UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Evolving Video Analysis: From Object Perception to Holistic Understanding

by Mingfei Han

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Professor Xiaojun Chang

University of Technology Sydney
Faculty of Engineering and Information Technology

August 2024

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Mingfei Han, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and IT at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature: Signature removed prior to publication.

Date: 14 August 2024

ABSTRACT

Evolving Video Analysis: From Object Perception to Holistic Understanding

by

Mingfei Han

The domain of video content analysis has experienced rapid advancements due to the proliferation of digital video content and the evolving capabilities of computer vision technologies. Despite these advancements, significant challenges remain in both video object perception and holistic video understanding, which are crucial for applications ranging from autonomous driving to interactive media. This thesis aims to address these challenges by developing innovative methodologies that enhance the accuracy and efficiency of video analysis systems.

In the area of video object perception, this research tackles the problem of accurately detecting, categorizing and referring objects within video frames under various conditions. Key contributions include the Hierarchical Video Relation Network (HVR-Net), which utilizes inter-video proposal relations to enhance object detection accuracy, and Progressive Frame-Proposal Mining (PFPM), which leverages sparse annotations to improve detection in a weakly supervised context. Additionally, the Hybrid Temporal-scale Multimodal Learning (HTML) framework is introduced to refine the segmentation of objects based on textual descriptions, effectively bridging the gap between visual content and language inputs.

For holistic video understanding, this thesis introduces methodologies and datasets that significantly improve the interpretation of complex video scenes and dynamics. The Dual-AI framework employs dual paths to innovatively combine spatial and temporal data, enhancing the analysis of individual actions and group dynamics for more accurate recognition of complex group activities. Additionally, we have de-

veloped specialized methodologies for Portrait Mode Video recognition, optimizing video analysis techniques for the vertical video format commonly found on social media, thereby addressing its unique challenges. Furthermore, the Shot2Story20K dataset establishes a new benchmark for multi-shot video understanding, facilitating detailed narrative synthesis across sequential shots to enrich the storytelling potential of video content analysis.

In conclusion, this thesis contributes a suite of methodologies and datasets that enhance both the foundational aspects of video object perception and the broader capabilities of video understanding. These innovations not only address the current limitations of video analysis technologies but also lay the groundwork for future advancements, suggesting a path forward for the integration of even more sophisticated machine learning models into video content analysis systems. Through these efforts, the thesis demonstrates significant progress in making video analysis more robust, adaptable, and context-aware, aligning more closely with human-level perception and interpretation.

Dissertation directed by Professor Xiaojun Chang,
Australian Artificial Intelligence Institute, University of Technology Sydney

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Xiaojun Chang, for his unwavering guidance and expertise throughout my Ph.D. journey. Together, we have pioneered advancements in video analysis, evolving from CNN-based approaches to the transformer era and the exploration of LLM-based video analysis. His support has been invaluable, not only in my academic pursuits but also in life and research. I also extend my sincere thanks to my associate supervisor, Prof. Yi Yang, for his passion for research and his enthusiasm for tackling new research challenges.

I am particularly grateful to Prof. Yu Qiao and Prof. Yali Wang for hosting me as a visiting student during the early stages of my academic journey. Working with them on various video object perception projects greatly benefited my research, and I deeply appreciate their dedicated research spirits and insightful guidance. My thanks also go to Dr. Heng Wang, Dr. Linjie Yang, and Dr. Xiaojie Jin for their support in exploring new topics, including pioneering work on new video formats such as portrait mode and multi-shot videos. The insightful discussions and teamwork we shared have greatly enriched my research experience.

I would like to acknowledge the company and thoughtful discussions I shared with my mates at Monash University and UTS: Dr. Mingjie Li, Dr. Changlin Li, Dr. Siyi Hu, Mrs. Rui Liu, Mrs. Yuetian Weng, Mr. Sihao Lin, Dr. Zizheng Pan, Dr. Haoyu He, Mr. Qizhou Wang, Dr. Guangrui Li, Mr. Liulei Li, and Dr. Tianqi Tang. Their support and companionship have been greatly valued. Additionally, I am deeply thankful to my friends and mates in China for their unwavering support during challenging times: Dr. Longxuan Kou, Ms. Qingling Jia, Mrs. Chenhui Wang, Dr. Chihang Yang, Mr. Yunda Sun, Dr. Haoran Liang, Dr. Shiyu Xuan,

Dr. Zhicheng Huang, Mr. Zhongwei Ren, Mr. Luting Wang, and Mr. Xin Gu.

Finally, my deepest thanks go to my parents and family members for their unconditional support throughout this journey, especially during the challenging times of visa issues, COVID-19, and all other hurdles. Your unwavering belief in me has been my greatest source of strength.

Mingfei Han

Sydney, Australia

July, 2024

List of Publications

Conference Paper:

C-1 **Mingfei Han**, Yali Wang, Xiaojun Chang, Yu Qiao. "Mining Inter-Video Relations for Video Object Detection", In *European Conference on Computer Vision (ECCV 2020).*

C-2 **Mingfei Han**, David Junhao Zhang, Yali Wang, Ruiyan, Lina Yao, Xiaojun Chang, Yu Qiao. "Dual-AI: Dual-path Actor Interaction Learning for Group Activity Recognition", In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022).*

C-3 **Mingfei Han**, Yali Wang, Zhihui Li, Lina Yao, Xiaojun Chang, Yu Qiao. "HTML: Hybrid Temporal-scale Multimodal Learning Framework for Referring Video Object Segmentation", In *IEEE/CVF International Conference on Computer Vision (ICCV 2023).*

C-4 **Mingfei Han**, Linjie Yang, Xiaojie Jin, Jiashi Feng, Xiaojun Chang, Heng Wang. "Video Recognition in Portrait Mode", In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024).*

C-5 **Mingfei Han**, Linjie Yang, Xiaojun Chang, Heng Wang. "Shot2Story20K: A New Benchmark for Comprehensive Understanding of Multi-shot Videos", In *submission.*

Journal Paper:

J-1 **Mingfei Han**, Yali Wang, Mingjie Li, Xiaojun Chang, Yi Yang, Yu Qiao. "Progressive Frame-Proposal Mining for Weakly Supervised Video Object Detection", In *IEEE Transactions on Image Processing (Volume 33, 1560-1573).*

Contents

List of Figures

Chapter 1

Introduction

In the realm of computer vision, video content analysis stands as a cornerstone, bridging raw visual data with actionable insights and interpretations. This discipline is pivotal not only in advancing academic research but also in powering a multitude of industrial applications ranging from autonomous driving and surveillance to content recommendation and interactive media. As digital video content proliferates at an unprecedented rate, driven by social media platforms and the increasing capabilities of consumer electronics, the demand for sophisticated video analysis technologies has surged. These technologies are tasked with providing precise and context-aware interpretations that can inform decision-making and automate complex processes.

Within the domain of video content analysis, this thesis primarily investigates two critical perspectives: object perception and holistic understanding. Object perception, which includes tasks like object detection and segmentation, focuses on identifying and delineating individual elements within video frames. This forms the basis for more complex operations, as it allows systems to recognize and track objects across sequences, crucial for applications such as traffic monitoring and activity recognition. On the other hand, holistic understanding encompasses a broader spectrum, aiming to interpret entire scenes and actions within videos, engaging areas such as video recognition, captioning, and summarization. This branch delves into the narrative and dynamic context of videos, striving to emulate human-like comprehension and responsiveness to audio-visual streams, thereby enabling deeper interactions between humans and machines.

In the domain of video object perception, my research begins by examining how to establish effective relationships among all candidate object proposals to mitigate categorical confusion. This effort stems from the understanding that while candidate proposals effectively cover the objects within a video, accurately categorizing these proposals remains a significant challenge. Recognizing the impracticality of densely annotating every frame in general video object detection due to the extensive labor and time required, I explore the less resource-intensive approach of weakly supervised video object detection. Here, I introduce a progressive frame-proposal mining method that constructs holistic proposal relations across frames, enhancing detection accuracy with minimal annotations.

Building upon the necessity for temporal dynamics in understanding video content, I propose a hybrid temporal relation modeling technique within the context of referring video object segmentation. This approach is designed to align textual instructions of varying lengths with the dynamics of video frames sampled at different rates, ensuring precise object segmentation in response to user queries.

In the domain of video holistic understanding, my research initially focuses on the enhancement of holistic comprehension through the meticulous construction and refinement of individual relationships within group activities. By analyzing and reinforcing individual actions and interactions, this study highlights the essential synergy between detailed individual relation construction and the broader understanding of group dynamics, facilitating deeper insights into collective behaviors.

Further exploring current trends in video content, my research shifts to the emerging prevalence of portrait mode videos on social media platforms, which present unique challenges due to their distinct data distribution. To address this, I introduce a study on portrait mode videos, focusing on video recognition tasks, supported by a newly developed dataset tailored specifically for this format. This initiative aims

to adapt video analysis techniques to better handle the aspect ratio and content focus unique to portrait mode videos.

Lastly, I address the complex problem of holistic understanding from the perspective of textual captioning and summarization in multi-shot videos. Recognizing the distinct narrative structures and interconnected events typical of multi-shot content, I propose a new benchmark and develop an innovative method for generating detailed textual summaries. This method leverages the capabilities of large language models to enhance narrative coherence and contextual relevance across various shots, thereby improving the interpretability and utility of video summaries.

The contributions of the thesis are listed below.

- I investigate the modeling of inter-proposal relations across different frames to capture temporal dynamics and categorical clues essential for video object perception. This research develops effective methodologies tailored to various tasks, including video object detection, weakly supervised video object detection, and referring video object segmentation. Extensive experiments demonstrate that these methods successfully capture the inter-frame and inter-proposal dynamics and categorical consistency across diverse perception tasks.

- I delve into recognizing and describing videos through holistic understanding. To facilitate this, I explore the dynamics between individuals to enhance group activity comprehension. Additionally, I concentrate on two specific video formats: portrait mode and multi-shot videos. For these, I propose new benchmarks and develop methodologies that address challenges in video recognition and textual summarization effectively.

The following introduces the background and developed methodologies for video object perception and holistic understanding.

1.1 Video Object Perception

In Chapter 3, I tackle video object detection challenges by exploiting inter-video proposal relations to mitigate categorical confusion. Traditional methods focus on intra-video dynamics, but my approach, the Hierarchical Video Relation Network (HVR-Net), integrates relations both within and across videos. This strategy enhances object distinction and accuracy by leveraging spatio-temporal contexts from multiple sources, providing a richer understanding of similar-looking objects across different categories. HVR-Net's effectiveness is validated on the ImageNet VID benchmark, where it significantly improves classification accuracy and sets new performance standards by utilizing comprehensive inter-video insights to reinforce categorical consistency and detection precision.

In Chapter 4, I explore the realm of weakly supervised video object detection, which presents the challenge of detecting objects with minimal manual annotation. Traditional methods rely heavily on densely annotated frames, which are time-consuming and often impractical for large-scale applications. To address this, I introduce a novel approach known as Progressive Frame-Proposal Mining (PFPM), which efficiently utilizes sparse labels to improve detection accuracy across video frames. PFPM innovatively constructs holistic proposal relations by progressively mining data from coarser to finer details, adapting the detection model to effectively utilize available annotations. This method not only reduces the dependency on extensive manual labeling but also enhances the model's ability to generalize from limited data. Validated on the ImageNet VID benchmark, PFPM demonstrates substantial improvements over existing weakly supervised methods, proving its efficacy in leveraging minimal supervision to achieve robust detection results.

In Chapter 5, I address the complex challenge of referring video object segmentation, where the objective is to segment objects from video frames based on a textual

description. Existing methods often struggle with variations in temporal dynamics and the multimodal nature of instructions. To overcome these limitations, I introduce the Hybrid Temporal-scale Multimodal Learning (HTML) Framework, which enhances segmentation accuracy by aligning dynamic visual content with varying lengths of textual descriptions. The HTML framework uniquely combines intra-scale and inter-scale multimodal learning, allowing for effective integration of textual and visual data across different temporal scales. This approach ensures that the model captures both the immediate and extended context of the object in question, improving the precision and relevance of segmentation in response to descriptive queries. Extensive testing on benchmarks like Ref-YouTube-VOS and Ref-DAVIS17 shows that HTML sets a new standard for accuracy in referring video object segmentation, outperforming existing state-of-the-art methods by effectively bridging the gap between language and vision.

1.2 Video Holistic Understanding

In Chapter 6, I address the sophisticated challenge of group activity recognition, which requires understanding both individual actions and their collective dynamics within a group. Traditional methods often struggle with capturing the complex interplay of spatial and temporal factors influencing group activities. To enhance the analysis, I introduce Dual-AI, a Dual-path Actor Interaction Learning framework that innovatively integrates spatial and temporal data to improve recognition accuracy. Dual-AI employs two complementary paths: the Spatial-Temporal (ST) path and the Temporal-Spatial (TS) path. Each path processes actor interactions differently, allowing the framework to adaptively capture diverse group dynamics. The ST path analyzes spatial relationships first, followed by temporal dynamics, while the TS path reverses this order, catering to different activity patterns. This dual-path strategy ensures a comprehensive understanding of both individual behaviors

and group interactions, significantly boosting the accuracy of activity recognition. Tested on challenging benchmarks like Volleyball and Collective Activity datasets, Dual-AI demonstrates its superiority by not only achieving state-of-the-art performance but also showcasing robust adaptability across various group activities.

In Chapter 7, I explore the unique challenges of video recognition in portrait mode, a format increasingly prevalent on social media platforms. Traditional video recognition algorithms are primarily optimized for landscape mode, which does not align well with the aspect ratio and content characteristics of portrait mode videos. These videos often focus more on subjects with limited background context and include distinctive spatial distributions. To address these specific challenges, I introduce the PortraitMode-400 dataset, specifically designed to support the development and evaluation of recognition algorithms tailored for portrait mode videos. This dataset features a diverse range of categories reflective of typical portrait video content, ensuring relevance and applicability to real-world scenarios. Building on this dataset, I develop and test various methodological enhancements that optimize recognition techniques for the vertical video format. These adaptations include specialized data augmentation strategies, tailored cropping approaches, and modified network training protocols that better accommodate the unique properties of portrait mode videos. The methodologies demonstrated in this chapter not only enhance the accuracy of video recognition in portrait mode but also provide a framework for future research in adapting video analysis tools to other non-standard video formats.

In Chaper 8, I address the complex challenge of understanding and summarizing multi-shot videos, which encompass multiple scenes and events within a single video stream. Traditional video captioning approaches often struggle to adequately represent the narrative complexity and temporal transitions inherent in multi-shot content. To enhance the capability of video analysis systems in capturing the detailed progression and interrelationships between shots, I introduce the Shot2Story20K

dataset. Shot2Story20K is specifically designed for the audio-visual understanding of multi-shot videos, featuring detailed annotations that describe both visual content and corresponding audio narratives across various shots. This dataset encourages the development of models that can integrate and contextualize the rich information presented in both visual and auditory domains. Utilizing Shot2Story20K, I propose a novel framework that leverages state-of-the-art large language models to generate comprehensive video summaries. This method innovatively combines shot-specific descriptions with overarching narrative synthesis, providing a cohesive and detailed summary of complex video content. Extensive experiments demonstrate that this approach not only outperforms existing captioning methods but also offers new insights into the effective integration of multimodal information for video understanding.

1.3 Thesis Organization

This thesis is structured as follows:

In Chapter 2, a survey of video object perception and holistic understanding is presented.

Chapters 3 to 5 sequentially explore the video object perception via inter-frame relation construction across different tasks. More specifically, Chapter 3 introduces the Hierarchical Video Relation Network (HVR-Net), which enhances object detection by leveraging inter-video proposal relations. Chapter 4 details a novel method for improving detection with minimal annotations. It addresses the challenges of weakly supervised learning in video object detection. Chapter 5 proposes a framework that aligns textual descriptions with video content across differing temporal scales, improving segmentation precision.

In Chapters 6 to 8, the investigation shifts towards video holistic understand-

ing from both video recognition and textual summarization. Chapter 6 explores dual-path learning to analyze individual actions and group dynamics simultaneously, enhancing the recognition of complex group activities. Chapter 7 focuses on adapting video recognition techniques to the portrait video format, highlighting the development of a new dataset and methodologies tailored for social media content. Chapter 8 presents a new benchmark and summarization method designed to address the complexities of multi-shot video content, integrating multimodal information for narrative synthesis.

Finally, Chapter 9 provides a brief summary of the thesis and discusses potential directions for future exploration.

Chapter 2

Literature Review

This chapter introduces a survey of related work in video analysis, mainly encompassing video object perception and video holistic understanding.

2.1 Video Object Perception

2.1.1 Video Object Detection

Object Detection in Still Images. The field of object detection in still images [33, 57, 58, 71, 116, 139, 141] has made considerable progress due to advancements in deep neural networks [69, 89, 154, 161, 202] and the availability of large-scale, well-annotated datasets [113, 145]. Existing methods generally fall into two categories: two-stage detectors like RCNN [58], Fast-RCNN [57], and Faster-RCNN [141] which prioritize accuracy, and one-stage detectors such as YOLO [139], SSD [116], and RetinaNet [112] which are optimized for computational efficiency. Additionally, recent advancements in anchor-free detection [44, 95, 230, 231] have demonstrated impressive performance. Nonetheless, adapting these image-based detection methods to video poses challenges, particularly due to motion blur and object occlusion inherent in video sequences.

Object Detection in Videos. Enhancing still image detection techniques for video applications often involves leveraging temporal information to manage the continuity and dynamics of objects across frames [51, 65, 77]. Methods like box-level association help form object trajectories, while feature aggregation techniques use adjacent frames to enrich the current frame's features [12, 35, 153, 200, 235].

These strategies have shown that understanding proposal relationships across different frames can alleviate detection challenges in videos [35, 153] through modeling long-term dependencies [174, 190]. Despite these advances, there's a tendency to overlook inter-video relationships, which can be critical when objects have similar appearances across different videos. To address this, we propose HVR-Net, a novel approach that integrates both intra-video and inter-video proposal relations to enhance detection accuracy.

2.1.2 Weakly Supervised Object Detection

In the domain of still images, weakly-supervised detectors have been investigated without bounding box annotations. Most approaches formulate this problem as multiple instance learning [14, 79, 176, 194, 224], i.e., each image is considered as a bag of candidate instances, and at least one instance belongs to the object class. In recent years, convolution network based MIL frameworks have been actively studied [14, 164, 165] to boost performance. Specifically, most weakly supervised object detection in the following two-stage manner.

The first stage is proposal generation. Since there are no bounding box annotations available, Selective Search [171] and/or EdgeBox [237] are widely used to generate off-the-shelf object proposals. Few recent works introduce additional proposal generation modules to obtain more qualified proposals. For example, [166] uses a small network [165] to further refine the coarse edge boxes. [38] uses the predicted map of CAM [227] as semantic guidance of proposal selection. [194] proposes an objectness score evaluation method for selection, based on the cascaded structure of [38].

The second stage is proposal mining and network training to obtain detection results. One of the most well-known approaches is two-stream weakly supervised deep detection network (WSDDN) [14], which simultaneously performs region selection

and classification towards end-to-end learning. Subsequently, a number of non-trivial extensions have been developed by refining instances. For example, [165, 164, 211] uses top-scoring proposals from the MIL network as supervision to train instance refinement classifiers. [164] further uses object clusters to assign pseudo labels for object proposals. [224] proposes to generalize detector by progressively increasing learning difficulty from easy to hard examples. [86] introduces a context classification loss to find a region covering the whole object to refine the classification. [109] designs a spatial and appearance graph within image to mine high quality proposals for classification refinement. Additionally, several approaches propose to optimize the overall network. For example, [177] regularizes detector with object proposal cliques to alleviate localization randomness during learning instances. [148] introduces generative adversarial learning to train a fast detector. [9] employs a discrete generative network to model annotation aware conditional distribution for proposal labeling. [176] optimizes a series of smoothed loss functions to alleviate the non-convexity problem of deep-MIL methods.

Due to the weakly supervised characteristics in video object detection, we follow the above two-stage manner to develop the detection framework. However, different from image-based detection, we work on video-based detection with complex object and/or camera motions, object disappearance among frames, etc.

2.1.3 Relevant Weakly Supervised Video Tasks

There exist several video tasks in the weakly-supervised settings, such as action detection [8, 155, 183], action-driven object detection [214, 222], video object segmentation [225, 223], object localization [75, 223] and video object grounding [149, 213]. However, these existing weakly-supervised tasks are either based on extra prior knowledge of specific tasks or with guidance of extra modality. For example, weakly supervised action detection [8, 155] and action-driven object de-

tection [214] need extra person box annotations, which are used to pre-train person detector for generating person proposals as prerequisites of detection. Moreover, action-driven object detection [214, 222] requires extra action phrase annotations, besides of object tags. Similarly, weakly-supervised video object segmentation [225] needs fully-annotated object detection benchmarks to provide object location information. Video object localization [75, 223] applies strong data scenario prior, where objects belong to a single category and appear in each frame of video and additional fully annotated data [29] is needed to pre-train the network[223]. Lastly, weakly supervised video object grounding [149, 213] uses extra natural language description to detect objects. Different from all these high-level video understanding tasks, this thesis considers a more fundamental and challenging task that has not been explored, i.e., detecting objects only with their tags in the video. Addressing such task can beneficial for both research and industry, due to its wide applications in practice.

2.1.4 Referring Video Object Segmentation

Vision-only Video Segmentation. Tasks such as video instance segmentation (VIS) [212] and video object segmentation (VOS) [135, 134] demand precise segmentation of objects within predefined semantic queries, often requiring sophisticated models to track each instance consistently across frames. Earlier approaches rely on heavy supervision and complex algorithms to manage instance associations across frames, whereas recent developments leverage transformer models [174, 19] for more integrated, end-to-end segmentation solutions.

Referring Video Object Segmentation. Referring Video Object Segmentation (RVOS) [147] involves segmenting objects based on detailed, open-world textual descriptions, posing unique challenges in the integration of visual and textual data. Most early methods in RVOS proposed to refer the object by applying image-level

methods on video frames separately and associate them with heuristic rules. However, they usually fail to utilize the temporal dynamic. [147] casts the task as a joint problem of referring segmentation in frame and mask propagation across frames by a memory attention module. [108] proposed a top-down pipeline by constructing exhaustive set of object tracklets and then selecting the target by matching the language features with the all the candidate tracklets. [226] proposed to model the temporal dynamic with an additional optical flow modality. [196] argued the importance of the structural information of video content and proposed to utilize the frame, object and video features simultaneously to obtain better representation. MTTR [17] introduced the DETR structure to RVOS area and [198] proposed to use language-conditional queries to simplify the referring pipeline and improve the performance, which serves as our baseline. Different from the previous works, this thesis raises the mismatch issue that the various descriptions of different objects are corresponding to different temporal scales of the video. Moreover, a concise HTML framework is proposed via multimodal interaction across different temporal scales to capture the core object semantics in the video.

2.2 Video Holistic Understanding

2.2.1 Group Activity Recognition

The increasing complexity of video content and its applications has driven research in Group Activity Recognition, which has transitioned from relying on basic hand-crafted features to more advanced deep learning models that better handle the spatial and temporal dynamics of group interactions [11, 74]. Early approaches are based on hand-crafted features and typically use probabilistic graphical models [1, 3, 2, 93, 94, 193] and AND-OR grammar methods [4, 151]. Recently, methods incorporating convolutional neural networks [11, 74] and recurrent neural networks [186, 207, 137, 11, 36, 152, 104, 74, 73] have achieve remarkable performance, due to the learning of temporal context and high-level information.

More recent group activity recognition methods [199, 54, 70, 209, 45, 136, 103, 220] often require the explicit representation of spatiotemporal relations, dedicated to apply attention-based methods to model the individual relations for inferring group activity. [199, 220] build relational graphs of the actors and explore the spatial and temporal actor interactions in the same time with graph convolution networks. These methods simulate spatiotemporal interaction of actors in a joint manner. Differently, [209] builds separate spatial and temporal relation graphs subsequently to model the actor relations. [54] encodes temporal information with I3D [22] and constructs spatial relation of the actors with a vanilla transformer. [103] introduces a cluster attention mechanism for better group informative features with transformers. Different from previous approaches, we propose to learn the actor interactions in complementary Spatial-Temporal and Temporal-Spatial views and further promote actor interaction learning with a designed self-supervised loss for effective representation learning.

2.2.2 Video Recognition Datasets

Video recognition research is heavily reliant on the quality and diversity of available datasets. Historically, datasets were often created in controlled environments, like KTH [146] and Weizmann [15], which allowed for focused study on specific actions under ideal conditions. More recently, the shift has moved towards using datasets compiled from internet sources such as YouTube, which present a more realistic and challenging set of conditions due to their variability and complexity, examples of which include UCF101 [158] and HMDB51 [90]. These datasets have spurred the development of sophisticated models capable of handling complex video data, facilitating significant advancements in video recognition technologies.

2.2.3 Video Description Datasets

The ability to describe video content accurately is fundamental to numerous applications, necessitating comprehensive video description datasets. Traditional

datasets like MSRVTT [204] and ActivityNet Captions [88] have provided platforms for benchmarking captioning and description algorithms. Compared to existing video description datasets, our contributed dataset is more challenging due to the explicit modeling of the multi-shot nature of web videos. Our textual description includes both shot-level captions and video-level summaries, combining visual and audio understanding, which provides a unique test bed for multi-modal video understanding. Most existing video captioning benchmarks, such as MSRVTT [204], YouCook2 [229] and ActivityNet Caps [88], also use multi-shot videos as annotation source, but they either annotate a holistic caption for the video (MSRVTT) or ask annotators to decide the boundary of different events. In our study, we observe that video shots naturally create a sequence of related events, motivating us to annotate distinct captions for each shot. Ego4D [59] only annotates dense visual captions but not audio captions for relatively long egocentric videos. Video Storytelling [98] is a small-scale dataset with annotations of multiple events in a videos and provides a summary of the video by concatenating all captions.

A recent work VAST [27] feeds generated video and audio captions into an LLM to generate video summary. However, it processes multi-shot video as a whole and lacks the granularity of the events in different shots. Moreover, VAST directly uses predicted captions without any human verification, leading to potentially noisy and biased summaries towards the captioning models. Our dataset stands out from VAST with its accurately annotated visual and narration shot captions. Although our video summary is also generated using an LLM, it is further verified by annotators to make sure there is no hallucinated details from the LLM. Our dataset has an average length of 218.3 words for the video summary, which is much longer than existing benchmarks, and is longer than the combined length of captions in one video in ActivityNet and YouCook2.

Chapter 3

Mining Inter-Video Proposal Relations for Video Object Detection

Expanding on the core concepts introduced in Chapter 2, this chapter delves into the specialized field of categorical object recognition within video content. It focuses on identifying and locating objects in video sequences, addressing the unique challenges they present. By employing the strategies of object proposal relation learning, this chapter aims to overcome these challenges and develop robust solutions for effective object detection in dynamic video environments.

3.1 Introduction

Video object detection presents unique challenges in computer vision [12, 35, 145, 153, 200, 235]. Traditional image-based object detectors [68, 116, 139, 140] often struggle with this task, primarily due to issues like motion blur, sudden occlusions, and unusual poses that are prevalent in video sequences. Recent research [35, 153] has demonstrated that modeling object proposal relations across different frames can effectively integrate spatio-temporal context and enhance representation for detection. However, these methods typically focus only on relations within the same video, facing difficulties in distinguishing objects with similar appearances or movements across different videos.

As depicted in Fig. 3.1(a), the detector incorrectly identifies *Cat* as *Dog* in the target frame t, despite leveraging spatio-temporal contexts from other support frames $t - s$ and $t + e$ to improve the current frame's proposal representation. The

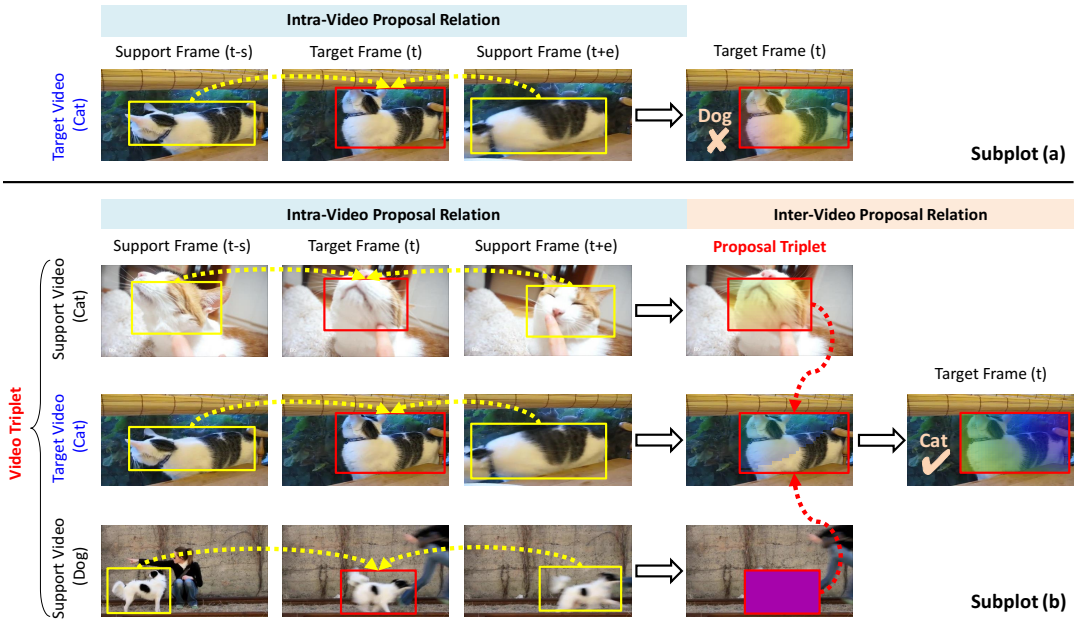


Figure 3.1 : Illustration of Motivation. Subplot (a): The intra-video proposal relationships capture the appearance and movement of *Cat* within a single video, providing limited information on variations across different videos. Consequently, the detector incorrectly identifies *Cat* as *Dog* in the target frame t, despite utilizing spatio-temporal contexts from supporting frames $t - s$ and $t + e$. Subplot (b): To address this issue, we develop a novel inter-video proposal relation module. This module is capable of adaptively identifying challenging object proposals (i.e., proposal triplet) from videos with high confusion (i.e., video triplet), enhancing the learning and correction of their relations to minimize confusion across videos. Support videos/frames provide contextual information for identifying the object of interest, while target videos/frames are the primary sequences where the detection tasks are performed.

primary issue is that intra-video relations provide limited insights on how this *Cat* compares to similar objects across various videos. For example, as shown in Fig. 3.1(b), the *Cat* in the target video resembles a *Dog* in the support video, but differs

from a *Cat* in another support video. This leads to confusion in distinguishing *Cat* from *Dog* when the analysis is restricted to individual videos without considering inter-video object relationships.

To overcome this challenge, we introduce a novel Inter-Video Proposal Relation method that effectively harnesses inter-video proposal relationships to develop discriminative representations for video object detection. Initially, we implement a multi-level triplet selection scheme to identify difficult training proposals from videos that frequently cause confusion. These selected proposal triplets are crucial for reducing confusion and are utilized to enhance object feature construction. Furthermore, we propose the advanced Hierarchical Video Relation Network (HVR-Net), which systematically integrates intra-video and inter-video proposal relation modules within a unified framework. This structure allows for the progressive utilization of both intra-video and inter-video contextual dependencies, significantly improving video object detection performance. Extensive experimental evaluations on the large-scale video object detection benchmark, ImageNet VID, demonstrate the superior performance of our HVR-Net, achieving **83.8** mAP with ResNet101 and **85.4** mAP with ResNeXt101 32x4d.

3.2 The Proposed Approach

Overview. In this section, we introduce our Hierarchical Video Relation Network (HVR-Net) designed to improve video object detection by utilizing both intra-video and inter-video contexts through a multi-level triplet selection scheme. The complete framework is illustrated in Fig. 3.2. **First**, we develop a video-level triplet selection module. For each target video, it flexibly selects two confused videos from a set of support videos—specifically, the most dissimilar video within the same category and the most similar video from different categories—based on their CNN features. This process results in a triplet of confused videos per training batch,

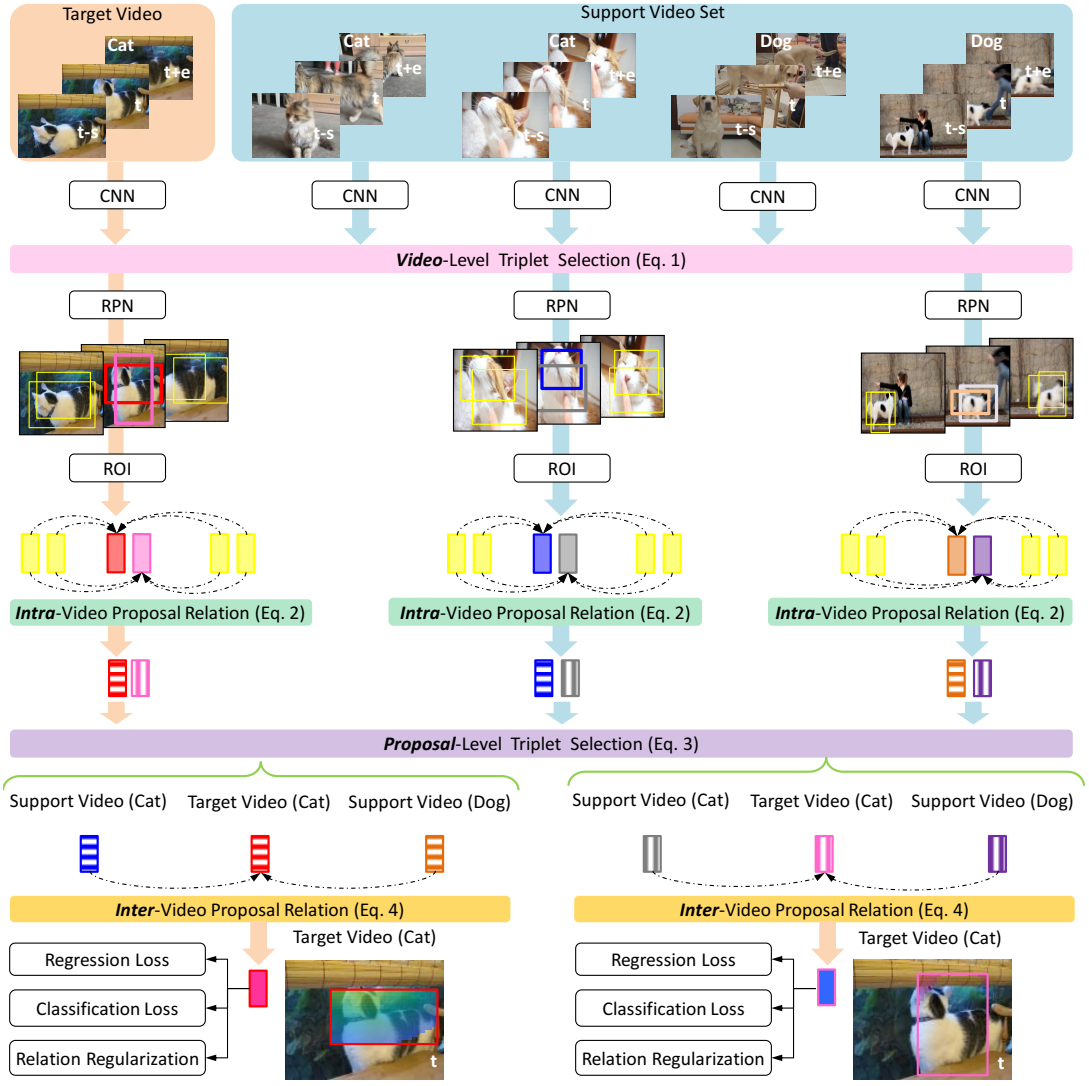


Figure 3.2 : Our HVR-Net Framework. This framework significantly enhances video object detection by progressively integrating intra-video and inter-video proposal relations within a multi-level triplet selection scheme. Further details are provided in Section 3.2.

guiding our HVR-Net to model object confusion across videos. **Second**, we introduce an intra-video proposal relation module. For each video in the triplet, we process its sampled frames (e.g., $t-s$, t, and $t+e$) through the RPN and ROI layers of Faster RCNN, generating feature vectors for object proposals in each frame. We

then aggregate proposals from support frames to enhance those in the target frame t, integrating long-term dependencies to address intra-video issues like motion blur and occlusion. **Third**, we create a proposal-level triplet selection module. While the intra-video-enhanced proposals contain object semantics for individual videos, they do not account for variations across videos. To model these variations, we select challenging proposal triplets from the video triplet based on the intra-video-enhanced features. **Finally**, we develop an inter-video proposal relation module. For each proposal triplet, this module aggregates proposals from support videos to enhance those in the target video, leveraging inter-video dependencies to mitigate object confusion.

3.2.1 Video-Level Triplet Selection

To effectively reduce inter-video confusions, we begin by identifying a triplet of challenging videos for training. Specifically, we randomly sample K object categories from the training set and then sample N videos per category, resulting in $K \times N$ videos in a batch. One video is randomly chosen as the *target video*, while the remaining $(K \times N - 1)$ videos form the set of *support videos*. For each video, we randomly sample one frame as the *target frame* t, and the other $T - 1$ frames as *support frames*, such as frame $t - s$ and frame $t + e$ in Fig. 3.2.

Each video's T frames are fed individually into the CNN backbone of Faster RCNN for feature extraction, producing a feature tensor of size $H \times W \times C \times T$, where $H \times W$ is the spatial size and C is the number of feature channels. We then perform global average pooling along the spatial and temporal dimensions, yielding a C-dimensional video representation. Based on cosine similarity between video representations, we identify the video triplet

$$\mathcal{V}^{triplet} = \{\mathcal{V}^{target}, \mathcal{V}^{+}, \mathcal{V}^{-}\}, \tag{3.1}$$

where $\mathcal{V}^{+}$ is the most dissimilar support video within the same class as $\mathcal{V}^{target}$, and

$\mathcal{V}^-$ is the most similar support video from different classes.

3.2.2 Intra-Video Proposal Relation

After identifying $\mathcal{V}^{triplet}$, we generate object proposals for each video in the triplet. We process the sampled T frames of each video through the RPN and ROI layers of Faster RCNN, producing M proposal features per frame.

Research has demonstrated that spatio-temporal proposal aggregation across different frames [35, 153] can enhance video object detection. Hence, we introduce an intra-video proposal relation module to build dependencies between target frame proposals and support frame proposals within each video. Specifically, we adapt a non-local-style relation module for $\mathcal{V}^v$ in the video triplet ($v \in \{target, +, -\}$),

$$\boldsymbol{\alpha}^v_{t,m} = \mathbf{x}^v_{t,m} + \sum_{i \in \Omega} \sum_j g(\mathbf{x}^v_{t,m}, \mathbf{x}^v_{i,j}) \times \mathbf{x}^v_{i,j}, \tag{3.2}$$

where $\mathbf{x}^v_{t,m}$ is the m-th proposal feature in the target frame t, $\mathbf{x}^v_{i,j}$ is the j-th proposal feature in the support frame i, and i belongs to the set of support frames Ω (e.g., $\Omega = \{t - s, t + e\}$). The similarity between $\mathbf{x}^v_{t,m}$ and $\mathbf{x}^v_{i,j}$ is measured using a kernel function $g(\cdot, \cdot)$, such as Embedded Gaussian [190]. We then aggregate $\mathbf{x}^v_{t,m}$ by weighted sum over all the support frame proposal features, resulting in $\boldsymbol{\alpha}^v_{t,m}$, an enhanced version of $\mathbf{x}^v_{t,m}$, which incorporates video-level object semantics to address challenges like motion blur and occlusion.

3.2.3 Proposal-Level Triplet Selection

The intra-video relation module integrates spatio-temporal object contexts within each video, but it does not account for inter-video object variations. To capture these variations, we further select challenging proposal triplets from the intra-video-enhanced proposals in the video triplet $\mathcal{V}^{triplet}$. We compare the cosine similarity between these proposals based on their features from Eq. (3.2). For a proposal

$\mathcal{P}_{t,m}^{target}$ in the target video, we identify its corresponding proposal triplet,

$$\mathcal{P}^{triplet} = \{\mathcal{P}_{t,m}^{target}, \mathcal{P}^{+}, \mathcal{P}^{-}\}, \tag{3.3}$$

where $\mathcal{P}^{+}$ is the most dissimilar proposal within the same category, and $\mathcal{P}^{-}$ is the most similar proposal from different categories.

3.2.4 Inter-Video Proposal Relation

After identifying all proposal triplets, we model the relationships among them to capture object variations across videos. We employ a non-local-style relation module for each proposal triplet,

$$\boldsymbol{\beta}_{t,m}^{target} = \boldsymbol{\alpha}_{t,m}^{target} + f(\boldsymbol{\alpha}_{t,m}^{target}, \boldsymbol{\alpha}^{+}) \times \boldsymbol{\alpha}^{+} + f(\boldsymbol{\alpha}_{t,m}^{target}, \boldsymbol{\alpha}^{-}) \times \boldsymbol{\alpha}^{-}, \tag{3.4}$$

where $f(\cdot, \cdot)$ is a kernel function (e.g., Embedded Gaussian) for similarity comparison, $\boldsymbol{\alpha}^{+}$ is the intra-video-enhanced feature of the positive proposal $\mathcal{P}^{+}$, and $\boldsymbol{\alpha}^{-}$ is the intra-video-enhanced feature of the negative proposal $\mathcal{P}^{-}$. This approach further aggregates the proposal $\mathcal{P}_{t,m}^{target}$ in the target video with inter-video object relationships. To minimize object confusions during detection, we introduce the following loss for the target video,

$$\mathcal{L} = \mathcal{L}_{detection} + \gamma \mathcal{L}_{relation}, \tag{3.5}$$

where $\mathcal{L}_{detection} = \mathcal{L}_{regression} + \mathcal{L}_{classification}$ is the traditional detection loss (i.e., bounding box regression and object classification) applied to the final proposal features $\boldsymbol{\beta}_{t,m}^{target}$ in the target frame. The coefficient γ is a weighting factor, and $\mathcal{L}_{relation}$ is a triplet-style metric loss to regularize Eq. (3.4),

$$\mathcal{L}_{relation} = \max(d(\boldsymbol{\alpha}_{t,m}^{target}, \boldsymbol{\alpha}^{-}) - d(\boldsymbol{\alpha}_{t,m}^{target}, \boldsymbol{\alpha}^{+}) + \lambda, 0). \tag{3.6}$$

This loss imposes a discriminative constraint on the relation computed in Eq. (3.4), emphasizing that the similarity between the target proposal $\boldsymbol{\alpha}_{t,m}^{target}$ and the positive

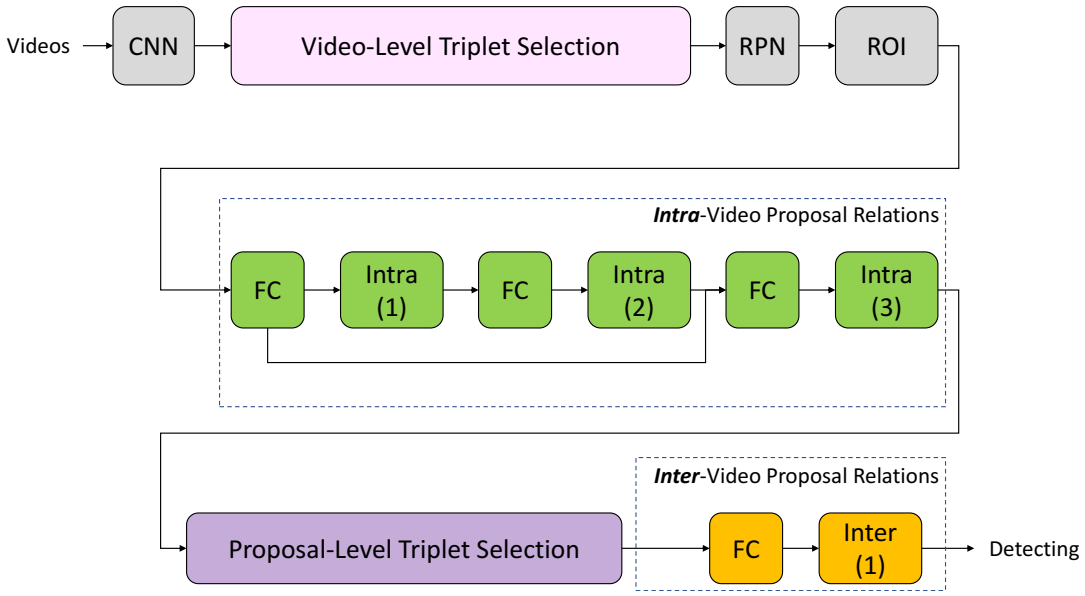


Figure 3.3 : HVR-Net Architecture. We flexibly adapt the widely-used Faster RCNN architecture as our HVR-Net. More implementation details can be found in Section 3.3.1.

proposal $\boldsymbol{\alpha}^+$ should be greater than its similarity to the negative proposal $\boldsymbol{\alpha}^-$ by a margin λ, where d is the Euclidean distance. By enforcing this condition, $\boldsymbol{\beta}_{t,m}^{target}$ becomes more discriminative, effectively reducing inter-video object confusion by enhancing relationships with similar objects and diminishing those with dissimilar ones.

3.3 Experiments

We mainly evaluate our HVR-Net on the large-scale ImageNet VID dataset [89]. It consists of 3862 training videos (1,122,397 frames) and 555 validation videos (176,126 frames), with bbox annotations across 30 object categories. Moreover, we train our model on intersection of ImageNet VID and DET dataset [35, 153], and report mean Average Precision (mAP) on validation set of VID.

3.3.1 Implementation Details

Architecture. We flexibly adapt Faster RCNN to our HVR-Net with the following details. The architecture is shown in Fig. 3.3. We use ResNet-101 [69] as backbone for ablation studies, and also report the results on ResNeXt-101-32x4d [202] for SOTA comparison. We extract the feature of each sampled frame after the *conv*4 stage, in order to select video triplet in a training batch. Region Proposal Network (RPN) is used to generate proposals from each frame of the selected video triplet, by using the feature maps after the *conv*4 stage. We introduce three intra modules in Fig. 3.3. Before each of them, we add 1024-dim fully-connected (FC) layer. Additionally, we use a skip connection between intra(1) and intra(3), to increase learning flexibility. In this case, the intra(3) module can use both initial and transformed proposals of support frames to enhance proposals in the target frame.

We introduce one inter module which is added upon a 1024-dim FC layer. Additionally, for both intra and inter modules, the kernel function is set as Embedded Gaussian in [190], where each embedding transformation in this kernel is a 1024-dim FC layer.

Training Details. We implement our HVR-Net on Pytorch, by 8 GPUs of 1080Ti. In each training batch, we randomly sample $K = 3$ object categories from training set, and randomly sample $N = 3$ videos per category. Hence, there are 9 videos in a batch. Then, we randomly select one video as target video, and use other 8 videos as the support video set. For each video, we randomly sample 3 frames, where the middle frame is used as target frame. λ is empirically set to 1.0 without generability.

3.3.2 Ablation Studies

Effectiveness of HVR-Net. We first compare our HVR-Net with the baseline architecture, i.e., Faster RCNN. As shown in Table 3.1, our HVR-Net significantly

Table 3.1 : Effectiveness of our HVR-Net.

Methods	Intra-Video	Inter-Video	mAP(%)
Baseline: Faster-RCNN	-	-	73.2
Our HVR-Net	✓	-	$80.6_{\uparrow \mathbf{7.4}}$
Our HVR-Net	✓	✓	$\mathbf{83.2}_{\uparrow \mathbf{10.0}}$

Table 3.2 : Multi-Level Triplet Selection of Our HVR-Net.

Multi-Level Triplet Selection	mAP(%)
Simple	81.0
Our	**83.2**

outperforms Faster RCNN, indicating its superiority in video object detection. More importantly, HVR-Net with both intra-video and inter-video is better than that with intra-video only (83.2 vs. 80.6). It demonstrates that, learning proposal interactions inside each single video is not sufficient to describe category differences among videos. When adding inter-video proposal relation module, our HVR-Net can flexibly select hard proposals from confused videos, and effectively build up relations among these proposals to distinguish object confusions.

Multi-Level Triplet Selection. Our HVR-Net is built upon a multi-level triplet selection scheme, including video-level and proposal-level proposal selection. To demonstrate the effectiveness, we replace these two selection modules with a simple approach, i.e., selecting random videos and using all proposals in each video. In Table 3.2, when using the straightforward selection, the performance of HVR-Net is getting worse. The main reason is that, blindly selected videos and proposals do not guide our HVR-Net to focus on object confusion in videos. Alternatively, when

Table 3.3 : Supervision in Our HVR-Net.

Detection Loss	Relation Regularization	mAP(%)
✓	-	80.0
✓	✓	**83.2**

Table 3.4 : Number of Intra and Inter Modules in Our HVR-Net.

No. of Intra	No. of Inter	mAP(%)
2	1	81.8
3	1	**83.2**
3	2	82.1

Table 3.5 : Number of Testing Frames in Our HVR-Net.

Testing Frames	5	11	17	21	31
mAP(%)	80.5	81.6	82.0	82.9	**83.2**

we add our video and proposal triplet selection, HVR-Net can effectively leverage hard proposals of confused videos to learn and correct inter-video object relations, in order to boost video object detection.

Table 3.6 : Comparison with the state-of-the-art methods on ImageNet VID (mAP). FRCNN stands for Faster-RCNN.

Methods	Backbone	Post-processing	Base detector	mAP(%)
D&T[51]	ResNet101	-	R-FCN	75.8
MANet[188]	ResNet101	-	R-FCN	78.1

Continued on next page

Table 3.6 continued from previous page

Methods	Backbone	Post-processing	Base detector	mAP(%)
LWDN[76]	ResNet101	-	R-FCN	76.3
RDN[35]	ResNet101	-	FRCNN	81.8
LongRange[153]	ResNet101	-	FPN	81.0
Deng [34]	ResNet101	-	R-FCN	79.3
PSLA [60]	ResNet101+DCN	-	R-FCN	80.0
THP [233]	ResNet101+DCN	-	R-FCN	78.6
STSN[12]	ResNet101+DCN	-	R-FCN	78.9
Ours	ResNet101	-	FRCNN	**83.2**
TCNN [78]	DeepID+Craft	Tublet Linking	RCNN	73.8
STMN [200]	ResNet101	Seq-NMS	R-FCN	80.5
FGFA[235]	Align. Inc-ResNet	Seq-NMS	R-FCN	80.1
D&T($\tau = 10$)[51]	ResNet101	Viterbi	R-FCN	78.6
D&T($\tau = 1$)[51]	ResNet101	Viterbi	R-FCN	79.8
MANet[188]	ResNet101	Seq-NMS	R-FCN	80.3
ST-Lattice[24]	ResNet101	Tublet-Rescore	R-FCN	79.6
SELSA[197]	ResNet101	Seq-NMS	FRCNN	82.5
Deng [34]	ResNet101	Seq-NMS	R-FCN	80.8
PSLA [60]	ResNet101+DCN	Seq-NMS	R-FCN	81.4
STSN+[12]	ResNet101+DCN	Seq-NMS	R-FCN	80.4
Ours	ResNet101	Seq-NMS	FRCNN	**83.8**
D&T[51]	ResNeXt101	Viterbi	FRCNN	81.6
D&T[51]	Inception-v4	Viterbi	R-FCN	82.1
LongRange[153]	ResNeXt101-32×8d	-	FPN	83.1
RDN[35]	ResNeXt101-64×4d	-	FRCNN	83.2

Continued on next page

Table 3.6 continued from previous page

Methods	Backbone	Post-processing	Base detector	mAP(%)
RDN[35]	ResNeXt101-64×4d	Seq-NMS	FRCNN	84.5
SELSA[197]	ResNeXt101-32×4d	-	FRCNN	84.3
SELSA[197]	ResNeXt101-32×4d	Seq-NMS	FRCNN	83.7
Ours	ResNeXt101-32×4d	-	FRCNN	**84.8**
Ours	ResNeXt101-32×4d	Seq-NMS	FRCNN	**85.5**

Supervision in HVR-Net. As mentioned in Section 3.2.4, we introduce a relation regularization in Eq. (3.6), in order to emphasize the correct relation constraint on inter-video relation module in Eq. (3.4). We investigate it in Table 3.3. As expected, this regularization can boost HVR-Net by a large margin, by enhancing similarity between proposals in the same category, and reducing similarity between proposals in the different categories.

Number of Intra and Inter Relation Modules. We investigate the performance of our HVR-Net, with different number of intra-video and inter-video proposal modules. When changing the number of intra modules (or inter modules), we fix the number of inter modules (or intra modules). The results are shown in Table 3.4. As expected, when increasing the number of both modules, the performance of HVR-Net is getting better and tends to become flat. Hence, in our experiment, we set the number of intra modules as three, and set the number of inter module as one.

Number of Testing Frames. We investigate the performance of HVR-Net, w.r.t., the number of sampled frames in a testing video. As expected, when increasing the number of testing frames, the performance of HVR-Net is getting better and tends to become stable. Hence, we choose the number of testing frames as 31 in our

Table 3.7 : Comparison with state-of-the-art methods in mAP.

Methods	Fast (mAP)	Medium (mAP)	Slow (mAP)
FGFA [235]	57.6	75.8	83.5
MANet [188]	56.7	76.8	86.9
Deng[34]	61.1	78.7	86.2
LongRange[153]	64.2	79.5	86.7
Ours	**66.6**	**82.3**	**88.7**

experiment. Besides, we test HVR-Net by unloading inter-video proposal relation module in the testing phase, which achieves the comparable mAP.

3.3.3 SOTA Comparison

We compare our HVR-Net with a number of recent state-of-the-art approaches on ImageNet VID validation set. As shown in Table 3.6 and Table 3.7, HVR-Net achieves the best performance among various settings and object categories.

In Table 3.6, we first make comparison without any video-level post-processing techniques. Under the same backbone, We significantly outperform the well-known approaches such as FGFA [235] and MANet [188], which uses expensive optical flow as guidance of feature aggregation. More importantly, our HVR-Net outperform the recent approaches [35, 153] that mainly leverage proposal relations among different frames for spatio-temporal context aggregation. This further confirms the effectiveness of learning inter-video proposal relation. Second, we equip HVR-Net with the widely-used post-processing approach Seq-NMS. Once again, we outperform other state-of-the-art approaches under the same backbone. It shows that, our HVR-Net is compatible and complementary with post-processing of video object detection,

which can further boost performance.

Additionally, we follow FGFA [235] to evaluate detection performance on the categories of slow, medium, and fast objects, where these three categories are divided by their average IoU scores between objects across nearby frames, i.e., Slow (score>0.9), Medium (score∈[0.7,0.9]), Fast (Others). As shown in Table 3.7, our HVR-Net boost the detection performance on all these three categories, showing the importance of inter-video proposal relation for confusion reduction.

3.3.4 Visualization

Detection Visualization. We show the detection result of HVR-Net in Fig. 3.4. Specifically, we compare two settings, i.e., baseline with only intra-video proposal relation module, and HVR-Net with both intra-video and inter-video proposal relation modules. As expected, when only using intra-video relation aggregation, baseline fails to recognize the object in the video, e.g., a female lion in Subplot (a) is mistakenly recognized as a horse with confidence larger than 0.9. The main reason is that, intra-video relation mainly focuses on what the object looks like and how it moves in this video. For the video in Subplot (a), the appearance and motion of this lion are quite similar to a horse, leading to high confusion. Alternatively, when we introducing inter-video proposal relation module, HVR-Net successfully distinguish such object confusion in videos. Hence, it is necessary and important to learn inter-video proposal relations for video object detection.

Video and Proposal Feature Visualization in HVR-Net. We visualize the proposal features of target frames in video triplets with t-SNE in Fig. 3.5. As expected, with inter-video proposal relation integrated, the proposal features of confusing objects can be clarified, while baseline, with intra-video proposal relations only, mistakenly clusters the proposals not belong to same category, e.g., in Fig. 3.5 (b), proposals of domestic cat mistakenly stay with proposals of fox together as a

Figure 3.4 : Detection Visualization. For each video, the first row shows the baseline with only intra-video proposal relation module. The second row shows HVR-Net with both intra-video and inter-video proposal relation modules. Clearly, our inter-video can effectively guide HVR-Net to tackle object confusion in videos. For example, a female lion in Subplot (a) looks quite similar to a horse, due to its color and its motion in this video. As a result, the baseline mistakenly recognizes it as a horse, when only using intra-video relation aggregation.

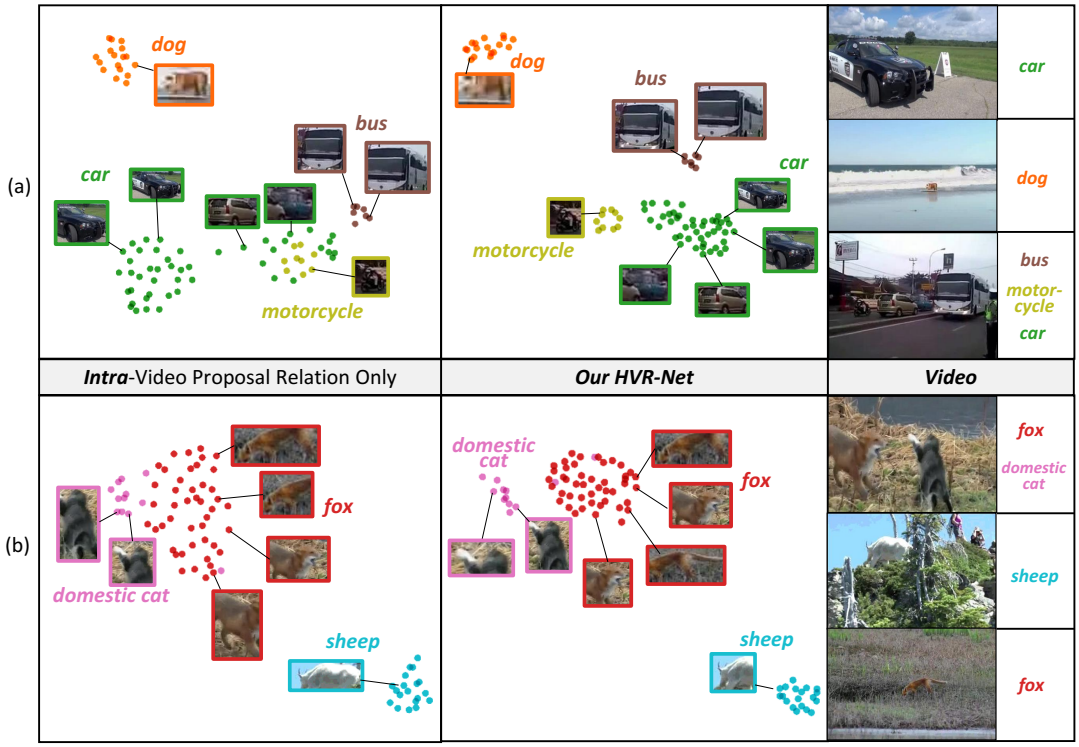


Figure 3.5 : Proposal Feature Visualization of Video triplet by t-SNE. With intra-video relation only, proposals of confusing objects mistakenly stay together as a cluster (i.e. domestic cats and foxes in (b), cars and motobikes in (a)). Our HVR-Net can learn the discriminative cues and clarify those proposals of confusing objects. For each video triplet, three target frames and their proposals are shown.

cluster, while our HVR-Net can learn a compact cluster (e.g., proposals of fox) and assign proposals of domestic cat correctly. The reason is that the object confusion is clarified with inter-video proposal relation integrated, leading to enlarged difference of confused proposals in feature embedding.

Performance Analysis on Object Categories. We show the accuracy (mAP) comparison of 10 categories with our HVR-Net and baseline with intra-video proposal relation only. Top-5 improved most categories and top-5 declined most categories are shown in Fig. 3.6. The proposed inter-video proposal relation module

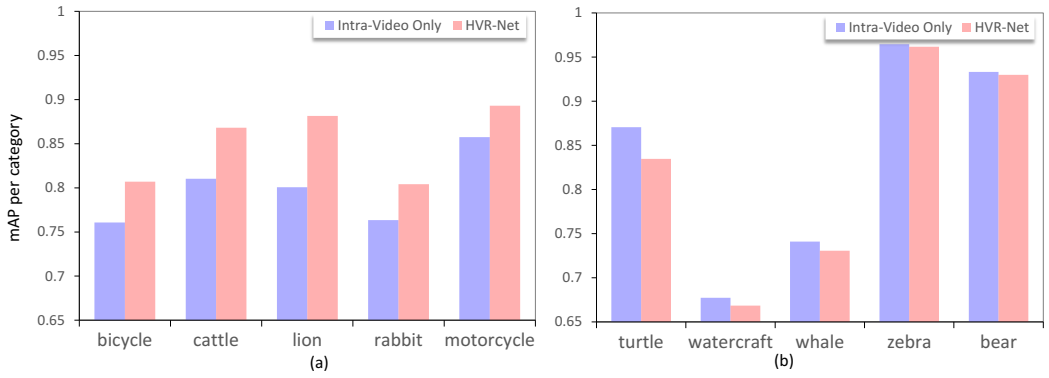


Figure 3.6 : Comparison of mAP per Category. Top-5 improved most categories and top-5 declined most categories are shown in subplot (a) and (b) separately. For each category, mAP is shown for baseline with only intra-video proposal relation module and our HVR-Net.

boosts performance a large margin in cattle, rabbit, lion and other mammal categories. The reason is that objects in those categories usually share similar motion and appearance characteristics. With the inter-video proposal relation integrated, the object confusion is clarified, as illustrated in Fig. 3.4.

3.4 Conclusion

In this chapter, we propose to learn inter-video object relations for video object detection. Based on a flexible multi-level triplet selection scheme, we develop a Hierachical Video Relation Network (HVR-Net), which can effectively leverage intra-video and inter-video relation in a unified manner, in order to progressively tackle object confusions in videos. We perform extensive experiments on the large-scale video object detection benchmark, i.e., ImageNet VID. The results show that our HVR-Net is effective and important for video object detection.

Chapter 4

Progressive Frame-Proposal Mining for Weakly Supervised Video Object Detection

This chapter progresses into the realm of weakly supervised video object detection. Here, we address the challenge of detecting objects with minimal supervision by leveraging sparse annotations more effectively. By developing a progressive mining technique that refines detection capabilities across frames, this chapter explores innovative strategies to maximize the utilization of available data, significantly reducing the reliance on extensive manual annotations while enhancing the model's performance in real-world scenarios.

4.1 Introduction

Recent years witnessed that deep learning methods have achieved great success in video object detection [12, 35, 153, 197, 235, 28, 64]. However, such remarkable performance heavily depends on large-scale video benchmarks with full object annotations [89], i.e., bounding boxes are densely annotated for all video frames which objects appear in. This is labor-intensive for real-world applications in practice. Alternatively, one can easily obtain a large number of weakly-annotated videos from internet. This fact inspires us to explore video object detection in a weakly supervised setting, i.e., *learning detector with only object tags in the video.*

In fact, weakly supervised approaches have been explored in image-based object detection, by mining informative proposals in the pipeline of multiple instance learning [9, 14, 118, 87, 164, 165, 91, 194, 211, 92]. In particular, [14] introduces

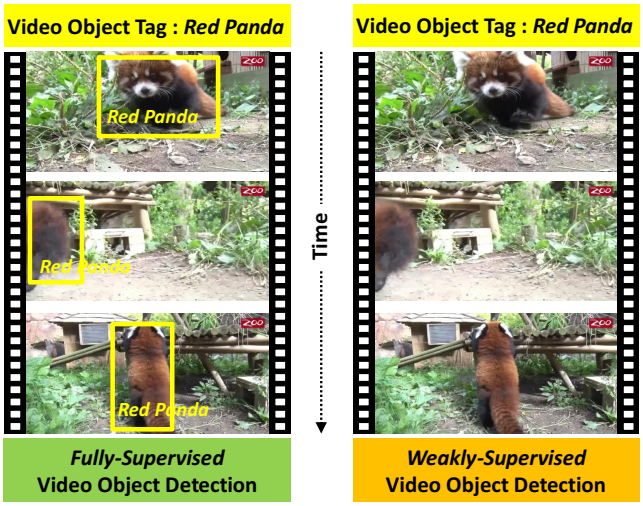


Figure 4.1 : Weakly Supervised Video Object Detection. It is often labor-intensive to annotate bounding boxes on tons of video frames in practice. Hence, we consider a novel and challenging weakly supervised video object detection problem, where each video is only tagged by object labels, without frame-level box annotations.

a popular two-stream weakly supervised deep detection network (WSDDN), which simultaneously performs region selection and classification in an end-to-end fashion. To further boost detection performance, several extensions have been introduced by effective proposal generation [166, 38, 194], instance refinement [164, 211, 86, 109], network optimization [177, 148, 9], etc. However, these approaches mainly focus on the domain of still images. Directly applying them on every single video frame would lead to unsatisfactory detection performance. First, different from still images, video frames may be blurred by object/camera motions, and some of them are redundant without any objects of interest. Using these noisy, even useless frames would increase learning difficulty of object detectors. Second, due to the lack of ground-truth bounding boxes, weak detectors are often equipped with a huge number of object proposals (e.g., from selective search). Applying all these proposals on all video frames would introduce expensive computation cost. Third, these image-

based weak detectors treat video frames as individual images. This ignores importance of different proposals among frames, which further harms the effectiveness to detect objects in videos.

Additionally, there exists some weakly-supervised approaches for high-level video understanding tasks, such as action detection [8, 155] and action-driven object detection [214], or video object grounding [149, 213]. However, these tasks often require extra careful annotations besides of video tags, due to their specific topics. For example, action tasks [8, 155, 214] need to pre-train a person detector with bounding box annotations of human, in order to discover human activities and relevant objects in the video. Object grounding tasks [149, 213] should be equipped with extra natural language descriptions. Alternatively, we target at a fundamental and novel task of video object detection. Since only object tags are given in each training video, such task brings new challenges and opportunities in object detection in both aspects of research and industry.

For these reasons mentioned above, we introduce the weakly-supervised video object detection problem, and design a Progressive Frame-Proposal Mining (PFPM) framework to tackle it. As shown in Fig. 4.2, PFPM can effectively leverage video tags as supervision, and progressively mine object proposals in a coarse-to-fine manner, i.e., from videos to frames, from frames to instances. Specifically, we first introduce a concise Multi-Level Selection (MLS) scheme. By taking advantage of both low-level and high-level visual clues, it can discover object-relevant frames from an input video, and subsequently exploit representative proposals on these frames. Via MLS, we can significantly reduce frame redundancy as well as improve proposal effectiveness. Second, we design a novel Holistic-View Refinement (HVR) scheme. By globally weighting proposals over video frames, it can correctly assign the importance score for each MLS-based proposals, and generate discriminative pseudo bounding boxes to boost video detection via self-training. Finally, we investigate

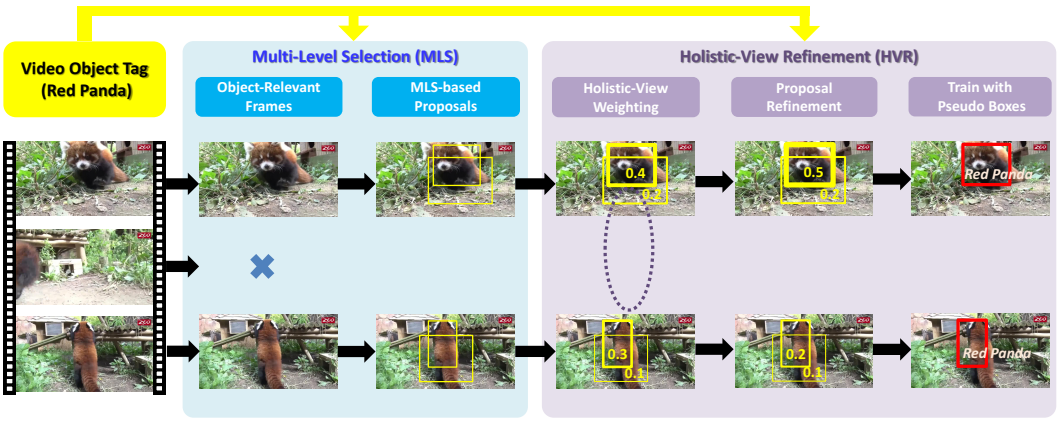


Figure 4.2 : Our Progressive Frame-Proposal Mining (PFPM) Framework. With the only supervision of object tags, our PFPM provides a novel coarse-to-fine mining pipeline to exploit discriminative proposals for object detection in videos. Specifically, it consists of two distinct mining phases, e.g., Multi-Level Selection (MLS) and Holistic-View Refinement (HVR). First, MLS can discover object-relevant frames by video object classification, and then integrate multi-level semantic clues to exploit discriminative proposals from these frames. Second, HVR can weight MLS-based proposals among video frames, and further refine them to generate pseudo object boxes for self-training. More explanation can be found in Section 4.2.

extensive experiments on the large-scale video object detection benchmark, i.e., ImageNet VID, without using bounding boxes annotations. Our PFPM shows its superiority, compared with the recent state-of-the-art weakly-supervised detectors. We summarize the contributions of this chapter as follows:

1. *New Problem Statement*: To the best of our knowledge, we are the first to propose the problem of weakly supervised video object detection, i.e., object detection by only video tags, without frame-level bounding box annotations and/or extra prior knowledge of objects.

2. *Distinct Framework*: We introduce an effective and efficient Progressive Frame-Proposal Mining (PFPM) framework to address weakly supervised video object detection. It consists of two distinct phases, i.e., Multi-Level Selection (MLS) and Holistic-View Refinement (HVR), which formulate a whole detection pipeline that can utilize video tags to discover object proposals and refine them in a novel progressive manner. By elaborating mining, our PFPM can largely boost video detection performance while alleviating computation cost with discriminative proposals.

3. *The State-of-the-Art Result*: We benchmark the state-of-art methods under the weakly supervised setup, e.g., we outperform WSDDN [14] with 21.7 mAP improvement on ImageNet VID. It shows the effectiveness and superiority of our PFPM for weakly supervised video object detection.

The rest of this chapter is organized as follows. We introduce our PFPM in Section 4.2, with detailed explanation of Multi-Level Selection (MLS) and Holistic-View Refinement (HVR). Finally, we perform extensive experiments in Section 4.3, and make conclusions in Section 4.5.

4.2 Progressive Frame-Proposal Mining

Overview. To tackle weakly supervised video object detection without bounding box annotations, we introduce a novel Progressive Frame-Proposal Mining (PFPM) framework in this section. Based only on the supervision of object tags in videos, PFPM can effectively generate pseudo bounding boxes by exploiting frames and instances in a coarse-to-fine manner. The whole framework is shown in Fig.4.2, which consists of two mining stages, i.e., Multi-Level Selection (MLS) and Holistic-View Refinement (HVR). First, we design a concise MLS scheme to reduce frame redundancy and improve proposal effectiveness, with guidance of both low-level and

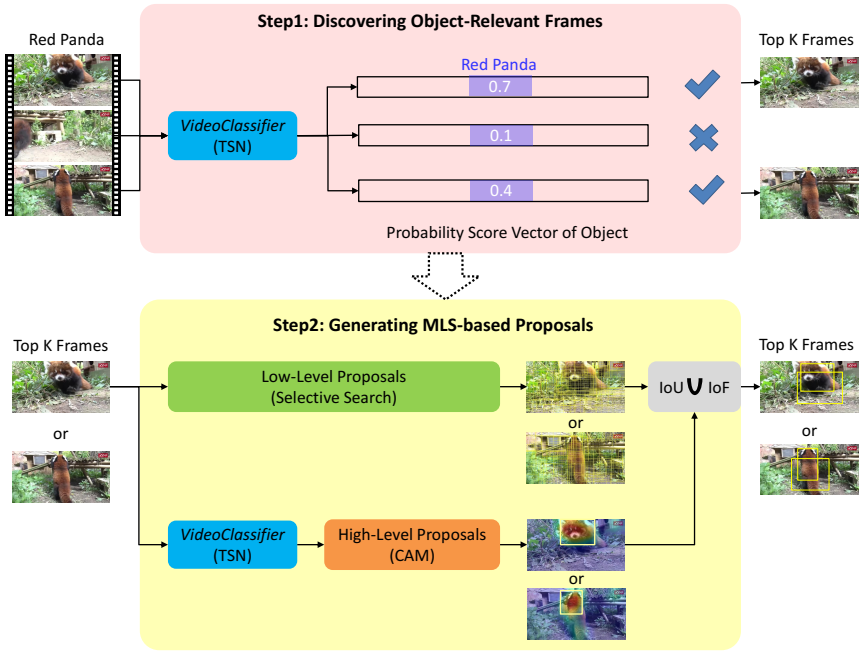


Figure 4.3 : Multi-Level Selection (MLS). Given a training video with object tags, we first train a video classifier to select top K object-relevant frames whose probability scores on object labels are high. For each selected frame, we then generate MLS-based proposals by integrating visual clues from both low-level Selective Search and high-level CAM. More details can be found in Section 4.2.1.

high-level clues. Second, we develop a robust HVR scheme to further refine MLS-based proposals by globally weighting proposal importance among video frames, and subsequently generate pseudo object boxes for robust self-training.

4.2.1 Multi-Level Selection (MLS)

As mentioned before, most previous approaches [165, 164] perform Selective Search [171] to generate object proposals of each image in an unsupervised manner. However, it is unsuitable to use such simplistic strategy in our weakly supervised video object detection problem, since each video contains a large number of frames. On the one hand, direct usage of Selective Search on every single frame will result

in unacceptable computation burden. On the other hand, many frames are useless and even noisy for detection, due to object absence, motion blur and out-of-focus disturbance.

Motivated by these observations, we design a Multi-Level Selection (MLS) scheme for proposal generation and selection, which is guided by both low-level texture information and high-level semantic information, as shown in Fig. 4.2-4.3. First, we design an object-relevant frame selection method in Section 4.2.1. according to high-level semantic information encoded in the video classifier. This would reduce redundant frames and thus alleviate computation burden and training difficulty for detection. Second, we propose to generate proposals from the selected frames in Section 4.2.1. To promote effectiveness, we propose to integrate both low-level and high-level object information to exploit discriminative proposals.

Discovering Object-Relevant Frames

As discussed before, a video may contain frames of reduced provided information either because of blur frames with fast object motion or background frames without any objects etc. We do not know which they are, since no frame-level annotations are given in the weakly-supervised setting. In this case, we first need to discover object-relevant frames from an input video. To achieve this goal, we propose to train a video classifier by using video frames as input and video object tags as supervision. With such classifier, we can find the high-score frames which often contain objects.

Specifically, we instantiate our video classifier as Temporal Segment Network (TSN) [185], due to its simplicity and practicality. Formally, given a training video $\mathcal{V}$, we divide it into T segments and randomly sample one frame from each segment to cover the entire video. TSN integrates these T sparsely-sampled frames $\{\mathcal{T}_1, \mathcal{T}_2,$

$\ldots, \mathcal{T}_T\}$ together, and make video-level prediction $\mathbf{s} \in \mathbb{R}^C$ via

$$\mathbf{s}_t = \mathcal{F}(\mathcal{T}_t), \tag{4.1}$$

$$\mathbf{s} = \mathcal{H}(\mathcal{G}(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_T)), \tag{4.2}$$

where $\mathbf{s}_t \in \mathbb{R}^C$ is the score vector of C object classes for frame $\mathcal{T}_t$. The feature encoding function $\mathcal{F}$ is Convolutional Neural Network (CNN). The segment consensus function $\mathcal{G}$ combines the outputs from multiple frames to obtain a consensus of class hypothesis among them, and then the prediction function $\mathcal{H}$ generates the score vector $\mathbf{s}$ for this training video $\mathcal{V}$. In the experiment, we use ResNet101 for $\mathcal{F}$, average pooling for $\mathcal{G}$ and softmax function for $\mathcal{H}$. Then, we compute the cross-entropy loss between video prediction $\mathbf{s}$ and video object tag $\mathbf{y} \in \mathbb{R}^C$ for training TSN.

After end-to-end learning, we use this trained classifier as an object score predictor to select object-relevant frames for each training video. Specifically, to prevent from selecting neighbor frames, we sample N frames of a video with the fixed temporal interval. For frame $\mathcal{T}_n$, we can use Eq.(4.1) to produce its object score vector $\mathbf{s}_n \in \mathbb{R}^C$ where $n = 1, ..., N$. It is worth mentioning that, the N frames for object-relevant selection are different from the T frames for training TSN, i.e., the N frames are densely sampled and fixed to cover the entire video for selection, while the T frames are sparse sampled for efficient training and they vary in each training epoch. Further details can be found in the implementation details of Section 4.3.1.

Finally, we use the score vector as guidance to select object-relevant frames from $\mathcal{T}_{1:N}$. Suppose that, the object category of this training video is c. We check all N sampled frames in the video, according to their probability scores on this category, i.e., $\mathbf{s}_1(c), ..., \mathbf{s}_N(c)$. Subsequently, we select K frames $\{\mathcal{T}_k\}$ that are highly relevant to object category c, based on top K scores on this category.

Generating MLS-Based Proposals

After discovering object-relevant frames $\{\mathcal{T}_k\}$ from each training video, we next generate proposals on these frames for training weakly-supervised detectors afterwards. Traditionally, Selective Search (SS) [171] is often applied for this task to generate proposals $\{\mathcal{P}_k^{ss}(v)\}$ for frame $\mathcal{T}_k$,

$$\{\mathcal{P}_k^{ss}(v)\} \leftarrow SS(\mathcal{T}_k), \tag{4.3}$$

where V is the number of SS-based proposals and $v = 1, ..., V$. However, SS is mainly based on low-level visual clues (such as color, texture, etc) in an unsupervised manner. This leads to two drawbacks. On the one hand, the number of proposals V is often large to preserve the high recall. But this would introduce expensive computation cost in both training and inference. On the other hand, the proposals lack guidance of object categories. Hence, many of them are irrelevant to objects with low precision. Based on these observations, we propose to further filter SS-based proposals, with guidance of high-level object semantics encoding in the video classifier (i.e., the trained TSN).

First, as shown in Fig.4.3, for each object-relevant frame $\mathcal{T}_k$, we use the CNN backbone of video classifier in Eq. 4.1 to generate its Class Action Maps (CAM) [227],

$$\mathbf{M}_k = CAM(\mathcal{F}(\mathcal{T}_k)), \tag{4.4}$$

where each of C channels in $\mathbf{M}_k \in \mathbb{R}^{H \times W \times C}$ refers to a probability map of an object category, where each pixel value of the image indicates the possibility of object occurrence.

Second, since the frame $\mathcal{T}_k$ refers to object category c, we use the c-th channel of CAM, $\mathbf{M}_k^c$, to produce semantics-relevant proposals for this frame. Specifically, we set a probability threshold θ_{cam} to transform $\mathbf{M}_k^c$ into a binary map, Apparently,

the connected regions (probability $> \theta_{cam}$) in this binary map indicate the highly-confident locations where object appears. Hence, we treat the minimum bounding boxes of the connected regions, *i.e.*, the minimum-area bounding box that covers the connected regions, as semantics-relevant proposals,

$$\{\mathcal{P}_k^{cam}(u)\} \leftarrow \mathit{Threshold}(\mathbf{M}_k^c, \theta_{cam}), \tag{4.5}$$

where U is the number of semantics-relevant proposals generated from CAM and $u = 1, ..., U$. The minimum bounding box captures the most informative areas while minimizing the inclusion of irrelevant ones.

Third, we use CAM-based proposals as high-level guidance to further select SS-based proposals. Specifically, given a CAM-based proposal $\mathcal{P}_k^{cam}(u)$ in Eq. (4.5), we compute its Intersection over Union (IoU) and Intersection over Foreground (IoF), with regards to all the SS-based proposals $\{\mathcal{P}_k^{ss}(v)\}$. IoU metric select low-level proposals overlapping with substantial portions of the high-level proposal. IoF metric is adept at identifying low-level proposals within the bounds of a high-level proposal. Foreground refers specifically to the region of interest, rather than the general semantic concept of foreground in an image. Subsequently, we preserve those SS-based proposals, which overlap with CAM-based proposals (IoU $> \theta_{iou}$) or located inside CAM-based proposals (IoF $> \theta_{iof}$),

$$\{\mathcal{P}_k^{mls}(r)\} \leftarrow \mathit{IoU}(\mathcal{P}_k^{ss}, \mathcal{P}_k^{cam}) \ \cup \ \mathit{IoF}(\mathcal{P}_k^{ss}, \mathcal{P}_k^{cam}). \tag{4.6}$$

We call them Multi-Level-Selection (MLS) proposals $\{\mathcal{P}_k^{mls}(r)\}$. From one hand, these proposals are discriminative to capture objects, since they integrate both low-level and high-level object clues from SS and CAM. From the other hand, the number of proposals is largely reduced via multi-level selection, leading to computation efficiency. Next, we apply these MLS-based proposals for self-training weakly-supervised detector.

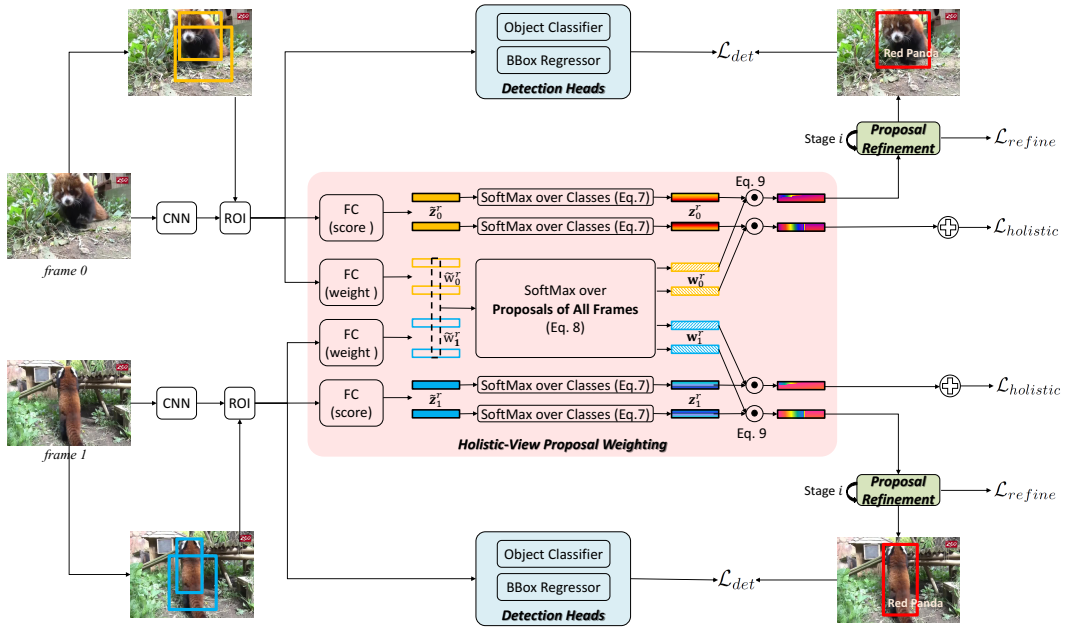


Figure 4.4 : Holistic-View Refinement (HVR). Given object-relevant frames (e.g., frame 0 and frame 1) with their MLS-based proposals (e.g., proposal r), we first weight all the proposals in a holistic video view. Then, we refine proposals for several times to generate pseudo boxes for training detection heads.

4.2.2 Holistic-View Refinement (HVR)

Given a training video, we have obtained object-relevant frames $\{\mathcal{T}_k\}$ and their MLS-based proposals $\{\mathcal{P}_k^{mls}(r)\}$ so far, in order to reduce frame redundancy and improve proposal effectiveness. Next, we apply these MLS-based proposals to produce pseudo bounding boxes on object-relevant frames, and subsequently use pseudo ground truth for training detection heads. To achieve this goal, we design a Holistic-View Refinement (HVR) scheme in Fig.4.4, based on multiple instance learning [14]. But different from these image-based works [14, 164, 165, 176], our HVR leverages object context among video frames in a holistic view.

Holistic-View Proposal Weighting

Note that, we only have video-level object tags without any instance-level supervision. Hence, we need to integrate prediction of all the proposals as video prediction for weakly-supervised learning. However, simply averaging all the score vectors is inappropriate, since the importance of different proposals varies. Traditionally, image-based approaches [14, 164, 165, 176] estimate the weight of proposals for each individual frame. Apparently, this would ignore object context among frames. To tackle such problem, we weight proposals in a holistic video view.

Object Prediction of MLS-Based Proposals. For each training video, we first feed its object-relevant frames into a CNN backbone (e.g., VGG16). For each MLS-based proposal, we can perform ROI pooling and FC layers to generate its score vector of object classification,

$$\mathbf{z}_k^r(c) = \frac{\exp[\tilde{\mathbf{z}}_k^r(c)]}{\sum_{j=1}^{C} \exp[\tilde{\mathbf{z}}_k^r(j)]}, \tag{4.7}$$

where $\tilde{\mathbf{z}}_k^r \in \mathbb{R}^C$ is the pre-softmax score vector of proposal $\mathcal{P}_k^{mls}(r)$ in frame $\mathcal{T}_k$.

Holistic Weight of MLS-Based Proposals. We add extra FC layers after ROI pooling, which generates a weight vector of each MLS-based proposal, e.g., $\tilde{\mathbf{w}}_k^r \in \mathbb{R}^C$ is the weight vector of proposal $\mathcal{P}_k^{mls}(r)$ in frame $\mathcal{T}_k$. Each entry in this vector refer to the importance of this proposal, w.r.t., an object class. As mentioned before, such weight vector is often unsatisfactory, since it dose not contain object context among frames. Hence, we further perform softmax over the weight vectors of all the R proposals in all the K frames, and comprehensively estimate the holistic weight of each proposal $\mathbf{w}_k^r \in \mathbb{R}^C$,

$$\mathbf{w}_k^r(c) = \frac{\exp[\tilde{\mathbf{w}}_k^r(c)]}{\sum_{k'=1}^{K} \sum_{r'=1}^{R} \exp[\tilde{\mathbf{w}}_{k'}^{r'}(c)]}. \tag{4.8}$$

Video Prediction. Subsequently, we average all the proposal predictions with

their weights, and produce the object prediction vector of a video $\boldsymbol{\phi} \in \mathbb{R}^C$,

$$\boldsymbol{\phi}(c) = \sum_{k=1}^{K} \sum_{r=1}^{R} \mathbf{w}_k^r(c)\mathbf{z}_k^r(c). \tag{4.9}$$

Since we have video-level object tag $\mathbf{y} \in \mathbb{R}^C$ as supervision, we apply the cross-entropy loss for training,

$$\mathcal{L}_{holistic} = CrossEntropy(\boldsymbol{\phi}, \mathbf{y}). \tag{4.10}$$

Proposal Refinement

Recent studies have shown that, weakly-supervised detection can be further enhanced by multi-stage refinement of instance classifier [164, 165]. For this reason, we choose a popular refinement module [164], and build it upon our holistic-view weighting module to generate pseudo boxes.

We use the i-th refinement stage as illustration. Specifically, we add ROI pooling and extra FC layers as instance classifier in this refinement stage, which can generate classification score vectors of all the proposals. For simplicity, we denote $\mathbf{G}_k^{(i)} \in \mathbb{R}^{(C+1)\times R}$ as score matrix of frame k at stage i, where each column of $\mathbf{G}_k^{(i)}$ refers to score vector of one MLS-based proposal in this frame. To achieve effective refinement, we follow [164] to use $\mathbf{G}_k^{(i-1)}$ at stage $i-1$ as supervision of $\mathbf{G}_k^{(i)}$, and apply cross entropy loss to train instance classifier at stage i,

$$\mathcal{L}_{refine}^{(i)} = \sum_{k=1}^{K} CrossEntropy(\mathbf{G}_k^{(i)}, \mathbf{G}_k^{(i-1)}). \tag{4.11}$$

To take holistic view of proposals into account, we set score matrix at the initial stage as the holistic-view score matrix obtained from Eq.(4.7) and (4.8),

$$\mathbf{G}_k^{(0)} = [\mathbf{w}_k^1 \odot \mathbf{z}_k^1, ..., \mathbf{w}_k^R \odot \mathbf{z}_k^R]. \tag{4.12}$$

Note that, except how to generate the initial score matrix $\mathbf{G}_k^{(0)}$ of frame k, our refinement follows the standard procedure in [164]. Hence, we suggest readers to

find more refinement details from [164] if necessary. Finally, we sum the loss of all refinement stages for training,

$$\mathcal{L}_{refine} = \sum\nolimits_i \mathcal{L}_{refine}^{(i)}. \tag{4.13}$$

Training Detection Heads with Pseudo Boxes

After refinement, the classification score of each proposal becomes reasonable. Hence, we generate the pseudo box in frame k, according to score matrix $\mathbf{G}_k^{(final)}$ at the final refinement stage. Specifically, for object category c, we pick the proposal whose score on class c is highest among all the proposals in frame k. Then, we tag this proposal as class c and treat its box as ground truth bounding box. After obtaining these pseudo boxes of each frame, we use them to train traditional detections heads,

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{reg}, \tag{4.14}$$

where $\mathcal{L}_{cls}$ is the cross entropy loss for object classifier, while $\mathcal{L}_{reg}$ is the smooth L1 loss for bounding box regressor. Finally, we combine all the losses to train our HVR scheme,

$$\mathcal{L}_{HVR} = \mathcal{L}_{holistic} + \mathcal{L}_{refine} + \mathcal{L}_{det}. \tag{4.15}$$

In the inference phase, we can simply obtain detection results from our trained heads of object classifier and bounding box regressor in the traditional and standard manner. In addition, as no labels are available during inference, we use the trained video classifier to select object-relevant frames in each test video, and obtain MLS-based proposals by Eq.(4.5) and (4.6) for detection.

4.3 Experiments

In this section, we first introduce the experimental setup including dataset, evaluation metric and implementation details. Then, we conduct extensive experiments

Table 4.1 : The Effectiveness of Our PFPM (mAP in %). We compare our PFPM with the image-based baseline. More explanation can be found in Section 4.3.2.

Methods	MLS	HVR	mAP	GFLOPs
Image-based Baseline	-	-	25.4	6293
PFPM with MLS	✓	-	$29.1_{\uparrow \mathbf{3.7}}$	5268
PFPM with MLS and HVR	✓	✓	$\mathbf{34.1}_{\uparrow \mathbf{8.7}}$	5268

Table 4.2 : The Influence of Multi-Level Selection (mAP in %). With frame selection and proposal selection, the detection results get improved. More explanation can be found in Section 4.3.2. Low-level and High-level mean SS and CAM selections respectively.

Frame Selection	Proposal Selection		mAP
	High-level	Low-level	
		✓	25.3
✓		✓	28.1
✓	✓	✓	29.1

to discuss our designs to show the effectiveness of the proposed method. Next, we compare our results with other recent works to show the superiority of our method. We also apply our MLS proposals to other approaches, demonstrating the versatility of our method. Finally, we visualize our detection results for further qualitative analysis.

4.3.1 Experiment Setup

We evaluate our PFPM framework on the large-scale benchmark for video object detection, on ImageNet VID [89]. It consists of 3862 training videos (1,122,397

Table 4.3 : The Influence of Holistic-View Refinement (HVR) (mAP in %). When taking holistic weighting into account, the detection performance can be significantly improved. More explanation can be found in Section 4.3.2.

Holistic Video	Detection heads	mAP
		29.1
	✓	30.2
✓	✓	34.1

Table 4.4 : The Influence of Training Batch Construction (mAP in %). Batch construction means the content of input data. Batch iterator means the way in which the batch data is sampled. More explanation can be found in Section 4.3.2.

Batch Construction	Batch Iterator	Results
Random one frame	Image-Based	30.2
Random one frame	Video-Based	32.9
Sampled frames	Video-Based	34.1

frames) and 555 validation videos (176,126 frames) across 30 object categories.To evaluate weakly-supervised video object detection, *we only use object tags of each training video in ImageNet VID*, without taking any bounding box annotations for training. We report mean Average Precision (mAP) at 0.5 IoU threshold on the validation set, as suggested in [35, 153, 197].

Implementation Details

Unless stated otherwise, we implement our PFPM framework as follows: For Multi-Level Selection (MLS) in Section 4.2.1, we use the training set of ImageNet VID (with video-level object tags) to train ResNet101-based Temporal Segment

Network (TSN), where the number of segment is $T = 5$, and all other details follow the official code of TSN with default settings [185]. Then, we use the trained TSN as video classifier. For each training video, we uniformly sample 15 frames, and select top $K = 2$ per category for object-relevant frames. This follows the common practice in the research community of video object detection [235, 197], in order to avoid the selection of adjacent frames. Moreover, we set thresholds as θ_{cam}=0.5, θ_{iou}=0.4, θ_{iof}=0.9 to select MLS-based proposals per frame. For Holistic-View Refinement (HVR) in Section 4.2.2, we use VGG16 as detection backbone to extract proposal features, as suggested in [164, 165]. The number of proposal refinement stages to generate pseudo boxes is 3. In all experiments, we train HVR for 70k iterations in total. The learning rate is 0.0005, and drops by a factor of 10 on iteration 40k. We implement on Pytorch by 8 GPUs of 2080Ti with one training video per GPU.

4.3.2 Ablation Studies

We first conducted experiments to discuss the influence of different components of our method in Tables 4.1-4.3, including MLS in Section 4.2.1 and HVR in Section 4.2.2. Then, we further explore the proposed method with different settings, such as training batch construction in Table 4.4, the threshold setting of θ_{cam}, θ_{iou} and θ_{iof} in Table 4.5, the number of selected frames K per category, the number of selected proposals R and the number of refinement stages I in Table 4.6, etc. Finally, we evaluate the inference efficiency of our proposed method in Table 4.7.

The Effectiveness of Our PFPM

We first compare our PFPM with image-based baseline [164]. In this baseline, we use selective search to generate proposals and perform proposal refinement individually for each frame. As shown in Table 4.1, our PFPM significantly outperforms the baseline, by using our proposed Multi-Level Selection (MLS) and Holistic-View Refinement (HVR). Utilizing MLS in our method reduces the number of proposals

Table 4.5 : The Influence of Different Thresholds (mAP in %). Our framework tends to be robust to these thresholds. In each subtable, other parameters are set to their state-of-the-art report values. More explanation can be found in Section 4.3.2.

Threshold θ_{cam} (fixed $\theta_{iou} = 0.4$, $\theta_{iof} = 0.9$)			
θ_{cam}	0.35	0.5	0.6
mAP of Our PFPM	33.7	34.1	33.8
Threshold θ_{iou} (fixed $\theta_{cam} = 0.5$, $\theta_{iof} = 0.9$)			
θ_{iou}	0.3	0.4	0.5
mAP of Our PFPM	34.0	34.1	33.8
Threshold θ_{iof} (fixed $\theta_{cam} = 0.5$, $\theta_{iou} = 0.4$)			
θ_{iof}	0.7	0.8	0.9
mAP of Our PFPM	33.9	33.9	34.1

needed during inference, resulting in lower GFLOPs consumption (6293 vs. 5268). Furthermore, implementing the HVR module improves the refinement ability of the detection head without increasing inference costs. This is because the fundamental inference process, which involves deriving detection results from trained detection heads, is consistent regardless of the HVR module's presence. It clearly shows that, our designs are effective and efficient to boost weakly-supervised video object detection, by mining frames and proposals in a progressive manner.

The Influence of Multi-Level Selection (MLS)

Given a training video, MLS mainly consists of discovering object-relevant frames (Select Frame) and selecting MSL-based proposals from these frames (Select Proposal). Hence, we investigate different settings of frame and proposal selection in Table 4.2, in order to check if both steps are necessary. As expected, our PFPM

Table 4.6 : The Influence of Selected Frames K, Proposals R, and Refinement Stages (mAP in %). In each subtable, other parameters are set to their state-of-the-art report values. More explanation can be found in Section IV-B6-IV-B8.

No. of Selected Frames K (fixed $R = 500$, $I = 3$)			
K	1	2	3
mAP of Our PFPM	33.1	34.1	34.4

Selecting K=2 for optimal efficiency and accuracy trade-off.

No. of Selected Proposals R (fixed $K = 2$, $I = 3$)			
R	200	500	1000
mAP of Our PFPM	32.9	34.1	32.7

Selecting R=500 for best performance.

No. of Refinement Stages I (fixed $K = 2$, $R = 500$)				
I	1	2	3	4
mAP of Our PFPM	32.9	33.5	34.1	33.8

Selecting I=3 for best performance.

is getting better when we select the training frames. It shows that, we should select object-relevant frames for weakly supervised detection, instead of using random frames in a blind way. Moreover, the performance can be further improved, when we apply multi-level proposal selection. It indicates that, MLS-based proposals can be more discriminative, by integrating visual clues from both high-level CAM (from TSN) and low-level Selective Search. It can be also discovered from our experiments that frame selection yielded a significantly higher performance improvement than CAM. While integrating Selective Search (SS) with frame selection substantially filtered out noisy proposals, enhancing model performance, CAM's addition provided a modest boost. This emphasizes the effectiveness of frame reduction.

Table 4.7 : Inference Efficiency of Our PFPM. More explanation can be found in Section 4.3.2.

Proposal Generation Method	No. of Proposals	GFlops of PFPM
PFPM with Selective Search (SS)	1700	6293
Multi-Level Selection (MLS)	850	$5268_{\downarrow 1025}$

Table 4.8 : Different initial low-level proposal generation algorithms applied with our PFPM framework. More explanation can be found in Section 4.3.2

Methods	mAP
Baseline with EdgeBoxes [237]	23.2
Baseline with Selective Search [171]	25.4
Our PFPM with EdgeBoxes [237]	31.5
Out PFPM with Selective Search [171]	**34.1**

The Influence of Holistic-View Refinement (HVR)

Based on our MLS scheme, we next evaluate HVR in Table 4.3. First, when we treat each frame as individual image and weight proposal importance per frame, the detection performance is 29.1 in mAP (%). Then, with adding detection head trained with pseudo ground truth, the detection performance is improved by 1.1 in mAP (%), from 29.1 to 30.2. Finally, with our holistic-view proposal weighting applied, the result is improved by a large margin of 3.9 in mAP (%), from 30.2 to 34.1. It shows that it is crucial to take object context among frames, when weighting the importance of proposals.

The Influence of Training Batch Construction

We then investigate different methods of training batch construction and batch iterator. All the settings are based on the frames selected by our MLS. (1) The first setting is to treat all the frames as individual images. In each batch, we randomly select frames to create a batch of 8 images. This is the traditional strategy in the weakly supervised setting of image-based object detection. (2) However, we target at video-based object detection. Hence, we also explore the video-based strategies. By doing so, the second setting of training batch construction is to select frames according to each video, i.e., we select 8 videos randomly. For each video, we randomly pick 1 frame to construct training batch. (3) The third setting is similar to the second one, except that we use all the MLS frames of each video. As shown in Table 4.4, the detection result of the first setting is 30.2 in mAP (%). Then, when we use the setting of video-based batch construction, the detection result is gradually improved. Especially, the third setting can leverage holistic weighting of proposals among frames. Hence, the result achieves the best with 34.1 in mAP (%). It shows the importance of exploiting object context among frames, for weakly supervised object detection in videos.

The Influence of Different Thresholds

As shown in Table 4.5, we investigate different threshold settings of θ_{cam}, θ_{iou} and θ_{iof} in Eq. (4.5)-(4.6), in order to reflect their influence on generating object proposals in MLS. First, when we change θ_{cam} from 0.35 to 0.6, the results are comparable with best result achieved at θ_{cam}=0.5. Second, when θ_{iou} is increased from 0.3 to 0.5, the result jitters by average 0.2 mAP (%) and achieve the best at θ_{iou}=0.4. Similarly, when θ_{iof} varies, the performance also varies by average 0.1 mAP (%). All these results show that, our framework tends to be robust to different θ_{cam}, θ_{iou} and θ_{iof}. Hence, we choose the best setting of $\theta_{cam} = 0.5$, $\theta_{iou} = 0.4$ and

$\theta_{iof} = 0.9$ in our experiments.

The Number of Selected Frames K

We evaluate K in Table 4.6. When $K = 1$ (i.e., only one frame per category is selected), the detection result of our PFPM is 33.1. When K increases to 3, the detection result is improved from 33.1 to 34.4. It demonstrates that, our PFPM can utilize the additional temporal context effectively. Incrementing K by 1 results in hundreds more proposals during holistic view refinement, significantly increasing temporal processing and, consequently, GPU memory usage. Therefore, we limit our experiments to a maximum of K=3 frames. To ensure a balance between computational demands and detection performance, we select K=2 as it represents the minimal frame count necessary to maintain video information that enhances detection capability.

The Number of Selected Proposals R

As shown in Table 4.6, we perform our PFPM with different numbers of selected proposals per frame in the video. When we increase R from 200 to 500, the performance is improved by 1.2 mAP (%). It shows that, it is vital to sample sufficient proposals to cover objects in the video. Then, when we increase R from 500 to 1000, the performance degrades by 1.4 mAP (%). Increasing the number of proposals per frame boosts memory demands, limiting our experiments to a maximum of 1000 proposals. We settled on R=500 for our experiments, avoiding excessive noise from irrelevant proposals, and striking an optimal balance for peak performance within our experimental scope.

The Number of Refinement Stages

As shown in Table 4.6, we perform our PFPM with different numbers of refinement stages in HVR. As expected, the performance tends to get better, when the

Table 4.9 : Comparison with The State-of-The-Art on Weakly-Supervised ImageNetVID (AP % per category). Our PFPM significantly outperforms the recent method. Additionally, we also equip the counterparts with our proposed MLS scheme, showing the generality of our method. * means our reimplementation on ImageNet VID.

Methods	airplane	antelope	bear	bicycle	bird	bus	car	cattle	dog	d-cat	elephant	fox	g-panda	hamster	horse	lion
WSDDN* [14]	6.8	5.7	28.0	0.9	14.0	21.2	19.9	9.2	6.1	6.6	22.2	22.0	22.1	8.1	3.8	0.1
OICR* [165]	4.3	10.4	41.8	2.4	13.6	35.1	33.1	19.0	18.2	20.5	45.6	24.2	37.4	3.9	5.7	0.2
PCL* [164]	18.4	12.0	38.2	10.3	21.5	37.6	32.7	31.6	23.3	18.7	**50.6**	41.3	33.4	2.1	11.0	0.5
Our MLS+WSDDN [14]	5.1	8.5	38.0	1.9	8.7	22.0	19.3	11.5	15.7	16.8	36.6	34.6	33.8	4.5	8.5	0.1
Our MLS+OICR [165]	34.9	28.5	38.3	12.2	26.8	21.1	32.5	24.8	23.6	**27.4**	47.6	47.3	38.2	4.3	11.1	**3.3**
Our MLS+PCL [164]	43.0	29.7	**42.0**	11.9	27.4	32.4	33.6	**37.5**	24.8	22.4	50.5	**53.4**	41.3	3.2	16.3	0.6
Our PFPM	**59.7**	**50.0**	41.4	**13.8**	**28.2**	**47.2**	**42.5**	36.9	**25.1**	21.8	49.8	52.9	**44.7**	**13.5**	**38.0**	0.7
Methods	lizard	monkey	motor	rabbit	r-panda	sheep	snake	squirrel	tiger	train	turtle	w-craft	whale	zebra	avg	
WSDDN* [14]	9.9	11.8	4.7	12.8	1.0	14.9	12.2	2.2	29.4	**18.3**	15.0	16.0	1.0	24.7	12.4	
OICR* [165]	**31.3**	9.9	50.1	13.7	3.2	40.2	4.0	1.9	54.7	11.8	29.8	25.0	1.6	49.8	21.4	
PCL* [164]	30.0	10.4	49.2	17.6	13.3	43.2	**12.4**	10.7	56.0	13.4	38.8	23.4	1.0	57.8	25.4	
Our MLS+WSDDN [14]	11.6	9.2	17.9	4.2	10.2	7.6	9.9	2.5	57.1	9.7	29.6	12.5	2.6	29.0	16.0	
Our MLS+OICR [165]	21.9	13.5	44.8	7.0	**14.1**	40.5	4.9	11.5	60.9	12.1	35.2	36.2	4.8	54.8	26.1	
Our MLS+PCL [164]	26.5	13.6	55.2	18.0	4.8	**44.9**	4.6	11.9	60.5	13.4	41.3	44.8	2.2	60.2	29.1	
Our PFPM	19.0	**20.3**	**58.6**	**23.8**	13.8	44.0	7.9	**14.2**	**62.9**	11.2	**56.8**	**44.8**	**16.0**	**63.4**	**34.1**	

number of refinement stages increases. When the refinement stages increases to 4, the performance degrades to 33.8, we argue that this due to the fact the generated pseudo proposal tag is noisy and accumulated with the increasement of refinement stages. Hence, we choose the best setting with three refinement stages in our experiments.

The Inference Efficiency

In Table 4.7, we evaluate the inference efficiency of our PFPM. The first line of this table, experimenting with proposals from Selective Search (SS), corresponds to the first line in Table 4.1. The second line of Table 4.7, which utilizes our MLS for proposals, aligns with both the second and third lines in Table 4.1. Since the number of MLS-based proposals is much smaller than the one of SS-based proposals (i.e., average number of proposals in validation set, MLS vs. SS: 850 vs. 1700), MLS can largely reduce computation cost in the inference phrase.

The low-level proposal generation:

As shown in Table 4.8, our proposed framework can be adapted to the other traditional low-level proposal generation algorithms, and different proposal generation algorithms would affect the final performance. Our PFPM can boost the baseline using EdgeBoxes [237] by 8.3 points in mAP. This is because our Multi-Level Selection (MLS) and Houlistic-View Refinement (HVR) methods are built upon the proposals generated by low-level vision clues. Our MLS is supposed to be able to integrate with the other traditional proposal generation algorithms and achieve a performance boost, as long as the proposal generation algorithm adopts low-level vision clues, such as texture information and intensity statistics.

4.3.3 Comparison with The State-of-The-Art

We compare our PFPM with the recent state-of-the-art weakly-supervised object detectors [14, 164, 165], where all the methods are trained by using only object tags in the training video, without using any bound box annotations. We show the detection result of each category in the validation set. First, since the SOTA approaches do not have frame and proposal selection, we apply our MLS for training these approaches. We can see in Table 4.9 that, all SOTA approaches get largely improved detection performance. The fact indicates that, it is necessary to discover object-relevant frames and select proposals for weakly-supervised video object detection. Moreover, our PFPM achieves the best performance among all these approaches, e.g., we significantly outperform WSDDN with 21.7 mAP improvement. It shows the superiority of our progressive frame and proposal mining framework.

Contrastingly, Ren et al. [142] report higher performance (36.6 mAP) on ImageNet-VID, using frame-level tags for both training and testing, which facilitates object existence identification per frame. However, our approach, while yielding lower mAP, offers greater annotation efficiency by relying solely on video tags, emphasizing its practicality in less annotated scenarios.

4.3.4 Visualization

We first show the detection results, by comparing our PFPM with the baseline method. Then, we visualize the proposal generated by our Multi-Level Selection scheme. Finally, we compare different proposal weighting mechanisms to understand our Holistic-View Refinement scheme.

Detection Visualization

We visualize detection results in Fig.4.5, where we compare our PFPM with PCL [164], a recent state-of-the-art weakly supervised detector that is also our baseline.

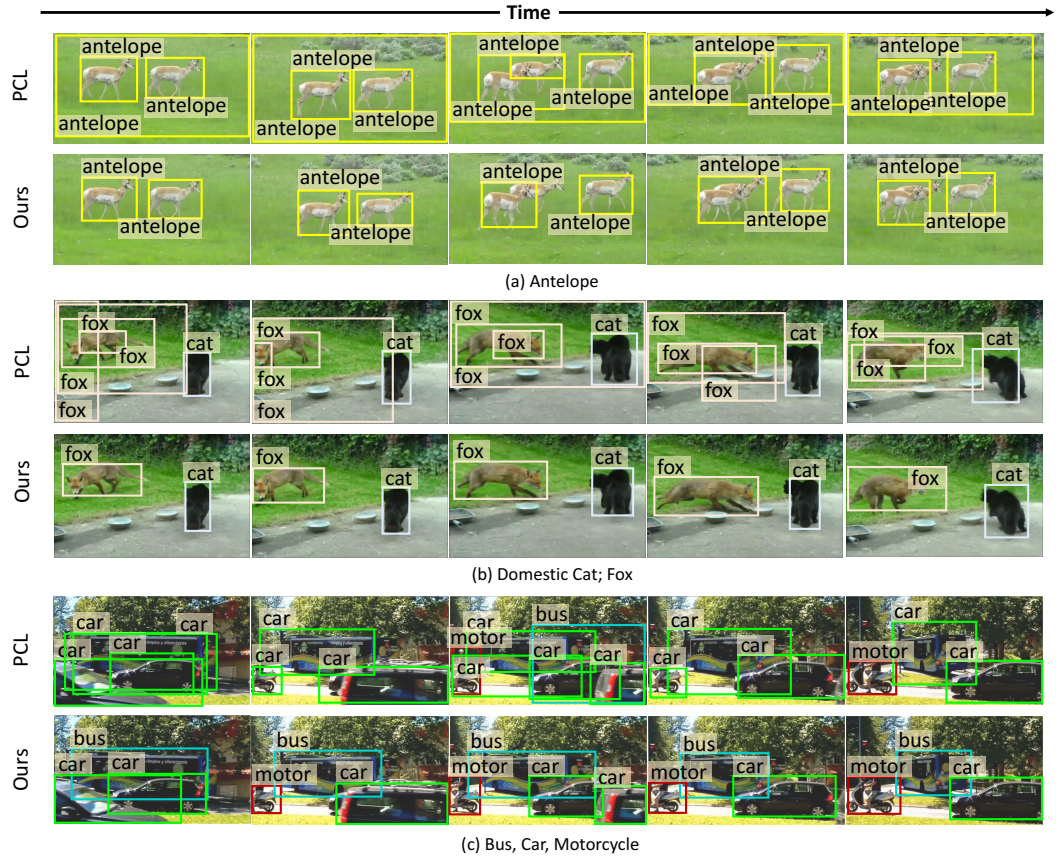


Figure 4.5 : We show detection results of our PFPM, compared with PCL [164] that is a recent SOTA weakly-supervisor detector. In (c), the first frame contains objects that are inconsistent with the video-level object tags. Our PFPM clearly achieves better detection performance with correct labels and accurate box predictions.

As expected, our PFPM achieves much better detection performance than PCL. For example, PCL mistakenly treat the background boxes as antelope in Fig. 4.5 (a) and fox in Fig. 4.5 (b), due to the lack of effective frame-proposal mining. With massive noisy frames and proposals, the baseline method is limited to learn appropriate proposal weighting, and subsequently results in false alarms of the background boxes. Moreover, PCL misclassifies the object categories of Bus, Car, Motorcycle in Subplot (c). This is mainly because that, PCL is an image-based weakly detector. Without

considering object context among frames, it is hard to learn the proposal-category mapping in videos. On the contrary, our PFPM can learn to clarify proposals which belong to different categories, by using holistic-view refinement.

MLS-based Proposal Generation

We visualize how to generate MLS-based proposals. As shown in Fig.4.6, Selective Search (SS) can generate proposals which capture the regions containing low-level texture clues. Alternatively, CAM can generate proposals which capture the regions containing high-level semantic clues. By IoU between the CAM-based and SS proposals, we can obtain the proposals that are rich in textures and overlap with the semantic regions. By IoF between the CAM-based and SS proposals, we can preserve proposals in the semantic regions. Subsequently, by merging IoU-threshold and IoF-threshold proposals, we discover the discriminative proposals around objects as shown in subplot (e) of Fig. 4.6.

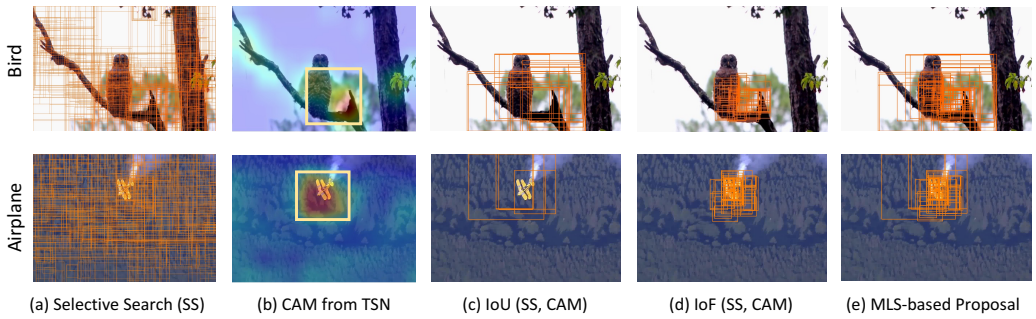


Figure 4.6 : MLS-based Proposal Generation. We use CAM as high-level guidance to select low-level proposals generated from selective search (SS). Via IoU and IoF between SS and CAM, we effectively select discriminative proposals around objects.

Holistic-View Proposal Weighting

We compare different proposal weighting mechanisms in Fig.4.7. As expected, individual-frame weighting often mistakenly assigns importance of proposals in different frames, e.g., it assigns the comparable importance on the proposals located around the front and back parts of car, while the front parts are more discriminative to recognize car. It is because this weighting ignores object context among video frames. Our holistic-view weighting can effectively tackle such problem via exploiting proposal importance among frames.

Figure 4.7 : Holistic-View Proposal Weighting. Individual-frame weighting mistakenly assigns comparable importance on the proposals located around the front and back parts of car, while the front parts are more discriminative to recognize car. Our holistic-view weighting can effectively tackle such problem via exploiting proposal importance among frames.

4.4 Discussion

In the realm of weakly supervised video object detection, our task faces challenges such as limited annotation and the need for accurate proposal generation. Our current progress has been marked by advancements in addressing these issues, yet there remains significant room for improvement. The emergence of large-scale

pretrained models like DINOv2 [129] and SAM [85], renowned for their robust detection and segmentation abilities, represents a promising trend. These models could greatly enhance proposal generation by reducing noise and refining segmentation. Concurrently, the growing popularity of large multi-modal models, such as CLIP [138] and LLaVA [114], introduces a new perspective in incorporating semantic information and reasoning capabilities. The combination of these advanced vision and multi-modal models offers a unique opportunity to tackle the challenges of our task more effectively. By utilizing these models, we envision a future direction where initial proposals are significantly reduced, enabling the inclusion of more temporal frames per batch and facilitating the integration of consistent semantic insights across frames. This approach not only aims to enhance detection accuracy but also aligns with the trend towards more annotation-efficient methodologies in weakly supervised video object detection.

4.5 Conclusion

In this chapter, we present a novel Progressive Frame-Proposal Mining (PFPM) framework for weakly supervised video object detection. It consists of Multi-Level Selection (MLS) and Holistic-View Refinement (HVR), First, MLS can discover object-relevant frames and select discriminative proposals, with guidance of both low-level and high-level visual clues. Hence, MLS can reduce frame redundancy as well as improve proposal effectiveness for training. Second, HVR weights the MLS-based proposals in a holistic video view. This can provide correct proposal importance to generate pseudo boxes by refinement. The comprehensive experiments have shown that our PFPM is effective for weakly supervised video object detection.

Chapter 5

HTML: Hybrid Temporal-scale Multimodal Learning Framework for Referring Video Object Segmentation

This chapter introduces the Hybrid Temporal-scale Multimodal Learning (HTML) Framework, designed to tackle the complexities of referring video object segmentation. By addressing the integration of temporal dynamics and multimodal inputs, HTML aims to accurately segment and track objects specified by textual descriptions across diverse temporal scales. This approach enhances the alignment between visual content and language, ensuring precise and context-aware segmentation to facilitate more intuitive and effective human-machine interactions in video analysis tasks.

5.1 Introduction

Referring Video Object Segmentation (RVOS) has witnessed the growing interest, due to its wide applications in visual editing, virtual reality, human-robotic interaction and so on. Different from the traditional vision-only VOS, RVOS aims to segment the object instance from an input video, according to an open-world description about the referred object. In this case, the model has to learn both visual and textual contents comprehensively, in order to discover the underlying object by multimodal interaction.

Recent studies [39, 108, 196, 198] have shown that, cross-modal attention is an effective way to bridge the gap between vision and language in RVOS. However, these approaches perform vision-language interactions with video frames sampled from a

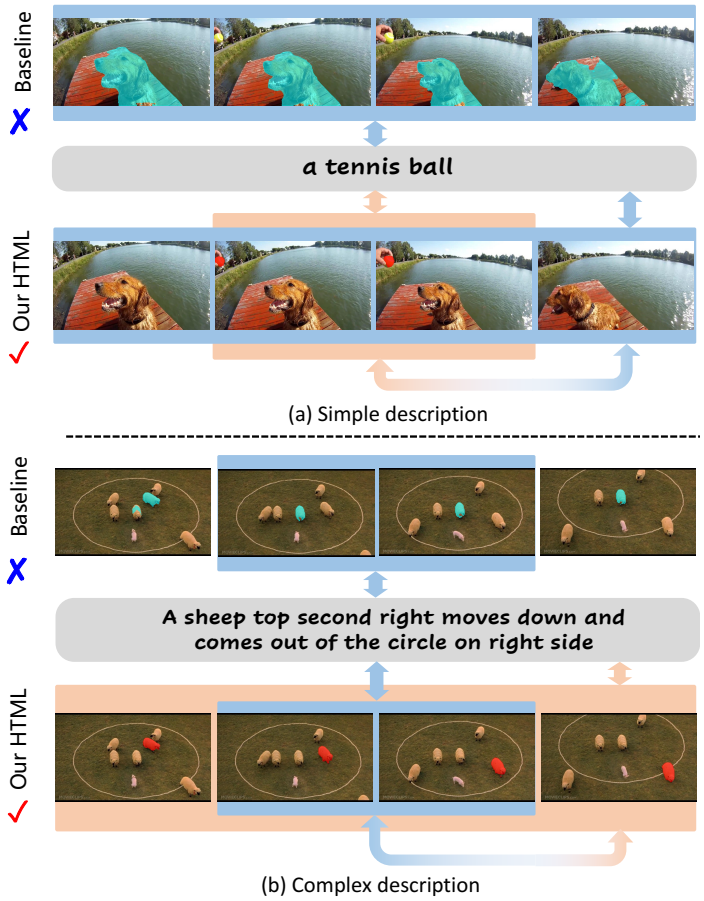


Figure 5.1 : **Referring descriptions in different lengths**. (a) The description is simple containing only the category name. (b) The description is complicated with movement and position of the object. Single-scale baseline (e.g., four frames in (a) and two frames in (b)) fails to segment the referred object, while our hybrid-scale HTML succeeds. More discussion can be found in introduction.

single temporal scale, which may limit their power to infer the referred object with accurate segmentation. The main reason is that, the open-world descriptions vary in length and contain rich semantics about the referred object, e.g., where it is, how it moves, which objects it interact with. Apparently, such diversified texts are corresponding to various temporal-scale snippets.

For example, the language query in Fig. 5.1 (a) is *a tennis ball*. Such a short

description is corresponding to the ball located at a small region in the middle two frames. If the single-scale baseline samples four frames as input, it will fail to segment the referred object. This is because it overlooks the *dog* in the center place among all these four frames, while lacking the detailed understanding in the middle two frames. Alternatively, the language query in Fig. 5.1 (b) is *a sheep top second right moves down and comes out of the circle*. Such a long description is corresponding to the particular sheep in the group, which moves across frames. If the single-scale baseline samples two frames as input, it will fail to segment the referred object. This is because it is misled by the subtle movement of sheep group in only two frames, without understanding how each sheep moves from the adjacent frames.

To tackle this difficulty, we propose a concise Hybrid Temporal-scale Multimodal Learning (HTML) Framework for RVOS, which can alleviate object confusion by language-vision interactions across different temporal scales. Specifically, we sample video frames according to different temporal scales. For each temporal scale, we introduce an intra-scale multimodal perception module, which can effectively exploit core visual semantics within the frames at this temporal scale, by mutual enhancement between textual and visual embeddings. Then, we design an inter-scale multimodal perception module, where linguistic embeddings dynamically interact with visual features across temporal scales. In this case, we can hierarchically leverage object context from all the temporal scales to boost RVOS. Finally, we evaluate our HTML on a number of benchmarks, including Ref-Youtube-VOS [147], Ref-DAVIS17 [83], A2D-Sentences and JHMDB-Sentences [53]. The extensive experiments have shown that, our HTML achieves the state-of-the-art performance on all of them.

Overall we make three contributions in this chapter:

- Concise and unified learning framework: our Hybrid Temporal-scale Multimodal

Learning (HTML) framework hierarchically constructs multimodal interactions via different strides of frame sampling, which can mutually enhance embeddings from both modalities for accurate segmentation.

- Effective multimodal perception module: our Cross-scale Multimodal Perception (CMP) module can effectively reduce complex object confusions with *intra-scale* and *inter-scale* multimodal perceptions, where linguistic and visual features interact across temporal scales.
- State-of-the-art performance on the widely-used benchmarks, which shows the superiority of our framework. Specifically, on Ref-Youtube-VOS [147], our method with ResNet-50 achieves **57.8** in $\mathcal{L}\&\mathcal{F}$, **outperforming the recent SOTA method** [198] **with ResNet-101.**

5.2 Method

To effectively align diversified descriptions and complex videos, we propose a distinct Hybrid Temporal-scale Multimodal Learning (HTML) framework for RVOS. In this section, we introduce our HTML in detail. First, we deliver an overview of HTML framework. Then, we explain how to build the hybrid temporal-scale multimodal learning paths, in the aid of vision-conditioned linguistic decoder and language-conditioned visual decoder. Next, we introduce a Cross-scale Multimodal Perception (CMP) module to align multimodal features across temporal scales. Finally, we describe the training objectives to optimize our HTML.

5.2.1 Framework Overview

As shown in Fig. 5.2, our HTML framework consists of three main parts. First, we need to extract visual and linguistic features from backbones. We adopt a visual backbone to extract frame features from T frames sampled from the given video. It can be either 2D CNN networks or 3D transformer networks. We then

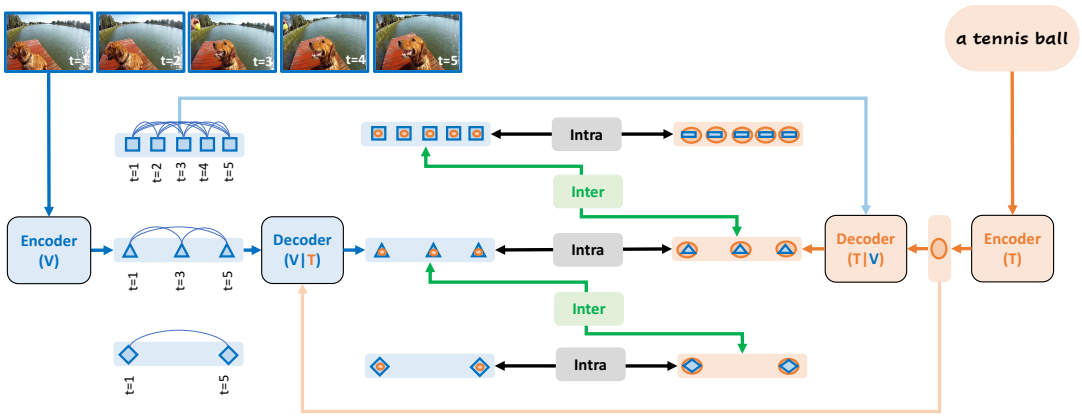


Figure 5.2 : Our Hybrid Temporal-scale Multimodal Learning framework. It aligns linguistic and visual features by learning hierarchical multimodal interactions with hybrid temporal scales, detailed in Sec. 5.2.2. Moreover, a Cross-scale Multimodal Perception (CMP) module is designed to enable interaction and cooperation among temporal scales, detailed in Sec. 5.2.3.

feed the extracted vision features into a deformable transformer encoder [234] to construct spatiotemporal relations between different frames. Meanwhile, to make a fair comparison with previous works in RVOS [198], we utilize the pretrained linguistic embedding model, RoBERTa [117], to extract textual features $\mathbf{s}_e \in \mathbb{R}^{L \times C}$ from language descriptions with L words. More details can be found in Sec. 5.3.2.

After extracting visual and linguistic features, we next construct multimodal interactions between the language descriptions and the videos. Different from the previous approaches [198, 196], we build L multimodal learning paths, where the linguistic embedding hierarchically interacts with visual features in different temporal scales. Then, we incorporate the mutually enhanced visual and linguistic features by a novel Cross-scale Multimodal Perception (CMP) module to align multimodal features across different scales. Finally, we design the training losses.

5.2.2 Hybrid Temporal-scale Multimodal Learning

To capture core object semantics, we propose a novel hybrid temporal-scale multimodal learning framework to learn multimodal relations. To start with, we build hybrid temporal scales via different sampling strides. Then, we construct basic multimodal relation learning units. Finally, we explain how to construct hybrid learning paths.

Hybrid Temporal Scale Construction

Since the texts of various objects may refer to different video parts, single temporal scale often fails to describe the diversified textual contents. To simulate such diversity and flexibility, we build hybrid temporal scales by periodicly sampling frames with different strides.

We first regard all the input frames as the first temporal scale, and then build other $L-1$ temporal scales upon it in a sequential manner. In order to ensure the diversity of sampled temporal scale, we randomly pick one frame from every h frames of last scale, where h denotes the predefined stride. Subsequently, we feed the sampled frames into visual encoder Encoder(V) to extract feature maps for each of the scales respectively. Specifically, for temporal scale l, we can obtain $\mathbf{M} \in \mathbb{R}^{T\times H\times W\times C}$, where T denotes the number of frames in the temporal scale.

Multimodal Relation Learning

In order to discover the core object semantics, we construct multimodal relations via vision-conditioned linguistic decoder and language-conditioned visual decoder to align semantics between different modalities.

Vision-Conditioned Linguistic Decoder. In order to align linguistic object semantics to the vision contents, we design a vision-conditioned linguistic decoder Decoder(L|V). Specifically, we have visual features $\mathbf{M} \in \mathbb{R}^{T\times H\times W\times C}$, and lin-

guistic embeddings $\mathbf{s}_e \in \mathbb{R}^{1\times C}$. The vision-conditioned multimodal relations $\mathbf{e} = \text{Decoder}(\mathbf{s}_e|\mathbf{M})$ are constructed by

$$\mathbf{e}_0 = \text{DeformAttn}(\mathbf{s}_e + \mathbf{q}, \mathbf{M}), \tag{5.1}$$

$$\mathbf{e}_k = \text{DeformAttn}(\mathbf{e}_{k-1}, \mathbf{M}), \tag{5.2}$$

where $k \in \{1, ..., K-1\}$. We first add $\mathbf{s}_e$ with learnable queries $\mathbf{q} \in \mathbb{R}^{N\times C}$ to represent candidate instances in the video. Then, we use deformable attention module [234] to reason the vision-conditioned multimodal relations where vision features serve as key and value to decompose linguistic features, as in Eq. (5.1). Finally, we stack the cross-attention module for K times, as in Eq. (5.2).

Language-Conditioned Visual Decoder. In order to align visual object semantics to linguistic contents, we design a language-conditioned visual decoder Decoder(V|L) (similar to Eqs. (5.1) and (5.2)), to enhance visual representation with the attendance of the language description. Differently, vision features are enhanced by multi-head self-attention (MHSA) modules at the first place, and then linguistic features $\mathbf{s}_e$ serve as key and value in cross-attention modules. In this case, it can reason the language-conditioned multimodal relations. Finally, we can get enhanced visual features by language-conditioned multimodal relations, as $\mathbf{F} = \text{Decoder}(\mathbf{M}|\mathbf{s}_e)$, where $\mathbf{F} \in \mathbb{R}^{T\times H\times W\times C}$.

Hierarchical Multimodal Learning As single-scale multimodal learning is insufficient to understand the relations between videos and texts, we propose to construct the multimodal relations hierarchically for the L hybrid temporal scales with the assistance of Decoder(L|V) and Decoder(V|L). We can obtain linguistic-attended visual features and visual-attended linguistic embeddings for different temporal scales, capturing core object semantics conditioned on different visual and linguistic contexts.

5.2.3 Cross-scale Multimodal Perception

The multimodal relations are constructed conditioned on different modalities with hybrid temporal scales. However, the modeling process of different scales is independent. To promote the cooperation and align the visual and linguistic semantics both within the scale and across the scales, we design Cross-scale Multimodal Perception (CMP) module.

Intra-scale Perception. Despite sharing visual and linguistic feature extraction, the multimodal relation construction via Decoder(V|L) and Decoder(L|V) are independent to each other. To promote the cooperation between modalities, we propose an intra-scale perception module.

Specifically, in each temporal scale l, we have visual attended linguistic embeddings $\mathbf{e} \in \mathbb{R}^{N\times T\times D}$ and linguistic attended visual features $\mathbf{F} \in \mathbb{R}^{T\times H\times W\times C}$. To achieve fine-grained semantic alignment, we measure the similarity by dot product between $\mathbf{e}$ and $\mathbf{F}$ on pixel level. Specifically, we obtain the similarity map via multimodal perception module, denoted as $\mathcal{I} = \mathrm{MP}(\mathbf{F}, \mathbf{e})$ by

$$\Omega = \mathrm{MaskHead}(\mathbf{e}), \tag{5.3}$$

$$\mathcal{I} = \Omega \cdot \mathbf{F}, \tag{5.4}$$

where MaskHead denotes three consecutive MLP layers for embedding conversion. Each value of $\mathcal{I}$ represents the relevance between visual-attended linguistic embeddings and linguistic-attended visual features, which can be interpreted as the existence of the referred object. As such, $\mathcal{I}$ is regarded as the object mask prediction with the context of current temporal scale. To this end, we achieve the multimodal perception in the same temporal scale.

Inter-scale Perception. The multimodal relations are constructed in different temporal scales. However, the process in each scale is independent and biased to-

wards the contained object semantics. To alleviate it, we propose to align multi-modal features from different temporal scales with a concise inter-scale perception module.

Specifically, suppose that the referred object appears in frame t in temporal scales l and $l+1$, the referred object can be segmented by measuring similarity between $(\mathbf{F}^l(t), \mathbf{e}^l(t))$ and $(\mathbf{F}^{l+1}(t), \mathbf{e}^{l+1}(t))$ simultaneously. Conditioned on same frame t, the visual-attended linguistic embedding $\mathbf{e}^l(t)$ from scale l is supposed to be relevant to the linguistic-attended visual features $\mathbf{F}^{l+1}(t)$ from scale $l+1$. Thus, similar to Eq. (5.4), the similarity cross different temporal scales can be measured by

$$\mathcal{I}^{l\to l+1}(t) = \mathrm{MP}(\mathbf{e}^l(t), \mathbf{F}^{l+1}(t)). \tag{5.5}$$

Without losing generality, the inter-scale similarity can also be measured by $\mathcal{I}^{l+1\to l}(t)$. More specifically, frame t can be any frame shared by the adjacent temporal scales, which is ensured by our hybrid temporal scales sampling strategy. Each value in $\mathcal{I}^{l\to l+1}(t)$ represents the referred object prediction with the prior of linguistic object semantic from temporal scale l. In the same way, values in $\mathcal{I}^{l+1\to l}(t)$ represent the opposite. To this end, we achieve multimodal perception across different temporal scales.

5.2.4 Training objectives

Our network can be trained in an end-to-end manner to locate and segment the target instance simultaneously. Specifically, the losses for intra-scale and inter-scale multimodal perception in temporal scale l are formed as

$$\mathcal{L}^l_{intra} = \sum \mathcal{L}_{cls}(\mathbf{y}, \hat{\mathbf{y}}) + \mathcal{L}_{box}(\mathbf{b}, \hat{\mathbf{b}}) + \mathcal{L}_{mask}(\mathcal{I}, \hat{\mathcal{I}}), \tag{5.6}$$

$$\mathcal{L}^l_{inter} = \sum \mathcal{L}_{mask}(\mathcal{I}^{l+1\to l}, \hat{\mathcal{I}}) \tag{5.7}$$

where time and instance subscripts are omitted for simplicity, $\mathbf{y}$ and $\mathbf{b}$ denote binary classification for instance existence and bounding box prediction respectively. Here

$\mathcal{L}_{cls}$ is the focal loss [112], $\mathcal{L}_{box}$ is the sum of L1 loss and GIoU loss [143], and $\mathcal{L}_{mask}$ is the combinition of DICE loss [126] and binary mask focal loss. We optimize the network by first finding the best prediction as the positive sample, via minimizing the matching cost $\mathcal{L}^{l}_{intra}$ and $\mathcal{L}^{l}_{inter}$ in each temporal scale l respectively. Then, we average the matching losses from different temporal scales and perception modules, and minimize it for positive samples.

5.3 Experiments

5.3.1 Datasets and Metrics

Datasets. We conduct experiments on four datasets: Ref-Youtube-VOS[147], Ref-DAVIS17[83], A2D-Sentences and JHMDB-Sentences[53], following the common practice[198].

Metrics. We follow the standard evaluation protocol [147, 196, 198] to adopt region similarity $\mathcal{L}(\%)$, contour accuracy $\mathcal{F}(\%)$ and mean $\mathcal{L}\&\mathcal{F}$ for Ref-Youtube-VOS and Ref-DAVIS17. For JHMDB-Sentences, we adopt mAP to evaluate the model. For A2D-Sentences, we use Precision@K, Overall IoU, Mean IoU and mAP for evaluation.

5.3.2 Implemented Details

We set the number of attention layers K to 4 and the hidden dimension C to 256. The number of learnable queries N is set to 5. The number of MaskHead output channels D is set to 8. During training, we first sample video clips by sliding windows and then generate $L = 3$ hybrid temporal scales with stride $h = 2$ for generalized multimodal representations and relations. We use same training recipes as in [198, 17]. All frames are downsampled by shorter side to 360 and limit the maximumsize for the long side to 640. Our model is pretrained on image referring segmentation datasets [217, 217].

Table 5.1 : Comparison with the SOTA methods on Ref-YTB-VOS.

Method	Backbone	Ref-Youtube-VOS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
CMSA [215]	ResNet-50	34.9	33.3	36.5
CMSA + RNN [215]	ResNet-50	36.4	34.8	38.1
URVOS [147]	ResNet-50	47.2	45.3	49.2
LBDT-4 [147]	ResNet-50	47.2	45.3	49.2
MLRL [196]	ResNet-50	49.7	48.4	51.0
ReferFormer [198]	ResNet-50	55.6	54.8	56.5
Ours	ResNet-50	**57.8**	**56.5**	**59.0**
PMINet [41]	ResNeSt-101	48.2	46.7	49.6
PMINet + CFBI [41]	ResNeSt-101	53.0	51.5	54.5
CITD [108]	ResNet-101	56.4	54.8	58.1
ReferFormer [198]	ResNet-101	57.3	56.1	58.4
Ours	ResNet-101	**58.5**	**57.3**	**59.8**
PMINet + CFBI [41]	Ensemble	54.2	53.0	55.5
CITD [108]	Ensemble	61.4	60.0	62.7
ReferFormer [198]	Swin-L	62.4	60.8	64.0
Ours	Swin-L	**63.4**	**61.5**	**65.3**
MTTR [17]	Video-Swin-T	55.3	54.0	56.6
ReferFormer [198]	Video-Swin-T	59.4	58.0	60.9
Ours	Video-Swin-T	**61.2**	**59.5**	**63.0**
ReferFormer [198]	Video-Swin-S	60.1	58.6	61.6
Ours	Video-Swin-S	**61.4**	**59.9**	**62.9**
ReferFormer [198]	Video-Swin-B	62.9	61.3	64.6
Ours	Video-Swin-B	**63.4**	**61.5**	**65.2**

During inference, we report results with all input frames in single temporal scale for fair comparisons. On Ref-DAVIS17, we directly inference on models trained on Ref-Youtube-VOS. Similarly, we reports JHMDB-Senteces results directly on models

Table 5.2 : Comparison with the SOTA methods on Ref-DAVIS17.

Method	Backbone	Ref-DAVIS17		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
CMSA [215]	ResNet-50	34.7	32.2	37.2
CMSA + RNN [215]	ResNet-50	40.2	36.9	43.5
URVOS [147]	ResNet-50	51.5	47.3	56.0
LBDT-4 [40]	ResNet-50	54.5	-	-
MLRL [196]	ResNet-50	58.0	53.9	62.0
ReferFormer [198]	ResNet-50	58.5	55.8	61.3
Ours	ResNet-50	**59.5**	**56.6**	**62.4**
ReferFormer [198]	Swin-L	60.5	57.6	63.4
Ours	Swin-L	**61.6**	**58.9**	**64.4**
ReferFormer [198]	Video-Swin-B	61.1	58.1	64.1
Ours	Video-Swin-B	**62.1**	**59.2**	**65.1**

trained with A2D-Sentences.

5.3.3 SOTA Comparisons

We compare our method with the state-of-the-art methods on Ref-Youtube-VOS, Ref-DAVIS17, A2D-Sentences and JHMDB-Sentences. On Ref-Youtube-VOS, our approach achieves 58.5 in $\mathcal{L}\&\mathcal{F}(\%)$ with ResNet-50, as shown in Tab. 5.1, which surpasses the recent SOTA method ReferFormer[198] with same backbone by 2.2 points. Moreover, it surpasses all the other SOTA methods with larger ResNet-101 on all evaluation metrics, which fully suggests the superiority of our method. When equipped with larger backbone, our method still show considerable superiority with accuracy gap of 1.2 points for ResNet-101 and 1.0 points for Swin-L. We also experiment our method with the well-known Video Swin Transformers [121]. Our method with Video-Swin-Tiny backbone surpasses the SOTA method with the same

Table 5.3 : Comparison with the state-of-the-art methods on A2D-Sentences.

Method	Backbone	Precision					IoU		mAP
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	Overall	Mean	
Hu et al. [72]	VGG-16	34.8	23.6	13.3	3.3	0.1	47.4	35.0	13.2
Gavrilyuk et al. [53]	I3D	47.5	34.7	21.1	8.0	0.2	53.6	42.1	19.8
CMSA + CFSA [216]	ResNet-101	48.7	43.1	35.8	23.1	5.2	61.8	43.2	-
ACAN [178]	I3D	55.7	45.9	31.9	16.0	2.0	60.1	49.0	27.4
CMPC-V [115]	I3D	65.5	59.2	50.6	34.2	9.8	65.3	57.3	40.4
ClawCraneNet [107]	ResNet-50/101	70.4	67.7	61.7	48.9	17.1	63.1	59.9	-
MTTR ($\omega = 8$) [17]	Video-Swin-T	72.1	68.4	60.7	45.6	16.4	70.2	61.8	44.7
MTTR ($\omega = 10$) [17]	Video-Swin-T	75.4	71.2	63.8	48.5	16.9	72.0	64.0	46.1
ReferFormer [198]	Video-Swin-T	82.8	79.2	72.3	55.3	19.3	77.6	69.6	52.8
ReferFormer [198]	Video-Swin-B	83.1	80.4	74.1	57.9	21.2	78.6	70.3	55.0
Ours	Video-Swin-T	82.2	79.2	72.3	55.3	20.1	77.6	69.2	53.4
Ours	Video-Swin-B	**84.0**	**81.5**	**75.8**	**59.2**	**22.8**	**79.5**	**71.2**	**56.7**

Table 5.4 : SOTA results comparison on JHMDB-Sentences.

Method	Backbone	mAP
Hu et al. [72]	VGG-16	17.8
Gavrilyuk et al. [53]	I3D	23.3
ACAN [178]	I3D	28.9
CMPC-V [115]	I3D	34.2
MTTR ($\omega = 8$) [17]	Video-Swin-T	36.6
MTTR ($\omega = 10$) [17]	Video-Swin-T	39.2
ReferFormer† ($\omega = 6$) [198]	Video-Swin-T	39.1
ReferFormer [198]	Video-Swin-T	42.2
ReferFormer [198]	Video-Swin-B	43.7
Ours	Video-Swin-T	42.7
Ours	Video-Swin-B	**44.2**

backbone by 1.8 points. With larger Video-Swin Transformers (Small and Base models), our method still achieves SOTA performance, which shows the generality of our method.

As shown in Tab. 5.2, our method surpasses the SOTA methods on Ref-DAVIS17 by over 1.0 points on both ResNet-50, Swin-L and Video-Swin-Base backbones, with new a SOTA record 62.1 in $\mathcal{J}\&\mathcal{F}(\%)$. We also experiment our method on A2D-Sentences and JHMDB-Sentences datasets and compare with other SOTA results as shown in Tab. 5.3 and Tab. 5.4. Our method achieves SOTA performances with new records on both the two datasets. On A2D-Sentences, our HTML surpass SOTA result by 1.7 points in mAP and higher recall by 1.6 on Precision@0.9. On JHMDB-Sentences, our method still achieves a new SOTA record with 44.2 in mAP. These results demonstrate the superiority of our method.

Table 5.5 : Ablation study on the components of our HTML.

	Components	#Frames	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
i.	Baseline	5	55.6	54.8	56.5
ii.	HTML w/o CMP	5	56.0	54.7	57.4
iii.	HTML	5	56.3	55.0	57.5
iv.	Baseline	8	56.2	55.0	57.3
v.	HTML w/o CMP	8	57.1	56.0	58.2
vi.	HTML	8	**57.8**	**56.5**	**59.0**

5.3.4 Ablation Study

In this section, we ablate core components of our HTML with Ref-Youtube-VOS based on ResNet-50.

Effectiveness of our HTML. To validate the effectiveness of our Hybrid Temporal-scale Multimodal Learning framework, we investigate each of our components by gradually adding them to the baseline [198]. First, comparing (i)&(iii) in Tab. 5.5, our HTML improves the performance of baseline to 56.3 when 5 frames are used for training, which is better than the baseline trained with longer input frames. Further, when longer temporal input is available, comparing (iv)&(vi), our HTML improves the model by 1.6 to 57.8 in $\mathcal{L}\&\mathcal{F}$. These prove the effectiveness of our method with frames in different lengths.

Second, comparing counterpart networks using different number of frames, *i.e.*(iii) and (vi), our method can benefit from longer temporal input (8 frames vs. 5 frames) with an improvement of 1.5 points. This conclusion holds among the other counterpart network pair, *i.e.*(ii)&(v). Further, each of our components brings an improvement of 0.3 when trained with total 5 frames, while the improvement can be doubled to 0.7-0.9 with total 8 frames. This indicates that our method can better

Table 5.6 : Ablation study on Hybrid Temporal Scales in (a) and (b), and Cross-scale Multimodal Perception in (c).

(a) Effect of No. of temporal scales.

#Scales	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	56.2	55.0	57.3
2	56.9	55.8	58.1
3	**57.8**	**56.5**	**59.0**

(b) Effect of No. of input frames.

Frames	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
5	56.3	55.0	57.5
8	**57.8**	**56.5**	**59.0**
12	57.5	56.4	58.6

(c) Effect of direction of CMP.

Inter-scale	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
None	56.2	55.0	57.3
$l \rightarrow l+1$	56.9	55.8	58.0
$l+1 \rightarrow l$	**57.8**	**56.5**	**59.0**

utilize the long temporal input.

Finally, taking 8 frames for instance, hierarchical multimodal learning and cross-scale multimodal perception improves $\mathcal{L}\&\mathcal{F}$ by 0.9 and 0.7 respectively. It proves the effectiveness of each component of our HTML.

Length of language descriptions. We explore the performance of our method with different sampled sets of language descriptions, to validate the ability of capturing diversified linguistic object semantics. Since the train set of Ref-Youtube-VOS contains an average of two language descriptions for each object, we sample the shorter ones to form the Short set and the others to form the Long set.

As shown in bottom line of Fig. 5.3, the Long set has longer sentences than the Short set in the average number of words. As shown in Fig. 5.3, when the objects are described by Short query set, more input frames in single temporal scale achieve

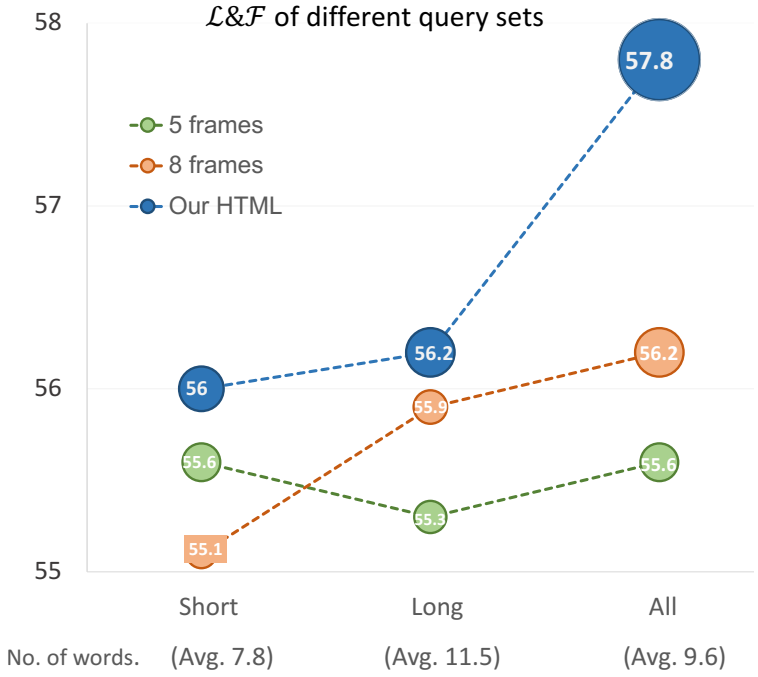


Figure 5.3 : Performance comparison of different query sets with different temporal scales on Ref-Youtbe-VOS.

inferior performance (5 frames vs. 8 frames: 55.6 vs. 55.1). It's interesting to note that more visual content (8 frames) fails to improve the performance when the linguistic content is insufficient (shorter queries). When the complexity of queries increases, *i.e.*with the Long query set, the quantitative relation is reversed: 8 frames guided model obtains better performance than the model with 5 frames, increasing by 0.8 (8 frames-Short query vs. 8 frames-Long query: 55.1 vs. 55.9). We infer that this is caused by the mismatch of visual and linguistic object semantics. On one side, shorter quires, *i.e.*relatively simple linguistic semantics, are insufficient to interpret longer videos. On the other side, the longer language descriptions contain more content irrelevant to the visual input.

Differently, our models trained with either the Short set or the Long set all surpass the single temporal scale guided model. Impressively, when the objects

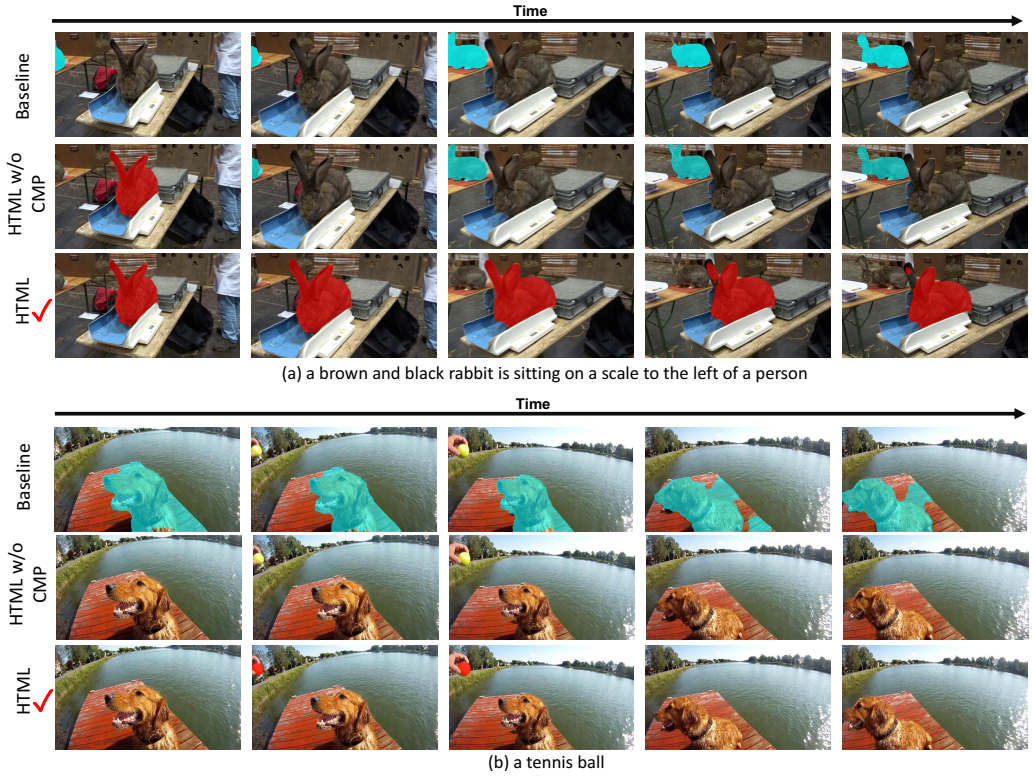


Figure 5.4 : Visualization results of complex and simple language descriptions on Ref-Youtube-VOS. Red masks indicate positive segmentation results and blue masks indicate the negatives. Our HTML is able to clarify such object confusion.

are described by queries flexible in lengths, *i.e.*with the All set, our method gets a performance boost by 1.6, while single temporal scale baseline (both 8 frames and 5 frames guided networks) improves by 0.3. This shows that our method has the strong ability to solve the mismatch issue between visual content and diversified linguistic contents.

Number of hybrid temporal scales. We investigate the effectiveness of our hierarchical multimodal learning, by exploring the number of hybrid temporal scales. For fair comparison among different settings, the number of total input frames is set to 8 in this subsection. As shown in Tab. 5.6 (a), when two temporal scales are constructed, our method brings an improvement of 0.6 in $\mathcal{L}\&\mathcal{F}$. When the number of

temporal scales is increased to three, continuous improvement of 0.9 (56.9 vs. 57.8) is observed. It proves that our model benefits from the increasing visual diversity constructed by hybrid temporal scales.

Number of input frames. We explore the effect of total input frames here. In this subsection, hybrid temporal scales are constructed by default following Sec. 5.2.2. As shown in Tab. 5.6 (b), more input frames bring an improvement by 1.5 in $\mathcal{L}\&\mathcal{F}$ (5 frames vs. 8frames: 56.3 vs. 57.8). When the input frames increase continuously to 12 frames, the performance saturates and drops slightly to 57.5. We conjecture that it is caused by insufficient video-language pairs (8 frames vs. 12 frames: 49k vs. 32k) compared to largely increased computation complexity (8 frames vs. 12 frames: 21.5 GFLOPs vs. 33.6 GFLOPs for transformers).

Direction of inter-scale perception. We explore the effect of direction of inter-scale multimodal perception in CMP. Comparing first two lines of Tab. 5.6 (c), $l \rightarrow l+1$ perception improves the baseline by 0.7 points; As in first and last lines, $l+1 \rightarrow l$ perception improves the baseline by 1.6 points. These prove the effectiveness of our inter-scale perception. We choose the latter one for better performance.

5.3.5 Visualizations

We visualize the segmentation results of complex and simple language descriptions in Fig. 5.4. Specifically, we compare three settings, *i.e.*, baseline with only single temporal scale, HTML w/o CMP which constructs hybrid temporal scales, and our final HTML which dynamically construct multimodal relations cross temporal scales. As expected, when only single temporal scale is adopted, baseline fails to segment the target in the video, *e.g.*, the background rabbit in the subplot (a) is mistakenly referred. The main reason is that the rabbit shares similar appearance to the target object and also locates *to the left of a person* in last three frames. The model is misled and confused by the single temporal scale. When applied with hy-

brid temporal scales, the language description can interact with both long and short temporal scales. Thus, the previous false prediction in the first frame is corrected. Further, when applied with CMP, our model is able to clarify the object confusion by discovering the core semantics *on a scale* and make correct predictions. Similarly, in subplot (b), baseline is misled by the video clip where tennis ball only appears in the middle two frames. When gradually applying our proposed modules, the mistakenly predicted dog is clarified and further the target tennis ball is correctly segmented.

5.4 Discussion

While our proposed approach achieves state-of-the-art results in referring video object segmentation (RVOS), it still faces limitations that highlight opportunities for further advancements. One key limitation is its focus on short-term to medium-term video sequences, where performance may degrade when applied to long-term videos with substantial scene changes, complex object interactions, or evolving contextual information. Additionally, the reliance on predefined temporal scales for frame sampling and processing constrains the model's adaptability to varying video dynamics, potentially limiting generalization in diverse scenarios. Furthermore, the approach does not fully leverage the capabilities of large-scale pretrained multimodal models, which have demonstrated remarkable success in capturing intricate cross-modal interactions and generalizing across datasets. To address these challenges, future research should explore extending the framework to handle long-term video segmentation by incorporating memory-augmented networks or global-attention-based transformers to maintain temporal coherence and capture long-range dependencies. Developing adaptive temporal scaling mechanisms that dynamically adjust frame sampling and processing based on video content or linguistic descriptions could further enhance the system's robustness and flexibility. Additionally, integrating

large vision-language models as feature extractors or fine-tuning them specifically for RVOS could boost performance and scalability across diverse datasets. Finally, exploring interactive RVOS, where systems incorporate user feedback or external contextual cues to refine segmentation, represents a promising direction for real-world applications. By addressing these limitations and pursuing these directions, RVOS systems can evolve to meet the demands of increasingly complex and diverse tasks.

5.5 Conclusion

In this work, we develop a HTML framework to align linguistic and visual features by learning multimodal relations hierarchically in different temporal scales. Moreover, we introduce an inter-scale multimodal perception module to construct dynamic multimodal interactions across temporal scales. We conduct experiments on four datasets and establish new state-of-the-art results. Particularly, our method with ResNet-50 backbone surpasses the recent methods with ResNet-100. The comprehensive ablation experiments and visualization results show that our method is able to discover core object semantics in the different modalities.

Chapter 6

Dual-AI: Dual-path Actor Interaction Learning for Group Activity Recognition

This chapter introduces the Dual-AI framework, which advances the understanding of group activity recognition. Dual-AI innovatively combines spatial and temporal data in dual paths to analyze both individual actions and their interactions within a group, enhancing the accuracy of recognizing complex group dynamics. This dual-path approach not only addresses the nuances of spatial and temporal data integration but also highlights how tailored actor interaction models can significantly improve the interpretation of collective activities in varied settings.

6.1 Introduction

Group Activity Recognition (GAR) is an important problem in video understanding. In this task, we should not only recognize individual action of each actor but also understand collective activity of multiple involved actors. Hence, it is vital to learn spatio-temporal actor relations for GAR.

Several attempts have been proposed to model actor relations by building visual attention among actors [70, 199, 209, 220, 54, 103, 10]. However, it is often difficult for joint spatial-temporal optimization [170, 13]. For this reason, the recent approaches in group activity recognition often decompose spatial-temporal attention separately for modeling actor interaction [54, 103, 209]. But single order of space and time is insufficient to describe complex group activities, due to the fact that different group activities often exhibit diversified spatio-temporal interactions.

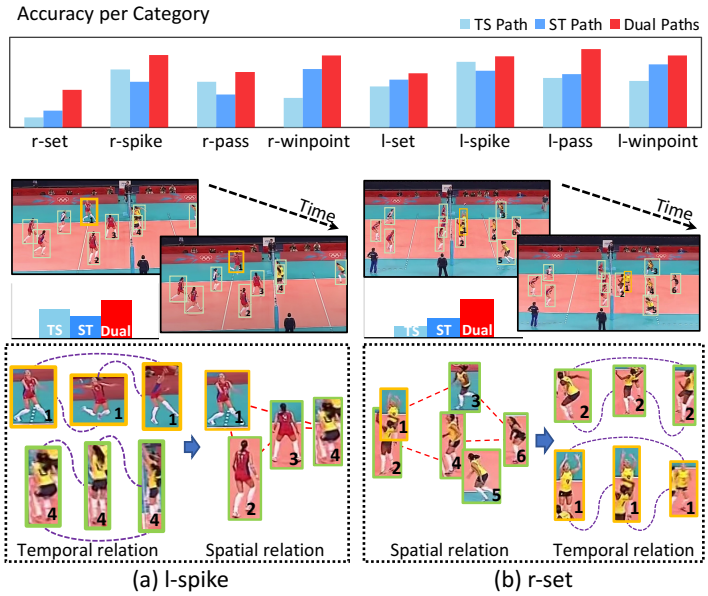


Figure 6.1 : Accuracy per Category and Example of *left spike* and *right set* group activity. Red dashed line and Violet dashed line below show spatial and temporal actor interaction respectively. With spatial and temporal modeling applied in different orders, ST path and TS path learn different spatiotemporal patterns and thereby are skilled at different classes, supported by the accuracy plot.

For example, Fig. 6.1 (a) refers to the *l-spike* activity in the volleyball, where the hitting player (actor 1) and the defending player (actor 4) move fast to hit and block the ball, while other accompanying players (*e.g.*, actor 2 and actor 3) stand without much movement. Hence, for this group activity, it is better to first understand temporal dynamics of each actor, and then reason spatial interaction among actors in the scene. On the contrary, Fig. 6.1 (b) refers to the *r-set* activity in the volleyball, where most players in the right-side team are moving cooperatively to tackle the ball falling on different positions, *e.g.*, actor 1 jumps and sets the ball, while actor 2 jumps together to make a fake spiking action. Hence, for this group activity, it is better to reason spatial actor interaction first to understand the action scene, and then model temporal evolutions of each actor. In fact, as shown in the

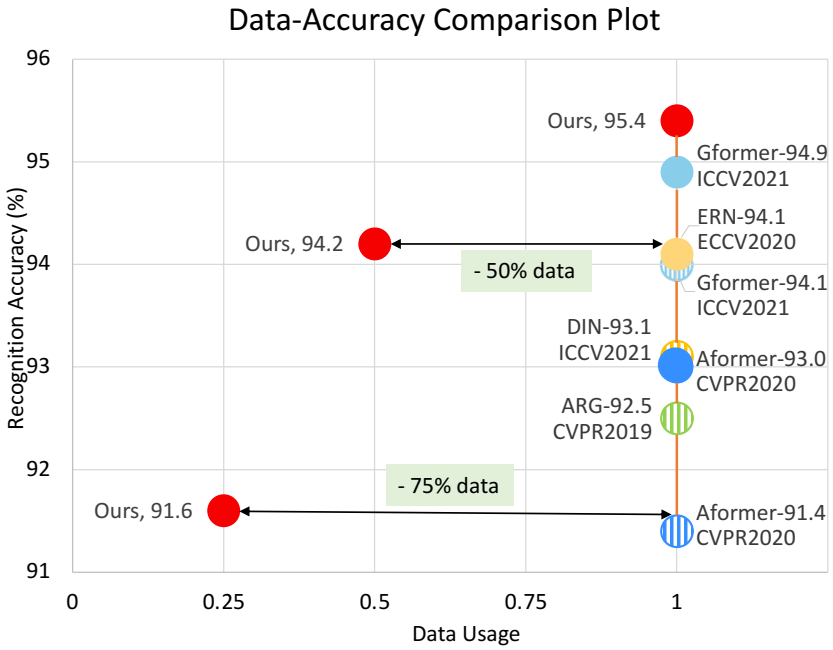


Figure 6.2 : **Accuracy comparison with data in different percentage on Volleyball dataset**. Our method achieves SOTA performance, and achieves 94.2% with 50% data, which is competitive to a number of recent approaches [136, 54, 199] trained with 100% data. Solid point means result with additional optical flow input.

accuracy plot of Fig. 6.1, the order of space and time interaction varies for different activity categories.

Based on these observations, we propose a distinct Dual-path Actor Interaction (Dual-AI) framework for GAR, which can effectively integrate two complementary spatiotemporal views to learn complex actor relations in videos. Specifically, Dual-AI consists of Spatial-Temporal (ST) and Temporal-Spatial (TS) Interaction Paths, with assistance of spatial and temporal transformers. ST path first takes spatial transformer to capture spatial relation among actors in each frame, and then utilizes temporal transformer to model temporal evolution of each actor over frames. Alternatively, TS path arranges spatial and temporal transformers in a reverse order to describe complementary pattern of actor interaction. In this case, our Dual-AI

can comprehensively leverage both paths to generate robust spatiotemporal contexts for boosting GAR.

Furthermore, we introduce a novel Multi-scale Actor Contrastive Loss (MAC-Loss), which is a concise but effective self-supervised signal to enhance actor consistency between two paths. Via such actor supervision in all the frame-frame, frame-video, video-video levels, we can further reduce action confusion between any two individual actors to improve the discriminative power of actor representations in GAR.

Finally, we conduct extensive experiments on the widely-used benchmarks to evaluate our designs. Our Dual-AI simply achieves state-of-the-art performance on all the fully-annotated datasets, such as Volleyball, Collective Activity. More interestingly, our Dual-AI with 50% training data is competitive to a number of recent approaches with 100% training data in Volleyball as shown in Fig. 6.2, which clearly demonstrates the generalization power of our Dual-AI. Motivated by this, we further investigate the challenging setting with limited actor supervision [209], where Dual-AI also achieves SOTA results on Weak-Volleyball-M and NBA datasets. All these results show that our Dual AI is effective for learning spatiotemporal actor relations in GAR.

6.2 Related Work

Group activity recognition has attracted a large body of work recently due to its wide applications. Early approaches are based on hand-crafted features and typically use probabilistic graphical models [1, 3, 2, 93, 94, 193] and AND-OR grammar methods [4, 151]. Recently, methods incorporating convolutional neural networks [11, 74] and recurrent neural networks [186, 207, 137, 11, 36, 152, 104, 74, 73] have achieve remarkable performance, due to the learning of temporal context and high-level information.

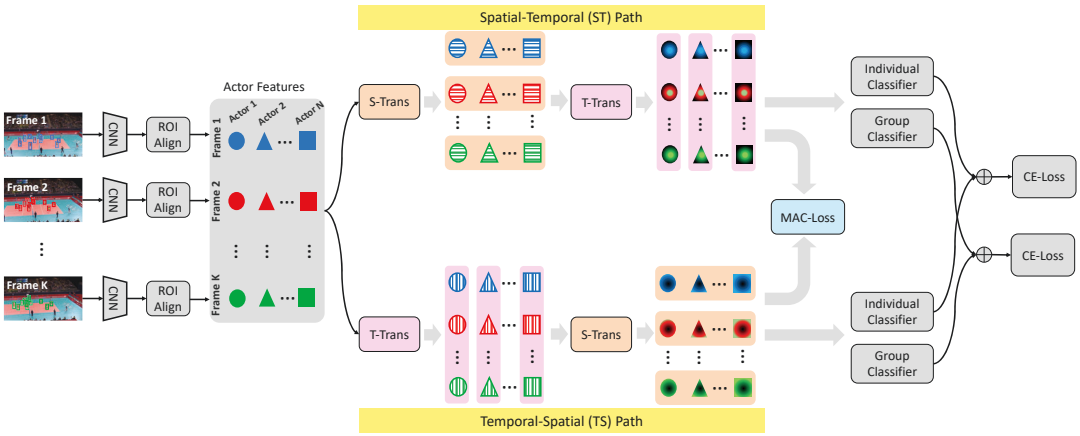


Figure 6.3 : Our Dual-path Actor Interaction (Dual-AI) learning framework, where S-Trans and T-Trans denote Spatial-Transformer and Temporal-transformer respectively. It effectively explores actor evolution in two complementary spatiotemporal views, *i.e.*, ST path and TS path, detailed in Sec. 6.3.2. Moreover, a Multi-scale Actor Contrastive loss is designed to enable interaction and cooperation of the two paths as in Sec. 6.3.3.

More recent group activity recognition methods [199, 54, 70, 209, 45, 136, 103, 220] often require the explicit representation of spatiotemporal relations, dedicated to apply attention-based methods to model the individual relations for inferring group activity. [199, 220] build relational graphs of the actors and explore the spatial and temporal actor interactions in the same time with graph convolution networks. These methods simulate spatiotemporal interaction of actors in a joint manner. Differently, [209] builds separate spatial and temporal relation graphs subsequently to model the actor relations. [54] encodes temporal information with I3D [22] and constructs spatial relation of the actors with a vanilla transformer. [103] introduces a cluster attention mechanism for better group informative features with transformers. Different from previous approaches, we propose to learn the actor interactions in complementary Spatial-Temporal and Temporal-Spatial views and

further promote actor interaction learning with a designed self-supervised loss for effective representation learning.

Vision Transformer has gradually become popular for computer vision tasks. In image domain, ViT[42] firstly introduces a pure transformer architecture without convolution for image recognition. Following works [102, 221, 120, 189] make remarkable progress on enabling transformer architecture to become a general backbone on various kinds of downstream computer vision tasks. In video domain, many works[64, 6, 101, 13, 46, 131] explore spatial and temporal self-attention to learn efficient video representation. TimeSformer[13] investigates the different space and time attention mechanisms to learn spatial-temporal representation efficiently. MViT[46] utilizes the multi-scale features aggregation to enhance the spatial-temporal representation. Motionformer[131] presents a trajectory-focused self-attention block, which essentially tracks space-time patches for video transformer. The above transformer architectures are designed for general video classification task. It has not been fully explored to tackle the challenging GAR problem with transformers. We propose to construct dual spatiotemporal paths with transformers to flexibly learn actor interactions for group activity recognition.

6.3 Method

To learn complex actor relations in the group activities, we propose a distinct Dual-path Actor Interaction (Dual-AI) framework for GAR. In this section, we introduce our Dual-AI in detail. First, we describe an overview of Dual-AI framework. Then, we explain how to build the interaction paths, with assistance of spatial and temporal transformers. Next, we introduce a Multi-scale Actor Contrastive Loss (MAC-Loss) to further improve actor consistency between paths. Finally, we describe the training objectives to optimize our Dual-AI framework.

6.3.1 Framework Overview

As shown in Fig. 6.3, our Dual-AI framework consists of three important steps. First, we need to extract actor features from backbone. Specifically, we sample K frames from the input video. To make a fair comparison with the previous works in GAR [11, 103, 199, 220, 219], we choose ImageNet-pretrained Inception-v3 [162] as backbone to extract feature of each sampled frame. Then, we apply RoIAlign [68] on the frame feature, which can generate actor features in this frame from bounding boxes of N actors. After that, we adopt a fully-connected layer to further encode each actor feature into a C dimensional vector. For convenience, we denote all the actor vectors as $\mathbf{X} \in \mathbb{R}^{K \times N \times C}$. More details can be found in Sec. 6.4.2.

After extracting actor feature vectors, we next learn spatiotemporal interactions among these actors in the video. Different from the previous approaches [199, 220, 209, 208, 54], we disentangle spatiotemporal modeling into consecutive spatial and temporal interactions in different orders. Specifically, we design spatial and temporal transformers as basic actor relation modules. By flexibly arranging these transformers in two reverse orders, we can enhance actor relations with complementary integration of both spatial-temporal (ST) and temporal-spatial (TS) interaction paths. Finally, we design training losses to optimize our Dual-AI framework. In particular, we introduce a novel Multi-scale Actor Contrastive Loss (MAC-Loss) between two paths, which can effectively improve discriminative power of individual actor representations, by actor consistency in all the frame-frame, frame-video, video-video levels. Subsequently, we integrate actor representations of two paths to recognize individual actions and group activities.

6.3.2 Dual-path Actor Interaction

To capture complex relations for diversified group activities, we propose a novel dual path structure to describe actor interactions. To start with, we build basic

spatial and temporal actor relation units, with assistance of transformers. Then, we explain how to construct dual paths for spatiotemporal actor interactions.

Spatial/Temporal Actor Relation Units

To understand spatiotemporal actor evolution in videos, we first construct basic units to describe spatial and temporal actor relations. Since there is no prior knowledge about actor relation, we propose to use transformer to model such relation by the powerful self-attention mechanism.

Spatial Actor Transformer. In order to model the spatial relation of the actors in single frame, we design a concise spatial actor transformer (S–Trans). Specifically, we denote $\mathbf{X}^k \in \mathbb{R}^{N\times C}$ as the feature vectors of N actors in the k-th frame. The spatial relation among these actors are modeled by $\hat{\mathbf{X}}^k = \text{S–Trans}(\mathbf{X}^k)$, which consists of three modules as follows,

$$\mathbf{X}' = \text{SPE}(\mathbf{X}^k) + \mathbf{X}^k, \tag{6.1}$$

$$\mathbf{X}'' = \text{LN}(\mathbf{X}' + \text{MHSA}(\mathbf{X}')), \tag{6.2}$$

$$\hat{\mathbf{X}}^k = \text{LN}\big(\mathbf{X}'' + \text{FFN}(\mathbf{X}'')\big). \tag{6.3}$$

First, we use spatial position encoding (SPE) to add spatial structure information of the actors in the scene, as in Eq. (6.1). We represent spatial position of each actor with center point of its bounding box and encode the spatial positions with PE function in [54, 20]. Second, we use multi-head self-attention (MHSA) [174] module to reason the spatial interaction of the actors in the scene, as in Eq. (6.2). Finally, we use feed-forward network (FFN) [174] to further improve learning capacity of the spatial actor relation unit, as in Eq. (6.3).

Temporal Actor Transformer. In order to model the temporal evolution of single actor across frames, we design a temporal actor transformer (T–Trans) following the way in Eqs. (6.1) to (6.3). Differently, we use the input as the feature

vectors of the n-th actor across K frames, *i.e.*, $\mathbf{X}^n \in \mathbb{R}^{K \times C}$. In this case, the MHSA module can reason the evolution of actor n in different time steps. Moreover, to add temporal sequence information of actor n, temporal position encoding (TPE) is used instead of SPE, which encodes frame index $\{1, ..., K\}$ with PE function in [174]. Finally, we can get actor features enhanced by temporal interactions, as $\hat{\mathbf{X}}^n = \text{T–Trans}(\mathbf{X}^n)$.

Dual Spatiotemporal Paths of Actor Interaction

Once the spatial and temporal relations of actors are built, we can further integrate them to construct spatiotemporal representation of the actor evolution. As discussed in Sec. 6.1, the single order of space and time is insufficient to understand the complex actor interactions, leading to the failure of inferring group activities. Thus, we propose a dual spatiotemporal paths framework for GAR to capture the complex interaction of the actors.

It consists of two complementary spatiotemporal modeling patterns for actor evolution, *i.e.*, Spatial-Temporal (ST) and Temporal-Spatial (TS), by switching the order of space and time as:

$$\mathbf{X}_{\text{ST}} = \text{T–Trans}(\mathbf{X} + \text{MLP}(\text{S–Trans}(\mathbf{X}))) \tag{6.4}$$

$$\mathbf{X}_{\text{TS}} = \text{S–Trans}(\mathbf{X} + \text{MLP}(\text{T–Trans}(\mathbf{X}))), \tag{6.5}$$

where we adopt a residual structure to enhance the actor representation. MLP with parameters in shape $C \times C$ is used to add non-linearity. By reshaping the frame and actor dimension as batch dimension, S–Trans and T–Trans reason about spatial and temporal actor interaction respectively.

By stacking spatial and temporal transformers in different orders, the actor representation is reweighted and aggregated according to different spatiotemporal context. ST path first reasons about the interaction of different actors in the scene of

each frame. Then, the temporal evolution is modeled to reweight the built actor interaction across different frames. As such, ST path is skilled at recognizing activities with distinct spatial arrangement, such as *set* in volleyball games. This activity requires the player to move to a new position and set the ball, usually accompanied by other players moving or jumping for fake spiking. Complementarily, TS path reasons about the actor evolution, in the opposite order of ST path. It considers temporal dynamics of each actor in the first place, and then reasons about spatial actor interaction to understand the scene. Hence, it is skilled at recognizing activities with distinct actor evolution patterns, such as *spike* in volleyball games, which requires hitter to jump and quickly hit the ball.

Subsequently, to fully take advantage of such complementary characteristic, we feed the representation of actors from ST and TS paths to generate individual actions and group activity predictions, and fuse them as final predictions of dual spatiotemporal paths.

6.3.3 Multi-scale Actor Contrastive Learning

The actor representation is reweighted and aggregated by dual spatiotemporal paths, however, the modeling process is independent. To promote cooperation of these two complementary paths, we design a self-supervised Multi-scale Actor Contrastive loss (MAC-loss). As dual spatiotemporal paths model evolution of each actor in different patterns, we define a pretext task of actor consistency. Specifically, we design such constraints in multiple scales of frame and video levels.

Frame-Frame Actor Contrastive Loss. The frame representation of the actor in one path should be similar with its corresponding frame representation in the other path, while different from other frame representation of this actor in the path. As shown in Fig. 6.4 (a), taking actor n in ST path as an example, we attract frame representation in k-th frame ($\mathbf{X}_{\mathrm{ST}}^{n,k}$) to its corresponding representation from

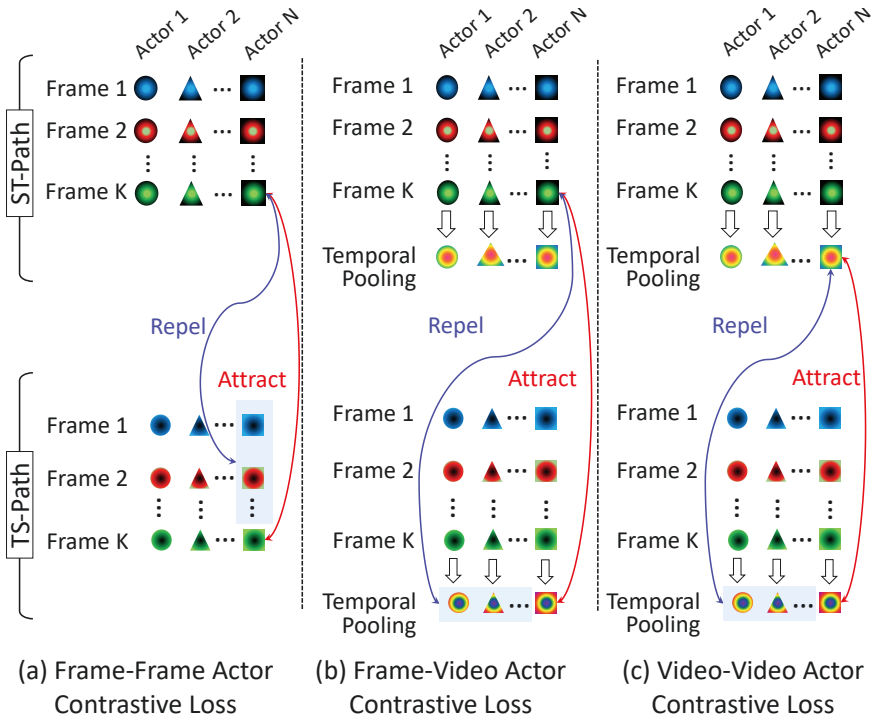


Figure 6.4 : Illustration of MAC-loss for Actor N. It consists of three levels, *i.e.*, frame-frame, frame-video and video-video. The blue block means the source of negative pairs. For simplicity, we only show the constraints from ST path to TS path. It is similar for the constraints from TS path to ST path.

TS path ($\mathbf{X}_{\mathrm{TS}}^{n,k}$). Meanwhile, we repel the representation of actor n in other frames from TS path ($\mathbf{X}_{\mathrm{TS}}^{n,t}$, where $t \neq k$),

$$\mathcal{L}_{ff}(\mathbf{X}_{\mathrm{ST}}^{n,k}, \mathbf{X}_{\mathrm{TS}}^{n,k}) = -\log \frac{h(\mathbf{X}_{\mathrm{ST}}^{n,k}, \mathbf{X}_{\mathrm{TS}}^{n,k})}{\sum_{t=1}^{K} h(\mathbf{X}_{\mathrm{ST}}^{n,k}, \mathbf{X}_{\mathrm{TS}}^{n,t})}, \tag{6.6}$$

where $h(\mathbf{u}, \mathbf{v}) = \exp(\frac{\mathbf{u}^\top \mathbf{v}}{||\mathbf{u}||_2 ||\mathbf{v}||_2})$ is the exponential of cosine similarity measure. Vice versa, the loss for actor n in TS path can be obtained by $\mathcal{L}_{ff}(\mathbf{X}_{\mathrm{TS}}^{n,k}, \mathbf{X}_{\mathrm{ST}}^{n,k})$.

Frame-Video Actor Contrastive Loss. The frame representation of the actor in one path should be consistent with its video representation in the other path, while different from video representation of other actors in the path. As shown in Fig. 6.4 (b), taking actor n in ST path as an example, we attract its frame representation

$\mathbf{X}_{\mathrm{ST}}^{n,k}$ to its video representation $\tilde{\mathbf{X}}_{\mathrm{TS}}^{n}$ from TS path, which is obtained by pooling frame representation $\mathbf{X}_{\mathrm{TS}}^{n,1:K}$. Meanwhile, we repel the video representation of other actors in the minibatch from TS path ($\tilde{\mathbf{X}}_{\mathrm{TS}}^{i}$, where $i \neq n$),

$$\mathcal{L}_{fv}(\mathbf{X}_{\mathrm{ST}}^{n,k}, \tilde{\mathbf{X}}_{\mathrm{TS}}^{n}) = -\log \frac{h(\mathbf{X}_{\mathrm{ST}}^{n,k}, \tilde{\mathbf{X}}_{\mathrm{TS}}^{n})}{\sum_{i=1}^{B\times N} h(\mathbf{X}_{\mathrm{ST}}^{n,k}, \tilde{\mathbf{X}}_{\mathrm{TS}}^{i})}, \tag{6.7}$$

where B denotes the minibatch size. Vice versa, the loss for actor n in TS path can be obtained by $\mathcal{L}_{fv}(\mathbf{X}_{\mathrm{TS}}^{n,k}, \tilde{\mathbf{X}}_{\mathrm{ST}}^{n})$.

Video-Video Actor Contrastive Loss. Furthermore, we constrain the consistency of video representation of each actor across dual paths, as shown in Fig. 6.4 (c). We achieve this by minimizing cosine similarity measure $\mathcal{L}_{w}$ of corresponding video representation ($\tilde{\mathbf{X}}_{\mathrm{TS}}^{n}, \tilde{\mathbf{X}}_{\mathrm{ST}}^{n}$). Our proposed MAC-loss is then formed as

$$\mathcal{L}_{MAC} = \lambda_{ff}\mathcal{L}_{ff} + \lambda_{fv}\mathcal{L}_{fv} + \lambda_{w}\mathcal{L}_{w}, \tag{6.8}$$

where $\lambda_{\{\cdot\}}$ denote weights for the different components.

6.3.4 Training objectives

Our network can be trained in an end-to-end manner to simultaneously predict individual actions of each actor and group activity. Combining with standard cross-entropy loss, the final loss for recognition is formed as

$$\mathcal{L}_{cls}{=}\mathcal{L}_{CE}(\frac{\hat{y}_{\mathrm{ts}}^{G}+\hat{y}_{\mathrm{st}}^{G}+\hat{y}_{scene}^{G}}{3}, y^{G}) + \lambda\mathcal{L}_{CE}(\frac{\hat{y}_{\mathrm{ts}}^{I}+\hat{y}_{\mathrm{st}}^{I}}{2}, y^{I}), \tag{6.9}$$

where $\hat{y}_{\{ts,st\}}^{I}$ and $\hat{y}_{\{ts,st\}}^{G}$ denote individual action and group activity predictions from TS and ST paths. y^{I} and y^{G} represent the ground truth labels for the target individual actions and group activity. $\hat{y}_{scene}^{G}$ denotes the scene prediction produced by separate group activity classifier, using features directly from backbone. λ is the hyper-parameter to balance the two items. Finally, we combine all the losses to train our Dual-AI framework,

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{MAC}. \tag{6.10}$$

During inference, we infer the individual actions and group activity by averaging the predictions from the dual spatiotemporal paths.

Table 6.1 : Comparison with state-of-the-art methods on **Volleyball dataset** in term of Acc.%.

Method	Backbone	Data Ratio	Optical Flow	Individual Action	Group Activity
HDTM[74]	AlexNet	100%		-	81.9
CERN[150]	VGG16	100%		-	83.3
StageNet[137]	VGG16	100%		-	89.3
HRN[73]	VGG19	100%		-	89.5
SSU[11]	Inception-v3	100%		81.8	90.6
AFormer[54]	I3D	100%		-	91.4
ARG[199]	Inception-v3	100%		83.0	92.5
TCE+STBiP [219]	Inception-v3	100%		-	93.3
DIN [220]	ResNet-18	100%		-	93.1
GFormer[103]	Inception-v3	100%		83.7	94.1
Ours	Inception-v3	25%		82.1	89.7
	Inception-v3	50%		83.0	92.7
	Inception-v3	100%		**84.4**	**94.4**
SBGAR[104]	Inception-v3	100%	✓	-	66.9
CRM[10]	I3D	100%	✓	-	93.0
Aformer[54]	I3D	100%	✓	83.7	93.0
JLSG[45]	I3D	100%	✓	83.3	93.1
ERN[136]	R50-FPN+I3D	100%	✓	81.9	94.1
GFormer[103]	I3D	100%	✓	84.0	94.9
Ours	Inception-v3	25%	✓	83.0	91.6
	Inception-v3	50%	✓	84.0	94.2
	Inception-v3	100%	✓	**85.3**	**95.4**

Table 6.2 : Comparisons with previous state-of-the-art methods on **Collective Activity datatset**.

Method	Backbone	MPCA
HDTM[74]	AlexNet	89.7
PCTDM[207]	AlexNet	92.2
CERN-2[150]	VGG-16	88.3
Recurrent[187]	VGG-16	89.4
stagNet[137]	VGG-16	89.1
SPA+KD[167]	VGG-16	92.5
PRL[70]	VGG-16	93.8
CRM[10]	I3D	94.2
ARG[199]	ResNet-18	92.3
HiGCIN[208]	ResNet-18	93.0
DIN[220]	ResNet-18	95.3
TCE+STBiP[219]	Inception-v3	95.1
Ours	ResNet-18	**96.0**
	Inception-v3	**96.5**

6.4 Experiments

6.4.1 Dataset

Volleyball Dataset. This dataset [74] consists of 4,830 labeled clips (3493/1337 for training/testing) from 55 volleyball games. Each clip is annotated with one of 8 group activity classes. Middle frame of each clip is annotated with 9 individual action labels and their bounding boxes.

Collective Activity Dataset. This dataset [31] contains 44 short videos with every ten frames annotated with individual action labels and their bounding boxes. The group activity class of each clip is determined by the largest number of the individual action classes. We follow [208, 207, 220] to merge the *crossing* and *walking*

Table 6.3 : Comparision with state-of-the-art methods on **NBA and Weak-Volleyball-M dataset** following metrics adopted in [209]. * means the results are from [209].

Method	Backbone	Mod-ality	NBA Acc./Mean Acc.	Weak Vlb. -M Acc.
TSN*[184]	Incep-v1	RGB	– / 37.8	–
I3D*[22]	I3D	RGB	– / 32.7	–
Nlocal*[190]	I3D-NLN	RGB	– / 32.3	–
ARG*[199]	Incep-v3	RGB	– / –	90.7
SAM[209]	Res-18	RGB	– / –	93.1
SAM[209]	Incep-v3	RGB	49.1 / 47.5	94.0
Ours	Incep-v3	RGB	51.5 / 44.8	95.8
	Incep-v3	Flow	56.8 / 49.1	96.1
	Incep-v3	Fusion	**58.1** / **50.2**	**96.5**

Table 6.4 : Comparison with state-of-the-art methods trained with Volleyball dataset of different data ratios in term of group activity recognition Acc.%.

Method	5%	10%	25%	50%	100%
PCTDM[207]	53.6	67.4	81.5	88.5	90.3
AFormer[54]	54.8	67.7	84.2	88.0	90.0
HiGCIN[208]	35.5	55.5	71.2	79.7	91.4
ERN[136]	41.2	52.5	73.1	75.4	90.7
ARG[199]	69.4	80.2	87.9	90.1	92.3
DIN[220]	58.3	71.7	84.1	89.9	93.1
Ours	**76.2**	**85.5**	**89.7**	**92.7**	**94.4**

into *moving*.

Weak-Volleyball-M Dataset. This dataset [209] is adapted from Volleyball dataset while merging *pass* and *set* categories to have total 6 group activity classes,

and discarding all individual annotations (including individual action labels and bounding boxes) for weakly supervised GAR.

NBA Dataset. This dataset [209] contains 9,172 annotated clips (7624/1548 for training and testing) from 181 NBA game videos, each of which belongs to one of the 9 group activities. No individual annotations, such as individual action labels and bounding boxes, are provided.

6.4.2 Implementation Details

We select the Inception-v3 model as our CNN backbone, following widely used settings [11, 103, 199, 220, 219] in GAR. We also use ResNet-18 model as backbone for Collective Activity Dataset, following widely used settings [208, 220]. We apply the ROI-Align with crop size 5×5 and a linear embedding to get actor features with dimension $C = 1024$. Each Spatial or Temporal transformer has one attention layer with 256 embedding dimension. The $\lambda_{ff}, \lambda_{fv}, \lambda_{vv}$ in MAC-Loss are all set 1. More details for K and N can be found in supplementary material.

6.4.3 SOTA Comparison

Full Setting. This setting allows us to train our model with all data fully annotated with group activities and individual annotations. We compare our method with the state-of-the-art approaches on Volleyball and Collective Activity dataset. As shown in Tab. 6.1, our approach (94.4%) with only RGB frames and Inception backbone has already outperformed other SOTA methods with computationally high backbones (I3D, FPN) and additional optical flow input. Furthermore, equipped with RGB and optical flow late fusion, our method can improve the SOTA result by a large margin to 95.4%. Remarkably, even with only 50% data, our method still surpasses the vast majority of the SOTA methods with 100% data, *e.g.*, Ours (50%) vs. SARF (100%): 94.2 vs. 93.1. As shown in Tab. 6.2, our approach also

achieves state-of-the-art performance on Collective Activity dataset. These results demonstrate the effectiveness of our method.

Weakly Supervised Setting. Under this setting we use all raw data and group activity annotations, without any individual annotations. We follow the [209] to report results on Weak-Volleyball-M dataset and NBA dataset. As shown in Tab. 6.3, our method surpasses all the existing methods by a good margin, establishing new state-of-the-art results. Specifically, our approach improves the previous SOTA [209] by 2.5% on Weak-Volleyball-M and by 9% on NBA dataset in term of Acc.%. It indicates that our Dual-AI framework can enhance the learning ability of the model to obtain robust representation and achieve promising performance in the case individual annotations missing.

Limited Data Setting. In this setting, we train our method with random sampled data in different ratios to show the generalization power of our method. To compare the results under this setting, we implement a number of previous SOTA methods that have the officially-published codes available. As shown in Tab. 6.4, our method surpasses previous SOTA methods in all data ratios. Moreover, with the available training data decreasing, the performance of our method remains promising and the gain against other methods gets enlarged, which demonstrates the robustness of our method.

6.4.4 Ablation Study

Dual Spatial Temporal Paths. To validate the effectiveness of our Dual Spatiotemporal Paths, we investigate six settings. Particularly, we experiment with 50% data for limited Volleyball. In addition to T-S and S-T introduced in Section Sec. 6.3.2, other two paths, *i.e.*, S-S and T-T are introduced to validate in a broader range. S-S/T-T means that features go through two successive Spatial/Temporal-Transformer, respectively. As shown in Tab. 6.5, our Dual Paths achieves the best

Table 6.5 : Effectiveness of our Dual Path Actor Interaction.

Dual-Path	Weak Volleyball-M	Limited Volleyball	Full Volleyball
S-S	88.9	88.4	91.2
T-T	91.6	87.9	90.9
S-T	93.0	89.3	92.2
T-S	92.6	89.5	92.1
ST-TS Fusion	**94.2**	**90.8**	**93.3**

Table 6.6 : Effectiveness of our MAC-loss. Different components are ablated on Volleyball dataset in term of Acc.%.

Components of MAC-loss			Data Ratio	
F-F	F-V	V-V	50%	100 %
			90.8	93.3
✓			91.2	93.5
	✓		91.0	93.3
		✓	91.6	93.6
✓	✓	✓	**92.1**	**94.0**

result under different setting. The reason is that, dual-path TS and ST are good at inferring different group activities and the learned representation from ST and TS can complement each other, leading to a better performance. This demonstrates that our dual path ST-TS is a preferable way to comprehensively leverage both paths to generate robust spatiotemporal contexts for boosting group activity recognition.

Multi-scale Actor Contrastive Loss. We explore the performance of our network with different components of MAC loss. As shown in Tab. 6.6, with different component of consistent loss (frame-frame, frame-video, video-video), our network consistently outperforms w/o consistent loss. By utilizing all components

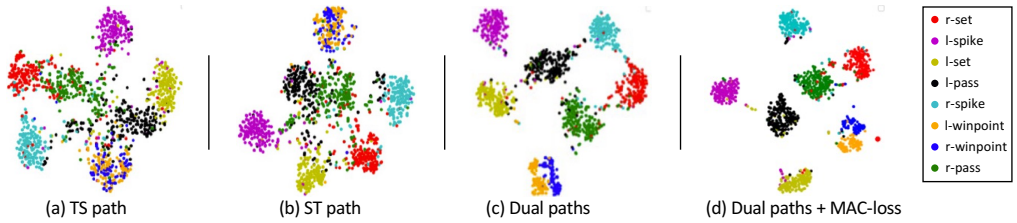


Figure 6.5 : t-SNE [173] visualization of video representation on the Volleyball dataset learned by different variants of our Dual-AI model: ST path only, TS path only, Dual spatiotemporal paths, and final Dual-AI model.

Scene Fusion	Data Ratio	
	50%	100%
w/o	92.1	94.0
Early	92.0	93.9
Middle	92.2	94.0
Late	**92.7**	**94.4**

Table 6.7 : Effectiveness of scene information.

of MAC-loss, our network can achieve the best results. Note that, given less available training data, the loss can help network get a larger accuracy improvement. It demonstrates that the MAC-loss can enable cooperation of the dual complementary modeling process, thereby enhancing the learned representation from ST and TS paths, especially with limited available data.

Scene Information. We investigate the effectiveness of scene information, by exploring the way to fuse scene context in a early, middle and late fusion manner. As shown in Tab. 6.7, late scene context fusion is the best choice. Regardless of the available data ratio, the scene information can improve the performance by around 0.6 in term of Acc.%. This is because that scene information can provide global-level context, which can supplement the actor-level relation modeling and is crucial to GAR.

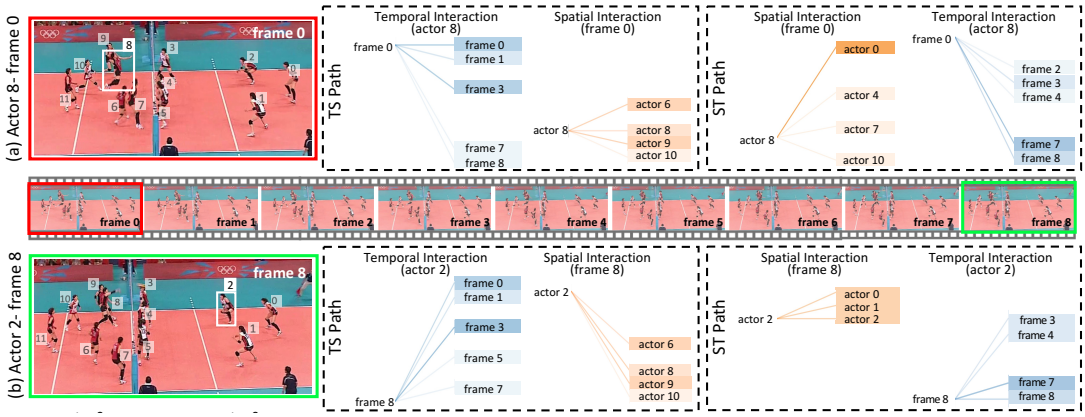


Figure 6.6 : Actor interaction visualization for *l-spike* activity with connected lines. Brighter color indicates stronger relation. (a) For actor 8 in frame 0, we visualize the temporal interaction with same actors in different frames for ST and TS paths; similarly, we visualize the spatial interaction with different actors in frame 0. (b) We visualize the actor interaction for actor 2 in frame 8 in the same way.

6.4.5 Visualization

Group Feature Visualization. Fig. 6.5 shows the t-SNE [173] visualization of the learned representation. We project video representation extracted from Volleyball validation dataset to 2-D dimension using t-SNE. We can see that learned representation from Dual Path transformer (c) can be grouped better than single Temporal-Spatial path (a) and Spatial-Temporal path (b). Furthermore, equipped with MAC-loss, our Dual-AI network (d) is able to differentiate group representations much better. These results demonstrate the effectiveness of our Dual-AI framework.

Spatial/Temporal Actor Attention Visualization. We visualize the actor interaction of *l-spike* activity in Fig. 6.6. The attention weight between actors is represented by connected lines, and the brightness of the lines represents the scale of the attention weight. Orange and Blue lines correspond to the Spatial and Temporal

interaction, respectively. As shown by spatial interaction in Fig. 6.6 (a), the spiking player (actor 8) is more related with accompanying players in TS path, who are "moving" (actor 6 and 10) and "standing" (actor 9). Differently, in ST path, actor 8 has wider connections with accompanying players (*e.g.*, actor 7 and actor 10) and defending players (*e.g.*, actor 0 and actor 4). Similarly, as shown by spatial interaction in Fig. 6.6 (b), the actor 2 is related to different accompanying and defending players in TS path and ST path respectively, showing complementary patterns. As for temporal interaction in both (a) and (b), the anchor actor is more related with early frames (frame 0 and frame 3) in TS path, while more related with late frames (frame 7 and frame 8) in ST path, showing highly complementary patterns.

6.5 Conclusion

In this work, we develop a Dual-AI framework to flexibly learn actor interactions in Spatial-Temporal and Temporal-Spatial views. Furthermore, we design a distinct MAC-loss to enable cooperation of dual paths for effective actor interaction learning. We conduct experiments on three datasets and establish new state-of-the-art results under different data settings. Particularly, our method with 50% data surpasses a number of recent methods trained with 100% data. The comprehensive ablation experiments and visualization results show that our method is able to learn actor interaction in a complementary way.

Chapter 7

Video Recognition in Portrait Mode

This chapter shifts focus to the emerging trend of portrait mode video recognition. With the increasing prevalence of portrait mode videos on social media platforms, this format presents unique challenges due to its distinct aspect ratio and content characteristics. This chapter explores the development of specialized methodologies and the introduction of the PortraitMode-400 dataset, designed specifically to optimize recognition techniques for portrait mode videos. These advancements aim to better accommodate the vertical orientation and subject-centric content typical of this video format, enhancing both accuracy and applicability in contemporary media environments.

7.1 Introduction

Most efforts in video recognition have focused on improving the accuracy and efficiency of different models and architectures on public benchmarks. Over the past two decades, there has been a dramatic shift in the types of video recognition models, starting from bags of features [156, 180, 133, 172, 179, 132, 163], moving on to convolutional neural networks [205, 80, 49, 184, 50, 182, 210, 111, 170, 169, 22], and more recently, vision transformers [7, 6, 101, 106, 47, 13, 121, 18, 119, 221, 131]. With the evolution of various models, video datasets have played a crucial role in driving each generation of models. The introduction of each video dataset has guided the research community to focus on new challenges. We have moved from using datasets collected in controlled environments (*e.g.*, KTH [146], Weizmann [15]) to more realistic videos (*e.g.*, UCF101 [158], HMDB51 [90]), and now to large-scale

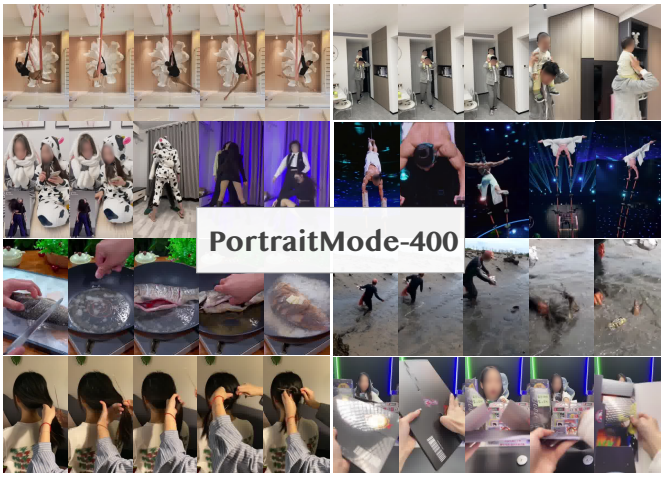


Figure 7.1 : A glance of PortraitMode-400, which is the first dataset dedicated to portrait mode video recognition. It covers videos from 9 domains and 400 specific categories. We show video samples (left to right, top to down) for *aerial yoga*, *riding neck*, *partner dancing (pop music)*, *acrobatics*, *cooking fish soup*, *catching crab*, *styling hair with hairpins* and *opening mystery card packs*, from different domains of our dataset.

web video datasets (*e.g.*, Kinetics-700 [21], HowTo100M [125]).

While existing video datasets are mostly built on landscape mode videos, portrait mode videos have become increasingly more popular on major social media applications. The shift from landscape mode to portrait mode is not just changing the aspect ratios of the videos. It has significant implications for the types of content that are created and the spatial bias inherent in the data. Portrait mode videos bring in distinct challenges for video recognition as well. For example, they tend to focus more on the subject (*i.e.*, typically humans) with much less background context, and include more egocentric content. In addition, they contain a lot of verbal communication that is essential to understand the video content. There is a pressing need for portrait mode video datasets to explore these new research problems.

This chapter introduces the first dataset dedicated to portrait mode video recognition, named PortraitMode-400 (abbreviated as PM-400), shown in Figure 7.1. The dataset consists of 76k videos collected from Douyin*, a popular short-video application, and annotated with 400 categories. The dataset's taxonomy is built in a data-driven way by aggregating search queries and covers a wide range of categories, including sports, food, music, handicrafts, and daily activities, among others. Many of the categories are fine-grained, as shown in Figure 7.2 (a). The data annotation was performed by professionally trained human annotators, and additional quality assurance was conducted to improve the annotation accuracy and consistency. We built PortraitMode-400 as a single-label dataset, and removed videos that can be tagged with multiple labels during annotation. While the recent 3Massiv [63] dataset also includes a significant percentage of portrait mode videos, it is mostly built for multi-lingual and multi-modal research, and only has 34 coarse visual concepts, unlike PortraitMode-400.

In addition to introducing the PortraitMode-400 dataset, we have also made preliminary attempts to investigate several critical research problems related to portrait mode video recognition:

- How well does a model trained on landscape mode videos perform on portrait mode videos, and vice versa? We investigate this question by constructing a subset from the Kinetics-700 dataset [21] for a rigorous comparison and visualize classification heatmaps (shown in Figure 7.3 and Figure 7.4) to reveal the differences in spatial bias resulting from the change in video format.

- What are the optimal training and testing protocols for portrait mode video recognition? We delve into various components of state-of-the-art deep learn-

* Douyin is a popular social media application built for smartphones and primarily features portrait mode short-form videos. https://www.douyin.com/

ing systems, such as data augmentation, evaluation cropping strategies, *etc.* Our discoveries challenge the existing conventions for landscape mode videos, thus necessitating further exploration into portrait mode videos.

- How important is temporal information for portrait mode videos? Can we recognize the actions from single frames [61] or do we need to utilize temporal information for accurate results? We explore different temporal utilization strategies and find that integrating temporal information substantially improves video recognition in portrait mode.

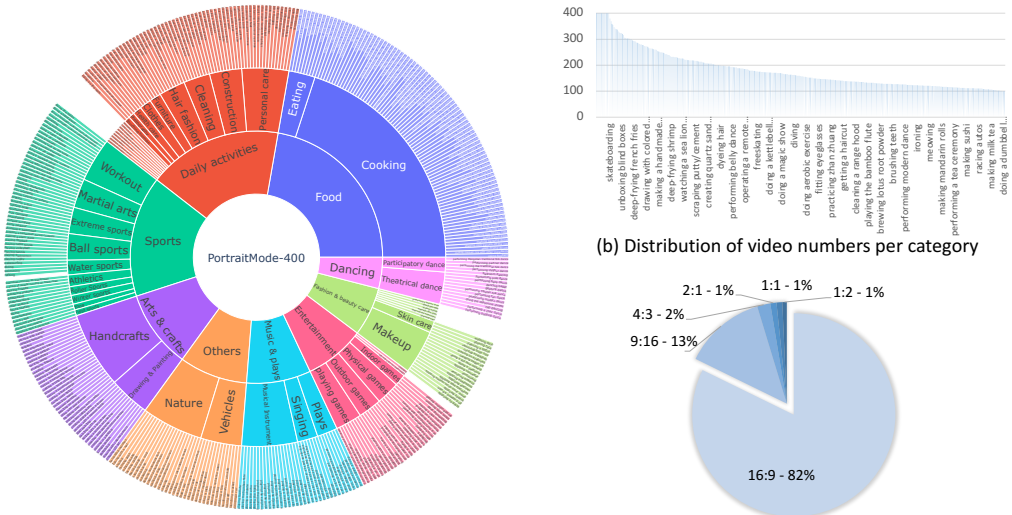
- Audio is a critical modality for video understanding [52, 82]. Does audio contribute to video recognition in portrait mode? Our experiments show that even simple audio integration can improve recognition accuracy, indicating possibilities for multimodal video analysis.

7.2 The PortraitMode-400 dataset

In this section, we provide a comprehensive overview of the process behind constructing our PortraitMode-400 dataset. We begin by discussing our data-driven approach to building a taxonomy, which is based on user queries. Next, we detail our rigorous annotation process and the criteria we applied to ensure high-quality and consistent annotations. Finally, we compare PortraitMode-400 with existing datasets that are relevant to our work, highlighting the unique contributions and advantages of our dataset.

7.2.1 Taxonomy

The videos in PortraitMode-400 were sourced from Douyin*. To better capture the various types of content that portrait mode videos can provide, we created a new taxonomy for PortraitMode-400 instead of reusing categories from existing datasets.



(a) Hierarchical structure of our taxonomy

(b) Distribution of video numbers per category

(c) Distribution of aspect ratios for retrieved videos

Figure 7.2 : Overview of our dataset. (a) We construct our taxonomy in a three-level hierarchical structure, which contains 9 domains and 400 leaf-node categories. (b) We show the distribution of video numbers per category of our dataset, which contains a relatively balanced distribution of categories. (c) We plot the distribution of aspect ratios for the retrieved videos via search queries. The majority of videos (over 85%) are in portrait mode, with 16:9 being the dominant format.

Our approach involved building the taxonomy based on popular search queries from Douyin users, which often include text descriptions about the corresponding videos. However, we found that many search queries lacked visual semantic meaning, such as celebrity names or song names. To address this, we manually selected candidate queries containing verbs (*e.g.*, "eating cakes") or nouns indicating potential actions (*e.g.*, "concealer" which often leads to videos about how to use a concealer). After manually examining approximately 38k search queries, we identified about 2.4k usable queries with corresponding videos that might contain actions or motions, as we aimed to incorporate more temporal information in the final dataset.

With the initial set of selected search queries, our second step is to recursively

Table 7.1 : Comparison of different portrait mode video datasets. S100-PM is a portrait-mode-only subset sampled from Kinetics-700, as detailed in Section 7.2.3. 3Massiv contains 5% landscape mode videos and is targeted for video classification in 34 coarse categories. Our PortraitMode-400 contains portrait mode videos only and has more videos in a diversified taxonomy (400 classes).

Dataset	% of PM	# of Classes	# of Videos	Duration	Avg. Duration	Year
S100-PM [21]	100%	100	20k	1s-10s	9s	'19
3Massiv [63]	95%	34	50k	5s-2min	20s	'21
PortraitMode-400	100%	400	76k	2s-1min	27s	'23

aggregate the queries in a bottom-up manner. This process generates increasingly abstract concepts, resulting in a hierarchical tree structure taxonomy, as illustrated in Figure 7.2 (a). In addition to producing the final taxonomy, we have two other objectives in this step: 1) merging similar queries into a final leaf node category of the taxonomy; 2) splitting or removing queries that may overlap with existing categories, so that all final categories are mutually exclusive. For example, we merge *tutorials for fitness*, *exercises for weight loss* and *fat-burning fitness exercises* to *aerobics*; we split *calligraphy exercise* into *pen calligraphy* and *brush calligraphy*. After completing the second step, we obtained about 500 candidate categories derived from the 2.4k selected search queries, which are organized in a three-layer hierarchy as depicted in Figure 7.2.

The taxonomy used in the Kinetics-400 [81] dataset is built through a combination of reusing categories from previous datasets and crowdsourcing. In contrast to Kinetics-400, our taxonomy is developed using a data-driven approach that better reflects the current trends in social media. Besides, our taxonomy covers a wider

range of content, including everyday activities (*food*, *beauty care*, *entertainment*, *etc*) , natural phenomena (*raining*, *snowing*, *etc*) as well as transportation-related activities (*airplane taking off*, *launching rocket*, *etc*). This is in contrast to existing datasets that mostly focus on human actions. Furthermore, our taxonomy offers more fine-grained categories compared to 3Massiv [63], which is designed for coarse visual concept classification. For instance, while 3Massiv has only one class for food, our taxonomy includes 89 distinct categories under the food parent node, covering various types of food and food-related activities such as cooking and eating.

7.2.2 Sampling and annotation

For each of the 500 candidate categories in the taxonomy, we have about 2 to 50 selected search queries associated with it, as described in Section 7.2.1. We retrieve 1.2k to 740k videos for each query from Douyin* depending on how frequently the query has been searched. Subsequently, we create a pool of videos for each category by aggregating all the retrieved videos from their corresponding queries. Figure 7.2 (c) illustrates the distribution of the aspect ratio of the retrieved videos. Although 16:9 is the dominant aspect ratio, there are also other aspect ratios for portrait mode videos, such as 4:3. For the video pool, we use a few criteria to sample target videos for annotation: 1) we select videos whose aspect ratios (height/width) are greater than 1 to ensure that PortraitMode-400 includes only portrait mode videos; 2) we select videos whose duration is shorter than 1 minute to limit annotation costs; and 3) we select videos that have been viewed over 700 times by Douyin users to ensure that our dataset better reflects the typical types of content for portrait mode videos.

Finally, we perform deduplication on the video pool to eliminate duplicated or similar videos. To achieve this, we extract feature vectors of each video using Uniformer-Base [101] pretrained on Kinetics-700 dataset [21]. Next, we build a graph by connecting video pairs with feature vectors having a cosine similarity greater

than 0.98. We then apply the Louvain algorithm [16] on the graph to identify video clusters and discard all the videos in each cluster except one. About 25% of videos are removed through deduplication, and only videos that meet all the aforementioned criteria move on to the next stage for human annotation.

The human annotation task is straightforward. An annotator is presented with a given category and its video pool, and is asked to confirm or deny whether the category name is a good match for the content of each video. Before starting annotation, annotators undergo training to learn the annotation criteria for all candidate categories, and they are required to pass a quality check test. Only annotators with an accuracy greater than 95% are qualified for annotation to ensure the accuracy and consistency of their annotations. During annotation, annotators discard videos that may be confused with multiple categories of our taxonomy, ensuring that PortraitMode-400 is a strictly single-label dataset. Under our restricted rules, approximately 65% of videos are rejected. To ensure annotation quality, approximately 20% of annotations are reviewed by two additional examiners.

7.2.3 Comparisons with existing datasets

After finishing annotating all the videos, we keep all the categories that have at least 100 videos. We keep at most 400 videos per category so that the distribution of videos across different categories are more or less balanced, as shown in Figure 7.2 (b). Our dataset contains 76k videos in total, spanning over 400 categories. We randomly sample 50 videos per category for testing, and the rest are used for training. Table 7.1 compares the statistics of PortraitMode-400 with other relevant datasets. Though 3Massiv mostly includes portrait mode videos, it is a multi-lingual and multi-modal dataset designed for concept recognition with only 34 coarse concepts. PortraitMode-400 has a more diversified and fine-grained taxonomy that is dedicated for portrait mode video recognition.

To conduct a rigorous comparison between landscape mode and portrait mode video recognition, we created two subsets from the Kinetics-700 dataset: a portrait mode subset and a corresponding landscape mode subset. The details of these subsets are shown in Table 7.1. We first constructed the portrait mode subset, named **Selected-100** Portrait Mode (S100-PM), using the top 100 categories with the most portrait mode videos in Kinetics-700. Each category in S100-PM contains 160 to 352 portrait mode videos, resulting in a total of 20k videos. To build a counterpart landscape mode version from Kinetics-700, we sampled the same number of landscape mode videos as S100-PM for each category, resulting in a landscape mode subset named Selected-100 Landscape Mode (S100-LM). Therefore, S100-PM and S100-LM have the same taxonomy and the same video distribution per category. Although the video content of S100-PM and S100-LM may differ due to different video formats, we believe that they are still useful benchmarks for illustrating and validating the difference between landscape mode and portrait mode video recognition. We have also tried AutoFlip[†] to convert landscape mode videos to portrait mode, thereby ensuring the same video content in both subsets. However, the converted portrait mode videos had unsatisfactory data quality. Thus, building S100-PM and S100-LM from Kinetics-700 remains the best option for rigorously comparing different video formats on recognition tasks.

7.3 Landscape Mode *vs.*Portrait Mode

Landscape and portrait mode videos, often shot in different ways and purposes, display unique content and biases. This affects subjects' action patterns and overall visual dynamics. Therefore, models trained on one mode may struggle in the other. This section examines how models adapt across these different modes, focusing on their spatial information and cross-mode generalizability.

[†]https://ai.googleblog.com/2020/02/autoflip-open-source-framework-for.html

Table 7.2 : Cross mode evaluation with different models on Selected-100. Evaluation results performed on the PM subset correspond to the last column of Table 7.3. Views during inference are shown by the multiplication of # of spatial crops and # of temporal views. Rows highlighted perform best for the corresponding model.

Model	Train	Val.	Acc.	GFLOPs×views
X3D-M[49]	PM	PM	**52.0**	4.9×3×10
		LM	41.2	4.9×3×10
	LM	PM	**44.5**	4.9×3×10
		LM	43.5	4.9×3×10
Uniformer-S[101]	PM	PM	**42.0**	41.8×1×4
		LM	36.2	41.8×1×4
	LM	PM	40.1	41.8×1×4
		LM	**40.8**	41.8×1×4
MViTv2-S[106]	PM	PM	**41.0**	64.0×1×5
		LM	35.7	64.0×1×5
	LM	PM	33.7	64.0×1×5
		LM	**36.3**	64.0×1×5

7.3.1 Cross Mode Evaluation

To show the impact of the different domain priors of landscape and portrait mode videos on video recognition tasks, comparisons need to be made between the same video content shot in portrait mode and landscape mode. Ideally, for each action or event, we should shoot it with both portrait mode and landscape mode cameras. However, such a process is time-consuming and hard to achieve. Therefore, we opt for sampling original portrait mode videos and landscape mode videos with the same distribution and taxonomy from Kinetics-700 [21], as detailed in Section 7.2.3.

To explore the impact of the different priors to video recognition models, we conducted extensive experiments using different subsets of S100 (S100-PM and S100-LM). We trained various models on different subsets and evaluated their performance on landscape mode videos and portrait mode videos, by randomly selecting 25% videos as the validation set for each subset. For example, evaluated on S100-PM, models trained with S100-PM and S100-LM respectively can be fairly compared to see which video type is more effective to train models for videos in portrait mode. We conduct the experiments on three models, i.e. a CNN model X3D [49], a hybrid transformer model Uniformer [101], and a pure transformer model MViTv2 [106] to show the impact of video formats on different model architectures. During training and testing, we resize frames based on the shorter side while preserving aspect ratios and crop them into 224×224 pixel squares for input. We train all models from scratch without pretraining to avoid the impact of pretraining dataset. Popular pretraining datasets such as ImageNet [89] are biased towards landscape images which may add additional bias to our analysis.

We summarize all results as in Table 7.2. By comparing results in each row, we find that models trained on PM videos has a larger performance gap on the PM and LM testsets than models trained with LM videos. Moreover, models trained on PM data usually have better performance on PM testset compared to the models trained with LM videos. For example, evaluated on S100-PM, X3D trained with PM videos outperforms the model trained with LM videos by a large margin of 8% (51.2% vs. 44.5%). When evaluated on S100-LM, X3D achieves relatively comparable performance either trained with PM videos or LM videos (41.2% vs. 43.5%). This indicates that training videos in portrait mode are necessary to achieve satisfying performance on portrait mode videos.

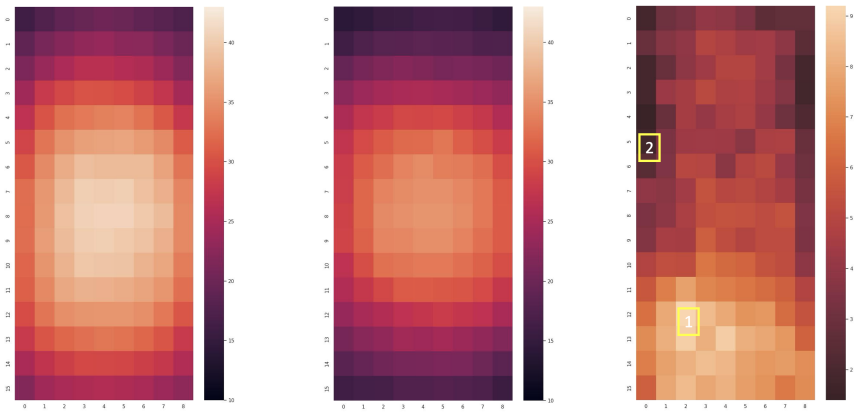


Figure 7.3 : The heatmaps of evaluating the Probing-P (a) and Probing-L (b) at different spatial locations on the validation set of S100-PM. (c) shows the accuracy differences between Probing-P and Probing-L.

7.3.2 Spatial priors

To investigate the different spatial data priors of portrait mode videos and landscape mode videos, we extensively evaluate the models trained on S100-PM and S100-LM on different frame positions to show the importance of frame features at different locations.

Specifically, we first train Uniformer-S [101] with 112×112 crops and shorter-side resized (set to a random value between 256 and 320) frames on either S100-PM or S100-LM. We name the resulted two models Probing-P and Probing-L. Then we evaluate the models with crops of 112×112 on different locations in a sliding window at the shorter-side resized video clips. The sliding strides vary for portrait mode and landscape mode videos in both height and width. For portrait mode videos, the stride in height is set to 1/16 of the frame height and the stride in width is set to 1/9 of the frame width. Sliding strides of landscape mode videos are adjusted vice versa.

Using Probing-P and Probing-L, we compose an accuracy map of size 16×9 from the accuracies obtained from the different evaluation positions on the S100-PM validation set as shown in Figure 7.3 (a) and Figure 7.3 (b). We further compute the difference between the two heat maps in Figure 7.3 (a) and (b) and obtain the difference map as in Figure 7.3 (c). Here, the difference value in each position indicates the gap of recognition abilities of the same model trained on landscape mode videos and portrait mode videos, respectively. If a value on the different map is greater than 0, it indicates that Probing-P achieves higher accuracy than Probing-L. For example, as outlined by the yellow boxes in Figure 7.3 (c), mark 1 indicates the model trained with PM videos is stronger to recognize the video categories at this location, while mark 2 indicates models trained by PM and LM videos have similar performance at this location. In general, it can be inferred from the brighter areas in Figure 7.3 (a) that informative areas in PM videos are more densely concentrated at the middle to lower half of the video. It can also be inferred from Figure 7.3 (c) that the bottom part of the PM videos contains specific domain knowledge that does not exist in the LM videos, leading to bad performance of models trained on LM videos in this region.

Similarly, we show the accuracy heat maps of the Probing-L and Probing-P evaluated on the LM videos in Figure 7.4 (a) (b), with the difference of the two heat maps shown in Figure 7.4 (c). It can be seen that the informative areas in LM videos are in the center part of the video, and the left and right sides on the video frame contain specific domain knowledge that cannot be learned from PM videos. For example, some actions with a wide background in LM videos may not have similar visual cues in the PM videos.

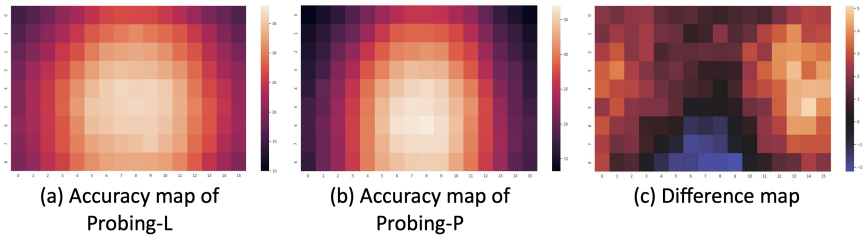


Figure 7.4 : The heatmaps of evaluating the Probing-L (a) and Probing-P (b) at different spatial locations on the validation set of S100-LM. (c) shows the accuracy difference between Probing-L and Probing-P.

7.4 Comparison of data preprocessing recipes

Effective data preprocessing is essential for achieving high performance in video classification tasks. In this section, we investigate the impact of different data preprocessing strategies on the performance of portrait mode video recognition. We hypothesize that videos in different aspect ratios may require different crop resolutions for optimal performance. To test this hypothesis, we perform extensive experiments on various portrait mode video datasets, using different crop resolutions and data augmentation techniques. Through our experiments, we identify the best recipes for portrait mode videos when using CNN or transformer models, which are different from that of landscape mode videos.

7.4.1 Resizing and area sampling

Resizing and cropping are critical steps in the data preprocessing pipeline for video recognition, as they allow videos to be processed efficiently and are also important ways of data augmentation. Different models in various architectures adopt different strategies. The two popular strategies are the Inception-style method [161, 50, 47, 101, 168, 121], and the shorter-side resizing method [154]. In this subsection, we will explore these two methods in more detail and investigate their

effectiveness for portrait mode video recognition.

The shorter-side resizing method is widely used in video recognition methods [22, 190, 169, 49, 181, 13, 127, 23, 236, 195, 160, 105, 182, 191, 201, 210]. It involves resizing the video frames so that the shorter side of the frame is set to a length that is fixed [23, 201] or randomly sampled within a range [22, 190, 169, 49, 181, 13, 127, 236, 195, 160, 105, 182, 191, 210], while the longer side is scaled proportionally. Then the frames are centre-cropped to a square shape, typically 224×224 and passed into the model. This approach ensures that the input frames have a consistent aspect ratio and are cropped without distortion. In contrast, the Inception-style method augments the shorter-side resizing method with two additional random sampling steps. The first one is to sample a target pixel number from the whole-size video frame by the random ratio between 8% and 100%. Then, it randomly samples an aspect ratio between 3/4 and 4/3 and reshapes the crop area accordingly. Finally, it crops the frames at a random position and resizes them to a fixed resolution in squares (*e.g.*, 224×224) without keeping the aspect ratio. This approach can sample a diverse set of inputs and is designed to adapt the model to videos in different sizes.

We carry out extensive experiments on models of different architectures with the two resizing strategies in Table 7.3. To alleviate the bias introduced by mixed-orientation data, the models are trained from scratch and we keep any other training setup identical to their original papers, except for learning hyper-parameters, such as batch size and learning rate. During inference, identical augmentation and sampling methods are adopted for different recipes. We guide the readers to supplemental materials for more details.

As shown in Table 7.3, each model is evaluated on three different portrait mode video benchmarks. For the CNN-based model, *i.e.*, X3D-M [49], the random scaling strategy from the Inception-style method brings an improvement of 2.2% (54.2% vs.

Table 7.3 : Comparison of top-1 accuracy (%) of different resizing and area sampling strategies for portrait mode videos, *i.e.*, inception style (Incep.) and shorter-side style(Short.). Views during inference are shown by the multiplication of # of spatial crops and # of temporal views.

Model	Data	Incep.	Short.	GFLOPs×views
X3D-M[49]	S100-PM	**54.2**	52.0	4.9×3×10
	3Massiv	**53.7**	52.6	4.9×3×10
	PM-400	**61.7**	61.2	4.9×3×10
Uniformer-S[101]	S100-PM	39.7	**42.0**	41.8×1×4
	3Massiv	42.8	**43.6**	41.8×1×4
	PM-400	50.2	**50.4**	64.0×1×5
MViTv2-S[106]	S100-PM	36.9	**41.0**	64.0×1×5
	3Massiv	50.4	**52.1**	64.0×1×5
	PM-400	61.7	**62.0**	64.0×1×5

52.0%) on S100-PM [21], 1.1% (53.7% vs. 52.6%) on 3Massiv [63] and 0.5% on PM-400. Differently, as for the transformer-based models, *i.e.*, Uniformer-S [101] and MViTv2-S [106], randomly scaled input crops bring down the accuracy by a large margin. For example, the random scaling reduces the performance of Uniformer-S by 2.3% (42.0% vs. 39.7%) on S100-PM, 0.8% (43.6% vs. 42.8%) on 3Massiv and 1.3% (72.1% vs. 70.8%) on PM-400. MViTv2-S also shows performance drops from 0.3% to 4.1% across benchmarks. This suggests that optimal strategies diverge from those used in mixed orientation benchmarks like Kinetics[81].

It may be hard to determine the cause of the interesting phenomenon, but we can make a reasonable assumption that it is due to the different data priors in portrait mode only video benchmarks, such as S100-PM and PM-400. With portrait mode

Table 7.4 : Top-1 accuracy (%) of different training crop resolutions. The models are always tested with the same square crops in **224×224** to ensure the same inference cost across different training crop resolutions.

Model	Data	Training crops		
		224×224	256×192	288×192
X3D-M[49]	S100-PM	**52.0**	51.6	50.8
	3Massiv	**52.6**	52.5	50.8
	PM-400	**61.2**	61.0	60.8
Uniformer-S[101]	S100-PM	42.0	43.3	**45.4**
	3Massiv	43.6	44.6	**45.8**
	PM-400	50.4	50.8	**51.6**
MViTv2-S[106]	S100-PM	41.0	40.0	**45.5**
	3Massiv	52.1	52.3	**53.8**
	PM-400	62.0	61.4	**62.8**

videos, the object and its movement are typically limited to a vertical space, which may result in unique visual patterns that are not present in hybrid orientation benchmarks, such as Kinetics. While the cause requires further investigation, these results suggest that there may be unique characteristics of portrait mode videos that require specialized recognition methods.

7.4.2 Shape of frame crop

In this subsection, we explore the impact of different crop strategies on model performance in portrait mode video recognition. Specifically, we investigate the performance of models trained and tested on crops of varying sizes and aspect ratios.

Table 7.5 : Top-1 accuracy (%) of using the same resolution for both training and testing crops. We also report the performance difference compared with using 224×224 testing crops from the first column of Table 7.4, where ↑ means higher result.

Model	Data	Testing crops	
		256×192	288×192
X3D-M[49]	S100-PM	$51.4_{0.2\downarrow}$	$50.4_{0.4\downarrow}$
	3Massiv	$52.6_{0.1\uparrow}$	$52.0_{1.2\uparrow}$
	PM-400	$62.9_{1.9\uparrow}$	$63.1_{2.3\uparrow}$
Uniformer-S[101]	S100-PM	$44.4_{1.1\uparrow}$	$46.5_{1.1\uparrow}$
	3Massiv	$45.7_{1.1\uparrow}$	$47.3_{1.5\uparrow}$
	PM-400	$51.9_{1.1\uparrow}$	$53.3_{1.7\uparrow}$
MViTv2-S[106]	S100-PM	$39.8_{0.2\downarrow}$	$46.8_{1.30\uparrow}$
	3Massiv	$52.7_{0.4\uparrow}$	$54.8_{1.0\uparrow}$
	PM-400	$62.1_{0.7\uparrow}$	$63.7_{0.9\uparrow}$

Traditional methods typically use square frame crops to ensure even coverage of object and movement in both vertical and horizontal directions. However, we argue that this approach may not be optimal for portrait mode videos, which typically contain object and movement information in vertical directions. Cropping the frames into squares could potentially result in a loss of critical information and more background noise. As shown in Figure 7.3, portrait mode videos possess more informative content distributed vertically, and cropping into squares may not effectively capture this information.

To comply with the unique information distributive characteristics, we propose to crop the areas in vertical rectangles and input them directly into models without

Table 7.6 : **Temporal information importance**: Effect of utilizing temporal information for video recognition on different benchmarks.

Data	Model	# of Frames	Top1-Acc.
K400 [81]	Uniformer-frames	16×4	72.1
	Uniformer [101]	16×4	$76.6_{4.5\uparrow}$
3Massiv [63]	Uniformer-frames	16×4	41.9
	Uniformer [101]	16×4	$42.8_{0.9\uparrow}$
PM-400	Uniformer-frames	16×4	45.7
	Uniformer [101]	16×4	$50.3_{4.6\uparrow}$

Table 7.7 : **Audio importance**: Comparison of different modalities with offline feature embeddings.

Data	Modality	Top1-Acc.
3Massiv [63]	Visual	52.7
	Audio	31.6
	Visual+Audio	54.9
PM-400	Visual	54.6
	Audio	15.2
	Visual+Audio	57.0

distortion. We experiment with crops in different aspect ratios and in similar pixel numbers to the square input, *i.e.*, 256×192 and 288×192, in order to fairly compare the models under different input resolutions. With input shape changed, we only modify the last global pooling layer. We keep any other training details identical to the setup using square inputs.

As shown in Table 7.4, we train models with different input crops on portrait mode video benchmarks and test with square crops, *i.e.*, 224×224 to ensure identical

inference cost. It is thrilled to see that increase in aspect ratio introduces continuing performance improvement for transformer-based models, *i.e.*, Uniformer-S and MViTv2-S. We also observe that change in aspect ratio degrades the performance of X3D-M, showing different behaviour to transformer-based models. The potential reason could be due to the fixed square receptive field of convolution networks regardless of the input resolutions, which is not compatible with the elongated image shape.

In order to further validate the benefits of rectangular input, we evaluate the performance of X3D-M [49], Uniformer-S [101] and MViTv2-S [106] on non-square training resolutions and tested them on three portrait mode video benchmarks. We find that the three models achieve higher accuracies on 3Massive and PM-400 with both crops in 256×192 and 288×192. On S100-PM, Uniformer-S and MViTv2-S achieve better testing results with 288 × 192 resolution, with FLOPs increased by around 15% (47.5G vs 41.8 for Uniformer-S; 72.7G vs. 64.5G for MViTv2-S). Note that FLOPs of 256×192 are smaller than square 224×224 (single clip inference cost: 40.6G vs. 41.8G for Uniformer-S; 62.9G vs. 64.5G for MViTv2-S). The performance boost further supports the potential benefits of rectangular input for video recognition in portrait mode.

7.5 The importance of temporal information

In this subsection, we investigate the importance of utilizing temporal information for portrait mode video recognition. We show that the PortraitMode-400 is a valuable resource for evaluating video models in the challenging setting of portrait mode video recognition.

We design two baselines with different temporal utilization approaches and extensively evaluate the models trained on Kinetics-400 [81], 3Massiv [63] and our PortraitMode-400. Specifically, we build our baselines with Uniformer-S and train

the models with 224×224 crops. Uniformer-frames is constructed with image-based Uniformer-S and temporal aggregation of predicted logits using mean pool. It serves as a naive baseline since the temporal information is incorporated simply by merging the predicted logits across frames. For more advanced temporal correspondance, we train a video-based Uniformer-S endowed with self-attention on temporal dimension, building and learning temporal relations in different levels.

As shown in Table 7.6, by leveraging temporal self-attention, Uniformer-S obtain accuracy improvement by 4.5% and 4.6% on Kinetics-400 and PortraitMode-400 respectively. Interestingly, the 3Massiv dataset, most of which videos are in portrait mode, does not show as large of a performance gain from using temporal information as our PM-400. In contrast, our PortraitMode-400 dataset shows a significant performance gain from using temporal information, attributable to its diverse collection of videos rich in intricate temporal dynamics.

7.6 The importance of the audio modality

In this section, we aim to explore the significance of audio information in portrait mode video recognition. To achieve this, we adopt the R3D-50 [66] backbone trained on Kinetics-700 [21] for spatio-temporal modeling and the VGG [67] model trained for sound classification [56] for audio modeling, following the practice in 3Massiv [63]. We freeze the audio-visual backbones and train the classifier and multimodal fusion layers.

Our findings, as presented in Table 7.7, reveal that the model trained with audio consistently outperforms the model trained without audio on both the PM-400 and 3Massiv by approximately 2.4 points. This indicates that audio information plays a crucial role in portrait mode video recognition. Incorporating audio information can significantly enhance the performance of the model. We argue that audio cues can provide additional information about the subject's actions, emotions, and the

surrounding environment, which poses unique challenges for video recognition in portrait mode.

7.7 Discussions

In this work, we advocate conducting research on portrait mode videos. To this end, we introduce the PortraitMode-400 dataset dedicated for portrait mode video recognition with a fine-grained taxonomy. We also make initial attempts to explore the specific properties of portrait mode videos, including their spatial bias, and the optimal training and evaluation protocols, with effects of the temporal information and audio modality. We believe our dataset can serve as a testbed to facilitate further research such as novel architecture designs and multi-modality modeling on portrait mode videos.

Chapter 8

Shot2Story20K: A New Benchmark for Comprehensive Understanding of Multi-shot Videos

This chapter introduces Shot2Story20K, a new benchmark for the comprehensive understanding of multi-shot videos. Recognizing the complexity of videos that contain multiple, distinct shots, this chapter addresses the challenge of generating cohesive narrative summaries that accurately link these separate events. The Shot2Story20K dataset facilitates this by providing detailed annotations for both visual content and audio narratives across sequential shots. Here, we explore methodologies that leverage large language models to synthesize these multimodal inputs into coherent summaries, pushing the boundaries of video captioning and storytelling in dynamic and complex video environments.

8.1 Introduction

Video captioning is a long-standing video understanding task to facilitate open-world video analysis with the help of human-annotated captions. Since a video may contains multiple events, dense captioning benchmarks (Ego4D [59], YouCook2 [229], ActivityNetCaps [88]) are tailored to capture the information of multiple events in a video ranging from 3-20 minutes. However, even within seconds, we find that there are already more than one single event in a lot of daily videos such as news broadcast, tutorial videos, and movies. Specifically, shot transition, which is a common technique to transfer from one event to another, or to switch the viewpoint of a single event, happens less than every 4s for average English movies after 2010 [32].

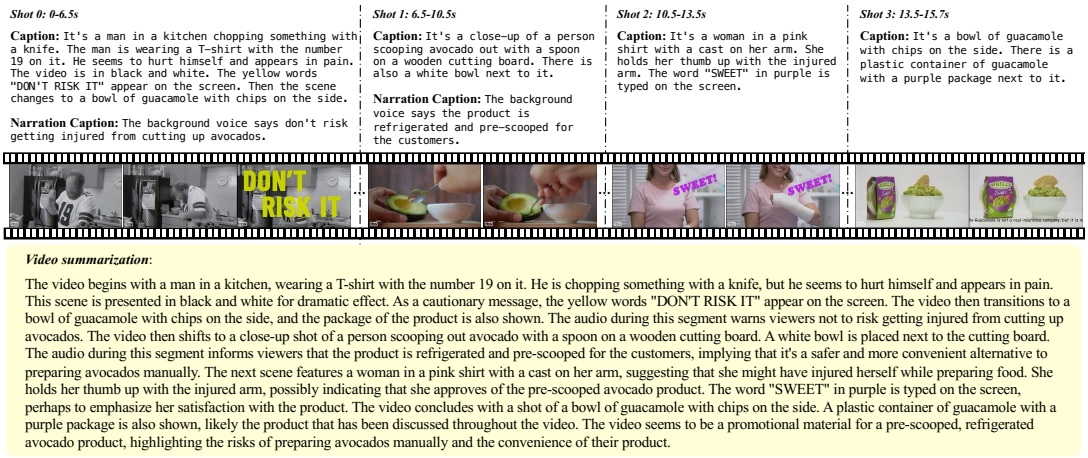


Figure 8.1 : An annotated example of our PortraitMode-400 with sing-shot visual captions and narration captions. Moreover, we provide coherent and reasonable video summaries to facilitate comprehensive understanding of multi-shot videos.

Although some existing captioning benchmarks [204, 88, 229] already use multi-shot videos, they often annotate the captions in a coarse-grained manner, either providing a holistic caption or asking annotators to subjectively choose the boundary of each event. To better accommodate the multi-shot formation of videos, we believe a new video benchmark with rich textual descriptions based on video shots is favored in the research community.

On the other hand, multi-shot videos are often accompanied by rich narrations that relates to the different events happening in the video. A model needs to capture both the visual and audio signals to understand the underlying story. Specifically, narrations may contain key information that cannot be inferred from pure visual information only. See Figure 8.1, without the narration, a viewer is unable to capture the relationship between the man's action and the avocado product in the first shot.

In this work, we propose a new benchmark Shot2Story for audio-visual under-

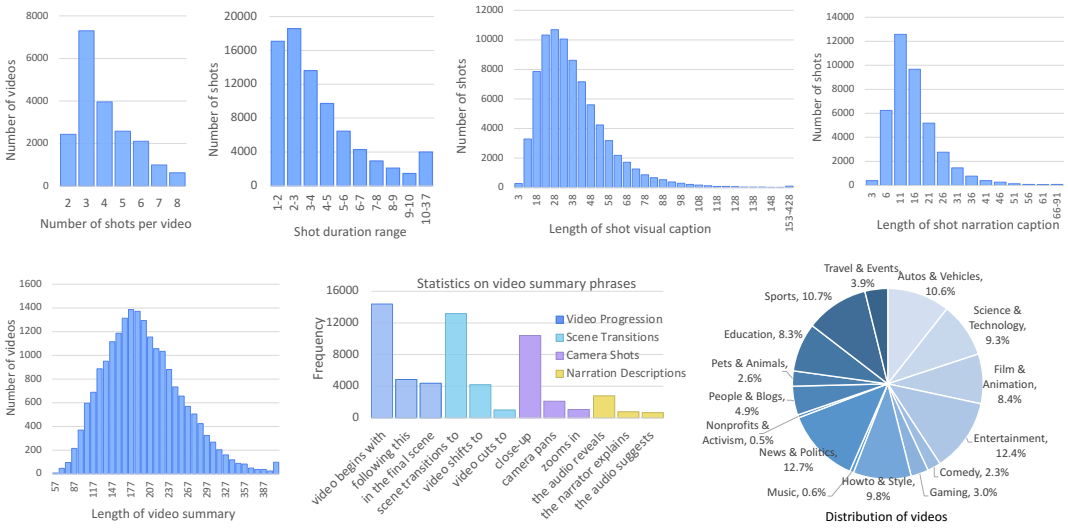


Figure 8.2 : Statistics of Shot2Story. Our dataset comprises videos with 2 to 8 shots each. Most shots range from 1 to 5 seconds, accompanied by detailed visual captions and narration captions. It features extensive summaries, highlighting video progressions, transitions, camera cuts and narration descriptions, with statistics of frequent expressions depicted in the figure.

standing of multi-shot videos. We collected a dataset of 20,023 short videos where the average number of shots in each video is 4.0. For each video shot, we annotate a detailed textual description for the video frames and another textual description for the human speech. We also leverage a state-of-the-art large language model (LLM) GPT-4 [128] to generate a long textual video summary from the annotated clip descriptions, which are further verified by human annotators. The summary includes additional details such as transitions of different shots, progression of multiple events, and mapping of the subject identities in different scenes. An overview of our dataset can be seen in Figure 8.2.

To benchmark the advances of multi-modal video understanding, we designed several distinctive tasks using our dataset, including single-shot captioning, multi-shot summarization, and video retrieval with shot description. We design and im-

plemented several baseline models using a frozen vision encoder and an LLM, by prompting the LLM with frame tokens and ASR (Automatic Speech Recognition) text. Through extensive experiments, we show that: (1) the ASR text is critical to understand the complex multi-shot scenario, (2) processing the video as a whole without the shot-structure degenerates the model's capacity of understanding the multi-shot video, (3) the summarization model trained on our benchmark can be generalized to other datasets with longer durations (ActivityNet) and out-of-domain topics (MSRVTT). Without any bells and whistles, we attains competitive results on zero-shot video question-answering by converting the problem into pure text-based QA with the generated video summaries.

8.2 The Shot2Story benchmark

8.2.1 Overview

Our new benchmark Shot2Story contains 20,023 videos. The length of each each video is ranging from 10s to 40s. For each video, we first use a off-the-shelf shot detection method TransNetV2 [159] to split it into shots. For each video shot, we annotate captions for both visual and audio information. Then we further annotate video summaries based on the annotated shot captions. Figure 8.2 shows an overview of our dataset with some key statistics. An example of one annotated video is shown in Figure 8.1.

8.2.2 Data preparation

We source videos for our dataset from the public video benchmark HDvila100M [206]. It offers a large collection of narrative videos, comprising 3M YouTube videos segmented into 100M clips, each about 13 seconds long. We choose this data source for its concise yet complex multi-shot formats, diverse topics, and abundant ASR content. Since we prefer videos with both rich visual and ASR information, we de-

Table 8.1 : High-level comparison of our dataset to previous ones. The summary length of ActivityNet and YouCook2 are their combined length of captions in one video. M and G stands for manual and generated, respectively.

Dataset	Annotation	Multi-shot Video	Multi-event Descriptions	Audio Captions	Detailed Summary	Summary Length	#Videos	Avg. Duration
MSRVTT [204]	M	✓	✗	✗	✗	-	10K	15s
ActivityNet Caps [88]	M	✓	✓	✗	✗	52.4	20K	3min
VideoStorytelling [98]	M	✓	✓	✗	✓	162.6	105	12.5min
Ego4D [59]	M	✗	✓	✗	✗	-	10K	23min
YouCook2 [229]	M	✓	✓	✗	✗	67.8	2K	6min
VAST [26]	G	✓	✗	✓	✓	32.4	27M	5∼30s
Shot2Story	M+G	✓	✓	✓	✓	201.8	20K	16.7s

sign several filtering techniques to exclude those videos with either low visual-ASR correlation or static visual content.

We start with keeping video clips with durations between 10 to 40 seconds, since we observe that the majority of the video clips from HDvila100M fall in this range. Then we remove videos with more than 8 shots due to the heavy annotation cost. We also notice that the video segments with too many shots in HDvila100M tend to be slideshows or image collages that deviates from our focuses. Further, to harvest videos with rich visual-ASR correlations, we set up a metric between video shots and ASR texts. Specifically, we uniformly sample 4 video frames for each shot and obtain the cosine similarity score between the video shot embedding and the text embedding using CLIP [138] encoders. We only keep the videos containing at least one shot that is visually correlated to ASR with a threshold of 0.25. In the next step, in order to obtain videos with diverse shot contents, we set up an inter-shot metric to filter out the videos with similar adjacent shots. We compute the cosine similarities between embeddings of adjacent shots and keep the videos with all inter-shot similarity scores smaller than 0.9. Finally, to further remove the videos with static contents, we adopt an intensity-based scene changes filter in PySceneDetect* with a threshold 11 in our segmented shots. If the filter is unable to detect new segments at a low threshold, it is conceivable that the shot contains static contents. We only keep the video clips in which all shots contain no static content based on our filtering method.

As a result, from a total of 1.1M sampled video clips from HDvila100M, we obtain $20,023$ video clips that meet our quality standard. The number of shots in each video is from 2 to 8. These videos are then shared with our annotators for further annotations.

*https://www.scenedetect.com/

8.2.3 Annotation of single-shot captions

After using TransNetV2 to divide the target videos into video shots, we ask annotators to annotate both visual-only captions and audio-related captions for each shot. We split the annotation of these two captions to facilitate separate modeling of these two types of information source.

For visual-only caption, we require annotators to describe the major subjects and events in the video. Since it is an open-world setting, the videos can be quite diverse and hard to describe. In order to reduce the difficulties of annotating a caption from scratch, we generate an initial video caption using MiniGPT-4 [232] by sampling 4 image frames from the video clip and prompting the model using the prompt below.

MiniGPT-4 prompt: ###Human:<Img>Frame1</Img ><Img>Frame2</Img> <Img>Frame3</Img><Img>Frame4</Img>Please describe this video. Do not include details that you are not sure of. For example, if there is text in the image, do not include the content of the text if they are not clearly shown. ###Assistant:

Although MiniGPT-4 is originally designed for image understanding, empirically it is able to generate captions for videos, both comprehensively and reasonably. It is able to describe different subjects including person, animals, food, tools, and virtual objects like animated characters. We ask annotators to correct any mistakes they find in the generated captions, and to add missing details to the captions. The mistakes include incorrect description of the object categories, attributes, actions, facial expressions etc. Also, there might be some subjective description generated by MiniGPT-4 such as emotion and atmosphere. We ask annotators to remove all these subjective descriptions. For example, the annotator corrects the caption from "standing in front of the car" to "getting close to the car", and adding a missing detail of "a close-up shot of the front". In this way, we find the annotation speed

significantly faster ($\sim 3\times$) compared to writing a caption from scratch. On the other hand, we find the captions generated this way has more coherent style and tend to cover more details of the video.

In contrast to the traditional video captioning benchmarks [204, 88, 229], we also annotate narration captions in addition to the visual-only captions. Different from existing audio captioning benchmarks [55], we focus more on human speeches rather than acoustic events. The annotators are required to associate the human speech with the video content and summarize the main idea of the speech. We require annotators to describe the source of the speech using visual information. For example, if someone is talking, the annotator needs to describe which person in the video is talking. If the human speech refers to some object in the video, the annotator is required to describe which object in the video the speaker is referring to. Note that the speaker identity and reference of visual objects are critical information for understanding a video that cannot be trivially obtained using existing algorithms. There are existing research on speaker identification [84] and visual grounding [5, 228], but they only work well on constraint scenarios. Given this annotation process, our narration captioning task requires a joint understanding of visual and audio signals.

8.2.4 Annotation of video summary

To create video summaries with the annotated video-shot captions, we leverage an LLM-based approach. Specifically, we form a text prompt with incorporating all shot captions and ASR text included, and uses GPT-4 [128] to generate a cohesive summary. The quality is assured through further review and correction by our annotators.

We prompt GPT-4 to produce coherent, fluent text summaries with transition expressions such as “the video begins”, “following this”, and “in the final scene”

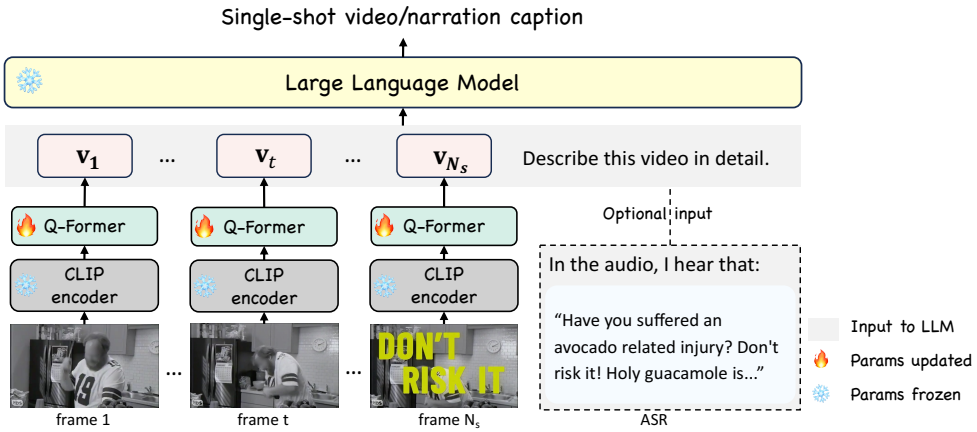


Figure 8.3 : Model structure for video-shot captioning. Visual tokens from the CLIP [138] visual backbone and Q-Former [97, 232], along with text prompts, form the input to the LLM [30]. ASR input is optional for single-shot video captioning.

to connect video-shot descriptions. The generated annotations also encompass a higher-level understanding of shots, using key phrases such as "scene shifts back to" and "returns to the scene" to denote recurring scenes across different shots. Notably, GPT-4 often identifies and links the same subjects across scenes without relying on explicit re-identification models. It draws on descriptive and attributive text from our shot captions like "a newsroom" or "a man wearing a black suit" to infer scene or subject identity. To ensure quality, annotators carefully review and correct any inconsistencies in scene or subject references within these summaries.

8.2.5 Comparison to existing benchmarks

Compared to existing video description datasets, our dataset is more challenging due to the explicit modeling of the multi-shot nature of web videos. Our textual description includes both shot-level captions and video-level summaries, combining visual and audio understanding, which provides a unique test bed for multi-modal video understanding. Table 8.1 shows a high-level comparison of our new dataset with existing video captioning benchmarks. Most existing video captioning bench-

marks, such as MSRVTT [204], YouCook2 [229] and ActivityNet Caps [88], also use multi-shot videos as annotation source, but they either annotate a holistic caption for the video (MSRVTT) or ask annotators to decide the boundary of different events. In our study, we observe that video shots naturally create a sequence of related events, leading us to annotate distinct captions for each shot. Ego4D [59] only annotates dense visual captions but not audio captions for relatively long egocentric videos. Video Storytelling [98] is a small-scale dataset with annotations of multiple events in a videos and provides a summary of the video by concatenating all captions.

A recent work VAST [26] feed generated video and audio captions into an LLM to generate video summary. However, their work processes a multi-shot video as a whole and lacks the granularity of the events in different shots. Additionally, VAST directly uses predicted captions without any human verification, which indicates their video summaries can be noisy and containing biases from the captioning models. Our dataset stands out from VAST with its more detailed visual and audio shot captions. These captions, averaging 35.3 words for visual and 17.8 words for audio, are the result of a thorough manual annotation process. Although our video summary is also generated using an LLM, it is further verified by annotators to make sure there is no hallucinated details from the LLM. Our dataset has an average length of 201.8 wprds for the video summary, which is much longer than existing benchmarks, and longer than the combined length of captions in one video in ActivityNet and YouCook2.

8.3 Tasks and Experiments

8.3.1 Basic settings

For all the tasks described in this section, we follow the same training/validation/test split. Specifically, the number of videos for training, validation, and test

Table 8.2 : Performance of single-shot video captioning task. V and A means Visual and ASR.

Modalities	B	M	R	C
V	10.5	16.0	30.1	38.8
V+A	10.7	16.2	29.6	37.4

set are 14016, 1982 and 4025, respectively. We resize the frames to 224×224. We employ ViT-G/14 from EVA-CLIP [48] and Q-Former from BLIP-2 [97] as visual encoder, and Vicuna v0-7B [30] as the language model. We load pretrained Q-Former from MiniGPT-4 [232]. In training, we update only Q-Former parameters, keeping the ViT and LLM frozen. We adopt AdamW [122] as our optimizer and use a learning rate of 8e-5. We train the models for 40 epochs with a batch size of 128 for single-shot video captioning and narration captioning. We finetune our video summarization models on the single-shot captioning model with a batch size of 16.

8.3.2 Single-shot video captioning

To understand the visual content of each video shot, we introduce the single-shot video captioning task. Note that the task is to generate descriptions for individual video shots, while ASR information can be leveraged to improve the accuracy of the captions. For this task, we adapt the framework of MiniGPT-4 [232], with the model structure depicted in Figure 8.3. Specifically, we adopt the similar structure as we generate pseudo captions for data annotation in Sec. 8.2.3. First, we sample N_s frames from a video shot, and encode them using a fixed vision encoder, then feed the encoded features to a Q-Former to produce visual tokens. The visual tokens are appended into a text prompt and the LLM is asked to generate a caption for this video shot.

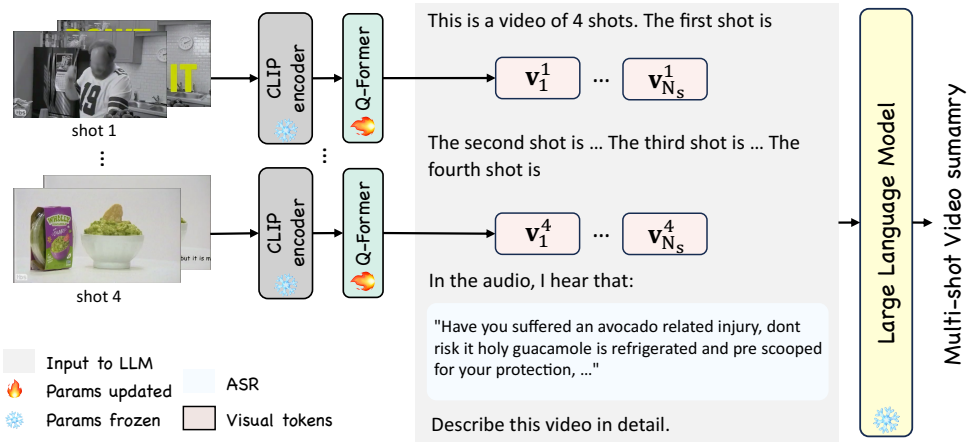


Figure 8.4 : Model structure for multi-shot video summarization model SUM-shot. We arranges visual tokens in a multi-shot format to encapsulate multi-shot information. Additionally, ASR text is incorporated for audio-visual video summarization.

We compare two model variants on this task. One is with the ASR text as additional context cues in the text prompt and the other is without the ASR information. We evaluate our models using BLEU@4 [130] (abbreviated to B), METEOR [37] (abbreviated to M), ROUGE [110] (abbreviated to R) and CIDEr [175] (abbreviated to C), and show the results in Table 8.2. It shows that inclusion of ASR-derived texts yields a modest enhancement in the B and M by 0.2. Conversely, it incurs a decrement of 0.5 and 1.4 in R and C, respectively. These results imply that ASR text complements visual data without introducing discrepancies, yet posing integration challenges for augmenting single-shot video captioning performance. Figure 8.5 (a) displays output examples of our model's single-shot video captioning. It accurately details visual elements within the shot, effectively capturing actions like "gesturing with her hands" and articulates secondary elements within a scene like "a doll on the couch".

Table 8.3 : Performance of single-shot narration captioning task. V and A means visual and ASR.

Method	Modalities	B	M	R	C
VALOR [25]	Audio	6.6	10.0	23.9	13.5
Ours	A	4.7	17.1	30.3	130.9
	V+A	18.8	24.8	39.0	168.7

8.3.3 Single-shot narration captioning

Human narration is another critical factor to understand a multi-shot video. It often provides information of the background knowledge and commentator's view on visual events. We conduct experiments to predict the narration caption of a video-shot and name this task single-shot narration captioning. We adopt the same model structure as single-shot video captioning with the ASR text as additional input, except that the prediction target is the narration caption. We compare with existing audio captioning model VALOR [25]. We finetune VALOR on our single-shot narration captions and show the results in Table 8.3. We also add another baseline model that only takes ASR text as input and predicts the narration captions using Vicuna [30].

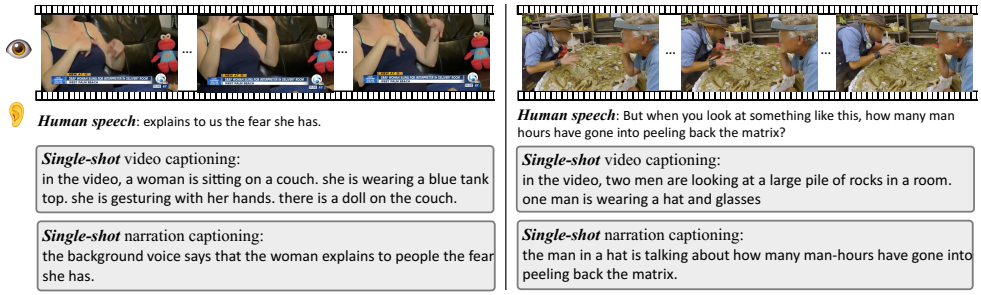
Since our narration captions contain descriptions about the related visual information as well, e.g. the subject, referred objects etc, using only ASR text does not produce satisfactory results. The baseline model VALOR is unable to capture the rich ASR text information with only the raw audio, leading to a weak performance of 13.5 in CIDEr. Our model combining visual and ASR text can generate reasonable narration captions on most cases. As shown in Figure 8.5 (a), our model identifies narration sources and aptly describes spoken content, as highlighted by phrases like

"the background voice says" and "the man in a hat is talking".

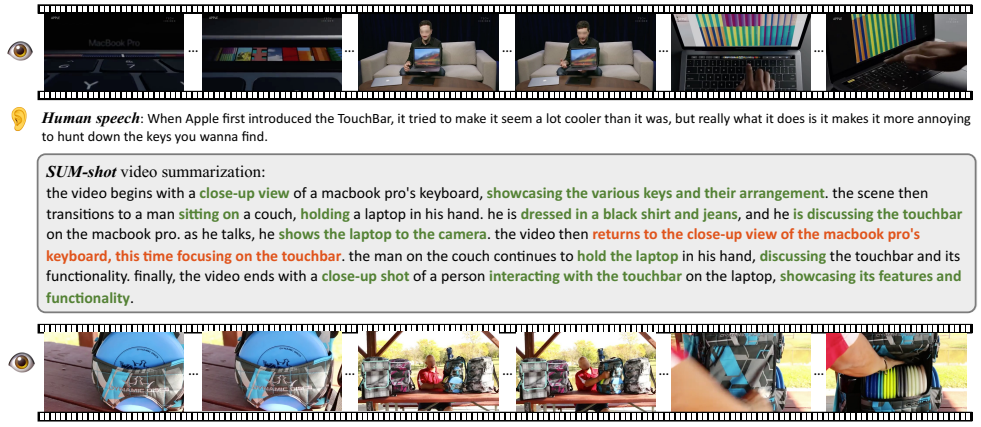
8.3.4 Multi-shot video summarization

Multi-shot video summarization is a new task that is distinct from existing video description tasks. It requires the model to understand the shot structure of the given video and to provide a coherently paragraph to describe the progression of events in the different shots. In this section, we experiment with three model variants. The first model SUM-text uses a two-stage approach, first generating captions using our video-shot captioning model for each video shot, then embed the generated captions into a text prompt as the input to the LLM (Vicuna-v0 [30]) to generate a video summary. The second model SUM-holistic uses similar model as Figure 8.3. We uniformly sample 16 frames from the full video clip and prompt the LLM with frame tokens and ASR text. The third model SUM-shot uses a more refined framework by sampling 4 frames in each video shot and prompting the LLM with frame tokens from different shots, as is shown in Figure 8.4. Compared to SUM-holistic and SUM-shot, SUM-text is not trained end-to-end and may loss critical information with the captioning step, for example, it cannot capture the correspondence of the same subject in two shots. SUM-holistic does not have the shot information explicitly and rely on the LLM to parse the video shots using the provided frame features. SUM-shot is given the shot structure as input, which makes it easier to generate descriptions based on the different shots. We compare with Video-ChatGPT [124] by instruction-tuning their model on our video summary data without the ASR input.

Table 8.4 shows the results of the three models. It is shown that SUM-text achieves the overall best performance, although it is a two-stage model with pre-generated shot captions. SUM-shot is slightly worse than SUM-text, indicating that better model design needs to be explored for end-to-end video summarization. SUM-



(a) Example predictions of single-shot video captioning and narration captioning.



(b) Example predictions of multi-shot video summarization.

Figure 8.5 : Example predictions of our models. (a) demonstrates our model's single-shot video captioning, producing precise descriptions and identifying narration speakers, e.g., *gesturing with hands*, *a man in a hat speaking*. (b) shows multi-shot video summarization, with accurate captions in green and errors in red, illustrating the model's ability to narrate event sequences and maintain subject consistency, as seen in the progression from *close-up of a backpack* to *transitions to a man* and *return to the backpack*.

Table 8.4 : Performance of models on video summarization. E2E means whether the model is trained in an end-to-end approach.

Model	E2E	ASR	B	M	R	C
Video-ChatGPT [124]	✓	✗	5.0	14.0	19.7	1.2
SUM-shot w/o ASR	✓	✗	9.8	18.4	24.9	4.7
SUM-text	✗	✓	12.2	20.4	27.1	9.2
SUM-holistic	✓	✓	10.9	18.3	26.2	6.3
SUM-shot	✓	✓	11.7	19.7	26.8	8.6

holistic is consistently worse than SUM-shot, showing the importance of the shot structure in predicting a video summary matching the transition of shots. SUM-shot w/o ASR underperforms compared to SUM-shot and SUM-holistic, highlighting ASR's significance in multi-shot understanding. Video-ChatGPT is not able to match the performance of our models, potentially due to their weakness in processing multiple scenes and lack of ASR input. Video-ChatGPT directly encodes the whole video into a sequence of tokens and may loss a lot of details in the frames, while ours directly feeding frames tokens into LLM without compressing them.

Figure 8.5 (b) illustrates our SUM-shot model's predictive capabilities. The model adeptly narrates event sequences with appropriate emphasis. For instance, in the MacBook-example, it not only details the keyboard but also rationalizes the display of various keys, aligning with the ASR data about the touchbar discussion, thus crafting a coherent summary. Nonetheless, some predictions, marked in red, are erroneous, such as the non-existent "returns to a close-up view of the macbook" shot. These inaccuracies likely stem from the LLM's tendency to "hallucinate" plausible yet non-factual details. Despite these errors, the model demonstrates a proficiency in generating consistent and nuanced summaries, highlighting both the potential of

Table 8.5 : Comparison of performance for text-to-video (T2V), text-to-shot (T2S), and video-to-text (V2T) retrieval tasks.

Method	Pretrain Datasets	T2V			T2S			V2T		
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
Alpro [96]	WebVid-2M+CC-3M	46.3	69.5	78.4	50.3	76.3	83.2	45.2	69.8	78.2
Clip4clip [123]	CLIP400M	47.2	70.4	77.6	52.4	78.2	85.4	48.9	70.2	78.1
UMT [100]	CLIP400M+UMT25M	66.3	81.8	85.8	68.6	88.4	92.0	64.9	82.3	86.2

our model and the challenges that our dataset presents for future research.

8.3.5 Video question-answering with summary

Since the generated summaries are long and complex, the traditional captioning metrics (B, M, R, C) may not reflect the true quality of the generated summaries. We thus adopt another video understanding task, zero-shot video question-answering (QA), to further evaluate the quality of our generated summaries. Existing work [62] directly uses image captions as input to an LLM to generate question response. However, no such work has been done for videos.

Specifically, we directly apply our video summarization model on video QA benchmarks MSRVTT-QA [203] and ActivityNet-QA [218] by splitting the testing videos into video shots and feeding them into the SUM-shot model. The generated summaries and the associated questions are then fed into a Vicuna model to derive the answers. Note there is no adaptation or finetuning conducted for the Vicuna model. Since the original answers in the QA benchmarks are very short and the generated responses from LLM tend to be a long sentence, we levarge the `gpt-3.5-turbo` model to generate a binary decision of whether the answer is correct, following Video-ChatGPT [124]. We compare our results with Video-ChatGPT [124], MovieChat [157] and VideoChat [99] as in Table 8.2. Note that

Video-ChatGPT and VideoChat both use large amount of instruction tuning data to learn to directly generating answers from visual features and the text prompt, while ours bypasses instruction tuning by distilling the video information into a video summary. Additionally, for a direct comparison, we evaluate Video-ChatGPT on question-answering in the same methodology as ours. As shown in Table 8.6, our model outperforms Video-ChatGPT by a large margin. Our model also follows the zero-shot QA settings since the model only uses Shot2Story as training data. Note that MSRVTT contains a large portion of videos with out-of-domain topics such as tv shows and food, while ActivityNet has much longer videos than our training videos. This validates the the robustness and transferability of our model across different topics and longer videos. This surprisingly good result indicates that a comprehensive and detailed video summary is a high quality abstraction of the video, facilitating a wide range of tasks including video QA and video-based conversation.

8.3.6 Video retrieval with shot description

Text-based video retrieval is another task to evaluate multi-modal video representations. Traditional video retrieval often utilizes highly condensed text descriptions with benchmarks such as MSRVTT [204], LSMDC [144], and VATEX [192]. Retrieval models can simplify the problem by leveraging key objects / actions in the video withoutunderstand more complex details such spatial-temporal information and user intent. We present a distinct setting for retrieval with only descriptions of one video shot. Specifically, we design three settings: (1) using a shot description as query source to query the corresponding video (T2V). (2) using a shot description as query source to query the specific shot (T2S). (3) using a video as source to query a randomly sampled shot description in this video (V2T).

We report results on the testing set of our benchmark with 4025 videos, in-

Table 8.6 : Performance on video question answering. IT means whether the model uses video-text instruction tuning data. All methods follow the zero-shot manner.

Model	IT	QA Input	MSRVTT QA	ActivityNet QA
VideoChat [99]	✓	V+T	45.0	26.5
Video-ChatGPT [124]	✓	V+T	49.3	35.2
MovieChat [157]	✓	V+T	49.7	**51.5**
Video-ChatGPT [124]	✓	T	53.7	37.4
SUM-shot+Vicuna	✗	T	**56.8**	47.4

cluding 15913 shots. We evaluate several baseline models including Alpro [96], CLIP4clip [123], and UMT [100] and show the results in Table 8.5. In the three retrieval tasks, Alpro underperforms relative to Clip4clip by approximately 2.0 points in R@1, while UMT outperforms Clip4clip significantly, with an R@1 improvement of 19.1 for T2V and 16.2 for T2S. The performance comparison confirms that refined video-language alignment is crucial for retrieval accuracy. While Alpro employs regional token alignment and CLIP4clip uses global video-text matching, UMT advances the field with its R@1 improvements, utilizing masked modeling and distilling a ViT [43] for more detailed alignment. In light of these findings, our Shot2Story, enriched with ASR information that closely aligns with visual elements, presents an opportunity to harness ASR as a natural linkage for improving video-text alignment, potentially guiding future enhancements in this domain. Additionally, a comparison between video and shot retrieval tasks reveals that T2V presents a greater challenge than T2S, aligning with our hypothesis that retrieving a full video using a shot caption necessitates a more detailed understanding of the video. It confirms the capacity of our dataset to facilitate detailed and complex video understanding tasks.

8.4 Conclusion

In this work, we present a large-scale video understanding benchmark with annotations based on video shots. We provide detailed textual descriptions for each shot as well as a comprehensive video summary for the whole video. With the rich and diverse descriptions, our benchmark serves as a playground for more powerful multi-modal video understanding models, ready to be extended for a range of other video understanding tasks, such as video question answering, visual grounding, and video-based conversation.

Chapter 9

Conclusion and Future Works

Concluding the exploration, this thesis has delved deep into the complexities of video content analysis, an area increasingly pivotal as video data becomes more ubiquitous and varied. This journey started with addressing foundational challenges in video object perception and extended into the nuanced demands of holistic video understanding, reflecting the dual need for precision in detection and depth in interpretation.

In the realm of video object perception, we significantly advanced the field with the Hierarchical Video Relation Network (HVR-Net) and Progressive Frame-Proposal Mining (PFPM), enhancing object detection and leveraging sparse annotations to tackle practical issues of scalability and efficiency. Additionally, the Hybrid Temporal-scale Multimodal Learning (HTML) framework refined the ability to integrate textual descriptions with video content for precise referring video object segmentation.

Transitioning to holistic video understanding, the introduction of the Shot2Story benchmark and Dual-AI framework marked substantial improvements in narrative synthesis and group activity recognition. Moreover, the development of methodologies for Portrait Mode Video (PMV) recognition adapted video analysis techniques to the emerging trends of social media content, showcasing the adaptability of video analysis in various formats. These contributions have significantly broadened our understanding of complex video interactions and the contextual interpretation of scenes.

Overall, this thesis contributes a suite of methodologies that significantly advance the state of the art in video content analysis. The strategies developed herein for both detecting intricate object details and unravelling complex video narratives offer robust pathways for future research. These innovations hold the promise of enhancing various applications, from automated surveillance systems to advanced multimedia content curation, driving forward the capabilities of computer vision systems in handling the ever-increasing complexity of video data environments.

Looking ahead, the convergence of large vision-language models and large language models presents an exciting opportunity to unify video object perception and holistic understanding into a single, powerful framework. Such integration could significantly enhance video analysis by leveraging the strengths of these models to interpret complex multimodal data seamlessly. Instances and their relationships are crucial for understanding video content in depth. Large models, with their extensive capabilities for contextualizing information, can improve how machines perceive and interpret narratives and scenes. Furthermore, specialized single-object perception models, essential for customized scenes, could also benefit from the advancements in large models, particularly in terms of robustness and adaptability. Exploring these integrations could lead to more sophisticated, efficient, and context-aware video analysis systems, pushing the boundaries of what is currently possible in video understanding technology.

In conclusion, my PhD research has rigorously explored video content analysis, specifically tackling challenges in video object perception and holistic video understanding through innovative frameworks. Looking forward, embracing large models tailored for enhanced human-machine interaction, developing more generalized scenarios, and creating robust data collection pipelines for customized scenes are pivotal. These steps will significantly advance our capability to analyze and interpret complex video environments. This thesis has laid a crucial foundation for

these endeavors, marking substantial progress in the field and charting a course for future innovations that will further refine the generalizability and effectiveness of video content analysis systems.

Bibliography

[1] Mohamed Rabie Amer, Peng Lei, and Sinisa Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. pages 572–585. Springer, 2014.

[2] Mohamed R Amer and Sinisa Todorovic. Sum product networks for activity recognition. 38(4):800–813, 2015.

[3] Mohamed R Amer, Sinisa Todorovic, Alan Fern, and Song-Chun Zhu. Monte carlo tree search for scheduling activity recognition. pages 1353–1360, 2013.

[4] Mohamed R Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song-Chun Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. pages 187–200. Springer, 2012.

[5] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017.

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021.

[7] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6816–6826, 2021.

[8] Anurag Arnab, Chen Sun, Arsha Nagrani, and Cordelia Schmid. Uncertainty-aware weakly supervised action detection from untrimmed videos. pages 751–

768. Springer, 2020.

[9] Aditya Arun, C. V. Jawahar, and M. Pawan Kumar. Dissimilarity coefficient based weakly supervised object detection. 2019.

[10] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. pages 7892–7901, 2019.

[11] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. pages 4315–4324, 2017.

[12] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. pages 331–346, 2018.

[13] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? 2021.

[14] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. pages 2846–2854, 2016.

[15] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *10th IEEE International Conference on Computer Vision (ICCV 2005), 17-20 October 2005, Beijing, China*, pages 1395–1402, 2005.

[16] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[17] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. *arXiv preprint arXiv:2111.14821*, 2021.

[18] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. *Advances in Neural Information Processing Systems*, 34:19594–19607, 2021.

[19] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. pages 213–229. Springer, 2020.

[21] João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *CoRR*, abs/1907.06987, 2019.

[22] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. pages 6299–6308, 2017.

[23] Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6165–6175, 2021.

[24] Kai Chen, Jiaqi Wang, Shuo Yang, Xingcheng Zhang, Yuanjun Xiong, Chen Change Loy, and Dahua Lin. Optimizing video object detection via a scale-time lattice. pages 7814–7823, 2018.

[25] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023.

[26] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *arXiv preprint arXiv:2305.18500*, 2023.

[27] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

[28] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. pages 10337–10346, 2020.

[29] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. 37(3):569–582, 2014.

[30] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[31] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *ICCV Workshops*, pages 1282–1289. IEEE, 2009.

[32] James E Cutting, Kaitlin L Brunick, Jordan E DeLong, Catalina Iricinschi, and Ayse Candan. Quicker, faster, darker: Changes in hollywood film over 75 years. *i-Perception*, 2(6):569–576, 2011.

[33] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[34] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Object guided external memory network

for video object detection. pages 6678–6687, 2019.

[35] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. 2019.

[36] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. pages 4772–4781, 2016.

[37] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

[38] Ali Diba, Vivek Sharma, Ali Mohammad Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. 2017.

[39] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021.

[40] Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4964–4973, 2022.

[41] Zihan Ding, Tianrui Hui, Shaofei Huang, Si Liu, Xuan Luo, Junshi Huang, and Xiaoming Wei. Progressive multimodal interaction network for referring video object segmentation. *The 3rd Large-scale Video Object Segmentation Challenge*, page 7, 2021.

[42] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Trans-

formers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[43] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[44] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. pages 6569–6578, 2019.

[45] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Saleh, Javen Shi, Ian Reid, and Hamid Rezatofighi. Joint learning of social groups, individuals action and sub-group activities in videos. pages 177–195. Springer, 2020.

[46] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227*, 2021.

[47] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227*, 2021.

[48] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

[49] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020.

[50] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slow-

fast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[51] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. pages 3038–3046, 2017.

[52] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020.

[53] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966, 2018.

[54] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. pages 839–848, 2020.

[55] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[56] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 776–780. IEEE Press, 2017.

[57] Ross Girshick. Fast r-cnn. pages 1440–1448, 2015.

[58] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature

hierarchies for accurate object detection and semantic segmentation. pages 580–587, 2014.

[59] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[60] Chaoxu Guo, Bin Fan, Jie Gu, Qian Zhang, Shiming Xiang, Veronique Prinet, and Chunhong Pan. Progressive sparse local attention for video object detection. pages 3909–3918, 2019.

[61] Guodong Guo and Alice Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014.

[62] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877, 2023.

[63] Vikram Gupta, Trisha Mittal, Puneet Mathur, Vaibhav Mishra, Mayank Maheshwari, Aniket Bera, Debdoot Mukherjee, and Dinesh Manocha. 3massiv: Multilingual, multimodal and multi-aspect dataset of social media short videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21064–21075, 2022.

[64] Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao. Mining inter-video proposal relations for video object detection. In *European conference on computer vision*, pages 431–446. Springer, 2020.

[65] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang.

Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465*, 2016.

[66] 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 (CVPR)*, pages 6546–6555, 2018.

[67] David Harwath, Galen Chuang, and James Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech, 2018.

[68] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. pages 2961–2969, 2017.

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016.

[70] Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. pages 980–989, 2020.

[71] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. pages 3588–3597, 2018.

[72] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016.

[73] Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. pages 721–736, 2018.

[74] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. pages 1971–1980, 2016.

[75] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Efficient video object co-localization with co-saliency activated tracklets. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3):744–755, 2018.

[76] Zhengkai Jiang, Peng Gao, Chaoxu Guo, Qian Zhang, Shiming Xiang, and Chunhong Pan. Video object detection with locally-weighted deformable neighbors. In *AAAI*, volume 33, pages 8529–8536, 2019.

[77] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. Object detection in videos with tubelet proposal networks. pages 727–735, 2017.

[78] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. 28(10):2896–2907, 2017.

[79] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. pages 350–365. Springer, 2016.

[80] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[81] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[82] 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.

[83] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision*,

pages 123–141. Springer, 2018.

[84] You Jin Kim, Hee-Soo Heo, Soyeon Choe, Soo-Whan Chung, Yoohwan Kwon, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Look who's talking: Active speaker detection in the wild. *Interspeech*, 2021.

[85] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[86] Satoshi Kosugi, Toshihiko Yamasaki, and Kiyoharu Aizawa. Object-aware instance labeling for weakly supervised object detection. pages 6064–6072, 2019.

[87] Satoshi Kosugi, Toshihiko Yamasaki, and Kiyoharu Aizawa. Object-aware instance labeling for weakly supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[88] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[89] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. pages 1097–1105, 2012.

[90] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011.

[91] Baisheng Lai and Xiaojin Gong. Saliency guided end-to-end learning for weakly supervised object detection. In *Proc. IJCAI*, 2017.

[92] Baisheng Lai and Xiaojin Gong. Saliency guided end-to-end learning forweakly supervised object detection. In *IJCAI*, pages 2053–2059. AAAI Press, 2017.

[93] Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. pages 1354–1361. IEEE, 2012.

[94] Tian Lan, Yang Wang, Weilong Yang, Stephen N Robinovitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. 34(8):1549–1562, 2011.

[95] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. pages 734–750, 2018.

[96] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *CVPR*, pages 4953–4963, 2022.

[97] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[98] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, 22(2):554–565, 2019.

[99] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding, 2023.

[100] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. *arXiv preprint arXiv:2303.16058*, 2023.

[101] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition, 2022.

[102] Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Xiaoyun Zhao, Cong Wang,

Xin Chen, Zhong Liu, Caineng Pan, Mengke Li, et al. Ffa-ir: Towards an explainable and reliable medical report generation benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[103] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. 2021.

[104] Xin Li and Mooi Choo Chuah. Sbgar: Semantics based group activity recognition. pages 2876–2885, 2017.

[105] Xianhang Li, Yali Wang, Zhipeng Zhou, and Yu Qiao. Smallbignet: Integrating core and contextual views for video classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1092–1101, 2020.

[106] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022.

[107] Chen Liang, Yu Wu, Yawei Luo, and Yi Yang. Clawcranenet: Leveraging object-level relation for text-based video segmentation. *arXiv preprint arXiv:2103.10702*, 2021.

[108] Chen Liang, Yu Wu, Tianfei Zhou, Wenguan Wang, Zongxin Yang, Yunchao Wei, and Yi Yang. Rethinking cross-modal interaction from a top-down perspective for referring video object segmentation. *arXiv preprint arXiv:2106.01061*, 2021.

[109] Chenhao Lin, Siwen Wang, Dongqi Xu, Yu Lu, and Wayne Zhang. Object

instance mining for weakly supervised object detection. volume 34, pages 11482–11489, 2020.

[110] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[111] 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.

[112] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. pages 2980–2988, 2017.

[113] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. pages 740–755. Springer, 2014.

[114] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[115] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[116] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. pages 21–37. Springer, 2016.

[117] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[118] Yuxuan Liu, Pengjie Wang, Ying Cao, Zijian Liang, and Rynson W. H. Lau. Weakly-supervised salient object detection with saliency bounding boxes.

IEEE Trans. Image Process., 30:4423–4435, 2021.

[119] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

[120] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. 2021.

[121] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.

[122] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[123] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

[124] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[125] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019.

[126] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[127] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video trans-

former network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3163–3172, 2021.

[128] OpenAI. Gpt-4.

[129] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[130] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[131] Mandela Patrick, Dylan Campbell, Yuki M Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, Jo Henriques, et al. Keeping your eye on the ball: Trajectory attention in video transformers. In *NeurIPS*, 2021.

[132] Xiaojiang Peng, LiMin Wang, Zhuowei Cai, Yu Qiao, and Qiang Peng. Hybrid super vector with improved dense trajectories for action recognition. In *ICCV Workshops*, volume 13, pages 109–125, 2013.

[133] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109–125, 2016.

[134] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016.

[135] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object

segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[136] Rizard Renanda Adhi Pramono, Yie Tarng Chen, and Wen Hsien Fang. Empowering relational network by self-attention augmented conditional random fields for group activity recognition. pages 71–90. Springer, 2020.

[137] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. pages 101–117, 2018.

[138] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[139] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. pages 779–788, 2016.

[140] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[141] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 2017.

[142] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10598–10607, 2020.

[143] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss

for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[144] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015.

[145] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. 115(3):211–252, 2015.

[146] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 23-26, 2004*, pages 32–36, 2004.

[147] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020.

[148] Yunhan Shen, Rongrong Ji, Shengchuan Zhang, Wangmeng Zuo, and Yan Wang. Generative adversarial learning towards fast weakly supervised detection. 2018.

[149] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. pages 10444–10452, 2019.

[150] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. Cern: confidence-energy recurrent network for group activity recognition. pages 5523–5531, 2017.

[151] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song

Chun Zhu. Joint inference of groups, events and human roles in aerial videos. pages 4576–4584, 2015.

[152] Xiangbo Shu, Jinhui Tang, Guojun Qi, Wei Liu, and Jian Yang. Hierarchical long short-term concurrent memory for human interaction recognition. 2019.

[153] Mykailo Shvets, Wei Liu, and Alexander C. Berg. Leveraging long-range temporal relationships between proposals for video object detection. In *ICCV*, 2019.

[154] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[155] Parthipan Siva and Tao Xiang. Weakly supervised action detection. volume 2, page 6, 2011.

[156] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France*, pages 1470–1477, 2003.

[157] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023.

[158] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[159] Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020.

[160] Ximeng Sun, Rameswar Panda, Chun-Fu Richard Chen, Aude Oliva, Roge-

rio Feris, and Kate Saenko. Dynamic network quantization for efficient video inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7375–7385, 2021.

[161] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. pages 1–9, 2015.

[162] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. pages 2818–2826, 2016.

[163] Kevin Tang, Bangpeng Yao, Li Fei-Fei, and Daphne Koller. Combining the right features for complex event recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2696–2703, 2013.

[164] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. PCL: Proposal cluster learning for weakly supervised object detection. 2018.

[165] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. pages 2843–2851, 2017.

[166] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan L. Yuille. Weakly supervised region proposal network and object detection. 2018.

[167] Yansong Tang, Zian Wang, Peiyang Li, Jiwen Lu, Ming Yang, and Jie Zhou. Mining semantics-preserving attention for group activity recognition. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1283–1291, 2018.

[168] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre

Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[169] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019.

[170] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. pages 6450–6459, 2018.

[171] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. 104(2):154–171, 2013.

[172] Muhammad Muneeb Ullah, Sobhan Naderi Parizi, and Ivan Laptev. Improving bag-of-features action recognition with non-local cues. In *BMVC*, volume 10, pages 95–1. Citeseer, 2010.

[173] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[174] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[175] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[176] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-MIL: continuation multiple instance learning for weakly supervised object detection. 2019.

[177] Fang Wan, Pengxu Wei, Zhenjun Han, Jianbin Jiao, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. 41(10):2395–2409, 2019.

[178] Hao Wang, Cheng Deng, Junchi Yan, and Dacheng Tao. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3939–3948, 2019.

[179] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103:60–79, 2013.

[180] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[181] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. Video modeling with correlation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 352–361, 2020.

[182] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1895–1904, 2021.

[183] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. pages 4325–4334, 2017.

[184] 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. pages 20–36. Springer, 2016.

[185] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang,

and Luc Van Gool. Temporal segment networks for action recognition in videos. 41(11):2740–2755, 2018.

[186] Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. pages 3048–3056, 2017.

[187] Minsi Wang, Bingbing Ni, and Xiaokang Yang. Recurrent modeling of interaction context for collective activity recognition. July 2017.

[188] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection. pages 542–557, 2018.

[189] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[190] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. pages 7794–7803, 2018.

[191] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018.

[192] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019.

[193] Zhenhua Wang, Qinfeng Shi, Chunhua Shen, and Anton Van Den Hengel. Bilinear programming for human activity recognition with unknown mrf graphs. pages 1690–1697, 2013.

[194] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas S. Huang. TS2C: tight box mining with surrounding

segmentation context for weakly supervised object detection. 2018.

[195] Chao-Yuan Wu, Ross Girshick, Kaiming He, Christoph Feichtenhofer, and Philipp Krahenbuhl. A multigrid method for efficiently training video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 153–162, 2020.

[196] Dongming Wu, Xingping Dong, Ling Shao, and Jianbing Shen. Multi-level representation learning with semantic alignment for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4996–5005, 2022.

[197] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. 2019.

[198] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022.

[199] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. pages 9964–9974, 2019.

[200] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. pages 485–501, 2018.

[201] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. Learning from temporal gradient for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3252–3262, 2022.

[202] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. pages 1492–1500, 2017.

[203] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017.

[204] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. CVPR, June 2016.

[205] Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. A discriminative CNN video representation for event detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1798–1807, 2015.

[206] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, pages 5036–5045, 2022.

[207] Rui Yan, Jinhui Tang, Xiangbo Shu, Zechao Li, and Qi Tian. Participation-contributed temporal dynamic model for group activity recognition. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1292–1300, 2018.

[208] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Higcin: hierarchical graph-based cross inference network for group activity recognition. 2020.

[209] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. pages 208–224. Springer, 2020.

[210] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 591–600, 2020.

[211] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. 2019.

[212] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019.

[213] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1939–1947, 2020.

[214] Zhenheng Yang, Dhruv Mahajan, Deepti Ghadiyaram, Ram Nevatia, and Vignesh Ramanathan. Activity driven weakly supervised object detection. pages 2917–2926, 2019.

[215] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10502–10511, 2019.

[216] Linwei Ye, Mrigank Rochan, Zhi Liu, Xiaoqin Zhang, and Yang Wang. Referring segmentation in images and videos with cross-modal self-attention network. *arXiv preprint arXiv:2102.04762*, 2021.

[217] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[218] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[219] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. In *AAAI*, volume 35, pages 3261–3269, 2021.

[220] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. 2021.

[221] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.

[222] Yuan Yuan, Xiaodan Liang, Xiaolong Wang, Dit-Yan Yeung, and Abhinav Gupta. Temporal dynamic graph lstm for action-driven video object detection. pages 1801–1810, 2017.

[223] Dingwen Zhang, Junwei Han, Le Yang, and Dong Xu. Spftn: a joint learning framework for localizing and segmenting objects in weakly labeled videos. 42(2):475–489, 2018.

[224] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. pages 4262–4270, 2018.

[225] Yu Zhang, Xiaowu Chen, Jia Li, Chen Wang, and Changqun Xia. Semantic object segmentation via detection in weakly labeled video. pages 3641–3649, 2015.

[226] Wangbo Zhao, Kai Wang, Xiangxiang Chu, Fuzhao Xue, Xinchao Wang, and Yang You. Modeling motion with multi-modal features for text-based video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11737–11746, 2022.

[227] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. pages 2921–2929, 2016.

[228] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, pages 6578–6587, 2019.

[229] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, volume 32, 2018.

[230] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. pages 850–859, 2019.

[231] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. pages 840–849, 2019.

[232] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[233] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Towards high performance video object detection. pages 7210–7218, 2018.

[234] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[235] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. pages 408–417, 2017.

[236] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815, 2022.

[237] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. pages 391–405. Springer, 2014.