



**HAL**  
open science

## Human activity recognition in videos

Mohammed Guermal

► **To cite this version:**

Mohammed Guermal. Human activity recognition in videos. Computer Science [cs]. Université Côte d'Azur, 2024. English. NNT : 2024COAZ4015 . tel-04680002

**HAL Id: tel-04680002**

**<https://theses.hal.science/tel-04680002v1>**

Submitted on 28 Aug 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# THÈSE DE DOCTORAT

## Compréhension de l'activité humaine dans des vidéos

**GUERMAL Mohammed**

Centre Inria d'Université Côte d'Azur - STARS

**Présentée en vue de l'obtention  
du grade de docteur en :**

AUTOMATIQUE TRAITEMENT DU SIGNAL  
ET DES IMAGES d'Université Côte d'Azur

**Dirigée par :** Francois BREMOND

**Soutenue le :** 28 mai 2024

**Devant le jury, composé de :**

Thierry CHATEAU, Professeur, Université Clermont  
Auvergne

Bertrand LUVISON, Ingénieur de recherche, CEA  
LIST

Francois BREMOND, Professeur, Université Côte  
d'Azur

Ezio MALIS, Directeur de recherche, Centre Inria  
d'Université Côte d'Azur



# Compréhension de l'activité humaine dans des vidéos

Human activity recognition in videos

Jury :

Directeur de these  
Francois BREMOND, Professeur, Université Côte d'Azur

Rapporteurs

Thierry CHATEAU, Professeur, Université Clermont Auvergne  
Bertrand LUVISON, Ingenieur de recherche, CEA LIST

Examineurs

Ezio MALIS, Directeur de recherche, Centre Inria d'Université  
Côte d'Azur Sophia-Antipolis

*“The greatest challenge to any thinker is stating the problem in a way that will allow a solution.”*

Bertrand Russell

## Résumé

---

Mots clés : Reconnaissance d'activités humaines, apprentissage approfondie, Robotique, intelligence artificielle.

La compréhension des actions dans les vidéos est un aspect crucial de la vision par ordinateur avec des implications profondes dans divers domaines. Alors que notre dépendance aux données visuelles continue de croître, la capacité à comprendre et interpréter les actions humaines dans les vidéos est essentielle pour faire progresser les technologies dans la surveillance, les soins de santé, les systèmes autonomes et l'interaction homme-machine. La vision par ordinateur a connu d'énormes progrès avec l'avènement de méthodes d'apprentissage profond telles que les réseaux neuronaux convolutionnels (CNN) et plus récemment les transformers. Ces méthodes ont permis à la communauté de la vision par ordinateur d'évoluer dans de nombreux domaines tels que la segmentation d'image, la détection d'objets, la compréhension de scènes, etc. Cependant, en ce qui concerne le traitement vidéo, il reste encore limité par rapport aux images statiques. La reconnaissance des activités humaines repose sur une analyse vidéo approfondie. Dans cette analyse, il est essentiel de prendre en considération différents aspects de la vidéo, tels que les informations spatiales (comme la couleur RGB, la pose, la détection d'objets, etc.) ainsi que les informations temporelles. Il est ensuite nécessaire de combiner ces deux types d'entrées pour prédire avec précision l'activité humaine qui se déroule dans la vidéo. Dans cette thèse, nous nous concentrons sur la compréhension des actions que nous divisons en deux parties principales : la reconnaissance des actions et la détection des actions. Principalement, les algorithmes de compréhension des actions font face aux défis suivants : 1) l'analyse temporelle et spatiale, 2) les actions détaillées, et 3) la modélisation temporelle.

Cette thèse, introduit les différents défis liés à la reconnaissance des activités humaines. Nous présenterons également les méthodes et solutions existantes, en mettant en évidence leurs limites. Ensuite, nous exposerons notre propre travail et nos contributions dans ce domaine spécifique. En conclusion, nous discuterons des perspectives futures et des extensions envisageables pour nos solutions.



# Abstract

---

Keywords: Action understanding, Robot vision, Deep Learning, CNNs, Transformers.

Understanding actions in videos is a pivotal aspect of computer vision with profound implications across various domains. As our reliance on visual data continues to surge, the ability to comprehend and interpret human actions in videos is necessary for advancing technologies in surveillance, healthcare, autonomous systems, and human-computer interaction. Moreover, There is an unprecedented economical and societal demand for robots that can assist humans in their industrial work and daily life activities. Hence, understanding human behaviour and its activities would be very helpful and would facilitate development of such robots. The accurate interpretation of actions in videos serves as a cornerstone for the development of intelligent systems that can navigate and respond effectively to the complexities of the real world. In this context, advancements in action understanding not only push the boundaries of computer vision but also play a crucial role in shaping the landscape of cutting-edge applications that impact our daily lives. Computer Vision has known huge progress with the rise of deep learning methods such as convolutional neural networks (CNNs) and more lately transformers. Such methods allowed computer vision community to evolve in many domains such image segmentation, object detection, scene understanding and so on. However, when it comes to video processing it is still limited compared to static images. In this thesis, we focus on action understanding and we divide it into two main parts: **action recognition** and **action detection**. Mainly, action understanding algorithms faces following challenges : 1) **temporal and spacial analysis**, 2) **fine grained actions**, and 3) **temporal modeling**.

In this thesis we introduce with more details the different aspects and key challenges of action understanding. After that we are going to introduce our contributions and solution on how to deal with these challenges. We are going to focus mainly on recognising fine-grained action using spatio-temporal objects semantics and their dependencies in space and time, we are going also to tackle action detection in real-time and anticipation by introducing a new joint model of action anticipation and online action detection for a real life scenarios applications of action detection. We are going also to introduce a new method of efficiently training networks, specifically transformers and also a more efficient use of multi-modalities (RGB, Optical-Flow, Audio...). Finally, we will discuss some ongoing and future works. All our contributions where extensively evaluated on challenging benchmarks and outperformed previous works.





## *Acknowledgements*

- First of all, I would like to thank my supervisor Francois Bremond. Thank you for giving me the chance for pursuing my PhD. I also appreciate your patience in the numerous discussions. Your guidance really helps me in making critical decisions. In the last three years, I learnt many things from you. The most important one is to remain calm and positive in any situation.
- Secondly, I would like to acknowledge my PhD jury members. Thanks to my thesis reviewers, LUVISON Bertrand and CHATEAU Thierry, who kindly agreed to review my PhD manuscript. Also, thanks to Ezio Malis for serving as members of my thesis committee.
- I would like to give special thanks to RUI Dai. Your initial suggestions led me to be a curious man and take up academic research seriously, and provided me with help and guidance whenever needed.
- I want to express my gratitude to Université Côte d'Azur for sponsoring my doctoral thesis and offering essential resources and technical assistance for my research. A special acknowledgment goes to my colleagues at the STARS team in Inria for their support during my onboarding period. I am also deeply thankful to my friends, particularly Abid, Tomasz, Michal, Tanay, and Snehasis, for their unwavering support throughout my Ph.D. journey. Your presence has added vibrant and intriguing dimensions to my life in Sophia. I appreciate all of you for the meaningful discussions and encouragement over the past four years.
- I would also like to express a special thanks to my firends ( Youssef, Yahya, Ayoub, Abdellah, Souhail, Nazih, Soufiane, Othmane and Sanaa...) outside the lab, who helped me through times of difficulties and made me believe in myself while I was down and doubting my capacities, thank you all for the support and I hope you are all proud of me.
- Last but not least, I extend my gratitude to my family, who recommended that I pursue a doctoral degree following my master's studies. My parents, in particular, have consistently supported my life choices. Their support has been instrumental in enabling me to pursue studies in France and successfully complete my Ph.D. research.


Thank you all for the support



# Contents

<b>Abstract</b>	<b>vii</b>
<b>Acknowledgements</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem statement	2
1.1.1 Action recognition	3
1.1.2 Action detection	4
1.2 Applications	5
1.3 Scientific challenges	6
1.3.1 Fine-grained activity recognition	8
1.3.2 Time handling	8
1.3.3 Video representation	8
1.3.4 Dataset generalization	9
1.3.5 Multi-modalities	9
1.3.6 Subtle activity recognition	9
1.4 Contributions	10
1.4.1 Semantic reasoning for fine-grained action recognition	11
1.4.2 Temporal reasoning for real-world scenarios action detection	11
1.4.3 Multi-modal and multi-dataset training	11
1.4.4 Subtle activity recognition	11
1.5 Thesis structure	12
<b>2 Related work</b>	<b>13</b>
2.1 Basic concepts	13
2.2 Methods prior to deep learning	15
2.3 Human object interaction actions (HOI)	15
2.4 Temporal modeling and online action detection	16
2.5 Efficient transformers and cross-dataset training	21
2.6 Multi-modal fusing	22
<b>3 THORN: Temporal Human-Object Relation Network for Action Recognition</b>	<b>25</b>
3.1 Introduction	26
3.2 Related work	28
3.2.1 3D-CNNs	29
3.2.2 Graph convolutions	29
3.3 Proposed method	30
3.3.1 Visual encoder	31
3.3.2 Object representation filter	31
3.3.3 Object relation reasoning module	32
<b>Graph reasoning</b>	<b>33</b>
<b>TCN</b>	<b>34</b>

3.3.4	Predictions . . . . .	34
3.4	Experiments . . . . .	36
3.4.1	Ablation study . . . . .	37
3.4.2	Comparison with the State-of-the-Art . . . . .	38
3.4.3	Qualitative study . . . . .	39
3.5	Conclusion . . . . .	42
<b>4</b>	<b>JOADAA: joint online action detection and action anticipation</b>	<b>45</b>
4.1	Introduction . . . . .	45
4.2	Related work . . . . .	48
4.3	Proposed method . . . . .	49
4.3.1	Past processing block . . . . .	50
4.3.2	Anticipation prediction block . . . . .	51
4.3.3	Online action prediction block . . . . .	51
4.4	Experiments . . . . .	52
4.4.1	Datasets . . . . .	53
4.4.2	Implementation details . . . . .	53
4.4.3	Comparison with the SoTA . . . . .	53
	OAD comparison on the simple dataset (THUMOS'14) . . . . .	53
	OAD comparison on densely annotated datasets . . . . .	54
	OAD comparison using off-line methods . . . . .	55
	AA SoTA comparison . . . . .	55
4.4.4	Ablation study . . . . .	55
	Ablation on the past processing block . . . . .	55
	Ablation on the action anticipation module . . . . .	56
	Ablation on the OAD prediction layer . . . . .	57
4.4.5	Qualitative analysis . . . . .	57
4.5	Conclusion . . . . .	58
<b>5</b>	<b>Robust and Efficient Multimodal Multi-dataset Multitask Learning</b>	<b>59</b>
5.1	Introduction . . . . .	59
5.2	Related work . . . . .	62
5.2.1	Transfer learning . . . . .	62
5.2.2	Parameter efficient task adaptation . . . . .	62
5.2.3	Multimodal learning . . . . .	63
5.3	CM3T framework . . . . .	63
5.3.1	Basic concepts . . . . .	63
	Adapters . . . . .	63
	Prefix tuning . . . . .	64
5.3.2	Choosing a pretrained model . . . . .	64
5.3.3	Fine-tuning or using Adapters . . . . .	65
5.3.4	Adding other modalities . . . . .	68
5.4	Experiments . . . . .	69
5.4.1	Datasets . . . . .	69
5.4.2	Training details . . . . .	69
5.4.3	Results and Observations . . . . .	71
5.4.4	SoTa comparison . . . . .	71
	Baseline comparison . . . . .	71
	Comparison with traditional adapters . . . . .	71
	SoTa comparison on Epic-Kitchen 100 . . . . .	72
	SoTa comparison on Udiva v0.5 and MPIIGroupInteraction . . . . .	72

	Cross attention module . . . . .	72
	Time and resources . . . . .	72
5.4.5	Ablation studies . . . . .	72
	MHVA / PT . . . . .	73
	Different backbone . . . . .	73
5.4.6	Cross-attention adapter behaviour with different modalities at different levels . . . . .	73
5.5	Adding adapters to cross-attention adapters . . . . .	74
5.6	Conclusion . . . . .	75
<b>6</b>	<b>MultiMediate'23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions</b> . . . . .	<b>77</b>
6.1	Introduction . . . . .	77
6.2	Related work . . . . .	78
	6.2.1 Engagement estimation . . . . .	78
	6.2.2 Bodily behaviour recognition . . . . .	78
6.3	Challenge description . . . . .	79
	6.3.1 Engagement estimation task . . . . .	79
	6.3.2 Bodily Behavior Recognition Task . . . . .	80
6.4	Experiments and Results . . . . .	81
	6.4.1 Engagement estimation . . . . .	81
	Approach . . . . .	81
	Results . . . . .	82
	6.4.2 Bodily behaviour recognition . . . . .	82
	Approach . . . . .	82
	Results . . . . .	83
6.5	Conclusion . . . . .	83
<b>7</b>	 <b>Uncovering Near-Future Abnormal Behaviour via Human Interactions in Real-world Videos</b> . . . . .	<b>85</b>
7.1	Introduction . . . . .	86
7.2	Related Work . . . . .	88
	Supervised: . . . . .	89
	Unsupervised: . . . . .	89
	Semi-supervised: . . . . .	89
7.3	Preliminaries . . . . .	91
7.4	Proposed Method . . . . .	91
	7.4.1 Feature Encoder . . . . .	92
	7.4.2 Interaction Modules (TIM/OIM) . . . . .	92
	7.4.3 Co-Attention Encoder(CAE) . . . . .	93
	7.4.4 Anticipation Decoder . . . . .	94
7.5	Experiments . . . . .	94
	7.5.1 Criminal Human Behaviour dataset (CHB) . . . . .	95
	7.5.2 Evaluation Metrics . . . . .	95
7.6	Preliminary results . . . . .	96
7.7	State-of-the-art Comparison . . . . .	96
7.8	Conclusion . . . . .	96

<b>8 Discussion and Future Work</b>	<b>97</b>
8.1 Contribution summarization . . . . .	97
8.2 Limitations and Perspectives . . . . .	100
8.2.1 fine-grained activity recognition . . . . .	100
8.2.2 Action detection . . . . .	101
Scene and features encoding . . . . .	102
Towards unsupervised action detection . . . . .	102
Action detection for real-world scenarios . . . . .	103
<b>Bibliography</b>	<b>105</b>

# List of Figures

1.1	Video data growth and availability due to different sources, internet videos, surveillance cameras, media, etc. All this data makes it easier to study action understanding in videos. . . . .	3
1.2	In action detection (left) the goal is to map an untrimmed video into different time steps for different actions. While in action recognition (right) the objective is to categorize (give a class label to) a clip. . . . .	4
1.3	This figure is from the Epic-Kitchen55 [52] dataset. The frame shown belongs to an example of a fine-grained action <i>wash plate</i> . To recognize such action it is important to focus on different semantics such as the relevant objects <i>plate, hands, and tap</i> and also to model their relations and dependencies. . . . .	7
2.1	Different techniques for integrating information across the temporal dimension in [108]. Convolutional, normalization, and pooling layers are denoted by red, green, and blue boxes, respectively. In the Slow Fusion model, the columns depicted in the illustration share parameters.	17
2.2	Key action recognition approaches based on Deep Learning aim to model temporal information in videos. These approaches utilize (a) 2D CNN (left), (b) 2D CNN + RNN (middle), and (c) 3D CNN (right) to consolidate temporal information for action classification. . . . .	17
2.3	This figure illustrates how CNNs work on images, at early stages we learn low level features such edges and texture and deeper in the network we learn more explicit semantics. This showcase strength of CNNs on image like data. . . . .	18
2.4	Class-specific visualization results from ViT with attention maps in contrast to class-activation maps in CNNs. Source <a href="https://www.researchgate.net/publication/355693348_Transformers_in_computational_visual_media_A_survey">https://www.researchgate.net/publication/355693348_Transformers_in_computational_visual_media_A_survey</a> . . . . .	20
2.5	Example of lateral fusion in SlowFast Network [68]. The two streams could either take a high frame rate and low frame rate such in the figure or input different modalities. . . . .	23
3.1	An example of the Human-Object Interactions of <i>wash plate</i> in a first-view video. Green arrows represent interactions at the same time step (i.e., spatial relation) while black arrows represent interactions across time. In practice, the model captures all the objects detected. For simplicity, here we highlight only the relevant objects to <i>wash plate</i> . . .	28



3.2	THORN architecture contains three main components: (1) a <b>Visual encoder</b> (i.e., X3D) encodes the input RGB clip into a primary spatio-temporal representation. (2) The obtained representation is fed to the <b>Object Representation Filter</b> , which maps the previous representation into object-class representation. To ensure a discriminative object representation, an object classifier is added on top of the object-class representation. This classifier is trained with the pseudo-object ground truth provided by an object detector. (3) The object-class representation is also sent to the <b>Object Relation Reasoning</b> module to model the temporal-object relation in a dissociated manner. Finally, two classifiers are used to predict the verbs and nouns relevant to the action. . . . .	30
3.3	Schema of our Object Representation Filter (ORF). The input is the feature map from the 3D encoder reshaped to $T \times H'W'D$ and the duplicated $C_o$ times, where $C_o$ is the number of classes. Finally, we have a representation specific to each object class. . . . .	31
3.4	Overview of one layer of the Object Relation Reasoning module, using a graph architecture [183]. The input is a graph representation between different classes and the output is an updated representation of the graph. The $\times N_{block}$ stands for the number of blocks used in total, while the $\times 3$ at the bottom in blue stands for the number of used multi-head attentions. . . . .	35
3.5	Accuracy improvement on nouns (right) and verbs (left) w.r.t X3D. . . . .	39
3.6	Example of the learned adjacency matrix of the action from Epic-Kitchen55 dataset. We notice a strong correlation between the classes <i>knife</i> and <i>water</i> for the action <i>wash knife</i> . Thus, we are able to collect high inter-class relation to recognize the right verb and its relevant objects. Moreover, the irrelevant classes such as <i>fish</i> are not activated, showing robustness of the learned attention. . . . .	41
3.7	Example of action <i>wash leaf</i> . the highest activated classes were leaf and tap and when inferring the class activation map we can see that most activated pixels are around the objects of interest. Hence, the features extracted are more significant, which makes it easier to predict the right action. . . . .	42
3.8	Example of action <i>put leaf</i> . In this example the most activated object was leaf and its activation map shows that the focused-on pixels actually belongs the leaf, proving the strength and robustness of our approach. . . . .	42
3.9	The action in this figure is <i>mix meat</i> , and looking at the figure we notice that the highlighted pixels are the ones corresponding to the spatula and the meat. Therefore, it is easier to predict the right action. . . . .	43
4.1	An example of human non-sequential dependencies. For instance, the actions <i>RUN</i> and <i>OneHanded Catch</i> are highly correlated but distant in time. Also the same start action <i>RUN</i> can lead to many different actions and scenarios. Therefore, it is very hard for online action detection or action anticipation to detect such relations without access to the future. In JOADAA, we propose to tackle this limitation by introducing a pseudo-future information by combining action anticipation and online action detection in the same task. . . . .	46

4.2	Proposed JOADAA architecture with three units i) Past processing, ii) Anticipation prediction, and iii) Online Action prediction. Each stage is highlighted by a color for better understanding. . . . .	49
4.3	Action anticipation accuracy improvement on six actions w.r.t. TesTra model. This is performed on the Multi-THUMOS dataset, using 4 frames as anticipation length. . . . .	57
5.1	Representation of existing parameter efficient transfer learning techniques and CM3T. Backbones pretrained using self-supervised learning provide good general features, thus all methods of fine-tuning work well. In the case of supervised learning, adapters fail to perform well (shown in red) and CM3T is introduced to solve this (shown in green.) . . . . .	60
5.2	the left of the figure, the standard Transformer is used with an additional adapter layer, added after each sub-layer and before adding the skip connection back. The output of the adapter layer is then forwarded to the layer normalization. . . . .	64
5.3	Adjusting (top) fine-tunes all Transformer parameters (illustrated by the red Transformer box) and necessitates preserving a complete model replica for every task. Prefix-tuning (bottom), wherein the Transformer parameters are fixed, and optimization exclusively targets the prefix (depicted by the red prefix blocks). Source <a href="https://arxiv.org/pdf/2101.00190.pdf">https://arxiv.org/pdf/2101.00190.pdf</a> . . . . .	65
5.4	Detailed architecture of CM3T. Coloured parts are the ones that are fine tuned and the rest are frozen. It has three separate blocks added to it which are shown in three different colours. Prefix tuning is complicated to show in detail, so only a diagram is shown. Comparing eq 5.2 and 5.4 gives how the upscaling and downscaling weights are computed. The rest of the details are described in section 5.3 . . . . .	66
6.1	Snapshots of scenes of a participant in the NOXI corpus being disengaged (left), neutral (center) and highly engaged (right). . . . .	80
6.2	Setup of the MPIIGroupInteraction dataset. Reproduced with permission from the authors of [150]. . . . .	81
7.1	The differences between anomaly anticipation with offline and online anomaly detection task, where $f(\theta)$ is the functionality of the respective methods. . . .	86
7.2	Illustration of complexity in abnormal human behavior. Notice the three different cases of Interactions with divergent spatio-temporal cues. Abnormal human-to-human interactions (e.g. arrest) have often significant appearance and motion change whereas human-to-object interactions (e.g. shoplifting) are often subtle. However, there can be abnormal human-to-(object & human) interactions like <i>protest</i> that have a unique spatiotemporal blend with large people density. . . . .	87
7.3	Comparison of previous anticipation-based methods with ours in early trend modeling. Previous methods consider only scene-level features ( <i>i.e. from the whole frame</i> ) to encode joint spatiotemporal embeddings, thereby they have a partial understanding of complex abnormal behavior. In contrast, our dissociatively learn the scene-level temporal consistencies and object-level spatial interaction to obtain a better understanding of early trends. . . . .	88

7.4	<b>Spatial Interaction aware Transformer (SIaT):</b> It has three key modules <i>i.e.</i> (i) Drift Augmenter, (ii) Human-Scene Selective Transformer, (iii) Text Inducer, and a MLP Detector to detect anomalies in weak-supervision. This figure shows the training regime of HSDaT. However, during inference, the modules indicated by dashed lines can be excluded for reduced overhead. . . . .	91
7.5	Overview of architecturally identical <b>Temporal Interaction Module (TIM)</b> and <b>Object Interaction Module (OIM)</b> . Note that the figure is color coded. Here, $F_{txt}$ , $F_O$ are the temporal-pooled textual and object mask features and $F_O$ is the spatial pooled object mask feature. . . . .	93
7.6	<b>ED-Crime Properties:</b> (a) Abnormal video distribution and train-test sample count on a K-fold evaluation, (b) Diversity of abnormal categories, (c) Human, scene-related anomaly distribution, and (d) long and short anomalies distribution. . . . .	95

# List of Tables

3.1	Ablation study on different settings. This evaluation is on the EPIC-KITCHEN dataset. Temporal nodes means using the final output of X3D of size $T \times 2048$ to create nodes, while spatio-temporal nodes means using a mid layer of size $T \times 7 \times 7 \times 432$ with more spatial information. Finally ADJ-matrix stands for using the adjacency matrix for predicting the verbs instead of using only nodes for nouns and verbs. . . . .	35
3.2	Ablation study on fusing the scores of THORN with the scores from the object detector (Faster RCNN). This evaluation is on EPIC-KITCHEN dataset. Fusing both scores brings significant improvement on top-1 accuracy. For the object detector, we use an average pooling on all the video clip frames object detection scores and add a threshold of 0.3 . . . . .	38
3.3	Comparing THORN model with other state-of-the-art methods on the validation set. Even though some of these comparisons are not fair since these models are using multi-modalities, we still hold the overall best accuracy, which shows the strength of our model . . . . .	39
3.4	Comparing THORN model with other state-of-the-art methods on EGTEA Gaze+ split1. We hold the best accuracy on actions . . . . .	39
4.1	State of the art comparison for OAD on THUMOS'14, Multi-THUMOS, and CHARADES. Due to the lack of available OAD methods for CHARADES and Multi-THUMOS datasets, we compare also with two offline methods PDAN and MSTCT, adapted to an online setting. . . . .	54
4.2	Comparison with SoTA for the action anticipation task. 1, 2, 4, and 6 represent the number of anticipated frames. We notice that our method is more robust w.r.t. the number of anticipated frames compared to other methods where accuracy drops dramatically. . . . .	54
4.3	Effect of action anticipation prediction and <b>online action detection</b> using long-short-term knowledge. 1, 2, 4, and 6 are the number of anticipated frames. Best viewed in color. . . . .	55
4.4	Results of using only short-term past information on multiple datasets for <b>online action detection</b> and action anticipation. 2, 4, and 6 are the number of anticipated frames. . . . .	55
4.5	Comparison of JOADAA with LSTR method using long-past information. JOADAA is more robust to utilize long-past information. . . . .	56
4.6	Comparing two techniques for past information processing. We use a transformer encoder and a set of LSTM blocks with a convolution layer. . . . .	56
4.7	Analyzing the JOADAA behavior with and without action anticipation. . . . .	57
4.8	Effect of fusing local and global information on OAD. FC stands for fully-connected layer. As expected capturing different type of dependencies provides better results. . . . .	57

5.1	SOTA comparison on EK100. We show main comparisons in same colors. . . . .	70
5.2	SoTa comparison on UDIVA 0.5 . . . . .	71
5.3	SoTa comparison on MPIIGroupInteraction . . . . .	71
5.4	Ablation on multimodality attention. The abbreviations used are MHVA: multi-head vision adapters, PT: prefix tuning, CAA: cross attention adapters, MmCA: multi-modality cross-attention. These results are reported on Epic-Kitchen 100. . . . .	73
5.5	Ablation study on different components of our proposed architecture. . . . .	73
5.6	Results of our method using different backbone. Experiments were done on Epic-Kitchen 100 . . . . .	73
5.7	Results for ablation study in section 5.4.6. The entries show cross attention removed from a particular block. Block 1 is closest to input and Block 4 is the last block before classification head. . . . .	74
5.8	Result for adding adapters to cross-attention adapters . . . . .	74
6.1	Social interaction datasets with engagement annotations, excluding MOOC and school settings and children as participants. <i>Screen</i> indicates whether the interaction was screen-mediated, <i>Group size</i> the number of humans per interaction, <i>length</i> the total duration of interactions, and <i>part.</i> the total number of human participants. . . . .	79
6.2	Concordance correlation coefficient (CCC) of our baseline on engagement detection validation and test sets. . . . .	83
6.3	Validation and test results for the random baseline and different variants of the Video Swin Transformer. . . . .	84
7.1	State-of-the-art comparisons on UCF-C, IITB-C, and ED-C datasets. . . . .	96

*Dedicated to ...*



## Chapter 1

# Introduction

Artificial Intelligence encompasses the field of computer vision, aiming to replicate aspects of the human visual system and empower computers to extract meaningful information from diverse inputs such as images and videos. With the rise of smartphones and omnipresence of cameras generating vast amounts of video and media content daily see Figure 1.1, the importance of video comprehension and analysis has surged. Consequently, delving into video analysis has become a pivotal area of research within computer vision. Video analysis involves a comprehensive approach to interpreting scenes, identifying objects, discerning actions, events, attributes, and grasping concepts from a sequence of frames constituting a video. Despite the remarkable achievements of deep learning techniques in various computer vision tasks, such as image classification and object detection, video understanding remains a challenging frontier with substantial room for improvement. Among the facets of video understanding, the analysis of actions within a video stands out as one of the most crucial and complex tasks. Notably, human presence plays a prominent role in videos, with statistics indicating that 35%, 34%, and 40% of pixels in movies, TV, and YouTube videos are associated with humans. Consequently, delving into the study of human actions and behavior within videos becomes paramount for comprehending their content. Action understanding, as a vital component of video analysis, contributes significantly to the advancement of real-world applications, including smart home systems, sports analysis, and human-robot interaction. On a daily basis, humans effortlessly and thoroughly perform a range of activities. We have the ability to understand and interpret these activities considering not only the context around us, but also the subtle gestures of other people. However, when it comes to analyzing situations, computers underperform humans. Toward better-performing robots and computers, researchers in computer vision have been continuously working to develop solutions that not only mimic, but potentially even surpass human capabilities in action analysis. This endeavor represents a step towards achieving a future where cutting-edge technologies like advanced robots, self-driving cars, and smart cities come together to enhance our daily lives.

In this domain of video and action understanding, **action recognition** plays a pivotal role, it aims to classify and categorize actions in pre-segmented or trimmed videos containing one action. As far as this thesis is concerned, trimmed video refers to clips with only action in it, therefore moments before the action or after are removed. Nevertheless, in real-life scenarios, videos are untrimmed containing instances of the actions as well as other moments before and after acting as background. Moreover, these videos are long videos containing many actions instances as well as co-occurring actions. This leads to another focus of this thesis, which is **action detection**. Temporal action detection involves the skill to identify and categorize action instances within specific time intervals. Recently, there has been a notable



surge in interest in this task due to its capacity to furnish insights into the nature of actions and their temporal occurrences. The difficulty arises from the visual resemblance between the moments immediately preceding or following an action and the actual commencement or conclusion of the action, posing a significant challenge in accurately localizing action intervals.

Whether it is **action recognition** or **action detection** most existing methods target high-level semantics and video actions with a sparse set of actions. However, as explained earlier, actions usually are dense in the wild, and also low-level action (**fine-grained activities**) represents the majority of human daily activities and has many interesting applications such as collaborative robots. In this thesis, we focus on video analysis that targets fine-grained actions and videos with sparse as well as dense occurrences of actions. Finally, although researchers have explored the temporal action detection task under both full and limited supervision settings, the methods employed for action detection in videos with densely occurring actions continue to heavily lean towards full supervision. This preference arises due to the complicated temporal relationships existing among action instances, densely populated action regions, and the multitude of action categories within videos. Therefore, this thesis exclusively focuses on the investigation of fully supervised action detection methods, with the objective of predicting action labels for every frame in a video. In this thesis, our research is not limited to only, the coarse and fine-grained activities, we will also shed light on another type of action recognition, which aims at understanding human behaviors in social interactions. This type of human understanding goes a long way into building assisting robots, and has many applications such human-robot interactions, class-teaching... .

Finally, in this thesis, our focus lies in finding ways and techniques to overcome existing limitations and challenges in the field of recognition of human actions. (Section 1.1) introduces the problem statement and different aspects of action understanding, mainly: action recognition (classification) and action detection. (Section 1.2) introduces the applications and impacts of video understanding in real life, while (Section 1.3) discusses key challenges within human activity recognition. (Section 1.4) briefly presents our contributions. (Section 1.5) retraces the structure of the thesis.

## 1.1 Problem statement

Action classification and action detection are the foundational components of action understanding in the domains of computer vision and video analysis. Both tasks play a pivotal role in the fields of computer vision and video analysis. A video that has been pre-segmented or otherwise edited to separate individual actions is usually the starting point for action recognition, which aims to classify or categorize specific activities within the video. Action detection broadens its scope by identifying actions and localizing the temporal bounds inside untrimmed movies, thereby catching the exact moments when certain actions take place. Figure 1.2 serves as a visual aid, distilling the core of both action recognition and action detection tasks, to provide the reader with a brief summary of these tasks. Through this doctoral thesis, we will explore new approaches, state-of-the-art algorithms, and cutting-edge strategies in this academic endeavor with the goal of improving the precision, resilience, and effectiveness of action recognition and detection systems. Our contributions are

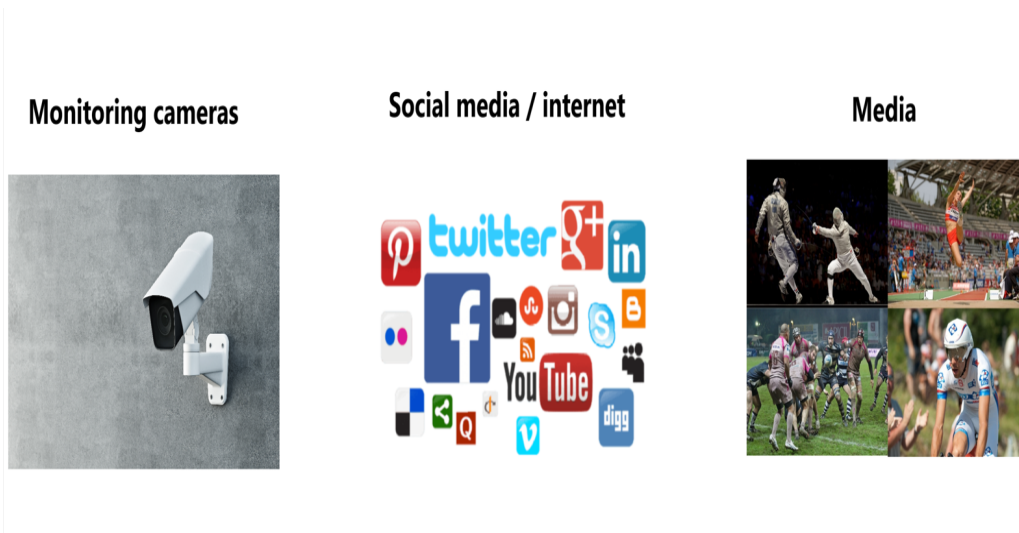


FIGURE 1.1: Video data growth and availability due to different sources, internet videos, surveillance cameras, media, etc. All this data makes it easier to study action understanding in videos.

centered around enhancing the state-of-the-art in action understanding. Our primary focus lies in extracting hidden cues essential for decoding videos. Within the scope of this thesis, an array of innovative solutions is proposed, each designed to unravel distinct layers of complexity. One facet involves the extraction of human interactions within their environmental milieu, deciphering their significance in the larger context. Furthermore, our exploration extends to deciphering crucial interdependencies across varying temporal intervals, encompassing past, present, and future frames, thereby constructing a comprehensive understanding. Digging even deeper, we aim at decoding intricate social cues embedded within human bodily behaviors and nuanced actions, unraveling subtleties that often remain veiled. Our work extends to bridge the gap between theoretical advancements and their tangible application in real-world industrial scenarios. This synthesis between theory and practicality stands as a testament to our commitment to not only advancing conceptual frameworks but also implementing them in pragmatic settings, enriching the industry landscape with our insights and innovations.

### 1.1.1 Action recognition

Action recognition or action classification is the task of assigning labels to video clips. These video clips are usually trimmed (pre-segmented: clips contain only actions with no background), they are usually short (around 10 seconds), and at a frame rate of 30 FPS (frames per second); still, these settings could vary depending on the data structure. Action classification in videos comes with several challenges, including variability in lighting conditions, background clutter, occlusions, and the need to distinguish between subtle action differences. Furthermore, handling long-duration videos can be computationally intensive and require efficient methods for temporal modeling. Action categories vary from simple activities in the form of a verb (e.g. running, jumping, clapping...) to more complex and fine-grained ones, verb+noun (cutting bread, picking trash, opening a door). Additionally, a video clip could contain more than one action; this task is called multi-label video classification.

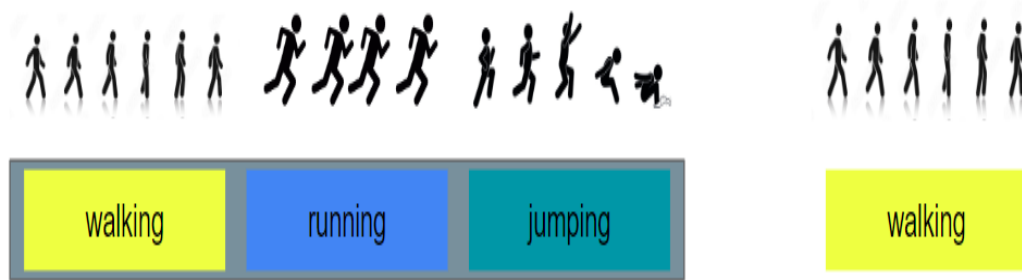


FIGURE 1.2: In action detection (left) the goal is to map an untrimmed video into different time steps for different actions. While in action recognition (right) the objective is to categorize (give a class label to) a clip.

In conclusion, the classification of actions in videos is a crucial task in computer vision with a myriad of practical applications. It involves the identification of human actions within video data, making it a valuable tool for enhancing security, enabling natural human-computer interaction, improving sports analysis, and providing personalized content recommendations. Advances in deep learning coupled with the availability of large annotated datasets have significantly improved the accuracy and applicability of action classification models in recent years, making them a critical component of modern computer vision systems.

### 1.1.2 Action detection

Action detection is an extension of action recognition. Action detection is the task of giving labels to all actions present in a video clip, as well as their temporal boundaries (starting and ending time). Action detection itself has different aspects. First, we mention *action localization* where video clips are sparsely annotated, as in the case of [96]. The other aspect of action detection is called *temporal action detection or segmentation*, where video clips contain fine-grained and densely annotated actions, such as [247]. Many other tasks are related to action detection; for this thesis, we focus on:

- **Online action detection:** Online action detection involves the real-time identification and tracking of actions as they unfold within a video sequence. Unlike offline action detection, where the entire video is available for analysis, online action detection requires models to make predictions in real-time, introducing challenges associated with handling partial information, coping with occlusions, and adapting to variable frame rates. The dynamic nature of real-world scenarios makes online action detection a challenging yet crucial task, with applications ranging from video surveillance to human-computer interaction.
- **Action anticipation:** Action anticipation, as an extension of online action detection, takes the task a step further by involving the prediction of future actions before they actually occur. In this task, models not only recognize ongoing actions in real-time but also anticipate and prepare for potential future actions based on current contextual cues. This proactive element adds an intriguing layer to the challenges posed by online action detection, opening up avenues for applications in diverse fields such as autonomous systems,

human-robot interaction, and sports analytics. The intersection of online action detection and action anticipation represents a compelling area of research aimed at advancing the capabilities of intelligent systems in dynamic environments.

Both these subtasks of action detection are very interesting, due to their many applications in real-world scenarios.

In this thesis, we are going to focus on both aspects of action understanding, action recognition and detection, to capitalize on the challenges of each of them, and through our contributions to provide some useful solutions on how to handle these challenges and tasks.

## 1.2 Applications

Action understanding and video analysis are very important, as it can be of great benefit to humans in their daily life. In fact, it has many real-life applications:

### **Video Surveillance:**

In our modern society, surveillance cameras are much more commonplace than ever. This pervasiveness is primarily related to the critical role that safety and security play in our daily activities. The implementation of surveillance systems has emerged as a crucial instrument in protecting not only our physical assets but also the well-being of humans within a variety of situations as society struggles with the ever-evolving panorama of potential threats and vulnerabilities such as *vandalism, robberies, violence*. . . . Recently, the effectiveness of security cameras has greatly increased due to the incorporation of cutting-edge technology, notably action recognition algorithms. By enabling real-time analysis and comprehension of video feeds from surveillance cameras, these algorithms have paved the way to a new era of surveillance capabilities. The ability to recognize and analyze human activity in these video streams has enormous potential to improve security and safety. In conclusion, action recognition algorithms have further helped improve security cameras. These algorithms can provide real-time analysis and help understand videos from surveillance cameras.

### **Human-robot interaction:**

The industrial landscape has seen an unprecedented increase in demand for intelligent robots, signaling a paradigm shift in how we see automation and human-machine collaboration. The main goal of these sophisticated robotic systems, which have the ability to supplement or perhaps completely replace human involvement in some tasks, is at the center of this technological revolution. The desire for higher efficiency, improved precision, and increased safety in a wide range of industrial processes is one of the causes driving this paradigm change. These advanced robots, empowered by action recognition algorithms, exhibit a significant ability to read a variety of human cues. These robots have a unique ability to explore the world of human emotions, using complex algorithms to recognize and react to subtly expressed emotional cues in body language, facial expressions, and vocal intonation. A new era of human-robot collaboration has begun as a result of this newly discovered depth of interaction, which gives these machines a certain level of empathy and responsiveness. Practical applications of action recognition algorithms in the context of human-robot interaction have already begun to reshape the industrial landscape. A notable example of this transformative trend can be found in Amazon warehouse

robots, which operate in concert with human personnel to optimize the order fulfillment process. These robots, equipped with advanced action recognition capabilities, seamlessly navigate the warehouse environment, interpreting the actions and needs of human workers. They work collaboratively, efficiently locating and transporting items to human pickers, thereby streamlining the order-processing workflow.

### **Healthcare:**

The COVID-19 pandemic, with its overwhelming influx of patients and the resultant strain on healthcare resources, has increased the urgency of innovative solutions that can alleviate the burden on healthcare professionals. Action understanding algorithms offer a multifaceted approach to enhance healthcare capabilities. One notable application resides in the realm of surgical robotics, where these algorithms can be harnessed to create surgical robots capable of seamlessly collaborating with human surgeons during intricate medical procedures. These robotic assistants have the potential to improve surgical precision and reduce the margin of error, ultimately improving patient outcomes. Moreover, the application of action understanding algorithms extends to the realm of patient monitoring, particularly among the elderly population. By deploying such algorithms, healthcare providers can institute robust home monitoring systems that give them real-time insights into the well-being and conditions of elderly patients. This proactive approach enables physicians to better assess and manage the healthcare needs of their elderly patients, providing a vital lifeline for those who prefer to age in the comfort of their homes while receiving high-quality medical attention. Another dimension of action-understanding algorithms is their profound impact on the understanding of patients with autism spectrum disorders. These algorithms, because of their ability to decode actions and emotions, facilitate a deeper understanding of the behavior and emotional states of autistic children. This newfound understanding translates into more effective and tailored therapeutic interventions, thereby enhancing the quality of life for autistic individuals and their families. In times of crisis, such as pandemics or other emergencies, the deployment of action understanding algorithms equips hospitals with a heightened ability to assess and manage critical situations. By augmenting the decision-making process with real-time insights derived from these algorithms, healthcare institutions can manage crisis scenarios more effectively, optimize resource allocation, and ensure that patients receive the care they need when they need it. Thanks to such solutions, hospitals can better assess situations in crisis times, older people could be home monitored, and autistic children could also be better understood and have an easier life.

## **1.3 Scientific challenges**

In recent years, the field of image analysis has undergone a profound transformation, largely driven by the extraordinary progress witnessed in deep learning algorithms. These algorithms have emerged as formidable tools, showcasing enhanced robustness and delivering exceptional performance across an expansive spectrum of tasks. From the fundamental challenges of object detection and precise image classification to the intricate domain of semantic segmentation, these advances have significantly boosted the capabilities of image analysis techniques [83]. However, despite these remarkable strides in image analysis, the domain of video processing presents a distinct set of challenges and hurdles that remain largely unresolved. Videos, characterized by their dynamic and temporal nature, introduce a layer of

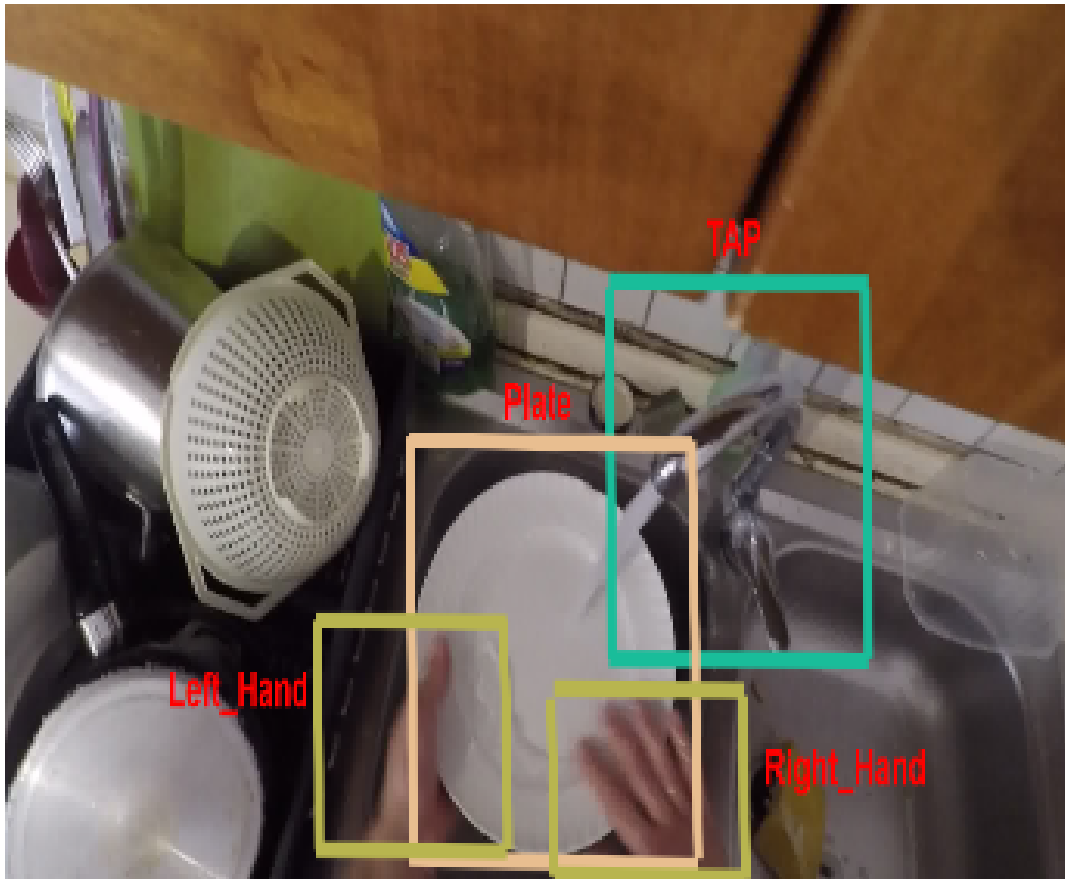


FIGURE 1.3: This figure is from the Epic-Kitchen55 [52] dataset. The frame shown belongs to an example of a fine-grained action *wash plate*. To recognize such action it is important to focus on different semantics such as the relevant objects *plate*, *hands*, and *tap* and also to model their relations and dependencies.

complexity that demands a deeper and more comprehensive understanding than what current methodologies offer. The inherent dynamics within videos pose multifaceted challenges that necessitate innovative approaches and fresh ideas. The complexities embedded within the temporal sequences, the intricacies of motion, and the evolving relationships between elements within the visual data, demand novel strategies and creative solutions to be effectively addressed. This need for a paradigm shift in tackling the challenges of video processing serves as a call to action for researchers and innovators. It beckons the exploration of unconventional methodologies and the development of inventive solutions that can effectively navigate the complexities inherent in video data. In essence, while deep learning has revolutionized image analysis, its translation to video processing requires a divergence toward novel concepts and innovative techniques. Embracing these complexities as opportunities for innovation, the field stands at the precipice of exploration, awaiting fresh ideas and inventive solutions to unlock the full potential of video analysis and processing.

### 1.3.1 Fine-grained activity recognition

For a comprehensive understanding of human activities, it is crucial to decode the hidden details that define them. Unlike simpler actions, such as jumping or drinking, fine-grained actions present a unique challenge due to their complexity, demanding a higher level of contextual awareness. These nuanced actions often exhibit minimal inter-class variance, exemplified by activities such as drinking from a bottle or drinking from a cup. Consequently, accurate detection and recognition of these actions require a deep understanding of the surrounding human environment.

The significance of grasping these fine-grained details goes beyond mere recognition. Consider, for instance, a scenario where prior knowledge of actions involving picking up an onion and wielding a knife has been acquired. Armed with this context, a predictive model becomes better equipped to identify the action of cutting an onion, as it is now more likely to occur. This underscores the importance of handling spatial and temporal semantic dependencies with precision.

In the forthcoming thesis, our primary focus will revolve around the intricate realm of fine-grained actions. We will endeavor to construct a robust framework that addresses the intricate web of dependencies that govern these actions, both in the spatial and temporal dimensions. Through this endeavor, we aim to shed light on the intricacies of human activity recognition, providing valuable insights into the realms of context awareness and predictive modeling.

### 1.3.2 Time handling

Unlike static images, videos are defined by a space-time dimension. Hence, one of the important challenges in video analysis is **how to handle time?** Actions can have a high intra-class variance. Let us take the example of *preparing a dish*, different persons can take different steps in different orders. Hence, it is very important to model the temporal order and relations of the fine-grained actions leading to a coarse activity. In untrimmed videos, this becomes even more challenging, as we have to deal with questions like when an action starts and ends. Moreover, in these scenarios, actions are not always sequential as some actions can be dependent and yet happen at different time steps, as it is, it is already hard to infer long videos, therefore it is even harder for models to remember relevant data in order to extract or model long-range dependencies. Therefore, managing time also includes modeling long-range temporal dependencies.

### 1.3.3 Video representation

3D CNN architectures such as I3D [33] or X3D [65] have proven to be effective in handling spatio-temporal information. Nevertheless, these CNNs perform poorly when datasets are more complex and require deeper understanding. Therefore, 3D-CNNs are most of the time used to extract a global understanding of scenes in video clips, which we pass later on to methods that could better handle temporal dependencies, semantic relations, etc. Hence, having a video input on top of big models can be memory and resource consuming. Therefore, action recognition and detection are usually a two-stage model. The first stage consists of feature extraction, which is done offline, using a pre-trained model or by fine-tuning a model on the new distribution. The second stage consists of a module to handle mainly the temporal information, like graph convolutions [183], TSN [225] or LSTMs [198]. This

disparity between feature extraction and temporal study limits model capacities as we lose coherence in the semantic modeling of the actions. Furthermore, this becomes more critical in action detection, as untrimmed videos are usually long, and snipping these videos into smaller clips to extract features affects the temporal dependencies. As mentioned earlier, long-range dependencies are one important key to solving action detection and recognition. Since in our thesis, we focus more on action recognition and temporal modeling of action, we adopt this two-stage method.

#### 1.3.4 Dataset generalization

An optimal model or artificial intelligence (AI) robot is the one that can handle any scenario from arbitrary real-world videos. However, deep learning models are still not general enough to handle large distribution variances between datasets. One closer step towards such a model is to partially finetune pre-trained foundation models on new datasets without losing learned modelizations from previous datasets. Lately, transformers [217] have seen great success in computer vision in its many applications. Nevertheless, such networks require large datasets to train, so it is harder to fine tune them on smaller datasets. In this thesis we provide a new efficient way to fine-tune these transformers on smaller and new distributions, keeping the main framework intact and only adding a few linear layers (10% - 20% of the main module weights). We believe that this would be an important contribution towards building generalized models.

#### 1.3.5 Multi-modalities

In the contemporary landscape of scene recording and analysis, the scope has expanded far beyond the conventional RGB imagery. We are now privileged to have a plethora of sensors at our disposal, each with the capability to capture a diverse range of modalities, including but not limited to skeletons, depth, audio, and motion data. These varied modalities present a treasure trove of information that can greatly enhance our ability to recognize and understand complex human actions.

However, the true challenge lies in the efficient fusion of these multi-modal data. Integrating information from different sources is like putting together a complex puzzle, and finding the most effective method to do so represents a significant hurdle. Moreover, it's essential to bear in mind that utilizing all these modalities simultaneously can be resource-intensive. Hence, it becomes imperative to devise optimal strategies for their inclusion, ensuring that computational resources are utilized judiciously.

In the forthcoming thesis, we focus on harnessing these diverse modalities efficiently. Our mission is to leverage these rich sources of data, and to devise intelligent methodologies for their fusion, ensuring that the resulting insights are greater than a simple fusion or concatenation. We recognize that this endeavor holds immense potential in advancing the field of scene analysis and action recognition, ultimately contributing to a deeper understanding of human activities in various contexts.

#### 1.3.6 Subtle activity recognition

Human daily activities are not limited to one. One of the key challenges is to recognize activities that have infrequent occurrence and less distinctive patterns, which



can have a significant implication in the creation of machines that can effectively interact with and support humans in social interactions. Interactive intelligent agents acting as artificial mediators, engaging conversationally in a manner similar to humans, possess the capacity to exert a positive impact on the trajectory and results of human interactions. Research has extensively explored these agents in diverse domains such as collaborative teamwork, mental health, and education. A crucial requirement for successful and adaptable artificial mediation lies in the ability to fully recognize and understand the wide array of social signals conveyed by individuals. Currently, the effective resolution of this challenge remains predominantly unresolved.

## 1.4 Contributions

The contributions within this thesis are fundamentally rooted in responding to the tangible challenges encountered in real-world scenarios. These challenges serve as the driving force behind our effort to propose innovative solutions to effectively navigate these complexities.

Our primary contribution aims to tackle the challenge of fine-grained activity recognition. The proposed solution is capable of achieving fine-grained action understanding by extracting dynamics and interactions prevalent within video sequences.

Furthermore, our second significant contribution centers on a novel action detection framework. This framework diverges from conventional approaches by placing a focus on real-world applications, specifically targeting online action detection and action anticipation. These aspects of action detection hold immense practical relevance, aligning closely with real-time scenarios where the ability to predict and detect actions as they unfold in dynamic environments is crucial.

Our third contribution introduces a new efficient training process across multiple modalities and datasets. Through this work, we present a vital step towards streamlining the training process, facilitating the integration of diverse data sources and modalities, thereby enhancing the model's adaptability and robustness.

Also with MultiMediate'23, we introduce the inaugural challenge focusing on assessing engagement and recognizing physical behaviors during social interactions. We delineate the specific tasks, establish evaluation criteria, and provide insights into novel annotations derived from the NOvice eXpert Interaction (NOXI) database [27] and undisclosed test recordings of MPIIGroupInteraction [150]. Additionally, we present baseline methodologies for the challenge tasks and share the outcomes of the evaluation process.

Finally, our last contribution focuses on an important task of action understanding, which is abnormal and criminal activity anticipation. In this task, we propose a novel dataset for this task as well as a benchmark.

Each of these contributions reflects a deliberate and strategic approach aimed at tackling distinct yet interconnected facets within the realm of video analysis and

action recognition. These advances collectively serve as building blocks toward addressing the challenges prevalent in real-world scenarios, fostering innovation and progress within the field of video processing and action understanding.

#### 1.4.1 Semantic reasoning for fine-grained action recognition

**THORN: Temporal Human-Object Relation Network for Action Recognition.** In this work, we focus on tackling two of the previously mentioned challenges: **Fine-grained activity** and **Time handling**. THORN focuses on learning the semantics of objects and their relations to predict actions. In practice, it extracts spatio-temporal representations of objects and it learns their cross-relations by leveraging a graph-like structure. With THORN, we achieved competitive state-of-the-art performance in egocentric view action recognition datasets.

#### 1.4.2 Temporal reasoning for real-world scenarios action detection

**JOADAA: joint online action detection and action anticipation.** In this framework, we focus on online action detection. We propose to combine action anticipation and online action detection. In JOADAA, we add a middle stage in online action detection. This stage anticipates the upcoming action ahead of time, and we then use this information as a pseudo-future to make predictions on the ongoing frames.

#### 1.4.3 Multi-modal and multi-dataset training

**Robust and Efficient Multimodal Multi-dataset Multitask Learning.** In this work, we propose mainly two contributions. First, an efficient way to fine-tune pre-trained models on new datasets. Second, a more adequate use of multimodalities for network training.

##### Cross dataset training

Our method is specific to transformers. These transformer networks have known great success across different domains (NLP, image processing, etc.). In our work, we mainly add a few linear layers (10%-20% of the total weights of the transformer) to pretrained transformers (which we keep frozen), and only learn the weights on those linear layers. Finally, we are able to achieve the same accuracies w.r.t. fully-fine tuned models.

##### Multi-modality training

In previous works, multi-modal fusion is done after downsampling the input features, which leads to loss of information and poor cross-modality relations. We propose to use cross-attention added to each block of a transformer architecture. This gives our model more flexibility and allows it to benefit from the full information in different modalities. To show the flexibility of our proposed framework, we use different datasets, with different modalities (RGB, optical-flow, audio and transcript).

#### 1.4.4 Subtle activity recognition

**MultiMediate '23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions.** In this work, we focus on two tasks, namely, engagement estimation and bodily behavior recognition in social interactions. As part of the MultiMediate '23 challenge we present a novel set of annotations for both tasks. For

engagement estimation, we collected novel annotations on the NOvice eXpert Interaction (NOXI) database. For recognition of bodily behavior, we annotated the test recordings of the MPIIGroupInteraction corpus with the BBSI annotation scheme. Additionally, we present baseline results for both challenge tasks.

## 1.5 Thesis structure

- **Chapter 1 "Introduction"** presents action recognition and describes the key challenges and possible approaches. Then, our contribution is introduced.
- **Chapter 2 "Related work"** introduces the state-of-the-art action recognition and detection models. In this chapter, we present the proposed solutions to the different challenges in this domain. We mainly focus on recent deep learning approaches. Next, we introduce cross-data training and existing approaches to ways of efficiently training models. We also discuss multi-modalities and their case uses, and state-of-the-art approaches to using them.
- **Chapter 3 "Action recognition for fine-grained action"** This chapter we detail our introduced framework that aims at doing semantic reasoning on spatio-temporal representation of videos.
- **Chapter 4 "Online action detection"** Introduces a new way to combine online action detection and anticipation of actions. We show that by joining both approaches, they can benefit from each other and improve performance on both tasks.
- **Chapter 5 "Efficient transformer training"** details a new approach to transformer training, with minimum resources while having comparative results to traditional methods based on fully trained networks. This chapter also introduces a new way of using full information from different types of modalities and their cross relations.
- **Chapter 6 "Action recognition for social interaction"** tackles another aspect of action recognition and video analysis and deals with subtle action. These types of action are usually related to social interactions.
- **Chapter 7 "Action anticipation for abnormal activities"** Focuses on anticipating abnormal and criminal activities. We aim at analysing different temporal and spacial cues,
- **Chapter 8 "Conclusion and future works"** summarizes our thesis work and contributions, and opens a discussion of possible suggestions on future work and directions, with short- and long-term perspectives.

## Chapter 2

# Related work

In this chapter, we review the methods for action recognition and detection published in recent years. This literature study revolves around how action recognition has been approached in recent years for generic videos and what are their limitations. We are also going to talk about related work in the training of deep learning models. More related work to each of our contributions is discussed in the corresponding chapters.

We start with a discussion on basic concepts of *Deep Learning* and their definitions.

### 2.1 Basic concepts

- **Deep Neural Network (DNN)** - A DNN consists of three main blocks or layers: *input layer, hidden layers and output layer*. Each layer computes specific features. The input layer receives the raw input and extracts low-level features such as lines, edges, and corners. The hidden or mid-layers receive the low-level features and compute high-level semantics that the output layer could use to make better predictions.
- **Convolutional Neural Network (CNN)** - CNNs are derived from DNNs; the main difference is that CNNs as their name suggests are based on convolutions. The hidden layers in CNNs are typically structured as a fusion of convolutions followed by pooling operations. Convolution layers are orchestrated by specialized filters, also known as kernels. These filters extract information from neighboring pixels within a feature map. This process facilitates the computation of high-level semantics, allowing CNNs to extract complex patterns and features from the input data. In conjunction with convolution, pooling operations play an important role in shaping the network functionality. These operations, which can encompass various strategies such as average pooling or min-max pooling, contribute to the efficient management of feature maps. By downsampling local regions of the feature maps, pooling operations enable the network to focus on essential details while effectively reducing the spatial dimensions of the data. CNNs represent a shift in deep learning, harnessing the power of convolution and pooling to extract hierarchical and context-rich information from input data. This unique architecture enables CNNs to excel in a wide range of tasks, from image classification to object detection, and to learn patterns and relationships within complex datasets.
- **Recurrent Neural Network (RNN)** - These networks can be described as message passing networks, they are a stack of the same network. Their nature

allows them to treat sequential data more properly compared to CNNs or standard DNNs; hence they are used in video analysis to capture temporal dependencies for better video classification tasks. However, RNNs are not without drawbacks. One significant limitation lies in their sensitivity to the problem of vanishing gradients, which hampers their ability to effectively capture long-range dependencies. This can result in a degradation of performance when dealing with sequences of substantial length. Furthermore, RNNs tend to be computationally expensive, which can pose challenges in real-time applications and large-scale datasets. Additionally, the sequential processing nature of RNNs can hinder parallelization, affecting their efficiency in training and inference. In summary, RNNs offer valuable strengths in modeling sequential data and capturing contextual dependencies but must contend with challenges related to vanishing gradients, computational demands, and limited parallelization capabilities. Researchers and practitioners continue to explore solutions and variations, such as long-short-term memory (LSTM) and gated recurrent unit (GRU) networks, to mitigate these limitations and to harness the full potential of recurrent neural networks. Refer to [184] for a more detailed study of RNN.

- **Long Short term Memory (LSTM)** - LSTMs represent a significant evolution from traditional recurring neural networks (RNNs). One of the key distinctions lies in the enhanced ability of LSTM to capture and maintain long-range dependencies within sequential data. Unlike standard RNNs, which often struggle with vanishing gradients, LSTMs incorporate specialized gating mechanisms. These mechanisms, consisting of input, forget, and output gates, allow LSTMs to selectively control the flow of information through the network, effectively mitigating the problem of vanishing gradients. However, despite their considerable advantages, LSTMs are not without drawbacks. Their increased complexity compared to traditional RNNs results in a higher computational burden, making them more resource intensive in both the training and inference phases. This computational overhead can be a limiting factor in real-time applications or when dealing with large datasets. Moreover, while LSTMs are better at handling vanishing gradients, they are not entirely immune to the issue, especially when faced with extremely long sequences. For more information on LSTMs, the reader can refer to [198].
- **Attention mechanisms** The concept of attention in artificial intelligence draws inspiration from the way humans naturally focus on distinct regions in an image or specific words within a sentence. Human visual attention allows us to emphasize a particular area with "high resolution," perceiving the surrounding context in "low resolution," and adapt our focus or inferences accordingly. This natural phenomenon has been adapted into an attention mechanism in the field of artificial intelligence. In simple terms, attention in deep learning can be broadly understood as a vector of importance weights. When predicting or inferring elements such as a pixel in an image or a word in a sentence, we leverage the attention vector to gauge their significance. Recently, two categories of attention have emerged: hard and soft attention. Hard attention adheres to the principle of making decisive choices when selecting specific portions of input data. This decision-making process serves to simplify the task, particularly in object recognition, by strategically placing the Region of Interest (RoI) at the center of fixation. This focused approach ensures that irrelevant features beyond the designated region in the visual environment are naturally

excluded from consideration. On the other hand, Soft attention takes the entire input (image or video) and then softweights the RoI as per their relevance for the end task.

In the next sections of this chapter we dive into previous work on action understanding and their limitations to previously mentioned challenges.

## 2.2 Methods prior to deep learning

Before delving further into methods of deep learning, we review in this section some handcrafted approaches of handling action recognition. The fundamental concept underlying video analysis revolves around extracting distinctive features from a localized spatiotemporal representation of a video. An image, essentially a 2-dimensional dataset, results from projecting a 3-D real-world scene and encapsulates spatial configurations such as shapes and appearances of humans and objects. A video, on the other hand, is a sequence of these 2-D images arranged chronologically. Consequently, a video input depicting an action's execution can be depicted as a specific 3-D XYT space-time volume formed by concatenating 2-D (XY) images over time (T).

Space-time approaches involve recognizing human activities by scrutinizing the space-time volumes of action videos. A typical methodology for human action recognition in the space-time domain operates as follows: leveraging training videos, the system constructs a model for each action. Upon receiving an unlabeled video, the system creates a 3-D space-time volume corresponding to the new video, comparing it with each action model (i.e., template volume) to gauge the similarity in shape and appearance. The system then infers that the new video corresponds to the action with the highest similarity. This example illustrates a standard space-time approach utilizing the 3-D space-time volume representation and a template-matching algorithm for recognition.

Below, we elaborate on some prevalent pre-deep learning era methods based on space-time approaches, categorized into (A) space-time volumes, (B) space-time local features.

**(A) Space-Time Volumes:** The central concept in recognition using space-time volumes lies in measuring the similarity between two volumes. To calculate accurate similarities, an array of space-time volume representations has been developed. Some approaches only stack regions of a person (i.e., silhouettes) to track shape changes explicitly [24]. [113] focuses on extracting segmented 3-D XYT volume segments that corresponds to a moving human.

**(B) Space-Time Local Features:** This intuition in such methods is if a system is able to extract semantics describing characteristics action's 3-D volume, the action is now an object-matching problem.

## 2.3 Human object interaction actions (HOI)

In the ever-evolving landscape of action recognition, the prevailing state-of-the-art methods predominantly train their gaze on coarse and straightforward activities.

Such activities, which include walking and running, have undeniably garnered substantial attention due to their prevalence in our daily lives. However, the intricate world of fine-grained actions, although brimming with rich complexities, often finds itself in the shadows, relatively underexplored and underrepresented.

Within this uncharted territory, we unearth a captivating array of activities that defy simple categorization. Activities such as assembling furniture or food preparation stand as prime examples of this fine-grained landscape. What distinguishes these actions is not only their subtlety but also their composite nature. They transcend the boundaries of isolated human movements, weaving intricate narratives of interaction between the human agent and the surrounding environment, often populated by objects and artifacts.

This unique blend of finesse and complexity makes these actions particularly challenging to accurately discern and classify. Their study holds immense promise not only for the advancement of action recognition, but also for shedding light on the profound interactions between humans and their surroundings.

**CNNs** CNNs have achieved great results in video analysis and action recognition, typically two-stream networks [194, 66, 67] and 3D-CNNs [101, 34, 226]. However, these networks are limited to video-level label datasets. In fact, the local nature of filters in CNNs limits their range of dependency capturing in space as in time, not only that CNNs capture local features, but they share the weights of their kernels across all pixels, hence they cannot capture specific semantics to model fine-grained and complex actions. In order to improve CNNs performances on HOI datasets, some methods, such as [230], propose to mix object detection to capture object features and then fuse with 3D-CNN to have a richer description of the clips. However, this does not bring a significant improvement as there is no modeling of objects' interactions and their dependencies in space-time.

**Graph convolutions** Graphs have also been visited in the action recognition task. Graphs are good at modeling interactions and dependencies. The nodes and edges can describe the relationship between semantics. However, they are built on top of extracted features from CNNs, and hence they are very dependent on the feature quality. Not only that, but usually these methods use object detection and ROI-Align to capture detected object features, in crowded scenarios, it becomes hard to capture specific semantics to different elements in the human surroundings, which limits graphs capacity to learn different dependencies and relations.

## 2.4 Temporal modeling and online action detection

Action detection can be approached in different ways. Some methods [63, 95] choose to approach it as a frame-level action classification. In a way, it is similar to semantic segmentation but with frames instead of pixels. The most critical part in action detection is how to model time. In the following, we briefly introduce some proposed architectures to handle temporal information.

**RNNs** As mentioned earlier, due to their message passing architecture, RNNs are suitable to handle sequential data. Therefore, they can capture temporal dynamics

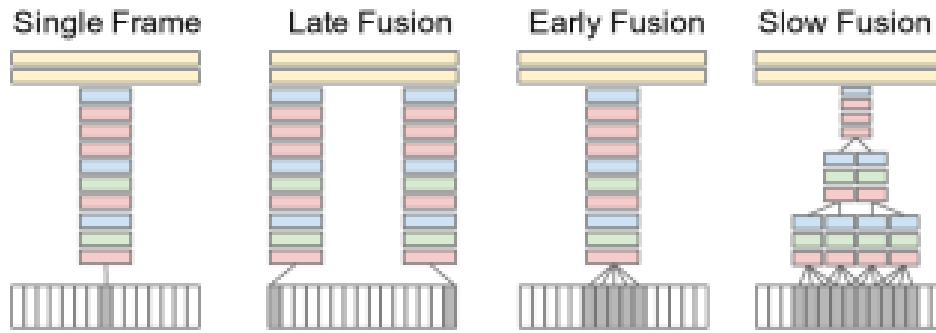


FIGURE 2.1: Different techniques for integrating information across the temporal dimension in [108]. Convolutional, normalization, and pooling layers are denoted by red, green, and blue boxes, respectively. In the Slow Fusion model, the columns depicted in the illustration share parameters.

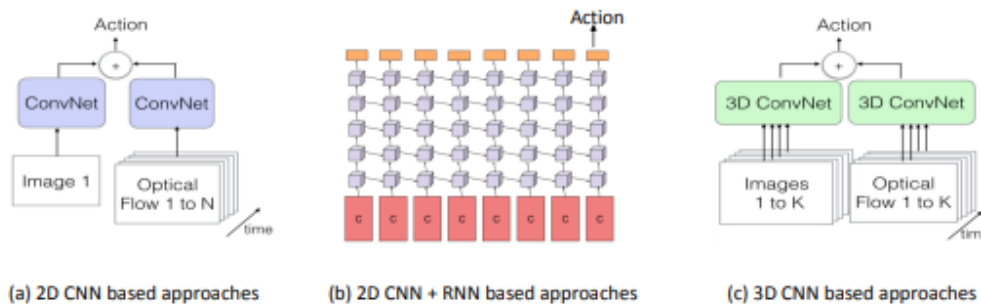


FIGURE 2.2: Key action recognition approaches based on Deep Learning aim to model temporal information in videos. These approaches utilize (a) 2D CNN (left), (b) 2D CNN + RNN (middle), and (c) 3D CNN (right) to consolidate temporal information for action classification.

in videos. Nevertheless, it has been proven that such networks are limited to capturing dependencies between actions that have apparent movements and motion. Moreover, as discussed previously, these networks suffer from vanishing gradient; hence, they cannot capture very long temporal dependencies.

CNNs convolutional neural networks, these architectures are suitable for image like data as they can extract useful semantics. Basically, they are a stack of convolutions and grouping operations that enable models to capture different semantics from images at different layers. See Figure 2.3 for an illustration. In the following, we discuss more details on CNNs and their application to video processing.

- **2D CNN based approaches** [108] expanded the temporal connectivity of a CNN to leverage local spatio-temporal information. They investigated various strategies to integrate the input data across different temporal dimensions in CNNs, as illustrated in Figure 2.1. These strategies include: (i) a model based on a single frame, (ii) an early fusion model that combines information across an entire time window immediately at the pixel level, (iii) a late fusion



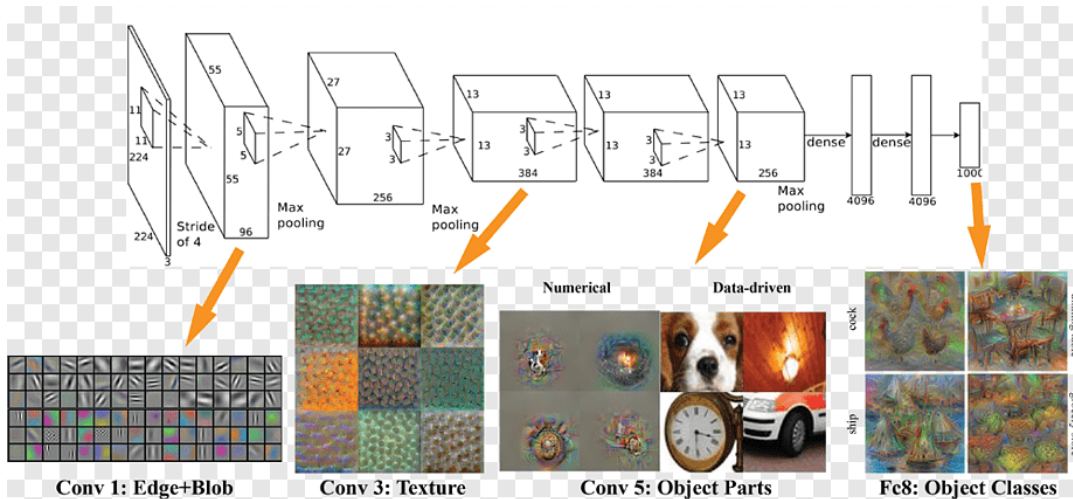


FIGURE 2.3: This figure illustrates how CNNs work on images, at early stages we learn low level features such as edges and texture and deeper in the network we learn more explicit semantics. This showcase strength of CNNs on image like data.

model that employs two separate single-frame networks with shared parameters and merges the two streams through a fully connected layer, and (iv) a slow fusion model, which represents a balanced blend of the two approaches. The slow fusion model integrates temporal information throughout the network, allowing higher layers to access information in both spatial and temporal dimensions. Lastly, they proposed a multi-resolution CNN architecture for action recognition. The input frames are directed into two distinct processing streams: a context stream modeling low-resolution images and a fovea stream handling the high-resolution center crop. This multi-resolution CNN architecture, combined with slow fusion along the temporal domain, demonstrated effectiveness in classifying actions in sports videos characterized by dissimilar backgrounds.

- 2D CNN + RNN based approaches** Authors in [58, 241, 124, 197, 203] employed the concept of an encoder + decoder framework for action recognition. As depicted in Figure 2.2(b), the fundamental approach in these methodologies involves encoding frame-level features using a 2D CNN (encoder) and subsequently subjecting these features to complex temporal pooling using sequential networks like LSTM (decoder) before carrying out action classification. These encoders typically comprise image classification networks pretrained on ImageNet [57]. Additionally, the core idea remains consistent, with variations such as stacked LSTMs, Gated Recurrent Unit (GRU), and bi-directional LSTMs being employed in [124, 197, 203].
- 3D CNN based approaches** Du et al. [68] introduced the concept of 3D (XYT) convolution to capture spatio-temporal patterns within actions. The utilization of 3D kernels facilitates a close integration of spatial and temporal dimensions, leading to improved action classification. Current research on 3D ConvNets highlights their effectiveness as descriptors due to their generality, compactness, simplicity, and efficiency [68]. These convolutional deep networks in 3D

can concurrently model appearance and motion. Unlike 2D ConvNets, where convolution and pooling operations are performed solely in the spatial domain, 3D ConvNets execute these operations spatio-temporally.

Carreira and Zisserman [33] recently developed I3D, a 3D CNN-based fully convolutional network designed for action classification. The unique architecture of I3D allows it to benefit from pre-training on ImageNet [57] by inflating 2D kernels to 3D kernels. Asymmetric operations are introduced along space and time; for instance, initial layers apply  $1 \times 3 \times 3$  convolutional operations compared to the traditional  $3 \times 3 \times 3$ , addressing the higher dimension along the spatial domain. I3D, featuring 9 inception modules and multiple bottlenecks to reduce parameter complexity, is well-suited for video classification problems after being pre-trained on both ImageNet [57] and Kinetics [110].

The success of I3D has led to the development of holistic methods like the slow-fast network [68] and MARS [47] for generic datasets such as Kinetics [110] and UCF-101 [195]. The slow-fast network [68] incorporates the concept of fovea and context stream from [109], utilizing a 3D CNN as the visual backbone. With the slow pathway capturing spatial semantics of image frames and the fast pathway focusing on motion at a fine temporal resolution, this network operates videos at low and high frame rates, respectively. To optimize the network, the fast pathway (with a high frame rate) reduces channel capacity.

- **TCN** or temporal convolutional networks are basically one-dimensional convolution networks. These networks use convolution across the temporal dimension, to capture temporal behaviors and relations. Unlike RNNs, these networks can process longer videos. However, and as discussed in 2.1, CNNs use kernels that share weights between different local regions, making it harder to capture specific key information.

**Transformers** Transformers are more recent networks; they were first introduced in NLP (natural language processing) [211]. Due to their huge success, some researchers sought to use them for video analysis and action recognition [8]. These architectures can efficiently model the dependencies between different parts of an input. Unlike the previously mentioned methods, which are local operations, transformers can attend to full information and capture global and long-range dependencies. Moreover, such networks are scalable to big data and models. Deep learning advances have effectively adapted the transformer design for computer vision applications such as image classification, resulting in vision transformers. Due to these advantages, transformers have been more successful and used in computer vision in general.

- **Transformers in computer vision** Attention mechanisms are commonly employed in computer vision alongside Convolutional Neural Networks (CNNs). However, there are limited instances where the transformer architecture is exclusively utilized to address computer vision challenges. Attention mechanisms can also be employed to replace specific components of CNNs while preserving the overall network structure.

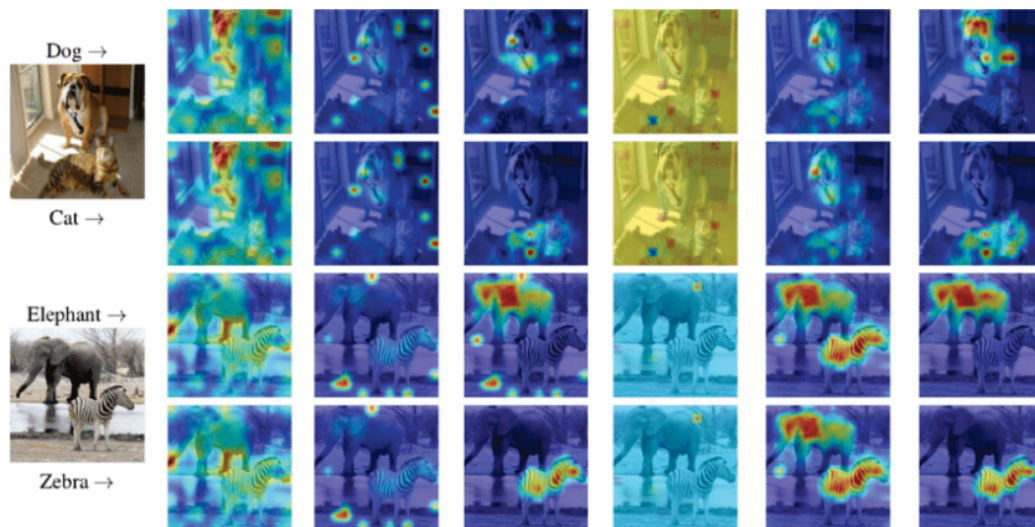


FIGURE 2.4: Class-specific visualization results from ViT with attention maps in contrast to class-activation maps in CNNs. Source [https://www.researchgate.net/publication/355693348\\_Transformers\\_in\\_computational\\_visual\\_media\\_A\\_survey](https://www.researchgate.net/publication/355693348_Transformers_in_computational_visual_media_A_survey)

While CNNs are integral to many traditional computer vision models, recent developments in the field have illustrated that transformer models applied directly to sequences of image patches can perform exceptionally well in image classification tasks.

The Vision Transformer (ViT) [114] model, founded on a transformer encoder, has exhibited highly competitive performance across various computer vision applications, including image classification, object recognition, and semantic image segmentation. This model underscores the adaptability of transformers in the realm of computer vision.

The ViT model incorporates a self-attention layer, enabling the global embedding of information across the entire image. Through training, the model acquires the ability to represent the relative locations of image patches, effectively reconstructing the image's structure. The transformer encoder in ViT comprises three main components:

**Multi-Head Self-Attention Layer:** This layer concatenates attention outputs linearly and employs multiple attention heads to train both local and global dependencies within an image.

**Multi-Layer Perceptrons:** This component consists of a two-layer system with a Gaussian Error Linear Unit (GELU) activation function.

**Layer Norm:** Implemented before each block, it restricts the formation of new dependencies between training images, contributing to reduced training time and improved overall performance.

Additionally, residual connections are integrated into the ViT architecture after each block to facilitate information flow throughout the network without encountering non-linear activations. The MLP layer functions as the classification head for image classification tasks, featuring one hidden layer for pre-training and a single linear layer for fine-tuning. See figure 2.4 for a clear visualization of vit [114] architecture.

- **Transformers for video analysis** As example, we cite ViViT [8]. We can say that “Video Vision Transformer (ViViT)” is an extension of ViT that works on videos. It is also a type of neural network architecture that is used to process video data. It combines the ideas behind both the transformer model and the vision transformer to create an architecture that can effectively process both spatial and temporal information in video data. As addition, there is a temporal convolutional layer which is used to model the temporal structure of the video data. It works by applying convolutional filters to the video frames, which allows the model to learn spatiotemporal features of the frames.
- **3D-CNNs vs. Transformers** Video Vision Transformers (ViViT) and 3D Convolutional Neural Networks (3D CNNs) are neural networks for video recognition, differing in their data processing.

3D CNNs are tailored for 3D data, using 3D convolutional filters on video receptive fields to learn spatiotemporal features. In contrast, ViViT, based on the transformer architecture, divides videos into frames and employs a multi-head self-attention mechanism for frame importance, capturing both spatiotemporal and temporal context features from non-grid data.

Distinctively, 3D CNNs focus on spatio-temporal features from grid-like video data, while ViViT excels in capturing contextual features from non-grid video data. ViViT’s efficiency and parallelizability make it suitable for large-scale video recognition tasks.

All of these methods can achieve good results in modeling temporal information. However, directly applying such methods is not efficient in solving real-world scenarios such as online action detection. Contrary to offline action detection, online action detection (OAD) suffers from limited information as it has access to only past and present information, hence even with good networks such as transformers it is hard to accurately predict action in a streaming manner. This becomes even more challenging in densely annotated datasets such as [239]. We are going to detail more challenges and related work on this part in the next section.

## 2.5 Efficient transformers and cross-dataset training

In computer vision, it is common to use pre-trained models from large datasets. These models are then fine-tuned on smaller datasets, as they carry in them good hyperparameter and feature extraction weights. Some other techniques are **Generative learning** methods that commonly involve good data augmentation techniques and learn good feature representation using variation in input, or **Contrastive learning methods** which aim to learn a better space for the features learned by the model. Model sizes have been increasing lately, such models need more and more data to train, moreover, transformers are known to require a big dataset to train. This presented us with two challenges. First, large models and datasets are resource consuming. Second, it is hard to fit these transformer models on a small down-task dataset. To tackle such challenges, some methods propose only updating new parameters added to the model or input [90, 107, 123, 125, 106], or updating some of the model parameters in a sparse manner [255, 205, 242], or finally, low-rank factorization of weight matrices to reduce the number of parameters to be updated while keeping

the weight matrix approximately the same [93]. As this is part of our contributions, it will be detailed in the upcoming chapters.

## 2.6 Multi-modal fusing

With the advancement of recent technologies, we now have available sensors that can capture other modalities than **RGB and audio**, such as **skeletons or gaze**. Moreover, there exist deep learning algorithms that can infer **optical flow, skeletons, or gaze** from RGB frames. Such algorithms are time and resource costly; however, they help increase the amount of available modalities and data. The question that arises here is *How do we mix these modalities together for better performance?*. Some of the existing approaches are:

**Early fusion** in early fusion we concatenate all inputs (from different modalities) in early stages of the network. It is argued that the earlier the fusion, the better it is. However, this kind of approach requires big data amount for training.

- **Advantages:**

- **Comprehensive Representation:** Early fusion provides a holistic representation of information by combining modalities at the input level, allowing the model to consider all modalities simultaneously.
- **Simplified Model:** The model architecture is often simpler compared to other fusion methods, which can lead to easier training and understanding.

- **Disadvantages:**

- **Fixed Fusion:** It assumes equal importance for all modalities, which may not be suitable for tasks where modalities contribute differently.
- **Increased Dimensionality:** The input data dimensionality can become large, leading to potential challenges in training large-scale models.

**Late fusion** involves processing each modality independently before combining their respective features in the later stages of the model.

- **Advantages:**

- **Modularity:** Late fusion allows for modularity, as each modality can be processed independently before fusion, making it easier to update or modify individual components.
- **Adaptability:** Different modalities can be processed using specialized networks, adapting to the characteristics of each modality.

- **Disadvantages:**

- **Information Loss:** Late fusion may lead to information loss as it processes modalities independently before combining features, potentially missing correlations between modalities.
- **Complexity:** The model architecture can become more complex, requiring careful design to ensure effective fusion without introducing redundancy.

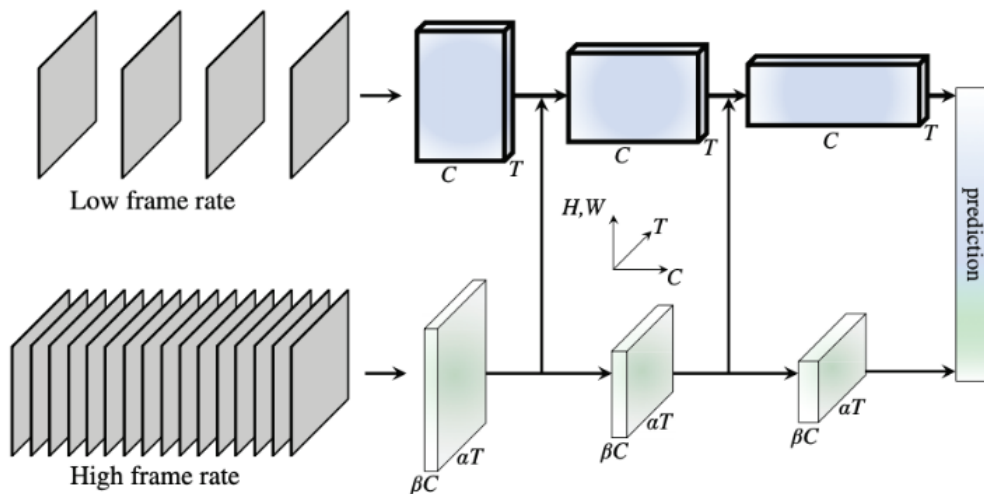


FIGURE 2.5: Example of lateral fusion in SlowFast Network [68]. The two streams could either take a high frame rate and low frame rate such in the figure or input different modalities.

**Lateral fusion** is used in two stream networks such as [68], where there are connections between the two streams to pass information.

- **Advantages:**

- **Temporal-Spatial Integration:** Lateral connections in networks like SlowFast [68] facilitate the integration of temporal and spatial information by connecting feature maps across different layers and speeds.

- **Disadvantages:**

- **Complex Architecture:** Implementing lateral connections introduces complexity to the model architecture, requiring careful design and parameter tuning.
- **Increased Computational Cost:** Lateral connections may increase computational requirements during both training and inference, affecting the overall efficiency of the model.

**Attention mechanisms** dynamically weigh the significance of different modalities based on the context, allowing the model to selectively focus on the most relevant information.

- **Advantages:**

- **Contextual Relevance:** Attention mechanisms dynamically weigh the importance of different modalities based on contextual information, allowing the model to focus on the most relevant features.
- **Flexibility:** The attention mechanism is flexible and adaptive, making it suitable for tasks where certain modalities may be more informative in specific contexts.

- **Disadvantages:**

- **Training Complexity:** Implementing attention mechanisms can introduce additional complexity during training, requiring careful tuning of hyperparameters.
- **Computational Overhead:** Attention mechanisms may introduce computational overhead, especially in large-scale models, affecting inference speed.

Another approach involves the implementation of a teacher-student network [47], as detailed in the experiments. This network mimics the motion stream at inference time without actually computing them, specifically the optical flow. In this setup, a teacher network (motion stream) is independently trained for the end task of action classification. Subsequently, the RGB stream is trained for action classification while also mimicking the features learned by the motion stream. This is achieved through a distillation loss that minimizes the Euclidean distance between the features learned by both streams. The experiments demonstrate that MARS is more effective at test time than individual streams, even with a significant reduction in test time.

Some other approaches involves searching for a neural architecture to fuse the RGB and optical flow modalities. In [174] introduced a Neural Search Architecture (NAS) designed to combine RGB and optical flow streams. This search mechanism explores questions such as how to combine RGB and optical flow (through concatenation or summation) and at which layers these modalities should be combined.

However, these types of fusion are usually done after downsampling, and hence, they limit the number of shared information. which leads to loss of information and poor cross-modality relation.

There exist other approaches such as [181] that use minimal downsampling. However, the remaining limitation is that most of these methods that use fusion lack flexibility, as they need to handle each modality differently. In this thesis, we introduce a more flexible way to fuse different modalities.

## Chapter 3

# THORN: Temporal Human-Object Relation Network for Action Recognition

The exploration of human actions within the task of understanding interactions, especially interactions between humans and objects, serves as a fundamental aspect in various domains. In this thesis, the focus is on proposing a methodology for identifying and understanding human actions by delving into the intricate set of interactions that define each action.

The crux of our proposed approach lies in the development of an end-to-end network called THORN. This network is designed to harness the significant challenges embedded within human-object interactions and object-object interactions, thereby enabling an accurate prediction of actions. The architectural foundation of THORN rests on a robust 3D backbone network, which forms the basis for its functionality.

THORN comprises several crucial components essential to its effectiveness. First, an object representation filter is integrated into the model, enabling the modeling of objects within the interactions. Second, an object-relation reasoning module allows the complete capture of intricate relationships between objects involved in the actions. Finally, a classification layer is incorporated to facilitate the prediction of action labels based on the learned interactions and representations.

To substantiate the resilience of THORN, rigorous evaluations were conducted on two extensive and demanding datasets: EPIC-Kitchen55 and EGTEA Gaze+. These datasets are recognized as among the largest and most challenging repositories for first-person and human-object interaction analyses. Remarkably, THORN demonstrates state-of-the-art performance in both datasets, showcasing its robustness and efficacy in accurately predicting actions within complex interaction scenarios.

This achievement underlines the potential and applicability of THORN in real-world scenarios where understanding human actions through interactions is paramount, such as robotics, surveillance, human-computer interaction, and beyond. The success of this model opens doors to further advances in understanding and interpreting human behavior within diverse contexts, thereby contributing significantly to various fields that rely on action recognition and understanding.



### 3.1 Introduction

Human activity recognition in video is a fundamental problem in computer vision, due to its wide field of applications, such as human-computer interaction [103] or video surveillance [151]. In the expansive domain of action recognition, the strides made by machine learning and computer vision models are undeniably impressive. However, the landscape still presents a critical gap: the prevailing state-of-the-art methods predominantly focus on deciphering relatively straightforward activities, such as walking or drinking. The real challenge lies in decoding the complexity of multifaceted longer-term activities, such as assembling furniture or the complex steps involved in food preparation. Astonishingly, these complex activities remained largely unexplored territory within the realm of recognition methodologies. One significant limitation of existing methods is their reliance on end-to-end models. Although these models excel in generating video-level labels, they falter when it comes to explicitly dissecting actions into their hierarchical components or the intricate web of subactions and interactions they entail. This obvious gap in the management of complex and composite activities presents a significant opportunity for advancement in the field. The current focus on simple actions leaves a vast unexplored terrain, one with challenges and opportunities. Tackling the recognition of complex activities requires a paradigm shift, moving beyond surface-level recognition towards a granular understanding that dissects actions into their nuanced subactions and captures the hidden interactions between various elements involved. By embracing a more hierarchical perspective that disentangles actions into their constituent parts and captures the interplay between these elements, the field can progress towards more robust and comprehensive action recognition systems. Addressing this limitation not only expands the scope of action recognition, but also holds immense potential for real-world applications across numerous domains; it can revolutionize how machines perceive and interact within our environment, paving the way for a new era of intelligent systems capable of comprehending and responding to multifaceted human actions. Furthermore, neuroscience [16, 15] has shown that human perception of action is actually based on the decomposition of an action into different groups of interactions that allow him to understand other human behaviors. In this thesis, we decide to visit this composite action that we refer to as actions of Human-Object Interaction (HOI). Not only that, we also focus on first-person view HOI action recognition.

First-person action recognition introduces a unique set of challenges that significantly impact the accuracy and efficiency of recognition systems. One primary challenge lies in the constrained field of view inherent in first-person perspective videos. Often, crucial actions to understanding a situation occur outside of the narrow viewing range of the camera. This limitation poses a hurdle in capturing the entirety of actions, leaving crucial parts undocumented and complicating the recognition process. Moreover, the substantial ego-motion induced by the rapid movements of the camera adds another layer of complexity. These swift camera movements result in considerable motion within the video frame, making it inherently challenging to accurately discern and recognize actions. The dynamism introduced by these movements creates complexities in action recognition, often requiring sophisticated algorithms to decipher and interpret actions amidst the blur of motion or rapid transitions. The perspective of ego vision, which typically encompasses human hands and an array of surrounding objects, further complicates the recognition process. Actions in this context are predominantly characterized by interactions

between the individual and the surrounding objects. Consequently, a significant challenge emerges in distinguishing between relevant objects central to the action and elements that serve as distractors within the field of view. Addressing these challenges requires progress on several fronts. Enhancing recognition systems to anticipate and infer actions occurring beyond the immediate visual scope, mitigating the impact of ego-motion on action perception, and developing robust methodologies to discern relevant objects amidst a cluttered field of view are pivotal areas for improvement. These advances hold the key to overcoming the limitations posed by the first-person perspective, thereby enabling more accurate and comprehensive action recognition systems. Solving these challenges not only augments the accuracy of recognition, but also broadens the applicability of first-person action recognition across numerous domains. From improving personal assistive technologies to revolutionizing immersive and surveillance experiences, overcoming these obstacles unlocks the potential for a more nuanced understanding of human actions from the first-person point of view.

In Human-Object Interaction (HOI) recognition, actions manifest themselves as combinations of verbs and nouns. For example, consider the action of "cutting bread with a knife." This action encapsulates the verb "cut" alongside the nouns "knife" and "bread." Recognizing HOIs lies in dissecting these actions into their constituent parts, a process akin to visual relationship detection. In this context, the task extends beyond mere object recognition (identifying the nouns involved) to a more complex endeavor of inferring the relationships and motions (the verbs) occurring between various objects and the human. This perspective frames HOI recognition as a multifaceted challenge that necessitates not only the identification of objects within a scene but also an intricate understanding of how these objects interact with the human agent. Essentially, HOI recognition involves translating visual cues that denote not just the presence of objects, but also the dynamic relationships and interactions unfolding between them and the human entity. This requires systems that can discern subtle visual cues that indicate interactions, motions, and relationships within a scene, enabling a deeper understanding of human-object interactions. By addressing these complexities, HOI recognition transcends simple object identification, offering profound implications for various domains. Fig. 3.1 represents an example of an object-based action: *wash plate*. Such action requires highlighting objects like the *hand*, the *plate* and the *tap* while giving less attention to other objects that are not important to the action.

Previous works such as two-stream CNNs [194, 66, 67] or 3D CNNs [101, 34, 226] have achieved very good results on third view and video level label datasets [111, 120, 109, 196]. However, when it comes to HOI actions, they still lack in performance. That is due mainly to the fact that CNNs capture shareable local features in the image/videos and they cannot handle complex or fine-grained actions. Another major challenge is the fact that such activities can often be performed in a variety of ways, making it harder for CNNs to learn significant patterns.

Thus, our intuition is to build a model that can extract detailed and object specific semantics in the videos, as well as explore the cross-object relation at different time steps. By doing so we can, firstly, improve object recognition in actions of HOI (the noun). Moreover, we can refine the motion recognition (the verb) by having a clearer idea about the interactions of these objects and their roles in the action. Finally, by encoding the scenes into a graph of object interactions, we make it easier

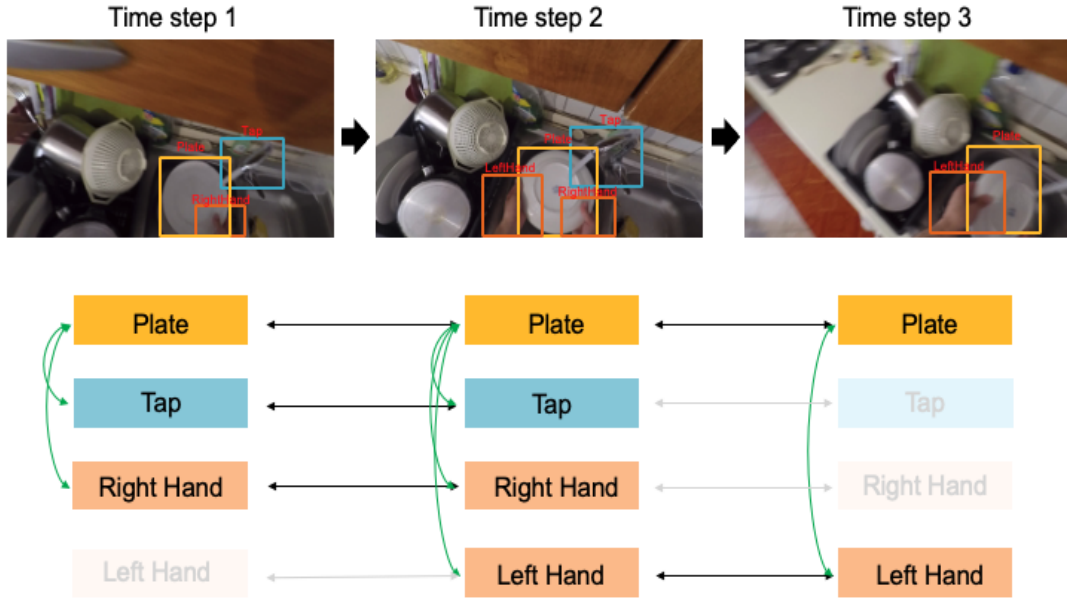


FIGURE 3.1: An example of the Human-Object Interactions of *wash plate* in a first-view video. Green arrows represent interactions at the same time step (i.e., spatial relation) while black arrows represent interactions across time. In practice, the model captures all the objects detected. For simplicity, here we highlight only the relevant objects to *wash plate*.

to learn patterns for actions even if they have many variations, since the interactions are usually the same.

To address the aforementioned challenges, we propose a new module built on top of 3D-CNNs; this module is divided into two sub-parts. Firstly, we design an **Object Representation filter**. This first submodule acts as a filter that retrieves specific and object-related semantics from the overall and mixed representation (extracted from the 3D-CNN). Secondly, we add an **Object Relation Reasoning** module that uses the detailed representations to explore cross-object relations (interactions). Finally, we obtain an object-centric model that can predict actions of HOI by exploring human-object and object-object interactions.

To summarize, our main contributions are:

1. A model that can find and extract detailed semantics of specific objects;
- 2- A graph-based module capable of exploring interactions between different objects.

## 3.2 Related work

Recognition of HOI actions became the focus of many research subjects lately, especially with the development of important datasets such as [53, 6, 64]. Several approaches have been proposed to tackle this problematic. In the following, we review some of these approaches.

### 3.2.1 3D-CNNs

3D-CNNs methods focus on getting the overall appearance of the videos without considering the objects interactions. Since these methods cannot capture specific or detailed semantics, they are limited in case of actions of HOI. Making this architectures more adequate to video level labels. We cite as an example I3D [34], The design choice of I3D enables it to leverage pre-training from ImageNet [57]. This is done by inflating the 2D kernels to 3D kernels. Moreover, asymmetric operations are imposed along space and time, for example, initial layers apply  $1 \times 3 \times 3$  convolutional operations compared to  $3 \times 3 \times 3$  to handle the higher dimension along the spatial domain. I3D with 9 inception modules, multiple bottlenecks [39] to reduce parameter complexity, and pre-trained on ImageNet [57] and Kinetics [110], is well engineered for video classification problems. Although it achieves good results on many action recognition datasets, its performance is still poor on actions of HOI. To improve the performances on these 3D-CNNs, Long Features Bank [230] for instance, tries to capture HOI actions by extracting and fusing features from local clips as well as globally from the whole video. This method uses object detection and ROI-Align to capture the features of the detected object. Although they successfully capture richer features and more temporal information, they fail to do any object interaction modeling. Hence, they cannot improve much on HOI actions. In the same direction, Temporal Binding Networks (TBN) [112] proposes to capture local clip features from different clips and fuse them for later prediction. In addition to that, TBN uses multi-modalities as it captures audio-visual features using audio, RGB, and optical flow. However, we believe that this multi-modality will not always bring much information about the objects. Sounds can be very noisy and very similar which can confuse the prediction. Moreover, fusing multi-modalities can be hard and requires lot of efforts that may not lead to significant improvements. Finally, other works such as [224] also use multi-modality reasoning. However, we argue that HOI actions recognition requires more focus on objects and their interactions.

### 3.2.2 Graph convolutions

Recently, graphs have also been considered a way for solving action recognition [229, 183, 238, 48].

As for human-object interaction, videos as a space-time region graph [229] propose to model the interaction between objects and humans in two steps as they build two different graphs. This allows to correlate objects across space-time. Similarly, in [74], the authors construct the nodes of the graph with consideration to the node class. For instance, the node for the scene is computed using the aforementioned I3D. While for objects, they use the Faster-RCNN network [42] trained on MS COCO. All these methods mentioned above try to define their nodes by using ROI-Align. However, this is not optimal as, in most cases, multiple objects are present at the scene and some of them are too close to each other. In this case, the projected coordinates of different objects tend to be in the same set of pixels. Therefore, extracting an object specific feature from a feature map with low resolution becomes difficult. Not only that these methods rely on pre-trained object detectors, hence they can not leverage only objects relevant to the action. Whereas in our work, we learn to filter only relevant objects and learn specific representation to different object-classes in an end-to-end way.

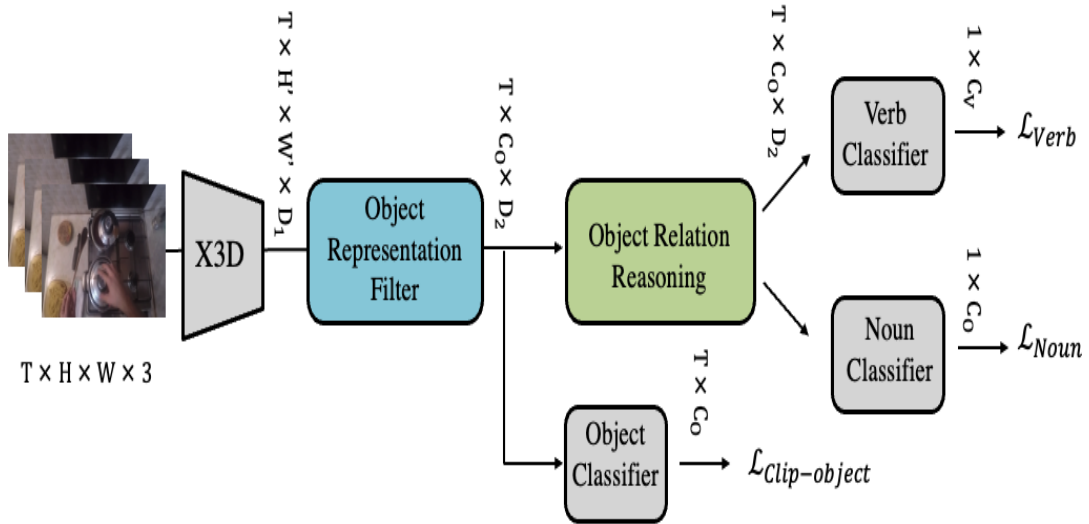


FIGURE 3.2: THORN architecture contains three main components: (1) a **Visual encoder** (i.e., X3D) encodes the input RGB clip into a primary spatio-temporal representation. (2) The obtained representation is fed to the **Object Representation Filter**, which maps the previous representation into object-class representation. To ensure a discriminative object representation, an object classifier is added on top of the object-class representation. This classifier is trained with the pseudo-object ground truth provided by an object detector. (3) The object-class representation is also sent to the **Object Relation Reasoning** module to model the temporal-object relation in a dissociated manner. Finally, two classifiers are used to predict the verbs and nouns relevant to the action.

In the domain of semantic modelling, Class Temporal Relational Network(CTRN) [48], is proposed for the action detection tasks. However, CTRN is a two steps method, which is built on top of pre-extracted flattened 1-dimensional features. The dissociation between the visual encoder and the temporal module makes the model overlook the appearance and spatial information in the video, while such information is critical to the HOI action recognition. In this work, we propose a one-step method, THORN, for HOI action recognition. Different from CTRN, our method leverages the object detector to extract the object semantics directly from the spatio-temporal features. After that, graph reasoning is applied to refine the object representation and to jointly model inter-object relations. This design allows the model to capture the latent relations among the objects in the videos, which results in higher accuracy in HOI action recognition.

### 3.3 Proposed method

In this section, we detail each sub-part of the proposed model, THORN. The main components in this model are: a **3D Visual Encoder** which encodes the video into a spatio-temporal embedding. Then, the previously extracted embeddings are passed to the **Object Representation Filter** (ORF). This filter extracts class-specific features. Finally, the **Object Relation Reasoning** module, computes the relation between the different objects to predict the action. Fig. 3.2 provides an overview of the model.

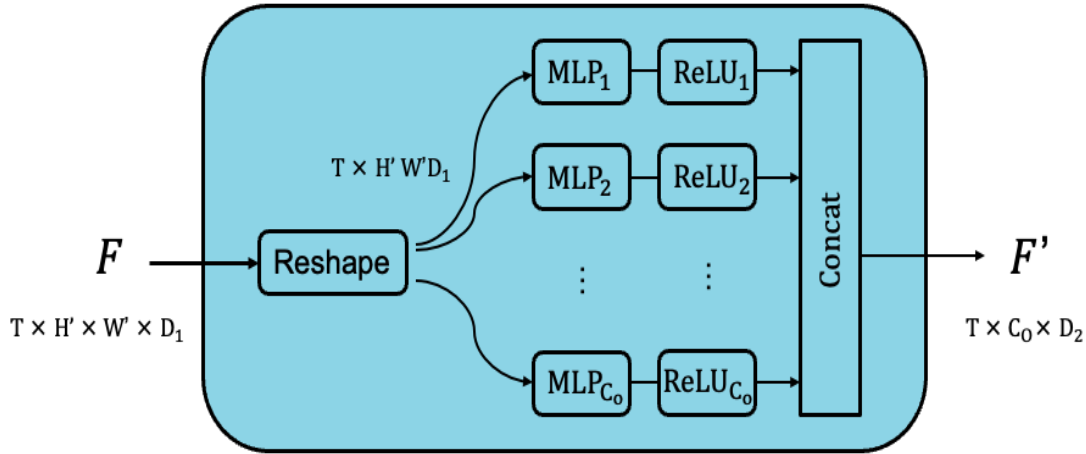


FIGURE 3.3: Schema of our Object Representation Filter (ORF). The input is the feature map from the 3D encoder reshaped to  $T \times H' \times W' \times D_1$  and the duplicated  $C_0$  times, where  $C_0$  is the number of classes. Finally, we have a representation specific to each object class.

### 3.3.1 Visual encoder

We start by using a visual encoder to extract an embedding that serves as a full understanding of the scene, and carries the global information of the input frames. We choose X3D [65] as our visual encoder. X3D has many advantages as it does not do any temporal pooling and it keeps the full temporal information, providing richer temporal information. Moreover, X3D is a lighter model compared to other architectures such as I3D [34]. The input to the 3D encoder is a set of video-clip frames. The output is a spatio-temporal representation  $F$  of shape  $(T \times H' \times W' \times D_1)$ , where:  $H' = W' = 7$ ,  $D_1 = 432$ , and  $T$  is the same as the input.

This embedding carries both spatial and temporal information. The spatial information is important as it provides object related information, such as its appearance, shape and position (e.g. drawers usually appear at the bottom of the image). That is why instead of using the X3D final output of shape  $(T \times 2048)$  to construct our nodes, we use a finer spatial representation of shape  $(T \times 7 \times 7 \times 432)$ , making nodes of our graph contain more and finer information about the objects. We provide more details on this in the ablation study, by comparing both settings. Finally, as X3D is a light-weighted model it is easier to train the *Visual Encoder* jointly with the following modules.

### 3.3.2 Object representation filter

In pursuit of object-centered reasoning, our main goal is to provide representations of objects within a scene. To accomplish this, we proposed the 'Object Representation Filter' (ORF) module. This module is designed to extract distinct semantic representations tailored to individual object classes from the comprehensive scene representation obtained previously. The ORF module functions as a filtering mechanism, isolates object-specific representations from the visual encoder output. In practical terms, the initial step involves reshaping the visual encoder representation, denoted as  $F$ , into a format of  $(T \times H' \times W' \times D_1)$ . Following this, the reshaped features  $F'$  are duplicated a total of  $C_0$  times, where  $C_0$  represents the number of object classes within the dataset. For each object class, a dedicated channel-mixer MLP (Multi-Layer Perceptron) is employed, it comprises a linear transformation layer, followed

by non-linear activation and dropout procedures. An overview of the ORF module is depicted in Fig. 3.3, illustrating how each MLP layer learns to filter features specific to a particular object class. The functionalities within this module can be succinctly represented by the following equations:

$$F'_i = \text{ReLU}(\text{MLP}(F)) \quad (3.1)$$

$$F' = \text{DropOut}([F'_1, F'_2, F'_3, \dots, F'_{C_o}]) \quad (3.2)$$

With  $F' \in \mathbb{R}^{T \times C_o \times D_2}$ . Where  $D_2$  is smaller than  $D_1$  to shallow the channel size. To ensure an adequate supervision of the object representations, an additional MLP layer is introduced atop  $F'$ , representing the object classifier as depicted in Fig. 3.2. The transformation is described by Equation :

$$F'' = \text{ReLU}(\text{MLP}(F')) \quad (3.3)$$

Where  $F'' \in \mathbb{R}^{T \times C_o \times 1}$ . Due to the absence of frame-level object labels within the dataset, the object classifier is trained using pseudo labels generated by an object detector, in our case we used Fast-RCNN [42]. Given the possibility of multiple objects appearing within a single frame in the video, the object classifier is trained by utilizing binary cross-entropy loss denoted as  $\mathcal{L}_{clip-objects}$ . Although the ORF module successfully outputs representations for each object class, further steps are essential to interrelate and refine these object representations, to facilitate the exploration of their interactions and action modeling. This necessitates the introduction of the subsequent module, *Object Relation Reasoning Module*, within our pipeline, detailed in the subsequent section.

### 3.3.3 Object relation reasoning module

In the preceding section, the transformation of clip representations into class-specific representations marks a pivotal step. This transformation lays the groundwork for the subsequent mapping of these representations onto a graph-like structure. This graph is characterized by vertices that correspond to individual object classes across different time steps, with each vertex (node) representing the previously derived embedding of a specific class. This graph comprises a total of  $C_o \times T$  nodes, where  $C_o$  stands for the number of object classes and  $T$  denotes the number of time steps. Its structure topology is intricately defined by its vertices and an adjacency matrix denoted as  $A'_{C_o}$ . The significance of this adjacency matrix lies in its representation of the interconnections or relationships among the various nodes (objects) within the graph. The adjacency matrix serves as a critical descriptor of connectivity, outlining the relationships between different nodes at diverse time steps. Notably, its weights encode the strength or intensity of these relationships at distinct points in time. This intricate matrix captures the evolving dynamics of interactions between object classes over the temporal dimension, offering insights into how their relationships wax and wane throughout the sequence. Essentially, this graph-based representation method encapsulates the evolving relationships among object classes over time, creating a framework that not only delineates their connections but also quantifies the intensity or significance of these associations. By encapsulating these

complex temporal relationships within a structured graph format, this methodology provides a powerful framework for understanding the dynamics and interplay between object classes across sequential data. Fig. 3.4 represents an overview of this module.

### Graph reasoning

The essence of graph reasoning lies in its endeavor to conduct cross-class reasoning within the framework of the constructed graph. This methodology is purposefully designed to navigate the intricate web of relationships between various object classes encoded within the graph structure. The object relations, crucial to understanding the dynamics of the video sequences, exhibit a high level of dependency on the specific content of the video. To tackle this intricate dependency, multiple Graph Convolutional Network (GCN) blocks are strategically stacked within the architecture. This stacking facilitates the learning of diverse levels of semantic understanding by iteratively processing and refining the representations derived from the graph. Each GCN block contributes to capturing different facets of the relationships among object classes, enabling a comprehension of the video-specific object interactions. Additionally, the adjacency matrix, a fundamental component defining the interconnections between nodes in the graph, is parameterized for adaptability. This adaptability enables the matrix to be optimized and fine-tuned throughout the training process, dynamically adjusting its structure to best suit the data it encounters. Through this process, the adjacency matrix evolves to adapt to the unique characteristics and nuances presented by different datasets, thereby enhancing its efficacy in encapsulating the intricacies of class relations within diverse videos. Furthermore, self-attention mechanisms are employed, enriching the learning process by enabling the adjacency matrix to focus on and weigh different class relations dynamically. This adaptive attention mechanism empowers the adjacency matrix to discern and prioritize class relations based on the inherent characteristics and complexities embedded within the video data. Consequently, the adaptive nature of our adjacency matrix, honed through the integration of GCN blocks and self-attention mechanisms, fosters a more sophisticated understanding of class relations. This adaptability enables the matrix to differentiate and discern nuanced class relations across varying video contexts, enhancing the model ability to capture and interpret the intricate dynamics of object interactions specific to different video sequences. This adaptability and sophistication form the bedrock for more robust and context-aware reasoning within the constructed graph, promising advancements in understanding complex relationships within video data across diverse scenarios and contexts. Fig. 3.4 represents a block of the graph convolution reasoning.

As the object relations are complex, it is hard to predefine the inter-object relations for each video. Therefore, by leveraging the self-attention mechanism [216, 183], our graph adjacency matrix is learnable and can vary with the videos. In practice, the adjacency matrix  $A_{C_o}$  is initialized with a fully connected matrix. Finally, the full topology of the graph is  $A_{C_o} \in \mathbb{R}^{C_o \times C_o}$  and the vertexes representation  $G_{in} \in \mathbb{R}^{D_2 \times T \times C_o}$ . First, we embed the input  $G_{in}$  using a bottleneck convolutional layer (i.e.  $1 \times 1$ ), then the output feature maps are rearranged into  $\mathbb{R}^{D_2 \times T \times C_o}$  and  $\mathbb{R}^{C_o \times D_2 \times T}$  followed by a matrix multiplication. The value of the resultant matrix is then normalized by a softmax activation. Now, the superimposed adjacency matrix  $A'_{C_o}$  can be formulated as:



$$A'_{C_o} = A_{C_o} + \text{softmax}(W_1^T G_{in}^T W_2 G_{in}) \quad (3.4)$$

Where  $W_1$  and  $W_2$  are learnable weights of the bottleneck convolutions, and  $G_{in}$  being  $F'$  the stacked class representations in section B.  $G_{out}$ , the output of the graph layer is passed to the next graph layer and follows the same equations. In this work, we use 5 blocks of graph convolutions. As for the  $A'_{C_o}$ , each value represents an edge between two nodes (objects). We learn a graph that is shared across different time-steps but depends on each layer and for each video, as we said earlier we learn different semantics at each level.

After bottleneck convolutions, we do the graph convolution operation with the formulation in [117]:

$$G_{out} = A'_{C_o} G_{in} W_3 \quad (3.5)$$

$W_3$  is a learnable parameter where  $W_3 \in \mathbb{R}^{D_2 \times D_2}$ . The equation 3.5 represents the message passing and node feature updating, and finally  $G_{out}$  is rearranged to  $\mathbb{R}^{D_2 \times T \times C_o}$ .

From equation 3.5, we can understand how graph convolutions work. The graph convolutional layer represents each node as an aggregate of its neighborhood, hence each node gathers information from its neighborhood and adapts itself accordingly. In other words, at each graph block, each object collects information about other objects and finally finds to which ones it is most correlated, and thus whether there is an interaction or not. That is why we judge that the use of graphs is a promising idea in this domain.

## TCN

TCN stands for Temporal Convolution Network. The graph reasoning is capable of extracting the relation between objects. However, in our study, we aim at modelling the spatio-temporal interactions in a large time span. To do so, we add a 1D convolution layer on top of the previous output of the graph reasoning (i.e.,  $G_{out}$ ). As shown in Fig. 3.4, each *Object Relation Reasoning Module* contains a TCN. This 1D-convolution layer is used to aggregate the information across time. While stacking multiple object relation reasoning blocks, each block is used to model the object relation in a specific temporal scale. Finally, the output of the *Object Relation Reasoning Module* is:

$$G_{out} = \text{Conv1D}(G_{out}) + G_{in} \quad (3.6)$$

As mentioned earlier, the output of each block  $G_{out}$  is the input  $G_{in}$  to the next block.

### 3.3.4 Predictions

Predictions are based on the learned nodes and adjacency matrix. However, since in our case the actions are composed of verbs and nouns, we show that using the adjacency matrix for predicting the verb and the object feature representation for noun prediction is more effective. This makes sense since the adjacency carries more information about how different objects interact with each others, while the nodes carry a refined objects representations, after been processed through the different graph

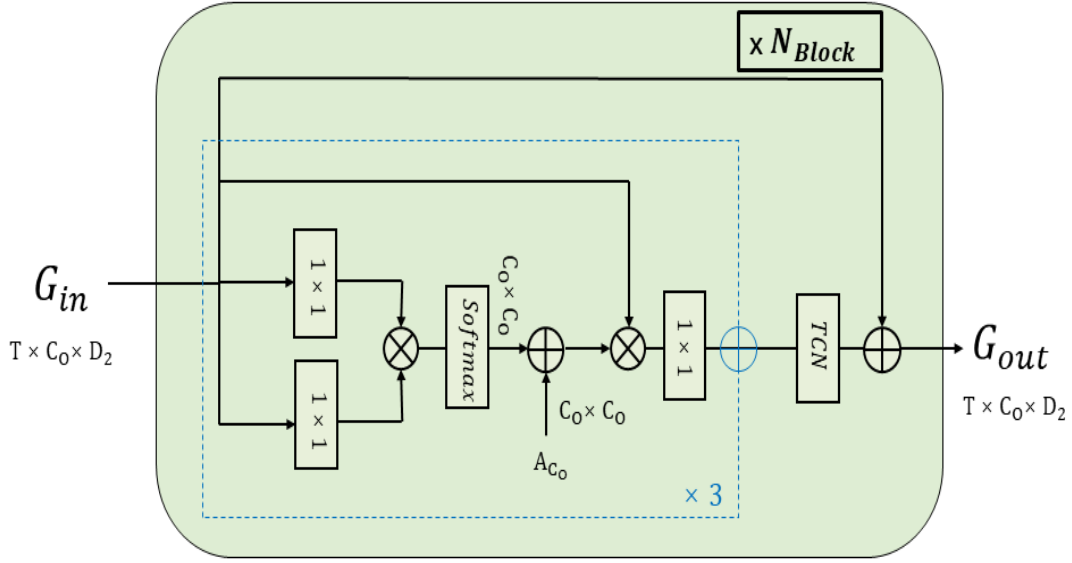


FIGURE 3.4: Overview of one layer of the Object Relation Reasoning module, using a graph architecture [183]. The input is a graph representation between different classes and the output is an updated representation of the graph. The  $\times N_{block}$  stands for the number of blocks used in total, while the  $\times 3$  at the bottom in blue stands for the number of used multi-head attentions.

convolutions blocks. Our final layers are two fully-connected layers one projecting  $G_{out}$  from  $\mathbb{R}^{D_2 \times C_o}$  to  $\mathbb{R}^{1 \times C_o}$ , and the other fully-connected layer projecting  $A'_{C_o}$  from  $\mathbb{R}^{C_o \times C_o}$  into  $\mathbb{R}^{1 \times C_v}$ , where  $C_o$  and  $C_v$  stand for the number of object classes and verb classes respectively.

Since we have 3 outputs, our loss is a sum of three losses and can be formulated as:

$$\mathcal{L} = \mathcal{L}_{verbs} + \mathcal{L}_{nouns} + \mathcal{L}_{clip-objects} \quad (3.7)$$

$$\mathcal{L} = \mathcal{L}_{anticip} + \mathcal{L}_{likelihood} \quad (3.8)$$

TABLE 3.1: Ablation study on different settings. This evaluation is on the EPIC-KITCHEN dataset. Temporal nodes means using the final output of X3D of size  $T \times 2048$  to create nodes, while spatio-temporal nodes means using a mid layer of size  $T \times 7 \times 7 \times 432$  with more spatial information. Finally ADJ-matrix stands for using the adjacency matrix for predicting the verbs instead of using only nodes for nouns and verbs.

	verbs		nouns		actions	
	top1	top5	top1	top5	top1	top5
X3D	46.5	79.8	34.3	65.3	21.0	38.7
THORN/temporal nodes	55.8	82.86	39.9	66.37	26.8	44.0
THORN/temporal nodes + ADJ-matrix	60.3	86.0	41.1	66.9	30.1	47.3
THORN/spatio-temporal nodes + ADJ-matrix	61.0	85.9	42.9	67.9	30.5	47.5

$$\mathcal{L}_{likelihood} = \text{MSE}(TIM_{out}, Likelihood_p) + \text{MSE}(TIM_{out}, Likelihood_{1-p}) \quad (3.9)$$

Where  $\mathcal{L}_{verbs}$  and  $\mathcal{L}_{nouns}$  are the negative log-likelihood losses (since each action is composed of one verb and one noun). As described earlier, the  $\mathcal{L}_{clip-objects}$  is the loss to ensure the semantics of the object representation.

### 3.4 Experiments

**Dataset.** We evaluate our model on two of the largest and challenging datasets for first-view and human-object interaction action recognition.

**Epic-Kitchen55** [54] is a multi-faceted non-scripted recordings in native environments - i.e. the wearers' homes, capturing all daily activities in the kitchen over multiple days. Annotations are collected using a novel 'live' audio commentary approach. It contains 55 hours of recording of 32 different kitchens in 4 cities. This dataset has a total of 125 verbs and 352 nouns. Also have a multi-language narrations and a total of 39,594 action segments

**EGTEA Gaze+** [127] EGTEA Gaze+ is a large-scale dataset for FPV actions and gaze. It subsumes GTEA Gaze+ and comes with HD videos (1280x960), audios, gaze tracking data, frame-level action annotations, and pixel-level hand masks at sampled frames. Specifically, EGTEA Gaze+ contains 28 hours (de-identified) of cooking activities from 86 unique sessions of 32 subjects. These videos come with audios and gaze tracking (30Hz). We have further provided human annotations of actions (human-object interactions) and hand masks. The action annotations include 10325 instances of fine-grained actions, such as "Cut bell pepper" or "Pour condiment (from) condiment container into salad".

In both datasets, each action is a combination of a verb and a noun. Actions are relevant to different steps of preparing food (e.g. *clean kitchen*, *cut vegetables*, *prepar table*).

**Implementation.** We implement our method using X3D as the visual encoder where  $D_1 = 432$ ,  $H' = W' = 7$  and  $D_2$  is 128. We input a clip of 16 RGB frames for Epic-Kitchen and 25 frames for EGTEA Gaze+. We use a dropout probability of 0.3. For the *object relation reasoning* module,  $N_{Block}$  is 5 blocks.

For the temporal convolution network, we run our model with different values of the kernel size. As there was no impact on the results, we kept a kernel size of 9. In the training phase, we utilized Adam [116] to optimize the model with an initial learning rate of 0.00005. We scaled the learning rate by a factor of 0.1 with the patience of 5 epochs. The network was trained on a 4-GPU machine for 30 epochs. We evaluated our model using top1 and top5 accuracy on verbs and nouns for Epic-Kitchen, while for EGTEA Gaze+ we evaluated directly on actions using top 1 accuracy.

### 3.4.1 Ablation study

In this section, we validate our model design for the modules in the THORN. The evaluation is conducted on the EPIC-Kitchen dataset. We propose different settings, and see how each setting can improve the performance. In table 3.1, we can notice different results:

First, we compare our baseline model X3D with THORN. Note that, in THORN, the graph nodes can be constructed either using the output of the last layer of X3D (temporal nodes) or using its intermediate layer (spatio-temporal nodes). Here, we first compare X3D with THORN (temporal nodes), i.e., we construct the nodes by the features in shape  $T \times 2048$ . In this setting, nodes serve to predict both verbs and nouns. In this scenario, we improve nouns prediction by +5.6%, while, the verbs accuracy increased by +9.3%. This result proves the importance of the cross-object reasoning, compared to only capturing visual information from 3D-CNNs. This proves our intuition that Graph convolutions are particularly well-suited to capture relational information within complex structures, such as the intricate networks of interactions between humans and objects. By treating objects as nodes and their relationships as edges in a graph, graph convolutions allow the model to propagate information efficiently across the interconnected elements of a scene. This graph-based approach inherently aligns with the nature of human-object interaction datasets, where the relationships between objects often exhibit non-linear and context-dependent patterns. Unlike 3D-CNNs, which primarily focus on extracting features from volumetric data.

Secondly, we study the importance of the adjacency matrix for predicting the verbs. To do so, we use the adjacency matrix (ADJ-matrix) to predict verbs, while keeping the nodes to predict the nouns. In this setting, the verb prediction improves by +4.5% compared to the previous setting and by +13.8% to the baseline X3D. Utilizing the edges of a graph proves to be more effective in predicting actions than focusing solely on nodes due to the critical role that relationships play in understanding dynamic interactions. In the context of graph-based models for action prediction, edges represent the connections or dependencies between entities, offering a pathway to capture the contextual information crucial for inferring actions. While nodes may encapsulate individual entities, it is often the relationships and interactions between these entities that dictate the unfolding events. By emphasizing edges, the model can effectively encode the temporal and spatial dependencies, enabling a more nuanced grasp of how actions propagate through a system. This approach inherently acknowledges the importance of context and sequencing in action prediction, allowing the model to discern subtle variations in behavior based on the dynamic interplay between entities. Consequently, prioritizing edges in a graph-based model offers a more comprehensive and contextually rich representation, leading to improved accuracy and interpretability in predicting actions within complex systems.

Thirdly, we study the effect of changing the temporal nodes with the spatio-temporal nodes. Spatio-temporal nodes are the nodes constructed by the middle layer of X3D which contains the spatial information ( $T \times 7 \times 7 \times 432$ ). With spatio-temporal nodes, THORN improves +1.8% on nouns. This is because, with spatial dimensions, the ORF can better capture the object relative locations and the size of

TABLE 3.2: Ablation study on fusing the scores of THORN with the scores from the object detector (Faster RCNN). This evaluation is on EPIC-KITCHEN dataset. Fusing both scores brings significant improvement on top-1 accuracy. For the object detector, we use an average pooling on all the video clip frames object detection scores and add a threshold of 0.3

Faster-RCNN scores	THORN	Nouns
✓	×	31.5
×	✓	32.8
✓	✓	42.9

the object, then embed them in the node representation. As a result, the noun accuracy improves. This setting also brings +0.7% improvement on verbs.

Our overall architecture obtains +13.8% more accuracy on verbs and +8.6% on nouns w.r.t. vanilla X3D. This reflects the importance of our proposed modules in THORN and how an object-centric method can improve results on human-object interaction actions.

We then study the components for predicting the nouns in our model. In table 3.2, we show that fusing scores of object detection and the scores obtained by the THORN nodes representation works better than using only one of them. We also find that predictions using only our model are better than the object detector itself. This shows that our model can refine the objects represented by the other objects (nodes) using our graph-based module.

### 3.4.2 Comparison with the State-of-the-Art

We compare our proposed method with the state-of-the-art methods on EPIC-Kitchen and EGTEA Gaze+ in table 3.3 and 3.4.

In Table 3.3, we compare our results with the state-of-the-art methods. Among these methods, Long Features Bank (LFB) [230] proposes to use global as well as local features for action recognition. To do so, they extract features on both clip and video levels, and combine them to have a better understanding of the scene. Nevertheless, this method still lacks accuracy for the objects. Moreover, LFB is a two step method which trains separately an object and verb recognizer modules. For our THORN, we train a single model for predicting both entities. As a result, we have a +8.5% improvement on top 1 nouns and a +4.9% w.r.t. LFB on action recognition.

Our method achieves the overall best performance. We claim that AssembleNet++ utilizes additional modality such as optical flow in both training and inference time. Even though, we still have the lead in top 1 accuracy for the verbs, nouns and actions, which proves again that having an object-centric and specific reasoning on object interactions is a key solution for having a better action recognition on HOI datasets. Finally, our results prove that using only RGB with an object-centric model achieves better or similar results compared to methods relying on heavy multi-modality reasoning.

TABLE 3.3: Comparing THORN model with other state-of-the-art methods on the validation set. Even though some of these comparisons are not fair since these models are using multi-modalities, we still hold the overall best accuracy, which shows the strength of our model

Model	Obj	RGB	Flow	Audio	Verbs	Nouns	Actions
					top1	top1	top1
Baradel [13]	×	✓	×	✓	40.9	-	-
3D-CNN	×	✓	×	×	49.8	26.1	19.0
STO [230]	✓	✓	×	×	51.0	26.6	19.5
LFB [230]	✓	✓	×	×	52.6	31.5	22.8
AssembleNET++ ODF+SDF [224]	✓	✓	✓	×	<b>60.0</b>	37.1	25.2
<b>THORN</b>	✓	✓	×	×	<b>61.0</b>	<b>42.9</b>	<b>30.5</b>

TABLE 3.4: Comparing THORN model with other state-of-the-art methods on EGTEA Gaze+ split1. We hold the best accuracy on actions

	Two-stream	I3D [34]	TSN [225]	ego-rnn [201]	LSTA [200]	SAP [228]	<b>THORN</b>
ACC %	43.8	54.2	58.0	62.1	62.0	64.1	<b>67.5</b>

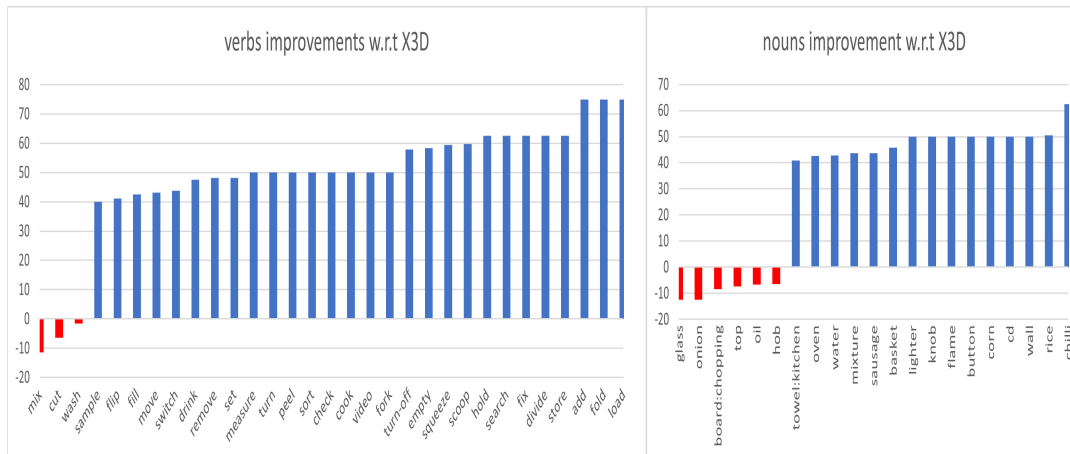


FIGURE 3.5: Accuracy improvement on nouns (right) and verbs (left) w.r.t X3D.

In table 3.4, we compare our method with the state-of-the-art on EGTEA Gaze+ dataset. We have the best accuracy w.r.t. the others methods, which shows the generalization and robustness of our model on actions of HOI.

To sum up, compared to other methods, ours is lightly weighted as we use X3D, while other methods rely on heavy 3D-CNNs such as I3D. THORN is trained jointly on nouns and verbs as opposed to other methods such as LFB [230], and we only need RGB frames and object classes per-frame.

### 3.4.3 Qualitative study

In this section, we conduct a qualitative study of THORN. In Fig. 3.5, we show the impact on some classes after adding our proposed module w.r.t. vanilla X3D. In

EPIC-Kitchen, we significantly improve accuracy on 28 verb classes. Only the accuracy of 3 out of 125 verbs decreases, while the decrease is negligible. This improvement on verbs shows that understanding the inter-relation of different objects is important for HOI. X3D outperforms our object-centric method on the action mix, particularly in the case of motion actions, owing to its specialized architecture designed for capturing temporal dynamics. Motion actions, such as those involving body movements or gestures, often exhibit distinctive patterns that can be effectively captured by models emphasizing temporal information. X3D excels in this regard by incorporating spatiotemporal convolutions that specifically focus on modeling the evolution of features over time. Unlike our object-centric method, which places emphasis on capturing cross-object relationships, X3D's strength lies in its ability to discern and exploit the temporal dependencies inherent in motion-based actions. Since these actions primarily involve changes in position and appearance over consecutive frames, the temporal modeling capabilities of X3D prove to be instrumental in achieving superior performance on the action mix dataset. In scenarios where cross-object relation reasoning may not be as critical, the efficiency of X3D in capturing nuanced temporal patterns positions it as a more adept solution for tasks centered around motion-based actions within the action mix dataset.

As for nouns recognition, remarkably, THORN, leveraging graph convolutions and object-centric reasoning, exhibits a notable capability to predict certain classes like "water" and "wall" that were previously challenging for the object detector alone. The traditional object detector struggles to effectively identify these classes, often resulting in low detection rates. The breakthrough achieved by THORN lies in its unique reasoning process, particularly in the intricate handling of cross-object classes. By engaging in cross-object reasoning, THORN refines the representation of nodes within the graph, allowing for a more nuanced understanding of the scene. Consequently, the model can leverage these refined representations to predict classes that might be overlooked by conventional 3D-CNNs, showcasing the superior capacity of graph convolutions and object-centric reasoning in capturing complex relationships and nuances within a dataset for enhanced noun recognition.

We show also the strength of using the adjacency matrix and the attention mechanism.

In Fig. 3.6 we show an example of the learned adjacency matrix for the action *wash knife*. In this figure, we find that there is a high correlation between the classes *knife* and *water* in both directions. Whereas the classes *tap*, *fish* and *sponge* are only correlated to themselves since they are not directly relevant to the objective action class *wash knife*. This example shows the effectiveness of THORN to capture the inter-object relations in the clipped HOI videos.

The object representation filter is one of the main parts of our architecture as it allows to extract a good representation for different objects related to the actions. To make sure our filtering works, we extract the activation maps for the different object and see what do they highlight in the scene.

Figures 3.7, 3.8, 3.9, represent different actions with their Class Activation Map (CAM). The example in Fig. 3.7 represents the action *wash leaf*, when looking at the output of the object representation filter the highest activations were on the classes *leaf* and *tap*. As discussed, we want to learn features specific to each class. The CAM of tap and leaf in this example clearly shows that only the pixels of the relevant

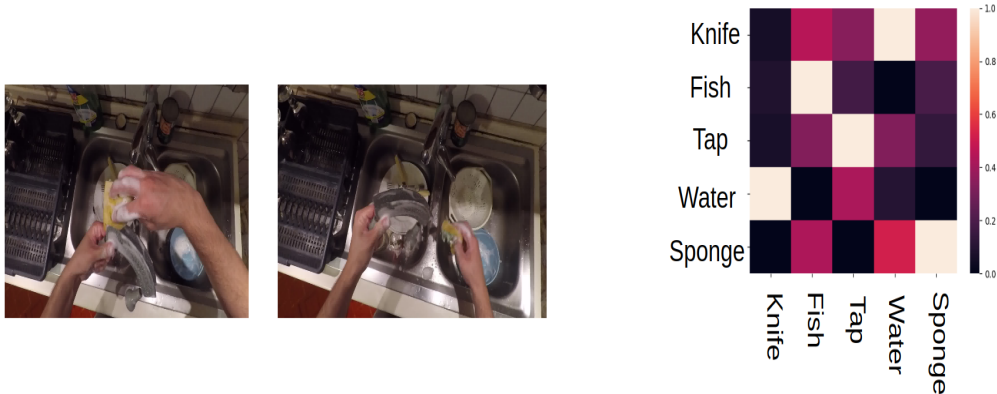


FIGURE 3.6: Example of the learned adjacency matrix of the action from Epic-Kitchen55 dataset. We notice a strong correlation between the classes *knife* and *water* for the action *wash knife*. Thus, we are able to collect high inter-class relation to recognize the right verb and its relevant objects. Moreover, the irrelevant classes such as *fish* are not activated, showing robustness of the learned attention.

objects were highlighted, hence, the features in the nodes are more representative of the objects of interest. Moreover, this result shows that our work does similar work to unsupervised object segmentation. Hence, unlike other methods that rely on pre-trained object detectors and tracking methods to extract object and then use ROI-Align to extract objects features, our method is capable of yielding the same result in a unsupervised manner and in a simplified way. Besides that, our THORN model learns to only focus on objects of interest. Our model distinguishes itself by prioritizing relevant objects in a scene, even when confronted with the presence of similar objects, for instance in Fig. 3.8 where we have the salad and the leaf are semantically very similar. This targeted focus on objects directly associated with the performed actions enhances the precision of action recognition in complex scenarios. In situations where multiple similar objects coexist, traditional models may struggle to discern the most pertinent entities, leading to potential confusion and misclassification. Our model, however, employs sophisticated object-centric reasoning to identify and emphasize the objects most relevant to the ongoing actions. By dynamically adapting its attention to salient entities within the context of the action, the model minimizes the impact of scene complexity and irrelevant objects. This strategic focus not only bolsters the accuracy of action recognition but also ensures that the model’s predictions align more closely with the subtle interplay between objects and actions, thereby enhancing the overall robustness and effectiveness of the recognition process.



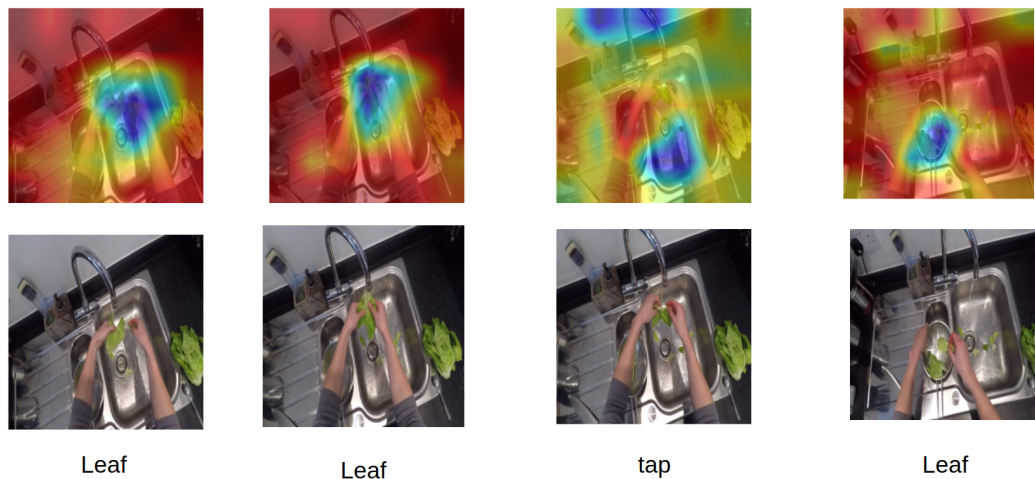


FIGURE 3.7: Example of action *wash leaf*. the highest activated classes were leaf and tap and when inferring the class activation map we can see that most activated pixels are around the objects of interest. Hence, the features extracted are more significant, which makes it easier to predict the right action.

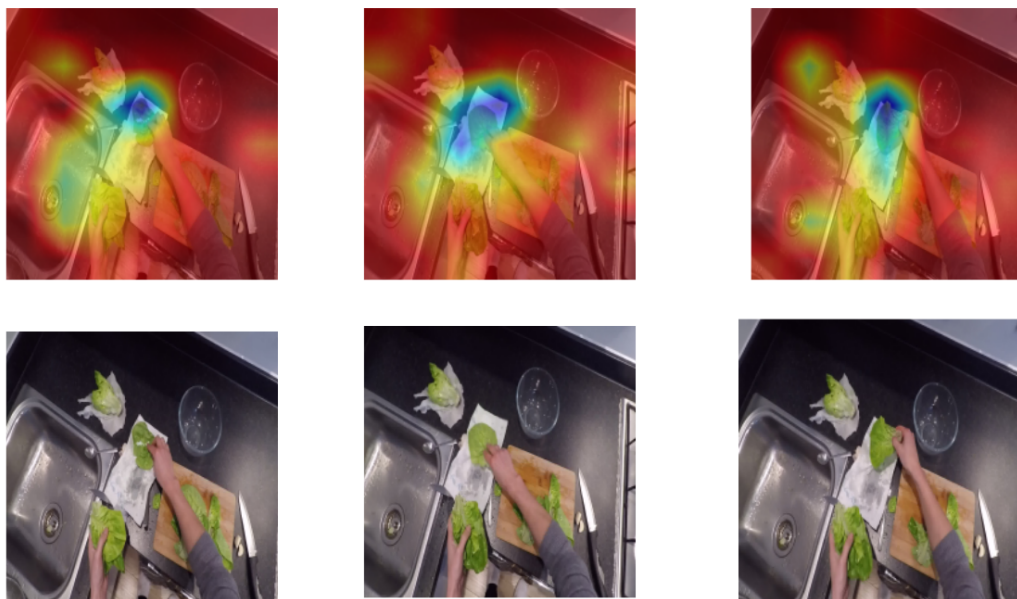


FIGURE 3.8: Example of action *put leaf*. In this example the most activated object was leaf and its activation map shows that the focused-on pixels actually belongs the leaf, proving the strength and robustness of our approach.

### 3.5 Conclusion

First-person view action recognition heavily relies on capturing and comprehending the visual connections existing between various objects and the human subject. In our study, we propose an innovative object-centric model that involves a sequential process: initially projecting the conventional CNN (Convolutional Neural Network) features into features specific to object classes. This step enables a more focused understanding of the scene by emphasizing the individual objects present.

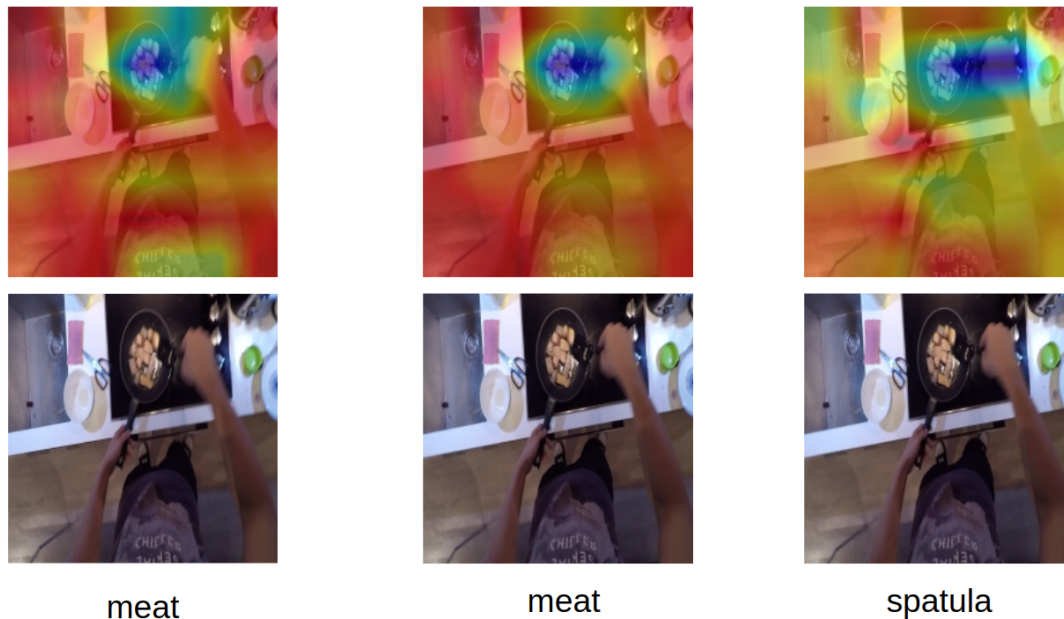


FIGURE 3.9: The action in this figure is *mix meat*, and looking at the figure we notice that the highlighted pixels are the ones corresponding to the spatula and the meat. Therefore, it is easier to predict the right action.

Subsequently, we engage in graph reasoning, a pivotal phase where we analyze the interrelations between different objects. Here, each object class is represented as a node within a graph, and the connections or edges between these nodes signify the relationships existing between distinct objects. This approach allows us to model and understand the contextual dependencies and interactions between objects within the scene. To validate the efficacy of our model, we conducted evaluations using two extensive and demanding datasets. The results demonstrate that our method, named THORN, achieves state-of-the-art performance on both datasets. This success underscores the effectiveness and robustness of our approach in comprehending and recognizing actions within the first-person perspective. Considering that our method performance heavily relies on the precision of object detection, our forthcoming efforts will concentrate on crafting an architecture that seamlessly integrates both object detection and action recognition tasks. This integration aims to enhance the overall performance and accuracy of our model by leveraging the synergy between these interconnected tasks. Furthermore, an area of interest for our research involves extending our model capabilities to encompass first-view action detection within untrimmed videos. This expansion would enable our model to not only recognize actions but also accurately detect their occurrence within continuous and unsegmented video streams, further broadening the application potential of our approach.



## Chapter 4

# JOADAA: joint online action detection and action anticipation

In the previous chapter we discussed a very important aspect of action understanding which was action recognition in fine-grained activities. To have a full view of action understanding, this chapter introduces action detection. As a real world application, we are going to focus on online action detection and action anticipation. Action anticipation involves forecasting future actions by connecting past events to future ones. However, this reasoning ignores the real-life hierarchy of events which is considered to be composed of three main parts: past, present, and future. We argue that considering these three main parts and their dependencies could improve performances on AA(action anticipation).

On the other hand, online action detection is the task of predicting actions in a streaming manner. In this case, one has access only to the past and present information. Therefore, in online action detection (OAD) the existing approaches miss semantics of future information which limits their performance. To sum up, for both tasks, the complete set of knowledge (past-present-future) is missing, which makes it challenging to infer action dependencies, therefore leading to low performances. To address this limitation, we propose to fuse both tasks into a single uniform architecture. By combining action anticipation and online action detection, our approach can cover the missing dependencies of future information in online action detection. This method, referred to as JOADAA, presents a uniform model that jointly performs action anticipation and online action detection. We validate our proposed model on three challenging datasets: THUMOS'14, which is a sparsely annotated dataset with one action per time step, CHARADES, and Multi-THUMOS, two densely annotated datasets with more complex scenarios. JOADAA achieves SOTA results on these benchmarks for both tasks.

### 4.1 Introduction

Envisioning upcoming occurrences plays a vital role in human intelligence as it aids in making choices while engaging with the surroundings. Humans possess an inherent skill to predict future happenings in diverse situations involving interactions with the environment. Likewise, the capacity to anticipate events is imperative for advanced AI systems operating in intricate settings, including interactions with other agents or individuals. The goal of online action detection (OAD) is to accurately pinpoint ongoing actions in streaming media, by predicting impending events. Action anticipation can help improve OAD as it imitates the capacity of human cognition to anticipate events before they occur. Therefore, OAD and action

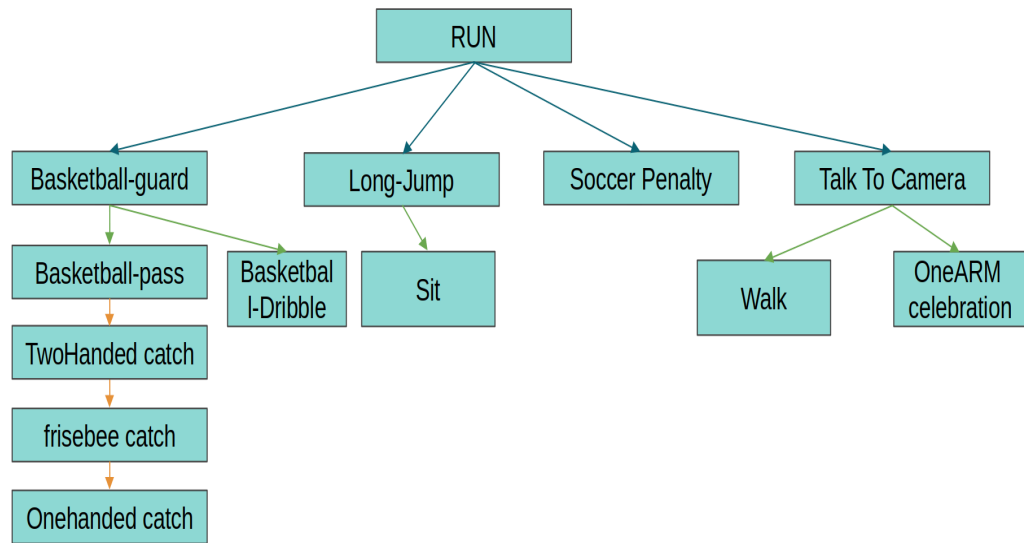


FIGURE 4.1: An example of human non-sequential dependencies. For instance, the actions *RUN* and *OneHanded Catch* are highly correlated but distant in time. Also the same start action *RUN* can lead to many different actions and scenarios. Therefore, it is very hard for online action detection or action anticipation to detect such relations without access to the future. In JOADAA, we propose to tackle this limitation by introducing a pseudo-future information by combining action anticipation and online action detection in the same task.

anticipation are two important areas of research in computer vision, which have numerous applications in security surveillance, home-care, sports analysis, self-driving cars, and online danger detection. Human perception of actions can be viewed as a continuous cycle in which prior knowledge is used to forecast future behavior, and present knowledge is used to revise and update future predictions.

To tackle action detection, we propose a unified framework of action anticipation and online action detection. Our predictions are in two steps, first we anticipate upcoming actions based on past information. Second, we update the anticipation by introducing the present information. By doing so, we gain in the online action detection by introducing the anticipated actions as pseudo-future information. In addition, it improves the action anticipation by comparing the prediction to the present information, thus combining them to improve both tasks.

Transformer networks such as [218, 134, 8] have had a significant success in computer vision and video understanding. This is due to their ability to capture long-range dependencies, which enables them to handle more effectively sequential data. LSTR [236], TesTra [249], or FUTR [77] have benefited from the transformer backbones to address the tasks of OAD and AA (action anticipation). However, OAD and AA tasks suffer from limited information as they don't have access to future information and global knowledge of the scene. This limited information restricts the ability of transformers to capture long-range dependencies and to learn significant relations between events. This can be demonstrated by comparing the effectiveness of models for offline action detection with online action detection. Offline, one has access to all pieces of information and a clear knowledge of the past, present, and

future. Furthermore, complex densely annotated datasets (such as Multi-THUMOS [239]) have not been explored for online action detection and anticipation. It is challenging to recognize and foresee activities in such datasets. Most OAD architectures are only validated on sparsely-annotated activity datasets. Such simple annotated datasets are less challenging. First, these datasets do not have co-occurring actions. Second, they rarely have dependencies between actions in distant time steps. Furthermore, actions in densely annotated datasets have many possible outcomes. An example of these complex dependencies is given in Figure 4.1. Due to these challenges, OAD methods are only validated on simple datasets. Therefore, even with the help of transformers, it is difficult to build knowledge of these long-range dependencies without having access to complete information.

In the past, OAD and action anticipation have been treated as separate tasks. However, to tackle the above challenges, we propose JOADAA (Joint Online Action Detection and Action Anticipation) to tackle OAD and AA together. We create a pseudo-future when performing online action detection. By leveraging cross-attention between the real frame features and the anticipated frames, we enhance the quality of the features, thus improving the accuracy of the predictions by making the present aware of a pseudo-future.

We propose to extract two types of information from these updated features: local dependencies using TCNs (temporal convolution networks) and global dependencies using MHA (multi-head attention). Finally, we fuse both pieces of information to make online action detection predictions.

Following previous work, we extract features from video clips using 3D convolution neural networks (3D CNNs). We use I3D [33] as a pre-trained backbone on the Kinetics dataset [110]. We store these extracted features in a memory bank. JOADAA consists of three main parts i) **Past Processing Block**, ii) **Anticipation prediction Block**, and iii) **Online action prediction Block**. First, we capture past information using a transformer encoder. The encoder output is first passed through a classification layer, which helps improve the quality of the embedding by making it class-dependent. Next, in the anticipation prediction part, we assume that we have not yet got the current frame. A transformer decoder is employed to learn from the last layer of the past embeddings to anticipate the upcoming actions in the next frame. This is carried out by introducing a set of learnable queries, called *anticipation queries*. Finally, the online action prediction part uses anticipation embedding and current frame features to enhance the quality of the current frame. The new enhanced present frame features are fused with past features. Finally, global and local information is extracted using MHA and TCN layers respectively, achieving a new enhanced feature map. Based on the challenges discussed, we propose the following main contributions:

- We design a new architecture **JOADAA**, to jointly perform online action detection and action anticipation.
- We tackle both tasks for two different types of datasets, a densely annotated dataset and a simple activity dataset.
- We validate our proposed method on three benchmark datasets and achieve new SOTA results for online action detection and action anticipation.

## 4.2 Related work

**Online Action Detection** Action detection is the task of localizing action instances in time steps. We distinguish two types of action detection i.e., offline and online. In off-line action detection, the model has access to the entire video [187, 189, 232, 250, 50]. Online action detection, on the other hand, occurs in real-time and has access to the past and the present only. [55] explicitly introduced online action detection task for the first time and proposed TVSeries dataset. After that, they also proposed a two-stream feedback network with LSTM [20] to model temporal structure [11]. RED [70] uses reinforcement loss to encourage early recognition of activities. IDN [61] learns discriminative features and stores only knowledge that is relevant in the present. [237] uses LSTM to predict future information recursively and combine it with past observations to identify actions. Note that the aforementioned methods adopt RNN to model input action sequences, which are inefficient and lack of interaction between features, resulting in poor modeling capabilities for long-term dependence. To achieve optimal features, LAP-Net [166] presents an adaptive sampling technique. PKD [248] uses curriculum learning to transfer information from offline to online models. Shou et al. [190], like early action detection, focus on online detection of actions starts (ODAS). StartNet [72] divides ODAS into two stages and learns using a policy gradient. WOAD [73] employs video-level labeling and weakly-supervised learning. LSTR [236] uses a set of encoder-decoder architectures to capture the relations between long-term and short-term actions. They achieve state-of-the-art results on sparsely-annotated datasets, but perform poorly on densely labeled datasets such as Multi-Thumas [239].

**Action Anticipation** is the task of predicting future actions given the limited observation of a video. In the past, many strategies have been proposed to solve the next action anticipation, forecasting a single future action in a matter of seconds. Recently, the idea of anticipating long-term activities from a long-range video has been put out. Girdhar and Grauman [75] introduced the anticipative video transformer (AVT), which anticipates the following action using a self-attention decoder, which was further improved by FUTR [77] for minutes-long future actions. However, their architecture is suitable only for simple activities and simple datasets, which is not applicable to real-world scenarios that have multiple actions occurring at the same time.

Whether it is action anticipation or online action detection, unlike TAL(Temporal Action Localization) [30, 71, 129, 132, 188] and TAP(Temporal Action Proposal) [129, 130, 207] which are both offline tasks having access to the entire video, online action detection does not have access to the future and requires causal understanding from history to present, Hence the remaining challenge is leveraging history is that for long untrimmed videos, its length becomes intractably long over time. To make it computationally feasible, some methods [61, 73, 166, 227] make the online prediction conditioned only on the most recent frames spanning less than a minute. This way the history beyond this duration that might be informative to current frame predictions is left unused.

Finally, in the study of mixing action anticipation and online action prediction, the authors in [249] use the same architecture for both action anticipation and online action detection tasks. However, they dissociate these tasks, while we tackle both tasks jointly to improve both of them. Furthermore, the architecture in [249] is very

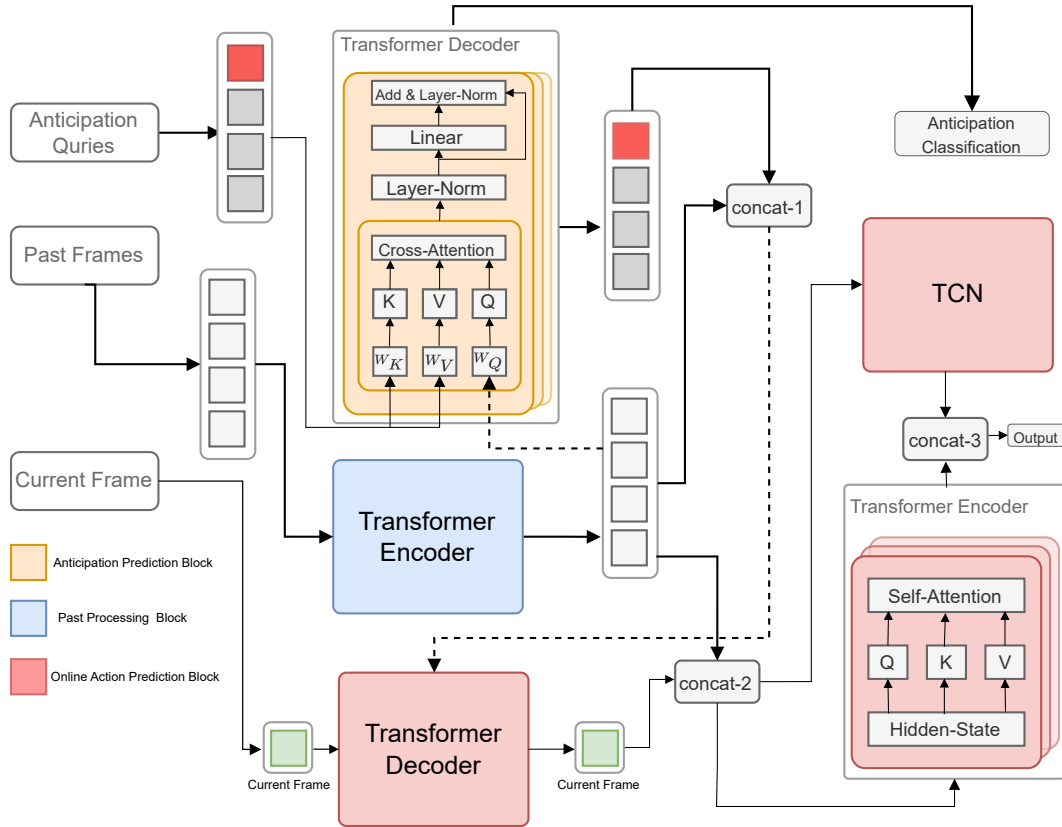


FIGURE 4.2: Proposed JOADAA architecture with three units i) Past processing, ii) Anticipation prediction, and iii) Online Action prediction. Each stage is highlighted by a color for better understanding.

similar to [236], therefore, the same limitations apply here as well.

In summary, to have adequate predictions, we need to build a well-descriptive hierarchy of information consisting of past, present, and future. Unfortunately, tasks such as online action detection or action anticipation do not have access to this global knowledge. In our work, we suggest to combine OAD and AA in order to create a pseudo-full knowledge that can improve action anticipation accuracy and produce comparable results for online action detection.

### 4.3 Proposed method

The entire architecture comprises three main components which are the Past Processing Block, the Anticipation Prediction Block, and the Online Action Prediction. These components are depicted in Figure 4.2. Let us delve into each part in more detail. First, the Past Processing Block utilizes a short-term past transformer-encoder to enhance the features. This process focuses on optimizing the quality and representation of the features extracted from the past. The second part is the Anticipation Prediction Block. Here, a transformer-decoder is employed for anticipation. It predicts the actions that will occur in the upcoming frames. To make these predictions, the block utilizes the output from the previous block in conjunction with a set of



learnable queries known as anticipation queries. This enables the system to anticipate the actions that are likely to happen in the future frames. Finally, the Online Action Prediction is performed by another transformer-decoder module. It takes into account the results obtained from the anticipation block as well as the past information.

### 4.3.1 Past processing block

In order to improve the accuracy when predicting ongoing actions, our model begins by extracting prior information. To accomplish this, we incorporate a transformer encoder into our framework. This encoder takes the embedding of previous frames as input, allowing us to emphasize significant and reliable frames through the use of attention mechanisms. By doing so, the features become more informative and reflect the activities that occurred prior to the current frame. It can be quite challenging to determine the specific activity that a person is engaged in solely based on the original embedding or the current frame. For instance, if the current frame shows the person *hold bottle*, we are not sure if the ongoing action will be *picking up the bottle*, *placing the bottle*, *drinking water*, or *pouring water*. However, if we know from the past that one of the previous actions was *opening the bottle*, we can be more confident that the person is more likely to *drink water*. These features are later used to anticipate future actions. Following [218], the equations below sum up the first block of our architecture:

$$F' = \text{ATTENTION}(F) \quad (4.1)$$

$$\text{ATTENTION}(F) = \text{Softmax}(QK^T / \sqrt{d_k})V \quad (4.2)$$

$$Q = W_q \times X, K = W_k \times X, V = W_v \times X \quad (4.3)$$

$$X = F + \text{PE}(F) \quad (4.4)$$

PE stands for positional encoding, and  $F \in \mathbb{R}^{T \times D}$  are the extracted features using the pre-trained I3D model [33], and  $W_q$ ,  $W_k$  and  $W_v$  are learnable weights.

In addition, we present various approaches to incorporate past information into our proposed method. Based on the research conducted by [236], our method utilizes both long-term and short-term past information. However, the effectiveness of employing these different types of past information is heavily influenced by the specific characteristics of the dataset used. Initially, the common belief is that providing more information to a neural network is beneficial as it allows for a more comprehensive understanding of the events occurring in a video. Particularly with the utilization of transformers, we are able to capture long-range dependencies and to learn the sequential progression of actions leading up to the current state. Nevertheless, through our own investigation, we have discovered that this assumption is

not always valid. For instance, the very long-past knowledge may sometimes harm performances, especially for densely annotated datasets. In scenarios where many actions co-occur, it is challenging to learn significant long-term relations, and thus these long-term features may act as noise to the model. Further experimental details are provided in Section 4.4.4.

### 4.3.2 Anticipation prediction block

Inspired by [77], the module takes a feature map  $F' \in \mathbb{R}^{T \times D}$  and a set of anticipation queries (learnable)  $LQ \in \mathbb{R}^{N_q \times D}$ , as inputs. Here,  $N_q$  represents the number of queries and  $D$  is the embedding dimension, which is the same as the feature map dimension. Action anticipation can be achieved in two different ways. The first way is to proceed directly with a transformer encoder and to learn to predict the future. An encoder sees only a glimpse of the past and learns to predict the future. On the contrary, another way is to utilize a transformer decoder. In this approach, the strength of using learnable queries with a transformer decoder is that each query learns a specific feature for a specific frame in the future. The positional encoding indicates to the transformer the order of these learnable queries and helps the model relate each query to a corresponding point in the future. Additionally, by having these learnable queries in our model, it learns to adapt to each clip, since the queries are based on the past information of each clip. Therefore, these learnable queries learn to be aware of the past. JOADAA uses these learnable queries as a link between past events and possible future ones.

$$N_q = 1 + N_f \quad (4.5)$$

Where in Eq. 4.5, 1 is for the upcoming frame that represents the ongoing action (represented in red in Figure 4.2). Since we do not have access yet to this frame, it is also anticipated.  $N_f$  is the number of frames to anticipate in the future to which we have no access. Information from the past, present, and future are connected by these learnable queries to improve both tasks efficiently. Later, these anticipation queries act as a pseudo-future to do the prediction of the ongoing action, see Section 4.3.3.

### 4.3.3 Online action prediction block

At this crucial stage of our methodology, we introduce a sophisticated approach that leverages both the features extracted from the current frame and the previously acquired knowledge of potential actions in the current time step and subsequent time steps. These combined features are thoughtfully integrated into a decoder, where our model gains the ability to classify the current frame with heightened precision, thanks to its access to what we refer to as 'pseudo-future knowledge'. The ingenious use of pseudo-future knowledge manifests in two significant ways. Firstly, the prediction of the current frame is intricately optimized by employing anticipation queries, allowing our model to anticipate potential developments and to make more informed decisions. This anticipatory aspect contributes immensely to the model ability to accurately classify the current frame. Furthermore, our approach also harnesses the power of the current frame itself, serving as a rich source of contextual information. We believe that, by enhancing the learned queries based on this real-time frame, our anticipation module receives a substantial boost in its capabilities, leading to even more refined predictions and a superior performance overall, this is

proven in the section 4.4.3 as we notice improvement in SoTA comparison on action anticipation . To further enhance the robustness of the JOADAA framework, we incorporate what we term ‘local-to-global layers’. These layers are pivotal in capturing nuanced local information, ensuring that our model doesn’t overlook any subtle yet crucial cues in the data, by using 1D temporal convolution layer, known as a TCN layer, which effectively captures temporal dependencies, allowing the model to understand the temporal evolution of actions. In the realm of global and long-range dependencies, we turn to transformers, a proven and powerful tool. These components within our framework excel at capturing intricate global patterns and dependencies within the data, further bolstering the model capacity to make nuanced predictions. In summary, our research not only leverages anticipation queries and real-time frame information but also integrates local-to-global layers, TCN layers, and transformers to create a comprehensive and high-performing JOADAA framework. This multifaceted approach empowers our model to excel in object detection and action anticipation tasks, making it a promising solution for a wide range of applications in computer vision and beyond. However, as explained earlier, this huge amount of information is not always helpful and may act as noise. Therefore, by mixing transformers with TCNs, our model learns complementary information from an updated feature map that we pass through an FC (fully connected) layer for classification. Notably, we utilize a Softmax layer for basic datasets with only one action at a time for validation and a Sigmoid layer for datasets with co-occurring actions in all categorization layers (past, future, and present).

Note that we use three different concatenation layers in our architecture. The first concatenation is between past frames features and anticipated frames features, the aim of this concatenation is to provide the decoder with a pseudo full information (past and pseudo future), which is the main idea of this work (use AA to enhance OAD). The second concatenation is between past frames and the currently updated feature (since the current frame feature is now aware of past and possible future actions). Here we only concatenate past and present because online action detection is our main objective, which is why there is no more need for future information. The last concatenation is to use both local information learned through the TCNs and global information from the transformer decoder, which allows us to have better predictions as shown in the ablation studies Table 4.8.

We also use the same decoder for future frame anticipation and current frame prediction. Experiments have been conducted that showed that using different decoders does not improve the accuracy and sometimes leads to a slight decrease in accuracy. Hence, to keep the model lighter and have better prediction we keep the same weights. As for the encoders, the two of them are different; the last encoder is part of our proposed classification head, where we use a TCN to capture local dependencies and a transformer encoder to capture long-range dependencies. Therefore, our intuition was not to share the weights between the encoders as they have a separate function in our architecture.

## 4.4 Experiments

In this section, we discuss experiments carried out for online action detection and action anticipation tasks on two different types of datasets. First, we briefly describe the datasets used and explain the implementation of the experiments carried out. Second, we compare JOADAA with existing SOTA methods for both online action

detection and action anticipation. Finally, we explore the effectiveness of each module of our approach by performing an ablation study. More qualitative results are provided in the supplementary materials.

#### 4.4.1 Datasets

We first briefly explain the datasets used in our experiments. We experiment on two types of datasets, i) a sparsely annotated dataset (THUMOS'14 [96]), and ii) densely annotated datasets (Multi-THUMOS [239] and CHARADES [247]). Each of them is described below.

**THUMOS'14** contains 413 untrimmed videos with 20 categories of actions. The dataset is divided into two subsets: the validation set and the test set. The validation set contains 200 videos, and the test set contains 213 videos. Following common practice, we use the validation set for training and report the results in the test set. More details are available in [96].

**Multi-THUMOS** contains dense, multilabel frame-level action annotations for 30 hours across 400 videos from the THUMOS'14 [96] action detection dataset. It consists of 38,690 annotations of 65 action classes, with an average of 1.5 labels per frame and 10.5 action classes per video. More details can be found in [239].

**CHARADES** is composed of 9,848 videos of daily indoor activities with an average length of 30 seconds, involving interactions with 46 object classes in 15 types of indoor scenes and containing a vocabulary of 30 verbs leading to 157 action classes. Readers can find more details in [247].

#### 4.4.2 Implementation details

We implement our proposed model in PyTorch [165]. All experiments are performed on a system with 3 Nvidia V100 graphics cards. For all Transformer units, we set their number of heads to 16 and hidden units to 1024 dimensions. To learn the weights of the model, we use Adam Optimizer [116] with weight decay  $5 \times 10^{-5}$ . The learning rate increases linearly from zero to  $5 \times 10^{-5}$  in the first 40% training iterations and then decreases to zero using a cosine warm-up. Our models are optimized with a batch size of 16, and trained for 25 epochs. **Evaluation protocol:** We follow previous work and use mean average precision per frame (mAP) to evaluate performances.

#### 4.4.3 Comparison with the SoTA

##### OAD comparison on the simple dataset (THUMOS'14)

Table 4.1 presents the results of online action detection. For the THUMOS'14 [96] dataset we achieve state-of-the-art results by a margin of 1.4%. GateHUB[38] was SoTA results for OAD on the THUMOS'14 dataset. However, they provide two results on this dataset, one with TSN as the backbone feature extractor and one with Timesformer[20]. Upon careful examination, we noticed the following points: 1) Our accuracy still surpasses theirs. 2) The GateHUB method was not compared with TesTra, which demonstrated better accuracy with the same settings. 3) GateHUB achieves SOTA results only when TimeSformer[20] is used as an RGB feature

	THUMOS'14	Multi-THUMOS	CHARADES
FATS[115]	59.0	-	-
IDN[61]	60.3	-	-
PKD[248]	64.5	-	-
WOAD[73]	67.1	-	-
LFB[230]	64.8	-	-
TRN[237]	62.1	39.5	18.3
PDAN[50]	62.2	32.6	16.0
MSTCT[49]	70.5	41.4	19.5
LSTR[236]	69.5	43.0	20.0
TesTra[249]	71.2	41.7	19.9
GateHUB[38]	70.7	-	-
JOADAA	<b>72.6</b>	<b>45.2</b>	<b>21.5</b>

TABLE 4.1: State of the art comparison for OAD on THUMOS'14, Multi-THUMOS, and CHARADES. Due to the lack of available OAD methods for CHARADES and Multi-THUMOS datasets, we compare also with two off-line methods PDAN and MSTCT, adapted to an on-line setting.

	THUMOS'14				Multi-THUMOS			CHARADES		
	1	2	4	6	2	4	6	2	4	6
TTM[227]	46.8	45.5	43.6	41.1	-	-	-	-	-	-
LSTR[236]	60.4	58.6	53.3	48.9	-	-	-	-	-	-
TesTra[249]	66.2	63.5	57.4	52.6	28.0	22.4	19.8	18.1	13.7	13.5
JOADAA	<b>67.7</b>	<b>63.9</b>	<b>62.9</b>	<b>59.3</b>	<b>42.5</b>	<b>37.7</b>	<b>35.2</b>	<b>20.2</b>	<b>19.5</b>	<b>19.0</b>

TABLE 4.2: Comparison with SoTA for the action anticipation task. 1, 2, 4, and 6 represent the number of anticipated frames. We notice that our method is more robust w.r.t. the number of anticipated frames compared to other methods where accuracy drops dramatically.

extractor, making it difficult to determine whether the results are due to the extractor or to their proposed solution. In conclusion, while the GateHUB paper argues for capturing relevant information from the past to the present, our JOADAA method, which employs a simple implementation of transformers, outperforms it along with TesTra[249].

### OAD comparison on densely annotated datasets

We evaluate JOADAA on more complex datasets such as Multi-THUMOS[239] and CHARADES [247]. We utilize LSTR [236], TesTra[249], and TRN[237] to train on these datasets to build baseline methods, as there are no validated online methods to compare JOADAA to these datasets. JOADAA improves the baselines by 1.5% on CHARADES[247] and 2.2% on Multi-THUMOS [239] dataset. The main difference between our approach and baseline methods [236] and [249], is the introduction of pseudo-future knowledge to our online action prediction. It helps make more precise predictions by having a knowledge of different possible outcomes.

Dataset	1	2	4	6
THUMOS'14	70.5 / 67.7	71.5 / 63.9	72.2 / 62.9	72.6 / 59.3
CHARADES	20.0 / 20.7	21.4 / 20.2	21.5 / 19.5	21.4 / 19.0
Multi-THUMOS	44.5 / 42.8	45.2 / 42.5	45.0 / 37.7	45.2 / 35.2

TABLE 4.3: Effect of **action anticipation** prediction and **online action detection** using long-short-term knowledge. 1, 2, 4, and 6 are the number of anticipated frames. Best viewed in color.

Dataset	2	4	6
THUMOS'14	70.6 / 64.4	70.0 / 63.0	70.6 / 58.2
CHARADES	21.8 / 20.4	21.4 / 19.5	21.3 / 19.0
Multi-THUMOS	45.1 / 36.9	45.3 / 39.2	45.1 / 37.3

TABLE 4.4: Results of using only short-term past information on multiple datasets for **online action detection** and **action anticipation**. 2, 4, and 6 are the number of anticipated frames.

#### OAD comparison using off-line methods

For further comparison, we adapt offline methods to online settings. We use PDAN[50] and MSTCT[49] two SoTA methods on CHARADES and Multi-THUMOS in off-line action detection. We outperform these two methods on all three datasets THUMOS'14, Multi-THUMOS, and CHARADES.

#### AA SoTA comparison

Similarly, our model achieves SOTA results on action anticipation as noted in Table 4.2. When increasing the anticipated frames from 1 to 6, TesTra [249] accuracy drops by **13.6%** on the THUMOS'14 dataset, whereas our model decreases by only **8.4%**, which shows robustness of our proposed solution. Also, JOADAA performs much better in more complex datasets (CHARADES and Multi-THUMOS).

In Table 4.3, we demonstrate how far we can foresee the future. We notice that, in general, the further we anticipate, the better the accuracy of the online action detection (**blue**) until it reaches a level where the accuracy stops increasing. Such a behavior makes sense because the model can learn more action dependencies by inferring more information about upcoming events. On the other hand, action anticipation results (**red**) decrease when the anticipation period increases, because the model has more space to explore.

#### 4.4.4 Ablation study

In this section, we discuss how the different modules contribute to JOADAA.

##### Ablation on the past processing block

First, we analyze the use of long-range past features on different datasets. As discussed in Section 4.3, past information can be used in two manners, either using only short-term past (32 frames) or long-short-term past (512+32 frames). This past information is used to infer the pseudo-future in our approach. In tables 4.4 and

Dataset	long term past + short term past		short term past	
	LSTR	JOADAA	LSTR	JOADAA
THUMOS'14	69.5	<b>72.6</b>	65.4	<b>70.6</b>
Multi-THUMOS	42.0	<b>45.2</b>	40.0	<b>45.1</b>
CHARADES	20.0	<b>21.4</b>	19.8	<b>21.3</b>

TABLE 4.5: Comparison of JOADAA with LSTR method using long-past information. JOADAA is more robust to utilize long-past information.

Module	THUMOS'14
Transformer encoder	71.5
LSTM+Conv	54.2

TABLE 4.6: Comparing two techniques for past information processing. We use a transformer encoder and a set of LSTM blocks with a convolution layer.

4.5, we observe that our model is more robust when it comes to using only short-term past information (decreases by 2%) on the THUMOS'14[96], unlike LSTR [236] where the accuracy decreases by 4.1%. One important result of our study is that long-past knowledge is more important for simple actions (single-action datasets) than for complex actions (densely annotated datasets). This is because numerous actions may occur simultaneously without being connected in densely annotated datasets, making it more challenging to infer relations from them. As a result, including information from the distant past can skew model predictions.

Recently, transformers have been widely used, since they outperformed the existing approaches such as 3D-CNNs and RNNs. In fact, 3D-CNNs are known to be good general feature extractors as they can capture overall visual appearances in a video. However, their CNN filters capture pixel-level information in a local neighborhood but struggle with long-term dependencies. Therefore, we limit the use of 3D-CNNs to extract video clip features for our architecture. Furthermore, action detection tasks require a strong grasp of long-range temporal dependencies, and transformers excel at capturing long-term information compared to RNNs. Therefore, the transformers are the best choice for OAD and AA tasks. However, most papers lately use transformers based on the previous intuition without any justification.

Table 4.6 presents a comparison study between RNNs (LSTMs[198]) and transformers. We replace our first encoder for past information processing with 3 blocks of LSTM and a convolution layer to reduce the feature map size. Results show that transformers are better suited for capturing long-range dependencies and produce far more better results which justifies our design choice.

### Ablation on the action anticipation module

Another ablation study is done in Table 4.7. We conduct two main experiments: one with the full JOADAA model and the other one without the Action Anticipation (AA) module. We can see that the AA module enhances online action detection, which supports our claim that combining AA and OAD leads to better results.

Dataset	OAD+AA	OAD
THUMOS'14	72.6	71.2

TABLE 4.7: Analyzing the JOADAA behavior with and without action anticipation.

#### Ablation on the OAD prediction layer

Dataset	TCN+TR. Encoder	FC
THUMOS'14	72.6	69.7

TABLE 4.8: Effect of fusing local and global information on OAD. FC stands for fully-connected layer. As expected capturing different type of dependencies provides better results.

Table 4.8 shows the effect of fusing local and global knowledge, in contrast to using directly the output of the decoder on the current frame which carries only global information in it. By doing so, our results increase by **2.9%**. As argued earlier, this is due to the fact that TCNs can extract local changes and better detect relations in neighboring frames, whereas baseline transformers capture long-range dependencies that sometimes are not adapted to predicting the current frame events.

#### 4.4.5 Qualitative analysis

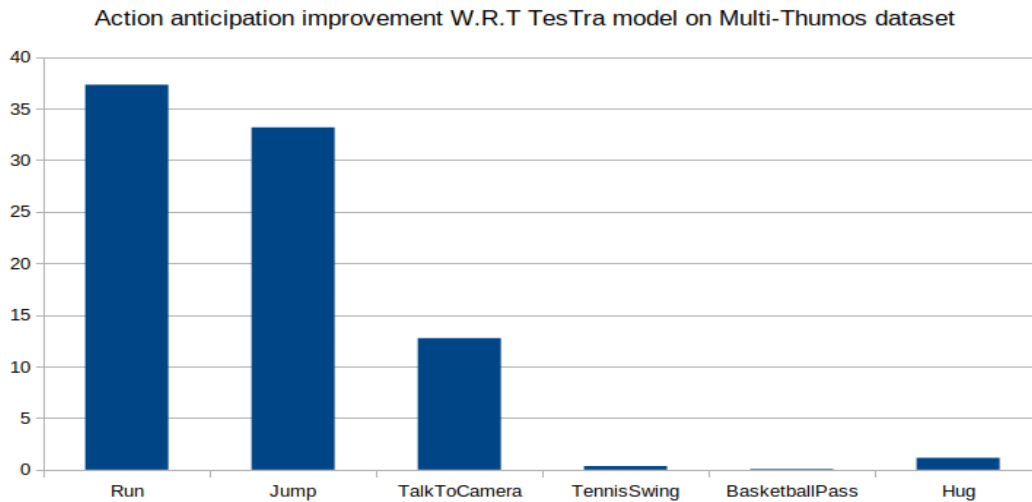


FIGURE 4.3: Action anticipation accuracy improvement on six actions w.r.t. TesTra model. This is performed on the Multi-THUMOS dataset, using 4 frames as anticipation length.

In this section, we analyze the effectiveness of our method on densely annotated datasets. We study anticipation improvement on six different actions, from the Multi-THUMOS dataset, according to their complexity as shown in Figure 4.3. We observe that the gain in some of these actions can reach 37%, while in some other actions, it is almost zero.



In fact, our prediction block anticipates the upcoming frame alongside future frames. By having access to the current frame, our model can correlate the anticipated action to the real action, hence we can learn to better anticipate the current frame, leading to a better-performing anticipation module.

Upon closer examination of these actions, we find that the improvement is particularly important for activities where there are multiple dependencies, or if the activity is interconnected with many other actions. The **RUN** action for instance, has correlations with up to seven other activities, as illustrated in Figure 4.1.

The qualitative results in Figure 4.3 demonstrate the robustness of JOADAA for complex correlated activities. This opens doors for future studies to analyze OAD and action anticipation on complex dense datasets.

## 4.5 Conclusion

Online action detection and anticipation are important fields in computer vision that have many real-world applications. These two tasks are highly correlated, and that is why we design JOADAA to address both tasks jointly improving one using the other and vice versa. Furthermore, we discuss the limitations of OAD and action anticipation for sparsely and densely annotated datasets.

Our model is limited in terms of effectively using long-range past features, especially for densely annotated datasets. Past knowledge undoubtedly adds to current knowledge and should lead to improvements. However, as demonstrated in this study, just adding pre-extracted features to transformers can also introduce noise. In the future, we are interested in tackling this limitation by modeling past features more effectively. One possible solution is to use an intermediate filter to learn only important features [48], or to learn the dependencies using the graph model to model only relevant features following [80].

## Chapter 5

# Robust and Efficient Multimodal Multi-dataset Multitask Learning

In the previous chapters we tackled the two main parts of action understanding (action recognition and action detection). Both our contributions were validated on selected datasets. However, one could ask how to transfer knowledge from some datasets to other datasets and have a generalized framework without the need of heavy and resource consuming fine-tuning. This study introduces a new model agnostic architecture for cross-learning, called CM3T, applicable to transformer based models. Challenges in cross-learning involve inhomogeneous or even inadequate amount of training data and lack of resources for retraining large pretrained models. Inspired from transfer learning techniques in NLP (adapters and prefix tuning), we introduce a plugin architecture that makes the model robust towards new or missing information. We also show that the backbone and other plugins do not have to be fine-tuned with these additions which makes training more efficient, requiring less resources and training data. We introduce two adapter blocks called multi-head vision adapters and cross-attention adapters for transfer learning and multimodal learning respectively. Through experiments and ablation studies on three diverse datasets - Epic-Kitchen-100, MPIIGroupInteraction and UDIVA v0.5 - with different recording settings and tasks, we show the efficacy of this framework. With only 12.8% trainable parameters as compared to the backbone for video input and 22.3% trainable parameters for two additional modalities, we achieve comparable or even better results compared to the state-of-the-art. Compared to similar methods, our work achieves this result without any specific requirements for pretraining/training and is a step towards bridging the gap between research and practical applications for the field of video classification.

### 5.1 Introduction

Computer vision vast subject of video categorization includes numerous sub-tasks and datasets for each of these tasks. Datasets, tasks, and recorded modalities have all increased recently. The majority of the effort is specialized to the job, corresponding datasets, or a subset of these modalities, and adapting it to a new input protocol is time-consuming. Additionally, there is a ton of overlapping work already done that may be merged to provide effective outcomes. Consequently, a method that can handle this increase in data with significant structural variability is required, as well as one that learns robust relations that can be shared between tasks and datasets. To address this issue, parameter efficient transfer learning (PETL) is becoming more and more appreciated.

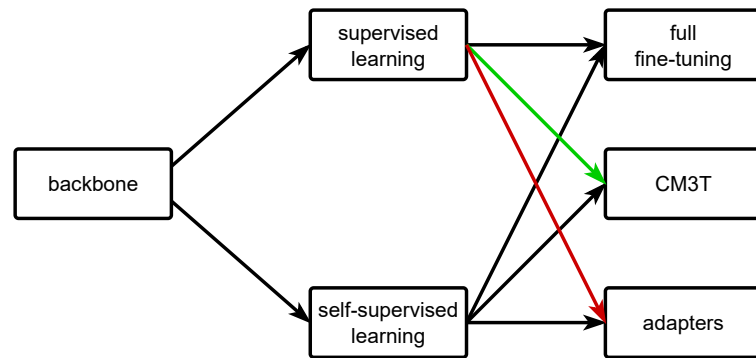


FIGURE 5.1: Representation of existing parameter efficient transfer learning techniques and CM3T. Backbones pretrained using self-supervised learning provide good general features, thus all methods of fine-tuning work well. In the case of supervised learning, adapters fail to perform well (shown in red) and CM3T is introduced to solve this (shown in green.)

PETL was founded on a single fundamental tenet, which is presented here. Modern universal feature extractors called transformer-based backbones [134, 8] are rigorously trained on large datasets. However, fine-tuning these models requires a lot of resources and fails for tiny datasets. While the fundamental spatial understanding of the video remains the same, fine-tuning any general feature extractor requires learning the surroundings in which the new data is recorded and the nuances of the new task. Therefore, we may use the fundamental knowledge offered by these pretrained models and merely require a few extra parameters to learn the fresh data supplied for fine-tuning.

There has been a lot of work on the above idea in the field of NLP [90, 92, 126]. These methods get similar results while adding less than 10% parameters to existing models which are trained to learn the new task while the pretrained weights are frozen. These have been extended to computer vision for a specific subdomain or involving complex pretraining/training methods [204, 40, 102, 141] [add refs], but we aim to provide an easier to use method, CM3T.

Similar to ours, existing PETL efforts rely on the many pretraining techniques like VideoMAE [209] and CLIP [167] that give foundation models. We require CM3T, however, if specific models that were pretrained using these techniques are not readily available. This circumstance might occur in a variety of contexts, such as when using custom models for specific downstream tasks or tiny datasets. The aforementioned issue is more difficult since getting starting models to perform similarly to foundation models with significant training on new datasets is more difficult.

The primary building components of CM3T are depicted in Figure 5.4. Multi-head vision adapters (in blue) are an improvement over scaled parallel adapters that parallelize modules in transformers with a set of bottleneck linear layers. With a minor modification, prefix tuning (in red) is employed as in [85]. In addition, we discovered that the aforementioned concept may be further expanded to cross-modal learning, where the weights of the pretrained model do not need to be altered in order to account for new modalities, just as the original model does not require

alteration in order to accommodate new datasets. This makes it simpler to use previous efforts to create more complicated systems. We now present the third module in the figure 5.4, cross-attention adapters for multimodal learning (in green).

Recently, most datasets have multi-modal information and our work leverages the above idea to utilise this additional information. Challenges in processing multimodal data include inhomogeneity of the present modalities, lack of correlation between modalities, and a need for a lot of training samples for convergence. CM3T addresses these challenges. Since cross-attention has been established as a good way for multimodal learning, we use it in place of linear layers in adapters, allowing their use for multimodal learning as well. It allows CM3T to learn the relationship between vision and other modalities. In addition, we add another module to capture the relationships between modalities other than vision, when available. The down-sampling layer in adapters provides a good embedding to use for cross-attention as it makes training easier compared to using original embedding from large transformer models. Also, training cross-modal adapters across datasets helps performance and provides a good pretrained feature extractor for small datasets. Another problem with multimodal models is that adding a new modality is cumbersome, but with our framework, it would just be a new plugin. It is clean and easy to use.

To show that our work is suitable for multi-dataset and multitask learning, we experiment on three vastly different datasets with different recording scenarios and tasks: EPIC-KITCHENS-100 (EK100), MPIIGroupInteraction (MPIIGI) and UDIVA v0.5.

The next section briefly discusses similar work and the comparison stating why it is needed. Section 5.3 describes the scientific details about all the additions. Section 5.4 lists the technical details and the results and observations for all the experiments. We show that we achieve comparable accuracy to state-of-the-art for all the datasets using only 12.8% trainable parameters as compared to the backbone for video input and 22.3% trainable parameters for two additional modalities. We perform multiple additional experiments to study how CM3T works in different scenarios and we explore the reasons for the results obtained.

The summary of our contributions are:

- We study how to apply transfer learning techniques popular in NLP to computer vision tasks with videos, without any specific pretraining requirements, and we extend their use to multimodality domain. We introduce multi-head vision adapters and cross-attention adapters to serve this purpose.
- Our method targets small datasets and custom models and we provide a framework for training these models on datasets or tasks with vision as the main modality and with other modalities available as input. This framework benefits from weight-sharing across tasks and datasets. It does this by storing relations between vision and other modalities and reusing it later. This improves performance and is useful for fine-tuning on small datasets.

## 5.2 Related work

### 5.2.1 Transfer learning

There are three types of transfer learning methods based on the availability of labels: inductive, transductive and unsupervised. Inductive transfer learning requires a source and target domain, both with labels. Our work falls under this category. We hypothesize that our framework is also extendable to transductive transfer learning, where target labels are sparse or missing. Other methods which fall under this category are briefly described in this section.

In computer vision, using pretrained models obtained from big training datasets [8, 134, 118] is a common approach. Another popular approach is teacher student networks, which involve training a model with few parameters using another larger model [37, 254, 2]. If these models are not available or to make these methods more robust, unsupervised techniques are utilised. These techniques can be divided into two main categories: generative learning [209, 26, 206] and contrastive learning [167, 41]. Generative learning methods commonly involve good data augmentation techniques and learn good feature representation using variation in input. Contrastive learning methods aim to learn a better space for the features learnt by the model. There are various task adaptation techniques like preventing catastrophic forgetting [119, 240], avoiding negative transfer [192, 86], and parameter efficient task adaptation [90, 126, 92]. With increasing size of models recently, these techniques are crucial. When tackling challenging downstream vision tasks, different methods work well for different specific settings. Also, they are not all extendable to include additional modalities. Thus, we choose to use parameter efficient task adaptation techniques. These can be easily modified to include any new modality as shown in this work, which is why they are chosen over the others mentioned here.

### 5.2.2 Parameter efficient task adaptation

There are three recent methods which show good results: (1) only updating new parameters added to the model or to the input [90, 107, 123, 125, 106]; (2) updating some of the parameters of the model in a sparse manner [255, 205, 242]; and (3) low-rank factorisation of weight matrices to reduce the number of parameters to be updated while keeping the weight matrix approximately the same [93]. [85, 140] combine these approaches to propose a unified parameter efficient training framework. Among these approaches, adapters, which belong to the first category, have been used in computer vision [171, 170] and natural language processing [90, 139, 106]. While adapters add additional parameters into models, prompt-based approaches instead add trainable parameters into the inputs [79, 123, 125], and experiments have shown their value in language tasks. We use both of these techniques like in [85] as inspiration to introduce the CM3T framework. VL\_Adapters [204] compare various adapter techniques [90, 107, 106] applied to question answering tasks, but not to pure vision tasks. Their work aims to use adapters to project vision and language pretrained model embeddings into the language model space whereas we show that it is possible to do it across vision datasets and also use it to add new modalities. Chen *et al.* introduce AdaptFormer, [40] which uses adapters with only the linear layers of a transformer and achieves better results than fully fine-tuning. But it uses VideoMAE [210] for pretraining ViT [114] which is not feasible if resources are limited and cannot be used to make a generalised framework. Their method fails with

models not carefully pretrained as described in 5.3. Jia *et al.* introduce visual prompt tuning (VPT), [102] which use prompt tuning for images, but prompts alone do not work well for videos as it is also mentioned by [40]. The paper [141] shows that adapters only work for vision if the bottleneck dimension is large. They introduce a pruning technique to reduce the size of these adapters. We introduce multi-head vision adapters as an alternative that works well even with a small bottleneck dimension and without any specific pretraining method.

### 5.2.3 Multimodal learning

There is an inherent difference between videos and other modalities like audio and text and thus it is challenging to combine them into one model. VATT [5] uses early fusion, where they concatenate all input modalities. Although the earlier the fusion, the better the results, there is a trade-off with the amount of data required for training as it is harder for models with early stage fusion to converge and this leads to tedious self-supervised learning. Some works design a specialised architecture for fusion at feature level [3, 159]. They work better but there are limitations as the fusion is done after downsampling the input features which leads to loss of information and poor cross-modality relations. [56, 121, 182] have feature level fusion with minimal downsampling, but they lack in handling specific modalities differently. So, there is a need for a model which can benefit from cross-modality learning at different levels. To answer this, we propose to use cross-attention added to each block of a transformer architecture. This gives the above mentioned flexibility to the model. State-of-the-art methods like M&M Mix [231] and MuMu [98] are either modality specific or have a rigid architecture making it hard to add/remove modalities. This work addresses these drawbacks by having a flexible architecture that can accommodate any type of input.

## 5.3 CM3T framework

The main motivation for our work is to define an easy way to use existing multimodal data and pretrained models when approaching a video classification or video understanding task. This will assist in bridging the gap between research and practical applications. This section lists the methodology of our work.

Before diving further into the architecture of our proposed solution, let's introduce some basic concepts.

### 5.3.1 Basic concepts

#### Adapters

Adapter-based tuning finds application with widely used Transformers, recognized for their ability to achieve state-of-the-art (SoTA) performance in various NLP tasks, including machine translation and text classification problems. The architecture is depicted in the figure 5.2.

The adapters perform a two-step process: initially projecting the original feature size to a smaller dimension and subsequently projecting them back to the original size. This design choice guarantees that the number of parameters remains significantly smaller than that of the original model, as illustrated in the procedure on the right of the figure 5.2. The reduction in parameters introduces an evident trade-off

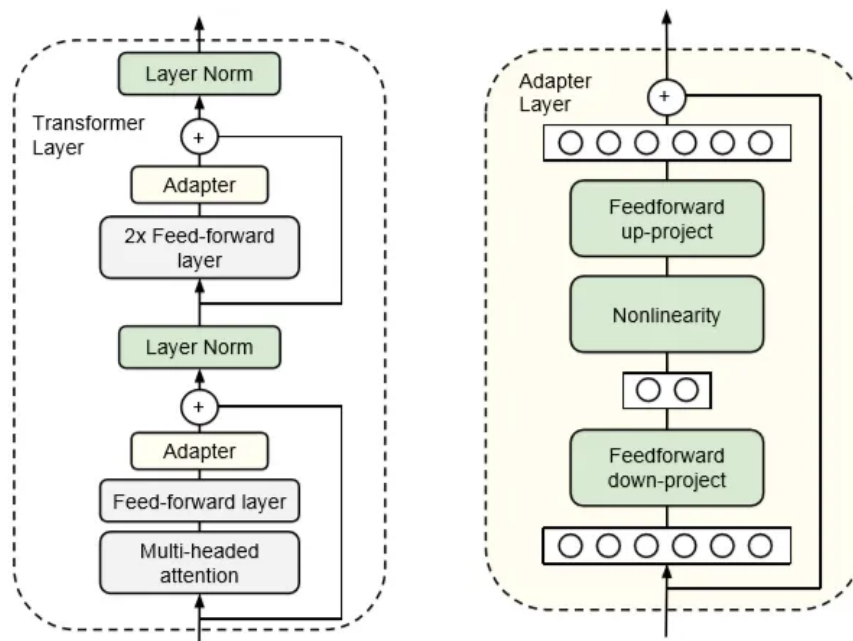


FIGURE 5.2: the left of the figure, the standard Transformer is used with an additional adapter layer, added after each sub-layer and before adding the skip connection back. The output of the adapter layer is then forwarded to the layer normalization.

between performance and parameter efficiency, an aspect that is discussed through this paragraph of our thesis and its experiments.

### Prefix tuning

The key motivation behind prefix tuning is that providing the right context or “prompt” to a language model can steer it to perform a downstream NLG task without needing to modify the model’s parameters.

Specifically, prefix tuning aims to learn a continuous prompt that can be optimized end-to-end, rather than relying on manual prompt engineering. When prepended to the input, the learned prefix provides the context needed to guide the model’s behavior towards the task objective.

By leveraging prompting while enabling end-to-end optimization of a continuous prompt, prefix tuning provides a way to adapt language models without extensive parameter tuning. The prefix allows injecting task-specific knowledge into the pretrained model in a lightweight way.

### 5.3.2 Choosing a pretrained model

Our method is focused on transformer based backbones which have produced state-of-the-art results for various vision tasks. We use Video Swin transformers [134], but the following steps of the framework are model invariant and the backbone can be chosen according to the need. The reason for choosing Video Swin is that different blocks process the input at different spatial resolutions. Depending on the side input

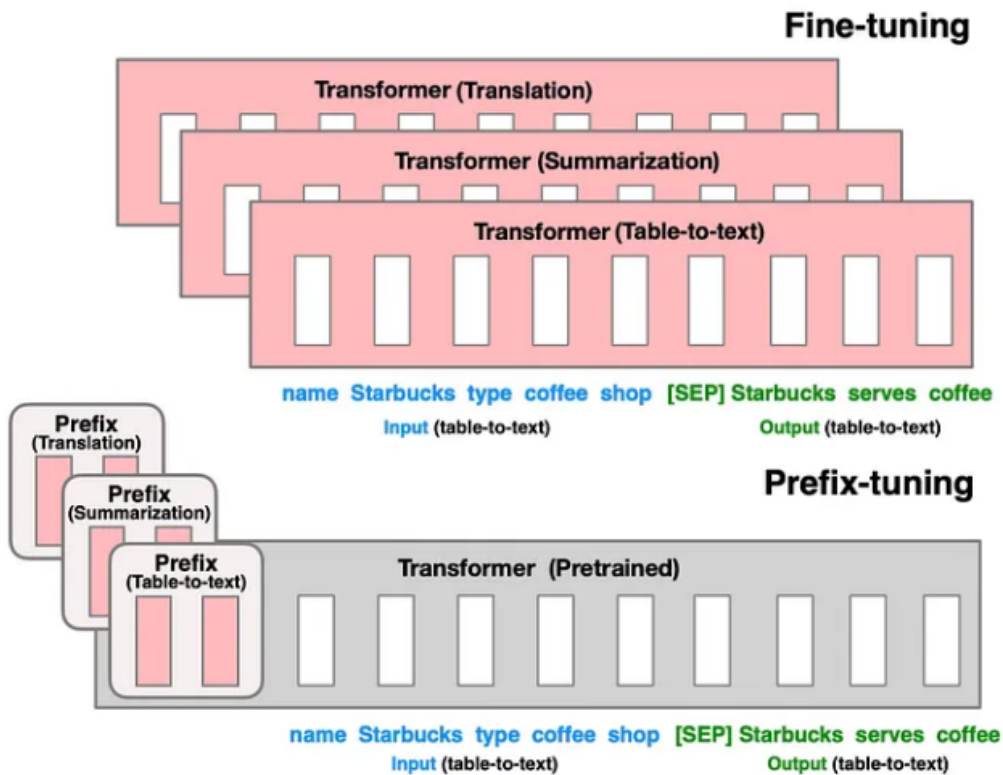


FIGURE 5.3: Adjusting (top) fine-tunes all Transformer parameters (illustrated by the red Transformer box) and necessitates preserving a complete model replica for every task. Prefix-tuning (bottom), wherein the Transformer parameters are fixed, and optimization exclusively targets the prefix (depicted by the red prefix blocks). Source <https://arxiv.org/pdf/2101.00190.pdf>

(other modalities), cross-attention performs well with different blocks *i.e.* different spatial resolutions.

### 5.3.3 Fine-tuning or using Adapters

The next stage is to fine-tune the vision model on the target dataset once it has been pretrained. If there is a shortage of computational power or time, substituting adapters and prefix tuning for full fine-tuning results in results that are comparable but require a lot less training parameters. For end-to-end learning, this step might also be combined with the stages that follow (in this section and the following one), however we do each step independently to assess how well it performs in comparison to the state-of-the-art. In section 5.3, the outcomes for end-to-end training are also displayed.

We take inspiration from scaled parallel adapters and prefix tuning as used by He *et al.*[85]. Figure 5.4 shows all the additions to the pretrained model along with our modifications. Multi-head vision adapters (in blue) and prefix tuning (in red) are discussed in this subsection and cross-attention adapters (in green) are discussed in the next one.



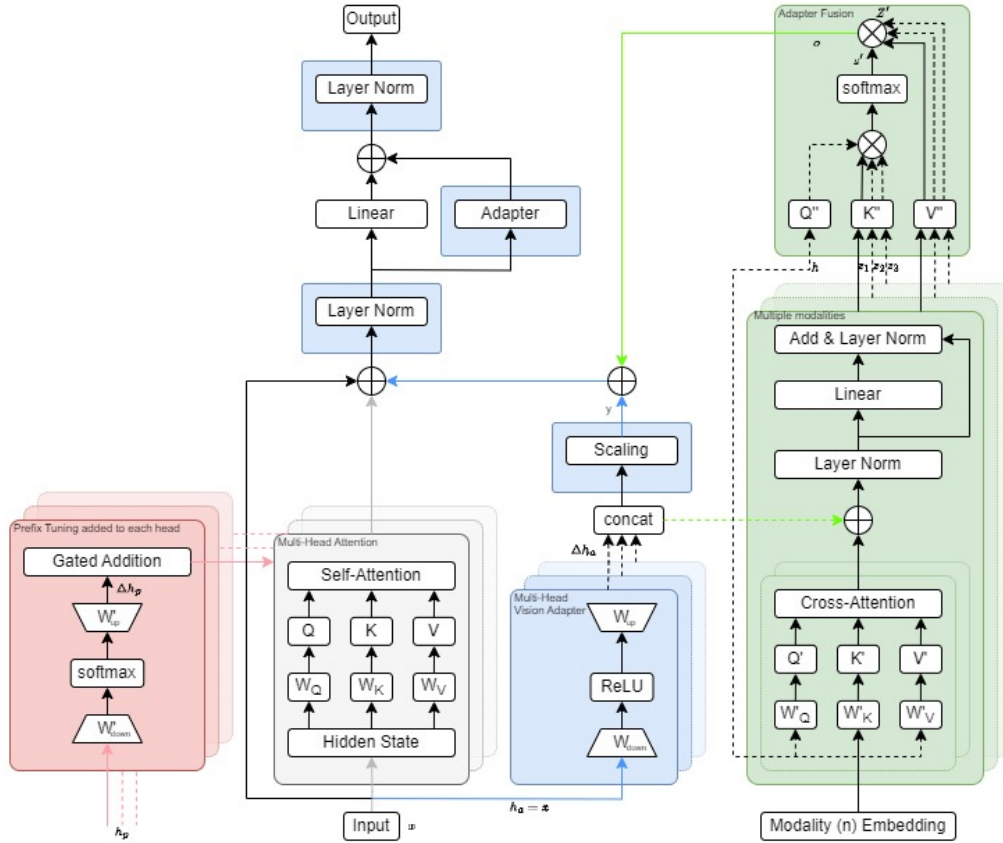


FIGURE 5.4: Detailed architecture of CM3T. Coloured parts are the ones that are fine tuned and the rest are frozen. It has three separate blocks added to it which are shown in three different colours. Prefix tuning is complicated to show in detail, so only a diagram is shown. Comparing eq 5.2 and 5.4 gives how the upscaling and downscaling weights are computed. The rest of the details are described in section 5.3

Mathematically, adapters from Houlsby *et al.* are defined as,

$$y = s \cdot \Delta h_a \quad (5.1)$$

$$\Delta h_a = \text{ReLU}(h_a W_{down}) \cdot W_{up} \quad (5.2)$$

where  $h_a = x$  is the input of size  $d$ ,  $W_{down} \in \mathbb{R}^{d \times r}$  is the weight matrix for the down-projection layer with bottleneck dimension  $r$ ,  $W_{up} \in \mathbb{R}^{r \times d}$  is the up-projection layer, and  $s$  is the scaling factor. We use this in parallel instead of sequential similar to He *et al.* We also use their definition for prefix tuning (in Figure ??, we do not show recurrent connection for prefix tuning, to simplifying it, but the equation shows the correct usage).

$$h_p \leftarrow (1 - \lambda) \cdot h_p + \lambda \cdot \Delta h_p \quad (5.3)$$

$$\Delta h_p = \text{softmax}(h_p W'_{down}) \cdot W'_{up} \quad (5.4)$$

$$W'_{down} = W_q P_k^T; W'_{up} = P_v \quad (5.5)$$

where  $W_q$  is the weight matrix for getting query vector from the input  $h_p$ ,  $P_k$  and  $P_v$  are prefix tuning vectors which are learned in the same way as in the paper [126], and  $\lambda$  is the factor used for gated addition. The red part of figure 5.4 shows prefix tuning added to transformers, it is added in parallel to each head of multi-head attention.

Using adapters for vision tasks is more challenging than NLP, as language understanding does not change with the task or dataset, but video datasets have a wide variety of settings like indoor or outdoor recordings, different views and camera angles, lighting changes, etc. Fine-tuning allows the networks to overcome these changes, but it is hard for adapters owing to less capability to change the original model activations. But with a few changes, adapters can show performance comparable to fully fine-tuned models. The blue part of figure 5.4 shows the architecture of the adapters used.

First, we make the scaling factor for adapters ( $s$  in equation 5.1) added to linear layers learnable, allowing greater change to activations. Attention in pretrained models might focus on features that are not relevant to the new downstream task or dataset, but this change allows adapters to overcome this.

AdaptFormer [210] adds traditional adapters in parallel to linear layers only and achieves better results than fine-tuning owing to a sophisticated pretrained ViT model using VideoMAE [210]. We achieve very poor performance with the same method without this specific pretraining, even when coupled with prefix tuning. So, this leads to our second change: inspired by multi-head attention we use modified multi-head adapters. This is different from multi-head attention as we divide the input along the window dimension of Video Swin transformers and not the channel dimension. So, we have different linear layers for different sets of windows. We saw that increasing the bottleneck dimension in adapters only increased the performance slightly (as shown by [141]), but adding the above change allowed the network to learn better even with a smaller bottleneck dimension. Overall, we do not increase the parameters by a big margin compared to traditional adapters as we use a smaller bottleneck dimension. To define the change mathematically, we divide the input  $h$  along the windows dimension to get  $\{h_1, h_2, h_3, \dots\}$ . Each has its own parallel adapter and the output is concatenated along the same dimension before scaling and addition. Extending equation 5.2,

$$\{h_{a1}, h_{a2}, \dots\} \leftarrow \{h_{a1}, h_{a2}, \dots\} + s \cdot \Delta\{h_{a1}, h_{a2}, \dots\} \quad (5.6)$$

where each operation is performed element-wise. Without the two changes mentioned above, adapters have very poor performance for the domain of computer vision with traditionally available pretrained models. We call these adapters multi-head vision adapters. These are specific to video swin transformers, but we hypothesise that the same concept can be applied to modify adapters for any model - using different linear layers in adapters for different sets of windows to which attention is applied.

The third change, is that we use ReLU activation in place of tanH with a lower dropout for prefix tuning and that works better.

### 5.3.4 Adding other modalities

Cross-attention adapters are used for adding modalities to the model received from the previous step. Cross attention-adapters are simply obtained by replacing the two linear layers in the adapters with a cross-attention module. Each added modality has its own adapter. The query and value inputs to this adapter are taken from the concatenation of hidden states from the bottleneck hidden state in the multi-head vision adapter *i.e.*  $Q = V = ReLU(xW_{down})$ , where  $h$  is the input to the swin transformer block and  $W_{down}$  is the same as that in equation 5.2. The key is taken as the feature embedding from the new modality.

To merge all the adapters trained for different modalities, in place of simple addition, we use AdapterFusion [176] which captures the interactions between different side inputs *i.e.* modalities other than vision. It is an attention block where each head has the same query as that of attention in cross-attention adapter for each modality, described above, let's say  $h$ . The key and value for each head are taken from the output of each cross-attention adapter respectively, let's say  $z_n$ , where  $n$  signifies the  $n^{th}$  modality. Mathematically, the module can be expressed as

$$s' = softmax\left(h^T W_Q \otimes z_n^T W_K\right), n \in \{1, \dots, N\} \quad (5.7)$$

$$z'_n = z_n W_V, \quad n \in \{1, \dots, N\} \quad (5.8)$$

$$Z'_n = [z'_0, \dots, z'_N] \quad (5.9)$$

$$o = s'^T Z' \quad (5.10)$$

where  $o$  is the output and  $W_Q, W_K$  and  $W_V$  are weight matrices for query, key and value respectively, and  $N$  is the number of side modalities.

To incorporate a new modality into the model, there are two additions, a new cross-attention adapter and a new concatenation to  $s$  and  $Z'_n$  vectors above. One disadvantage of this is that model size keeps increasing with more modalities. To alleviate this, we use the cross attention module proposed by Agrawal *et al.* [4] and we show results in section 5.4. It makes adding new modalities hard, but it is a trade-off

between flexibility and optimising resource usage.

## 5.4 Experiments

### 5.4.1 Datasets

To show the robustness of our approach we present three types of datasets, with very different tasks and modalities. First, an egocentric **Epic-Kitchen100** [51] consisting of three modalities RGB, optical-flow, and audio for actions of human-object interactions. Second, **MPIIGroupInteraction** [10] which is a body language dataset aiming at understanding human behavior in human-to-human interactions. For this dataset, we use the following modalities: RGB and audio. Finally, we have **UDIVA v0.5** [160] which tackles the task of human personality analysis, using also different modalities such as RGB, transcript, and audio.

Our approach shows effectiveness on all three tasks, proving our theory of bringing adapters mechanisms into vision problems to tackle all the challenges mentioned in the previous parts.

**Epic-Kitchen100** [51] is a first-view and human-object interaction action recognition dataset. Its a collection of 100 hours, 20M frames, 90K actions in 700 variable-length videos, capturing long-term unscripted activities in 45 environments, using head-mounted cameras. Compared to its previous version (EPIC-KITCHENS-55), EPIC-KITCHENS-100 has been annotated using a novel pipeline that allows denser (54% more actions per minute) and more complete annotations of fine-grained actions (+128% more action segments).. It consists of a total of 97 verbs and 300 nouns, each action is a combination of a verb + noun and has in total of 3806 action classes.

**MPIIGroupInteraction** [10] The MPIIGroupInteraction dataset is 26 hours of spontaneous human behavior with 15 distinct body language classes. This dataset presents a novel set of actions not very explored in computer vision and human-behavior understanding. It consists of subtle body language behaviours such as gesturing, grooming, or fumbling.

**UDIVA v0.5** [160]: this dataset is 90.5 hours of dyadic interactions among 147 participants distributed in 188 sessions, recorded using multiple audiovisual and physiological sensors. But only half of the data has been released. UDIVA v0.5 main task is personality recognition. It has 5 main classes Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (OCEAN).

### 5.4.2 Training details

For multi-head vision transformers, the bottleneck dimension used is,  $1/4$  multiplied by the channel dimension of the embedding for the respective block of Video Swin-B. A smaller dimension size produces worse results, and a larger size produces similar results. For EK100, we use a slightly larger bottleneck dimension ( $3/8$  times in place of  $1/4$ ) for the last block of Video Swin-B. For prefix tuning, the prefix channel dimension used is a minimum of 64 and  $1/8$  multiplied by the channel dimension of the embedding for the respective block of Video Swin-B. The bottleneck dimension for the generation of the prefixes is  $1/4$  multiplied by the channel dimension of the prefix. Smaller prefixes provide worse results. Larger prefixes make the

TABLE 5.1: SOTA comparison on EK100. We show main comparisons in same colors.

	Multimodality	Top-1 accuracy (%)	Epochs	Trained Parameters
M&M Mix	✓	49.6	50	>300%
MTV-B	✗	46.7	80	>100%
Video Swin-B (fully finetuned)	✗	41.7	49	100%
Video Swin-B (fully finetuned) +CAA	✓	48.9	56	109%
Video Swin-B (frozen) + Adapters + PT	✗	28.7	6	7.2%
Video Swin-B (frozen) + MHVA + PT	✗	39.8	14	12.8
CM3T	✓	48.2	22	22.3%

networks overfit very fast. The model starts overfitting after 5-6 epochs with a big prefix. Prefix tuning is a shortcut for the network to force attention to focus on particular features by learning fixed additional inputs to keys and values. This allows it to easily learn patterns in the inputs or activations of the training set and thus overfit.

For cross-attention, we use feature embeddings extracted from side modalities. We use trill-distilled [185] to obtain audio features. Roughly 15 time steps of the audio features correspond to 128 frames in the videos (we use a stride of 4 for our input and Video Swin takes 32 frames as input). We use TVL1 optical flow estimation and bninception [97] is used for its feature extraction. The features corresponding to each frame in the RGB video are taken, so the input temporal dimension is the same as RGB videos. Conv-1D is used for temporal pooling all side modalities as Video Swin uses small voxels to divide each input embedding and applies attention to each of them homogeneously. For simplicity, we give context from the side modalities for the whole input to each voxel. Also, we use performers [45] for cross-attention to make it more efficient.

The cross-attention adapters are added to the first two and last two layers of each block of Video Swin. For the last block, we change the input for the value to be the same as the key and use the cross-attention adapters for late fusion in place of modifying attention. This provides slightly better results.

For training, we use 8/16 Tesla V100 GPUs with a batch size of 3 per GPU for adapters and 2 for full fine-tuning. These are the largest batch sizes we can fit on one GPU for each case. We train for varying number of epoch, stopping if performance does not increase for 6 epochs. The learning rate is modified according to the batch size. Video Swin transformers use a batch size of 8 per GPU and a starting learning rate of  $3e-4$ , we use  $1.5e-4$  for batch size of 2 and  $1.8e-4$  for batch size 3. The weight decay is 0.05. Weight initialisation for downscaling weights of is used as kaiming initialization, zero initialisation for upscaling. Rest weight initialisations are either from the pretrained model or default initialisation from PyTorch. We use Video Swin-B pretrained on Something-Something v2 (SSv2) dataset for experiments on EK100 dataset. For the experiments on the other datasets, we use the same model pretrained on Kinetics 400 dataset. SSv2 is an egocentric dataset, similar to EK100 and pretrained Video Swin-B uses a larger window size for it, so we chose it for EK100. Kinetics 400 is more similar to the other datasets, so we use it for experiments on the others.

TABLE 5.2: SoTa comparison on UDIVA 0.5

	Multimodality	Mean MSE	Epochs	Trained Parameters
Video Swin-B (full fine-tuned)	X	0.82	51	100%
Video Swin-B (full fine-tuned) + CAA	✓	0.69	32	109.5%
CM3T (Swin-B backbone)	✓	0.69	27	22.3%
FAt transformers	✓	0.72	30	
Video Swin-T (full fine-tuned)	X	1.10	29	
CM3T (Swin-T backbone)	✓	0.81	18	

TABLE 5.3: SoTa comparison on MPIIGroupInteraction

	map	Epochs	Trained Parameters
Video Swin-B (full fine-tuned)	0.887	17	100%
Video Swin-B (full finetuned) + CAA	0.901	18	109.5%
FAt transformers	0.899	18	-
CM3T	0.901	9	22.3%

### 5.4.3 Results and Observations

#### 5.4.4 SoTa comparison

In this section we compare our results to the existing SoTa methods to show efficacy of our approach.

##### Baseline comparison

First of all, we compare our method to the fully fine tuned backbone which is video swin-B [135]. Video Swin transformer is one of the SoTa transformers on different datasets and tasks, hence we chose it as our backbone. For Epic-Kitchen100 dataset [51], we use top-1 accuracy for actions as a metric for all experiments and achieve accuracy just 0.7% less than the fully fine-tuned model. Note that we add CAA(cross attention adapters) to both Swin-B fully fin-tuned and our CM3T to have a fair comparison. Nevertheless our method achieves comparable results with only 22.3% parameters whereas Swin-B combines CAA goes up to 109% parameters. Moreover for the UDIVA 0.5 [160] and MPIIGroupInteraction [10] we achieve the SoTa results and again with only a fraction of 22.3% of the total number of parameters, see tables 5.2 and 5.3 for results on UDIVA 0.5 and MPIIGroupInteraction.

##### Comparison with traditional adapters

To show the robustness of our proposed adapters implementation we compare between basic adapters from AdaptFormer [40] and our proposed MHVA(multi-head vision adapters). We compare the results of our CM3T and adapters without any multi-modalities. With typical adapters + PT (prefix tuning) we achieve 28.7% whereas with our CM3T we achieve 39.8. This shows that our implementation of adapters is more robust and work better than typical adapters. The motivation for the change discussed in the methodology section is thus justified from these results. Results are in table 5.1.

### SoTa comparison on Epic-Kitchen 100

For Epic-Kitchen 100 [51] M&M Mix [231] holds the best results. M&M Mix [231] processes each of the three modalities using three branches of ViViT at different spatial resolutions using different sizes of input tubelets and different variants of ViViT. They use additional modules to share information across views and models for different modalities. One branch has more parameters than video swin-B, so the total number of parameters is more than three times the number of parameters of swin-B. We achieve comparable accuracy with significantly lesser parameters. Finally, we achieve only 1.4% less than the state-of-the-art on Epic-Kitchen 100 with a tiny number of parameters in comparison, see table 5.1.

### SoTa comparison on Udiva v0.5 and MPIIGroupInteraction

For UDIVA 0.5 and MPIIGroupInteraction we compare to FAt transformers [4]. FAt transformers have a lot of additions specifically for UDIVA v0.5 which is the reason for their good performance. They have additional input branches with face crops and contextual videos and a complex method of preprocessing too. We compare against the results published by them as is.

As for MPIIGI We achieve better results with transfer learning techniques than fully fine-tuning. There are two reasons for this, one is that MPIIGI is a small dataset and it is easier for these techniques to converge. The second reason is that Kinetics400 is very close to MPIIGI and the networks are initialised very well. This enables adapters to work better.

### Cross attention module

An interesting thing to note is that MTV-B which is the base model for M&M Mix and uses only RGB videos as input, achieves 46.7% accuracy and there is only a 2.9% accuracy increase when optical flow and audio are added to it in M&M Mix. Whereas we achieve an increase of 8.4% increase with CM3T when the two modalities are added. This might be because MTV-B is a better backbone as compared to video-swin-B and captures most of the information present in optical flow already as optical flow is also a visual feature, thus adding optical flow does not increase performance for them as much as for us. This proves the efficacy of cross-attention adapters. Moreover, in table 5.4 we compare two methods of using different modalities the typical cross attention between different modalities and our proposed CAA and we observe that with our proposed solution we can achieve 0.5% more accuracy, showing robustness and efficacy of the proposed CAA.

### Time and resources

We state that CM3T saves time and computational resources and we already discussed a reduction in trainable parameters. Tables ?? show that fewer epochs are required for the convergence of our models. For just fine-tuning RGB models, multi-head vision adapters and prefix tuning require a third of the time as compared to fully fine-tuning.

#### 5.4.5 Ablation studies

In this section we look at different component of the proposed solution and how they contribute to the results.

TABLE 5.4: Ablation on multimodality attention. The abbreviations used are MHVA: multi-head vision adapters, PT: prefix tuning, CAA: cross attention adapters, MmCA: multi-modality cross-attention. These results are reported on Epic-Kitchen 100.

	Top-1 Accuracy (%)
Video Swin-B (Frozen) + MHVA + PT	39.8
Video Swin-B (Frozen) + MHVA + PT + CAA	<b>48.2</b>
Video Swin-B (Frozen) + MHVA + PT + MmCA	47.7

### MHVA / PT

MHVA works well by itself and works even better when combined with PT, as shown in Table 5.5. But, PT alone does not work very well as prefix tuning tries to find learnable fixed inputs to be added to the actual input to provide context, but since supervised pretrained models do not give good relevant features for a different dataset, these inputs are not very useful unless combined with MHVA which provide a way for the model to learn good relevant embeddings under the new settings.

### Different backbone

Table 5.6 proves one of our previous claims, which that our method can be implemented with any backbone. In this study we implement our proposed solution with a ViViT-B and we observe the same results as with the Swin-B transformer.

TABLE 5.5: Ablation study on different components of our proposed architecture.

	Epic-Kitchen 100
Video Swin-B (Frozen) + MHVA	36.8
Video Swin-B (Frozen) + PT	23.3
Video Swin-B (Frozen) + MHVA + PT	39.8
Video Swin-B (Frozen) + MHVA + PT + CAA	<b>48.2</b>

TABLE 5.6: Results of our method using different backbone. Experiments were done on Epic-Kitchen 100

	top1-accuracy %
ViViT-B (Full fintuned)	37.4
ViViT-B (Frozen) + MHVA + PT	38.1
ViViT-B + MHVA + PT +CAA	44.3

### 5.4.6 Cross-attention adapter behaviour with different modalities at different levels

We apply cross-attention to the first and last two layers of each block in video swin transformers [134]. If the blocks have only two layers, we apply them to both layers. In this section, we study the importance of cross-attention at different levels for different modalities, by removing adapters from different blocks. Table 5.7 shows the results. This can be used to prune the architecture for specific modalities. We see that for audio and transcript, later layers are more important, whereas for optical



Method	Performance
Audio(MSE)	
UDIVA(CM3T)	0.69
UDIVA(CM3T - Block 1)	0.72
UDIVA(CM3T - Block 2)	0.73
UDIVA(CM3T - Block 3)	0.81
UDIVA(CM3T - Block 4)	0.78
Audio(Top-1 Accuracy)	
EK100(CM3T)	48.2%
EK100(CM3T - Block 1)	47.8%
EK100(CM3T - Block 2)	47.5%
EK100(CM3T - Block 3)	46.4%
EK100(CM3T - Block 4)	47.1%
Transcript(MSE)	
UDIVA(CM3T)	0.69
UDIVA(CM3T - Block 1)	0.70
UDIVA(CM3T - Block 2)	0.73
UDIVA(CM3T - Block 3)	0.82
UDIVA(CM3T - Block 4)	0.79
Optical Flow(Top-1 Accuracy)	
EK100(CM3T)	48.2%
EK100(CM3T - Block 1)	45.3%
EK100(CM3T - Block 2)	45.8%
EK100(CM3T - Block 3)	44.2%
EK100(CM3T - Block 4)	46.6%

TABLE 5.7: Results for ablation study in section 5.4.6. The entries show cross attention removed from a particular block. Block 1 is closest to input and Block 4 is the last block before classification head.

flow, earlier layers are more important. Block 3 is the biggest block and is needed for good results for all side modalities.

## 5.5 Adding adapters to cross-attention adapters

Since modalities are repeated across tasks and datasets, we see that training the entire cross-attention adapter module is not necessary. We can simply add scalable parallel adapters to the cross-attention modules. For this, the initial embedding is directly taken from the pretrained model and not the multi-head vision adapters, the rest stays the same. Table 5.8 shows the results for this experiment. We train cross-attention adapters for audio using EK100 and show results for UDIVA with normal cross-attention adapters and adapters added to cross-attention adapters.

Method	Performance
UDIVA(MSE)	
CM3T	0.690
CM3T (with adapters added to CA)	0.689

TABLE 5.8: Result for adding adapters to cross-attention adapters

## 5.6 Conclusion

In this work, we presented CM3T, a framework for using common pretrained video classification models with a transformer based architecture. The framework consists of three modules, two introduced by us, multi-head vision adapters and cross-attention adapters, and one already existing, prefix tuning. We show that these work well without specific pretraining or training methods and we study different variants. This work helps bridge the gap between research and practical applications of video classification models by making it easier to adapt existing work to new datasets and tasks, and also utilise any new modalities that might be present. The limitation of this approach is that if the dataset used for pretraining is very dissimilar to the target one, the results will not be good. The frozen pretrained model needs to have the relevant information for the target task or dataset. Using various data augmentation or self-learning methods might help in this cases or using a smaller model and fully fine-tuning might give better results. For future work, combining adapters with selective fine-tuning of the model might resolve the above issue while keeping the number of trainable parameters low.



## Chapter 6

# MultiMediate'23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions

Our main objective in this thesis is to provide a system capable of understanding human behaviors, thus to facilitate human-robots collaborations. Automatic analysis of human behaviour is a fundamental prerequisite for the creation of machines that can effectively interact with-and support humans in social interactions. In MultiMediate'23, we address two key human social behaviour analysis tasks for the first time in a controlled challenge: engagement estimation and bodily behaviour recognition in social interactions. This work describes the MultiMediate'23 challenge and presents novel sets of annotations for both tasks. For engagement estimation we collected novel annotations on the NOvice eXpert Interaction (NOXI) database. For bodily behaviour recognition, we annotated test recordings of the MPIIGroupInteraction corpus with the BBSI annotation scheme. In addition, we present baseline results for both challenge tasks.

### 6.1 Introduction

Artificial mediators [162], i.e. interactive intelligent agents that actively engage in a conversation in a human-like way have the potential to positively influence the course and/or outcomes of human interactions. They have been studied in a variety of contexts, including collaborative teamwork [25, 186], mental health [23], and education [136, 60]. A central prerequisite for effective and context-aware artificial mediation is the ability to comprehensively detect- and interpret the diverse set of social signals expressed by humans. At present, this challenge is still largely unsolved, and research on artificial mediators often has to rely on Wizard-of-Oz paradigms [136, 213, 60, 23, 155, 178].

With the multi-year MultiMediate challenge we contribute to realising the vision of autonomous artificial mediators by facilitating measurable advances on central conversational behaviour sensing and analysis tasks. The first iteration of the challenge in 2021 [147] has addressed eye contact detection and next speaker prediction while MultiMediate'22 has focused on backchannel analysis [148, 7]. In two separate tracks, MultiMediate'23 addresses the recognition of complex bodily behaviours, as well as the estimation of a person's engagement level. Bodily behaviours such as fumbling, folded arms, or gesturing are a key social signal and were shown to be connected to many important high-level phenomena including stress regulation, attraction, or social verticality [31, 214, 82, 144]. As a result, an accurate recognition

of bodily behaviours can serve as a building block for the recognition of such more abstract phenomena. Knowing how engaged participants are, individually or as a group, is important for a mediator whose goal it is to keep engagement at a high level. Engagement is closely linked to the previous MultiMediate tasks of eye contact detection [153, 163] as well as backchanneling [78].

With MultiMediate'23 we present the first challenge on engagement estimation and the recognition of bodily behaviours in social interaction. We define the tasks and evaluation criteria and describe new annotations collected on the NOvice eXpert Interaction (NOXI) database [27], as well as on unreleased test recordings of MPIIGroupInteraction [150]. Furthermore, we present baseline approaches for both challenge tasks and report evaluation results. We make all collected annotations, baseline implementations, and raw feature representations publicly available for further use, even beyond the scope of MultiMediate'23.<sup>1</sup>

## 6.2 Related work

We review previous works on methods and datasets for engagement estimation and bodily behaviour recognition in social interaction.

### 6.2.1 Engagement estimation

Engagement has been investigated from various research angles, e.g. how to define, annotate, or automatically predict it. Rich et al. [172] introduced a module for the recognition of engagement in human-robot interaction based on backchannels. Sanghvi et al. [175] predicted engagement based on body posture features. Bednarik et al. [18] focused on recognizing conversational engagement with gaze data. Research in detecting engagement in students is prolific and promising [105, 76]. Engagement is also often studied in children [168] and, more particularly, in children interacting with an artificial agent [154, 161, 99]. Guhan et al. [81] researched engagement in mental health patients, based on videos of the patient. Some datasets also offer engagement ratings, such as RECOLA [173], MHHRI [35], and [91] with annotations from [19]. In Table 6.1 we provide an overview over the existing social interaction datasets with engagement annotations. The NoXi dataset annotated for MultiMediate '23 is significantly larger compared to previous datasets.

### 6.2.2 Bodily behaviour recognition

Bodily behaviours are key signals in social interactions and are related to many higher-level attributes. For example, displacement behaviours (e.g. fumbling, face-touching, or grooming) are associated with anxiety and stress regulation [14, 145, 144]. Leaning towards the interlocutor is connected with rapport [180] and crossed arms can be indicative of emotion expressions [221]. Further connections were found between bodily behaviours and liking [142, 143], attractiveness [214], and social verticality [82].

Despite this importance, little previous work addressed the recognition of bodily behaviors such as fumbling, grooming, crossing arms, or gesturing in social interactions [11, 131]. While impressive progress was made in the estimation of body and

---

<sup>1</sup><https://multimediate-challenge.org>

Corpus	Screen	Group size	Length	Part.
Guhan et al. [81]	✓	2	1h5m	13
RECOLA [173]	✓	2	3h50m	46
Bednarik, Eivazi, and Hradis [19]	✓	4-7	6h	9 groups
MMHRI [35]	✗	2	6h	18
NOXI (ours)	✓	2	25h	87

TABLE 6.1: Social interaction datasets with engagement annotations, excluding MOOC and school settings and children as participants. *Screen* indicates whether the interaction was screen-mediated, *Group size* the number of humans per interaction, *length* the total duration of interactions, and *part.* the total number of human participants.

hand pose [29, 193], it is not a trivial task to establish the connection between low-level keypoint detections and complex bodily behaviors relevant to the interaction. Furthermore, only a limited number of bodily behavior recognition datasets containing spontaneous behaviour in social interactions are available. The PAVIS Face-Touching dataset [21] consists of a single annotated behaviour (face touching) in group discussions. The iMiGUE dataset [131] contains annotations of 32 behaviour classes annotated for speakers at sports press conferences. For the purpose of MultiMediate, the recently published BBSI dataset [11] is most relevant, it consists of 15 behaviour classes annotated for all participants of 3-4 person group conversations. Such group conversations are one of the main application domains of artificial mediators.

## 6.3 Challenge description

In the following, we present the two challenge tasks and the datasets used. For both tasks, test samples (without ground truth) are released to participants before the challenge deadline. The participants in turn submit their predictions for evaluation.

### 6.3.1 Engagement estimation task

**Task definition** The job is the frame-by-frame prediction of each participant’s level of conversational participation on a continuous scale from 0 (lowest) to 1 (highest). Investigating the multimodal and reciprocal behavior of both interlocutors in the Novice-Expert Interaction corpus is encouraged. To assess the predictions in the test set, we use the Concordance Correlation Coefficient (CCC) [128].

**Dataset** The NOvice eXpert Interaction (NOXI) database [27] is a corpus of dyadic, screen-mediated face-to-face interactions in an expert-novice knowledge sharing context. In a session, one participant assumes the role of an expert and the other participant the role of a novice. Figure 6.1 shows two users during interaction. Interactions from NOXI were captured in three different countries -France, Germany, and the UK— and in eight different languages —English, French, German, Spanish, Indonesian, Arabic, Dutch, and Italian— discussing a variety of subjects. The collection includes synchronized audio, video (25 frames per second), and motion capture data (collected with a Kinect 2.0) recordings of dyadic interactions in natural situations for more than 25 hours (x2). A portion of this corpus consisting of 48 training sessions and 16 testing sessions (75/25 split) will be used. We sought information on a range of conversation topics ranging from spontaneous behavior

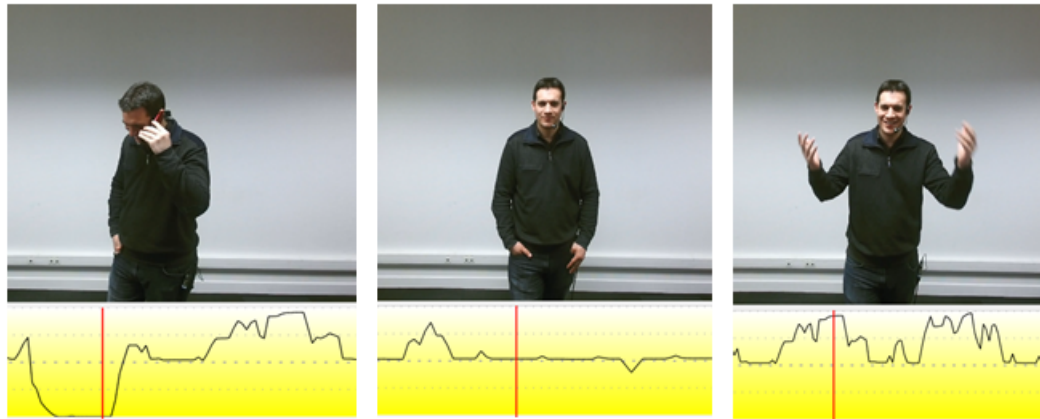


FIGURE 6.1: Snapshots of scenes of a participant in the NOXI corpus being disengaged (left), neutral (center) and highly engaged (right).

in a natural environment. As a result, matching recorded participants based on their shared interests was one of the key design objectives. This means that we first gathered potential experts willing to share their knowledge about one or more topics they were knowledgeable and passionate about, and secondly we recruited novices willing to discuss or learn more about the available set of topics offered by experts. The corpus further introduces interruptions of the novices in order to provoke experts' reactions when conversational engagement gets interrupted. Each session has been continuously annotated for this challenge, which means that each video frame has a score between 0 and 1. At least two (and occasionally as many as seven) annotators completed each rating (3.6 raters on average each session). By finding the mean across all raters, we produced gold standard annotations. The NOXI dataset can be obtained from the website<sup>2</sup>.

### 6.3.2 Bodily Behavior Recognition Task

**Task definition** We model the recognition of bodily behavior as a multi-label classification task. Predicting which of 15 behavior types will be present in a 64 (2.13 second) frame input window is required of challenge participants. We present a frontal image of the target participant as well as two side views (left and right) for each 64-frame window. Due to the extremely uneven behavior classes of the task, we use average precision calculations for each class to aggregate results using macro averaging, which gives each class the equal weight. This encourages participants to challenge themselves to create cutting-edge techniques to boost performance in difficult low-frequency sessions.

As in MultiMediate'21 [147], our challenge is based on the MPIIGroupInteraction dataset [150, 149]. This dataset has served as a basis for diverse tasks, including emergent leadership detection [146], eye contact detection [149, 69, 138], next speaker prediction [22], backchannel analysis [179, 7], and body language detection [11]. The MPIIGroupInteraction corpus consists of 22 group discussions between three to four people, each lasting for 20 minutes [150]. This year's bodily behaviour task is based on the recently collected BBSI annotations [11], consisting

<sup>2</sup>[https://multimediate-challenge.org/datasets/Dataset\\_NoXi/](https://multimediate-challenge.org/datasets/Dataset_NoXi/)

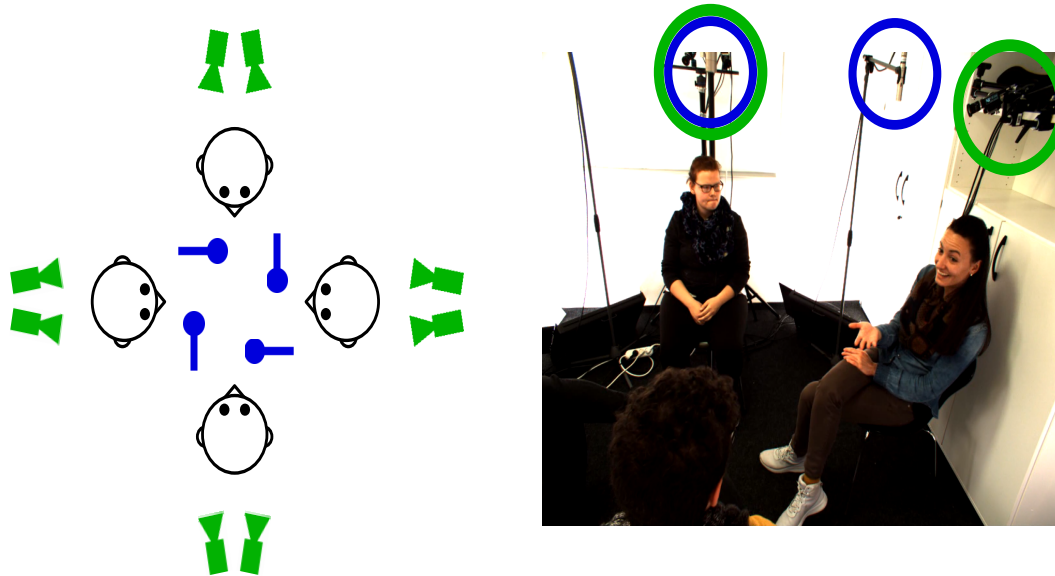


FIGURE 6.2: Setup of the MPIIGroupInteraction dataset. Reproduced with permission from the authors of [150].

of 15 bodily behaviour classes annotated on the whole MPIIGroupInteraction corpus. For MultiMediate’23, we excluded “Lean towards” as inter-annotator agreement was reported to be very low on this class. We collected bodily behaviour annotations for the remaining 14 classes on 996 samples obtained from six unpublished test recordings of MPIIGroupInteraction following the BBSI protocol [11]. To reach high-quality annotations on the test set, we obtained consensus decisions from three annotators. All classes except the “Stretching” class were present on the test set. The MPIIGroupInteraction dataset can be obtained from the website<sup>3</sup>.

## 6.4 Experiments and Results

We are providing a baseline model for each task. This section describes the training methodology as well as the utilized features and results achieved for both tasks.

### 6.4.1 Engagement estimation

#### Approach

We use a set of multimodal variables, including body posture, face features, and vocal features, for the engagement assessment task, followed by a fully connected neural network with three hidden layers. We use a dropout layer with a dropout rate of 0.25 after the second hidden layer to avoid overfitting. The Adam optimizer and mean squared error loss function were used to train the network. Using the KerasTuner framework hyperband search algorithm, all hyperparameters have been optimized [156].

*Head Features.* We extracted features from participants’ head and face using OpenFace 2.0 [12]. All features were extracted for each video frame. The resulting feature

<sup>3</sup>[https://multimediate-challenge.org/datasets/Dataset\\_MPII/](https://multimediate-challenge.org/datasets/Dataset_MPII/)



vectors consist of 68 3D facial landmarks, 56 3D eye landmarks, presence and intensity of 18 action units as well as markers for detection success, detection certainty facial position and rotation. Furthermore, we also use 17 action units provided by the Microsoft Kinect sensor.

*Pose Features.* We extract body pose estimates using OpenPose [29] as well as the Microsoft Kinect sensor data. Each result in the estimation of 350 data points comprises information about the location of various joints as well as their rotation.

*Voice Features.* We retrieved two feature sets over a one-second sliding window with a stride of 40ms to match the frame rate of the video stream for the paralinguistic assessment of engagement. The Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [62] is the first feature set. This set, which consists of 54 audio parameters, is frequently used for tasks like recognizing emotions, mood, and depressive states [215]. Secondly, we used a pretrained version of Soundnet [9] to extract sound embeddings from the raw signal. Soundnet is a deep convolutional neural network that has already been shown to provide effective features for vocal social signal analysis [220]. In our baseline approach, we fused the feature vectors of all modalities into one feature vector. As a large number of features can lead to overfitting we applied a PCA, reducing the number of features to 83 principal components.

## Results

The results are depicted in Table 6.2. Among the single modalities, the vocal features are clearly outperforming the body and head features on the validation set as well as on the test test. However, the multimodal feature fusion shows that the combination of all features still outperforms just using vocal features substantially. We believe that these different modalities are complementary, in fact just as gestures can describe excitement and motivation of speakers, some speakers tends to be calmer but still highly engaged in the discussions and that can be detected in their voice such as the tone they use how often and long they speak. This analysis goes both way, as some people are monotone whereas their gestures are more descriptive. The additional value added by head and body features indicates that the expression of engagement is not clearly bound to one modality but should be analyzed considering multiple modalities.

### 6.4.2 Bodily behaviour recognition

#### Approach

As our baseline solution, we chose the Video Swin Transformer [133], which produced recent state-of-the-art results in action recognition tasks. It operates on fixed inputs of length 32 frames and size of  $224 \times 224$  pixels. Given the input videos of length 64 frames and of larger resolutions, we set the stride to 2, that is we took every second frame, and we resized the video accordingly. We assigned input clips with multiple corresponding behavior class labels and clips of different viewpoints are treated as independent samples during training. To the clips with no labels, we assigned a new behavior class called *Background* and, instead of the 14 classes, we trained the model in a 15-class multi-label setup. To aggregate predictions across views at test time, we averaged the scores obtained from all three views. We used the Swin Base model that is pre-trained on ImageNet and Kinetics-400, and we fine-tuned it on the MPIIGroupInteraction dataset for only one epoch with learning rate

Features	Val CCC	Test CCC
<i>Head</i>		
openface	0.23	0.21
AUs	0.31	0.22
<i>Body</i>		
skeleton	0.47	0.43
openpose	0.53	0.43
<i>Voice</i>		
gemaps	0.58	0.55
soundnet	0.54	0.49
<i>Multimodal</i>		
feature fusion + pca	<b>0.71</b>	<b>0.59</b>

TABLE 6.2: Concordance correlation coefficient (CCC) of our baseline on engagement detection validation and test sets.

$10^{-3}$  and with AdamW optimizer. Our implementation uses the open-source toolbox MMAction2 [46] built on top of PyCharm.

## Results

Results of multiple ablations are reported in Table 6.3. We evaluated our approach against ablations that operate on single views, against an aggregation strategy using the maximum across views, and against not using an additional background class during training. The best mean average precision (MAP) on both validation and test sets was achieved by averaging across views and training with a background class. Although the inclusion of the background class only led to minor improvements, averaging across views yielded consistent improvements. The best single view was the frontal view, and side views resulted in a significant performance drop. All results clearly outperformed the random baseline. The results in the test set are systematically higher, likely as a result of higher quality annotations, and the lack of the “Stretching” class in the test set, which as a result is always evaluated with 1. For this task, we focus only on the RGB input, since bodily behaviors mainly reside in the visual appearance, unlike the task of engagement estimation. For this, we make use of a direct application of the swin transformer [134] as a baseline. The results seem very promising, as the test set has an accuracy of up to 56%. Such results show great potential for many fields. For example, it can assist physicians in their psychological analysis of patients.

## 6.5 Conclusion

We introduce MultiMediate’23, the first challenge addressing engagement estimation and bodily behaviour recognition in social interactions in well-defined conditions. We present publicly available datasets and evaluation protocols for both tasks, and evaluated baseline approaches. The evaluation server will remain accessible to researchers even beyond the MultiMediate challenge, contributing to continuing progress on both tasks. Some important conclusions were drawn through our experiments and results, one of which is the importance of the usage of multimodality in the task of engagement estimation. Results show the complementary nature of voice,

Approach	Val MAP	Test MAP
random baseline	0.0884	0.2355
w/o bkgd class, frontal view	0.3974	0.5315
w/o bkgd class, side view 1	0.3030	0.4341
w/o bkgd class, side view 2	0.3628	0.4893
w/o bkgd class, max of views	0.4087	0.5333
w/o bkgd class, mean of views	0.4084	0.5402
w/ bkgd class, frontal view	0.4051	0.5498
w/ bkgd class, side view 1	0.3096	0.4451
w/ bkgd class, side view 2	0.3686	0.4641
w/ bkgd class, max of views	0.4062	0.5443
w/ bkgd class, mean of views	<b>0.4099</b>	<b>0.5628</b>

TABLE 6.3: Validation and test results for the random baseline and different variants of the Video Swin Transformer.

gestures, and facial expressions. Finally, experiments have proven that computer vision algorithms are capable of capturing social cues in videos, shedding light on new applications of video analysis such as medical assistance for doctors, and assistance with people with social disorders, such as social anxiety and high introversion.

## Chapter 7



# Uncovering Near-Future Abnormal Behaviour via Human Interactions in Real-world Videos

In deep learning and computer vision, the anticipation of abnormal and criminal activities is of significant importance for various applications, particularly in enhancing security measures. By analyzing human behaviors in videos, researchers aim to optimize and enhance human life through effective feedback mechanisms.

One crucial application of such video analysis lies in security measures, where real-time or anticipatory analysis of abnormal and criminal behavior plays a key role in ensuring human safety. Understanding criminal behaviors requires a comprehensive study, as these activities differ significantly from normal daily or sports-related behaviors.

To effectively anticipate such behaviors, a thorough understanding of the entire scene, individual behaviors, correlations between individuals and their surroundings, and interactions among individuals is essential. The analysis should be able to capture both soft cues (e.g., abnormal gaits, gazes) and hard cues (e.g., weapons like knives, bats) as well as the interactions between these various cues. This detailed analysis facilitates the early detection of potential threats and helps ensure proactive security measures.

Abnormal behavior forecasting aims to automatically anticipate unusual behaviors in advance by carefully understanding the early trends of human interactions. Thus, it is the most significant task in surveillance systems that can empower preventive decision making in serious crimes (like abuse, or vandalism) and ensure actionable steps towards perceiving the anomaly (like closing the door during stealing, or robbery). However, due to the existence of complex human behavior and interactions with entities like objects, humans, or both (human and object) in real-world diverse scenarios, abnormal human behavior anticipation is challenging and still underexplored in current research.

In pursuit of this, we present a comprehensive benchmark dataset consisting of critical scenarios and diverse people density. Furthermore, we propose a novel transformer framework as a baseline that dissociatively encodes the temporal correlations and learns human-interaction spatial reasoning to better understand the early human trends and thereby effectively anticipate the near future abnormal human behaviors.



## 7.1 Introduction

Abnormal human behavior often occurs in diverse scenarios in the real world that can trigger chaotic situations and lead to irreversible loss of human life and property. These behaviors have sparse occurrences and are typically characterized by complex and unique spatio-temporal clues, thereby demanding special attention to understand them. Recently, proliferated video anomaly understanding methods have majorly aimed at automatically comprehending abnormal scenarios either in an offline or online manner.

Although online and offline anomaly comprehension methods can assist in timely alarming and the post-anomaly investigation, respectively, they fail to facilitate any anomaly preventive measures. For this, abnormal behavior forecasting/anticipation in real-world scenarios has a high social impact and the greatest need of the hour to minimize casualties and damages through mitigation measures. The functional difference between video anomaly anticipation with corresponding online and offline detection tasks is clearly illustrated in Figure 7.1.

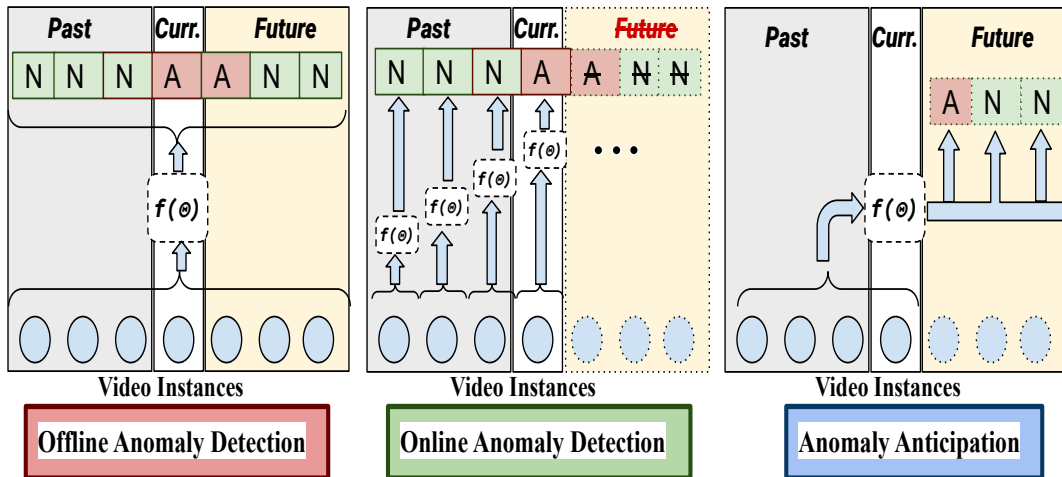


FIGURE 7.1: The differences between anomaly anticipation with offline and online anomaly detection task, where  $f(\theta)$  is the functionality of the respective methods.

Considering that real-world scenarios are often unbounded and dynamic, analyzing complex human behavior, interaction, and their influence on other entities (objects and humans) for abnormal behavior forecasting is challenging and is vastly divergent from those in constrained daily life situations. Further, it becomes more complex due to the existence of large variability in people's densities and their interaction in abnormal scenarios with a sophisticated involvement of sharp and subtle spatio-temporal cues. For example, as shown in Figure 7.2 shoplifting can be characterized by subtle signals with few human-to-object interactions involved, while protest abnormalities have strong signals with dense human-to-human and human-to-object interactions. Unlike human actions in daily life, these abnormal behaviors are often compounded by the fact that they can occur sparsely in untrimmed real-world footage, thereby making them harder to anticipate. To the best of our knowledge, there exist no benchmark methods that can handle such sparsity, disparity among abnormal behaviors, and anticipate in real-world scenarios.

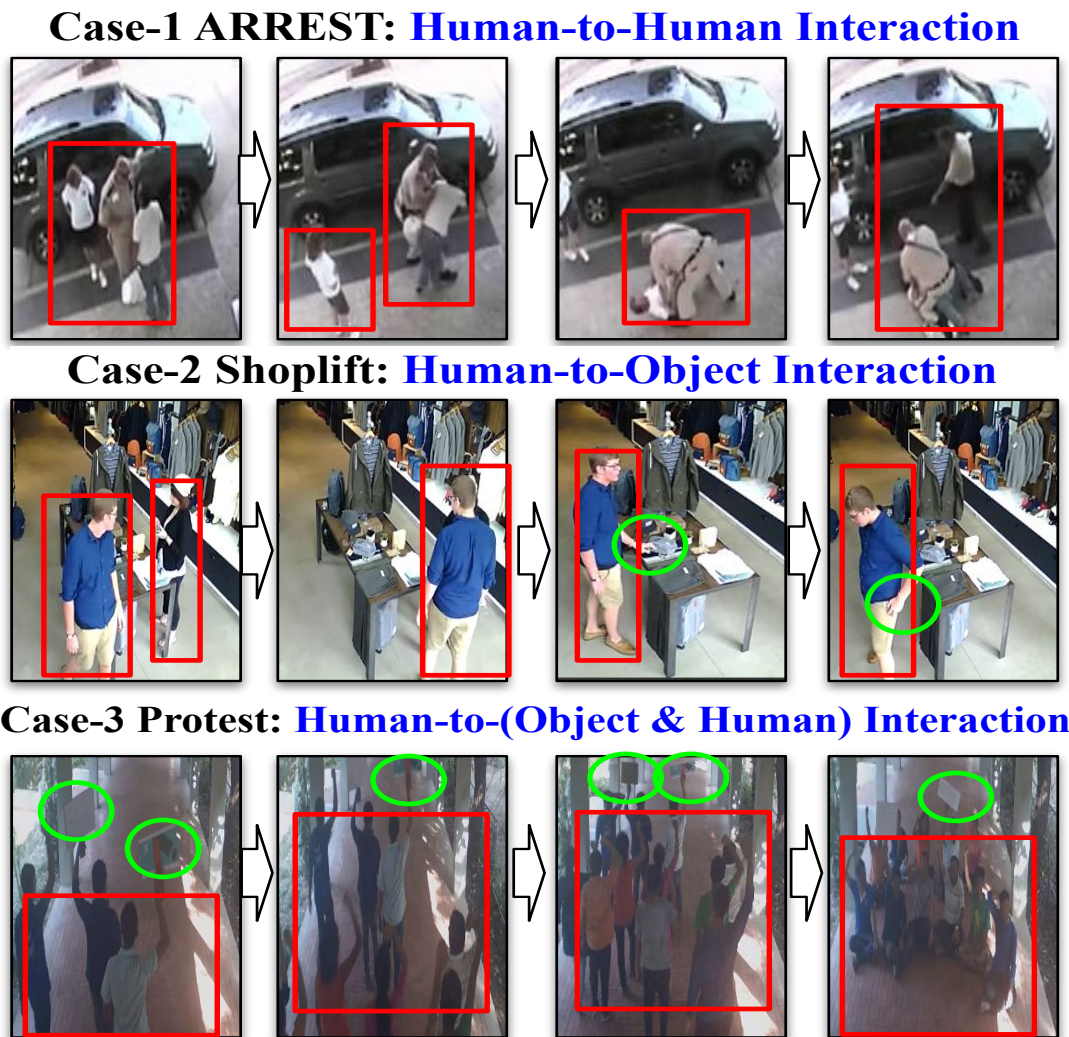


FIGURE 7.2: Illustration of complexity in abnormal human behavior. Notice the three different cases of Interactions with divergent spatio-temporal cues. Abnormal human-to-human interactions (e.g. arrest) have often significant appearance and motion change whereas human-to-object interactions (e.g. shoplifting) are often subtle. However, there can be abnormal human-to-(object & human) interactions like *protest* that have a unique spatiotemporal blend with large people density.

Forecasting abnormal human behavior requires a holistic understanding of early behavioral tendencies. Along this direction, there exist several methods developed for action anticipation in daily living that first encode the contextual representation of early trends and then predict the future. However, most of the previous methods rely on the global scene-level temporal representations to encode the early/observatory context. Since real-world anomalies have large diversities in subtle and sharp cues among objects, humans, and scene-localized regions, focusing only on global temporal dynamics for early trend context modeling leads to a partial understanding of complex scenarios which is a major drawback of recent methods. Thus, combining both scene-level temporal and object-level spatial semantics is critical, as future abnormal behavior-relevant cues may pertain to either one or both.

Motivated by this, we propose a novel “transformer encoder”, namely Spatial



Interaction aware Transformer (SIaT), that comprises two major building blocks: (i) temporal reasoning module (TRM), and (iii) spatial interaction module (SIM) to foster early trends of human behavior modeling. Unlike the previous method, the TRM and SIM dissociatively encode the scene-level temporal consistencies and object-level spatial interaction reasoning, respectively, to promote coarse-to-fine contextual understanding of the early behavioral trends.

The key difference between the previous methods and our approach is presented in Figure 7.3. In this work, we utilize panoptic object masks and raw RGB frames to represent object- and scene-level agents, respectively. By this, both the modalities become spatially coherent and thus it allows the SIaT to explicitly encode the correlation between scene-level temporal dynamics and object-level spatial interactions. This correlation among scene- and object-level semantics is encoded by projecting the latent feature distributions of TRM and SIM in a joint activation space by exploiting the contrastive likelihood.

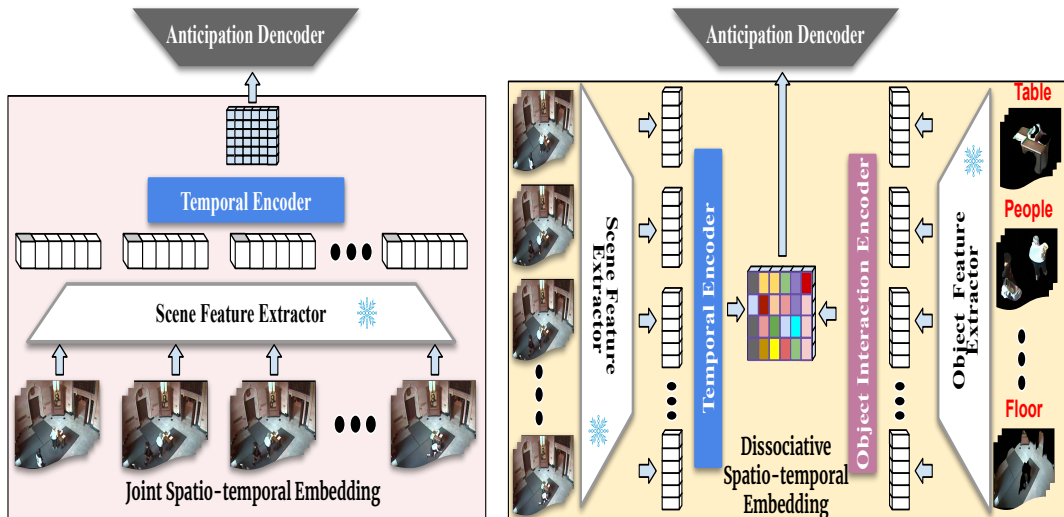


FIGURE 7.3: Comparison of previous anticipation-based methods with ours in early trend modeling. Previous methods consider only scene-level features (*i.e.* from the whole frame) to encode joint spatio-temporal embeddings, thereby they have a partial understanding of complex abnormal behavior. In contrast, our dissociatively learn the scene-level temporal consistencies and object-level spatial interaction to obtain a better understanding of early trends.

To summarize, our contributions are in three-folds:

- A novel task of abnormal activity anticipation (AAA)
- A new benchmark for the proposed task called ED-Crime
- A baseline model namely Spatial Interaction aware Transformer (SIaT)

## 7.2 Related Work

Based on the availability of video data and annotations, anomaly detection methods can be divided into three broad categories *i.e.* supervised, unsupervised, and semi-supervised.

**Supervised:**

Supervised methods of anomaly detection task assume that both normal and anomaly video patterns along with the frame-level annotations are available for learning. These methods are only applicable to detect a specific pre-defined set of anomaly patterns. Authors in [152] aim to detect violence activity such as *fighting* in hockey playing environment. As the dataset used in [152] contains trimmed videos from both normal and anomaly video patterns, it becomes a binary (*fighting vs. non-fighting*) video classification task. This method is only applicable to detect anomaly categories where fighting is involved. Similarly, the authors in [191, 208, 199] detect violent activities inside an elevator, loitering activity, and falls of elderly people in a supervised fashion.

Although the supervised method of anomaly detection can detect a specific anomaly category, it requires frame-level annotations for training video data. Obtaining these annotations is laborious and time-consuming. Along with this, defining all possible anomaly patterns during training is also a difficult task. An ideal anomaly detection method should learn to maximize the separation between normal and anomaly video patterns during training so that anomaly video patterns can be detected during testing based on the line of separation from training.

**Unsupervised:**

Unsupervised methods of anomaly detection tasks need no labeling of training data. It aims at extracting statistical properties from the unlabeled video data. The authors in [28] utilize the K-means clustering to group vehicular trajectories and use the hidden Markov model (HMM) to establish intra-cluster path patterns. Similarly, Hu *et al.* [94] represented the trajectory patterns of vehicles with a chain of Gaussian distribution and used hierarchical clustering for traffic abnormality detection. They have also proposed a novel Dirichlet process mixture model (DPMM) for trajectory representation and clustering. Moreover, authors in [104, 17] have obtained Harris corner and object-based trajectories respectively for clustering of normal and anomaly patterns.

It may be noted that unsupervised approaches for anomaly detection do not require any prior knowledge of data. However, these methods are based on the assumption that *anomaly patterns are rare compared to that of normal ones*. This assumption of an unsupervised approach may not hold *true* in all cases. Furthermore, many unsupervised approaches utilize hierarchical and probabilistic clustering, which may lead to unreliable results. In addition, unsupervised approaches have more computational complexity compared to other methods.

**Semi-supervised:**

Semi-supervised approaches for the anomaly detection task do not need densely labeled video data, unlike the supervised approach for model training. Videos with minimum prior knowledge are sufficient for the anomaly detection task. Based on the availability of video data during training, the semi-supervised approach can be divided into two subcategories *i.e. one-class learning and weakly-supervised*. In *one-class learning* approach only normal video data are required for training, which is easy to obtain. However, the *weakly-supervised* approach requires videos containing both normal and anomaly patterns with video-level supervision termed as *weak-supervision* for training. A detailed description of both approaches is given below.





**One-class learning:** In one-class learning, training is performed on normal video patterns, and at test time, deviation from normal pattern is treated as an anomaly. In these methods, obtaining normal videos is a relatively easy task because in surveillance videos normal patterns occur very frequently. Anomaly detection using one-class learning methods is defined for a specific scene. Authors in [169] formulate the anomaly detection task in a street scene using auto-encoders and a dictionary-based approach. The dictionary-based approach of Lu *et al.* [137] learns each spatial region of normal videos independently and treats anomaly patterns as an outlier. Similarly, the auto-encoder based approach of Hasan *et al.* [84] uses a deep auto-encoder network trained on pixel reconstruction error for anomaly detection tasks. This method is based on the assumption that *video clips containing anomaly patterns will not be reconstructed, unlike normal patterns*. Similarly, authors in [223] aim at detecting out-of-position of drivers pose in a vehicle environment. This method is inspired by the method of one-class SVM (OC-SVM) [177] and least square OC-SVM [246] where multiple hyperplane leads to maximal margin of separation. To combat this, Wang *et al.* [223] have proposed a one-class discriminative subspace (BODS) classifier that uses a pair of hyperplanes that is optimized through a non-convex optimization technique.

Since one-class learning methods require all normal patterns defined for specific scenes for model training, these methods suffer from the fact that *it is often difficult to learn feature representations for a wide diversity of normal patterns* and hence these methods are not suitable for general applications of anomaly detection task.

**Weakly-supervised:** To overcome the drawback of earlier methods, recent approaches [202, 245, 253, 252] learn feature representation from normal and anomaly videos. This formulation of anomaly detection as the binary class has been introduced with the dataset UCF-Crime [202]. The success of deep learning models [233, 251] in action detection motivates the researchers [202, 252] to make use of 3D convolutional networks (ConvNets) as the visual backbone for segment-wise feature extraction. By segments, we mean short video clips partitioned from an untrimmed video. Training 3D ConvNets like C3D, I3D pre-trained on huge datasets like Kinetics [110] and Youtube-8M [1] requires strong supervision for learning spatio-temporal patterns in a short video clip. Whereas obtaining temporal annotation for anomaly activities is a laborious task. Thus, [202] addresses the task of anomaly detection under weak supervision which makes use of video-level annotation for untrimmed anomalous and normal videos. [202] have proposed a MIL-based model to map video-segment based feature vectors to an anomaly score. This mapping is learned through a ranking loss which optimizes the separation of the anomaly and normal segments in a video. Inspired by previous studies, authors [253] use the motion-aware features with the MIL model to improve anomaly detection. In addition, they also employ an attention block to incorporate temporal context while detecting anomalies. Besides relying completely on the input features, the attention is applied only at the score level. Thus, the attention block does not modulate the feature maps leading to non-optimal detection. [245] model the temporal relation through Temporal Convolutional Network (TCN) [122] and also proposed a novel complementary loss to maximize the margin of separation between inter and intra-class instances. However, obtaining only long-range temporal dependency is not sufficient for detecting anomalies in video. Another approach [252] formulates the weakly-supervised problem as a supervised learning refining noise labels iteratively. But such methods are costly in terms of inference and are also data dependent, for instance, they rely too much on strong motion for detecting anomalies.





Module that constitutes two identical modules with different functionalities, namely Temporal Interaction Module (TIM) and Object Interaction Module (OIM) to dissociatively capture the scene-level global temporal interactions and object-level local spatial interactions respectively; (II) Contrastive Attention Encoder combines the distinctive interactions encoded local spatial and global temporal embedding by exploiting the contrastive likelihood. Next, we proceed to provide a concise description on the feature encoder and each building block of SIaT.

### 7.4.1 Feature Encoder

For a given temporal observation duration ( $t$ ), we extract scene and Object features from the Scene Encoder (SE) and Object Encoder (OE). The **OE** first extracts the frame-level panoptic masks with the corresponding text labels of  $k$  objects from Mask2former [43] and stacks them along the temporal dimension ( $t$ ). Then we extract object-level  $d_0$  dimensional vision language features from CLIP [CLIP] Image and Text encoder to obtain  $F_O \in \mathbb{R}^{t \times k \times d_0}$  and  $F_{txt} \in \mathbb{R}^{t \times k \times d_0}$  respectively. Then the **SE** extracts  $d_0$  dimensional frame-level spatial features from CLIP [CLIP] Image encoder and stacks them along the  $t$  dimension to obtain a global scene feature map  $F_S \in \mathbb{R}^{t \times d_0}$ .

### 7.4.2 Interaction Modules (TIM/OIM)

The aim of interaction modules is to learn low-level discriminative representations for future AHB w.r.t normal events by effectively encoding the global and local interactions in the temporal interaction module (TIM) and object interaction module (OIM) respectively. This is enforced by dissociatively encoding scene and object-relevant early anomaly sharp and subtle clues via TIM and OIM. For this, TIM first aims to highlight the temporal saliencies of the observation by encoding the cross temporal interactions among coarse-grained scene ( $F_S \in \mathbb{R}^{t \times d_0}$ ) and fine-grained object  $F_O \in \mathbb{R}^{t \times k \times d_0}$  level feature maps. While processing the  $F_O \in \mathbb{R}^{t \times k \times d_0}$  in TIM, a spatial-pooling operation is applied on  $k$  dimension of  $F_O$  to suppress the object appearance features and encourage the object-specific fine motion features. Next, OIM aims to promote the salient object features out of many irrelevant ones by encoding their spatial interactions with the surroundings. Although individual object mask features ( $F_O \in \mathbb{R}^{t \times k \times d_0}$ ) are empowered with fine-grained representations, they are contextually sparse. Further, encoding the object interaction with sparse context leads to a partial understanding of the complex interactions (e.g. ambiguity between arrest and fighting w/o a policeman as context). Due to this, CLIP pre-trained object-level textual feature  $F_{txt} \in \mathbb{R}^{t \times k \times d_0}$  is taken into consideration for infusing rich contextual information while encoding critical object interactions in OIM. When processing both fine-grained  $F_O$  and contextual  $F_{txt}$  in OIM, a temporal-pooling operation is applied on  $t$  dimension of  $F_O$  and  $F_{txt}$  to suppress the object motion features and focus on the object appearance and spatial location features. Although TIM and OIM encodes two distinct representations, they are functionally and architecturally identical.

**Functionality of TIM/OIM** Primarily, the TIM/OIM learns the temporal and object level spatial interaction by encoding the cross-correlation between the respective fine-grained and contextual representations. Figure 7.5 shows a detailed view of TIM/OIM with their respective input feature maps. First the, the input feature maps are projected to parallel 1D-conv layers, each having  $m_1$  1D conv filters with kernel

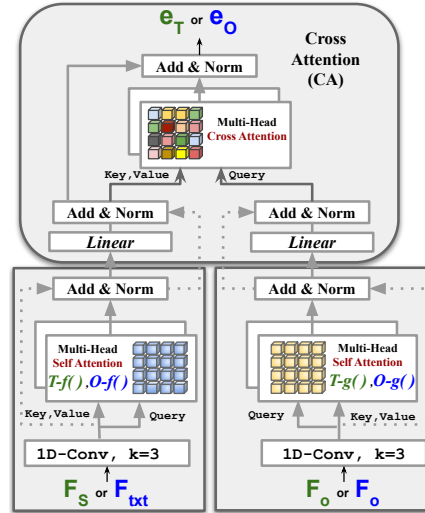


FIGURE 7.5: Overview of architecturally identical Temporal Interaction Module (TIM) and Object Interaction Module (OIM). Note that the figure is color coded. Here,  $F_{txt}$ ,  $F_O$  are the temporal-pooled textual and object mask features and  $F_O$  is the spatial pooled object mask feature.

size  $k \in \{3\}$ . These local projections are made to enhance the low-level semantics of scene dynamics and object spatiality by embedding the temporal and spatial localities in TIM and OIM. Next the locality-aware feature maps are processed in parallel to encode the all-pair self-correlation via a multi-head self attention. Next, the correlation maps generated from all heads are combined separately. The resultant is then *added* and *normalized* through a skip-connection to the respective *value* for retaining the local inductive bias of temporal and spatial input feature maps. Next, the individual fine-grained and contextual self-correlation-encoded maps are first latent activated by a *linear* layer and then fed to the standard multi-head cross attention (MHCA) [MHA] for computing the cross-interactions between coarse-to-fine features of temporal and object level features in TIM and OIM. The cross-interaction is captured by treating the contextual latent features as the *key* and *value* and the fine-grained latent features as the *query* of MHCA. The temporal and object interaction encoded outputs of TIM and OIM is represented by  $e_T \in \mathbb{R}^{t \times d_1}$  and  $e_O \in \mathbb{R}^{k \times d_1}$ , where  $t$  is the observation length  $k$  corresponds to associated objects and  $d_1$  is the embedding dimension.

### 7.4.3 Co-Attention Encoder(CAE)

To enhance feature mixing in HS<sup>2</sup>T during training the Text Inducer (TI) shown in Figure ??(c) semantically associates the CLIP based textual feature and human-scene augmented visual features. Such an association is learned by computing the video-to-category ( $M_c$ ) and video-to-object ( $M_o$ ) maps. For this, TI first separately inputs a pre-defined **text codebook** that has  $D_1$  dimensional embedding for each text feature. We utilize the pre-trained frozen CLIP text encoder to construct the text codebook. The text codebook has three types of embeddings: (i) abnormal category text  $E_C \in \mathbb{R}^{N_0 \times D_1}$ , (ii) object text  $E_O \in \mathbb{R}^{N_1 \times D_1}$ , (iii) learnable text  $E_L \in \mathbb{R}^{N_2 \times D_1}$ , where  $D_1$  is the embedding vector dimension and  $N_0, N_1, N_2$  are the number of abnormal categories, objects and learnable queries present in text codebook. The goal of learnable queries  $E_L$  is to iteratively update the text codebook with the missing object information in the predefined  $E_O$ . Then, it inputs the human-scene augmented feature



map  $F^{**} \in \mathbb{R}^{T \times nD}$  and projects it to a  $FC$  layer with  $D1$  units for making the vision and text embedding dimensions analogous. The output of  $FC$  layer is denoted by  $E_V \in \mathbb{R}^{T \times D1}$ . Now, to define the correspondence between  $T$  temporal regions and  $N_0$  abnormal category features, we construct the video-to-category map  $M_c \in \mathbb{R}^{T \times N_0}$ .  $M_c$  is computed by  $\text{softmax}(E_V \otimes E_C)$ ,  $\otimes$  being the Kronecker product. Similarly, we construct the video-to-object map  $M_o \in \mathbb{R}^{T \times N_3}$  to define the correspondence between  $T$  temporal regions and  $N_3$  object features, where  $N_3 = \text{concat}(N_1 + N_2)$ .  $M_o$  is computed by  $\text{softmax}(E_V \otimes \text{concat}(E_O, E_L))$ . The  $M_c$  and  $M_o$  learn the vision-text correspondence maps dissociatively while being abnormally agnostic. To learn anomaly-aware features, we apply  $\mathcal{L}_A$  that semantically binds  $M_c$  and  $M_o$  by focusing on the abnormal temporal segments obtained from the detector score map  $S \in \mathbb{R}^{T \times 1}$ .

#### 7.4.4 Anticipation Decoder

The decoder takes learnable tokens as input, referred to as *anomaly queries* and the outputs of CAE i.e.  $\theta_1, \theta_2$  to predict the future labels. It also learns the long-term action relation between the observed and future anomaly via self attention and cross attention. The anomaly queries are embedded with  $M$  learnable tokens  $Q \in \mathbb{R}^{M \times d1}$ . The temporal orders of the queries are fixed to be equivalent to that of the future anomalies, i.e., the  $i$ th query corresponds to the  $i$ th future anomaly. The decoder consists of two sequential multi-head cross attention(MHCA), layer norm (LN) and MLP. The final output of decoder is computed by following (7.1) and the output logits  $\hat{A}$  are then *softmax* activated.

$$\hat{A} = \text{MLP}(\text{LN}(\text{MHCA}(\theta_2, \text{MHCA}(\theta_1, Q)))) \quad (7.1)$$

**Training Objective** The  $M$  number of action queries are matched to the  $N$  number of ground-truth actions to apply action anticipation loss  $\mathcal{L}^{\text{anticipate}}$ . The  $\mathcal{L}^{\text{anticipate}}$  loss is defined with cross entropy between action  $A$  and logits  $\hat{A}$ .

## 7.5 Experiments

The experiments are conducted on our Criminal Human Behaviour dataset (CHB), as well as on two additional datasets that include training samples from both, human and scene-centric anomalies, namely UCF-Crime (UCF-C) [202], and IITB-Corridor (IITB-C) [IITBC]. There exist four major limitations in previous datasets, diverging from real-world settings: **(I)** Events performed by actors with simple background limit the generalization capabilities to real-world anomalies, **(II)** Single-type anomaly datasets such as only fighting or accident limit the scalability of the model, **(III)** Real-world datasets like UCF-Crime do not carry sufficient evaluation samples to cover various human and scene anomaly categories, and **(IV)** lack of temporally annotated videos. Although weakly-supervised settings can work on video-level labels, having precise temporal annotations for a complete dataset enables the model to evaluate multiple aspects (such as K-Fold evaluation). Motivated by this, we create ED-Crime to overcome above limitations. A detailed comparison with public datasets is described in the **Appendix**.





## 7.6 Preliminary results

## 7.7 State-of-the-art Comparison

Methods	Short (mAP)				Long (mAP)		
	1 sec.	2 sec.	3 sec.	Avg	4 sec.	8 sec.	Avg
===== Backbone with Scene							
ViT	-	-	-	-	-	-	-
Swin	-	-	-	-	-	-	-
CLIP	-	-	-	-	-	-	-
SoTA with Scene Feature							
OADTR	62.37	61.58	62.11	62.02	61.58	56.10	58.84
FUTR	62.53	59.89	60.42	60.94	61.21	55.67	58.44
LSTR	-	-	-	-	-	-	-
JOADAA	61.21	61.21	61.47	61.29	60.02	55.67	57.84
TesTra	62.53	61.21	63.32	62.35	62.00	58.00	60.00
SoTA with Object Feature							
OADTR	50.65	51.18	51.18	51.00	50.65	46.96	48.80
FUTR	59.10	58.31	55.40	57.60	55.40	50.65	53.02
LSTR	-	-	-	-	-	-	-
JOADAA	55.93	56.46	55.40	55.93	55.14	49.07	52.10
TesTra	55.40	56.20	54.35	55.31	55.14	51.48	53.31
SoTA with Scene+Object Feature							
OADTR	63.37	62.90	62.58	62.95	62.06	59.13	60.59
FUTR	61.47	61.21	61.74	61.47	62.79	56.46	59.62
LSTR	63.06	62.00	63.06	62.70	63.06	59.63	61.34
JOADAA	62.79	62.79	62.53	62.70	62.00	57.51	59.75
TesTra	63.85	63.32	62.80	63.32	62.53	59.10	60.81
===== SLaT (ours)	65.96	64.64	63.85	64.81	63.85	59.63	61.74

TABLE 7.1: State-of-the-art comparisons on UCF-C, IITB-C, and ED-C datasets.

This is an ongoing work, and the presented results are preliminary; nevertheless, the results are promising and motivate us more to invest on this work. We achieve state of the art results compared to different existing methods, which proves that our proposed solution is more suitable for such scenarios.

## 7.8 Conclusion

Security presents an important aspect of human daily life. Therefore, building preventive systems for criminal and abnormal activities is very important. In this work, we present a novel task of abnormal activity anticipation. We also present a novel ED-Crime dataset, and a new transformer-based baseline SLaT and benchmark dataset. Our intuition is to tackle criminal activities as a human-to-environment interaction problem. However, unlike our first work on fine-grained and HOI (human object interactions) actions, we added two other types of interaction, H-H and H-A (human-to-human and human-to-animal interactions). Moreover, these action presents other challenges, one of which is that these actions happen in crowded areas; hence our use of the contrastive likelihood that helps highlight non-salient features that may carry crucial information about the action. The presented results are preliminary; however, ever they are promising. We believe that our work will greatly impact the community and opens up a new research area.

## Chapter 8

# Discussion and Future Work

In the final chapter, we summarize the contributions of this thesis and depict the future work directions.

### 8.1 Contribution summarization

In this concluding chapter, we reflect on the journey undertaken in exploring action detection and recognition within the scope of this thesis. The attempt to decipher human actions from visual data has led to a comprehensive understanding of the challenges and nuances associated with these tasks. Through rigorous investigation, experimentation, and analysis, we have navigated the intricacies of real-time action detection, unraveling the complexities of identifying and tracking dynamic activities in video sequences. At the same time, the pursuit of action recognition has involved delving into the finer details of discerning and categorizing these actions, contributing to the broader landscape of computer vision and artificial intelligence. As we draw the curtain on this research venture, we acknowledge not only the accomplishments achieved but also the inherent intricacies that continue to propel these fields forward. The insights gained from this thesis serve as a foundation for future research endeavors, highlighting the ever-evolving nature of action detection and recognition in the pursuit of intelligent systems capable of understanding and responding to human activities. The summarization of each of our contributions follows.

**THORN: Temporal Human-Object Relation Network for Action Recognition:** In this work, we focus on actions of human-object interactions. Actions of the type Human-Object Interaction (HOI) hold particular significance in the context of real-life and human daily activities, making them inherently more representative of our day-to-day experiences. Unlike isolated actions, HOIs capture the dynamic interplay between humans and objects, reflecting the complexity of human interactions with the surrounding environment. Human daily activities are often characterized by the relationships between individuals and the objects they interact with, such as cooking, driving, or working. Recognizing and understanding these HOIs is crucial for developing artificial intelligence systems that can comprehend and respond to human behaviors. The challenges of recognizing HOI arise from the need to discern subtle visual cues that indicate interactions, motions, and relationships within a scene. The potential variability in object usage and the intricate choreography of human-object interactions add layers of complexity to the recognition process. To answer such challenges, we developed a new architecture that can capture relevant visual cues in videos, and then recognize the complex relationships between humans and objects, our approach employs graph convolutions to model the complex dependencies and contextual nuances within a scene. By treating objects as nodes



and their interactions as edges in a graph, we enable the model to discern and propagate information efficiently across the interconnected elements. This strategic use of graph convolutions not only refines the representation of objects, but also captures the dynamic interplay between them, laying the foundation for precise action predictions in HOI scenarios. In this work, we prove that object-centric reasoning is important for the recognition of action of human-object interaction in contrast to 3D-CNNs that can only capture overall motion and features in video clips.

**JOADAA: Joint Online Action Detection and Action Anticipation:** In this work, we discuss the other aspect of action analysis in videos, which is action detection. Action detection and localization are of paramount importance in the domains of deep learning and computer vision, as they enable machines to comprehend and respond to dynamic human activities within visual data. Accurate identification and localization of actions within a video sequence are fundamental for various applications, including video analysis, surveillance, and human-computer interaction. Challenges in this domain include handling occlusions, diverse viewpoints, and the temporal dynamics inherent in real-world scenarios. Moreover, the need for real-time processing adds an additional layer of complexity. In our work, we narrow our focus to the specialized realms of online action detection and anticipation, recognizing their increased relevance to real-life applications. Online action detection involves identifying actions as they unfold in real-time, allowing for timely responses and intervention. On the other hand, action anticipation takes a proactive approach, predicting future actions before they occur, enhancing the adaptability and efficiency of intelligent systems in dynamic environments. In this work, we propose a new venue based on joint learning action anticipation and online action detection. In fact, one of the utmost limitations of online action detection is limited information as one has access to only past and present information; hence with no knowledge of future, it has to infer actions compared to offline action detection. Therefore, our intuition is to bring as much knowledge of the future as possible to online action detection models. Hence, we implement a midlayer of action anticipation that anticipates future time steps and use them as a pseudo-future to improve accuracies on online action detection. Moreover, we explore another aspect and challenge of online action detection, we study two types of datasets, densely annotated datasets and sparsely annotated datasets. We show that handling temporal information is very important, for instance, one would think that bringing as much knowledge from the past is always good. However, we prove that in the case of sparsely annotated datasets, too much information can act as noise to our predictions. Many other results have been discussed in the corresponding chapter. Nevertheless, this work proves that our new approach is very interesting and opens new ventures for researchers in online action detection and anticipation.

**Robust and Efficient Multimodal Multi-dataset Multitask Learning:** Transfer learning has become increasingly essential due to the vast availability of data and the prevalence of multi-modalities in various domains. The need for transfer learning arises from several factors:

**Data Abundance:** In many fields, there is an abundance of data available. Transfer learning allows models pre-trained on large datasets to leverage the knowledge gained from that data when faced with a new, possibly smaller dataset. This is particularly beneficial as collecting labeled data for a specific task is expensive and time-consuming.

**Multi-Modalities:** In modern applications, data often come in various modalities, such as images, text, and audio. Transfer learning enables models to understand and leverage knowledge across different modalities. For example, a model trained on image data may be fine-tuned for a related task using textual information.

**Robust Feature Learning:** Transfer learning promotes the development of robust models that can learn general features from one domain and adapt them to another. Instead of starting from scratch, models initialized with pre-trained weights can capture common patterns, enabling them to learn more efficiently and effectively, especially when labeled data are limited.

**Domain Adaptation:** Transfer learning helps address the challenge of domain adaptation, where the source and target domains may differ. Pre-training on a source domain allows models to learn generic features, making them more adaptable and facilitating fine-tuning on a target domain.

**Reduced Training Time:** Training deep neural networks from scratch on large datasets can be computationally expensive and time-consuming. Transfer learning accelerates the training process by leveraging pre-existing knowledge, making it more feasible to apply deep learning techniques to real-world problems.

In summary, transfer learning is crucial in contemporary machine learning scenarios due to the wealth of available data, the diversity of multi-modal information, and the efficiency gained by initializing models with learned features. Robust models capable of learning inherent features through transfer learning not only improve performance but also make it easier to fine-tune them for specific tasks, providing a practical and effective approach to handling complex real-world challenges. To this end, and with the rise of transformers, we introduce CM3T, a framework designed for incorporating commonly pre-trained video classification models into a transformer-based architecture. This framework comprises three modules: two introduced by us, namely multi-head vision adapters and cross-attention adapters, and one pre-existing module, prefix tuning. In particular, we demonstrate the effectiveness of these modules without relying on specific pre-training or training methods, exploring various variants. Our work narrows the gap between research and practical applications in video classification models by simplifying the adaptation of existing approaches to new datasets and tasks. Additionally, it facilitates the incorporation of any emerging modalities that may be present in the data.

**MultiMediate'23: Engagement Estimation and Body-Behavior Recognition in Social Interactions:** Recognition of human activity in videos is not limited to high-motion actions. In order to build AI and robot systems capable of assisting humans in their daily lives, it is crucial to understand social cues, and also human behaviors and social interactions. To this end, we propose the MultiMediate'23 challenge, where we propose two tasks and a new dataset.

The first task is **Engagement Estimation Task**. The task at hand involves the frame-by-frame prediction of individual participants' conversational participation levels, quantified on a continuous scale from 0 (lowest) to 1 (highest). Emphasis is placed on exploring the multimodal and reciprocal behaviors exhibited by both interlocutors within the Novice-Expert Interaction corpus. The evaluation of the predictions in the test set is conducted using the Concordance Correlation Coefficient (CCC) [128].

The second task is **Bodily Behaviour Recognition**. We approach the recognition of bodily behavior as a multi-label classification task, where participants in the challenge are asked to predict the presence of 15 behavior types within a 64-frame (2.13 seconds) input window. Each window includes a frontal image of the target

participant and two side views (left and right). Given the notable imbalance among behavior classes, we employ average precision calculations for individual classes, aggregating results through macro-averaging to ensure equal weight for each class. This methodology encourages the challenge participants to devise innovative techniques aimed at enhancing performance, particularly in challenging low-frequency sessions. Finally, for the datasets we collected novel annotations on the NOvice eXpert Interaction (NOXI) database. For bodily behavior recognition, we annotated test recordings of the MPIIGroupInteraction corpus with the BBSI annotation scheme. Additionally, we present baseline results for both challenge tasks.



**Uncovering Near-Future Abnormal Behavior via Human Interactions in Real-World Videos.** In this work, we dive into an application centered on human activity recognition, as we focus on abnormal activities. To this length, we present three contributions: a novel task of abnormal activity anticipation (AAA), a new benchmark for the proposed task called ED-Crime, and a baseline model, namely Spatial Interaction-Aware Transformer (SIaT). Given the considerable diversity in subtle and pronounced cues among objects, humans, and scene-localized regions in real-world anomalies, a sole focus on global temporal dynamics for early trend context modeling results in a limited understanding of complex scenarios. This limitation constitutes a significant drawback of recent methods. Hence, the integration of both scene-level temporal and object-level spatial semantics is imperative. Future cues relevant to abnormal behavior may pertain to either one or both, and thus, a holistic approach is essential for a more thorough and accurate understanding of complex scenarios. To tackle such limitations, we introduce a new "transformer encoder" known as Spatial Interaction-aware Transformer (SIaT), which consists of two key components: (i) the temporal reasoning module (TRM) and (ii) the spatial interaction module (SIM). These elements are designed to enhance the modeling of early trends in human behavior. In contrast to previous methods, the TRM and SIM autonomously encode scene-level temporal consistencies and object-level spatial interaction reasoning, respectively. This approach aims to advance a comprehensive contextual understanding of early behavioral trends, promoting a nuanced perspective from coarse to fine. Our work is still in progress; however, we present interesting results and a new approach. We aim to improve these results even further.

## 8.2 Limitations and Perspectives

In this chapter, we analyze the limitations of the proposed methods and shed light on possible improvements and future directions.

### 8.2.1 fine-grained activity recognition

In our earlier efforts in **fine-grained activity recognition**, the efficacy of our work was impeded by the limited availability of object features. During that period, the absence of annotations on object bounding boxes presented a significant challenge, making it challenging to capture crucial object-centric information. To overcome this limitation, we integrated a pseudo-supervised object detection solution into our

framework. However, the landscape has evolved with the introduction of innovative methods like DINO (self-distillation with **no** labels) [32]. DINO makes the following observations: first, self-supervised ViT features contain explicit information about the semantic segmentation of an image, which does not emerge as clearly with supervised ViTs, nor with convnets. Second, these features are also excellent k-NN classifiers, reaching 78.3% top1 on ImageNet with a small ViT [114]. While DINO does not directly contribute to object detection, its advances in self-supervised learning open new possibilities for improving feature representation learning, including object-centric features. By enhancing the model's ability to understand and differentiate between diverse visual patterns, DINO indirectly supports the extraction of relevant object information, contributing to more comprehensive fine-grained activity recognition. Moreover, vision-language models, such as CLIP (Contrastive Language Image Pre-training) [167], have demonstrated substantial advancements in understanding visual content and textual information, offering potential benefits for improving the recognition of human-object interactions, particularly in the context of action recognition. Understanding actions often requires considering contextual information. Vision language models capture contextual relationships between objects, scenes, and actions, providing a more holistic understanding of the visual context. Incorporating such models into a comprehensive action recognition system can lead to improved performance and a deeper understanding of human-object interactions in diverse visual contexts.

Action recognition in human-object interaction (HOI) datasets is highly dependent on capturing the salient interactions between objects and understanding their dependencies. Information such as "object1 on top of object2", "object1 near object2" or "human looking at objectx" when incorporated in our models can help improve better classification and detection of such actions. In fact, having an adjacency matrix that carries and can learn such information is a big step in modeling human interaction with its surroundings. Some datasets like **Action genome** [100] are starting to provide such annotations where the aim is to model actions as compositions of spatio-temporal scene graphs. Nevertheless, there has been no direct work to learn specific and concrete interactions. Hence, this opens a new perspective of research, where we can incorporate and learn concrete interactions in our networks. Another way to strengthen fine-grained action recognition is to introduce object tracklets or bounding boxes over time. With such semantics, we can enrich the input of the modules. For instance, the introduction of a two-dimensional feature map ( $Object \times T$ ) to the temporal module provides a means of investigating object-temporal relations. However, the prevailing object detection datasets lack the necessary generality for a comprehensive grasp of semantic action understanding. This deficiency arises because numerous objects integral to fine-grained actions may be absent from current extensive object detection or image classification datasets. Additionally, the challenges associated with imperfect object detection, particularly in the context of low-resolution videos, further compound the difficulty of precisely identifying objects within the video stream.

### 8.2.2 Action detection

The effectiveness of recent **action detection** approaches, as demonstrated in works such as [234, 235, 243], falls short of expectations when applied to widely used benchmarks. This is particularly true for datasets characterized by dense occurrences of actions, as seen in Charades and TSU [36, 164]. Despite the demonstrated

efficacy of the proposed temporal models in various temporal reasoning tasks [219, 88, 157], the ultimate action detection results remain suboptimal due to constraints posed by an unoptimized spatiotemporal visual encoder.

### Scene and features encoding

The first issue lies in the feature extraction step. Feature extractors such as 3D-CNNs [33, 65, 226] or video transformers such as [135, 8] follow a window approach for temporal feature extractions. These feature extractors or visual encoders are designed for pre-segmented videos where the whole video represents one complete action. The video snippet with the same label should be represented similarly. Nevertheless, in practice, the input to these visual encoders are non-overlapping small windows, not showing the whole action instance. Hence, we end up with many tiny pieces of information or part of the action that can be taken from anywhere. These incomplete snippets increase extremely the data diversity at inference time resulting in an over-fitting issue for the current models. Recently, masked auto-encoder [87], shows big promises as it enables visual encoders to learn strong semantics from randomly masked instances. This opens doors to future directions where we can limit such drawbacks of visual encoders.

Another limitation of the action detection benchmarks is the limited information on the spatial or semantic information from the video because the input to the temporal module is a one-dimensional representation where each time step corresponds to a single feature vector. In fact, the separation of the visual encoder and the temporal module introduces a challenge where the visual encoder fails to efficiently extract features for the ultimate task, disrupting the end-to-end training process. This dissociation hampers the optimal performance of visual encoding, which consequently limits the effectiveness of action detection. Presently, our networks lack joint optimization of the visual encoder and temporal module due to hardware constraints. To establish a connection between these components, a potential solution involves inserting a momentum memory bank [44, 244] between the visual encoder and the temporal module. This dynamic bridge facilitates gradual access to spatial information in the video by the temporal module, enabling end-to-end training of the visual encoder and temporal module. Note that this approach differs from the previous one, as it involves a dynamically updated memory bank instead of utilizing a frozen one from the extracted snippet feature.

Besides the previously mentioned challenges focusing mainly on semantics extractions and visual encoders, some other challenges remain the subject of future perspectives for our research.

### Towards unsupervised action detection

Firstly, the methods used and mentioned in this thesis are all fully-supervised methods. Such settings require complete annotation of all action instances (i.e., temporal boundaries and categories) in training videos. However, such a supervised learning strategy is very time-consuming and costly. To eliminate the need for exhaustive annotations in the training phase, limited supervision is required. Consequently, our goal is to use video-level labels to disambiguate a set of actions occurring in a video.

### Action detection for real-world scenarios

Secondly, one of the challenges faced in this thesis was action detection on densely annotated datasets. In scenarios where co-occurring actions happen, it becomes more challenging to handle temporal information. The main question that needs to be addressed is *which past information is still relevant?* To answer these challenges, [48] proposes to study action-action relations and model their dependencies, in contrast to our work [80] where we do object-object interaction modeling. Finally, another challenge is multi-subject actions, as most works focus on a subject-agnostic approach for action recognition. However, we believe that for a full understanding of scenes, a subject-subject relation modeling on top of object-object and action-action is necessary. All of these three components are complementary, and as future work, we want to build a unified framework for all three.



# Bibliography

- [1] Sami Abu-El-Haija et al. "YouTube-8M: A large-scale video classification benchmark". In: *arXiv preprint arXiv:1609.08675* (2016).
- [2] Dhruv Agarwal et al. "From Multimodal to Unimodal Attention in Transformers using Knowledge Distillation". In: *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2021, pp. 1–8. DOI: [10.1109/AVSS52988.2021.9663793](https://doi.org/10.1109/AVSS52988.2021.9663793).
- [3] Tanay Agrawal et al. "Multimodal Personality Recognition using Cross-attention Transformer and Behaviour Encoding". In: *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP*. INSTICC. SciTePress, 2022, pp. 501–508. ISBN: 978-989-758-555-5. DOI: [10.5220/0010841400003124](https://doi.org/10.5220/0010841400003124).
- [4] Tanay Agrawal et al. "Multimodal Vision Transformers With Forced Attention for Behavior Analysis". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 3392–3402.
- [5] Hassan Akbari et al. "VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text". In: *CoRR abs/2104.11178* (2021). arXiv: [2104.11178](https://arxiv.org/abs/2104.11178). URL: <https://arxiv.org/abs/2104.11178>.
- [6] Karteek Alahari. "Actor and Observer: Joint Modeling of First and Third-Person Videos". In: *Proceedings of the 1st Workshop and Challenge on Comprehensive Video Understanding in the Wild*. 2018, pp. 3–3.
- [7] Ahmed Amer et al. "Backchannel Detection and Agreement Estimation from Video with Transformer Networks". In: *Proc. of the IEEE International Joint Conference on Neural Networks*. 2023.
- [8] Anurag Arnab et al. "Vivit: A video vision transformer". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 6836–6846.
- [9] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. "SoundNet: Learning sound representations from unlabeled video". In: *Advances in Neural Information Processing Systems*. 2016, pp. 892–900.
- [10] Michal Balazia et al. "Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 70–79.
- [11] Michal Balazia et al. "Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 70–79.
- [12] Tadas Baltrusaitis et al. "Openface 2.0: Facial behavior analysis toolkit". In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE. 2018, pp. 59–66.
- [13] Fabien Baradel et al. "Object level visual reasoning in videos". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 105–121.



- [14] M Bardi et al. "Behavioral and physiological correlates of stress related to examination performance in college chemistry students". In: *Stress* 14.5 (2011), pp. 557–566.
- [15] Roger G Barker and Herbert F Wright. "Midwest and its children: The psychological ecology of an American town." In: (1955).
- [16] Roger G Barker and Herbert F Wright. "One boy's day; a specimen record of behavior." In: (1951).
- [17] Faisal I Bashir, Ashfaq A Khokhar, and Dan Schonfeld. "Object trajectory-based activity classification and recognition using hidden Markov models". In: *IEEE Transactions on Image Processing* 16.7 (2007), pp. 1912–1919.
- [18] Roman Bednarik, Shahram Eivazi, and Michal Hradis. "Gaze and Conversational Engagement in Multiparty Video Conversation: An Annotation Scheme and Classification of High and Low Levels of Engagement". In: *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction. Gaze-In '12*. Santa Monica, California: ACM, 2012, 10:1–10:6. ISBN: 978-1-4503-1516-6.
- [19] Roman Bednarik, Shahram Eivazi, and Michal Hradis. "Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement". In: *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction. Gaze-In '12*. New York, NY, USA: Association for Computing Machinery, Oct. 2012, pp. 1–6. ISBN: 978-1-4503-1516-6. DOI: [10 . 1145 / 2401836 . 2401846](https://doi.org/10.1145/2401836.2401846). URL: <https://doi.org/10.1145/2401836.2401846> (visited on 12/26/2022).
- [20] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. "Is space-time attention all you need for video understanding?" In: *ICML*. Vol. 2. 3. 2021, p. 4.
- [21] Cigdem Beyan et al. "Analysis of face-touching behavior in large scale social interaction dataset". In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. 2020, pp. 24–32.
- [22] Chris Birmingham, Kalin Stefanov, and Maja J Mataric. "Group-Level Focus of Visual Attention for Improved Next Speaker Prediction". In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 4838–4842.
- [23] Chris Birmingham et al. "Can I Trust You? A User Study of Robot Mediation of a Support Group". In: *arXiv preprint arXiv:2002.04671* (2020).
- [24] Aaron F. Bobick and James W. Davis. "The recognition of human movement using temporal templates". In: *IEEE Transactions on pattern analysis and machine intelligence* 23.3 (2001), pp. 257–267.
- [25] Dan Bohus and Eric Horvitz. "Facilitating multiparty dialog with gaze, gesture, and speech". In: *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. 2010, pp. 1–8.
- [26] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- [27] Angelo Cafaro et al. "The NoXi Database: Multimodal Recordings of Mediated Novice-Expert Interactions". In: *Proceedings of the 19th International Conference on Multimodal Interaction*. ACM, In press. 2017.

- [28] Yingfeng Cai et al. "Trajectory-based anomalous behaviour detection for intelligent traffic surveillance". In: *IET Intelligent Transport Systems* 9.8 (2015), pp. 810–816.
- [29] Zhe Cao et al. "Realtime multi-person 2d pose estimation using part affinity fields". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7291–7299.
- [30] Nicolas Carion et al. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pp. 213–229.
- [31] Dana R. Carney Dana. "Beliefs about the nonverbal expression of social power." In: *Journal of nonverbal behavior*. 29.2 (2005-06-01). ISSN: 0191-5886.
- [32] Mathilde Caron et al. "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.
- [33] Joao Carreira and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.
- [34] Joao Carreira and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.
- [35] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. "Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement". In: *IEEE Transactions on Affective Computing* 10.4 (Oct. 2019). Conference Name: IEEE Transactions on Affective Computing, pp. 484–497. ISSN: 1949-3045. DOI: [10.1109/TAFFC.2017.2737019](https://doi.org/10.1109/TAFFC.2017.2737019).
- [36] Guang Chen, Can Zhang, and Yuexian Zou. "Afnet: Temporal locality-aware network with dual structure for accurate and fast action detection". In: *IEEE Transactions on Multimedia* 23 (2020), pp. 2672–2682.
- [37] Hanqing Chen et al. "Learning Student Networks in the Wild". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 6428–6437.
- [38] Junwen Chen et al. "Gatehub: Gated history unit with background suppression for online action detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 19925–19934.
- [39] Qiang Chen, Shuicheng Yan, et al. "Network in network". In: *International Conference on Learning Representations (ICLR)*. 2014.
- [40] Shoufa Chen et al. "Adaptformer: Adapting vision transformers for scalable visual recognition". In: *arXiv preprint arXiv:2205.13535* (2022).
- [41] Xinlei Chen\*, Saining Xie\*, and Kaiming He. "An Empirical Study of Training Self-Supervised Vision Transformers". In: *arXiv preprint arXiv:2104.02057* (2021).
- [42] Yuhua Chen et al. "Domain adaptive faster r-cnn for object detection in the wild". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3339–3348.
- [43] Bowen Cheng, Alex Schwing, and Alexander Kirillov. "Per-pixel classification is not all you need for semantic segmentation". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17864–17875.

- [44] Feng Cheng and Gedas Bertasius. “Tallformer: Temporal action localization with a long-memory transformer”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 503–521.
- [45] Krzysztof Choromanski et al. “Rethinking Attention with Performers”. In: *CoRR abs/2009.14794* (2020). arXiv: 2009.14794. URL: <https://arxiv.org/abs/2009.14794>.
- [46] MMAAction2 Contributors. *OpenMMLab’s Next Generation Video Understanding Toolbox and Benchmark*. <https://github.com/open-mmlab/mmaaction2>. 2020.
- [47] Nieves Crasto et al. “Mars: Motion-augmented rgb stream for action recognition”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 7882–7891.
- [48] Rui Dai, Srijan Das, and Francois Bremond. “CTRN: Class Temporal Relational Network For Action Detection”. In: *The British Machine Vision Conference*. Virtual, United Kingdom, Nov. 2021.
- [49] Rui Dai et al. “MS-TCT: multi-scale temporal convtransformer for action detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20041–20051.
- [50] Rui Dai et al. “Pdan: Pyramid dilated attention network for action detection”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 2970–2979.
- [51] Dima Damen et al. “Rescaling egocentric vision”. In: *arXiv preprint arXiv:2006.13256* (2020).
- [52] Dima Damen et al. “Scaling egocentric vision: The epic-kitchens dataset”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 720–736.
- [53] Dima Damen et al. “Scaling egocentric vision: The epic-kitchens dataset”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 720–736.
- [54] Dima Damen et al. “Scaling egocentric vision: The epic-kitchens dataset”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 720–736.
- [55] Roeland De Geest et al. “Online action detection”. In: *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*. Springer. 2016, pp. 269–284.
- [56] Jean-Benoit Delbrouck et al. “A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis”. In: *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*. Seattle, USA: Association for Computational Linguistics, July 2020, pp. 1–7. DOI: 10.18653/v1/2020.challengehml-1.1. URL: <https://www.aclweb.org/anthology/2020.challengehml-1.1>.
- [57] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [58] Jeffrey Donahue et al. “Long-term recurrent convolutional networks for visual recognition and description”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 2625–2634.

- [59] Shikha Dubey, Abhijeet Boragule, and Moongu Jeon. "3D ResNet with Ranking Loss Function for Abnormal Activity Detection in Videos". In: *arXiv preprint arXiv:2002.01132* (2020).
- [60] Olov Engwall and José Lopes. "Interaction and collaboration in robot-assisted language learning for adults". In: *Computer Assisted Language Learning* (2020), pp. 1–37.
- [61] Hyunjun Eun et al. "Learning to discriminate information for online action detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 809–818.
- [62] Florian Eyben et al. "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing". In: *IEEE Transactions on Affective Computing* 7.2 (2015), pp. 190–202. DOI: [10.1109/TAFFC.2015.2457417](https://doi.org/10.1109/TAFFC.2015.2457417).
- [63] Yazan Abu Farha and Jurgen Gall. "Ms-tcn: Multi-stage temporal convolutional network for action segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3575–3584.
- [64] Alireza Fathi, Yin Li, and James M Rehg. "Learning to recognize daily actions using gaze". In: *European Conference on Computer Vision*. Springer. 2012, pp. 314–327.
- [65] Christoph Feichtenhofer. "X3d: Expanding architectures for efficient video recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 203–213.
- [66] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. "Spatiotemporal multiplier networks for video action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4768–4777.
- [67] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1933–1941.
- [68] Christoph Feichtenhofer et al. "Slowfast networks for video recognition". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6202–6211.
- [69] Eugene Yujun Fu and Michael W Ngai. "Using Motion Histories for Eye Contact Detection in Multiperson Group Conversations". In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 4873–4877. DOI: [10.1145/3474085.3479230](https://doi.org/10.1145/3474085.3479230).
- [70] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. "Red: Reinforced encoder-decoder networks for action anticipation". In: *arXiv preprint arXiv:1707.04818* (2017).
- [71] Jiyang Gao et al. "Turn tap: Temporal unit regression network for temporal action proposals". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3628–3636.
- [72] Mingfei Gao et al. "Startnet: Online detection of action start in untrimmed videos". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 5542–5551.
- [73] Mingfei Gao et al. "WOAD: Weakly supervised online action detection in untrimmed videos". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1915–1923.

- [74] Pallabi Ghosh et al. "Stacked spatio-temporal graph convolutional networks for action segmentation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 576–585.
- [75] Rohit Girdhar and Kristen Grauman. "Anticipative video transformer". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 13505–13515.
- [76] Patricia Goldberg et al. "Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction". en. In: *Educational Psychology Review* 33.1 (Mar. 2021), pp. 27–49. ISSN: 1573-336X. DOI: [10.1007/s10648-019-09514-z](https://doi.org/10.1007/s10648-019-09514-z). URL: <https://doi.org/10.1007/s10648-019-09514-z> (visited on 01/31/2023).
- [77] Dayoung Gong et al. "Future Transformer for Long-term Action Anticipation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 3052–3061.
- [78] Mononito Goswami, Minkush Manuja, and Maitree Leekha. "Towards Social & Engaging Peer Learning: Predicting Backchanneling and Disengagement in Children". In: *arXiv:2007.11346 [cs]* (July 2020). arXiv: 2007.11346. URL: <http://arxiv.org/abs/2007.11346> (visited on 01/29/2022).
- [79] Yuxian Gu et al. "Ppt: Pre-trained prompt tuning for few-shot learning". In: *arXiv preprint arXiv:2109.04332* (2021).
- [80] Mohammed Guermal, Rui Dai, and François Brémond. "THORN: Temporal Human-Object Relation Network for Action Recognition". In: *arXiv preprint arXiv:2204.09468* (2022).
- [81] Pooja Guhan et al. *Developing an Effective and Automated Patient Engagement Estimator for Telehealth: A Machine Learning Approach*. arXiv:2011.08690 [cs]. Aug. 2022. URL: <http://arxiv.org/abs/2011.08690> (visited on 12/23/2022).
- [82] Judith Hall, Erik Coats, and Lavonia LeBeau. "Nonverbal Behavior and the Vertical Dimension of Social Relations: A Meta-Analysis." In: *Psychological bulletin* 131 (Dec. 2005), pp. 898–924. DOI: [10.1037/0033-2909.131.6.898](https://doi.org/10.1037/0033-2909.131.6.898).
- [83] Shijie Hao, Yuan Zhou, and Yanrong Guo. "A brief survey on semantic segmentation with deep learning". In: *Neurocomputing* 406 (2020), pp. 302–321.
- [84] Mahmudul Hasan et al. "Learning Temporal Regularity in Video Sequences". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
- [85] Junxian He et al. "Towards a unified view of parameter-efficient transfer learning". In: *arXiv preprint arXiv:2110.04366* (2021).
- [86] Kaiming He, Ross Girshick, and Piotr Dollar. "Rethinking ImageNet Pre-Training". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.
- [87] Kaiming He et al. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.
- [88] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. "Learning with side information through modality hallucination". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 826–834.
- [89] Rui Hou, Chen Chen, and Mubarak Shah. "Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos". In: *CoRR* abs/1703.10664 (2017). arXiv: [1703.10664](https://arxiv.org/abs/1703.10664). URL: <http://arxiv.org/abs/1703.10664>.

- [90] Neil Houlsby et al. "Parameter-Efficient Transfer Learning for NLP". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 2790–2799. URL: <https://proceedings.mlr.press/v97/houlsby19a.html>.
- [91] Michal Hradis, Shahram Eivazi, and Roman Bednarik. "Voice activity detection from gaze in video mediated communication". en. In: *Proceedings of the Symposium on Eye Tracking Research and Applications*. Santa Barbara California: ACM, Mar. 2012, pp. 329–332. ISBN: 978-1-4503-1221-9. DOI: [10.1145/2168556.2168628](https://doi.org/10.1145/2168556.2168628). URL: <https://dl.acm.org/doi/10.1145/2168556.2168628> (visited on 12/26/2022).
- [92] Edward Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: [2106.09685](https://arxiv.org/abs/2106.09685) [cs.CL].
- [93] Edward J Hu et al. "Lora: Low-rank adaptation of large language models". In: *arXiv preprint arXiv:2106.09685* (2021).
- [94] Weiming Hu et al. "Traffic accident prediction using 3-D model-based vehicle tracking". In: *IEEE Transactions on Vehicular Technology* 53.3 (2004), pp. 677–694.
- [95] Yifei Huang, Yusuke Sugano, and Yoichi Sato. "Improving action segmentation via graph-based temporal reasoning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 14024–14034.
- [96] Haroon Idrees et al. "The THUMOS challenge on action recognition for videos "in the wild"". In: *Computer Vision and Image Understanding* 155 (2017), pp. 1–23.
- [97] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.
- [98] Md Mofijul Islam and Tariq Iqbal. "Mumu: Cooperative multitask learning-based guided multimodal fusion". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 1. 2022, pp. 1043–1051.
- [99] Shomik Jain et al. "Modeling Engagement in Long-Term, In-Home Socially Assistive Robot Interventions for Children with Autism Spectrum Disorders". In: *Science Robotics* 5.39 (Feb. 2020). arXiv:2002.02453 [cs], eaaz3791. ISSN: 2470-9476. DOI: [10.1126/scirobotics.aaz3791](https://doi.org/10.1126/scirobotics.aaz3791). URL: <http://arxiv.org/abs/2002.02453> (visited on 01/31/2023).
- [100] Jingwei Ji et al. "Action genome: Actions as compositions of spatio-temporal scene graphs". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10236–10247.
- [101] Shuiwang Ji et al. "3D convolutional neural networks for human action recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012), pp. 221–231.
- [102] Menglin Jia et al. "Visual prompt tuning". In: *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*. Springer. 2022, pp. 709–727.
- [103] Xinghao Jiang, Ke Xu, and Tanfeng Sun. "Action recognition scheme based on skeleton representation with DS-LSTM network". In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.7 (2019), pp. 2129–2140.

- [104] Vasileios Karavasili, Konstantinos Blekas, and Christophoros Nikou. "A novel framework for motion segmentation and tracking by clustering incomplete trajectories". In: *Computer Vision and Image Understanding* 116.11 (2012), pp. 1135–1148.
- [105] Shofiyati Nur Karimah and Shinobu Hasegawa. "Automatic Engagement Recognition for Distance Learning Systems: A Literature Study of Engagement Datasets and Methods". en. In: *Augmented Cognition*. Ed. by Dylan D. Schmorrow and Cali M. Fidopiastis. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 264–276. ISBN: 978-3-030-78114-9. DOI: [10.1007/978-3-030-78114-9\\_19](https://doi.org/10.1007/978-3-030-78114-9_19).
- [106] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. "Com-pacter: Efficient low-rank hypercomplex adapter layers". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 1022–1035.
- [107] Rabeeh Karimi Mahabadi et al. "Parameter-efficient Multi-task Fine-tuning for Transformers via Shared Hypernetworks". In: *Annual Meeting of the Association for Computational Linguistics*. 2021.
- [108] Andrej Karpathy et al. "Large-scale video classification with convolutional neural networks". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014, pp. 1725–1732.
- [109] Andrej Karpathy et al. "Large-scale video classification with convolutional neural networks". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014, pp. 1725–1732.
- [110] Will Kay et al. "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950* (2017).
- [111] Will Kay et al. "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950* (2017).
- [112] Evangelos Kazakos et al. "Epic-fusion: Audio-visual temporal binding for egocentric action recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 5492–5501.
- [113] Yan Ke, Rahul Sukthankar, and Martial Hebert. "Spatio-temporal shape and flow correlation for action recognition". In: *2007 IEEE conference on computer vision and pattern recognition*. IEEE. 2007, pp. 1–8.
- [114] Salman Khan et al. "Transformers in vision: A survey". In: *ACM computing surveys (CSUR)* 54.10s (2022), pp. 1–41.
- [115] Young Hwi Kim, Seonghyeon Nam, and Seon Joo Kim. "Temporally smooth online action detection using cycle-consistent future anticipation". In: *Pattern Recognition* 116 (2021), p. 107954.
- [116] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).
- [117] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016).
- [118] Alexander Kolesnikov et al. "Large Scale Learning of General Visual Representations for Transfer". In: *CoRR abs/1912.11370* (2019). arXiv: [1912.11370](https://arxiv.org/abs/1912.11370). URL: <http://arxiv.org/abs/1912.11370>.

- [119] Zhi Kou et al. “Stochastic Normalization”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 16304–16314. URL: <https://proceedings.neurips.cc/paper/2020/file/bc573864331a9e42e4511de6f678aa83-Paper.pdf>.
- [120] Hildegard Kuehne et al. “HMDB: a large video database for human motion recognition”. In: *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.
- [121] Ayush Kumar and Jithendra Vepa. “Gated Mechanism for Attention Based Multimodal Sentiment Analysis”. In: *CoRR abs/2003.01043* (2020). arXiv: 2003.01043. URL: <https://arxiv.org/abs/2003.01043>.
- [122] Colin Lea et al. “Temporal convolutional networks for action segmentation and detection”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 156–165.
- [123] Brian Lester, Rami Al-Rfou, and Noah Constant. “The Power of Scale for Parameter-Efficient Prompt Tuning”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3045–3059. DOI: 10.18653/v1/2021.emnlp-main.243. URL: <https://aclanthology.org/2021.emnlp-main.243>.
- [124] Fu Li et al. “Temporal modeling approaches for large-scale youtube-8m video understanding”. In: *arXiv preprint arXiv:1707.04555* (2017).
- [125] Xiang Lisa Li and Percy Liang. “Prefix-tuning: Optimizing continuous prompts for generation”. In: *arXiv preprint arXiv:2101.00190* (2021).
- [126] Xiang Lisa Li and Percy Liang. “Prefix-Tuning: Optimizing Continuous Prompts for Generation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 4582–4597. DOI: 10.18653/v1/2021.acl-long.353. URL: <https://aclanthology.org/2021.acl-long.353>.
- [127] Yin Li, Miao Liu, and James M Rehg. “In the eye of beholder: Joint learning of gaze and actions in first person video”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 619–635.
- [128] Lawrence I-Kuei Lin. “A Concordance Correlation Coefficient to Evaluate Reproducibility”. In: *Biometrics* 45.1 (1989), pp. 255–268. ISSN: 0006341X, 15410420. URL: <http://www.jstor.org/stable/2532051> (visited on 02/14/2023).
- [129] Tianwei Lin et al. “Bmn: Boundary-matching network for temporal action proposal generation”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3889–3898.
- [130] Tianwei Lin et al. “Bsn: Boundary sensitive network for temporal action proposal generation”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.
- [131] Xin Liu et al. “iMiGUE: An identity-free video dataset for micro-gesture understanding and emotion analysis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 10631–10642.
- [132] Yuan Liu et al. “Multi-granularity generator for temporal action proposal”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3604–3613.



- [133] Ze Liu et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *arXiv preprint arXiv:2103.14030* (2021).
- [134] Ze Liu et al. "Video swin transformer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 3202–3211.
- [135] Ze Liu et al. "Video Swin Transformer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 3202–3211.
- [136] José Lopes, Olov Engwall, and Gabriel Skantze. "A first visit to the robot language café". In: *ISCA workshop on Speech and Language Technology in Education*. 2017.
- [137] Cewu Lu, Jianping Shi, and Jiaya Jia. "Abnormal event detection at 150 FPS in matlab". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 2720–2727.
- [138] Fuyan Ma et al. "TA-CNN: A Unified Network for Human Behavior Analysis in Multi-Person Conversations". In: *Proc. of the ACM International Conference on Multimedia*. 2022, pp. 7099–7103.
- [139] Rabeeh Karimi Mahabadi et al. "Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks". In: *arXiv preprint arXiv:2106.04489* (2021).
- [140] Yuning Mao et al. "Unipelt: A unified framework for parameter-efficient language model tuning". In: *arXiv preprint arXiv:2110.07577* (2021).
- [141] Imad Eddine MAROUF, Enzo Tartaglione, and Stéphane Lathuilière. *Tiny Adapters for Vision Transformers*. 2023. URL: <https://openreview.net/forum?id=V0Vo9eW2nzL>.
- [142] Albert Mehrabian. "Relationship of attitude to seated posture, orientation, and distance." In: *Journal of personality and social psychology* 10.1 (1968), p. 26.
- [143] Albert Mehrabian and John T Friar. "Encoding of attitude by a seated communicator via posture and position cues." In: *Journal of Consulting and Clinical Psychology* 33.3 (1969), p. 330.
- [144] Changiz Mohiyeddini, Stephanie Bauer, and Stuart Semple. "Displacement behaviour is associated with reduced stress levels among men but not women". In: *PloS one* 8.2 (2013), e56355.
- [145] Changiz Mohiyeddini, Stephanie Bauer, and Stuart Semple. "Neuroticism and stress: The role of displacement behavior". In: *Anxiety, stress, & coping* 28.4 (2015), pp. 391–407.
- [146] Philipp Müller and Andreas Bulling. "Emergent leadership detection across datasets". In: *2019 International Conference on Multimodal Interaction*. 2019, pp. 274–278.
- [147] Philipp Müller et al. "MultiMediate: Multi-modal Group Behaviour Analysis for Artificial Mediation". In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 4878–4882.
- [148] Philipp Müller et al. "MultiMediate'22: Backchannel Detection and Agreement Estimation in Group Interactions". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 7109–7114.
- [149] Philipp Müller et al. "Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour". In: *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. 2018, pp. 1–10.

- [150] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. “Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour”. In: *23rd International Conference on Intelligent User Interfaces. IUI '18*. Tokyo, Japan: Association for Computing Machinery, Mar. 2018, pp. 153–164. ISBN: 978-1-4503-4945-1. DOI: [10.1145/3172944.3172969](https://doi.org/10.1145/3172944.3172969). URL: <https://doi.org/10.1145/3172944.3172969>.
- [151] Tam V Nguyen and Bilal Mirza. “Dual-layer kernel extreme learning machine for action recognition”. In: *Neurocomputing* 260 (2017), pp. 123–130.
- [152] Enrique Bermejo Nievas et al. “Violence detection in video using computer vision techniques”. In: *International Conference on Computer Analysis of Images and Patterns*. Springer. 2011, pp. 332–339.
- [153] Catharine Oertel and Giampiero Salvi. “A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue”. en. In: *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*. ACM Press, 2013, pp. 99–106. ISBN: 978-1-4503-2129-7. DOI: [10.1145/2522848.2522865](https://doi.org/10.1145/2522848.2522865). URL: <http://dl.acm.org/citation.cfm?doid=2522848.2522865> (visited on 01/10/2021).
- [154] Catharine Oertel et al. “Engagement in Human-Agent Interaction: An Overview”. In: *Frontiers in Robotics and AI* 7 (2020). ISSN: 2296-9144. URL: <https://www.frontiersin.org/articles/10.3389/frobt.2020.00092> (visited on 01/31/2023).
- [155] N. Ohshima et al. “Neut: Design and evaluation of speaker designation behaviors for communication support robot to encourage conversations”. In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Aug. 2017, pp. 1387–1393. DOI: [10.1109/ROMAN.2017.8172485](https://doi.org/10.1109/ROMAN.2017.8172485).
- [156] Tom O'Malley et al. *KerasTuner*. <https://github.com/keras-team/keras-tuner>. 2019.
- [157] Aaron van den Oord et al. “Wavenet: A generative model for raw audio”. In: *arXiv preprint arXiv:1609.03499* (2016).
- [158] Halil İbrahim Öztürk and Ahmet Burak Can. “ADNet: Temporal Anomaly Detection in Surveillance Videos”. In: *arXiv preprint arXiv:2104.06653* (2021).
- [159] Cristina Palmero et al. “Context-Aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA Dataset”. In: *2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*. 2021, pp. 1–12.
- [160] Cristina Palmero et al. “Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 1–12.
- [161] Hae Won Park et al. “A Model-Free Affective Reinforcement Learning Approach to Personalization of an Autonomous Social Robot Companion for Early Literacy Education”. en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 687–694. ISSN: 2374-3468, 2159-5399. DOI: [10.1609/aaai.v33i01.3301687](https://doi.org/10.1609/aaai.v33i01.3301687). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/3846> (visited on 12/26/2022).
- [162] Sunjeong Park and Youn-kyung Lim. “Investigating User Expectations on the Roles of Family-shared AI Speakers”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–13.

- [163] Christopher Peters et al. "Engagement Capabilities for ECAs". In: *Autonomous Agents and Multi-agent Systems - AAMAS* (Jan. 2005).
- [164] AJ Piergiovanni and Michael Ryoo. "Temporal gaussian mixture layer for videos". In: *International Conference on Machine Learning*. PMLR, 2019, pp. 5152–5161.
- [165] Automatic Differentiation In Pytorch. *Pytorch*. 2018.
- [166] Sanqing Qu et al. "LAP-Net: Adaptive features sampling via learning action progression for online action detection". In: *arXiv preprint arXiv:2011.07915* (2020).
- [167] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763. URL: <https://proceedings.mlr.press/v139/radford21a.html>.
- [168] Shyam Sundar Rajagopalan et al. "Play with me — Measuring a child's engagement in a social interaction". In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 1. May 2015, pp. 1–8. DOI: [10.1109/FG.2015.7163129](https://doi.org/10.1109/FG.2015.7163129).
- [169] Bharathkumar Ramachandra and Michael Jones. "Street Scene: A new dataset and evaluation protocol for video anomaly detection". In: *The IEEE Winter Conference on Applications of Computer Vision*. 2020, pp. 2569–2578.
- [170] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. "Efficient Parametrization of Multi-Domain Deep Neural Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.
- [171] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. "Learning multiple visual domains with residual adapters". In: *Advances in neural information processing systems* 30 (2017).
- [172] C. Rich et al. "Recognizing engagement in human-robot interaction". In: *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2010, pp. 375–382.
- [173] Fabien Ringeval et al. "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Apr. 2013, pp. 1–8. DOI: [10.1109/FG.2013.6553805](https://doi.org/10.1109/FG.2013.6553805).
- [174] Michael S Ryoo et al. "Assemblenet: Searching for multi-stream neural connectivity in video architectures". In: *arXiv preprint arXiv:1905.13209* (2019).
- [175] Jyotirmay Sanghvi et al. "Automatic Analysis of Affective Postures and Body Motion to Detect Engagement with a Game Companion". In: *Proceedings of the 6th International Conference on Human-robot Interaction*. HRI '11. Lausanne, Switzerland: ACM, 2011, pp. 305–312. ISBN: 978-1-4503-0561-7.
- [176] Ipek Baris Schlicht, Lucie Flek, and Paolo Rosso. "Multilingual Detection of Check-Worthy Claims using World Languages and Adapter Fusion". In: *arXiv preprint arXiv:2301.05494* (2023).
- [177] Bernhard Schölkopf et al. "Estimating the support of a high-dimensional distribution". In: *Neural Computation* 13.7 (2001), pp. 1443–1471.

- [178] Sarah Sebo et al. "Robots in groups and teams: a literature review". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020), pp. 1–36.
- [179] Garima Sharma et al. "Graph-based Group Modelling for Backchannel Detection". In: *Proc. of the ACM International Conference on Multimedia*. 2022, pp. 7190–7194.
- [180] Christopher F. Sharpley and Anastasia Sagris. "When does counsellor forward lean influence client-perceived rapport?" In: *British Journal of Guidance & Counselling* 23.3 (1995), pp. 387–394. DOI: [10.1080/03069889508253696](https://doi.org/10.1080/03069889508253696). eprint: <https://doi.org/10.1080/03069889508253696>. URL: <https://doi.org/10.1080/03069889508253696>.
- [181] Aman Shenoy and Ashish Sardana. "Multilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation". In: *arXiv preprint arXiv:2002.08267* (2020).
- [182] Aman Shenoy and Ashish Sardana. "Multilogue-Net: A Context-Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation". In: *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)* (2020). DOI: [10.18653/v1/2020.challengehml-1.3](https://doi.org/10.18653/v1/2020.challengehml-1.3). URL: <http://dx.doi.org/10.18653/v1/2020.challengehml-1.3>.
- [183] Lei Shi et al. "Two-stream adaptive graph convolutional networks for skeleton-based action recognition". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 12026–12035.
- [184] B Shiva Prakash et al. "A survey on recurrent neural network architectures for sequential learning". In: *Soft Computing for Problem Solving: SocProS 2017, Volume 2*. Springer. 2019, pp. 57–66.
- [185] Joel Shor et al. "Towards Learning a Universal Non-Semantic Representation of Speech". In: *Interspeech 2020* (2020). DOI: [10.21437/interspeech.2020-1242](https://doi.org/10.21437/interspeech.2020-1242). URL: <http://dx.doi.org/10.21437/Interspeech.2020-1242>.
- [186] Elaine Short and Maja J. Mataric. "Robot moderation of a collaborative game: Towards socially assistive robotics in group interactions". In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2017, pp. 385–390.
- [187] Zheng Shou, Dongang Wang, and Shih-Fu Chang. "Temporal action localization in untrimmed videos via multi-stage cnns". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1049–1058.
- [188] Zheng Shou, Dongang Wang, and Shih-Fu Chang. "Temporal action localization in untrimmed videos via multi-stage cnns". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1049–1058.
- [189] Zheng Shou et al. "Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5734–5743.
- [190] Zheng Shou et al. "Online action detection in untrimmed, streaming videos-modeling and evaluation". In: *ECCV*. Vol. 1. 2. 2018, p. 5.
- [191] Guang Shu et al. "Violent behavior detection based on SVM in the elevator". In: *International Journal of Security and Its Applications* 8.5 (2014), pp. 31–40.
- [192] Yang Shu et al. "Zoo-Tuning: Adaptive Transfer from a Zoo of Models". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 9626–9637.

- [193] Tomas Simon et al. “Hand Keypoint Detection in Single Images using Multi-view Bootstrapping”. In: *CVPR*. 2017.
- [194] Karen Simonyan and Andrew Zisserman. “Two-stream convolutional networks for action recognition in videos”. In: *arXiv preprint arXiv:1406.2199* (2014).
- [195] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. “UCF101: A dataset of 101 human actions classes from videos in the wild”. In: *arXiv preprint arXiv:1212.0402* (2012).
- [196] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. “UCF101: A dataset of 101 human actions classes from videos in the wild”. In: *arXiv preprint arXiv:1212.0402* (2012).
- [197] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. “Unsupervised learning of video representations using lstms”. In: *International conference on machine learning*. PMLR. 2015, pp. 843–852.
- [198] Ralf C Staudemeyer and Eric Rothstein Morris. “Understanding LSTM—a tutorial into long short-term memory recurrent neural networks”. In: *arXiv preprint arXiv:1909.09586* (2019).
- [199] Erik E Stone and Marjorie Skubic. “Fall detection in homes of older adults using the Microsoft Kinect”. In: *IEEE Journal of Biomedical and Health Informatics* 19.1 (2014), pp. 290–301.
- [200] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. “Lsta: Long short-term attention for egocentric action recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9954–9963.
- [201] Swathikiran Sudhakaran and Oswald Lanz. “Attention is all we need: Nailing down object-centric attention for egocentric activity recognition”. In: *arXiv preprint arXiv:1807.11794* (2018).
- [202] Waqas Sultani, Chen Chen, and Mubarak Shah. “Real-world anomaly detection in surveillance videos”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6479–6488.
- [203] Chen Sun et al. “Temporal localization of fine-grained actions in videos by domain transfer from web images”. In: *Proceedings of the 23rd ACM international conference on Multimedia*. 2015, pp. 371–380.
- [204] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. “VI-adapter: Parameter-efficient transfer learning for vision-and-language tasks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5227–5237.
- [205] Yi-Lin Sung, Varun Nair, and Colin A Raffel. “Training neural networks with fixed sparse masks”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 24193–24205.
- [206] Teppei Suzuki. “TeachAugment: Data Augmentation Optimization Using Teacher Knowledge”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10904–10914.
- [207] Jing Tan et al. “Relaxed transformer decoders for direct action proposal generation”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 13526–13535.

- [208] Rafael Martínez Tomás et al. "Identification of loitering human behaviour in video surveillance environments". In: *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer. 2015, pp. 516–525.
- [209] Zhan Tong et al. "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training". In: *Advances in Neural Information Processing Systems*. 2022.
- [210] Zhan Tong et al. "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training". In: *arXiv preprint arXiv:2203.12602* (2022).
- [211] Amirsina Torfi et al. "Natural language processing advancements by deep learning: A survey". In: *arXiv preprint arXiv:2003.01200* (2020).
- [212] Du Tran et al. "Learning Spatiotemporal Features With 3D Convolutional Networks". In: *IEEE International Conference on Computer Vision (ICCV)*. 2015.
- [213] Dina Utami and Timothy Bickmore. "Collaborative user responses in multiparty interaction with a couples counselor robot". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 294–303.
- [214] Tanya Vacharkulksemsuk et al. "Dominant, open nonverbal displays are attractive at zero-acquaintance". In: *Proceedings of the National Academy of Sciences* 113.15 (2016), pp. 4009–4014.
- [215] Michel Valstar et al. "Avec 2016: Depression, mood, and emotion recognition workshop and challenge". In: *Proc. of the International Workshop on Audio/Visual Emotion Challenge*. 2016, pp. 3–10. DOI: [10.1145/2988257.2988258](https://doi.org/10.1145/2988257.2988258).
- [216] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.
- [217] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).
- [218] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).
- [219] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).
- [220] Johannes Wagner et al. "Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant?" In: *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*. Ed. by B. Yegnanarayana. ISCA, 2018, pp. 147–151. DOI: [10.21437/Interspeech.2018-1238](https://doi.org/10.21437/Interspeech.2018-1238). URL: <https://doi.org/10.21437/Interspeech.2018-1238>.
- [221] Harald G Wallbott. "Bodily expression of emotion". In: *European journal of social psychology*. 28.6 (1998). ISSN: 0046-2772.
- [222] Boyang Wan et al. "Anomaly detection in video sequences: A benchmark and computational model". In: *IET Image Processing* (2021).
- [223] Jue Wang and Anoop Cherian. "GODS: Generalized One-class Discriminative Subspaces for Anomaly Detection". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 8201–8211.

- [224] Lei Wang and Piotr Koniusz. "Self-supervising action recognition by statistical moment and subspace descriptors". In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 4324–4333.
- [225] Limin Wang et al. "Temporal segment networks: Towards good practices for deep action recognition". In: *European conference on computer vision*. Springer. 2016, pp. 20–36.
- [226] X Wang et al. "Non-local neural networks In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition". In: (2018).
- [227] Xiang Wang et al. "Oadtr: Online action detection with transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 7565–7575.
- [228] Xiaohan Wang et al. "Symbiotic attention with privileged information for egocentric action recognition". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 12249–12256.
- [229] Xiaolong Wang and Abhinav Gupta. "Videos as space-time region graphs". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 399–417.
- [230] Chao-Yuan Wu et al. "Long-term feature banks for detailed video understanding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 284–293.
- [231] Xuehan Xiong et al. "M&m mix: A multimodal multiview transformer ensemble". In: *arXiv preprint arXiv:2206.09852* (2022).
- [232] Huijuan Xu, Abir Das, and Kate Saenko. "R-c3d: Region convolutional 3d network for temporal activity detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5783–5792.
- [233] Huijuan Xu, Abir Das, and Kate Saenko. "R-C3D: Region convolutional 3D network for temporal activity detection". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5783–5792.
- [234] Huijuan Xu, Abir Das, and Kate Saenko. "R-c3d: Region convolutional 3d network for temporal activity detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5783–5792.
- [235] Mengmeng Xu et al. "G-tad: Sub-graph localization for temporal action detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10156–10165.
- [236] Mingze Xu et al. "Long short-term transformer for online action detection". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 1086–1099.
- [237] Mingze Xu et al. "Temporal recurrent networks for online action detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 5532–5541.
- [238] Sijie Yan, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition". In: *Thirty-second AAAI conference on artificial intelligence*. 2018.
- [239] Serena Yeung et al. "Every moment counts: Dense detailed labeling of actions in complex videos". In: *International Journal of Computer Vision* 126.2 (2018), pp. 375–389.

- [240] Kaichao You et al. “Co-Tuning for Transfer Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 17236–17246. URL: <https://proceedings.neurips.cc/paper/2020/file/c8067ad1937f728f51288b3eb986afaa-Paper.pdf>.
- [241] Joe Yue-Hei Ng et al. “Beyond short snippets: Deep networks for video classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4694–4702.
- [242] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. “Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models”. In: *arXiv preprint arXiv:2106.10199* (2021).
- [243] Runhao Zeng et al. “Graph convolutional networks for temporal action localization”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 7094–7103.
- [244] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. “Temporal query networks for fine-grained video understanding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 4486–4496.
- [245] Jiangong Zhang, Laiyun Qing, and Jun Miao. “Temporal Convolutional Network with Complementary Inner Bag Loss for Weakly Supervised Anomaly Detection”. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 4030–4034.
- [246] Jingjing Zhang and Ping Zhong. “Least Squares One-class Support Vector Machine on Fuzzy Set”. In: *International Journal of Control and Automation* 9.12 (2016), pp. 249–260.
- [247] Jingran Zhang et al. “Temporal reasoning graph for activity recognition”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 5491–5506.
- [248] Peisen Zhao et al. “Privileged knowledge distillation for online action detection”. In: *arXiv preprint arXiv:2011.09158* (2020).
- [249] Yue Zhao and Philipp Krähenbühl. “Real-Time Online Video Detection with Temporal Smoothing Transformers”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 485–502.
- [250] Yue Zhao et al. “Temporal action detection with structured segment networks”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2914–2923.
- [251] Yue Zhao et al. “Temporal action detection with structured segment networks”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2914–2923.
- [252] Jia-Xing Zhong et al. “Graph Convolutional Label Noise Cleaner: Train a Plug-And-Play Action Classifier for Anomaly Detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.
- [253] Yi Zhu and Shawn Newsam. “Motion-Aware Feature for Improved Video Anomaly Detection”. In: *arXiv preprint arXiv:1907.10211* (2019).
- [254] Yichen Zhu and Yi Wang. “Student Customized Knowledge Distillation: Bridging the Gap Between Student and Teacher”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 5057–5066.



- [255] Chengqing Zong et al. "Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021.