on the supervised fold (*S*), whereas the incremental models were self-trained for 5 epochs in each session (we also tried 2 epochs with no significant change observed).

When compared with classical papers in the crowd counting literature, the MAE values in Table 6.2 are noticeable higher. Concretely, a recent work that incorporates optical flow to improve crowd counting achieved MAEs of 1.76, 0.97, and 1.78 on the FDST, UCSD, and Mall dataset, respectively [219]. Also, the original MCNN implementation yielded MAEs of 3.77 and 1.07 on FDST and UCSD (performance was not reported on Mall) [224, 226]. The higher MAEs reported in our work are expected, due to the significantly different and more challenging training protocol (i.e., batch training in standard crowd counting work vs. continual learning here). For example, standard papers in crowd counting following the typical evaluation protocol employ 800 training images for the Mall and UCSD datasets; whereas, in our problem setting only 400 images are used for supervised training.

Previous batch results should be seen as upper bounds to our incremental task, whereas ours should be considered as new baselines for a new problem setting. While model updating does not seem to be effective (on average) on the validation streams, the effect is quite significant on the test fold (right side of the Table), with important performance gains.

6.5 Experimental Analysis of HMM latent Variable

6.5.1 Datasets and Setups

Continual activity recognition (CAR) dataset.

The CAR dataset (explained in Section 6.2.3) is used also for evaluating the latent variable and comparison with the previous challenge baseline. According to the protocol of the CAR dataset, each video sequence is divided into a supervised fold (*S*), corresponding to the first 5 minutes, a validation fold (*V*), corresponding to the sec-

Table 6.3: Time-varying HMM (TV HMM) vs. traditional HMM (HMM) based on one-off training, where P , R , and $F1$ are precision, recall, and $F1$ -score, respectively.

		HMM			TV HMM		
		P	R	F1	P	R	F1
CAR		0.86	0.18	0.24	0.83	0.37	0.49
50Salad_low	<i>timesplit</i>	0.22	0.05	0.08	0.226	0.052	0.077
	<i>cooksplitt</i>	0.076	0.023	0.018	0.076	0.023	0.018
50Salad_high	<i>timesplit</i>	0.39	0.23	0.27	0.4	0.24	0.26
	<i>cooksplitt</i>	0.407	0.236	0.256	0.405	0.238	0.26

ond 5-minute video, and a test fold (T), corresponding to the last 5-minute video. The last two folds are unsupervised.

In our experiments, for each sequence, the unlabelled data stream (i.e. the last 10 minutes, without distinction between validation and test folds) is partitioned into a number of sub-folds. Each sub-fold spans 1 minute in the CAR dataset, so that each unlabelled sequence is split into 10 sub-folds.

50 Salads activity detection dataset.

We also validate our method using the 50 Salads dataset, which consists of 50 videos by 25 cooks (actors) preparing salads, where each cook recorded two videos of themselves preparing salads by following different sequence of steps. The dataset contains two levels of annotations: “high-level” labels only span three classes, whereas “low-level” annotation involves a total of 51 classes. We added a background class to both settings, yielding a total number of classes of 4 and 52, respectively. To enforce a similar protocol to that of the CAR dataset, we introduced two splits to the 50 Salad dataset for training and unsupervised continual updates. The splits are defined as follows:

1. **Split by time** selects the first video for each cook as the training set and the second as the validation set. In this set-up, we do not have test set, so we only update our model continually over the validation set.
2. **Split by cook** divides the set of 25 cooks into 9, 8, and 8 for train, validation, and test sets, respectively. For evaluation, we perform 10-fold random split cross-validation and report the averaged results.

Session sub-folds span 1.28 minutes on average for the 50 Salads dataset, as the duration of each video is slightly longer than in the CAR dataset.

Table 6.4: Time-varying HMM (TV HMM) vs. traditional HMM (HMM) based on continual retraining, where P, R, and F1 are precision, recall, and F1-score, respectively.

		HMM			TV HMM		
		P	R	F1	P	R	F1
CAR		0.81	0.28	0.39	0.81	0.40	0.52
50Salad_low	<i>timesplit</i>	0.07	0.02	0.03	0.07	0.028	0.035
	<i>cooksplitt</i>	0.058	0.020	0.025	0.059	0.018	0.023
50Salad_high	<i>timesplit</i>	0.14	0.26	0.17	0.14	0.36	0.20
	<i>cooksplitt</i>	0.402	0.257	0.288	0.407	0.237	0.261

Table 6.5: Time-varying HMM (TV HMM) vs. traditional HMM (HMM) based on incremental update, where P, R, and F1 are precision, recall, and F1-score, respectively.

		Traditional HMM			Time-varying HMM		
		P	R	F1	P	R	F1
CAR		0.79	0.34	0.46	0.80	0.34	0.46
50Salad_low	<i>timesplit</i>	0.25	0.15	0.19	0.309	0.179	0.214
	<i>cooksplitt</i>	0.155	0.031	0.044	0.144	0.029	0.04
50Salad_high	<i>timesplit</i>	0.34	0.21	0.24	0.42	0.25	0.27
	<i>cooksplitt</i>	0.440	0.274	0.306	0.409	0.259	0.278

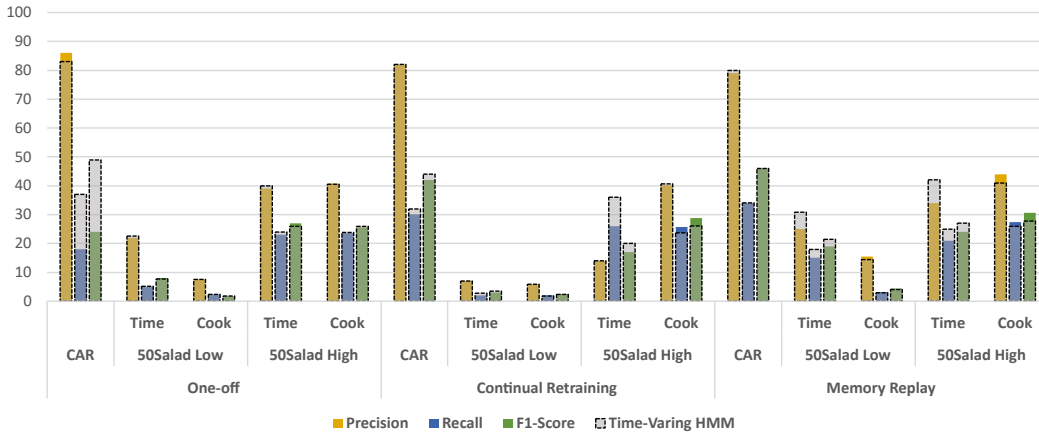


Figure 6.3: Time-varying HMM (dashed bars) vs. traditional HMM (colored bars).

6.5.2 Quantitative Analysis

The performance of Time-varying HMM (T-HMM) and traditional HMM (HMM) was assessed using three different experimental setups, namely: single, one-off training; continual retraining; and random sampling-based memory replay. In the first two settings, only the accuracy of the pseudo-labels generated by the latent model was assessed, whereas the third setting evaluated the whole framework in which the image classifier is incrementally updating using the pseudo-labels generated in sessions by the latent model, also incrementally updated. The quantitative comparison was performed using standard precision (P), recall (R) and F1-score (F1) performance

metrics. The quantitative analysis from the Table 6.3, 6.4 and 6.5 are further summarized in the Figure 6.3.

T-HMM vs. HMM It can be seen from Table 6.3, 6.4 and 6.5 that the T-HMM significantly performs better in terms of the F1 and R. T-HMM is better overall in all the three experimental setups. There were big gains in terms of R and F1 in the case of CAR dataset. The overall performance of T-HMM was quite low in case of 50Salad low as it comprises of relatively higher number of classes (51) as described in 6.5.1.

Continual retraining vs. One-off training It can be seen that the continual retraining improves the performance of not only T-HMM but also in the case of the HMM (Table 6.4). The HMM performance improves significantly in terms of F1 (+15%). Similarly, R also gained (+10%). The T-HMM significantly performs better than the HMM (+12% gain in R and F1 for CAR dataset). However, there was not a significant gain in terms of 50Salad dataset either low or high classes. The performance of T-HMM and HMM were close at some instances.

Incremental Update vs. Continual retraining The general trend of improvement can be seen in the case of incremental update via a random sampling memory replay as well. The incremental update helped improve the overall recall metric leading to better F1 score. In general, the overall improved F1 score can be found from +5-15% (for CAR dataset). While a similar trend of not much gain was observed for the 50Salad dataset.

6.6 Summary of the Chapter

In this chapter, we present CSSL workshop and challenge and a latent variable model. In the CSSL workshop and challenge, approach to the recently defined continual semi-supervised learning problem, in which the labels of the unsupervised data stream are estimated via a time-varying hidden Markov model. We formulated the continual semi-supervised learning problem and proposed suitable training and testing protocols for it. To encourage future work in the area we created the first two benchmark datasets, as the foundations of our IJCAI 2021 challenge. We proposed a simple strategy based on batch self-training a baseline model in sessions. The results show that, in both the activity recognition and the crowd counting challenge, the baseline appears in fact to be able to extract information from the unlabelled data stream. Nevertheless, the need for a more sophisticated approach leveraging the dynamics of the data stream is clear.

Next, we proposed a latent variable model approach to the defined continual semi-supervised learning problem, in which the labels of the unsupervised data stream are estimated via a time-varying hidden Markov model. Once the most likely label

assignment is estimated, the problem reduces to a continual supervised learning one, and can be tackled by, for instance, random sampling memory replay.

Employing a time-varying hidden Markov model as the latent model allows us to track the evolution of the target domain over time, under realistic assumptions, and therefore achieve better performance. The latent model was evaluated in three different experimental setups on two public datasets adapted to continual learning settings. The proposed solution showed a trend of better performance (a gain of 5-15% in F1 score) in all the experimental settings.

The model of the label dynamics we proposed makes some weak assumptions on the series of labels: in particular, that they can be described by a Markov chain and that emission probabilities can be described as multivariate Gaussian distributions. Our experiments support the view that these assumptions are fairly realistic – nevertheless, more general approaches should be explored. In the future, we plan to investigate the use of different logistic models of the dependency of the transition probabilities of the T-HMM from time, and extend the proposed latent variable approach to continual semi-supervised regression.

Chapter 7

Conclusions and Future Research Directions

In this thesis, we have presented a variety of novel methods for scene understanding divided into three parts. Part I 2.7.6: The ROad event awareness dataset for autonomous driving (Chapter 3), focuses on our novel approach to road event detection and the associated datasets and baselines. Part II 3.7 discusses complex activity detection (Chapter 4, Chapter 5) and the spatio-temporal modelling of the scene for holistic scene understanding. Finally, Part III 5.4.2 proposes a new paradigm for continual learning (Chapter 6) allowing us to update models incrementally by leveraging streaming unsupervised data.

To conclude the thesis, this chapter is divided into two sections. First, the methods proposed in this thesis are summarised in Section 7.1. Secondly, we illustrated possible research directions for future work in Section 7.2.

7.1 Summary of the Thesis

7.1.1 ROad event Awareness Dataset for Autonomous Driving

We proposed a strategy for situation awareness in autonomous driving based on the notion of road events, and contributed a new ROad event Awareness Dataset for Autonomous Driving (ROAD) as a benchmark for this area of research. The dataset, built on top of videos captured as part of the Oxford RobotCar dataset [126], has unique features in the field. Its rich annotation follows a multi-label philosophy in which road agents (including the AV), their locations, and the action(s) they perform are all labelled, and road events can be obtained by simply composing labels of the three types. The dataset contains 22 videos with 122K annotated video frames, for a total of 560K detection bounding boxes associated with 1.7M individual labels.

Baseline tests were conducted on ROAD using a new 3D-RetinaNet architecture, as well as a SlowFast backbone and a YOLOv5 model (for agent detection). Both frame-mAP and video-mAP were evaluated. Our preliminary results highlight the challenging nature of ROAD, with the SlowFast baseline achieving a video-mAP on the three main tasks comprised between 20% and 30%, at low localisation precision (20% overlap). YOLOv5, however, was able to achieve significantly better performance. These findings were reinforced by the results of the ROAD @ ICCV 2021 challenge, and support the need for an even broader analysis, while highlighting the significant challenges specific to situation awareness in road scenarios.

In another study we proposed a new learning framework, termed “learning with requirements”, and a new dataset for this task, called ROAD-R (the ROad event Awareness Dataset with logical Requirements), derived from ROAD. We showed that SOTA models most of the time violate the requirements, and how it is possible to exploit the requirements to create models that are compliant with (i.e., strictly satisfy) the requirements while improving their performance. We envision that requirement specification will become a standard step in the development of machine learning models, to guarantee their safety, as it is in any software development process. In this sense, ROAD-R may be followed by many other datasets with formally specified requirements.

7.1.2 Complex Activity Detection

Moving up in the scene understanding reasoning chain, we presented a spatiotemporal complex activity detection framework that leverages both part deformation and a heterogeneous graph representation. Our approach is based on three building blocks; action tube detection, part-based deformable 3D RoI pooling for feature extraction, and a GCN module that processes the variable number of detected action tubes to model the overall semantics of a complex activity. In an additional contribution, we temporally annotated two recently released benchmark datasets (ROAD and ESAD) in terms of long-term complex activities. Both datasets come with video-level action tube annotation, making them suitable benchmarks for future work in this area. We thoroughly evaluated our method, showing the effectiveness of our 3D part-based deformable model approach for the detection of complex activities.

This complex activity model works well with datasets endowed with spatiotemporal action annotations. This, however, is a serious limitation for the method cannot be applied to datasets without such ground truth. To overcome this limitation, in another study we proposed a novel hybrid graph neural network-based framework that combines both scene graph attention and a temporal graph to model activities of arbitrary duration. Our proposed framework is divided into three main building

blocks: agent tube detection and feature extraction; a local scene graph construction with attention; and a temporal graph for recognising the class label and localising each activity instance. We tested our method on three benchmark datasets, showing the effectiveness of our method in detecting both short-term and long-term activities, thanks to its ability to model their finer-grained structure without the need for extra annotation.

7.1.3 Continual Semi-Supervised Learning

Lastly, in the last part of this thesis, we formulated an original “continual semi-supervised learning” problem and proposed suitable training and testing protocols for it. To encourage future work in the area we created the first two benchmark datasets, as the foundations of our CSSL @ IJCAI 2021 challenge. We proposed a simple strategy based on batch self-training a baseline model in sessions. The results show that, in both the activity recognition and the crowd counting challenge, the baseline appears in fact to be able to extract information from the unlabelled data stream. Nevertheless, the need for a more sophisticated approach leveraging the dynamics of the data stream is clear.

To move in this direction we then formulated a latent variable model approach to the continual semi-supervised learning problem, in which the labels of the unsupervised data stream are estimated via a time-varying hidden Markov model. Once the most likely label assignment is estimated, the problem reduces to a continual supervised learning one and can be tackled by, for instance, random sampling memory replay.

Employing a time-varying hidden Markov model as the latent model allows us to track the evolution of the target domain over time, under realistic assumptions, and therefore achieve better performance. The latent model was evaluated in three different experimental setups on two public datasets adapted to continual learning settings. The proposed solution showed a trend of better performance (a gain of 5-15% in F1 score) in all the experimental settings.

The model of the label dynamics we proposed makes some weak assumptions on the series of labels: in particular, that they can be described by a Markov chain and that emission probabilities can be described as multivariate Gaussian distributions. Our experiments support the view that these assumptions are fairly realistic – nevertheless, more general approaches should be explored.

7.2 Future Research Directions

7.2.1 Extension of ROAD

Our ROAD dataset is extensible to a number of challenging tasks associated with situation awareness in autonomous driving, such as event prediction, trajectory prediction, continual learning, and machine theory of mind.

We are currently actively working to scale up our benchmark by at least one order of magnitude by extending ROAD-like annotation to major autonomous driving datasets such as PIE and WAYMO.

We also intend to propose to organise a follow-up Workshop (possibly a series of events) at the next suitable upcoming major computer vision (e.g. ICCV, CVPR) and robotics (e.g. ICRA, IROS) conferences, in order to start mobilising both communities towards the further development of the benchmark and associated array of tasks and baselines. An additional workshop on neurosymbolic learning in the real world is being considered as well.

7.2.2 Holistic Graph Approach for Complex Activity Detection and Prediction

In the future we intend to progress further from incremental inference to incremental training, by learning to construct activity graphs in an incremental manner, thus opening the way to applications such as future activity anticipation [205] and pedestrian intent prediction [206], while overcoming the need for a fixed-size snippet representation. A further exciting line of research is the modelling of the uncertainty associated with complex scenes, in either the Bayesian [207] or the full epistemic setting [208, 209].

7.2.3 Extension to Spatiotemporal Sentence Grounding

We also aim to expand our activity modelling to include *spatiotemporal sentence grounding*. More in detail, we plan to integrate our vision models for complex activity detection with language models, in order to localise the specific moment in video based on an input language query. The ultimate goal of our spatiotemporal sentence grounding will be to localise the object tubes in untrimmed videos, where the object may only exist in a very small segment of the video with the help of an input query sentence.

7.2.4 Continual Learning Extension to Regression

Just as in the situations awareness of autonomous driving area, in the future we intend to organise a follow-up Workshop and Challenge (possibly a series of events) at the next suitable upcoming major machine learning conferences (e.g. IJCAI, ICLR, CVPR) to attract researchers in the field of (semi-supervised or unsupervised) continual learning to develop and improve their methods using our proposed datasets and setup. Model-wise, we plan to investigate the use of different logistic models of the dependency of the transition probabilities of the T-HMM from time, and extend the proposed latent variable approach to continual semi-supervised regression.

References

- [1] S. Saha, G. Singh, and F. Cuzzolin, “Two-stream amtnet for action detection,” *arXiv preprint arXiv:2004.01494*, 2020.
- [2] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nusenes: A multimodal dataset for autonomous driving,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.
- [3] F. Camara, N. Bellotto, S. Cosar, F. Weber, D. Nathanael, M. Althoff, J. Wu, J. Ruenz, A. Dietrich, G. Markkula, *et al.*, “Pedestrian models for autonomous driving part ii: high-level models of human behavior,” *IEEE Transactions on Intelligent Transportation Systems*, 2020.
- [4] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, “Peeking into the future: Predicting future person activities and locations in videos,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5725–5734, 2019.
- [5] A. Zia, A. Hung, I. Essa, and A. Jarc, “Surgical activity recognition in robot-assisted radical prostatectomy using deep learning,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 273–280, Springer, 2018.
- [6] G. Hu, B. Cui, Y. He, and S. Yu, “Progressive relation learning for group activity recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 980–989, 2020.
- [7] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, “Action recognition in video sequences using deep bi-directional lstm with cnn features,” *IEEE access*, vol. 6, pp. 1155–1166, 2017.
- [8] G. SingH and F. Cuzzolin, “Recurrent convolutions for causal 3d cnns,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

-
- [9] S. Yun, S. J. Oh, B. Heo, D. Han, and J. Kim, "Videomix: Rethinking data augmentation for video classification," *arXiv preprint arXiv:2012.03457*, 2020.
 - [10] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Richly activated graph convolutional network for robust skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1915–1925, 2020.
 - [11] L. Meng, B. Zhao, B. Chang, G. Huang, W. Sun, F. Tung, and L. Sigal, "Interpretable spatio-temporal attention for video action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.
 - [12] M. Yu, M. Bambacus, G. Cervone, K. Clarke, D. Duffy, Q. Huang, J. Li, W. Li, Z. Li, Q. Liu, *et al.*, "Spatiotemporal event detection: A review," *International Journal of Digital Earth*, vol. 13, no. 12, pp. 1339–1365, 2020.
 - [13] L. Li, Y. Lin, B. Du, F. Yang, and B. Ran, "Real-time traffic incident detection based on a hybrid deep learning model," *Transportmetrica A: transport science*, vol. 18, no. 1, pp. 78–98, 2022.
 - [14] S. Khan, K. Muhammad, S. Mumtaz, S. W. Baik, and V. H. C. de Albuquerque, "Energy-efficient deep cnn for smoke detection in foggy iot environment," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 9237–9245, 2019.
 - [15] S. Khan, K. Muhammad, T. Hussain, J. Del Ser, F. Cuzzolin, S. Bhattacharyya, Z. Akhtar, and V. H. C. de Albuquerque, "Deepsnake: Deep learning model for smoke detection and segmentation in outdoor environments," *Expert Systems with Applications*, vol. 182, p. 115125, 2021.
 - [16] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, "Graph convolutional module for temporal action localization in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
 - [17] S. Khan and F. Cuzzolin, "Spatiotemporal deformable scene graphs for complex activity detection," in *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*.
 - [18] S. Xiao, Z. Zhao, Z. Zhang, X. Yan, and M. Yang, "Convolutional hierarchical attention network for query-focused video summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12426–12433, 2020.

- [19] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, “Streamlined dense video captioning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6588–6597, 2019.
- [20] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, “Bsn: Boundary sensitive network for temporal action proposal generation,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.
- [21] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, “Graph convolutional networks for temporal action localization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7094–7103, 2019.
- [22] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, “Bmn: Boundary-matching network for temporal action proposal generation,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3889–3898, 2019.
- [23] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, “Gaussian temporal awareness networks for action localization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 344–353, 2019.
- [24] Y. Liu, L. Ma, Y. Zhang, W. Liu, and S.-F. Chang, “Multi-granularity generator for temporal action proposal,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3604–3613, 2019.
- [25] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, “G-tad: Sub-graph localization for temporal action detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10156–10165, 2020.
- [26] K. Xia, L. Wang, S. Zhou, N. Zheng, and W. Tang, “Learning to refactor action and co-occurrence features for temporal action localization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13884–13893, 2022.
- [27] X. Liu, S. Bai, and X. Bai, “An empirical study of end-to-end temporal action detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20010–20019, 2022.
- [28] H.-Y. Hsieh, D.-J. Chen, and T.-L. Liu, “Contextual proposal network for action localization,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2129–2138, 2022.

- [29] W. Bao, Q. Yu, and Y. Kong, “Opental: Towards open set temporal action localization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2979–2989, 2022.
- [30] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, “Activitynet: A large-scale video benchmark for human activity understanding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 961–970, 2015.
- [31] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, “The thumos challenge on action recognition for videos “in the wild”,” *Computer Vision and Image Understanding*, vol. 155, pp. 1–23, 2017.
- [32] Z. Zhu, L. Wang, W. Tang, Z. Liu, N. Zheng, and G. Hua, “Learning disentangled classification and localization representations for temporal action localization,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 2, 2022.
- [33] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.
- [34] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, “The kinetics human action video dataset,” *arXiv preprint arXiv:1705.06950*, 2017.
- [35] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, “Rethinking the faster r-cnn architecture for temporal action localization,” in *proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1130–1139, 2018.
- [36] S. Khan and F. Cuzzolin, “Spatiotemporal deformable models for long-term complex activity detection,” *arXiv preprint arXiv:2104.08194*, 2021.
- [37] F. Cheng and G. Bertasius, “Tallformer: Temporal action localization with long-memory transformer,” *arXiv preprint arXiv:2204.01680*, 2022.
- [38] Z. Shou, D. Wang, and S.-F. Chang, “Temporal action localization in untrimmed videos via multi-stage cnns,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1049–1058, 2016.
- [39] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, “Scene graph generation with external knowledge and image reconstruction,” in *Proceedings of the*

- IEEE/CVF conference on computer vision and pattern recognition*, pp. 1969–1978, 2019.
- [40] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, “Action genome: Actions as compositions of spatio-temporal scene graphs,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10236–10247, 2020.
- [41] R. Koner, H. Li, M. Hildebrandt, D. Das, V. Tresp, and S. Günnemann, “Graphhopper: Multi-hop scene graph reasoning for visual question answering,” in *The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20*, pp. 111–127, Springer, 2021.
- [42] J. Fan, P. Zheng, and S. Li, “Vision-based holistic scene understanding towards proactive human–robot collaboration,” *Robotics and Computer-Integrated Manufacturing*, vol. 75, p. 102304, 2022.
- [43] V. Lomonaco, L. Pellegrini, A. Cossu, A. Carta, G. Graffieti, T. L. Hayes, M. De Lange, M. Masana, J. Pomponi, G. M. Van de Ven, *et al.*, “Avalanche: an end-to-end library for continual learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3600–3610, 2021.
- [44] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, and N. Díaz-Rodríguez, “Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges,” *Information fusion*, vol. 58, pp. 52–68, 2020.
- [45] G. Singh, S. Akrigg, M. Di Maio, V. Fontana, R. J. Alitappeh, S. Khan, S. Saha, K. Jeddisaravi, F. Yousefi, J. Culley, *et al.*, “Road: The road event awareness dataset for autonomous driving,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 1036–1054, 2022.
- [46] E. Giunchiglia, M. C. Stoian, S. Khan, F. Cuzzolin, and T. Lukasiewicz, “Road-r: The autonomous driving dataset with logical requirements,” *arXiv preprint arXiv:2210.01597*, 2022.
- [47] A. Shahbaz, S. Khan, M. A. Hossain, V. Lomonaco, K. Cannons, Z. Xu, and F. Cuzzolin, “International workshop on continual semi-supervised learning: Introduction, benchmarks and baselines,” in *International Workshop on Continual Semi-Supervised Learning*, pp. 1–14, Springer, 2022.

-
- [48] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.
 - [49] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of CVPR 2016*, pp. 3213–3223, 2016.
 - [50] P. Sun, H. Kretschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, “Scalability in perception for autonomous driving: Waymo open dataset,” 2019.
 - [51] P. Wang, X. Huang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, “The apollo-scape open dataset for autonomous driving and its application,” *IEEE transactions on pattern analysis and machine intelligence*, 2019.
 - [52] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, “Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6262–6271, 2019.
 - [53] S. Malla, B. Dariush, and C. Choi, “Titan: Future forecast using action priors,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11186–11196, 2020.
 - [54] Y. Liao, J. Xie, and A. Geiger, “KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d,” *arXiv.org*, vol. 2109.13410, 2021.
 - [55] Q.-H. Pham, P. Sevestre, R. S. Pahwa, H. Zhan, C. H. Pang, Y. Chen, A. Mustafa, V. Chandrasekhar, and J. Lin, “A* 3d dataset: Towards autonomous driving in challenging environments,” *arXiv preprint arXiv:1909.07541*, 2019.
 - [56] A. Patil, S. Malla, H. Gang, and Y.-T. Chen, “The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes,” in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 9552–9557, IEEE, 2019.

- [57] J. Chang, Ming-Fang D. Wang, P. Carr, S. Lucey, D. Ramanan, and others et al., “Argoverse: 3d tracking and forecasting with rich maps,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8748–8757, 2019.
- [58] L. Ding, J. Terwilliger, R. Sherony, B. Reimer, and L. Fridman, “MIT DriveSeg (Manual) Dataset,” 2020.
- [59] Y. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, “Thumos challenge: Action recognition with a large number of classes,” <http://cvc.ucf.edu/THUMOS14>, 2014.
- [60] C. Gu, C. Sun, S. Vijayanarasimhan, C. Pantofaru, D. A. Ross, G. Toderici, Y. Li, S. Ricco, R. Sukthankar, C. Schmid, *et al.*, “Ava: A video dataset of spatio-temporally localized atomic visual actions,” *arXiv preprint arXiv:1705.08421*, 2017.
- [61] Y. Li, L. Chen, R. He, Z. Wang, G. Wu, and L. Wang, “Multisports: A multi-person video dataset of spatio-temporally localized sports actions,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13536–13545, 2021.
- [62] Y. Bai, Y. Wang, Y. Tong, Y. Yang, Q. Liu, and J. Liu, “Boundary content graph neural network for temporal action proposal generation,” in *European Conference on Computer Vision*, pp. 121–137, Springer, 2020.
- [63] Y. Huang, Y. Sugano, and Y. Sato, “Improving action segmentation via graph-based temporal reasoning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14024–14034, 2020.
- [64] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang, “Autoloc: Weakly-supervised temporal action localization in untrimmed videos,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 154–171, 2018.
- [65] S. Paul, S. Roy, and A. K. Roy-Chowdhury, “W-talc: Weakly-supervised temporal activity localization and classification,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 563–579, 2018.
- [66] D. Zhang, X. Dai, and Y.-F. Wang, “Metal: Minimum effort temporal activity localization in untrimmed videos,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3882–3892, 2020.

-
- [67] P. Bharti, D. De, S. Chellappan, and S. K. Das, “Human: Complex activity recognition with multi-modal multi-positional body sensing,” *IEEE Transactions on Mobile Computing*, vol. 18, no. 4, pp. 857–870, 2018.
- [68] N. Thakur and C. Y. Han, “An improved approach for complex activity recognition in smart homes,” in *International Conference on Software and Systems Reuse*, pp. 220–231, Springer, 2019.
- [69] X. Zhou, W. Liang, I. Kevin, K. Wang, H. Wang, L. T. Yang, and Q. Jin, “Deep-learning-enhanced human activity recognition for internet of healthcare things,” *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6429–6438, 2020.
- [70] M. M. Arzani, M. Fathy, A. A. Azirani, and E. Adeli, “Switching structured prediction for simple and complex human activity recognition,” *IEEE transactions on cybernetics*, 2020.
- [71] H. Kwon, C. Tong, H. Haresamudram, Y. Gao, G. D. Abowd, N. D. Lane, and T. Ploetz, “Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–29, 2020.
- [72] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick, “Long-term feature banks for detailed video understanding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 284–293, 2019.
- [73] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, “Hacs: Human action clips and segments dataset for recognition and temporal localization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8668–8678, 2019.
- [74] S. Sudhakaran, S. Escalera, and O. Lanz, “Lsta: Long short-term attention for egocentric action recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9954–9963, 2019.
- [75] C. Luo and A. L. Yuille, “Grouped spatial-temporal aggregation for efficient action recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5512–5521, 2019.
- [76] L. Wang, Y. Qiao, and X. Tang, “Action recognition and detection by combining motion and appearance features,” *THUMOS14 Action Recognition Challenge*, vol. 1, no. 2, p. 2, 2014.

- [77] H. Xu, A. Das, and K. Saenko, “R-c3d: Region convolutional 3d network for temporal activity detection,” in *Proceedings of the IEEE international conference on computer vision*, pp. 5783–5792, 2017.
- [78] J. Gao, Z. Yang, K. Chen, C. Sun, and R. Nevatia, “Turn tap: Temporal unit regression network for temporal action proposals,” in *Proceedings of the IEEE international conference on computer vision*, pp. 3628–3636, 2017.
- [79] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei, and J. C. Niebles, “End-to-end, single-stream temporal action detection in untrimmed videos,” 2019.
- [80] Y. Huang, Q. Dai, and Y. Lu, “Decoupling localization and classification in single shot temporal action detection,” in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1288–1293, IEEE, 2019.
- [81] C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu, “Learning salient boundary feature for anchor-free temporal action localization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3320–3329, 2021.
- [82] X. Wang, S. Zhang, Z. Qing, Y. Shao, Z. Zuo, C. Gao, and N. Sang, “Oadtr: Online action detection with transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7565–7575, 2021.
- [83] Y. K. Wentao Bao, Qi Yu, “Opental: Towards open set temporal action localization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.
- [84] Z. Zhu, L. Wang, W. Tang, Z. Liu, N. Zheng, and G. Hua, “Learning disentangled classification and localization representations for temporal action localization,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 3644–3652, Jun. 2022.
- [85] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, “Dynamic graph cnn for learning on point clouds,” *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.
- [86] Z. Xie, J. Chen, and B. Peng, “Point clouds learning with attention-based graph convolution networks,” *Neurocomputing*, vol. 402, pp. 245–255, 2020.
- [87] G. Gkioxari, J. Malik, and J. Johnson, “Mesh r-cnn,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9785–9795, 2019.

- [88] X. Wang and A. Gupta, “Videos as space-time region graphs,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 399–417, 2018.
- [89] X. Liu, J.-Y. Lee, and H. Jin, “Learning video representations from correspondence proposals,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4273–4281, 2019.
- [90] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, “Graph-based global reasoning networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 433–442, 2019.
- [91] J. Li, X. Liu, Z. Zong, W. Zhao, M. Zhang, and J. Song, “Graph attention based proposal 3d convnets for action detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 4626–4633, 2020.
- [92] M. Nawhal and G. Mori, “Activity graph transformer for temporal action localization,” *arXiv preprint arXiv:2101.08540*, 2021.
- [93] Z. Yang, J. Qin, and D. Huang, “Acgnet: Action complement graph network for weakly-supervised temporal action localization,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 3090–3098, 2022.
- [94] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” *arXiv preprint arXiv:1710.10903*, 2017.
- [95] S. Farquhar and Y. Gal, “Towards robust evaluations of continual learning,” *ArXiv*, vol. abs/1805.09733, 2018.
- [96] A. Mallya and S. Lazebnik, “Packnet: Adding multiple tasks to a single network by iterative pruning,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.
- [97] R. Aljundi, M. Lin, B. Goujaud, and Y. Bengio, “Gradient based sample selection for online continual learning,” *CoRR*, vol. abs/1903.08671, 2019.
- [98] M. Pratama, A. Ashfahani, and E. Lughofer, “Unsupervised continual learning via self-adaptive deep clustering approach,” *CoRR*, vol. abs/2106.14563, 2021.
- [99] J. He and F. Zhu, “Unsupervised continual learning via pseudo labels,” *CoRR*, vol. abs/2104.07164, 2021.
- [100] X. Zhan, J. Xie, Z. Liu, Y. Ong, and C. C. Loy, “Online deep clustering for unsupervised representation learning,” *CoRR*, vol. abs/2006.10645, 2020.

- [101] W. Li and T. Liu, "Time varying and condition adaptive hidden markov model for tool wear state estimation and remaining useful life prediction in micro-milling," *Mechanical Systems and Signal Processing*, vol. 131, pp. 689–702, 2019.
- [102] N. J. Williams, I. Daly, and S. J. Nasuto, "Markov model-based method to analyse time-varying networks in eeg task-related data," *Frontiers in Computational Neuroscience*, vol. 12, p. 76, 2018.
- [103] M. Li and B. M. Bolker, "Incorporating periodic variability in hidden markov models for animal movement," *Movement ecology*, vol. 5, no. 1, p. 1, 2017.
- [104] E. Marhasev, M. Hadad, and G. A. Kaminka, "Non-stationary hidden semi markov models in activity recognition," in *Proceedings of the AAAI Workshop on Modeling Others from Observations (MOO-06)*, 2006.
- [105] Z. Wang, E. E. Kuruoglu, X. Yang, Y. Xu, and S. Yu, "Event recognition with time varying hidden markov model," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1761–1764, IEEE, 2009.
- [106] H. Chung, T. A. Walls, and Y. Park, "A latent transition model with logistic regression," *Psychometrika*, vol. 72, no. 3, p. 413, 2007.
- [107] F. X. Diebold, J.-H. Lee, and G. C. Weinbach, "Regime switching with time-varying transition probabilities," 1993.
- [108] E. Otranto, "A time varying hidden markov model with latent information," *Statistical Modelling*, vol. 8, no. 4, pp. 347–366, 2008.
- [109] S. Chib and M. Dueker, "Non-markovian regime switching with endogenous states and time-varying state strengths," *FRB of St. Louis Working Paper No.*, 2004.
- [110] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017.
- [111] V. S. Bawa, G. Singh, F. KapingA, I. Skarga-Bandurova, E. Oleari, A. Leporini, C. Landolfo, P. Zhao, X. Xiang, G. Luo, *et al.*, "The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods," *arXiv preprint arXiv:2104.03178*, 2021.
- [112] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the*

2013 ACM international joint conference on Pervasive and ubiquitous computing, pp. 729–738, 2013.

- [113] R. Behringer, S. Sundareswaran, B. Gregory, R. Elsley, B. Addison, W. Guthmiller, R. Daily, and D. Bevly, “The darpa grand challenge-development of an autonomous vehicle,” in *IEEE Intelligent Vehicles Symposium, 2004*, pp. 226–231, IEEE, 2004.
- [114] J. Winn and J. Shotton, “The layout consistent random field for recognizing and segmenting partially occluded objects,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, pp. 37–44, IEEE, 2006.
- [115] K. Korosec, “Toyota is betting on this startup to drive its self-driving car plans forward.” Available at: <http://fortune.com/2017/09/27/toyota-self-driving-car-luminar/>.
- [116] G. Pandey, J. R. McBride, and R. M. Eustice, “Ford campus vision and lidar data set,” *International Journal of Robotics Research*, vol. 30, no. 13, pp. 1543–1552, 2011.
- [117] M. e. a. Maurer, *Autonomous driving: technical, legal and social aspects*. Springer Nature, 2016.
- [118] M. Bertozzi, A. Broggi, and A. Fascioli, “Vision-based intelligent vehicles: State of the art and perspectives,” *Robotics and Autonomous Systems*, vol. 32, no. 1, pp. 1 – 16, 2000.
- [119] J. Redmon and A. Farhadi, “Yolo9000: Better, faster, stronger,” in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 6517—6525, 2017.
- [120] M. Codevilla, Felipe Dosovitskiy, “End-to-end driving via conditional imitation learning,” in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–9, 2018.
- [121] L. F. et al., “Arguing machines: Perception-control system redundancy and edge case discovery in real-world autonomous driving.” ArXiv preprint ArXiv:1710.04459, 2017.
- [122] F. Cuzzolin, A. Morelli, B. Cirstea, and B. J. Sahakian, “Knowing me, knowing you: Theory of mind in AI,” *Psychological Medicine*, vol. 50, no. 7, pp. 1057–1061, May 2020.

- [123] A. Rasouli and J. K. Tsotsos, “Autonomous vehicles that interact with pedestrians: A survey of theory and practice,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 900—918, 2020.
- [124] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, “Human motion trajectory prediction: A survey.” *arXiv preprint arXiv:1905.06113*, 2019.
- [125] S. Armstrong and S. Mindermann, “Occam’s razor is insufficient to infer the preferences of irrational agents,” in *Advances in Neural Information Processing Systems*, vol. 31, pp. 5603—5614, 2018.
- [126] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 year, 1000 km: The oxford robotcar dataset,” *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.
- [127] G. Singh, S. Saha, M. Sapienza, P. H. Torr, and F. Cuzzolin, “Online real-time multiple spatiotemporal action localisation and prediction,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3637–3646, 2017.
- [128] S. Saha, G. Singh, M. Sapienza, P. H. Torr, and F. Cuzzolin, “Deep learning for detecting multiple space-time action tubes in videos,” *arXiv preprint arXiv:1608.01529*, 2016.
- [129] G. Gkioxari and J. Malik, “Finding action tubes,” in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2015.
- [130] C. Feichtenhofer, H. Fan, J. Malik, and K. He, “Slowfast networks for video recognition,” in *Proceedings of the IEEE international conference on computer vision*, pp. 6202–6211, 2019.
- [131] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, “Bdd100k: A diverse driving video database with scalable annotation tooling,” *arXiv preprint arXiv:1805.04687*, 2018.
- [132] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [133] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, “Charades-ego: A large-scale dataset of paired third and first person videos,” 2018.

-
- [134] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *CVPR*, vol. 1, p. 4, 2017.
 - [135] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.
 - [136] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
 - [137] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *CVPR*, 2018.
 - [138] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, “Action tubelet detector for spatio-temporal action localization,” in *Proc. Int. Conf. Computer Vision*, 2017.
 - [139] Y. Li, Z. Wang, L. Wang, and G. Wu, “Actions as moving points,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
 - [140] G. Singh, S. Saha, and F. Cuzzolin, “Tramnet-transition matrix network for efficient action tube proposals,” in *Asian Conference on Computer Vision*, pp. 420–437, Springer, 2018.
 - [141] I. Sommerville, *Software Engineering*. 2011.
 - [142] S. Saha, G. Singh, M. Sapienza, P. H. S. Torr, and F. Cuzzolin, “Deep learning for detecting multiple space-time action tubes in videos,” in *British Machine Vision Conference*, 2016.
 - [143] P. Weinzaepfel, Z. Harchaoui, and C. Schmid, “Learning to track for spatio-temporal action localization,” in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, June 2015.
 - [144] Z. Li, Dong et al. proposal and recognition networks for action detection,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 303–318, 2018.
 - [145] M. Li, Y.-X. Wang, and D. Ramanan, “Towards streaming perception,” in *European Conference on Computer Vision*, pp. 473–488, Springer, 2020.
 - [146] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, “Video swin transformer,” *arXiv preprint arXiv:2106.13230*, 2021.

- [147] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, “Multiscale vision transformers,” *arXiv preprint arXiv:2104.11227*, 2021.
- [148] W. Liu, G. Kang, P.-Y. Huang, X. Chang, Y. Qian, J. Liang, L. Gui, J. Wen, and P. Chen, “Argus: Efficient activity detection system for extended video analysis,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pp. 126–133, 2020.
- [149] G. Singh, S. Akrigg, M. D. Maio, V. Fontana, R. J. Alitappeh, S. Saha, K. J. Saravi, F. Yousefi, J. Culley, T. Nicholson, J. Omokeowa, S. Khan, S. Grazioso, A. Bradley, G. D. Gironimo, and F. Cuzzolin, “ROAD: The road event awareness dataset for autonomous driving,” *IEEE TPAMI*, 2022.
- [150] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. of CVPR*, 2018.
- [151] Y. Hua, Z. Zhao, Z. Liu, X. Chen, R. Li, and H. Zhang, “Traffic prediction based on random connectivity in deep learning with long short-term memory,” in *Proc. of VTC-Fall*, 2018.
- [152] Y. LeCun, L. Bottou, G. Orr, and K. Müller, “Efficient backprop,” in *Neural Networks: Tricks of the Trade*, 2012.
- [153] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, “Action tubelet detector for spatio-temporal action localization,” in *Proc. of ICCV*, 2017.
- [154] D. Li, Z. Qiu, Q. Dai, T. Yao, and T. Mei, “Recurrent tubelet proposal and recognition networks for action detection,” in *Proc. of ECCV*, 2018.
- [155] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, “Scalability in perception for autonomous driving: Waymo open dataset,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2446–2454, 2020.
- [156] Y. Kong, Z. Tao, and Y. Fu, “Deep sequential context networks for action prediction,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1473–1481, 2017.
- [157] Y. Kong, Z. Tao, and Y. Fu, “Adversarial action prediction networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 539–553, 2018.
- [158] G. Singh, S. Saha, and F. Cuzzolin, “Predicting action tubes,” in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

- [159] C. Hubmann, M. Becker, D. Althoff, D. Lenz, and C. Stiller, “Decision making for autonomous driving considering interaction and uncertain prediction of surrounding vehicles,” in *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1671–1678, IEEE, 2017.
- [160] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. A. Eslami, and M. Botvinick, “Machine theory of mind,” in *International conference on machine learning*, pp. 4218–4227, PMLR, 2018.
- [161] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, “Continual lifelong learning with neural networks: A review,” *Neural Networks*, vol. 113, pp. 54–71, 2019.
- [162] V. Singh Bawa, G. Singh, F. KapingA, I. Skarga-Bandurova, E. Oleari, A. Leporini, C. Landolfo, P. Zhao, X. Xiang, G. Luo, *et al.*, “The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods,” *arXiv e-prints*, pp. arXiv–2104, 2021.
- [163] B. Laxton, J. Lim, and D. Kriegman, “Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2007.
- [164] M. S. Ryoo and J. K. Aggarwal, “Recognition of composite human activities through context-free grammar based representation,” in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06)*, vol. 2, pp. 1709–1718, IEEE, 2006.
- [165] Y. A. Ivanov and A. F. Bobick, “Recognition of visual activities and interactions by stochastic parsing,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 852–872, 2000.
- [166] S. Park and J. K. Aggarwal, “A hierarchical bayesian network for event recognition of human actions and interactions,” *Multimedia systems*, vol. 10, pp. 164–179, 2004.
- [167] W. Wolf and I. B. Ozer, “A smart camera for real-time human activity recognition,” in *2001 IEEE Workshop on Signal Processing Systems. SiPS 2001. Design and Implementation (Cat. No. 01TH8578)*, pp. 217–224, IEEE, 2001.
- [168] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, “Learning to localize actions from moments,” *arXiv preprint arXiv:2008.13705*, 2020.

- [169] B. Shi, Q. Dai, Y. Mu, and J. Wang, “Weakly-supervised action localization by generative attention modeling,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1009–1019, 2020.
- [170] G. Gong, X. Wang, Y. Mu, and Q. Tian, “Learning temporal co-attention models for unsupervised video action localization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9819–9828, 2020.
- [171] K. Soomro, A. R. Zamir, and M. Shah, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” *arXiv preprint arXiv:1212.0402*, 2012.
- [172] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, “Hollywood in homes: Crowdsourcing data collection for activity understanding,” in *European Conference on Computer Vision*, pp. 510–526, Springer, 2016.
- [173] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, “Scaling egocentric vision: The epic-kitchens dataset,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 720–736, 2018.
- [174] R. De Rosa, N. Cesa-Bianchi, I. Gori, and F. Cuzzolin, “Online action recognition via nonparametric incremental learning,” in *BMVC*, 2014.
- [175] S. Jetley and F. Cuzzolin, “3d activity recognition using motion history and binary shape templates,” in *Asian Conference on Computer Vision*, pp. 129–144, Springer, 2014.
- [176] S. Saha, G. Singh, and F. Cuzzolin, “Amtnet: Action-micro-tube regression by end-to-end trainable deep architecture,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4414–4423, 2017.
- [177] H. S. Behl, M. Sapienza, G. Singh, S. Saha, F. Cuzzolin, and P. H. Torr, “Incremental tube construction for human action detection,” *arXiv preprint arXiv:1704.01358*, 2017.
- [178] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, “Deformable convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 764–773, 2017.
- [179] X. Zhu, H. Hu, S. Lin, and J. Dai, “Deformable convnets v2: More deformable, better results,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9308–9316, 2019.

- [180] G. Singh, S. Akrigg, M. Di Maio, V. Fontana, R. J. Alitappeh, S. Saha, K. Jeddisaravi, F. Yousefi, J. Culley, T. Nicholson, *et al.*, “Road: The road event awareness dataset for autonomous driving,” *arXiv preprint arXiv:2102.11585*, 2021.
- [181] V. S. Bawa, G. Singh, F. KapingA, A. Leporini, C. Landolfo, A. Stabile, F. Setti, R. Muradore, E. Oleari, F. Cuzzolin, *et al.*, “Esad: Endoscopic surgeon action detection dataset,” *arXiv preprint arXiv:2006.07164*, 2020.
- [182] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [183] L. Chen, J. Lu, Z. Song, and J. Zhou, “Part-activated deep reinforcement learning for action prediction,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 421–436, 2018.
- [184] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.
- [185] M. Wang, L. Yu, D. Zheng, Q. Gan, Y. Gai, Z. Ye, M. Li, J. Zhou, Q. Huang, C. Ma, Z. Huang, Q. Guo, H. Zhang, H. Lin, J. Zhao, J. Li, A. J. Smola, and Z. Zhang, “Deep graph library: Towards efficient and scalable deep learning on graphs,” *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
- [186] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, “Heterogeneous graph neural network,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 793–803, 2019.
- [187] B. Lin, Y. Sun, X. Qian, D. Goldgof, R. Gitlin, and Y. You, “Video-based 3d reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey,” *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 12, no. 2, pp. 158–178, 2016.

- [188] Z. Yuan, J. C. Stroud, T. Lu, and J. Deng, “Temporal action localization by structured maximal sums,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3684–3692, 2017.
- [189] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, “Dense-captioning events in videos,” in *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.
- [190] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*, pp. 740–755, Springer, 2014.
- [191] “Bcewithlogitsloss function.” <https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html>. Accessed: 2022-11-01.
- [192] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, TaoXie, J. Fang, imyhxy, K. Michael, Lorna, A. V, D. Montes, J. Nadar, Laughing, tkianai, yxNONG, P. Skalski, Z. Wang, A. Hogan, C. Fati, L. Mammana, AlexWang1900, D. Patel, D. Yiwei, F. You, J. Hajek, L. Diaconu, and M. T. Minh, “ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference,” Feb. 2022.
- [193] M. Broström, “Real-time multi-object tracker using yolov5 and deep sort.” https://github.com/mikel-brostrom/Yolov5_DeepSort_Pytorch, 2020.
- [194] M. Zitnik and J. Leskovec, “Predicting multicellular function through multi-layer tissue networks,” *Bioinformatics*, vol. 33, no. 14, pp. i190–i198, 2017.
- [195] C.-L. Zhang, J. Wu, and Y. Li, “Actionformer: Localizing moments of actions with transformers,” in *European Conference on Computer Vision*, pp. 492–510, Springer, 2022.
- [196] P. Zhao, L. Xie, C. Ju, Y. Zhang, Y. Wang, and Q. Tian, “Bottom-up temporal action localization with mutual regularization,” in *European Conference on Computer Vision*, pp. 539–555, Springer, 2020.
- [197] H. Su, W. Gan, W. Wu, Y. Qiao, and J. Yan, “Bsn++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 2602–2610, 2021.

- [198] Z. Qing, H. Su, W. Gan, D. Wang, W. Wu, X. Wang, Y. Qiao, J. Yan, C. Gao, and N. Sang, “Temporal context aggregation network for temporal action proposal refinement,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 485–494, 2021.
- [199] X. Liu, Y. Hu, S. Bai, F. Ding, X. Bai, and P. H. Torr, “Multi-shot temporal event localization: a benchmark,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12596–12606, 2021.
- [200] C. Zhao, A. K. Thabet, and B. Ghanem, “Video self-stitching graph network for temporal action localization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13658–13667, 2021.
- [201] Z. Zhu, W. Tang, L. Wang, N. Zheng, and G. Hua, “Enriching local and global contexts for temporal action localization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13516–13525, 2021.
- [202] J. Tan, J. Tang, L. Wang, and G. Wu, “Relaxed transformer decoders for direct action proposal generation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13526–13535, 2021.
- [203] Q. Wang, Y. Zhang, Y. Zheng, and P. Pan, “Rcl: Recurrent continuous localization for temporal action detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13566–13575, 2022.
- [204] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, “Zero-shot temporal action detection via vision-language prompting,” *arXiv e-prints*, pp. arXiv-2207, 2022.
- [205] T. Liu and K.-M. Lam, “A hybrid egocentric activity anticipation framework via memory-augmented recurrent and one-shot representation forecasting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13904–13913, 2022.
- [206] P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, “Pedestrian graph+: A fast pedestrian crossing prediction model based on graph convolutional networks,” *IEEE Transactions on Intelligent Transportation Systems*, 2022.
- [207] A. Kendall and Y. Gal, “What uncertainties do we need in bayesian deep learning for computer vision?,” *Advances in neural information processing systems*, vol. 30, 2017.
- [208] S. K. Manchingal and F. Cuzzolin, “Epistemic deep learning,” *arXiv preprint arXiv:2206.07609*, 2022.

- [209] I. Osband, Z. Wen, M. Asghari, M. Ibrahimi, X. Lu, and B. Van Roy, “Episodic neural networks,” *arXiv preprint arXiv:2107.08924*, 2021.
- [210] Z. Liu, Z. Miao, X. Pan, X. Zhan, D. Lin, S. X. Yu, and B. Gong, “Open compound domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12406–12415, 2020.
- [211] T. Panagiotakopoulos, P. L. Dovesi, L. Härenstam-Nielsen, and M. Poggi, “Online domain adaptation for semantic segmentation in ever-changing conditions,” in *European Conference on Computer Vision*, pp. 128–146, Springer, 2022.
- [212] V. VS, P. Oza, and V. M. Patel, “Towards online domain adaptive object detection,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 478–488, 2023.
- [213] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2962–2971, 2017.
- [214] C. Rosenberg, M. Hebert, and H. Schneiderman, “Semi-supervised self-training of object detection models,” 2005.
- [215] C. Chen, W. Xie, W. Huang, Y. Rong, X. Ding, Y. Huang, T. Xu, and J. Huang, “Progressive feature alignment for unsupervised domain adaptation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 627–636, 2019.
- [216] V. Lomonaco and D. Maltoni, “Core50: a new dataset and benchmark for continuous object recognition,” in *Conference on Robot Learning*, pp. 17–26, PMLR, 2017.
- [217] K. Corona, K. Osterdahl, R. Collins, and A. Hoogs, “Meva: A large-scale multiview, multimodal video dataset for activity detection,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1060–1068, 2021.
- [218] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, “Crowdnet: A deep convolutional network for dense crowd counting,” in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 640–644, 2016.
- [219] M. A. Hossain, K. Cannons, D. Jang, F. Cuzzolin, and Z. Xu, “Video-based crowd counting using a multi-scale optical flow pyramid network,” in *Proceedings of the Asian Conference on Computer Vision*, 2020.

- [220] F. Xiong, X. Shi, and D.-Y. Yeung, “Spatiotemporal modeling for crowd counting in videos,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5161–5169, 2017.
- [221] W. Liu, M. Salzmann, and P. Fua, “Estimating people flows to better count them in crowded scenes,” *CoRR*, vol. abs/1911.10782, 2019.
- [222] K. Chen, C. C. Loy, S. Gong, and T. Xiang, “Feature mining for localised crowd counting,” in *Bmvc*, vol. 1, p. 3, 2012.
- [223] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, “Privacy preserving crowd monitoring: Counting people without people models or tracking,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, 2008.
- [224] Y. Fang, B. Zhan, W. Cai, S. Gao, and B. Hu, “Locality-constrained spatial transformer network for video crowd counting,” in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 814–819, 2019.
- [225] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International Conference on Machine Learning*, pp. 6105–6114, PMLR, 2019.
- [226] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-image crowd counting via multi-column convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 589–597, 2016.
- [227] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao, “Crowd counting and density estimation by trellis encoder-decoder networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6133–6142, 2019.
- [228] H. Xiong, H. Lu, C. Liu, L. Liu, Z. Cao, and C. Shen, “From open set to closed set: Counting objects by spatial divide-and-conquer,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8362–8371, 2019.
- [229] I. Triguero, S. García, and F. Herrera, “Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study,” *Knowledge and Information Systems*, vol. 42, pp. 245–284, March 2015.
- [230] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.

-
- [231] S. Thrun and L. Pratt, *Learning to learn*. Springer Science & Business Media, 2012.
- [232] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov models: estimation and control*, vol. 29. Springer Science & Business Media, 2008.
- [233] M. Bazzi, F. Blasques, S. J. Koopman, and A. Lucas, “Time-varying transition probabilities for markov regime switching models,” *Journal of Time Series Analysis*, vol. 38, no. 3, pp. 458–478, 2017.
- [234] A. Agresti, *Categorical data analysis*, vol. 482. John Wiley & Sons, 2003.
- [235] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.
- [236] I. Visser, M. Speekenbrink, *et al.*, “depmixs4: an r package for hidden markov models,” *Journal of Statistical Software*, vol. 36, no. 7, pp. 1–21, 2010.