

博 士 論 文

Human Activity Understanding by Multilateral Relation Mining (多面的関係性マイニングによる 人物行動理解)

48-207423

楊 麗錦

指導教員 佐藤 洋一 教授

東京大学大学院 情報理工学系研究科 電子情報学専攻

This dissertation is submitted for the degree of
Doctor of Philosophy

December 2022

© Copyright by Lijin Yang 2022.
All rights reserved.

Abstract

As one of the fundamental problems in computer vision, understanding human activity from videos is the key to the next-generation human-oriented assistive AI technology. It is also the core technique to various real-world applications such as surveillance, home-assistance robot, autonomous driving and VR/AR systems. In recent years, with the rapid development of deep learning techniques, remarkable progress has been made on this topic, accompanied by various strong deep backbone models that can extract powerful features to describe the undergoing activity in the video. However, most previous works only focus on increasing the representation ability by designing sophisticated network architectures, without fully leveraging the high-level relations hidden within the videos, which could largely enhance the human activity understanding performance.

This thesis focuses on mining multilateral relations to solve three remaining obstacles on the road toward the practical application of human activity understanding techniques in real-world environments. To be more specific, under the most common supervised setting, current models are not robust enough against the instability of videos (e.g. outlier frames in the video). Additionally, in real-world scenarios supervised learning methods struggle to generalize to unseen environments, while there are not enough labels for adapting the model to the new environments. Furthermore, due to the complexity and diversity of human activities in-the-wild, the capability of current models in dealing with complex activity described by natural language is still far from sufficient. To deal with the aforementioned obstacles, gradually relaxing the constraints in human activity understanding settings, we specifically mine the following relations that lie within videos: 1) temporal local-global relations between local video clips and a global video representation. With this type of relation, we can effectively find the most discriminative video snippet in a long video sequence, thus improving the performance of robust activity recognition. 2)

intrinsic relation among different modalities. Since videos are naturally accompanied by multiple modalities such as audio and optical flow, mining the relation among these modalities can help the action recognition when annotated labels are not present in a new domain. 3) relation between positive video proposal-sentence pair and negative video proposal-sentence pair. Since most real-world application does not limit the actions in a pre-defined set of classes, this third type of relationship can help to localize the temporal video segment that corresponds to a natural sentence, which further enhances the practicability of human activity understanding techniques.

Supervised human action understanding can be performed on videos taken from multiple perspectives, among which, first-person videos taken by wearable cameras record human behaviors from the same perspective as humans daily observation. This unique perspective enables a wide range of applications such as VR/AR, human computer interaction and home assistance techniques. Regarding the problem that most methods are designed for third-person videos and perform suboptimally on the first-person perspective, this thesis first focus on supervised activity recognition in first-person videos. One of the major reasons that previous works do not perform well on first-person action recognition is that the unique field of view makes actions sometimes happen outside the video viewing range. Thus, this thesis mines the relation between local and global deep features, leveraging a global knowledge of all the local clips to identify which clip is the most discriminative one and suppress the less important clip feature. In order to effectively leverage the local and global feature relations, we introduce a novel stacked temporal attention module, enabling a progressive relation mining and refinement.

To enhance the application of activity understanding algorithms in the real-world application, one major obstacles is that there does not exist sufficient labels to enable the adaptation of trained deep neural networks in the numerous in-the-wild scenarios. While unsupervised domain adaptation techniques are applicable to address this issue by

minimizing the domain gaps between seen and unseen scenarios, most previous works only focus on the appearance, without fully exploiting the characteristic of videos. Videos add one additional temporal dimension compared to images, which naturally provides multiple modalities such as optical flow and audio. In this thesis we mine the relation among these multiple modalities within videos to perform activity understanding across domains. We found that each modality has its strength in a certain aspect, and the interaction of these modalities can provide useful information for understanding activities in unseen environments.

To further escalate the practicability of human activity understanding in-the-wild, the ability of recognizing a pre-defined set of actions is far from sufficient. It is essential that any kind of activities described in natural languages can be effectively modeled. In this thesis, we also focus on the activity understanding in this open-set setting, where we aim to find the location of action in a video given a natural language sentence description as input. We further focus on a more challenging weakly-supervised setting where we do not have access to ground truth action location, but only video-level video-sentence correspondence. We formulate the relation in this problem as the ranking between positive location proposals and the negative ones, and learn to output correct proposals in a self-supervised manner.

Acknowledgements

I could never have completed this work without the support and assistance of many people. First and foremost, I would like to express my deepest gratitude to my adviser, Prof. Yoichi Sato, for his continuous support throughout my five-year study. His dedication to refinement and enthusiasm for research motivated me deeply. Under his patient guidance, I learned from scratch how to compose and express an idea more logically and how to tackle research problems. I could not imagine a better advisor than him. I also would like to express grateful thanks to Prof. Yusuke Sugano. With his help, I learned how to tackle research problems and how to think like a scientist rather than an engineer. Special thanks should also be expressed to Dr. Yifei Huang. Discussion with him is always pleasant and inspiring. I'm really grateful that he has been supporting me in learning how to formalize research problems; how to tackle research problems; and how to present work better. His patience and positiveness have taught me a lot.

This thesis would not have been possible without generous financial support from the SPRING GX Program of The University of Tokyo by Japan Science and Technology Agency (JST). These supports are gratefully acknowledged.

My sincere thanks also go to all members in Sato Lab, for the inspirational discussions and for the kindness of everyone. I will remember deeply in my heart how friendly they are and wish them all the best in the future.

Last but not least, I would like to express my deepest love and heartfelt thanks to my parents and all my family members. Without their uncondi-

tional support and continuous consideration, I would not be so optimistic about my life and study.

December 2022

Contents

Abstract	i
Acknowledgements	v
List of Figures	xiv
List of Tables	xvii
1 Introduction	1
1.1 Motivation	1
1.2 Overview	3
1.2.1 First-person Activity Recognition by Exploring Local- Global Relations (Chapter 2)	3
1.2.2 Domain Adaptive Activity Recognition by Exploring Relations among Different Modalities (Chapter 3)	4
1.2.3 Weakly-Supervised Temporal Grounding of Natural Lan- guage by Exploring Positive-Negative Relations (Chap- ter 4)	6
2 First-person Activity Recognition by Exploring Local-Global Relations	9
2.1 Introduction	9
2.2 Related Works	12
2.2.1 First-person Activity Recognition	12
2.2.2 Temporal Attention for Activity Recognition	13
2.3 Proposed Method	14

2.3.1	Backbone Encoder	15
2.3.2	Stacked Temporal Attention Module	15
2.3.3	Model Training	18
2.4	Experiment Results	18
2.4.1	Implementation Details	19
2.4.2	Initialization Options of Global Feature	19
2.4.3	Comparison of Temporal Attention Methods	20
2.4.4	Comparison with State-of-the-art	23
2.4.5	Exploration of STAM	23
2.4.6	Visualization	25
2.4.7	Experiments on Third-person Dataset	27
2.5	Conclusion	29
3	Domain Adaptive Activity Recognition by Exploring Relations among Different Modalities	31
3.1	Introduction	32
3.2	Related Works	35
3.2.1	Unsupervised Domain Adaptation (UDA) other than activity Recognition	35
3.2.2	Action Recognition and its UDA	35
3.3	Proposed Method	37
3.3.1	Mutual Complementarity (MC) Module	38
3.3.2	Spatial Consensus (SC) Module	40
3.3.3	Adversarial Domain Alignment	42
3.4	Experimental Results	42
3.4.1	Datasets	42
3.4.2	Implementation Details	43
3.4.3	Comparison with State-of-the-art	44
3.4.4	Visualization	49
3.4.5	Ablation Study	51
3.4.6	Contribution of Different Modalities	54
3.4.7	Analysis on Parameters and Computational Complexity	56
3.5	Conclusion	57

4	Weakly-Supervised Temporal Grounding of Natural Language by Exploring Positive-Negative Relations	59
4.1	Introduction	60
4.2	Related Works	63
4.2.1	Temporal Sentence Grounding with Strong Supervision	63
4.2.2	Weakly Supervised Temporal Sentence Grounding . . .	64
4.2.3	Self-training in Weakly Supervised Learning	65
4.2.4	Bayesian Deep Learning	65
4.3	Proposed Method	66
4.3.1	Problem Formulation and Overview	66
4.3.2	Uncertainty Estimation via Bayesian Teacher	66
4.3.3	Mutual Learning with Temporal Augmentation Cycle Consistency	69
4.4	Experimental Results	72
4.4.1	Implementation Details	73
4.4.2	Results and Comparisons	74
4.4.3	Ablation Study	75
4.4.4	Qualitative Results	82
4.4.5	Limitation and Future Work	82
4.5	Conclusion	83
5	Conclusion	85
5.1	Summary	85
5.2	Contributions	87
5.3	Future Directions	88
5.3.1	Compositional Human Activity Understanding	88
5.3.2	Long-tail Human Activity Understanding	90
	Bibliography	93
	Publications	119

List of Figures

1.1	Overview of this thesis. In this thesis, we aim at mining multilateral relations hidden with videos, for relaxing the constraints in the problem settings of human activity understanding. By this means, we can push forward the human activity understanding techniques to real-world applications. Specifically, by mining the three types of relations namely local-global relation (Chapter 2), multi-modal relation (Chapter 3), and positive-negative relation (Chapter 4), we step by step relax the constraints in the problem setting, going from fully-supervised action recognition to open-set action grounding.	2
2.1	Example of activity “turn off faucet” in the EGTEA dataset [LLR18]. Only the frames with green boundaries provide enough information for determining the activity. In all other frames, the activity happens outside the visible field-of-view.	11
2.2	Overview of our proposed STAM. Using the features from the backbone, a global feature \mathbf{g}^0 is first initialized. This global feature serves as aggregated global information of all input clips and is used for the following global attention layers. The stack of global attention layers progressively refines the attention weight as well as the global feature. Classification loss is applied to the output of each layer.	14

2.3	Visualization of temporal attention score of a few samples of the EGTEA dataset when stacking different numbers of global attention layers. The highest temporal attention score in each layer is underlined. In the case of global attention layer 0, initialization using the self-attention method is used. The frame with green background indicates that the backbone model can predict the correct activity class with this clip alone as input.	26
2.4	Visualization of temporal attention scores of three samples from the HMDB51 dataset when stacking different numbers of global attention layers. The highest temporal attention score in each layer is underlined. In the case of global attention layer 0, initialization using the self-attention method is used. The frame with green background indicates that the backbone model can predict the correct action class with this clip alone as input.	29
3.1	Different from existing UDA works that directly align the multi-modal inputs (a), we find that it is more effective to first enhance the transferability of each modality by cross-modal interaction, and then perform cross-domain alignment (b).	32
3.2	Overview of the proposed CIA model. We showcase three modalities RGB, Flow and Audio as input but it can be easily extended to add other modalities such as depth. In the figure, \oplus denotes element-wise summation, \otimes is element-wise multiplication, and \circledast means the correlation operation that calculates the Pearson correlation coefficient on each spatial position.	37
3.3	The Mutual Complementarity module (MC) showcased using modality M . M could be any modalities of RGB, Flow, and Audio, and also can be extended to other modalities if available, <i>e.g.</i> , depth.	39

3.4	Grad-CAM [SCD ⁺ 17] visualizations of features before and after cross-modality feature refinement by MC. The ground-truth activities are: (a-1) take spoon, (a-2) move spoon, (b-1) take garlic, (b-2) take oil. (a-1) and (a-2) show RGB activation maps (left) and the activation map of RGB modality refined by other modalities (right). Similarly, (b-1) and (b-2) depict the activation maps of the Flow modality alone and Flow refined by other modalities.	50
3.5	Grad-CAM [SCD ⁺ 17] visualizations of RGB, refined RGB, Flow, refined Flow, and fused modality after SC. The ground-truth activity labels are: (a) open cupboard, (b) put down spoon.	50
3.6	t-SNE plots of feature spaces produced by TA ³ N (a) and TA ³ N+CIA (b). Source is shown in blue and target in red. Our method better aligns source and target domains.	51
3.7	Per-class accuracy of several most frequent verbs of the E100 validation dataset. For verbs like “wash”, “turn-on” and “turn-off”, RGB modality interacted with Audio modality can significantly boost performance. Information from the Flow modality helps RGB in discriminating verbs like “open”, “cut” and “mix”.	56
3.8	Per-class accuracy of several most frequent nouns of the E100 validation dataset. For sound-related nouns like “tap” and “sponge”, the Audio modality can greatly aid the RGB modality in improving the per-class accuracy. At the same time, motion-related nouns like “lid” and “knife” can get improvement by interacting with the Flow modality.	57

4.1	(a) Existing methods [ZHCL22, ZHC ⁺ 22] find it hard to distinguish the two cases since they learn positive proposals purely based on negative proposals. (b) Our method provides extra supervision signals for learning positive proposals. (c) Performance of the backbone network [ZHC ⁺ 22], backbone network trained with existing self-training methods pseudo labeling [L ⁺ 13], Mean Teacher (MT) [TV17], and backbone network trained with our method. Directly applying self-training methods for semi-supervised learning negatively influences performance, while our self-training method can improve the backbone performance.	61
4.2	Overview of our proposed method. The teacher network takes as input weakly augmented data while the student network takes as input multiple strongly augmented data. Then teacher-student cycle consistency is used for mutual learning of the two networks, considering the uncertainty u into consideration. Gaussian masks are generated to represent the proposals, and we further use reconstruction loss L_{rec} and ranking loss L_{rank} to ensure high-quality proposals.	67
4.3	Model confidence (computed by $1-u$) and mIoU on the Charades-STA dataset are highly correlated, indicating that we can leverage the uncertainty estimation u to represent the quality of the network output.	68
4.4	Results on the ActivityNet Captions (left) and Charades-STA (right) datasets when our method is applied on different backbone networks.	79
4.5	Qualitative examples of the ground truth (GT), the backbone network (CPL), and the backbone method with our mutual learning (CPL+ours). Examples (a, b) are from the ActivityNet Captions dataset, and (c, d) are from the Charades-STA dataset.	81

List of Tables

2.1	Different design options for the global feature. Experiments are conducted on the EGTEA dataset using I3D as the backbone and 9 clips as input.	21
2.2	Comparison of different feature aggregation methods on EGTEA and EPIC-100 dataset. We investigate multiple aggregation methods: Max pooling, Average pooling (Avg pooling), Bi-directional GRU (Bi-GRU), 1D convolution (1D-conv) and self-attention (self-att). The middle block in this table shows result of previous works.	22
2.3	Results on the EGTEA dataset. * indicates optical flow is used, dark rows indicates gaze information is used.	24
2.4	Results on the EPIC-100 dataset. All methods use only RGB frames as input.	25
2.5	Analysis of the influence of input clip (temporal receptive field) and the number of stacked global attention layers. Experiments are done using I3D backbone on the EGTEA dataset. .	25
2.6	Results on the HMDB51 dataset. * indicate the method uses optical flow as input.	28
3.1	Performance comparison on the UCF-HMDB (U-H) dataset. We show the input modality and the backbone used by each method for better comparison. \diamond refers to averaging the outputs from each modality classifier, while * means concatenate features of different modalities. Under the same experiment setting, our method can clearly outperform previous methods.	45
3.2	Performance comparison on the EPIC-Kitchens-55 (E55) dataset.	46

3.3	Performance comparison on the EPIC-Kitchens-100 (E100) validation set. R, F and A refers to RGB, Flow and Audio modalities, respectively. We show each method together with its source only performance in the row above. Under the same experiment setting, our method can clearly outperform previous methods.	48
3.4	Ablation study on Mutual Complementarity module (MC) and Spatial Consensus module (SC) of our CIA model.	52
3.5	Performance comparison of our SC module with other approaches on the E100 validation set.	53
3.6	Results of single modality before and after interacting with different modalities on the E100 validation set are shown to validate the contribution of each modality.	55
3.7	Model parameter and computational complexity.	57
4.1	IoU@{ 0.3, 0.5, 0.7} and mIoU results on the Charades-STA dataset test split. CPL (ori) denotes the results reported in [ZHC ⁺ 22], while CPL (rep) is our replicated result. CPL (aug) is the backbone method CPL trained with original data but is directly inferenced with augmented data. CPL + aug is the result where CPL is both trained and inferenced with augmented data. backbone method CPL applied with the standard teacher-student self-training method Mean Teacher [TV17] is shown as CPL + MT. The bold numbers represent the top-1 result. Our proposed method outperforms previous works in all metrics.	76

4.2	IoU@{0.1, 0.3, 0.5} and mIoU results on the ActivityNet Captions dataset val_2 split. CPL (ori) denotes the results reported in [ZHC ⁺ 22], while CPL (rep) is our replicated result. CPL (aug) is the backbone method CPL trained with original data but is directly inferred with augmented data. CPL + aug is the result where CPL is both trained and inferred with augmented data. backbone method CPL applied with the standard teacher-student self-training method Mean Teacher [TV17] is shown as CPL + MT. The bold numbers represent the top-1 result. Our proposed method outperforms previous works in most metrics.	77
4.3	Ablation study on the Charades-STA dataset.	78
4.4	Results of our method when using different augmentation techniques. V: video temporal scaling and shifting; M: video temporal masking; D: decomposition of sentence queries with SRL.	80
4.5	IoU of different methods at Recall@5 on the Charades-STA dataset. Recall@1, IoU@0.3 is shown for reference.	83

Chapter 1

Introduction

1.1 Motivation

As one of the most fundamental research topics in computer vision, human activity understanding has been extensively studied in recent years. Several research directions include recognizing current activity, predicting human intention at the next timestamp, localizing the start and end timestamps of an action instance in the video, etc. Once reliably realized, the technique of human activity understanding can drive numerous real-world applications ranging from video recommendation systems to autonomous driving and home assistance robots.

However, several key obstacles still remain on the road toward the practical application of human activity understanding techniques in real-world environments. First of all, under the most common supervised setting, current human activity understanding models are not robust enough to handle the environmental change in the videos. Secondly, when it comes to real-world scenarios, supervised learning settings are not generalizable to unseen environments, while there are not enough labels for any models to be trained in these new environments. Thirdly, in terms of deploying the human activity understanding models into in-the-wild applications, the ability to recognize

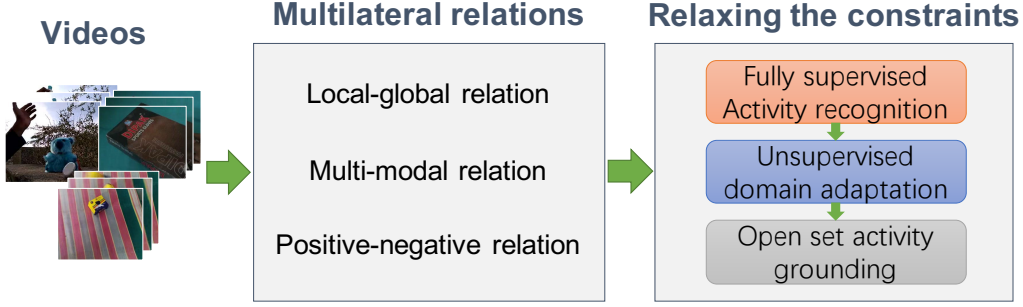


Figure 1.1: Overview of this thesis. In this thesis, we aim at mining multilateral relations hidden with videos, for relaxing the constraints in the problem settings of human activity understanding. By this means, we can push forward the human activity understanding techniques to real-world applications. Specifically, by mining the three types of relations namely local-global relation (Chapter 2), multi-modal relation (Chapter 3), and positive-negative relation (Chapter 4), we step by step relax the constraints in the problem setting, going from fully-supervised action recognition to open-set action grounding.

only a pre-defined set of actions is far from sufficient. The ability to model any kind of human activities, described in natural languages, is lacking in most current models for human activity understanding.

This thesis aims at solving the aforementioned obstacles, gradually relaxing the constraints in human activity understanding settings, making an essential step toward the real-world application of human activity understanding techniques. We address these issues by mining and leveraging multilateral relations hidden in the videos (Figure 3.1). Firstly to increase the model robustness in the supervised human activity understanding setting, we explore the local-global relations in videos, using the global information in the videos to better find the most discriminative video segment for human activity recognition (Chapter 2). Secondly, to address the problem that models perform sub-optimally in the new environments since no labels are provided, we apply domain adaptation techniques, mining the relations among differ-

ent modalities (Chapter 3). Last but not the least, to address the issue that learning only on the pre-defined action classes is not enough for real-world applications, we further explore the human activity understanding in an open-set environment, where we mine the relations between positive video-sentence pairs and the negative (unpaired) videos and sentences(Chapter 4). This thesis believes that several critical steps are made towards the real-world application of human activity understanding techniques by gradually relaxing the constraints in the problem setting.

1.2 Overview

1.2.1 First-person Activity Recognition by Exploring Local-Global Relations (Chapter 2)

As one of the fundamental task of human activity understanding, fully supervised human activity recognition takes trimmed video as input and need to accurately classify it into pre-defined fixed number of activity categories. Although remarkable progress has been made by exploring different deep learning methods, nowadays enhancing the robustness of recognition models against the instability of videos (e.g. outlier frames in the video) remains challenging. Unlike third-person videos captured by static cameras, first-person videos are captured by cameras mounted on the body of camera viewers, resulting in unique viewpoints that are the same as what we actually see with our eyes. On one hand, the unique self-center perspective of first-person videos naturally provide ground for us to better understand human intelligence, and thus help us to produce a smarter robot that could better imitate human. However, on the other hand, first-person videos also bring two major challenges: limited field of view and huge ego-motion.

Due to the unique field of view of the camera mounted on the body of camera wearers, the human activities of camera wearers are not always ob-

servable in the videos. For example, when the camera wearer drinks a bottle of water with cameras mounted on his/her head, the activity of “drink water” will not be able to be observed due to the motion of the head. And the sharp movements of camera wearers (e.g. suddenly turning head) are also likely to introduce purely background information which harms the recognition of actual activities. Due to these characteristics, compared to third-person videos, first-person videos are more likely to include temporal outliers in the videos. Therefore, the performance of activity recognition methods in first-person videos is still not comparable to that in third-person videos.

In order to deal with temporal outliers and thus improve first-person activity recognition accuracy, one natural solution is to apply temporal attention to the input videos, emphasizing the discriminative clips while suppressing the noisy ones. To better decide the importance of each clip of videos, we explore the multilateral relations: the relations within local-local clip features which only contain information about a part of videos, and the relations between local-global features for better deciding which part of the video is more discriminative. By effectively leveraging these relations through a stacked temporal attention module, we could enhance the robustness of the model against temporal outliers. Experiments on both first-person datasets and third-person datasets demonstrate the effectiveness and generability of the proposed model.

1.2.2 Domain Adaptive Activity Recognition by Exploring Relations among Different Modalities (Chapter 3)

One major obstacle to the real-world application of human activity recognition is that it’s time and labor-consuming to obtain annotations for training. Therefore, the lack of training data gives rise to increasing research on human activity understanding under weakly-supervised settings, unsupervised set-

tings, or few-shot settings. Among all these settings, unsupervised settings require no annotations of the target data, which could alleviate the need for annotations to the greatest extent. Unsupervised domain adaptive video activity recognition aims to recognize the activity of a target domain using a model trained with only out-of-domain (source) annotations, is applicable to address this problem, and shows great potential to promote the practical application of human activity understanding.

Compared with domain adaptive image classification task which takes images as input, the inherent complexity of videos makes domain adaptive video activity recognition task challenging but also provides ground for leveraging multi-modal inputs (*e.g.*, RGB, Flow, Audio). Most previous works on domain adaptive activity recognition utilize the multi-modal information by either aligning each modality individually or learning representation via cross-modal self-supervision. Different from previous works, we find that the cross-domain alignment can be more effectively done by cross-modal interaction exploring multilateral relations among modalities. Cross-modal knowledge interaction allows other modalities to supplement missing transferable information because of the **cross-modal complementarity**. Also, the most transferable aspects of data can be highlighted using **cross-modal consensus**. Extensive experiments against multiple baselines reveal the effectiveness of exploring multi-modal relations for improving the activity understanding task under the challenging unsupervised domain adaptation setting.

To enhance the transferability of deep models, we present a novel model that jointly considers these multilateral relations for domain adaptive action recognition. We achieve this by implementing two modules, where the first module exchanges complementary transferable information across modalities through the semantic space, and the second module finds the most transferable spatial region based on the consensus of all modalities.

1.2.3 Weakly-Supervised Temporal Grounding of Natural Language by Exploring Positive-Negative Relations (Chapter 4)

In addition to relaxing the restriction of human activity understanding setting from the perspective of demand of annotations (from fully-supervised to unsupervised), now it's also possible to challenge the activity understanding task from the perspective of dealing with more complex activities. Instead of understanding relative "simple" human activities which are a set of pre-defined, fixed number of categories consisting of verbs and nouns, human activity could also be described in more detail as a natural sentence. Compared to the activity that is described with one verb and one noun, describing an activity with natural language allows for the combination of multiple atomic "simple" activities, and is no longer limited by the pre-defined number of activity categories. This is an essential step if future AI assistive technologies could be widely used in our human society.

In this thesis, we concentrate on the temporal sentence grounding task which aims at finding the start and end of an activity in a video given a natural language sentence description as input. We further focus on a more challenging weakly-supervised setting without the start and end timestamp annotation of ground-truth activity, and only video-level video-sentence correspondence is provided. Previous works for this task are mainly formulated under the multiple instance learning framework, where they generate temporal segment proposals by the given sentences. However, they ignored the relations within these proposals which could provide valuable cues for this task since the level of supervision is low. To deal with this challenging task, we formulate the ranking relations both within positive proposals and between positive and negative ones and utilize the ranking relations to predict more accurate temporal timestamps. We design the first self-training-based approach for this problem, based on a teacher-student co-training frame-

work. Experiments on two datasets demonstrate that deeper exploration of such multilateral relations could better find the boundary of human activity.

Chapter 2

First-person Activity Recognition by Exploring Local-Global Relations

In this chapter, we introduce our method in the fully-supervised human activity understanding setting. Although extensive studies have been made in this direction, we especially focus on the robustness of the human activity understanding techniques. Since previous methods are not robust enough for addressing the changes of viewpoint in the videos, in this chapter we explore using local-global relations in the videos for more robust human activity recognition even under large viewpoint changes.

2.1 Introduction

The recognition of actions within videos represents a fundamental challenge within the field of computer vision. A significant amount of research has been devoted to the task of recognizing human activities within footage captured by stationary cameras, also known as third-person videos [SZ14, CZ17a, FFMH19]. In contrast to third-person videos, videos captured by

wearable cameras, known as first-person videos, provide a unique perspective on human behavior and have a wide range of applications, including the enhancement of human-computer interaction. Unfortunately, despite the potential benefits, the performance of current action recognition techniques in first-person videos remains inferior to that achieved in third-person videos [SEL19].

A significant obstacle to activity recognition on first-person videos is the limited field of view which results in the possibility of relevant actions occurring outside the frame of the recorded footage (Figure 2.1). Since the activity may not always be observable throughout the entire video sequence, although methods that involve sparsely sampling clips from the video, followed by computation of the clip consensus through pooling, have been shown to be effective for generating more discriminative features in third-person videos [WXW⁺16, LGH19], cannot yield comparable performance when applied to first-person videos.

An additional significant obstacle for first-person activity recognition is that the videos are commonly affected by substantial ego-motion resulting from sudden movements of the camera wearer, such as turning the head, which complicates the encoding of motion [SEL19]. The use of current clip sampling techniques can lead to the inclusion of clips that contain purely background information, which impairs recognition, even when using an enlarged temporal receptive field.

A potential solution to address both challenges is to employ temporal attention to the clips, thereby highlighting the informative clips while diminishing the impact of the noisy ones. Several previous studies have investigated this direction. For example, Pei *et al.* [PBTM17] propose to use gated LSTM for deciding the weight of each input frame, and Girdhar *et al.* [GCDZ19] leverage transformers [VSP⁺17] to compute spatiotemporal attention on humans in the video. However, when using techniques like LSTM



Figure 2.1: Example of activity “turn off faucet” in the EGTEA dataset [LLR18]. Only the frames with green boundaries provide enough information for determining the activity. In all other frames, the activity happens outside the visible field-of-view.

or transformers to compute clip-wise relationships, the model is still limited in its ability to make an optimal choice of attention without a comprehensive understanding of the entire input.

Intuitively thinking, having a complete understanding of all local clips can aid in determining which clip is the most informative and minimize the impact of less important clip features.

In this chapter, we seek to capitalize on this intuition by introducing a Stacked Temporal Attention Module (STAM) that fuses both local information obtained from individual clips and global information derived from the entire video to generate temporal attention. Our experiments show that this can be achieved by stacking self-attention [VSP+17] layers, with global information serving as the query vectors. It is worth highlighting that our STAM can be integrated with various existing activity recognition architectures. Through conducting experiments on multiple first-person datasets, we have found that the addition of our STAM leads to improved activity recognition performance. Additionally, our method has been demonstrated to generalize to third-person datasets through further experimentation on the HMDB51 dataset [KJG+11].

The main contributions of this chapter are summarized as follows:

- We propose a simple yet effective Stacked Temporal Attention Module

(STAM) that could be built on top of most existing backbones for improving first-person activity recognition.

- Experiments on multiple datasets demonstrate that our proposed module can improve the performance of various backbone models.

2.2 Related Works

2.2.1 First-person Activity Recognition

Recognizing activities through videos is an extensively researched area within the field of computer vision, and the recognition of activities within first-person videos has been an important topic of study. Prior efforts have primarily centered around the creation of various hand-crafted features [FLR12, LYR15] to encapsulate the complex spatiotemporal information present in video sequences. With the advancement of deep learning techniques, significant progress has been made in recent times [MFK16, SL18, LLL19b, KPvD⁺19, HSS20, HCS20, LNXG21, GG21]. Since temporal information is essential for recognizing the activities in the videos, a line of research [SEL19, LGG⁺18] uses recurrent networks for sequential modeling. However, one major drawback of using recurrent models is that the span of attention is limited [SMJ⁺16], which can impede performance when the input video is of considerable duration. Some previous works also tried to incorporate unique first-person cues such as hand and gaze [SAJ16, LLL19b, ZYP⁺18, HCL⁺20, HCLS18], or use multiple modalities [KNZD19] for improving first-person activity recognition. Other researchers have experimented with processing multiple frames, utilizing Spatio-temporal convolution networks to integrate features both spatially and temporally [LLR18, WZWY20]. A limitation of 3D convolution networks is that their restricted temporal receptive field leads to the lack of an understanding of the long-term temporal characteristics of

the input. In this study, we leverage these models as backbones and construct our Stacked Temporal Attention Module on top to further enhance their performances.

2.2.2 Temporal Attention for Activity Recognition

Previous studies have investigated various methods of temporal attention to aggregate the temporal features for longer-term video understanding and have demonstrated encouraging results in recognizing activities within third-person videos. One line of research has sought to utilize temporal aggregation techniques to implicitly assign varying weights to each feature [ZAOT18, HZG⁺19, WFF⁺19]. For example, TSN [ZAOT18] sparsely samples frames from videos and assigns uniform temporal attention by average pooling to combine features of each frame. TLE [DSVG17] designed a temporal linear encoding layer for generating weight for each temporal input. StNet [HZG⁺19] designed temporal convolution modules integrated into multiple blocks of CNN to implicitly aggregate the features both spatially and temporally. LFB [WFF⁺19] uses context beyond the action external memory called long-term feature bank to help classifying complex actions. The feature bank operation can implicitly generate temporal attention weights on the features.

An alternative line of research has focused on explicitly calculating a temporal attention score for each feature and utilizing a weighted average as the aggregation method. As for previous works, TAGM [PBTM17] modified the LSTM block and sequentially determine a weight for each input frame. One of the significant disadvantages of LSTM is its inability to incorporate global information, resulting in sub-optimal attention generated by only local features. Similar to our method, CatNet [WPQ20] also uses global information to aggregate the clip features. However, they simply concatenate the local and global features and use several fully connected layers for predicting the weights. The attention obtained by these fully connected layers is applied as

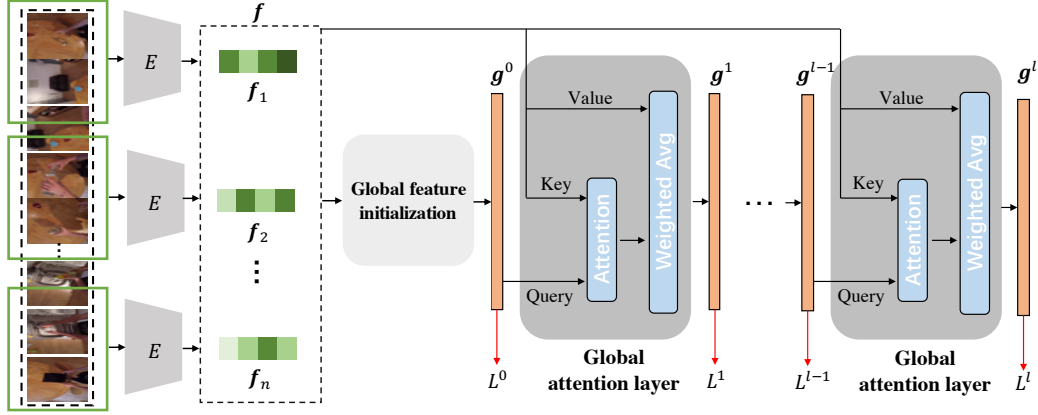


Figure 2.2: Overview of our proposed STAM. Using the features from the backbone, a global feature \mathbf{g}^0 is first initialized. This global feature serves as aggregated global information of all input clips and is used for the following global attention layers. The stack of global attention layers progressively refines the attention weight as well as the global feature. Classification loss is applied to the output of each layer.

temporal attention weights.

In this research, we present a straightforward yet effective module that can be incorporated into a variety of activity recognition models for computing temporal attention. Our results reveal that by simply stacking self-attention layers [VSP⁺17], with global feature vectors serving as queries, we are able to enhance recognition performance on a range of first-person datasets.

2.3 Proposed Method

In this section, we describe our proposed Stacked Temporal Attention Module (STAM). Different from existing methods for temporal attention [WXW⁺16, PBTM17, ZAOT18], we consider the global information across clips, and use the relation between a clip and the global representation to determine the significance of each clip. An overview of our model architecture is depicted

in Figure 4.2.

2.3.1 Backbone Encoder

Our STAM can cooperate with most of the existing backbones that use a neural network and back-propagation for training. Here we showcase a backbone encoder E that uses 3D convolution and takes as input a small video clip with a few video frames. Given a trimmed video containing an action, we follow the practice of TSN [WXW⁺16] to split the input into N segments. From each segment, we extract a clip of 16 consecutive frames as inputs to the backbone network. The backbone would encode the clips into clip features $\mathbf{f}_1, \dots, \mathbf{f}_N$. These clip features are used as inputs to our STAM.

2.3.2 Stacked Temporal Attention Module

With the features encoded by the backbone, our STAM finds attention for temporally aggregating the features using both local and global information. As shown in Figure 4.2, the STAM is a stack of several global attention layers. The input to the first global attention layer is the global feature vector \mathbf{g}^0 initialized by the local feature vectors \mathbf{f} . We explore multiple initialization methods in this work. In the l -th global attention layer, an attention weight vector \mathbf{a}^l is computed using the relation between the input global feature \mathbf{g}^{l-1} and each local feature \mathbf{f} . This attention weight is used to generate a new global vector \mathbf{g}^l as the output of this layer. For simplicity, we omit all the classifiers and normalization layers in this section.

Global Attention Layer

As in Figure 4.2, the global attention layer takes the encoded clip features and the global feature \mathbf{g} as inputs. Intuitively, knowing the global context can help to determine the importance of each local clip. Thus, the relationship

between the global feature and each of the clip features should be considered to emphasize discriminative clips for getting better temporal attention. This could be neatly done by taking advantage of the self-attention operation of the Transformers [VSP⁺17]. Take the first global attention layer as an example, the query vector is acquired using the global feature \mathbf{g}_0 , while the key and value vectors are from each local feature \mathbf{f} :

$$\mathbf{q}_i = W_q \mathbf{g}^0, \quad \mathbf{k}_i = W_k \mathbf{f}_i, \quad \mathbf{v}_i = W_v \mathbf{f}_i; \quad i \in [1, N] \quad (2.1)$$

where W_q, W_k, W_v are the weight matrices for projecting the features to query \mathbf{q} , key \mathbf{k} and value \mathbf{v} , respectively. The queries and keys are vectors with hidden dimension d , and the value vectors have the same dimension as the backbone features. We then compute a weight vector a_i^1 for the i -th clip using a modified scaled dot-product operation and softmax operation:

$$c_i = \frac{1}{N\sqrt{d}} \sum_{j=1}^N \mathbf{q}_i \cdot \mathbf{k}_j, \quad a_i^1 = \frac{e^{c_i}}{\sum_j e^{c_j}} \quad (2.2)$$

The generated weight vectors are used as temporal attention weights:

$$\mathbf{g}^1 = \sum_{i=1}^N a_i^1 \mathbf{v}_i \quad (2.3)$$

\mathbf{g}^1 is the output of the first global attention layer and will also be the input to the next layer. We add a classifier on top of \mathbf{g}^1 to recognize activities and omit it in the equations for simplicity.

Stacking several modules sequentially has shown significant improvements in many tasks like human pose estimation [NYD16] and activity segmentation [FG19]. The use of multiple models in a sequential manner allows for a gradual improvement of the previous output. In this study, drawing on the effectiveness of previous approaches, we propose the utilization of stacked

global attention layers to enhance the refinement of temporal attention and global features.

Denote the operation of the first global attention layer as:

$$\mathbf{g}^1 = \mathcal{G}^1(\mathbf{g}^0, \mathbf{f}_1, \dots, \mathbf{f}_N) \quad (2.4)$$

Similarly, the k-th global attention layer can be represented accordingly as follows:

$$\mathbf{g}^k = \mathcal{G}^k(\mathbf{g}_{k-1}, \mathbf{f}_1, \dots, \mathbf{f}_N) \quad (2.5)$$

We show in our experiments that as the global feature refines, the attention scores also change, giving more focus on the most critical clip from a global perspective.

Global feature initialization

Given the encoded clip features $\mathbf{f}_1, \dots, \mathbf{f}_N$ as inputs, the global feature vector \mathbf{g}^0 can be initialized in multiple ways. We experiment with the following choices while others are possible:

- **Pooling:** One of the most straightforward ways is to do pooling on all clip features to get the global feature: $\mathbf{g}^0 = \text{pool}(\mathbf{f}_1, \dots, \mathbf{f}_N)$. We experiment with both max-pooling and average-pooling in the experiment section.
- **Recurrent networks:** It is also feasible to use recurrent neural networks to process all clip features, and generate a global aggregated feature. Recurrent networks can be LSTM, GRU, *etc*, we choose bi-directional GRU as an example: $\mathbf{g}^0 = \text{GRU}(\mathbf{f}_1, \dots, \mathbf{f}_N)$.
- **Temporal convolution:** Apply a 1D temporal convolution layer to fuse all the clip features: $\mathbf{g}^0 = \text{Conv}(\mathbf{f}_1, \dots, \mathbf{f}_N)$.

- **Light-weight CNN:** Inspired by previous works [WXM⁺19, KTT19], we also tried to use a light-weight CNN for extracting the global feature. The lightweight CNN is a small 2D CNN [HZC⁺17] that takes one frame from each input clip as input. The output features of all inputs are averaged to form the global feature.
- **Self-attention:** We use the self-attention operation of Transformers [VSP⁺17] to compute the pairwise relationship between input clips, and use generated weight to form the global feature.

2.3.3 Model Training

To incrementally improve the global features at each layer, a classifier with cross-entropy loss is added to each attention layer during the training of the model. The final loss function is a combination of the loss from both the global feature initialization step and M global attention layers:

$$L = \sum_{i=0}^M \lambda_i L^i \quad (2.6)$$

Here $\lambda_0, \dots, \lambda_M$ is a set of model hyper-parameters to determine the contribution of the different losses.

2.4 Experiment Results

We conduct experiments on two publicly available first-person datasets: EGTEA [LLR18] and EPIC-Kitchens-100 [DDF⁺20] (EPIC-100). The EGTEA dataset contains 29 hours of video footage recorded from a first-person perspective, capturing meal preparation activities performed in a kitchen setting by 32 subjects. The dataset includes fine-grained annotations for 106 different activity classes. Following [SEL19], we report the average of three splits of the

dataset. The EPIC-100 dataset is currently the largest collection of first-person videos, featuring 100 hours of footage recorded in various kitchens by multiple subjects. The dataset includes fine-grained annotations for 97 verbs and 300 nouns. In accordance with the official protocol, performance is reported for verb, noun and the combined activities. For all datasets, the activity recognition accuracy is employed as the evaluation metric.

2.4.1 Implementation Details

We use PyTorch [PGM⁺19] for all the implementation. The hidden dim d is set as 512 in all experiments. For adding STAM on top of all backbones, we use the Adam optimizer with an initial learning rate 1e-4 for training, and decay the learning rate by a factor of 10 when the validation loss does not decrease for 3 consecutive epochs. We train a total of 40 epochs. As for the loss weight λ we empirically set all the λ to 1.

The hidden dim d is set as 512 in all experiments. For the number of global attention layer M , we empirically set different values to cooperate with different backbone encoders. We set $M = 2$ when using TSM [LGH19] and R3D-50 [HKS18] as backbone encoder. For the backbone encoder I3D [CZ17a], we utilize 3 global attention layers for better performance. Following the practice of [WXW⁺16], we use 9 clips with 16 frames as input for I3D backbones while 6 clips with 16 frames for R3D-50 backbones. For the 2D backbone TSM, we uniformly sample 16 frames as the input. The output temporal dimension is the same as the number of clip for both 2D and 3D backbones.

2.4.2 Initialization Options of Global Feature

One core component in our STAM is the use of global feature \mathbf{g} for temporal attention computation. It is possible to use multiple alternatives as the initialization of global feature as described in Section ?? . Table 2.1 shows the performance of our STAM when changed with different alternatives of global

feature \mathbf{g} . From Table 2.1, max-pooling does not produce good results while average-pooling serves as a strong baseline. The use of Bi-GRU, 1D-conv, or lightweight CNN increases the number of learnable parameters, however, the performance is inferior to that of self-attention. Based on these results, using self-attention for global feature initialization is the optimal choice. We posit that one potential explanation for this is that the utilization of weighted averages aids the flow of gradients in the self-attention operation of the global attention layer, thereby enhancing the effectiveness of the training process. Therefore, this option is employed in all subsequent experiments unless otherwise specified.

It is worthwhile to note that vanilla stacking self-attention can hardly take the advantage of stacking more layers, which shows the same trend as in [NBZA21]. However, other initialization options that leverage global information can see an improvement by stacking multiple layers. This is likely because although the vanilla stacking baseline uses updated local features as the key and query for the next layer, it still does not have access to all other local features at once. Additionally, simultaneously updating the query and key can lead to a buildup of errors, which may even lead to a decline in performance when increasing the number of layers. In contrast, our proposed method leverages global features and only refines the query in the stacking process, which allows the model to more accurately determine the relatively relevant clips, and benefit from the use of stacking. This suggests the usefulness of STAM for generating temporal attention, and shows potential applications in other tasks that deal with temporal information, *e.g.*, activity detection.

2.4.3 Comparison of Temporal Attention Methods

To evaluate our proposed temporal attention module STAM, we compare its performance with several baselines and state-of-the-art temporal attention

Global feature	Num of global att layers		
	1	2	3
Vanilla stacking	65.6	65.5	65.5
Max pooling	65.8	65.9	66.3
Avg pooling	65.9	66.3	66.7
Bi-GRU	65.8	66.3	65.9
1D-conv	65.9	66.1	65.9
Self-att	66.2	66.5	66.9
Light-weight CNN	64.7	66.0	67.1

Table 2.1: Different design options for the global feature. Experiments are conducted on the EGTEA dataset using I3D as the backbone and 9 clips as input.

methods. In this experiment, a fixed input of 6 clips with 16 frames is used for all methods. For the EGTEA dataset, the I3D model is used as the backbone, and for EPIC-100 the TSM model is employed. We specifically compare with the following clip feature aggregation methods:

- Global feature initialization baselines without the proposed STAM. These baselines serve as simple straightforward ways for generating temporal attention.
- **TLE** [DSVG17], **TAGM** [PBTM17], **CatNet** [WPQ20] and **TRN** [ZAOT18] are previous works that uses various algorithms for temporal aggregation.
- Our **STAM**. In this experiment, for EGTEA and EPIC-100 we use STAM with three global attention layers and two global attention layers respectively. Studies on the influence of the number of global attention layers are placed in Section 2.4.5.

Quantitative result comparison is shown in Table 2.2. Our STAM demonstrates a marked improvement over other feature aggregation methods, even

Method	EGTEA	EPIC-100		
		Verb	Noun	Action
Max pooling	63.6	61.8	45.6	33.7
Avg pooling	63.0	63.2	48.0	36.2
Bi-GRU	62.5	62.3	45.1	33.7
1D-conv	62.6	63.4	46.3	35.2
Self-att	65.5	62.6	48.6	36.7
<hr/>				
TAGM [PBTM17]	65.4	63.6	45.7	34.3
CatNet [WPQ20]	63.2	61.5	48.7	35.8
TLE [DSVG17]	63.4	63.6	45.4	34.3
TRN [ZAOT18]	64.0	64.6	47.9	36.4
MTRN [ZAOT18]	64.1	64.8	47.6	36.6
GSTA [SR21]	63.8	62.2	49.1	36.5
STA [LLZ ⁺ 20]	63.8	62.6	48.6	36.6
CTA [WBLB21]	62.6	62.3	47.2	35.4
<hr/>				
Our STAM	66.9	64.4	49.3	37.6

Table 2.2: Comparison of different feature aggregation methods on EGTEA and EPIC-100 dataset. We investigate multiple aggregation methods: Max pooling, Average pooling (Avg pooling), Bi-directional GRU (Bi-GRU), 1D convolution (1D-conv) and self-attention (self-att). The middle block in this table shows result of previous works.

when utilizing the same backbone model and input clip count. This provides strong evidence that our STAM is effective in utilizing both local and global information in the clip features to accurately predict temporal attention weights.

2.4.4 Comparison with State-of-the-art

In order to evaluate the enhancement in performance that our STAM brings to the backbones, we conducted experiments and compared its performance to other state-of-the-art methods. We compare the following methods for first-person activity recognition: Ego-RNN [SL18], LSTA [SEL19] and SAP [WZWY20]. We also add our proposed STAM on 3 commonly used backbones: TSM [LGH19], I3D [CZ17a], and 3D-Resnet50 [HKS18]. For the 2D backbone TSM we input multiple frames, and for other backbones the inputs are multiple 16-frame video clips.

Table 2.3 and Table 2.4 show the result comparison on the EGTEA dataset and the EPIC-100 dataset, respectively. From both tables, we can see that the performance of all backbones can be improved when adding our proposed STAM. On the EGTEA dataset, our method achieves the best performance among all the methods with RGB frames as inputs, and is even better than LSTA [SEL19] which uses optical flow as input. Only Min *et al.* [MC21] outperforms our method, but they use optical flow and human gaze as additional input. Similarly, our method can boost the performance of all backbones on the EPIC-100 dataset. This performance increase is significant given the intrinsic difficulty of the EPIC-100 dataset (*e.g.* non-scripted, unbalanced label, large variety of activity).

2.4.5 Exploration of STAM

For all temporal aggregation methods, the number of input clips (temporal receptive field) can have a huge influence on the final performance. We

Method	Split1	Split2	Split3	Average
Ego-RNN [SL18]	62.2	61.5	58.6	60.8
LSTA* [SEL19]	-	-	-	61.9
SAP [WZWY20]	64.1	62.1	62.0	62.7
Lu <i>et al.</i> [LLL19b]	63.7	61.1	59.0	61.3
Min* <i>et al.</i> [MC21]	69.6	-	-	-
TSM [LGH19]	63.8	61.8	60.2	61.9
I3D [CZ17a]	63.0	61.1	58.0	60.7
3D-ResNet50 [HKS18]	63.2	61.4	59.4	61.3

TSM + STAM	66.2	64.1	64.0	64.8
I3D + STAM	66.9	63.8	62.2	64.3
3D ResNet-50 + STAM	65.3	62.9	63.1	63.8

Table 2.3: Results on the EGTEA dataset. * indicates optical flow is used, dark rows indicates gaze information is used.

explore the performance of our STAM with different numbers of clips as input.

Table 2.5 shows the influence of different numbers of attention layers used in our STAM. We experiment with 3, 6 and 9 input clips. In both tables, *Avg* indicates direct averaging of all clip features, which serves as a baseline for understanding the usefulness of temporal attention. *Gain* emphasize the maximum performance gain with respect to the *Avg* method. Usually, since using more clips brings richer information, the recognition accuracy tends to become higher.

From the table, we can see that with the help of STAM, the performance gain of a small temporal receptive field (3 clips) is more obvious. This from one side proves that our STAM can learn to highlight the most important clip feature. The table also suggests that stacking two layers (with 6 clips as input) or three layers (with 3 or 9 clips as input) of global attention works the best, and stacking four layers saturates the performance.

Method	Top-1			Top-5		
	Verb	Noun	Action	Verb	Noun	Action
TSM [LGH19]	63.2	48.0	36.2	89.0	74.2	57.3
I3D [CZ17a]	55.5	43.3	29.3	86.3	70.1	50.6
3D ResNet-50 [HKS18]	56.7	45.3	31.5	86.8	71.7	52.7
<hr/>						
TSM + STAM	64.3	49.3	37.6	89.3	75.4	59.0
I3D + STAM	57.4	45.2	31.7	86.6	70.7	52.2
3D ResNet-50 + STAM	58.0	47.4	33.7	87.8	73.1	54.3

Table 2.4: Results on the EPIC-100 dataset. All methods use only RGB frames as input.

Clip Num	Avg	Num of global att layers					Gain
		0	1	2	3	4	
3	59.1	61.1	61.7	62.2	63.5	62.5	+4.4
6	62.3	64.3	65.1	65.6	65.1	64.9	+3.3
9	63.0	65.5	66.2	66.5	66.9	66.7	+3.9

Table 2.5: Analysis of the influence of input clip (temporal receptive field) and the number of stacked global attention layers. Experiments are done using I3D backbone on the EGTEA dataset.

2.4.6 Visualization

Figure 2.3 shows visualizations of several sample activities in the EGTEA dataset together with the temporal attention for each clip. This figure illustrates the model performance when using 6 clips as input, with one frame being used to represent each 16-frame clip. In each global attention layer, the highest temporal attention score is indicated by an underline. The frames with green boundaries signify that when only the corresponding clip is used as input, the network can correctly predict the activity class. For all samples, it can be observed that with the stacked global attention layers, the temporal attention gradually shifts and assigns greater weights to the correct clip. As




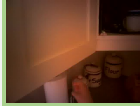
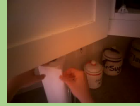


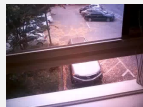


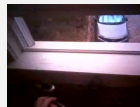
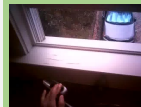

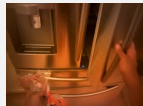





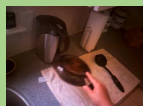
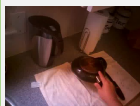
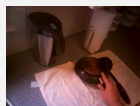


Take paper towel						
Global att layer 0	0.0952	0.1792	<u>0.2819</u>	0.1871	0.1587	0.0980
Global att layer 1	0.0901	0.0845	0.1540	<u>0.3078</u>	0.2965	0.0671
Global att layer 2	0.0677	0.0699	0.1091	0.2896	<u>0.4456</u>	0.0181
Global att layer 3	0.0835	0.0667	0.1024	0.2472	<u>0.4530</u>	0.0471
Turn off faucet						
Global att layer 0	0.1055	<u>0.4401</u>	0.2087	0.0858	0.1542	0.0056
Global att layer 1	0.1006	0.1467	0.1657	0.1728	0.1917	<u>0.2225</u>
Global att layer 2	0.0666	0.1296	0.1729	0.1755	0.1868	<u>0.2687</u>
Global att layer 3	0.0839	0.1422	0.1610	0.1693	0.1876	<u>0.2559</u>
Open fridge						
Global att layer 0	0.1663	0.0957	0.1886	0.1370	0.1837	<u>0.2288</u>
Global att layer 1	0.1569	0.2075	<u>0.2404</u>	0.1683	0.1055	0.1213
Global att layer 2	0.1382	0.2195	<u>0.2483</u>	0.1994	0.0883	0.1063
Global att layer 3	0.1449	<u>0.2030</u>	0.1914	0.1874	0.1335	0.1398
Put pan						
Global att layer 0	0.1694	0.1717	0.1571	<u>0.2277</u>	0.1572	0.1169
Global att layer 1	<u>0.2351</u>	0.1885	0.1789	0.1939	0.1182	0.0854
Global att layer 2	<u>0.3367</u>	0.2679	0.1976	0.1149	0.0547	0.0281
Global att layer 3	<u>0.2737</u>	0.2621	0.2127	0.1736	0.0433	0.0346

Figure 2.3: Visualization of temporal attention score of a few samples of the EGTEA dataset when stacking different numbers of global attention layers. The highest temporal attention score in each layer is underlined. In the case of global attention layer 0, initialization using the self-attention method is used. The frame with green background indicates that the backbone model can predict the correct activity class with this clip alone as input.

an example, for the activity “take paper towel”, the global layer 0 initialized using self-attention gives relatively more attention to a background clip. This outcome is expected, as self-attention on each clip’s feature tends to assign high values to clips with similar features. However, with the help of the global attention layer, the highest temporal attention is assigned to the 4th and 5th clip, which contain the necessary information for correct activity recognition and thus, should be emphasized. This provides strong evidence that the use of global features is crucial. The refinement of temporal attention through the stacking of global attention layers is demonstrated to be effective through the examination of several samples, including “turn of the faucet”, “open fridge” and “put pan”. As depicted in this figure, the optimal temporal attention score is achieved when utilizing two global attention layers, with a saturation point reached when stacking more than three layers, aligning with the results presented in Table 2.5.

2.4.7 Experiments on Third-person Dataset

Furthermore, to see how the proposed STAM performs on traditional third-person datasets, we add an experiment using the HMDB51 dataset [KJG⁺11]. HMDB51 is a widely used human motion dataset collected from YouTube that contains 6849 clips divided into 51 action categories. The average duration of each action is about 3s. Following the convention, we report the average result of three train/test splits.

Table 2.6 shows the result comparison on the HMDB51 dataset [KJG⁺11]. Although the performance of our STAM on TSM backbone is not as obvious as that in first-person datasets, clear improvement can still be validated when using I3D and R3D-50 as backbone encoders, which proves that our STAM is generalizable to third-person datasets. Moreover, I3D with STAM on the top achieves better accuracy than CatNet [WPQ20] which also uses I3D as the backbone encoder. This performance validates that refining temporal

attention by stacking global attention layers can generate more reasonable temporal attention.







With our proposed STAM, the performance of TSM (69.4 \rightarrow 70.3), I3D (73.1 \rightarrow 76.1) and 3D ResNet (67.8 \rightarrow 70.5) backbones are all improved. The improvement may not be as significant as in first-person videos, but it strongly proves the effectiveness of the proposed STAM, and demonstrates its generalization ability.







Method	Acc
TSN* [WXW ⁺ 16]	69.4
TLE* [DSVG17]	71.1
CatNet [WPQ20]	75.2
TSM [LGH19]	69.4
I3D [CZ17a]	73.1
R3D-50 [HKS18]	67.8

TSM + STAM	70.3
I3D + STAM	76.1
R3D-50 + STAM	70.5

Table 2.6: Results on the HMDB51 dataset. * indicate the method uses optical flow as input.

Figure 2.4 shows three examples from HMDB51 dataset together with the temporal attention value of each clip when using 6 clips as input. We use one frame to represent each 16-frame clip in this Figure. As can be seen from the examples, the scores produced by the initialization layer tend to highlight some repeated clips, thus resulting wrong prediction. However, after adding global attention layers, the temporal attention gradually shifts from wrong clips to discriminative clips. In this figure, the temporal attention score becomes optimal when stacking 2 global attention layers, and saturates when stacking more layers.

Cartwheel						
Global att layer 0	0.0347	0.1036	<u>0.4078</u>	0.2247	0.0931	0.1362
Global att layer 1	0.0582	0.0810	<u>0.2976</u>	0.2850	0.0689	0.2143
Global att layer 2	0.0497	0.0732	0.2074	<u>0.3894</u>	0.0665	0.2138
Global att layer 3	0.0624	0.0709	0.1695	<u>0.3295</u>	0.0908	0.2769

Brush hair						
Global att layer 0	0.1104	0.1948	0.0703	<u>0.3026</u>	0.2504	0.0715
Global att layer 1	0.1916	<u>0.2679</u>	0.1033	0.2474	0.1363	0.0535
Global att layer 2	0.2210	<u>0.2861</u>	0.1124	0.1467	0.1860	0.0508
Global att layer 3	0.1856	0.1716	0.0961	0.1325	<u>0.3560</u>	0.0581

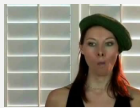
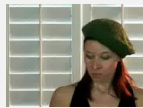
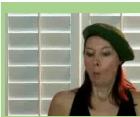
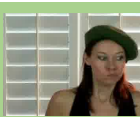
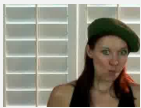
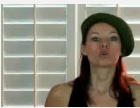
Chew						
Global att layer 0	0.1398	0.1648	0.1306	0.1786	0.1486	<u>0.2366</u>
Global att layer 1	0.1028	0.1029	<u>0.3163</u>	0.2472	0.1069	0.1239
Global att layer 2	0.1117	0.1216	<u>0.2541</u>	0.2286	0.1289	0.1551
Global att layer 3	0.1176	0.1469	0.2002	<u>0.2896</u>	0.1100	0.1357

Figure 2.4: Visualization of temporal attention scores of three samples from the HMDB51 dataset when stacking different numbers of global attention layers. The highest temporal attention score in each layer is underlined. In the case of global attention layer 0, initialization using the self-attention method is used. The frame with green background indicates that the backbone model can predict the correct action class with this clip alone as input.

2.5 Conclusion

In this chapter, we propose a novel Stacked Temporal Attention Module (STAM) that leverages multilateral relations between local-local video clip features and local-global clip features to focus on discriminative clips for more robust first-person activity recognition. Since most of the frames in the video contain irrelevant or repeated information that provides little cue for action recognition, it is necessary to focus on only discriminative clips rather than the background parts. For localizing the discriminative clips, it is essential to leverage the local-global clip relationship, since the global

knowledge is required to determine which clip is more discriminative than others from a more holistic perspective. Our proposed STAM is simple yet effective in generating temporal attention to enhance the aggregation of clip features for first-person activity recognition. As one of the good qualities, the STAM module in this chapter can be integrated with most existing models to improve their performance in activity recognition.

Through conducting extensive experiments on the egocentric (first-person) video datasets namely EPIC-100 and EGTEA, we have demonstrated the effectiveness of our STAM in boosting the performance of multiple backbones. Furthermore, our experiments on the third-person HMDB51 dataset indicate that our STAM is also applicable to general third-person videos. Results on the first-person datasets can best reflect the effectiveness of our proposed module, since first-person videos often contain large camera motion, the discriminative clips are often shorter than that of third-person datasets. As a future direction, we intend to investigate alternative design choices for constructing the global feature and apply the stacked temporal attention to other video-related tasks. Since this method requires full supervision where the labels are very labor-intensive to collect, we also plan to explore the action recognition problem in a weaker supervised setting. For example, we will demonstrate the unsupervised action recognition in the target domain in the next chapter, and a more open-set setting where no pre-defined action labels are used in Chapter 4.

Chapter 3

Domain Adaptive Activity Recognition by Exploring Relations among Different Modalities

In this chapter, compared to the supervised action recognition setting where the assumption is that enough labels are provided to train a decent action recognition model, we relax this restriction on the problem setting, focusing on a more challenging setting closer to the real-world application, *i.e.*, domain adaptation, where no labels on the new environment are provided for fine-tuning the model trained on known environments. We realize a state-of-the-art domain adaptation model by leveraging the relations hidden among different modalities, since videos are naturally accompanied by multiple modalities such as optical flow and audio.

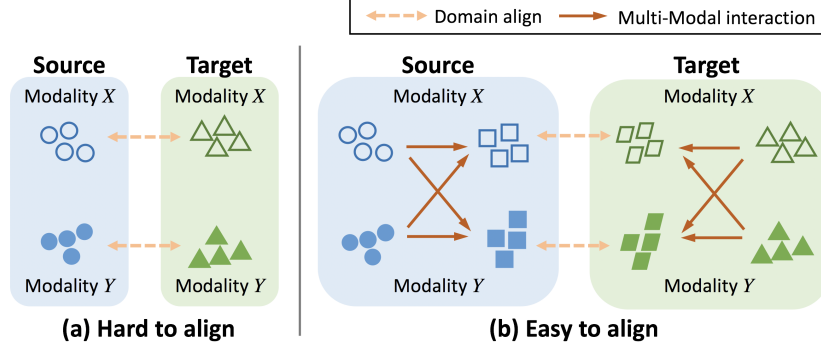


Figure 3.1: Different from existing UDA works that directly align the multi-modal inputs (a), we find that it is more effective to first enhance the transferability of each modality by cross-modal interaction, and then perform cross-domain alignment (b).

3.1 Introduction

The goal of unsupervised domain adaptation (UDA) models is to learn features on the source dataset that can be effectively utilized on the target dataset as well. Given its capacity to reduce the need for large-scale labeling, UDA has been widely studied for tasks such as image recognition [LCWJ15, SS16, THSD17, YWCH19], semantic segmentation [ZYKW18, CLS20] and object detection [CLS⁺18, CLZ⁺21].

Video data is more intricate than image data due to the additional temporal dimension, and the domain gap not only exists in the visual differences of the environments but also in the variations in motion when different individuals perform the same activity. This presents a challenge for the direct application of image-based domain adaptation methods for domain adaptive activity recognition [CKA⁺19, JNDV18]. One approach to addressing this complexity is to incorporate additional modality information, such as optical flow and audio.

In addition to directly combining multi-modal inputs [PCAN20], recent works employed self-supervised modality alignment to implicitly learn prop-

erties of source and target data [MD20, SZY⁺21, KTZ⁺21]. However, since the objectives of cross-modal alignment and cross-domain alignment are not perfectly consistent, simultaneously aligning *modality* and aligning *domain* can distract the primary learning objective, *i.e.*, minimizing the *domain* discrepancy.

To better exploit the cross-modal knowledge for enhancing the transferability of each modality, we propose to allow cross-modal interaction first for better cross-domain alignment. Each modality possesses distinct characteristics, thus the transferability of features across domains is different and complementary across modalities. For instance, in the activity “wash plate” on the target domain, the audio modality is more transferable in identifying the verb “wash” of the activity due to the similarity of water sounds across domains. On the other hand, the RGB modality may not perform as well in recognizing the verb on the target domain, but it can effectively recognize the noun “plate” on the target domain by utilizing its domain-transferable visual knowledge. If it is possible for these two modalities to interact and share their unique domain-transferable knowledge, both modalities can improve their transferability and accurately determine the activity “wash plate”. Based on this observation, we leverage this **cross-modal complementarity** and propose a Mutual Complementarity (MC) module that allows each modality to refine its feature by absorbing the transferable knowledge from other modalities, thus *the transferability of all modalities can be enhanced*.

The incorporation of multiple modalities also brings the aspect of **cross-modal consensus** into consideration. Since domain shift is frequently accompanied by changes in the scenario background, identifying and emphasizing more transferable foreground objects is crucial. Instead of applying spatial attention like previous works [LDZ⁺20, WLY⁺19] which introduce additional parameters that also suffer from domain gaps, we propose to use a correlation-based spatial consensus operation, which does not require any

additional parameters. By leveraging multi-modal features, our developed cross-modal Spatial Consensus module (SC) is able to identify and emphasize transferable regions that are consistent across different modalities. Our approach has been demonstrated to be more effective for domain adaptation in comparison to spatial attention methods in experimental evaluations.

We perform experiments on the widely-used UCF-HMDB dataset and EPIC-Kitchens-55 dataset. Our experiments reveal that through the integration of cross-modal knowledge, our proposed method demonstrates significant superiority over existing state-of-the-art methods. Additionally, our method is able to achieve a considerable improvement on the EPIC-Kitchens-100 dataset, which encompasses challenging fine-grained activities.

The main contributions of this chapter are summarized as follows:

- We propose a novel model to enhance multi-modality features for domain adaptive activity recognition. To our best knowledge, this is the first work to consider cross-modal interaction for increasing the feature transferability across domains.
- We propose to use a correlation-based operation to evaluate the transferability of spatial locations, which is proved to be simple and effective compared with spatial attention in the context of domain adaptation.
- Our proposed model achieves state-of-the-art performance on multiple datasets, including the challenging EPIC-Kitchens-100 dataset with fine-grained activities.

3.2 Related Works

3.2.1 Unsupervised Domain Adaptation (UDA) other than activity Recognition

For solving the domain gap problem which exists widely in various applications such as object recognition [GUA⁺16, SS16, SUH17], image segmentation [HHK18, CLS20, ZDG17, ZYKW18, GHXL21], and natural language understanding [SLM⁺18, PM13, TBL⁺19], domain adaptation has been extensively studied especially in recent years. Using a model trained on the source domain, the task of domain adaptation aims to improve its performance on the target domain. Some works focus on the input level and try to mitigate the domain gap by modifying the source input to become similar to the target domain via approaches like image-to-image translation [BSD⁺17, MKK⁺18]. Another direction focuses on the representation level and aligns the representations with Maximum Mean Discrepancy (MMD) [LZWJ17] or adversarial training [THSD17]. Very recently, a new trend of self-supervised training dominates the field of domain adaptation [CLS20, KWY⁺20, STDE19]. For example, Kang *et al.* [KWY⁺20] focuses on the task of domain adaptive semantic segmentation, performs self-supervised training from the input level, and builds the pixel-level cycle association between source and target pixel pairs. In recent years, incorporating multiple modalities for UDA has been investigated for the task of emotion recognition and image retrieval [QYX18]. These methods designed cross-modality attention-based single and multi-modal discriminators, proving that the use of multiple modalities can be more robust to domain shift compared with using only one single modality.

3.2.2 Action Recognition and its UDA

There has been a remarkable advance in the techniques of action recognition because of the advance of deep learning [CZ17b, FFMH19, LJS⁺20, HSS20,

[MXH⁺20a](#), [YHSS21](#)]. Recent methods use multiple modalities such as RGB frames, optical flow, and audio as input, and demonstrate the advantage of each modality [[KNZD19](#)]. A huge amount of research attention has also been attracted into domain adaptive action recognition, in par with the deep-learning brought advances in activity recognition. Most research works focus on domain adaptation across different viewpoints [[KDLF17](#)]. These works aim to adapt to the geometric transformations of a camera in the same environment, with optional additional information like the human pose [[LLC17](#)] and temporal correspondence [[SGS⁺18](#)].

Another group of research tries to adapt the models for activity recognition in different environments. Some of these methods try to design hand-crafted features for better aligning source and target domains [[ZS13](#), [FDdCW⁺11](#)]. More recently, most works begin to process the RGB frames base on deep neural networks [[ACDY20](#), [LLL⁺19a](#), [CLB⁺20](#)]. Recently, inspired by image-based domain adaptation works, several works [[MD20](#), [PCAN20](#), [ZXZO21](#), [SLLG20](#)] explored the use of multiple modalities (RGB and flow) for domain adaptive activity recognition. In the work of [[PCAN20](#)], the authors temporally align the different modalities independently and finally fuse modalities during inference. In [[MD20](#), [KTZ⁺21](#), [SZY⁺21](#)], self-supervised alignment of modality is adopted. However, self-supervised modality alignment has a different learning target with domain adaptation, and simultaneously learning a model with two targets distracts the model from the primary task – minimizing the domain discrepancy.

In this work, different from all previous works, we increase the transferability and allow cross-modal interaction by re-evaluating semantic transferability based on information from other modalities. We further use the cross-modal spatial consensus and design a consensus module to find the most transferable regions. The major difference between this chapter and previous methods [[MD20](#), [KTZ⁺21](#), [SZY⁺21](#)] lies in that our cross-modal interaction

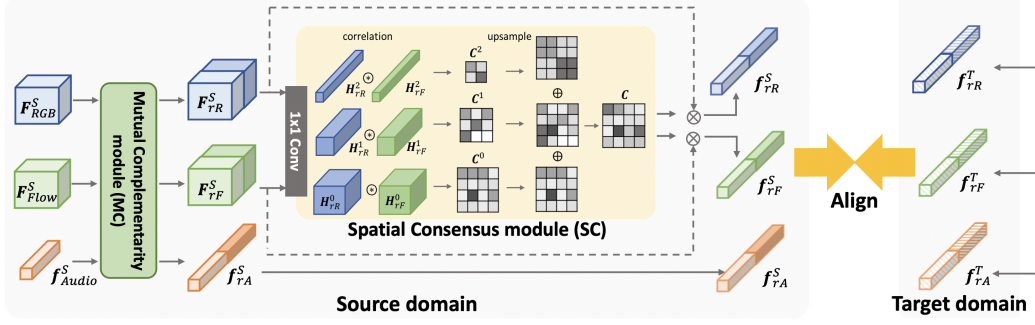


Figure 3.2: Overview of the proposed CIA model. We showcase three modalities RGB, Flow and Audio as input but it can be easily extended to add other modalities such as depth. In the figure, \oplus denotes element-wise summation, \otimes is element-wise multiplication, and \otimes means the correlation operation that calculates the Pearson correlation coefficient on each spatial position.

does not add additional self-supervision loss. By this means, the interaction between modalities can be optimized to purely improve the transferability of source and target domains.

3.3 Proposed Method

This chapter focuses on effectively leveraging the cross-modal complementarity and cross-modal consensus for domain adaptive activity recognition. As the proposed method, we design a Cross-modal Interactive Alignment (CIA) model. This model contains two major steps, where the first step supplements each modality with cross-modal transferable knowledge by mutual semantic refinement and the second step emphasizes transferable regions by exploiting the consensus of multiple modalities.

Figure 3.2 depicts the overview of the proposed CIA model. In both source domain S and target domain T , for each modality of RGB, Flow, and Audio, we first use backbone (omitted in the figure) networks to encode the input into frame-level features $F_{RGB}^S, F_{Flow}^S, f_{Audio}^S, F_{RGB}^T, F_{Flow}^T$ and f_{Audio}^T .

Note that when the operation in both domains are the same, we omit the notation of the domain identifier in the following part of this section. We then use two modules, named Mutual Complementarity module (MC) and Spatial Consensus module (SC), to allow feature interaction for exploiting the cross-modal complementarity and the cross-modal consensus, respectively. The MC module exploits cross-modal complementarity by enabling one modality to receive transferable semantic knowledge from other modalities, utilizing two gating functions (Sec. 3.3.1). Then the SC module emphasizes transferable spatial regions which share consensus among all modalities by a multi-scale correlation operation (Sec. 3.3.2). Finally, we adopt adversarial feature alignment on the SC outputs to minimize the discrepancy between source and target domains.

3.3.1 Mutual Complementarity (MC) Module

Different modalities excel in their unique perspectives for perceiving the input, and the MC module aims to leverage this cross-modal complementarity to enhance the transferability of each modality by selecting and absorbing domain-transferable knowledge from other modalities. Transferable semantic knowledge lies in the feature channels [ZKL⁺16], however, gaps between modalities prevent direct channel-wise fusion methods like max-pooling or summation. In our proposed MC, we instead use a “summarize and re-evaluate” operation to leverage cross-modal transferable information.

Figure 3.3 depicts the proposed MC by showcasing the workflow of modality M . The output of MC is a transferability-refined feature of modality M $\mathbf{F}_{rM} \in \mathbb{R}^{2c \times h \times w}$, which is the concatenation of two parts: a cross-refined feature \mathbf{F}_{cM} and a self-refined feature \mathbf{F}_{sM} .

\mathbf{F}_{cM} represents the feature of modality M refined by transferable knowledge from other modalities. For getting \mathbf{F}_{cM} , we first apply global average pooling on features of other modalities and concatenate them to obtain

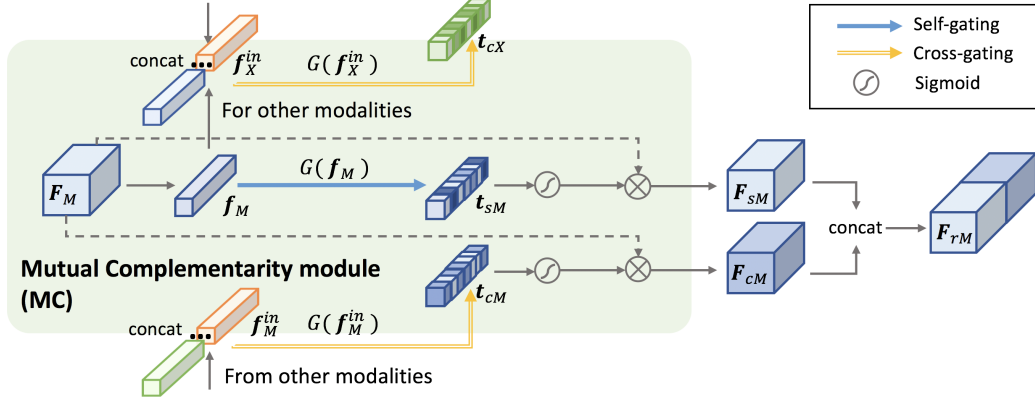


Figure 3.3: The Mutual Complementarity module (MC) showcased using modality M . M could be any modalities of RGB, Flow, and Audio, and also can be extended to other modalities if available, *e.g.*, depth.

a cross-modal knowledge representation \mathbf{f}_M^{in} . With \mathbf{f}_M^{in} , we summarize the domain-transferable knowledge and re-evaluate the semantic transferability of modality M by a cross-gating function [HSS18]:

$$\mathbf{t}_{cM} = \sigma \mathbf{W}_2^{in} (\delta(\mathbf{W}_1^{in} \mathbf{f}_M^{in})), \quad (3.1)$$

$$\mathbf{F}_{cM} = \mathbf{F}_M \cdot \mathbf{t}_{cM}, \quad (3.2)$$

where \mathbf{W}_1^{in} , \mathbf{W}_2^{in} are weight matrices, \cdot is channel-wise multiplication, σ and δ denotes the sigmoid and ReLU activations, respectively. Here \mathbf{t}_{cM} is the re-evaluation of semantic transferability of modality M by using cross-modal knowledge. \mathbf{t}_{cM} serves as the “advice” from other modalities to emphasize the channels of \mathbf{F}_M by channel-wise multiplication.

Although this gating mechanism is simple, it can learn nonlinear interaction between channels, while allowing multiple channels to be emphasized during the re-evaluation. This helps the gating operations to first summarize domain-transferable knowledge (using \mathbf{W}_1^{in}) and then re-weight the channels of \mathbf{F}_M utilizing the summarized knowledge (using \mathbf{W}_2^{in} and channel-wise multiplication).

When receiving complementary knowledge from other modalities, it is also important for modality M to preserve unique information and modality characteristics of itself. Thus, in addition to cross-gating, we use a self-gating operation to perform a self re-evaluation of modality M :

$$\mathbf{t}_{sM} = \sigma \mathbf{W}_2^M(\delta(\mathbf{W}_1^M \mathbf{f}_M)), \quad \mathbf{F}_{sM} = \mathbf{F}_M \cdot \mathbf{t}_{sM}, \quad (3.3)$$

To summarize domain-transferable knowledge while preventing the domain adaptation model from overfitting on the source domain, the MC only introduces a small number of model parameters by leveraging bottleneck during gating. In other words, we reduce the dimension by a ratio r via making $\mathbf{W}_1 \in \mathbb{R}^{\frac{c}{r} \times c}$ and $\mathbf{W}_2 \in \mathbb{R}^{c \times \frac{c}{r}}$. Finally, we get the transferability-refined feature of modality M by fusing the two refined features \mathbf{F}_{sM} and \mathbf{F}_{cM} via concatenation:

$$\mathbf{F}_{rM} = \text{Concat}(\mathbf{F}_{sM}, \mathbf{F}_{cM}). \quad (3.4)$$

We show the analysis of model parameters and computational complexity in the supplementary.

3.3.2 Spatial Consensus (SC) Module

To further enhance feature transferability by focusing on the most transferable spatial regions (*e.g.* foreground objects), previous works mainly use the spatial attention mechanism [WLY⁺19, LDZ⁺20, HCLS18]. However, this introduces additional model parameters which will also be affected by domain shift. Different from spatial attention, we propose a Spatial Consensus (SC) module to highlight the transferable regions that have shared consensus among modalities.

Our idea to find transferable locations is letting multiple modalities work with “collective wisdom”. Since features \mathbf{F}_{rR} and \mathbf{F}_{rF} encode different information, we first map these features into the same latent space to get

transferability estimations from their own perspective. Then we compute the feature similarity using a correlation operation to measure whether two modalities share the same opinion about spatial transferability. For each location, the feature similarity is high only if two modalities both find this location to be transferable.

Since transferable regions vary in size in different samples, we compute the correlation of feature maps at different scales [LDG⁺17]: the features \mathbf{H}_{rR} and \mathbf{H}_{rF} are first downsampled by a factor of 2^k times, resulting in two groups of feature maps $\{\mathbf{H}_{rR}^0, \mathbf{H}_{rR}^1, \dots, \mathbf{H}_{rR}^k\}$ and $\{\mathbf{H}_{rF}^0, \mathbf{H}_{rF}^1, \dots, \mathbf{H}_{rF}^k\}$. For each scale k , we compute the Pearson correlation coefficient on each spatial position (i, j) as:

$$\mathbf{C}^{k,(i,j)} = \frac{\mathbf{H}_{rR}^{k,(i,j)} * \mathbf{H}_{rF}^{k,(i,j)}}{\|\mathbf{H}_{rR}^{k,(i,j)}\|^2 \times \|\mathbf{H}_{rF}^{k,(i,j)}\|^2}, \mathbf{C}^k \in \mathbb{R}^{\frac{w}{2^k} \times \frac{h}{2^k}} \quad (3.5)$$

where $*$ indicates dot product. It is important that SC contains only a small number of parameters so that most of the representation is learned in the MC while also preventing overfitting. To this end, we choose to use correlation instead of spatial attention [WPLK18].

Finally, all the correlation maps $\{\mathbf{C}^0, \mathbf{C}^1, \dots, \mathbf{C}^k\}$ are upsampled to match the size as \mathbf{F}_{rR} and then summed together to form a consensus map \mathbf{C} . The consensus map \mathbf{C} is then used as a spatial weight map for the weighted average pooling of \mathbf{F}_{rR} and \mathbf{F}_{rF} . We also add residual connections following [WJQ⁺17, HZRS16], forming feature vectors \mathbf{f}_{rR} and \mathbf{f}_{rF} . Since MC already involves audio information and \mathbf{f}_{rA} does not contain spatial dimensions, \mathbf{f}_{rA} is not used in this module. During training, the SC module will encourage the network to extract features such that the spatial correlation becomes higher for locations more helpful for domain alignment.

3.3.3 Adversarial Domain Alignment

We apply adversarial domain alignment on three transferability enhanced features \mathbf{f}_{rR} , \mathbf{f}_{rF} and \mathbf{f}_{rA} , individually. Denote the two-layer MLP-based discriminator as D , the discriminator loss can be written as:

$$\begin{aligned} \mathcal{L}_{fd} = \sum_{M \in \{rR, rF, rA\}} \sum_{\mathbf{f}_M \in S, T} & -d \log(D_M(\mathbf{f}_M)) \\ & - (1 - d) \log(1 - D(\mathbf{f}_M)) \end{aligned} \quad (3.6)$$

where d is the binary domain label, S, T denotes the source and target domains respectively, and \mathbf{f}_M represents one of the features in $\{\mathbf{f}_{rR}, \mathbf{f}_{rF}, \mathbf{f}_{rA}\}$.

We average the frame-wise features to form video-level features \mathbf{v}_{rR} , \mathbf{v}_{rF} and \mathbf{v}_{rA} and fuse them as \mathbf{v}_{mm} . The domain alignment is also done on the video-level features \mathbf{v}_{rR} , \mathbf{v}_{rF} and \mathbf{v}_{rA} and its loss is denoted as \mathcal{L}_{vd} .

On the source domain, we apply the standard classification loss on the fused video-level feature \mathbf{v}_{mm} :

$$\mathcal{L}_y = - \sum_{\mathbf{v}_{mm} \in S} \mathbf{y} \log P(G_M(\mathbf{v}_{mm})), \quad (3.7)$$

where G_M represents the linear activity classifier for the corresponding feature.

As a result, our full loss function is a combination of \mathcal{L}_y , \mathcal{L}_{fd} and \mathcal{L}_{vd} :

$$\mathcal{L} = \lambda_y \mathcal{L}_y + \lambda_{fd} \mathcal{L}_{fd} + \lambda_{vd} \mathcal{L}_{vd} \quad (3.8)$$

3.4 Experimental Results

3.4.1 Datasets

We validate our proposed CIA model on three representative domain adaptive activity recognition datasets: UCF-HMDB [KJG⁺11, SZS12] (**U-H**) is

one widely used dataset that contains 12 action classes. We use the full version [CKA⁺19] in our experiments. $\mathbf{H} \rightarrow \mathbf{U}$ indicates the source dataset is HMDB while the target dataset is UCF, and vice versa. We also use the EPIC-Kitchens-55 (E55) as another benchmark dataset. To make a fair comparison with [MD20, SZY⁺21, KTZ⁺21], we follow the same setting as [MD20]. Class-wise action recognition accuracy is used as the evaluation metric on these two datasets.

Additionally, EPIC-Kitchens-100 [DDF⁺20] (E100) is a newly released dataset with fine-grained activities taken from the first-person perspective. This dataset is extremely challenging because (1) source and target activities are performed by different individuals in different kitchens. (2) The first-person viewpoint often makes the activity happen in a non-salient region, and (3) the annotation is fine-grained. There are 16115/26115 training videos for source/target domains and 7906 clips as the target-validation split. 97 verb classes and 300 noun classes form a total of 3369 fine-grained activity classes. We further add experiments on this dataset since its large-scale and fine-grained property makes it more suitable for analyzing model performance. Following the protocol in [DDF⁺20], we use the accuracy of verb, noun, and action as the evaluation metric.

3.4.2 Implementation Details

For a fair comparison, we use two backbones for feature extraction: I3D [CZ17b] pretrained on Kinetics and TBN [KNZD19] pretrained on Kinetics then fine-tuned on the source training set of the according dataset. The MC processes the feature with dimension $c = 1024$, and the ratio for gating bottleneck is $r = 16$. We use either average or concatenate as the late fusion methods based on datasets. For all experiments, we train the model on 4 NVIDIA-V100 GPUs.

U-H dataset: We first extract features using I3D [CZ17b] pretrained on Kinetics. For each action clip, we extract features from 25 uniformly sampled frames. We use the same strategy as TSN [WXW⁺16] to choose 5 frames from 25 frames. For training our STAM model, we apply Adam optimizer [KB14] with learning rate 3e-3. We empirically choose $\lambda_y = 1$, $\lambda_{vd} = 1$ and $\lambda_{fd} = 0.5$ for the experiments.

E55 dataset: On E55 dataset, we train I3D backbone together with our STAM model using Adam optimizer [KB14] with learning rate 1e-4. We uniformly sample 16 frames as the inputs. We empirically choose $\lambda_y = 1$, $\lambda_{vd} = 1$ and $\lambda_{fd} = 0.5$ for the experiments.

E100 dataset: For the experiments using I3D as backbone, we apply the same training method as for the E55 dataset.

For the experiments that use TBN [KNZD19] as backbone, we first extract features using TBN fine-tuned on the source dataset following [DDF⁺20]. For each action clip, we extract features from 25 uniformly sampled frames. We use the same strategy as TSN [WXW⁺16] to choose 5 frames from 25 frames. For training the model, we apply Adam optimizer [KB14] with learning rate 1e-4. Specifically, when using TRN [ZAOT18] as the temporal aggregation method, we train the model using SGD optimizer with learning rate 3e-3.

3.4.3 Comparison with State-of-the-art

We compare our CIA model with the following methods:

- **Multi-modal UDA activity recognition methods.** We compare with three recent methods MM-SADA [MD20], STCDA [SZY⁺21] and Kim *et al.* [KTZ⁺21]. These methods show state-of-the-art performance in the UDA activity recognition task.

Modality	Backbone	Method	U→H	H→U
RGB	R-TRN	TA ³ N [CKA ⁺ 19]	78.33	81.79
	R-TRN	TCoN [PCAN20]	87.24	89.06
	I3D	SAVA [CSSH20]	82.20	91.20
	I3D-TRN	TA ³ N [CKA ⁺ 19]	82.78	91.77
Flow	I3D-TRN	TA ³ N [CKA ⁺ 19]	82.50	90.89
R+F	I3D	Avg [◊]	83.61	91.07
	I3D	G-blend [WTF20]	84.72	91.24
	I3D	MMTM [JSIK20]	85.83	92.47
	I3D	MM-SADA [MD20]	84.20	91.10
	I3D	STCDA [SZY ⁺ 21]	83.10	92.10
	I3D	Kim <i>et al.</i> [KTZ ⁺ 21]	84.70	92.80
	I3D	CIA source only [◊]	86.11	92.47
	I3D	CIA (Ours) [◊]	88.33	94.05
	I3D	Concat [*]	86.11	92.99
	I3D	CIA source only [*]	85.83	93.52
	I3D	CIA (Ours) [*]	90.56	94.22
	I3D-TRN	TA ³ N [CKA ⁺ 19] [*]	89.17	92.81
	I3D-TRN	CIA (Ours) [*]	89.72	93.17
	I3D-TRN	CIA +TA ³ N [*]	91.94	94.57
	I3D	CIA target only [*]	96.83	99.12

Table 3.1: Performance comparison on the UCF-HMDB (**U-H**) dataset. We show the input modality and the backbone used by each method for better comparison. [◊] refers to averaging the outputs from each modality classifier, while ^{*} means concatenate features of different modalities. Under the same experiment setting, our method can clearly outperform previous methods.

Method	D1→D2	D1→D3	D2→D1	D2→D3	D3→D1	D3→D2	mean
Ours Source only	43.2	42.5	43.0	48.0	43.0	55.5	45.9
MMD [LCWJ15]	46.6	39.2	43.1	48.5	48.3	55.2	46.8
AdaBN [LWS+18]	47.0	40.3	44.6	48.8	47.8	54.7	47.2
MCD [SWUH18]	46.5	43.5	42.1	51.0	47.9	52.7	47.3
DAAA [JNDV18]	50.0	43.5	46.5	51.5	51.0	53.7	49.4
MM-SADA [MD20]	49.5	44.1	48.2	52.7	50.9	56.1	50.3
Kim <i>et al.</i> [KTZ+21]	50.3	46.3	49.5	52.0	51.5	56.3	51.0
STCDA [SZY+21]	52.0	45.5	49.0	52.5	52.6	55.6	51.2
CIA (Ours)	52.5	47.8	49.8	53.2	52.2	57.6	52.2
Ours target only	71.6	73.6	63.3	73.6	63.3	71.6	69.5

Table 3.2: Performance comparison on the EPIC-Kitchens-55 (**E55**) dataset.

- **Single-modal UDA activity recognition methods** [CKA+19, CSSH20, PCAN20, LCWJ15, LWS+18, SWUH18, JNDV18]. For better comparison, we follow [DDF+20] to enable **TA³N** [CKA+19] with multi-modality input and use TRN [ZAOT18] on the backbone for temporal feature fusion.
- **Multi-modal fusion methods for other tasks**. To better evaluate our CIA’s ability on using multi-modal information in the scope of domain adaptation, other than direct fusion via average (**Avg**) or concatenation (**Concat**), we add comparison with previous multi-modal fusion methods **G-blend** [WTF20] and **MMTM** [JSIK20]. Since [WTF20, JSIK20] are not originally designed for domain adaptation, we use their method on the same adversarial alignment framework with our method for a fair comparison.

Results on U-H dataset are shown in Table 3.1. From the table, because of the inherent difficulty of video data, multi-modal methods generally surpass single modality methods [CKA+19, PCAN20, CSSH20]. Meanwhile, previous multi-modal fusion works **G-blend** [WTF20] and **MMTM** [JSIK20] do not perform well in the domain adaptation setting, suggesting that our proposed CIA model better suits the task of domain adaptation. Our method

significantly outperforms previous state-of-the-art multi-modal works **MM-SADA**, **STCDA** and **Kim *et al.***. Compared with **Kim *et al.***, we can increase the accuracy from 84.70 to 88.33 on $\mathbf{U} \rightarrow \mathbf{H}$ and 92.80 to 94.05 on $\mathbf{H} \rightarrow \mathbf{U}$. This indicates the superiority of our CIA model in leveraging multi-modal interaction compared with self-supervised learning.

We also validate different late fusion methods by comparing average[◊] and concatenation^{*}. We found that using concatenation for late modality fusion can be more helpful. Using TRN [ZAO18] as a more sophisticated temporal aggregation method, our method outperforms TA³N on both datasets. Since our method can be flexibly fitted into any domain adaptation framework, we can further enhance TA³N by adding our model, achieving 91.94 and 94.57 on the two datasets.

Results on E55 dataset are illustrated in Table 3.2. We average the outputs of individual modality classifiers as the late fusion method for a fair comparison with prior works. Using cross-modal self-supervision, MM-SADA, STCDA, and Kim *et al.* cannot perform as well as our proposed method. This proves our assumption that simultaneously optimizing cross-modal alignment and cross-domain alignment can distract the model from minimizing the domain gap. However, by interacting before alignment, our method can better leverage the cross-modal complementarity and cross-modal consensus, thus boosting the mean accuracy by up to 1% compared with the previous state-of-the-art.

Results on E100 dataset Table 3.3 demonstrates the performance comparison with state-of-the-art methods on the challenging **E100** validation set. We average the scores of each modality for late fusion when implementing methods on the I3D backbone, while we use concatenation for methods on other backbones. Using RGB and Flow modalities and the same backbone, our proposed method performs favorably against the state-of-the-art method

Modality	Backbone	Method	Verb	Noun	Action
R+F	I3D	Source only	39.28	22.28	11.62
	I3D	MM-SADA [MD20]	40.41	23.92	12.80
	I3D	Source only	40.17	22.89	12.27
	I3D	CIA (Ours)	42.35	24.49	14.25
	TBN	Source only	42.41	27.26	16.03
	TBN	DAAA [JNDV18]	42.99	27.38	16.32
	TBN	Source only	42.98	27.49	16.44
	TBN	CIA (Ours)	43.93	27.54	17.01
	TBN-TRN	Source only	43.78	26.65	16.70
	TBN-TRN	TA ³ N [CKA ⁺ 19]	44.88	27.41	17.39
	TBN-TRN	Source only	44.12	27.12	16.86
	TBN-TRN	CIA (Ours)	45.23	27.75	18.02
	TBN-TRN	Source only	46.67	27.57	19.00
	TBN-TRN	TA ³ N [CKA ⁺ 19]	47.43	28.40	19.42
	TBN-TRN	Source only	47.69	28.48	19.61
	TBN-TRN	CIA (Ours)	48.34	29.50	20.30
R+F+A	TBN	Source only	47.10	28.30	18.66
	TBN	DAAA [JNDV18]	47.96	29.08	19.19
	TBN	Source only	48.22	29.86	19.73
	TBN	CIA (Ours)	49.08	30.36	20.49

Table 3.3: Performance comparison on the EPIC-Kitchens-100 (**E100**) validation set. R, F and A refers to RGB, Flow and Audio modalities, respectively. We show each method together with its source only performance in the row above. Under the same experiment setting, our method can clearly outperform previous methods.

MM-SADA [MD20] by 1.45% in terms of the accuracy of activity. When using RGB, Flow, and audio modalities, our method can show more significant improvements over previous works on all of the verb, noun, and action metrics.

3.4.4 Visualization

To better understand the proposed CIA model, in Figure 3.4 we show the Grad-CAM [SCD⁺17] visualizations of activation maps before and after cross-modality feature refinement by the MC module. From these cases we can clearly see the benefit of feature interaction with other modalities: in (a-1) and (a-2), other modalities help the RGB modality to put more focus on the hand by suppressing the attention on other objects. In (b-1), the refined Flow modality transfers its focus from foot to hand, and in (b-2) from left hand to right hand. These examples strongly prove that cross-modal transferable knowledge helps each modality to perform better in the target domain.

We also visualize the activation maps after the SC module to qualitatively evaluate its effectiveness. In the activity “put down spoon” shown in Figure 3.5(b), the RGB modality is guided by other modalities to ignore the tap, and the refined Flow feature becomes more focused in the center. And finally, our SC module can find the best focus by taking advantage of consensus from all modalities.

Figure 3.6 shows the t-SNE [VdMH08] visualization of the feature spaces produced by TA³N (a) and TA³N + CIA(ours) (b) on **U-H** dataset. Our CIA increased the accuracy of TA³N from 89.17 to 91.94, and the domain alignment is more visible, especially in the zoomed-in area, showing that our CIA increases feature transferability.

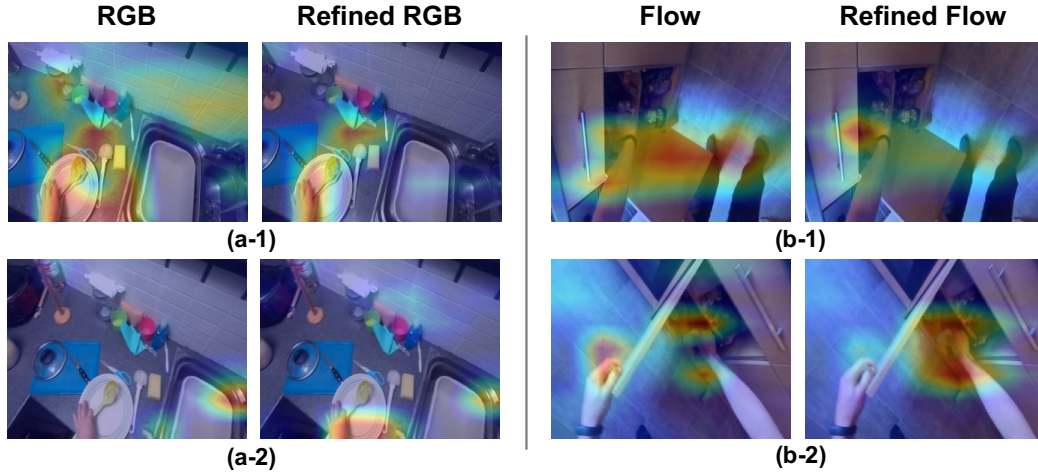


Figure 3.4: Grad-CAM [SCD⁺17] visualizations of features before and after cross-modality feature refinement by MC. The ground-truth activities are: (a-1) take spoon, (a-2) move spoon, (b-1) take garlic, (b-2) take oil. (a-1) and (a-2) show RGB activation maps (left) and the activation map of RGB modality refined by other modalities (right). Similarly, (b-1) and (b-2) depict the activation maps of the Flow modality alone and Flow refined by other modalities.

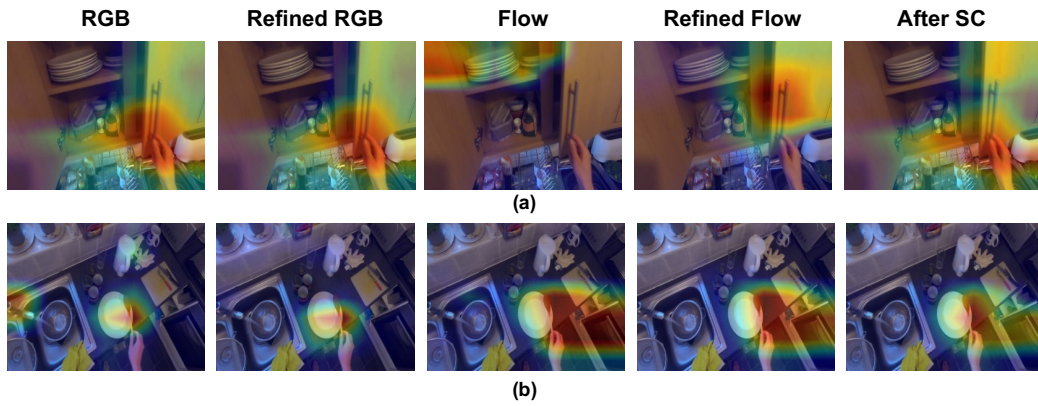
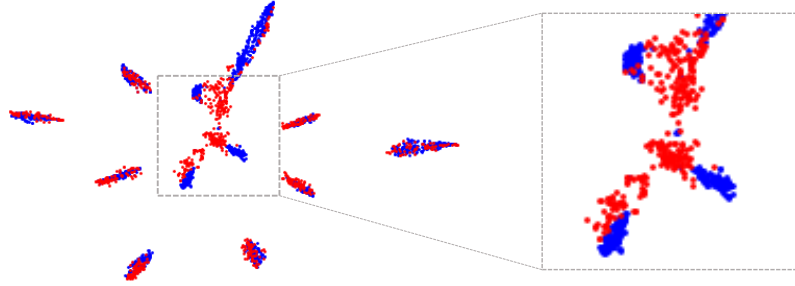
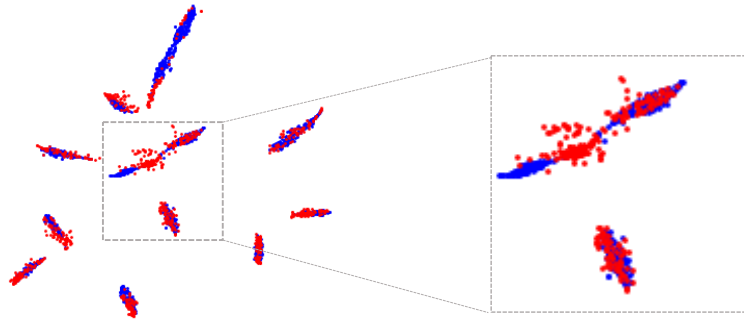


Figure 3.5: Grad-CAM [SCD⁺17] visualizations of RGB, refined RGB, Flow, refined Flow, and fused modality after SC. The ground-truth activity labels are: (a) open cupboard, (b) put down spoon.



(a) TA^3N



(b) $\text{TA}^3\text{N} + \text{CIA(ours)}$

Figure 3.6: t-SNE plots of feature spaces produced by TA^3N (a) and $\text{TA}^3\text{N} + \text{CIA}$ (b). Source is shown in blue and target in red. Our method better aligns source and target domains.

3.4.5 Ablation Study

Contribution of each module In this section, we conduct an ablation study on the **E100** validation set to examine the contribution brought by each module. We test our method w/o the MC or SC module, and also test whether to use self- or cross-refined features within the MC module.

Results can be seen in Table 3.4. Compared with the base setting (1st row), both self-refinement (2nd row) and cross-refinement (3rd row) benefit from the “summarize and re-evaluation” operation, while combining the

	MC	SC	Verb	Noun	Action
	×	×	47.96	29.08	19.19
Self-refinement	×	×	48.01	29.31	19.56
Cross-refinement	×	×	48.48	29.48	19.67
✓	×	×	48.62	29.96	19.98
×	✓	✓	48.66	29.79	19.83
✓	✓	✓	49.08	30.36	20.49

Table 3.4: Ablation study on Mutual Complementarity module (MC) and Spatial Consensus module (SC) of our CIA model.

self- and cross-refinement our MC (4th row) gets a more obvious increase in accuracy. This strongly proves that self- and cross-refinement provide mutual promotion to leverage multi-modal transferable information for better domain adaptation. With only MC or SC, the performance is not favorable against their combined version, indicating that our MC and SC can cooperate well to leverage both cross-modal complementarity and consensus for minimizing the domain gap.

Different design options of SC We also test different design options of our proposed SC. The SC module aims to spatially re-weight the features based on the transferability of each location. We compare with following design options for validation of our proposed SC:

- the most widely adopted feature fusion methods: spatial max pooling (**Max**) and average pooling (**Avg**).
- spatial attention mechanisms based methods: Other than these direct fusion methods, we consider two methods based on spatial attention mechanisms, one for general purpose (**Att** [WPLK18]) and one for domain adaptation (**TADA** [WLY+19]), to generate a spatial attention map for each modality. Other than using both modalities to generate the spatial map, recent researches [CWAS19, ZS19] found that the Flow

Setting	Module	Verb	Noun	Action
Source only	Avg	47.10	28.30	18.66
	Att [WPLK18]	47.32	28.85	19.21
	SC	47.85	29.18	19.55
Domain Adaptation	Avg	47.96	29.08	19.19
	Max	48.11	29.59	19.48
	Att [WPLK18]	48.08	29.46	19.39
	TADA [WLY ⁺ 19]	47.79	29.69	19.59
	Att*	48.29	29.56	19.62
	SC †	48.39	29.70	19.62
	SC	48.66	29.79	19.83
Action Recognition	Avg	72.43	51.36	40.90
	Att [WPLK18]	72.89	53.00	42.20
	SC	73.09	52.50	42.28

Table 3.5: Performance comparison of our SC module with other approaches on the **E100** validation set.

modality is stronger in encoding motion information and thus used Flow as the pivot to guide other modalities. We also experiment using a similar setting where we use Flow attention to guide the RGB attention (**Att***) as an additional comparison baseline. Weighted average is used to fuse the features based on the attention maps.

- SC and its simplified version: **SC †** is a simplified version of our SC which computes the correlation of feature maps only at a single scale. We construct this baseline to validate the effectiveness of multi-scale correlation.

Table 3.5 shows the comparison on the **E100** validation set. In the domain adaptation setting, simply replacing SC with max or average pooling on each spatial location negatively affects the performance. This indicates that max and average pooling do not do well in putting the focus on the

transferable regions. The usefulness of multi-scale correlation compared with single-scale correlation is proved, as SC can outperform SC \dagger . Without fully exploiting the multi-modal knowledge, Att and TADA with adversarial alignment cannot find transferable regions as good as our SC. Our SC gets the best performance among these options in the source-only and domain adaptation settings, showing that the spatial consensus among modalities is more domain-invariant.

Due to the lack of labels on the target domain, we cannot show target-only results. Instead, we show an “action recognition” setting by both training and testing models on the source domain. From Table 3.5, Att outperforms our SC in Noun accuracy since it can learn modality-specific spatial weights when no domain gap exists. From the comparison under different settings, we can see that when the domain gap hinders the learning of spatial weight, generating modality-specific spatial weight becomes even more challenging. In this case, our consensus-based SC shows superiority in highlighting transferable regions. However, when no domain gap exists, our SC becomes sub-optimal as we cannot emphasize different regions for different modalities, showing the limitation of our method.

3.4.6 Contribution of Different Modalities

To validate the contribution of each modality, in Table 3.6 we show the results of one modality before and after interacting with other modalities. From the table, we can clearly see the benefit brought by information interaction among multiple modalities. We can also see different modalities have different influences on verbs and nouns. For example, in the bottom block of Table 3.6, RGB brings more improvements for Audio in the noun accuracy, and Flow guides the Audio modality to better classify the verbs.

To further validate the enhancement brought by modality interaction, per-class accuracy for RGB modality interacted with different modalities can

Modality	Module	Verb	Noun	Action
RGB	-	30.88	22.98	10.23
(interact with Flow)	MC	39.17	24.94	13.88
(interact with Flow)	MC + SC	40.69	25.22	14.63
(interact with Audio)	MC	40.48	25.64	15.51
(interact with Flow, Audio)	MC	45.38	27.25	17.43
(interact with Flow, Audio)	MC + SC	45.21	27.85	17.80
Flow	-	42.02	21.15	12.90
(interact with RGB)	MC	42.52	24.54	15.32
(interact with RGB)	MC + SC	42.90	25.34	15.81
(interact with Audio)	MC	46.57	23.37	15.95
(interact with RGB, Audio)	MC	46.02	26.14	17.68
(interact with RGB, Audio)	MC + SC	46.28	26.30	17.75
Audio	-	33.34	14.82	8.64
(interact with RGB)	MC	40.10	22.26	13.80
(interact with Flow)	MC	43.80	21.20	14.26
(interact with RGB, Flow)	MC	45.11	24.66	16.27

Table 3.6: Results of single modality before and after interacting with different modalities on the **E100** validation set are shown to validate the contribution of each modality.

be seen in Figure 3.7 (referring to rows 1,2,4,5 of Table 3.6). In Figure 3.7, for the verbs like “wash”, “turn-on” and “turn-off”, RGB modality interacted with Audio modality can have a significant performance boost. We think this is because the unique sounds of water and switch are very similar in both source and target domains. Information from the Flow modality helps RGB in discriminating verbs like “open”, “cut” and “mix”. This is expected since Flow contains more transferable information about the motion and thus complements the RGB modality in predicting verbs. From Figure 3.8, a similar conclusion can be derived from the performance of noun classes, *e.g.* “tap” and “sponge” are usually related to the sound of water, thus audio modality can better help RGB modality in recognizing these sound-related nouns.

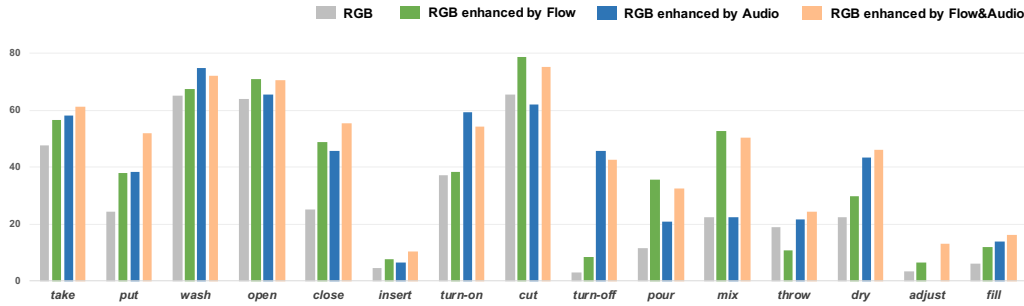


Figure 3.7: Per-class accuracy of several most frequent verbs of the **E100** validation dataset. For verbs like “wash”, “turn-on” and “turn-off”, RGB modality interacted with Audio modality can significantly boost performance. Information from the Flow modality helps RGB in discriminating verbs like “open”, “cut” and “mix”.

3.4.7 Analysis on Parameters and Computational Complexity

We show the parameter with and without our proposed CIA model on the I3D backbone in Table 3.7. The case of two-stream input (RGB and Flow)

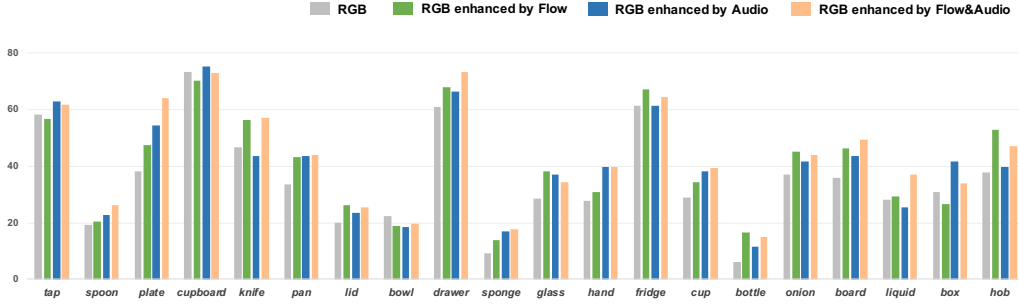


Figure 3.8: Per-class accuracy of several most frequent nouns of the **E100** validation dataset. For sound-related nouns like “tap” and “spoon”, the Audio modality can greatly aid the RGB modality in improving the per-class accuracy. At the same time, motion-related nouns like “lid” and “knife” can get improvement by interacting with the Flow modality.

is shown. Our proposed CIA model introduces a very small amount of additional parameters and computational complexity.

	Number of parameters	Computational complexity
I3D	25.57 M	53.46 GMac
I3D + Ours	27.64 M	53.51 GMac

Table 3.7: Model parameter and computational complexity.

3.5 Conclusion

In this work, we propose a novel CIA model for multi-modal domain adaptive activity recognition. Our CIA model uses two modules to enable the cross-modality feature interaction, which leverages both cross-modal complementarity and cross-modal consensus to accurately learn the most transferable features across the source and target domains. Our method shows considerable improvements on multiple datasets over a variety of previous methods. Our proposed method also has great potential in other domain adaptation

tasks, which we will explore in the future.

Chapter 4

Weakly-Supervised Temporal Grounding of Natural Language by Exploring Positive-Negative Relations

In this chapter, we take one more step forward toward the application of real-world human activity understanding. Both Chapter 2 and Chapter 3 solve the problem in a restricted setting, where only pre-defined action classes are considered. Since this strongly hinders the real-world application where action categories are open-set, in this chapter we relax the restriction of pre-defined action classes, focusing on the actions described by natural languages. Since this setting involves sentences that describe corresponding videos, we mine the relations between the corresponding video segment-sentence pairs, and the non-corresponding video segment-sentence pairs. We formulate a self-training method to leverage this relation, achieving state-of-the-art performance on the weakly supervised temporal sentence grounding task.

4.1 Introduction

One of the most important directions in video understanding is to temporally localize the start and end timestamp of a given sentence description. Also known as temporal sentence grounding, this task has a wide range of potential applications ranging from video summarization [ZCSG16, RYW18], video segmentation [LFV⁺17, HSS20], to recommendation systems [MCS00]. While most existing works deal with this task in a supervised manner, manually annotating temporal labels of the starting and ending timestamps of each sentence is extremely laborious, which harms the scalability and viability of this task in real-world applications. To escalate practicability, recent research attention has been drawn towards weakly supervised temporal sentence grounding, where video-language correspondence is given as annotation only at video-level for model training.

Previous weakly supervised temporal sentence grounding works [MPRC19, GDSX19, HLGJ21, MYK⁺20] mainly adopt the multiple instance learning (MIL) method. They generate mismatched video-language pairs as negative samples and train the model to distinguish the positive/negative samples, in order to learn a cross-modal latent space for a language feature to highlight a certain time period of the video. Some methods find negative samples by selecting sentences that describe another video [MPRC19, YZZW21], but these negative samples are often easy to distinguish and thus cannot provide strong supervision signals. Recent works [ZLZ⁺20, ZHCL22, ZHC⁺22] select negative samples by sampling video segments within the same video, allowing the model to distinguish more confusing video segments.

One major limitation of these methods is that they learn the models completely depending on negative samples, since the objectives of these methods are to generate positive proposals that are distinct from the negative ones, where the distance is usually measured by a certain metric such as the ability to reconstruct the query using only the video segment inside the pro-

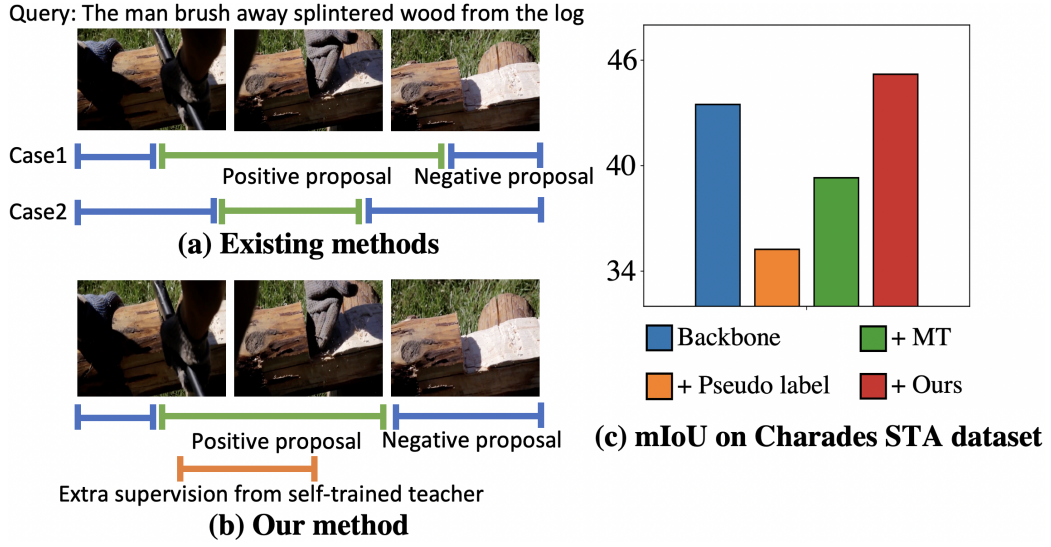


Figure 4.1: (a) Existing methods [ZHCL22, ZHC⁺22] find it hard to distinguish the two cases since they learn positive proposals purely based on negative proposals. (b) Our method provides extra supervision signals for learning positive proposals. (c) Performance of the backbone network [ZHC⁺22], backbone network trained with existing self-training methods pseudo labeling [L⁺13], Mean Teacher (MT) [TV17], and backbone network trained with our method. Directly applying self-training methods for semi-supervised learning negatively influences performance, while our self-training method can improve the backbone performance.

posals [LZZ⁺20, ZHC⁺22, ZHCL22]. However, due to the complex temporal structure of videos that often contain multiple events, being distinct from the negative proposals does not always guarantee the quality of the positive proposals. For example in Figure 4.1(a), it is hard for existing methods like [ZHC⁺22, ZHCL22] to distinguish the two cases since in both cases the positive proposals can better reconstruct the query sentence than the negative proposals.

However, in the absence of strong supervision, it is not straightforward to positively guide the process of temporal sentence grounding. To address this problem, our solution is to leverage self-training to produce extra supervision signals (Figure 4.1(b)). As for self-training, one may consider to directly apply existing techniques originally designed for semi-supervised learning such as pseudo label [L⁺13] or Mean Teacher with weak-strong augmentation [TV17]. However, as shown in Figure 4.1(c), our preliminary experiment suggests that the teacher’s supervision tend to be noisy and would degrade performance due to error accumulation. This is mainly because unlike semi-supervised learning [ZG09], no strong supervision is used for initializing the teacher network.

Following previous works [TV17, LMH⁺21, DLCD21, LDM⁺22], our method also apply the weak-strong augmentation technique, where the student network takes data with strong augmentation as input, while the teacher network gets as input weakly augmented data. Thus, compared to the student network, the teacher network can generate output less affected by heavy augmentation, providing supervisory guidance to the student network. To realize self-training in the weakly supervised temporal sentence grounding task, we specifically design the following two techniques: **(1)** As the teacher network itself is initially trained with only weak supervision and may generate erroneous supervision signals, we apply a Bayesian teacher network, enabling an uncertainty estimation of its output. The estimated uncertainty is used to

weigh the teacher supervision signal thus reducing the chance of error accumulation. (2) To efficiently update both networks, we develop cyclic mutual learning, where the forward cycle forces the student network to output temporally consistent representations with the teacher, and the backward cycle encourages the teacher’s output to be consistent with the average of multiple student outputs generated by inputs with different augmentations. This mutual-learning method allows the teacher to update more carefully than the student, preventing over-fitting to the low-quality supervision. On the other hand, a better teacher will provide reliable uncertainty measures for learning the student network. Our self-training technique can be applied to most existing methods and we observe performance improvement on multiple public datasets.

Our contributions can be summarized as follows: (1) We propose a novel method for temporal sentence grounding based on self-training. To the best of our knowledge, this is the *first attempt* to apply self-training to the weakly supervised temporal sentence grounding task. (2) To realize self-training for this task, we design a Bayesian teacher network to alleviate the negative effect of low-quality teacher supervision, and we use a mutual-learning strategy based on the consistency of the data augmentation to better update the teacher and student networks. (3) Our experiments on two standard datasets Charades-STA and ActivityNet Captions demonstrate that our method can effectively improve the performance of existing weakly supervised methods.

4.2 Related Works

4.2.1 Temporal Sentence Grounding with Strong Supervision

Many previous works focus on Temporal sentence grounding with strong supervision [AHWS⁺17, GGCN19, PMRC22, LQD⁺21, HZH⁺19, ZDW⁺19].

With precise start and end timestamps annotations for each video and query pair, TALL [GSYN17] makes the first attempt to directly regress the start and end timestamp with video and language inputs. LGI [MCH20] further used multi-granularity textual features and predict the timestamps considering local-global video-text interactions. However, these require manual annotation of temporal boundaries for each sentence, which is labor-consuming and subjective [ONRH20] (inconsistent among different annotators). This harms the potential of these approaches in real-world applications.

4.2.2 Weakly Supervised Temporal Sentence Grounding

To avoid laborious annotation and subjective annotation bias, methods for weakly supervised temporal sentence grounding do not use precise start and end timestamps, but only use video-level video-sentence correspondence during training [ZZL⁺20, WLHL20, CJ21, TXSP21]. Without explicit temporal annotations, one group of methods [MPRC19, GDSX19, HLGJ21, MYK⁺20] adopt the multi-instance learning (MIL) technique. These methods construct negative video-language pairs by selecting sentences from other videos, and learning the video-level visual-text correspondence by maximizing the matching scores of the positive pairs while suppressing that of the negative pairs. Then the learned correspondence is used to find the optimal temporal regions that best match the given queries during inference. However, generating negative pairs either from other videos [MPRC19] or within the same video [ZHCL22] can only encourage the models to output proposals that are distinct from the negative proposals. Since videos usually contain multiple complex temporal events, proposals distinct from the negative ones may just represent some other events but not correspond with the ground truth. In our method, we design a self-training method based on a teacher-student structure, where the teacher network can provide extra self-supervision sig-

nals to learn a better student network, and inversely the student network transfers learned knowledge to the teacher network by cycle consistency.

Another line of research aims to select the video segments which can best reconstruct the given query sentence [ZHCL22, LZZ⁺20, SWM⁺20]. The reconstruction result can also be used for contrastive learning [ZHC⁺22]. In our method, we also leverage the reconstruction performance to guide the mutual learning process of the teacher-student method.

4.2.3 Self-training in Weakly Supervised Learning

Self-training is originally proposed in semi-supervised learning and has been adopted in other scenarios such as domain adaptation [CLS20, LDM⁺22]. Many methods also use self-training to improve the model performance for weakly supervised tasks, for example, text classification [MSZH19], semantic segmentation [SY19, WZK⁺20, LQF21] or object/action detection [WLC⁺16, JWJ⁺17, ZWK⁺18, ZHCY21, YZY⁺21, CDZ⁺21, WHL⁺21, CGYX22]. To the best of our knowledge, we make the first attempt to explore the use of self-training on weakly supervised temporal sentence grounding.

4.2.4 Bayesian Deep Learning

To provide posterior uncertainty estimates, there has been a long presence of Bayesian inference in machine learning [BN06]. Since Bayesian inference on neural networks is difficult, early works explored a variety of methods such as Markov Chain Monte Carlo (MCMC) [Nea12] or variational inference [HvC93]. Bayesian deep learning has thus been applied to various tasks such as unsupervised domain adaptation [CLS20] and time series forecasting [JGK⁺22]. In this work, we utilize a Bayesian network to acquire uncertainty estimation for self-training.

4.3 Proposed Method

4.3.1 Problem Formulation and Overview

We first demonstrate the problem formulation before going into details of our proposed method. Given a set of N videos $\{v_1, \dots, v_N\}$ and their corresponding query sentences $\{q_1, \dots, q_N\}$ that describe each video, our goal is to ground each sentence to a specific temporal segment in video with start and end timestamps.

Figure 4.2 shows the overview of our method. Our self-training method consists of a Bayesian network as the teacher network and an identical network as the student network. We can apply the network architecture of most existing weakly supervised methods as the teacher/student. In the following part of this section, we showcase the backbone with the state-of-the-art method CPL [ZHC⁺22] which outputs Gaussian attention proposals. Following the weak-strong augmentation [LMH⁺21, OLF⁺22], the teacher network takes a weakly augmented video-language pair as input, whereas the student network takes a strongly augmented video-language pair as input. Both networks are first initialized with the training approach of the backbone. We then perform self-training and update both networks using uncertainty estimated by the teacher network (Section 4.3.2) and temporal augmentation cycle-consistency (Section 4.3.3).

4.3.2 Uncertainty Estimation via Bayesian Teacher

Since the teacher network itself is not learned by strong supervision, it may generate low-quality supervision signals even given weak augmentation. In fact, our preliminary experiment in Figure 4.1(c) shows that directly applying the output of the teacher network as supervision can even downgrade the overall performance. Thus, it is essential to suppress the influence of low-quality outputs. Inspired by the success of Bayesian deep learning [KG17],

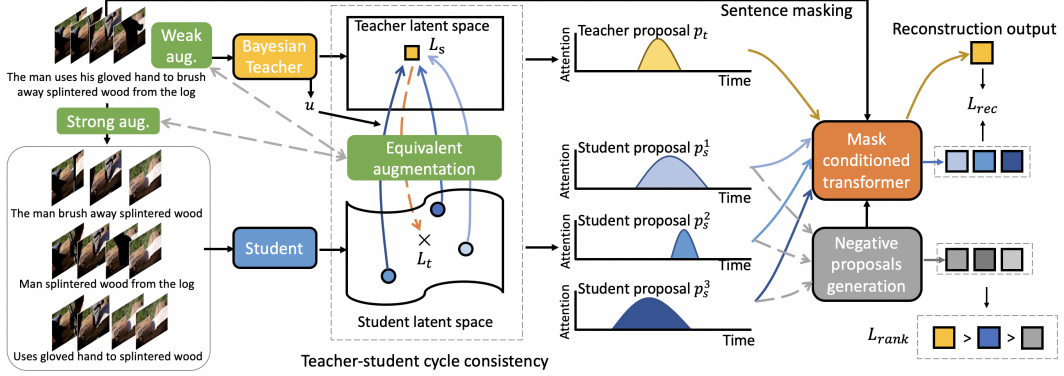


Figure 4.2: Overview of our proposed method. The teacher network takes as input weakly augmented data while the student network takes as input multiple strongly augmented data. Then teacher-student cycle consistency is used for mutual learning of the two networks, considering the uncertainty u into consideration. Gaussian masks are generated to represent the proposals, and we further use reconstruction loss L_{rec} and ranking loss L_{rank} to ensure high-quality proposals.

we propose to use Bayesian inference on the teacher network to get an uncertainty estimation.

Since all parameters are considered as random variables in a Bayesian network, obtaining the posterior distribution is often intractable. Recent works use variational inference as an approximation [BKM17]: given an input I , the predictive distribution of output O is acquired by D -time repeated stochastic forward passes with network parameters sampled from an approximating variational distribution $q(w)$:

$$\begin{aligned}
 p(O|I) &= \int p(O|I, w)q(w)dw \\
 &\approx \frac{1}{D} \sum_{i=1}^D p(O|I, w_i), \quad w_i \sim q(w),
 \end{aligned} \tag{4.1}$$

where $p(O|I, w_i)$ is one forward pass with model parameters w_i . In practice, we use the trick in [GG16] to perform Bayesian inference without changing

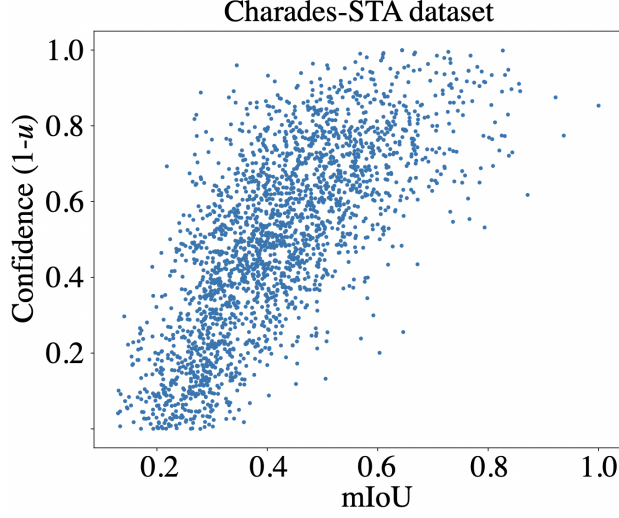


Figure 4.3: Model confidence (computed by $1-u$) and mIoU on the Charades-STA dataset are highly correlated, indicating that we can leverage the uncertainty estimation u to represent the quality of the network output.

model structure and parameter by sampling model parameters with dropout. Using a temporal sentence grounding model $F(v, q, w)$ with weights w which outputs a temporal segment proposal p given one video v and a sentence query q as input, the uncertainty estimation u can be computed as:

$$\bar{p} = \frac{1}{D} \sum_{i=1}^D F(v, q, w_i), \quad w_i \sim \text{dropout}(w) \quad (4.2)$$

$$u = \frac{1}{D} \sum_{i=1}^D p_i^2 - \bar{p}^2. \quad (4.3)$$

This uncertainty estimation is then used in the teacher-student mutual learning stage to alleviate the negative influence of the low-quality supervision signals from the teacher network.

As a proof of concept, we visualize the correlation between the outputs' uncertainty and their mean Intersection over Union (mIoU) with the ground truth segment in v that corresponds to the given query sentence q . For better

visualization, we transform uncertainty to confidence by $1 - u$, and visualize it with the mIoU score on the test set of the Charades-STA dataset in Figure 4.3. From the figure, we can see that samples with low IoU scores tend to also have low confidence scores. This study shows that we can leverage uncertainty measurement u to represent the quality of the teacher network’s output.

4.3.3 Mutual Learning with Temporal Augmentation Cycle Consistency

To effectively encourage the student to learn knowledge from teacher, and allow the learned knowledge to transfer back to the teacher, we design teacher-student mutual learning with temporal augmentation cycle consistency. We feed the student network a set of temporally augmented videos, including temporal scaling, shifting, and masking. We also add augmentation to the sentence queries by decomposition using semantic role labeling [GJ02] (details in Section 4.3.3). The teacher network is given only weakly augmented videos and sentence queries as input. Details of the mutual learning are as follows:

Model initialization. The first step is to initialize the teacher and student models. Initialization is a critical step for all self-training methods since we rely on the teacher to generate reliable supervision signals to optimize the student network. In our self-training method, it is possible to apply most existing weakly supervised temporal sentence grounding methods to initialize the model. We denote the initialization loss of the model as L_{init} . Note that the teacher and student models are initialized with the same parameters.

Data augmentation. We use the weak-strong augmentation strategy to allow the teacher network provide supervision signal to the student network.

For the data augmentation, we apply random temporal shifting, random temporal scaling, and random temporal masking on the input video. Specifically, we first randomly scale the temporal length of each video with a ratio of $l\%$, and crop the scaled video to its original size with a temporal shifting ratio $s\%$ (*i.e.*, the start timestamp of cropping is at $s\%$ of the scaled length). After this, we randomly choose $m\%$ of the timestamps and replace the feature on the corresponding timestamp with zero vectors. As for the sentence query, we randomly drop words at 50% probability while keeping the sentence containing all words in at least one semantic structure decomposed by semantic role labeling (SRL) [GJ02]. We repeat this augmentation k times, thus each pair of video-sentence input (v_t, q_t) is augmented into k video-subsentence pairs $\{(v_s^1, q_s^1), \dots, (v_s^k, q_s^k)\}$.

Teacher-student cycle consistency. We perform the teacher-student mutual learning leveraging the teacher-student cycle consistency. Denote the proposal output of the teacher network as p_t , we first apply an equivalent transformation \mathcal{T}^k to echo the student’s data augmentation (scaling and shifting) of the video. Since p_t represents a temporal segment (a start and an end timestamp) of the original video, this equivalent transformation is straightforward. Because of our augmentation strategy, the cycle consistency lies in that all the k student outputs p_s^k should be close to the teacher’s output p_t , and inversely, the average of student outputs $Avg(p_s^k)$ should cycle back to the original teacher output. This cycle consistency enables both student learning and teacher learning. The student output can be directly supervised by the transformed teacher output:

$$L_s = \frac{\sigma(\lambda_1 u)}{k} \sum_{i=1}^k |\mathcal{T}^k(p_t) - p_s^k|, \quad (4.4)$$

where σ denotes the sigmoid function, u is the teacher’s uncertainty, λ_1 is a hyper-parameter that controls the scale of the uncertainty measure, and p_s

is the temporal proposal of the student.

The teacher’s learning can be expressed as:

$$L_t = \sigma(\lambda_2 u) \left| \mathcal{T}^k(p_t) - \frac{1}{k} \sum_{i=1}^k (p_s^k) \right|, \quad (4.5)$$

where λ_2 is another hyper-parameter.

Enhancing self-training with reconstruction. To enhance the self-training, we additionally apply a triplet ranking loss based on masked reconstruction, as shown in the rightmost part of Figure 4.2. Different from previous methods that rank only between a proposal and its negative component sampled within or from other videos, we rank the reconstruction result based on the teacher’s proposal, student’s proposal, and a negative proposal taken by negative proposal mining from [ZHC⁺22]. To be specific, we randomly choose one of the sentences from either the teacher input or the student input, and then randomly replace 30% of the words in the sentence query with a mask token, and predict the next word using the prefix of the sentence and the visual features within each proposal by a mask conditioned transformer [ZHCL22]. Please refer to [ZHC⁺22] and [ZHCL22] for details of negative proposal mining and mask conditioned transformer. Denote the cross-entropy loss of the reconstruction by teacher proposal, student proposals, and the negative proposal as $L_{ce}(p_t), L_{ce}(p_s), L_{ce}(p_n)$, respectively, our ranking target is:

$$\begin{aligned} L_{rank} = & \max(L_{ce}(p_t, q_t) - L_{ce}(p_s, q_t) + m_1, 0) \\ & + \max(L_{ce}(p_s, q_t) - L_{ce}(p_n, q_t) + m_2, 0). \end{aligned} \quad (4.6)$$

To learn the reconstruction, we apply cross-entropy loss using the student and teacher’s proposals, without the negative proposals as [ZHC⁺22]:

$$L_{rec} = L_{ce}(p_t, q_t) + L_{ce}(p_s, q_t) \quad (4.7)$$

Updating teacher and student. In our method, the student and teacher are updated asynchronously. After initialization, we first fix the teacher network and learn the student by L_{init} , L_s , L_{rec} and L_{rank} , and then fix the student network and train the teacher by L_{init} , L_t , L_{rec} and L_{rank} . Details of training can be found in Section 4.4.1. We use the result of the teacher network as the final output in inference.

4.4 Experimental Results

Our experiments are performed on two publicly available datasets Charades-STA [GSYN17] and ActivityNet Captions [KHR⁺17], following the common practice of previous works. Charades-STA is a subset of the Charades dataset [SVW⁺16] with sentence annotations and temporal timestamp annotations. It contains 12,408/3720 video-query pairs in the training/testing set. We report our results on the test split. ActivityNet Captions is a subset of the ActivityNet dataset [CHEGCN15] which contains a number of 37,417/17,505/17,031 annotated video-sentence pairs in the train/val.1/val.2 split. Following the majority of the previous works, we also report our results on the val.2 split.

As for the evaluation metric, we adopt the "IoU@ n " metric under recall rate of top-1 prediction. A predicted proposal is considered correct if its Intersection over Union with the ground-truth proposal is greater than the predefined IoU threshold n . We choose $n = \{0.3, 0.5, 0.7\}$ for the Charades-STA dataset and $n = \{0.1, 0.3, 0.5\}$ on the ActivityNet Captions dataset.

4.4.1 Implementation Details

As for data pre-processing, we follow [ZHC⁺22] to use C3D [TBF⁺15] feature for ActivityNet Captions and I3D [CZ17b] feature for Charades-STA. The features are extracted by first downsampling each video at a rate of 8. Pre-trained GloVe word2vec [PSM14] are used to extract word embeddings. We follow [ZHCL22] to set the maximum sentence length as 20, the maximum video length as 200, and the vocabulary size for the Charades-STA and ActivityNet Captions datasets as 1,111 and 8,000, respectively.

For data augmentation, we use different parameters for different datasets. For student input, on Charades-STA, when generating each augmented data, l is a random number in $[100, 150]$, s is randomly chosen from $[-25, 25]$, and m is set to 10. We use $k = 2$ for Charades-STA since the sentences are typically short. On the ActivityNet Captions dataset, l is fixed as 100 and s is selected randomly from $[-50, 50]$. We set $m = 30$ and $k = 4$. On both datasets, when the augmentation causes an index out-of-range error, we repeat the feature on the nearest timestamp. We use the parameters l, s to perform the equivalent transformation \mathcal{T} . For the teacher input, we only apply random frame feature masking at 10%. All data augmentation is done only in the training stage, we use the original video-sentence pair as input to the teacher network to get results on the test sets.

As for the training, we first initialize the model with L_{init} for 15 epochs for the Charades-STA dataset and 30 epochs for the ActivityNet Captions dataset. After initialization, we repeat the following step 15 times: (1) fix the teacher network and train the student network for 3 epochs, using; (2) fix the student network and train the teacher network for 1 epoch. We use Adam optimizer with learning rate set to 0.0004 for training both networks, the learning rate is decayed with an inverse square root scheduler. We set $\lambda_1 = 1, \lambda_2 = 2$ for model training on both datasets. We give L_s and L_t a weight of 10 while giving other losses a weight of 1 during training.

Our method does not introduce additional parameters to the backbone network. When applying on networks that generate multiple proposals, we only use the top-1 proposal to compute reconstruction and ranking losses. When applying to backbones that do not contain reconstruction-based loss, we simply discard the L_{rec} and L_{rank} during training.

4.4.2 Results and Comparisons

The top block of Table 4.1 and 4.2 shows the performance of previous state-of-the-art weakly supervised temporal sentence grounding methods. Compared with our method in the last row of each table, we observe best performance is achieved on both datasets with our method.

In the bottom block of each table, we list the performance of the backbone method CPL trained with original data but is directly inferenced with augmented data (**CPL (aug)**), CPL both trained and inferenced with augmented data (**CPL + aug**). Also, we show the backbone method CPL applied with the standard teacher-student self-training method MT [TV17] (**CPL + MT**) with weak-strong augmentation. Note that in Table 4.2, we show both the results reported in the original CPL paper (**CPL (ori.)**) and the results of our replication (**CPL (rep.)**).

We note that, while the backbone network CPL [ZHC⁺22] performs worse when it is directly inferenced on augmented data, simply training with data augmentation already results in a good performance on both datasets. This implies the success of our self-training method, since if the backbone network performs consistently on strongly augmented data, no extra knowledge can be learned from the self-training. Compared to the backbone method CPL [ZHC⁺22], our method can consistently increase its performance on all of the metrics. This is proof that the student network learned useful knowledge from the positive guidance provided by the teacher and subsequently transferred the knowledge back to the teacher, thanks to the teacher-student

cycle consistency. Importantly, we observe larger improvement in IoU at higher thresholds (IoU@0.5 and 0.7 on Charades-STA, IoU@0.3 and 0.5 on ActivityNet Captions). This is because the backbone CPL judges each proposal using reconstruction error, thus tending to produce long proposals (see Section 4.4.4) to ensure a good reconstruction. The largest performance gap exists on the ActivityNet Captions dataset at IoU@0.3 and 0.5. We believe this is because our sentence augmentation technique makes the sentences shorter thus reconstruction task becomes easier to accomplish, which addressed the limitation stated in [ZHC⁺22], *i.e.*, worse performance on long sentences.

Comparing the performance of the backbone and our method with the standard self-training method MT [TV17], we can see that MT slightly degrades the backbone performance, while our method can increase the backbone performance. This is expected since (1) MT does not use uncertainty-guided training, resulting in the accumulation of errors, and (2) MT updates the teacher network via Exponential Moving Average (EMA), however, the student network does not take the whole sentence query as input like the teacher network, thus directly updating model weights to the teacher network performs unfavorably in our setting.

4.4.3 Ablation Study

We conduct ablation studies to show the effectiveness of each component in our method, as well as the influence of different augmentation techniques.

Comparison on self-training methods. To confirm the effect of each proposed component of our method, we apply standard self-training methods to see their usefulness in the weakly supervised temporal sentence grounding task. We also remove different components of our method to see the effect of each component. We specifically compare with the following baselines

Method	IoU@0.3	IoU@0.5	IoU@0.7	mIoU
TGA [MPRC19]	32.14	19.94	8.84	-
SCN [LZZ+20]	42.96	23.58	9.97	-
WSTAN [WDZL22]	43.39	29.35	12.28	-
VLANet [MYK+20]	45.24	31.83	14.17	-
MARN [SWM+20]	48.55	31.94	14.81	-
CRM [HLGJ21]	53.66	34.76	16.37	-
VCA [WCJ21]	58.58	38.13	19.57	38.49
LCNet [YZZW21]	59.60	39.19	18.87	38.94
RTBPN [ZLZ+20]	60.04	32.26	13.24	-
CNM [ZHCL22]	60.04	35.15	14.95	38.11
CPL [ZHC+22]	66.40	49.24	22.39	43.48
CPL (aug)	56.46	38.47	17.64	36.78
CPL + aug	67.35	50.09	23.75	44.39
CPL + MT [TV17]	65.17	32.55	11.40	39.31
CPL + Ours	69.16	52.18	23.94	45.20

Table 4.1: IoU@{ 0.3, 0.5, 0.7} and mIoU results on the Charades-STA dataset test split. CPL (ori) denotes the results reported in [ZHC+22], while CPL (rep) is our replicated result. CPL (aug) is the backbone method CPL trained with original data but is directly inferenced with augmented data. CPL + aug is the result where CPL is both trained and inferenced with augmented data. backbone method CPL applied with the standard teacher-student self-training method Mean Teacher [TV17] is shown as CPL + MT. The bold numbers represent the top-1 result. Our proposed method outperforms previous works in all metrics.

Method	IoU@0.1	IoU@0.3	IoU@0.5	mIoU
WS-DEC [DHG ⁺ 18]	62.71	41.98	23.34	28.23
VCA [WCJ21]	67.96	50.45	31.00	33.15
MARN [SWM ⁺ 20]	-	47.01	29.95	-
SCN [LZZ ⁺ 20]	71.48	47.23	29.22	-
RTBPN [ZLZ ⁺ 20]	73.73	49.77	29.63	-
CTF [CML ⁺ 20]	74.2	44.3	23.6	32.2
WSSLN [GDSX19]	75.4	42.8	22.7	32.2
LCNet [YZZW21]	78.58	48.49	26.33	34.29
WSTAN [WDZL22]	79.78	52.45	30.01	-
CRM [HLGJ21]	81.61	55.26	32.19	-
CNM [ZHCL22]	79.74	54.61	30.26	36.59
CPL (rep.) [ZHC ⁺ 22]	81.14	53.99	29.38	35.55
CPL (ori.)	79.86	53.67	31.24	-
CPL (aug)	78.52	51.32	28.69	34.53
CPL + aug	82.53	54.90	30.19	36.87
CPL + MT [TV17]	79.50	52.20	28.03	34.92
CPL + Ours	82.10	58.07	36.91	41.02

Table 4.2: IoU@{0.1, 0.3, 0.5} and mIoU results on the ActivityNet Captions dataset val_2 split. CPL (ori) denotes the results reported in [ZHC⁺22], while CPL (rep) is our replicated result. CPL (aug) is the backbone method CPL trained with original data but is directly inferenced with augmented data. CPL + aug is the result where CPL is both trained and inferenced with augmented data. backbone method CPL applied with the standard teacher-student self-training method Mean Teacher [TV17] is shown as CPL + MT. The bold numbers represent the top-1 result. Our proposed method outperforms previous works in most metrics.

Method	IoU@0.3	IoU@0.5	IoU@0.7	mIoU
Backbone alone	66.40	49.24	22.39	43.48
Pseudo label [L ⁺ 13]	61.59	44.93	19.85	40.24
Pseudo label + u	64.85	48.29	22.70	42.75
Mean Teacher [TV17]	65.17	32.55	11.40	39.31
Mean Teacher + u	66.02	39.68	14.44	40.43
w.o. u	68.62	48.89	21.88	42.88
w.o. L_{rank}, L_{rec}	68.24	48.80	22.07	43.39
Ours full	69.16	52.18	23.94	45.20

Table 4.3: Ablation study on the Charades-STA dataset.

on the Charades-STA dataset: **Pseudo-label** [L⁺13] is one straightforward technique for self-training, where the model iteratively refines its prediction based on the previous prediction. Since pseudo labeling often generates low-quality outputs causing error accumulation, we add another comparison where Bayesian inference is applied and the uncertainty u is utilized to weigh the pseudo labels. We denote this setting as **Pseudo label + u** . Also, we use **Mean Teacher** as a representative of standard self-training with weak-strong augmentation in which the teacher network is updated by exponential moving average (EMA), *i.e.*, L_t is not used for teacher update. To better see the effect of uncertainty guidance in self-training, we equip Mean Teacher with a Bayesian teacher network and denote this baseline as **Mean Teacher + u** . We further remove one of the other components, *i.e.*, the Bayesian inference of teacher (**w.o. u**) and the reconstruction loss (**w.o. L_{rank}, L_{rec}**), to indicate the effectiveness of each ingredient.

Results can be seen from Table 4.3. Not surprisingly, directly using the pseudo label technique downgrades the backbone performance due to the error accumulation. Mean Teacher’s EMA-based teacher update is not suitable to our method, due to the difference in input between the teacher and student networks. Adding uncertainty by Bayesian inference to Pseudo label and Mean Teacher can mitigate the error accumulation to some extent.

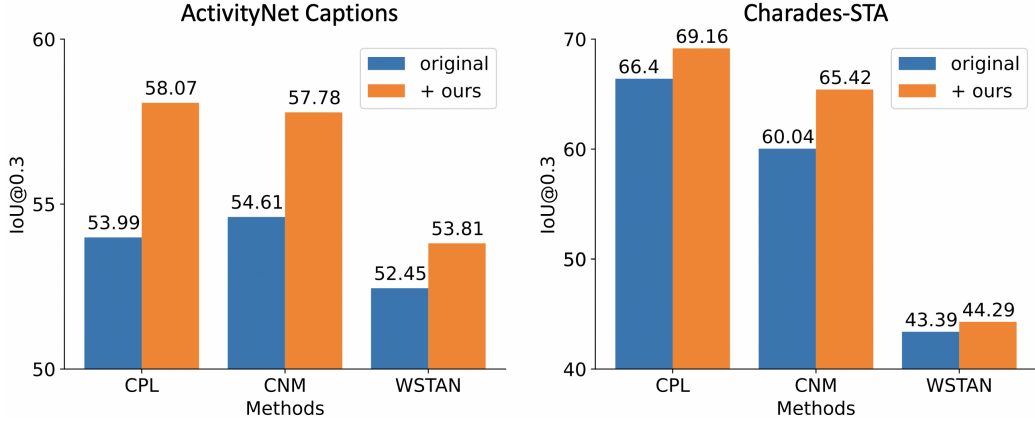


Figure 4.4: Results on the ActivityNet Captions (left) and Charades-STA (right) datasets when our method is applied on different backbone networks.

Our method cannot get optimal performance without the Bayesian inference of the teacher network, which proves our assumption that the student can learn better from the high-quality supervision signals of the teacher. Similarly with other weakly supervised temporal sentence grounding approaches [ZHCL22, ZHC⁺22], we can also observe that reconstruction loss contributes to the final result. Our full method with all the uncertainty measurement, cycle consistency, and reconstruction loss performs the best, indicating that the design of all these components is of great importance for our network.

Changing backbones. As explained before, our self-training method can work with multiple backbones and improve their performance. Here we show the performance of three backbone networks trained with our method. In Table 4.4, we demonstrate the IoU@0.3 performance of CPL [ZHC⁺22], CNM [ZHCL22] and WSTAN [WDZL22] before and after applying our method on two datasets. Our method can bring positive improvement to all the three backbones, indicating the generalizability of our proposed method.

Augmentation			Charades-STA		ActivityNet Captions	
V	M	D	IoU@0.3	IoU@0.5	IoU@0.3	IoU@0.5
			66.40	49.24	53.99	29.38
✓			68.02	51.49	54.40	31.13
	✓		68.59	51.39	54.83	30.15
		✓	66.44	49.73	56.96	34.06
✓	✓	✓	69.16	52.18	58.07	36.91

Table 4.4: Results of our method when using different augmentation techniques. V: video temporal scaling and shifting; M: video temporal masking; D: decomposition of sentence queries with SRL.

Discussion on different augmentation techniques. In our method, we use multiple augmentation techniques during the teacher-student mutual learning. We show the effect of each augmentation in Table 4.4. Here V stands for video temporal scaling and shifting, M denotes video temporal masking, and D denotes the decomposition of sentence queries with SRL. To better show the performance gap, we show the original backbone in the first row of Table 4.4 with gray background. We can see from the table that different augmentation techniques have different influences on each of the datasets: the augmentation on videos V and M are more effective on the Charades-STA dataset, while the augmentation of sentence decomposition works better on the ActivityNet Captions dataset. We think this is mainly because of the difference in the length of sentences. In the Charades-STA dataset, the sentences are mostly short with an average of 6.2 words per sentence, while in the ActivityNet Captions dataset the average number of words per sentence is 13.5. Thus, the effect of decomposition by semantic role labeling is more significant in the sentences of the ActivityNet Captions dataset.

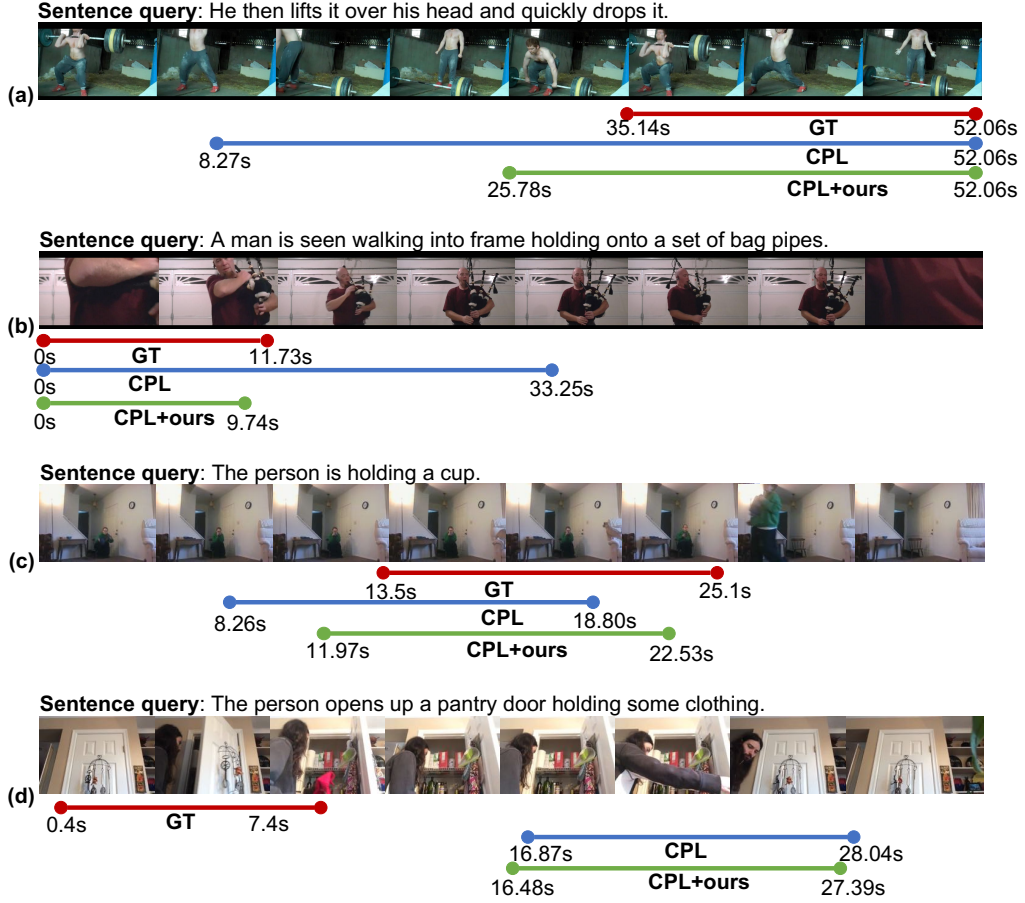


Figure 4.5: Qualitative examples of the ground truth (GT), the backbone network (CPL), and the backbone method with our mutual learning (CPL+ours). Examples (a, b) are from the ActivityNet Captions dataset, and (c, d) are from the Charades-STA dataset.

4.4.4 Qualitative Results

We show several qualitative examples in Figure 4.5. From this figure we can obtain several interesting observations:

- (1) As shown in Figure 4.5(a) (b) and (c), our method can achieve better results than the backbone CPL, proving that our self-training technique can positively provide extra guidance to the network.
- (2) As shown in Figure 4.5(a) (b) and (c), the backbone method CPL tends to output longer proposals, while our method can effectively reduce the length of the proposals to a reasonable range. This is mainly because CPL purely relies on reconstruction results as an indicator of the quality of each proposal, thus the proposals tend to be long in order to guarantee a successful reconstruction.
- (3) Fig. 5(d) shows that when the performance of CPL is too off, it is also hard for our method to refine this result. This reveals one limitation of our work, *i.e.*, relies on the performance of the backbone network.

4.4.5 Limitation and Future Work

As discussed in the previous section, although our method can be applied to multiple backbone networks and improve their performance, one limitation is that the performance relies on the backbone network. When the backbone network does not generate reliable results, our method can only marginally improve the overall performance.

Most methods generate multiple proposals for each sentence. However, we found that while our method increases the IoU performance of the top-1 proposal, our method can only marginally improve the results under recall rate of top-5 prediction (Recall@5). In Table 4.5 we show the Recall@5

Method	Recall@1	Recall@5		
	IoU@0.3	IoU@0.3	IoU@0.5	IoU@0.7
VLANet [MYK ⁺ 20]	45.24	95.70	82.85	33.09
VCA [WCJ21]	58.58	98.08	78.75	37.75
LCNet [YZZW21]	59.60	94.78	80.56	45.24
RTBPN [ZLZ ⁺ 20]	60.04	97.48	71.85	41.48
CPL [ZHC ⁺ 22]	66.40	96.99	84.71	52.37
CPL + aug	67.35	97.37	85.40	52.74
CPL + Ours	69.16	96.96	84.86	52.58

Table 4.5: IoU of different methods at Recall@5 on the Charades-STA dataset. Recall@1, IoU@0.3 is shown for reference.

performance on the Charades-STA dataset. We can see that the backbone network CPL directly trained with data augmentation can achieve consistently higher performance on Recall@5 compared to our method. We think the reason is that although our self-training method can generate more accurate top-1 proposals, it simultaneously harms the diversity of the output proposals. We leave the goal of increasing Recall@5 for future work.

4.5 Conclusion

In this chapter, we take one more step forward toward the application of real-world human activity understanding. We specifically target on the weakly-supervised temporal sentence grounding task and propose the first self-training-based method this task. Our self-training framework includes a pair of mutually learned teacher and student networks. We give weak-strong augmentation to the teacher-student networks and learn the networks by a teacher-student cycle consistency loss, a reconstruction-based loss and a ranking-based loss. Experiments on two public datasets demonstrate the outstanding performance of our method. We also show in ablation studies the

effectiveness of the components in our method and our method’s capability of working with different backbone networks.

Chapter 5

Conclusion

5.1 Summary

This thesis focus on mining multilateral relations for human activity understanding. Based the whether supervision is provided and whether the number of activities is fixed, human activity understanding can be divided into three settings: fully-supervised setting, unsupervised setting and open-set setting.

Start with a basic setting which defines the human activity understanding task as the classification of trimmed videos to a fixed number of activity categories with supervision during training. The main challenge under this setting is how to enhance the robustness of recognition performance against the instability of videos, e.g. temporal outliers. Unlike third-person videos that are usually taken by a fixed camera, first-person videos produced by first-person camera mounted on the body of camera-wearers suffer more from the temporal outlier problems. For dealing with temporal outliers to improve the first-person activity recognition performance, this thesis explores multilateral relations between local-local video features and local-global video features. By effectively leveraging these relations through a stacked temporal attention module (STAM), this work managed to significantly improve the first-person activity recognition accuracy on two first-person datasets.

In the real-world application of human activity recognition, there is not always sufficient supervision to train a model, which brings the demand for better human activity recognition under unsupervised setting. The second work of this thesis focuses on domain adaptive activity recognition setting to enable the adaptation of trained deep neural networks on new target domains without supervision. To enhance the transferability of deep models, this work explores two relations among multiple modalities: multi-modal complementarity and multi-modal consensus. The proposed CIA first enables information exchange across modalities, refining features of each modality by absorbing transferable knowledge from other modalities. After that, the other module leverage multi-modal consensus to further enhance feature transferability by focusing on the most transferable spatial regions. The proposed CIA achieves state-of-the-art recognition accuracy on three domain adaptive activity recognition datasets, which demonstrate the effectiveness of leveraging multilateral relations among different modalities.

In addition to relaxing the restriction of human activity understanding setting from the perspective of demand of annotations (from fully-supervised to unsupervised), now it's also possible to challenge the activity understanding task from the perspective of dealing with more complex activities. In addition to understanding relative "simple" human activities which are pre-defined, fixed number of categories and consist of verbs and nouns, human activity could also be described in more detail as a natural sentence. The third work of this thesis concentrates on finding the start and end of an activity in a video given a natural language sentence description as input. This work formulated the ranking relations both within positive proposals and between positive and negative ones, and utilize the ranking relations to predict more accurate temporal timestamps. Experiments on two datasets demonstrate that deeper exploration of such multilateral relations could better find the boundary of human activity.

5.2 Contributions

The main contributions in this thesis work can be summarized as follows:

- In this thesis, we propose a simple yet effective module that leverages multilateral relations between local-local video clip features and local-global clips features to focus on discriminative clips for more robust first-person activity recognition. The proposed module could be built on top of most existing video backbones for improving activity recognition accuracy. Performance improvements on both first-person and third-person datasets validate the potential of mining multilateral relations for dealing with temporal outliers to enhance the robustness of activity recognition.
- Propose a modal that leverages cross-modal complementarity and cross-modal consensus to enhance feature transferability for domain adaptive activity recognition. The proposed model uses two modules to enable the cross-modality feature interaction, which leverages cross-modal complementarity to absorb knowledge across modalities and cross-modal consensus to find and focus on spatial locations, respectively. This is the first work to consider cross-modal multilateral relations for increasing feature transferability across domains. Results comparison on three domain adaptive activity recognition datasets shows the superiority of exploring multilateral relations among different modalities over a variety of previous methods.
- Propose a novel method that jointly considers multilateral ranking relations both within positive proposals and between positive and negative proposals for natural sentence human activity temporal grounding task with only weak supervision. This is the first attempt to apply self-training to the weakly-supervised temporal sentence grounding task. The self-training framework includes a pair of mutually learned teacher

and student networks. Experiments on two public datasets demonstrate the outstanding performance of the proposed method. The proposed method is also capable of working with different backbone networks, further showing the effectiveness of exploring multilateral relations for finding more accurate activity boundaries.

5.3 Future Directions

5.3.1 Compositional Human Activity Understanding

As mentioned in Chapter 4, compared to activity composed of one verb and one noun, activity described by natural language allows for a combination of multiple atomic “simple” actions. In addition, on account of the characteristics of language, the total number of activity categories will no longer be restricted by a pre-defined number. With the help of natural language, now the boundaries of human activity understanding are further extended – providing the opportunity to explore more complex human activity as well as take one step towards truly practical AI.

For understanding complex human activity, recent works mainly treat it as monolithic events in an end-to-end manner. In other words, they produce a single label to densely describe a long video sequence, and treat it as an ordinary classification problem. Although treating the complex activity as a monolithic event could directly take advantage of most existing models proposed for simple activity understanding, they are difficult to scale up to more varied activity patterns. Fundamentally, most complex actions consist of a series of simpler events, and thus a dense activity can be treated as a composition of known event primitives[JKFN20, MXH⁺20b, MXH⁺20c]. Such a compositional representation can allow for a higher degree of model generalization. By learning from a finite number of action composites, and recombining their constituent event primitives in novel ways for unseen activity

descriptions, the representation can expand to large numbers of novel scenarios that have not been observed in the original activity space. In this case, the activity description is not treated as a single label but as a language modality that allows learning finer-grained video and language correspondence.

Several works show that compositionality is key to achieving generalization by combining known primitive elements, especially for handling novel composited structures. Take the temporal grounding task as an example, compositional temporal grounding is the task of localizing dense activity by using known words combined in novel ways in the form of novel query sentences for the actual grounding. In recent works, composition is assumed to be learned from pairs of whole videos and language embeddings through large-scale self-supervised pre-training. Alternatively, one can process the video and language into word-level primitive elements, and then only learn fine-grained semantic correspondences. In either case, the main challenge of utilizing datasets of limited action size to achieve generalization towards novel actions remains challenging due to its combinatorial complexity.

In fact, when talking about compositional activity understanding, the granularity of compositions is often ignored. For example, given a sentence “News reporter in blue are giving news about people cutting dogs’ hair and trimming dogs’ nails”. By analyzing and dividing the sentence, several phrases that constitute the complex activity can be generated, e.g. “News reporter in blue are giving news”, “people cuts dogs’ hair” and “people trims dogs’ hair”. Additionally, from the word-level perspective, this sentence can be further seen as the composition of verbs (give, cut, trim), nouns (reporter, news, dog, hair, *etc*) and adjectives (blue). Unfortunately, existing approaches do not consider the granularity of the compositions, where different query granularity corresponds to different video segments. However, a good compositional representation should be sensitive to different levels of granularity of both the action and the query language.

Since there is no ground truth to relate subsentences and their corresponding sequences, it's challenging to achieve accurate compositional human activity grounding. At the same time, the absence of correspondence between primitive actions and the timestamps demonstrates the similarity and association of compositional temporal grounding task and weakly supervised temporal grounding task. Inspired by the general weakly supervised temporal grounding task, the multilateral relations between vision-language correspondence at different granularity could be explored to help formulate the training of the model.

5.3.2 Long-tail Human Activity Understanding

Another challenge of real-world human activity understanding is the imbalanced / long-tailed activity category distribution, which means head-class occupies the majority of training data while the remaining tail-class contains only a few samples. Not only the first-person and third-person datasets mentioned in Chapter 2 under basic supervised human activity understanding setting, three datasets used in Chapter 3 under domain adaptation setting also show the obvious long-tail distribution of activity classes [DDF⁺20]. This kind of natural long-tail distribution leads to a high priority of better performance on head-class, which often means the sacrifice of the accuracy on tail-class. In other words, the learned feature space of head-class is larger than the learned feature space of tail-class.

Due to the long-tail distribution of training data, tail classes often obtain unsatisfactory performance, which hinders the practical application of deep learning models. In this thesis, for first-person human activity recognition task in chapter 2 and domain adaptive activity recognition task in chapter 3, we take the overall accuracy as the final metric instead of mean-class metric that could reveal the long-tail performance. Mining multilateral relations for human activity understanding while taking into consideration the long-tail

issues and reducing the performance gap between head-class and tail-class will be a promising future direction.

Bibliography

- [ACDY20] Nakul Agarwal, Yi-Ting Chen, Behzad Dariush, and Ming-Hsuan Yang. Unsupervised domain adaptation for spatio-temporal action localization. *Proceedings of the British Machine Vision Conference*, 2020.
- [AHWS⁺17] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5803–5812, 2017.
- [BKM17] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.
- [BN06] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Number 4. 2006.
- [BSD⁺17] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.
- [CDZ⁺21] Tianyue Cao, Lianyu Du, Xiaoyun Zhang, Siheng Chen, Ya Zhang, and Yan-Feng Wang. Cat: Weakly supervised object detection with category transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3070–3079, 2021.
- [CGYX22] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *Proceedings of the European Conference on Computer Vision*, pages 192–208, 2022.

- [CHEGCN15] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.
- [CJ21] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8425–8435, 2021.
- [CKA⁺19] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6321–6330, 2019.
- [CLB⁺20] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020.
- [CLS⁺18] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018.
- [CLS20] Minjie Cai, Feng Lu, and Yoichi Sato. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14392–14401, 2020.

- [CLZ⁺21] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2703–2712, 2021.
- [CML⁺20] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308*, 2020.
- [CSSH20] Jinwoo Choi, Gaurav Sharma, Samuel Schuler, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *European Conference on Computer Vision*, pages 678–695. Springer, 2020.
- [CWAS19] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019.
- [CZ17a] 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*, pages 6299–6308, 2017.
- [CZ17b] 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*, pages 6299–6308, 2017.
- [DDF⁺20] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020.

- [DHG⁺18] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in neural information processing systems*, 2018.
- [DLCD21] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021.
- [DSVG17] Ali Diba, Vivek Sharma, and Luc Van Gool. Deep temporal linear encoding networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2329–2338, 2017.
- [FDdCW⁺11] Nazli Faraji Davar, Teofilo de Campos, David Windridge, Josef Kittler, and William Christmas. Domain adaptation in the context of sport video action recognition. In *Domain Adaptation Workshop, in conjunction with NIPS*, 2011.
- [FFMH19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019.
- [FG19] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3575–3584, 2019.
- [FLR12] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327. Springer, 2012.
- [GCDZ19] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition, pages 244–253, 2019.

- [GDSX19] Mingfei Gao, Larry Davis, Richard Socher, and Caiming Xiong. WSLN: weakly supervised natural language localization networks. In *Empirical Methods in Natural Language Processing*, 2019.
- [GG16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, pages 1050–1059, 2016.
- [GG21] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. *arXiv preprint arXiv:2106.02036*, 2021.
- [GGCN19] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 245–253. IEEE, 2019.
- [GHXL21] Dayan Guan, Jiaying Huang, Aoran Xiao, and Shijian Lu. Domain adaptive video segmentation via temporal consistency regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8053–8064, 2021.
- [GJ02] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002.
- [GSYN17] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5267–5275, 2017.
- [GUA⁺16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of

neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

- [HCL⁺20] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020.
- [HCLS18] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European Conference on Computer Vision*, pages 754–769, 2018.
- [HCS20] Yifei Huang, Minjie Cai, and Yoichi Sato. An ego-vision system for discovering human joint attention. *IEEE Transactions on Human-Machine Systems*, 50(4):306–316, 2020.
- [HHK18] Haoshuo Huang, Qixing Huang, and Philipp Krahenbuhl. Domain transfer through deep activation matching. In *Proceedings of the European Conference on Computer Vision*, pages 590–605, 2018.
- [HKS18] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.
- [HLGJ21] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7199–7208, 2021.
- [HSS18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

- [HSS20] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14024–14034, 2020.
- [HvC93] GE Hinton and Drew van Camp. Keeping neural networks simple by minimising the description length of weights. In *Proceedings of COLT-93*, pages 5–13, 1993.
- [HZC⁺17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [HZG⁺19] Dongliang He, Zhichao Zhou, Chuang Gan, Fu Li, Xiao Liu, Yandong Li, Limin Wang, and Shilei Wen. Stnet: Local and global spatial-temporal modeling for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8401–8408, 2019.
- [HZH⁺19] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8393–8400, 2019.
- [HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [JGK⁺22] Xue-Bo Jin, Wen-Tao Gong, Jian-Lei Kong, Yu-Ting Bai, and Ting-Li Su. A variational bayesian deep network with data self-screening layer for massive time-series data forecasting. *Entropy*, 24(3):335, 2022.

- [JKFN20] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10233–10244, 2020.
- [JNDV18] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *Proceedings of the British Machine Vision Conference*, volume 2, page 4, 2018.
- [JSIK20] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13289–13299, 2020.
- [JWJ⁺17] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1377–1385, 2017.
- [KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [KDLF17] Yu Kong, Zhengming Ding, Jun Li, and Yun Fu. Deeply learned view-invariant features for cross-view action recognition. *IEEE Transactions on Image Processing*, pages 3028–3037, 2017.
- [KG17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.
- [KHR⁺17] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In

- Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 706–715, 2017.
- [KJG⁺11] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.
- [KNZD19] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019.
- [KPvD⁺19] Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas Noldus, and Remco Veltkamp. Multitask learning to improve egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.
- [KTT19] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6232–6242, 2019.
- [KTZ⁺21] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13618–13627, 2021.
- [KWY⁺20] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander G Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *Advances in Neural Information Processing Systems*, 2020.

- [L⁺13] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Proceedings of the International Conference on Machine Learning*, 2013.
- [LCWJ15] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105, 2015.
- [LDG⁺17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [LDM⁺22] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7581–7590, 2022.
- [LDZ⁺20] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 481–497. Springer, 2020.
- [LFV⁺17] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.
- [LGG⁺18] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. Videolstm convolves, attends and flows

for action recognition. *Computer Vision and Image Understanding*, 166:41–50, 2018.

[LGH19] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

[LJS⁺20] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2020.

[LLC17] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[LLL⁺19a] Yang Liu, Zhaoyang Lu, Jing Li, Tao Yang, and Chao Yao. Deep image-to-video adaptation and fusion networks for action recognition. *IEEE Transactions on Image Processing*, 29:3168–3182, 2019.

[LLL19b] Minlong Lu, Danping Liao, and Ze-Nian Li. Learning spatiotemporal attention for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[LLR18] 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*, pages 619–635, 2018.

[LLZ⁺20] Jun Li, Xianglong Liu, Wenxuan Zhang, Mingyuan Zhang, Jingkuan Song, and Nicu Sebe. Spatio-temporal attention networks for action recognition and detection. *IEEE Transactions on Multimedia*, 22(11):2990–3001, 2020.

- [LMH⁺21] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *Proceedings of the International Conference on Learning Representations*, 2021.
- [LNxG21] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021.
- [LQD⁺21] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, 2021.
- [LQF21] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1726–1736, 2021.
- [LWS⁺18] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.
- [LYR15] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015.
- [LZWJ17] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

- [LZZ⁺20] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- [MC21] Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1069–1078, 2021.
- [MCH20] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020.
- [MCS00] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, pages 142–151, 2000.
- [MD20] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 122–132, 2020.
- [MFK16] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016.
- [MKK⁺18] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018.
- [MPRC19] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from

- text queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2019.
- [MSZH19] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6826–6833, 2019.
- [MXH⁺20a] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020.
- [MXH⁺20b] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1046–1056, 2020.
- [MXH⁺20c] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.
- [MYK⁺20] Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 156–171, 2020.
- [NBZA21] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asseilmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021.

- [Nea12] Radford M Neal. *Bayesian learning for neural networks*, volume 118. 2012.
- [NYD16] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hour-glass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.
- [OLF⁺22] Takehiko Ohkawa, Yu-Jhe Li, Qichen Fu, Rosuke Furuta, Kris M Kitani, and Yoichi Sato. Domain adaptive hand key-point and pixel localization in the wild. *Proceedings of the European Conference on Computer Vision*, 2022.
- [ONRH20] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. *Proceedings of the British Machine Vision Conference*, 2020.
- [PBTM17] Wenjie Pei, Tadas Baltrusaitis, David MJ Tax, and Louis-Philippe Morency. Temporal attention-gated model for robust sequence classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6730–6739, 2017.
- [PCAN20] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11815–11822, 2020.
- [PGM⁺19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc,

E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019.

- [PM13] Barbara Plank and Alessandro Moschitti. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1498–1507, 2013.
- [PMRC22] Sudipta Paul, Niluthpol Chowdhury Mithun, and Amit K Roy-Chowdhury. Text-based temporal localization of novel events. In *Proceedings of the European Conference on Computer Vision*, pages 567–587, 2022.
- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.
- [QYX18] Fan Qi, Xiaoshan Yang, and Changsheng Xu. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 429–437, 2018.
- [RYW18] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *Proceedings of the European Conference on Computer Vision*, pages 347–363, 2018.
- [SAJ16] Suriya Singh, Chetan Arora, and CV Jawahar. First person action recognition using deep learned descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2620–2628, 2016.
- [SCD⁺17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via

- gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [SEL19] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9954–9963, 2019.
- [SGS⁺18] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7396–7404, 2018.
- [SL18] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. *arXiv preprint arXiv:1807.11794*, 2018.
- [SLLG20] Sijie Song, Jiaying Liu, Yanghao Li, and Zongming Guo. Modality compensation network: Cross-modal adaptation for action recognition. *IEEE Transactions on Image Processing*, 29:3957–3969, 2020.
- [SLM⁺18] Darsh Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. Adversarial domain adaptation for duplicate question detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1056–1063, 2018.
- [SMJ⁺16] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1961–1970, 2016.
- [SR21] Maitreya Suin and AN Rajagopalan. Gated spatio-temporal attention-guided video deblurring. In *Proceedings of the IEEE*

- Conference on Computer Vision and Pattern Recognition*, pages 7802–7811, 2021.
- [SS16] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.
- [STDE19] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.
- [SUH17] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 2988–2997. PMLR, 2017.
- [SVW⁺16] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision*, pages 510–526, 2016.
- [SWM⁺20] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*, 2020.
- [SWUH18] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.
- [SY19] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5208–5217, 2019.

- [SZ14] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.
- [SZS12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [SZY⁺21] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9787–9795, 2021.
- [TBF⁺15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
- [TBL⁺19] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558, 2019.
- [THSD17] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.
- [TV17] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 2017.

- [TXSP21] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2083–2092, 2021.
- [VdMH08] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [VSP⁺17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [WBLB21] Zachary Wharton, Ardhendu Behera, Yonghuai Liu, and Nik Bessis. Coarse temporal attention network (cta-net) for driver’s activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1279–1289, 2021.
- [WCJ21] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *ACM MM*, pages 1459–1468, 2021.
- [WDZL22] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. Weakly supervised temporal adjacent network for language grounding. *IEEE TMM*, 24:3276–3286, 2022.
- [WFF⁺19] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.
- [WHL⁺21] Liwei Wang, Jing Huang, Yin Li, Kun Xu, Zhengyuan Yang, and Dong Yu. Improving weakly supervised visual grounding

- by contrastive knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14090–14100, 2021.
- [WJQ⁺17] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.
- [WLC⁺16] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE TPAMI*, 39(11):2314–2320, 2016.
- [WLHL20] Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos. In *ACM MM*, page 1283–1291, 2020.
- [WLY⁺19] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5345–5352, 2019.
- [WPLK18] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision*, pages 3–19, 2018.
- [WPQ20] Jiaze Wang, Xiaojiang Peng, and Yu Qiao. Cascade multi-head attention networks for action recognition. *Computer Vision and Image Understanding*, 192:102898, 2020.
- [WTF20] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020.

- [WXM⁺19] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2019.
- [WXW⁺16] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.
- [WZK⁺20] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.
- [WZWY20] Xiaohan Wang, Linchao Zhu, Yu Wu, and Yi Yang. Symbiotic attention for egocentric action recognition with object-centric alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [YHSS21] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Stacked temporal attention: Improving first-person action recognition by emphasizing discriminative clips. *arXiv preprint arXiv:2112.01038*, 2021.
- [YWCH19] Chaohui Yu, Jindong Wang, Yiqiang Chen, and Meiyu Huang. Transfer learning with dynamic adversarial adaptation network. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 778–786. IEEE, 2019.
- [YZY⁺21] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 53–63, 2021.

- [YZZW21] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE TIP*, 30:3252–3262, 2021.
- [ZAOT18] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision*, pages 803–818, 2018.
- [ZCSG16] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *Proceedings of the European Conference on Computer Vision*, pages 766–782, 2016.
- [ZDG17] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030, 2017.
- [ZDW⁺19] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1247–1257, 2019.
- [ZG09] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.
- [ZHC⁺22] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15555–15564, 2022.

- [ZHCL22] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page 3, 2022.
- [ZHCY21] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE TPAMI*, 44(9):5866–5885, 2021.
- [ZKL⁺16] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
- [ZLZ⁺20] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *ACM MM*, pages 4098–4106, 2020.
- [ZS13] Fan Zhu and Ling Shao. Enhancing action recognition by cross-domain dictionary learning. In *Proceedings of the British Machine Vision Conference*. Citeseer, 2013.
- [ZS19] Jiaojiao Zhao and Cees GM Snoek. Dance with flow: Two-in-one stream action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9935–9944, 2019.
- [ZWK⁺18] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European Conference on Computer Vision*, pages 597–613, 2018.
- [ZXZO21] Weichen Zhang, Dong Xu, Jing Zhang, and Wanli Ouyang. Progressive modality cooperation for multi-modality domain adaptation. *IEEE Transactions on Image Processing*, 30:3293–3306, 2021.

- [ZYKW18] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Un-supervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision*, pages 289–305, 2018.
- [ZYP⁺18] Zheming Zuo, Longzhi Yang, Yonghong Peng, Fei Chao, and Yanpeng Qu. Gaze-informed egocentric action recognition for memory aid systems. *IEEE Access*, 6:12894–12904, 2018.
- [ZZL⁺20] Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in neural information processing systems*, pages 18123–18134, 2020.

Publications

Publications Related to the Thesis

- [1] **Lijin Yang**, Yifei Huang, Yusuke Sugano, and Yoichi Sato. “Stacked Temporal Attention: Improving First-person Action Recognition by Emphasizing Discriminative Clips”. in Proc. *British Machine Vision Conference (BMVC 2021)*, November 2021.
- [2] **Lijin Yang**, Yifei Huang, Yusuke Sugano, and Yoichi Sato. “Interact before Align: Leveraging Cross-Modal Knowledge for Domain Adaptive Action Recognition”. in Proc. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, June 2022.

Other Publications

- [1] Yifei Huang, Xiaoxiao Li, **Lijin Yang**, Lin Gu, Yinqying Zhu, Hirofumi Seo, Qiuming Meng, Tatsuya Harada, and Yoichi Sato. “Leveraging Human Selective Attention for Medical Image Analysis with Limited Training Data”. in Proc. *British Machine Vision Conference (BMVC 2021)*, November 2021.
- [2] **Lijin Yang**, Yifei Huang, Yusuke Sugano, and Yoichi Sato. “EPIC-KITCHENS-100 Unsupervised Domain Adaptation Challenge for Action

Recognition 2021: Team M3EM Technical Report”. *The Eighth International Workshop on Egocentric Perception, Interaction and Computing (EPIC 2021)*, June 2022.

- [3] Yifei Huang, **Lijin Yang** and Yoichi Sato. “Compound Prototype Matching for Few-shot Action Recognition”. in Proc. *European Conference on Computer Vision (ECCV 2022)*, October 2022.