Towards Better Accuracy and Efficiency in Classification and Generation using Deep Multi-Modal Learning

A thesis submitted in partial fulfilment of the requirements
for the degree of

Doctor of Philosophy
in
Analytics

by

Yuanzhi Liang

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

Jan 2024

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Yuanzhi Liang* declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Electrical and Data Engineering*, *Faculty of Engineering and Information Technology* 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.

SIGNATURE: _____

DATE: _____

ACKNOWLEDGMENTS

First and foremost, I extend my profound appreciation to my principal supervisor, Professor Yi Yang, for his invaluable guidance and unwavering support throughout my PhD journey. His expertise, insightful feedback, and inspiring mentorship have been pivotal in shaping my academic and research endeavors.

I am also deeply grateful to my co-supervisor, Dr. Linchao Zhu, for his exceptional dedication and patience in teaching and guidance. Dr. Zhu's expertise in experimental methodologies and paper writing, coupled with his constant encouragement, has significantly contributed to my development as a researcher. His willingness to share his vast knowledge and his approachable nature have greatly enhanced my learning experience.

My heartfelt thanks go to all my friends for their support and friendship during this journey. Their understanding, encouragement, and the joyful moments we shared have been a tremendous source of comfort and motivation through the challenging times of my research.

Lastly, I wish to express my deepest gratitude to my parents, Junsheng Liang and Shuang Li, for their unwavering love, sacrifice, and belief in me. Their constant support and encouragement have been the foundation of my resilience. This thesis is a culmination of not just my efforts, but the collective support and belief of all these remarkable individuals in my life. Thank you all for making this journey memorable and successful.

List of Publications

Related to the Thesis :

1. **Yuanzhi Liang**, Linchao Zhu, Xiaohan Wang and Yi Yang, "A simple episodic linear probe improves visual recognition in the wild," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022*

2. **Yuanzhi Liang**, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan and Yi Yang, "Seeg: Semantic energized co-speech gesture generation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022*

3. **Yuanzhi Liang**, Linchao Zhu, Xiaohan Wang and Yi Yang, "Penalizing the Hard Example But Not Too Much: A Strong Baseline for Fine-Grained Visual Classification," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS) 2022*

4. **Yuanzhi Liang**, Linchao Zhu, Xiaohan Wang and Yi Yang, "IcoCap: Improving Video Captioning by Compounding Images," *IEEE Transactions on Multimedia (TMM) 2023*

5. **Yuanzhi Liang**, Xiaohan Wang, Linchao Zhu and Yi Yang, "MAAL: Multimodality-Aware Autoencoder-Based Affordance Learning for 3D Articulated Objects," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2023*

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE **Page**

List of Tables

Abstract

In the rapid development of artificial intelligence, remarkable advancements have been achieved in diverse domains such as visual perception, visual-language understanding, virtual human generation, and robotics. These domains are intrinsically multi-modal, encompassing data from varied sources including images, text, audio, video, 3D pose, and robotic sensors. This multi-modal nature presents unique challenges and opportunities in AI research. One of the central challenges is effectively learning from this multi-modal data to build robust and versatile AI systems. To address this challenge, this thesis focuses on the crucial question: How can we accurately and efficiently extract valuable cues from multi-modal data to train effective deep neural networks? By exploring this question, the thesis aims to advance understanding and methodologies in the field, particularly in the areas of classification and generation, leveraging the complex and valuable multi-modal data.

This work bifurcates the scope of study into two primary areas: classification and generation. For classification tasks, we focus on visual modalities and improve the proposed methods' visual perception capabilities. The exploration can be separated into two folds. First, the thesis investigates a specialized and challenging task in classification,Äîfine-grained recognition. We propose a moderate hard example mining strategy to mitigate model overfitting and improve classification accuracy. This method prevents the network from merely memorizing hard examples, instead requiring the learning through moderated penalties. Then, beyond the specialized task, we go further with general classification tasks and explore a broader scope in visual modality. In this part, a novel method improving network generalizability and classification accuracy is presented. By introducing an episodic linear probing mechanism, this approach regulates network training, enhances representational discriminability, and bolsters neural network generalization across various perception tasks, including fine-grained, long-tail, and generic recognition.

In the domain of generative tasks, the thesis delves into text, pose, and affordance generation, each necessitating a deep understanding of multi-modal data. For text generation, an image-compounded captioning approach is introduced for video captioning, which effectively mines semantic cues from complex video data by jointly considering image and video properties. In pose generation, a semantic-energized method is proposed for virtual human pose generation. This approach enables networks to extract meaningful semantic cues from text and audio and generate accurate, expressive co-speech gestures aligned with speech semantics. Lastly, the thesis explores affordance learning in robotic

manipulation using 3D point clouds and robotic gripper data. To handle redundant data samples and multi-modal inputs, a multimodality-aware autoencoder framework is proposed. This framework facilitates efficient learning from sparse data samples, achieving comparable performance with limited positive samples and training epochs compared to previous works.

In conclusion, this thesis presents an exploration of various tasks in classification and generation, emphasizing multi-modal data properties. Extensive experiments across these tasks demonstrate the efficacy and efficiency of the proposed methods, consistently outperforming previous approaches.

Dissertation directed by Professor Yi Yang,
Australian Artificial Intelligence Institute, University of Technology Sydney

CHAPTER 1

INTRODUCTION

In the field of artificial intelligence, the pursuit of harnessing and comprehending multi-modal data stands as a crucial frontier. Humans perceive the world through multiple senses, with each contributing uniquely to our comprehensive understanding and intelligence. This ability to process and utilize multi-modal information has been fundamental in developing human intelligence [51, 138], complex societal structures [14, 73], and intricate organizations [73, 175]. Research highlights the significance of sensory input in human development, shaping the ways and possibilities for achieving intelligence. Treicher [99, 117, 118], a noted experimental psychologist, quantified human sensory input, finding that 83% of external information is visual, followed by 11% auditory. He further posited that humans remember 10% of what they read, 20% of what they hear, and 50% of what they see and hear combined.

Drawing inspiration from human sensory processing, numerous works [18, 20] in AI have explored the potential of using diverse modalities to construct AI systems. This exploration is more than a scientific endeavor; it is a journey toward unlocking AI's full potential, enabling machines to perceive [145, 179], integrate [85], interpret [2], and interact [84] with information from a data-rich world. This thesis aims to bridge the gap between AI systems and human perception, pushing AI to achieve a level of real-world data perception and understanding akin to human capabilities. As in prior works, we comprehensively explore vision, audio, and touch modalities, which are also the modalities machines can capture, process, and understand [18].

We start with vision, the most dominant human sensory. Our exploration spans from

specialized to general perception tasks, extending from single image understanding to complex video comprehension. This series of studies investigates how machines can better perceive domain-specific content, understand general content across various domains, and ultimately simulate human-like semantic responses to visual stimuli. After delving into vision, we add another vital sensory element - audio. We investigate machine understanding of audio information and behavior generation akin to human responses. The journey then advances to include the sense of touch, combining vision and tactile feedback to enable machines to comprehend real-world environments. We explore affordance learning for robot grippers, guiding robots in interacting with physical objects, akin to human vision and tactile abilities.

In detail, for visual sensory, we delve into visual classification, from specialized tasks like fine-grained visual recognition to various general visual representation learning tasks. These enable machines to perceive and understand visual data effectively. Moving from single-frame perception to multi-frame perception, we investigate video representation and how machines can simulate human reactions to visual information. Specifically, we explore video captioning, a task that involves describing observed video content, further broadening our research scope. Additionally, we explore auditory sensory processing, focusing on scenarios where responses are generated after listening. Co-speech gesture generation is identified as an ideal task for simulating human postural reactions during communication, enhancing our understanding of audio information processing and human-like response generation. Finally, we incorporate the sense of touch, focusing on affordance learning for robotic grippers, guiding robots to understand and manipulate real-world objects.

This thesis represents a comprehensive and diverse investigation, proposing novel methods that advance AI systems in perceiving and understanding multi-modalities. We offer new insights to the AI community and reorganize our content to narrate this research journey effectively. We categorize our research problems into two macroscopic AI tasks: classification and generation, each evaluated in terms of accuracy and efficiency. All these concretize our exploration as improving the accuracy and efficiency of classification and generation tasks in AI.

1.1 Improving Accuracy in Classification Tasks

Visual content, being among the most informative modalities in AI systems, presents crucial challenges and opportunities [81, 120, 143, 257]. The ability to perceive and interpret

visual data empowers AI systems to directly and intuitively understand their environments [85, 131, 145, 179]. This thesis first focuses on the visual modality, arguably the most informative yet complex, before addressing the intricacies of multiple modalities. In the diverse field of computer vision, tasks such as classification [120, 151, 253, 265], detection [143, 216], and segmentation [143, 156] continually elevate the requirements for models' visual perception abilities. Among these, classification stands as a foundational task, crucial for enabling machines to categorize and make sense of the vast array of visual information they encounter [81, 129, 253]. Moreover, the effectiveness of classification tasks serves as a direct benchmark for evaluating models' ability to perceive visual data [145, 253]. Thus, as a foundational step in exploring multi-modal data, this thesis initially investigates visual classification tasks.

For classification tasks, our work unfolds in two dimensions, progressing from specialized to general tasks. First, we focus on a specialized and challenging task, fine-grained visual classification (FGVC), a longstanding challenge involves dealing with visually similar classes that are difficult to generalize. FGVC is particularly crucial due to its applications in biodiversity [217], retail [115, 133], and medical diagnostics [66, 71, 150], where discerning subtly different categories is essential. FGVC's primary challenge lies in models' tendencies towards overfitting [257, 265] and poor generalization [81, 257] when handling closely related categories. To address this, the thesis introduces Moderate Hard Example Modulation (MHEM). MHEM moderately modulates the penalties for learning hard examples during training, mitigating overfitting and enhancing model generalization. This approach represents a significant advancement in refining models' ability to discern nuanced differences between similar categories, progressing visual perception capabilities of AI systems.

Additionally, the thesis tackles broader challenges in visual recognition, with a focus on network generalization and feature discrimination. To this end, Episodic Linear Probing (ELP) is introduced. ELP bolsters the generalizability of visual representations through a strategy of episodic re-initialization during training. This dynamic mechanism continuously refines the network,Äôs understanding of complex visual data, thereby enhancing feature discrimination and overall network robustness. The methodological rigor and efficacy of ELP present considerable improvements in various tasks visual perception, including fine-grained, long-tailed, and generic visual classification.

1.2 Enhancing Accuracy and Efficiency in Generation Tasks

Diving deeper into the realm of more diverse modalities, this thesis broadens its investigative scope to include generative tasks that encompass text [2, 105], audio [178], images [145, 253], videos [72, 181], 3D poses [4], and 3D point clouds [159, 180, 227], etc. These tasks introduce a broader and more intricate domain of study. In particular, this research delves into three distinct types of generation tasks, each leveraging multi-modal inputs: text generation derived from video inputs, pose generation from text and audio inputs, and affordance generation utilizing point cloud and robotic information inputs.

In text generation from video inputs, the focus is placed on video captioning. Confronted with challenges such as content density and ambiguity, this thesis introduces an Image-Compounded Learning method for video Captioner (IcoCap). IcoCap merges image semantics with video content, thereby facilitating more accurate and contextually relevant semantic extraction from videos. This method represents a significant advancement in the field of video captioning, yielding a more proficient captioner capable of producing higher accuracy captions.

Progressing to the generation of talking gestures, the thesis presents the innovative Semantic Energized Generation (SEEG) method. This approach is pivotal in the domain of virtual human generation, where co-speech gestures play a crucial role in conveying semantics. SEEG is designed to extract semantic cues from speech and text, enabling the generation of semantically-aware and vivid gestures. This method effectively improve the realism and expressiveness of virtual human gestures, pushing the boundaries of gesture generation.

Lastly, the thesis ventures into the real-world domain of robotics, targeting diverse modalities pertinent to robotic systems. The focus here is on affordance learning for 3D articulated objects in robotic applications. In this thesis, Multimodality-Aware Autoencoder-based Affordance Learning (MAAL) framework is introduced, integrating multi-modal data to enable efficient affordance learning. MAAL presents a novel approach in affordance generation for robotic manipulation, offering a process that is both data-efficient and robust.

Through these studies, this thesis presents a comprehensive research to addressing challenges in multi-modal AI learning. Each method contributes uniquely to the field, enhancing our understanding and capabilities in handling complex multi-modal data. These advancements demonstrate significant improvements over existing methods and

provide valuable insights for future exploration in artificial intelligence. This thesis not only addresses specific challenges within the realms of classification and generation but also contributes to the broader narrative of AI developments. By proposing novel and effective methods, it enhances the accuracy and efficiency of AI systems in processing multi-modal data.

In essence, this thesis marks a substantial contribution to the field of artificial intelligence. It underscores a significant step forward in creating effective and efficient AI systems capable of navigating the intricate landscape of multi-modal data. The insights and methodologies developed throughout this research pave the way for AI systems that are not only more effective in their current applications but also more versatile and capable of adapting to new and evolving challenges in the future.

1.3 Thesis Organization

In Chapter 3, we embark on the exploration of Fine-grained Visual Classification with the introduction of the Moderate Hard Example Modulation (MHEM) method. This foundational technique addresses model overfitting by modulating hard examples within training datasets, paving the way for improved generalization capabilities in specialized classification tasks. The success of MHEM sets a precedent for addressing overfitting and enhancing model performance, themes that are recurrent throughout the subsequent chapters.

Transitioning from the specialized context of Chapter 3, Chapter 4 expands our scope to general classification tasks with the introduction of Episodic Linear Probing (ELP). Building on the groundwork laid by MHEM's approach to model generalization, ELP furthers this narrative by specifically targeting network generalization and feature discrimination. This chapter not only demonstrates ELP's efficacy in enhancing visual recognition but also bridges the gap between overcoming overfitting in specialized tasks and advancing towards broader generalization in visual tasks. The methodologies developed here are instrumental in preparing the network for more complex and multimodal challenges ahead.

Chapter 5 takes the journey into the realm of multimodal learning, focusing on video captioning through our novel Image-Compounded Learning for Video Captioning (IcoCap) method. This chapter represents a pivotal shift from the visual classification tasks of the previous chapters to the generation of textual descriptions from video inputs. IcoCap's approach to compounding image semantics with video content illustrates the natural

progression from enhancing network generalization to applying these advancements in understanding and generating complex multimodal data. The evolution from visual recognition to video captioning underscores the thesis's exploration of deepening levels of content complexity and ambiguity.

In Chapter 6, we delve into the novel area of virtual human generation, specifically focusing on co-speech gesture generation through the Semantic Energized Gesture Generation (SEEG) method. This chapter builds on the multimodal advancements of IcoCap by applying learned lessons to the generation of semantically and rhythmically aligned gestures, highlighting a further application of our research in enhancing human-computer interaction. The progression from static images to video and now to interactive gestures exemplifies the thesis's overarching narrative of tackling increasingly intricate challenges through innovative methodologies.

Finally, Chapter 7 encapsulates the thesis's culmination in applying our progressively developed methods to real-world, multi-modal environments through the Multimodality-Aware Autoencoder-based Affordance Learning (MAAL) for robotic grippers. This chapter not only showcases the application of our research in a practical context but also represents the zenith of our exploration into multi-modality and the learning of complex affordances with minimal data. The narrative arc from addressing overfitting in fine-grained classification to enabling sophisticated interaction in diverse real-world environments through MAAL illustrates the thesis's comprehensive exploration of advancing AI methodologies for improving model generalization, feature discrimination, and multi-modal understanding. The final chapter summarizes the key findings of the thesis and offers insights into potential future research directions in the field.

CHAPTER 2

LITERATURE SURVEY

This chapter presents an extensive literature review within the realm of multi-modal learning in artificial intelligence. It delves into pivotal advancements and methodological developments that have shaped modern AI research. Multi-modal learning, essential for robust AI systems, encompasses domains from classification to generation. This review contains diverse approaches, challenges, and innovations in this diverse areas.

2.1 Classification Tasks

2.1.1 Fine-grained Classification

With the success of deep learning, the mainstream methods of fine-grained recognition shift from multi-stage frameworks based on hand-craft features [246, 265] to multi-stage frameworks with CNN features [114, 236]. The second-order bilinear feature interactions show significant improvements in representation learning [144]. Some methods based on metric learning are also efficient in capturing subtle details in images. Inspired by some weakly supervised learning methods [260, 262], Huang et al. [97] introduce additional localization module to improve performances and interpret attention areas. Moreover, ELP [135] provides a novel routine to improve the generalization in classification by regularizing the classifier's immediate suitability. Besides, Chen et al. [41] hierarchically predict the categories within different levels in the networks. However, these ideas introduce complex networks and high computational complexity.

Recently, the most popular methods in FGVC are part-based methods. These methods aim to find the discriminative part for classification. Among these, the attention strategy provides an approximation to the human visual system [69, 224, 270] and plays an important role in FGVC. Zheng et al. [270] propose to generate attention for multiple parts by a channel grouping network. Yao et al. [58] propose a novel sparse attention method to selective sample discriminative parts. The common attention may be misled by some biases or background context. With intensive augmentation, some methods impose the model to learn some robust and discriminative parts. Destruction and Construction Learning (DCL) [44] proposes a pipeline to learn with destructed samples and learn to re-construct the samples. The common samples apply the jigsaw [46, 203] operation for destruction. Then, a region alignment network is designed to recover destructed data. Besides the main classifier, the adversarial learning network is proposed to distinguish swapped or normal images. Considering the different degrees of jigsaw operations expose different properties of the data sample, to further investigate local contexts, Progressive Multi-Granularity Training (PMG) [61] takes multiple inputs with different scales and degrees of jigsaw operation. PMG designs the multi-stage training to learn multi-scale and multi-degree jigsaw data and introduces an associative learning method to jointly learn all information in various scales and stages. Look-into-Object (LIO) [277] explicitly and intrinsically learn the object structure, which provides a novel solution to mine structural features for recognition.

Another thought in FGVC is introducing external knowledge into training. Though expensive costs may be required, this kind of method improves the performance significantly. It has been shown that combing multiple knowledge sources often helps discriminative feature learning [244]. Additionally, [10, 62] point out the severe overfitting problem and propose methods to mitigate. Among these, the representative work is Pair Confusion [62]. It utilizes a siamese neural network and reduces the distances between the conditional probabilities of two networks.

Besides, the hard example contain valuable information. Mining the hard examples in training has been researched in many works [67, 235, 261]. Shrivastava et al. [67] propose the Online Hard Example Mining (OHEM) to boost performances of object detection. Moreover, works [96, 235] provides various online methods to select hard pairs or triplets for training. Rather than specifically select samples, choosing some relatively hard examples in the mini-batch and enhance the learning is also effective. Lin et al. [142] proposed Focal Loss to re-balance the loss weight of different samples according to the sample probabilities. Focal Loss is also a helpful hard-mining loss. The

hard examples with low confidence would be assigned higher weights. This loss works in many domains like detection, segmentation, classification, etc.

2.1.2 General Visual Classification and Representation Learning

Various works have been proposed to learn visual representation based on deep learning. In diverse recognition tasks in the wild, deep neural networks possess the powerful ability to learn and represent images to high-dimensional features. With the high-quality features, some simple classifiers [119, 233] are components to recognize the samples. Further, the quality of features is influenced by many factors. We roughly divided the factors into three aspects: data processing, network design, and training manner. Though the exact effect of representation learning [257] remains to be investigated, numerous researchers keep exploring and propose many valuable solutions.

For data processing, large-scale datasets provide considerable network samples and are the most straightforward way to improve representation. Benefiting from the powerful ability of networks, taking large-scale datasets as inputs lead the network to learn various samples and memorize plenty of properties for discriminating. Some diverse and hard examples may be difficult in a limited data scale [10, 147]. Under the view of larger scales of collections, it is always possible for the network to mine particular patterns. Besides directly collecting real data, pre-processing [47, 275] or generating data [274] are also equivalent. Various augmentations [198, 210] enforce the networks to solve problems with higher requirements and urge the network to be generalized to different conditions. However, the most straightforward way is also the most expensive. The storage, computing power, etc., should be concerned to handle the large datasets. Meanwhile, with the expansion of the data scale [214, 237], the efficacy and value of data should also be considered.

Moreover, well-designed network structures also dramatically boost representation and become the hottest direction in recent years. Diverse methods constantly emerge like skip-connection [88, 95], fusing channels [207], attention strategies [24, 158], architecture searching [27], transformers [212, 221], etc. With the same inputs, these methods explore different directions to boost the network's capacity. Meanwhile, almost all kinds of visual tasks [120, 143] develop further with better networks.

Furthermore, besides data processing and network designs, the training manner is also crucial for visual representation. It contains various aspects like the optimizer [89, 188], regularization [123, 142], learning manner [104, 199], etc. In this direction, regularization plays an important role. It can be reflected in the loss function [33, 142],

training strategies [87], etc., and is general to various networks and datasets. A proper regularization can leverage the network to learn better visual representation, for example, avoiding overfitting [142], explicit attention to the target [33], better diversity [50], etc. Vikash et al. [192] propose an interesting margin to describe the separability of features. Rather than focusing on the accuracy of the classifier, the quality of features can be reflected through immediate suitability. The more discriminative features are considered more than memorable by the classifier.

2.2 Generation Tasks

2.2.1 Video Captioning

Video captioning [2, 40, 240] is a challenging and complex task that aims to generate a natural language sentence to describe a given video sequence. Unlike image captioning, where the objective is to generate descriptions for static images, video captioning methods need to handle intricate video data that encapsulates diverse and dynamic semantics. The temporal dimension of video data adds a level of complexity that requires sophisticated approaches to capture and summarize the underlying content effectively.

In detail, the common approach in video captioning is the encoder-decoder framework, which employs a CNN to encode visual information and an RNN or LSTM to generate captions sequentially. Donahue et al. [59] proposed the Sequence-to-Sequence Video-to-Text model, which combined a 2D CNN with an LSTM to generate captions. Chen et al. [36] introduced the TDConvED network,Äîa convolutional sequence-to-sequence learning framework, specifically tailored to enhance video captioning. Most recent works [39, 126, 141, 267] also follow this framework and present various solutions to further boost the performances. Moreover, Chen et al. [45] propose to select frames in video for video captioning. Pan et al. [168] introduce a visual semantic embedding model to specifically consider the relationship between the semantics of the entire sentence and video content unexploited.

Video captioning [2, 105] is a challenging task that aims to produce a sentence to describe a video sequence. Rather than image captioning, methods in video captioning need to handle complicated video data with diverse semantics. Different from dense captioning [100, 195, 276], the typical methods [39, 126, 141, 197, 219, 267] for video captioning should summarize the diverse and ambiguous contents of the video into one sentence.

Moreover, another line of evolution is the video representation method. Works in video captioning apply features from some pre-trained models to represent videos. Models like bottom-up [11] in image representations, 3D CNNs [53, 238] in video representation, or generic large-scale pre-training models [141, 183, 194] are applied in video captioning to represent video data. Then, various methods [1, 19, 146, 167] are designed to investigate the semantic cues from well-trained representations and solve video captioning. Yang et al. [241] conducted a comparative analysis between CLIP features and ImageNet pre-trained features for video captioning. Additionally, they introduced an auxiliary task designed to discern the correspondence between video content and associated concepts. Some recent works [219, 267, 268] introduce complicated structures to mine detailed information from video features and achieve significant improvements. Besides, some works [141, 267] further propose end-to-end frameworks for representing videos from scratch and exploring the detailed instances and events in the video frames.

2.2.2 Co-speech Gesture Generation

Speech-driven gesture generation is an emerging issue that aims to generate vivid gestures based on the given speech data. Generally, methods for this problem take the speech data [249, 250] (audio, text, etc.) as input and produce corresponding gestures to simulate the real speaker. This requires various knowledge understanding [244] like human ethology [26, 154, 177, 229], linguistics [116, 174, 186], robotics [60, 165], graphics [7, 90, 249], vision [121, 182, 249], etc. Proposed methods should understand multi-modal and diverse information (speech rhythm from audio, text semantics, personal style from speakers' identities, semantic conveyed from motions, etc.), then generate reasonable and expressive gestures.

To overcome the above challenges, various works are proposed to explore. To understand the audio data and bridge the audio inputs to the gestures, Taras et al. [121] investigate the network structure to map speech acoustic and semantic features into the feature space of 3D gestures. Moreover, benefiting from an efficient modeling method MoGlow which is controllable for 3D motion synthesis, Alexanderson et al. [7] propose the style-controllable gesture generation model based on the MoGlow. The proposed method can generate diverse and plausible gestures just like the actual human. Ahuja et al. [3] propose Mix-StAGE, which disentangle the style feature with gesture features and encodes the gestures features to the style space. Mix-StAGE overcomes the challenge of style preservation and generates diverse styles of gestures for different people. As the multi-modalities involved in speech-driven gestures, Yoon et al. [249] explore the embed-

ding and representation of multiple modalities for gesture generation. They consider the trimodal context and construct holistic modeling for all the data.

In addition, the metrics for evaluating the generated gestures are also important and challenging. As the uncertainty of human behavior, evaluating the realistic level of generated gestures compared with the actual human maybe still an open question. Some works [3, 7, 75] rely on user studies to measure the quality of generated gestures. Rather than the subjective evaluation from an actual human, some works [3, 75, 182, 249] calculate the distances between generated gesture and the ground truth.

2.2.3 Affordance Generation

In the field of robotics, 3D object affordance is an important area of many practical applications. Before manipulating objects in reality, the robots need to understand what and where can be acted at first, which can be contributed to the exploration of affordance [74]. Recently, many works have emerged to explore this problem. [112] and [185] leverage the CNN network to produce the affordance area of the affordance map, which is used for indicating the grasping operations of robots. Jiang et al. [103] propose to constrain the consistency between hand contact points and object contact regions. The contact points of the robot hand are required to be close to the shape of the object's surface. Then, Mo et al. [159] provide a large-scale dataset and benchmark. The authors also predict affordance maps to indicate the actionability of robots at every point of objects. 3DAffordanceNet [54] explore another interesting problem and introduces a dataset for the functional understanding for 3D objects. Moreover, AdaAfford [227] goes further with the affordance predictions, considers the information hidden in the 3D shapes, and mines important kinematic and dynamic factors in 3D interactions. Through better modeling of the kinematic uncertainties, AdaAfford improves the performance of manipulating objects within fewer action steps. The significant advancements in [159] and [227] should be admired, but these works also contain defeats. All previous works utilize multiple decoders or critics to predict the probability of actionability (separately training three networks in [159] and four networks in [227]). The method design is complex and requires many data samples for training. In this work, we propose an AE-based pipeline to solve the problem efficiently.

2.3 Multi-modal Learning

2.3.1 Video-language Representation Learning

Representation of video [17, 82, 128, 141, 225, 278] is a long-standing problem in the representation learning [83, 230, 242, 243]. Numerous works have emerged, proposing diverse architectures and approaches that focus on exploiting the unique characteristics of video data to achieve effective and robust representations. In representation, the intuitive idea behind video representation is to extend the principles of image-based CNNs, which have demonstrated remarkable success in tasks such as object recognition and image classification.

One notable approach to incorporate temporal information into the original CNN framework is by introducing 3D kernels [30, 157]. These kernels extend the receptive field in the time dimension, thereby enabling the network to capture the relationships between sequential frames. This extension results in 3D Convolutional Neural Networks (3D CNNs) [30], which are specifically designed to process video data by jointly learning spatial and temporal features and have demonstrated considerable improvements in video representation tasks compared to their 2D counterparts. However, one drawback of 3D CNNs is the increased computational complexity and memory requirements, which can pose challenges in terms of scalability and efficiency. I3D [238] inflated the filters and pooling layers of 2D CNNs into 3D, enabling the network to learn richer spatio-temporal features. The I3D model achieved significant improvements in action recognition tasks and demonstrated the potential of incorporating pre-trained 2D CNN knowledge into video representation learning. More variations [53] of 3D CNNs further provide many video-based designs to boost the performances of representations in various tasks.

Moreover, recent works [12, 183, 278, 279] pay more attention to the large-scale pre-training. Motivated by the success of Bert [110] in NLP, many works [205, 279] propose to leverage the similar pre-training strategies to videos. Significant improvements occur in video tasks after applying the large-scale pre-training [17, 148] and various transformer-based networks [8, 12, 141]. Besides, tasks like mask-modeling [110], contrastive learning [43, 223], etc., further empower the representation ability of networks. CLIP [183], as a typical pre-training model, has also been proven that possesses remarkable ability in correlating language semantics and has already been widely used in various domains [149, 209]. These video-based designs have contributed to the evolution of video representation learning, enabling more effective and discriminative representations for various tasks. Despite the progress made thus far, video representation remains

an active area of research, with ongoing efforts to develop more efficient and accurate models capable of handling the ever-increasing complexity and scale of video data.

2.3.2 Multi-modal Fusion and Learning

Many tasks (e.g., VQA [16, 129], gesture generation [130, 249], video representation [128]) involve multi-modal inputs and require the network to handle the multi-modal problems [122, 220]. These problems usually entail the understanding of various knowledge [244] and require the proper handling of diverse inputs. Generally, the network needs to handle data samples with various modalities, which may possess different distributions and semantics. Methods usually need to fuse data or features for further learning. Formally, there are three kinds of strategies [111, 122] to fuse multi-modal data: early fusion, late fusion, and inter-media fusion. Early fusion means fusing data samples before specific learning. Methods [6, 78, 122] with early fusion usually combine raw data without considering the connection between data samples or fuse embedded features in low dimensional space. This strategy may be useful if the multi-modal data are conditionally independent [166, 176, 191]. However, the performances for highly correlated data samples or features would be lower [153]. Moreover, late fusion [106, 111, 200, 231] indicates the independent learning data sample before the last module, which is used for decision-making (e.g., classifier, retrieval projector). This leads the network can understand each modality better and avoid accumulating uncorrelated errors [184]. However, the advantages of late fusion in multi-modal tasks are insignificant [78, 184, 202] compared with early fusion. Finally, intermediate fusion [25, 122] is the most commonly used strategy in recent multi-modal learning. It flexibly fuses different data samples at different levels and designs explicit modules to model different modalities adaptively. Many works [57, 108, 248] with intermediate fusion achieve better performances in various multi-modal tasks.

CHAPTER 3

TOWARD BETTER ACCURACY FOR FINE-GRAINED VISUAL CLASSIFICATION

3.1 Introduction

Fine-grained visual classification (FGVC) is one of the long-standing recognition problems in computer vision [140, 196, 258]. Differ from the typical large-scale visual classification (LSVC) task like [120], classes in FGVC are visually similar to each other. FGVC may require to differentiate a Laysan Albatross (a large seabird) from a Sooty Albatross (a species of bird in the Albatross family). This task imposes difficulties even for a human. To solve the challenging task, various methods [44, 58, 61, 144, 271] are proposed and these methods target at learning more discriminative features. These research works achieve remarkable improvements on FGVC.

However, Anderson et al. [10] have pointed out that current methods and even the ensembles of some state-of-the-art methods still misclassify some 'hard examples' in the testing set. These examples are extremely hard to be recognized given the intrinsic intra-class similarities among fine-grained categories. According to the i.i.d. assumption, when splitting the data into training and testing sets, there are also a few extremely hard examples in the training set, e.g., Black Tern and Caspian Tern. The existing methods [44, 61, 70] can accurately classify these hard examples even though they are almost visually identical for non-expert human observers. Taking a simple ResNet50

This chapter is based on joint work [134] with Linchao Zhu, Xiaohan Wang, and Yi Yang, presented primarily as it appears in the TNNLS 2022.

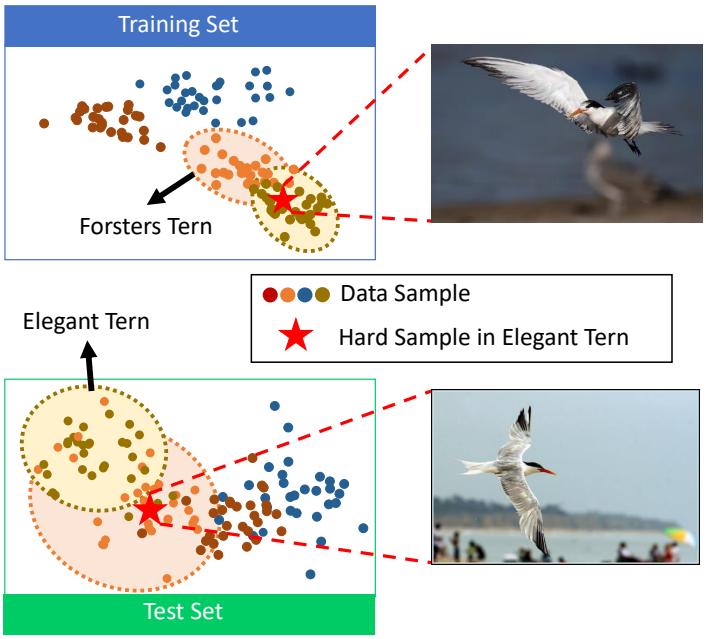


Figure 3.1: Feature visualization of Foresters Tern and Elegant Tern (hard classes in FGVC [10]). The features are extracted from a ResNet50 network trained with 20 epochs. In the training set, all examples including hard ones are correctly classified. However, the network fails to classify a few hard examples in the test set.

baseline as an example, the average softmax probability of training samples rises to 0.95 rapidly, and the training accuracy comes to 100% within 20 epochs. This demonstrates that the network can quickly classify the hard examples because deep convolution neural networks are remarkably powerful to well fit the training data [257]. However, in the test set, the network shows its inability to classify some hard examples (Fig. 3.1). The performance gap between the training set and the test set reveals that the network can not be generalized to recognize 'hard examples' in the test set, even if it easily solves some extreme cases in the training set.

The key challenge of this phenomenon is that network ***memorizes*** and overfits the hard examples in the training set, but it does not learn to generalize well to new hard examples in the test set. The learned representation is not discriminative even though the optimization process in the training set is empirically easy. However, it remains to be an open research problem to understand why the feature is not discriminative and discover the root cause of the generalization gap [257].

Motivated by the observations of inferior generalization on hard examples [10], in this chapter, we attempt to reduce the generalization gap in FGVC and aim to improve

the generalization through a proper modulation of hard examples. We empirically find that it is not effective to amplify the importance of hard examples during the training of FGVC models. We note that emphasizing the importance of hard examples induces even severe overfitting. Therefore, a better modulation for hard examples in FGVC should be concerned.

To this end, we propose a Moderate Hard Example Modulation (MHEM) strategy to moderately penalize the hard examples by considering their loss scales. **First**, we propose three conditions for MHEM, i.e., the hard mining condition, the moderate condition and the moderately sensitive condition. **Second**, we formulate a modulating function to generate proper weights for the training samples. The modulating function consists of a few hyper-parameters. We present their mathematical relations to satisfy the aforementioned conditions. Notably, we find that the typical Focal Loss [125, 254] does not satisfy the MHEM conditions. We further discuss the differences between Focal Loss and our modulating function in Section 3.2.4. **Third**, we construct a Moderate Modulation Baseline (M2B) to facilitate the network for a better generalization on hard examples. Without bells and whistles, this strong baseline shows that applying proper modulation weights to hard examples brings significant improvements to FGVC. Without introducing any extra computational overhead, the performances of a naive ResNet50 backbone can be competitive to some state-of-the-art methods. It demonstrates that our simple baseline is efficient and effective to overcome the overfitting problem and promote the network's generalization. Quantitatively, we evaluate our moderate modulation baseline as a new loss on the existing FGVC methods. We achieve consistent improvements across the typical FGVC benchmarks, validating that M2B can serve as a strong baseline for the future FGVC research works.

The main contributions of this chapter are summarized below:

1. We propose Moderate Hard Example Modulation (MHEM) to enable the generalization of hard examples in FGVC. We introduce three conditions for a moderate modulation, allowing the network to properly learn from hard examples and alleviate the overfitting problem.

2. We formulate a modulating function and present the mathematical relations of its hyper-parameters. We quantitatively validate its flexibility for fine-grained visual recognition.

3. We instantiate a strong moderate modulation baseline (M2B) that satisfies the MHEM condition. M2B does not introduce any computational overhead and achieves significant improvements. Notably, a naive backbone network with M2B achieves 2.7%,

1.5% and 2.6% improvements in three datasets respectively, which can outperform or be competitive to some state-of-the-art methods. When applying M2B on the existing state-of-the-art methods, we also obtain notable improvements, validating its broad applicability on the FGVC task.

3.2 Method

In this section, we first discuss three conditions of Moderate Hard Example Modulation (MHEM) (Section 3.2.1). These conditions serve as the foundation of MHEM and allow us to design specific modulation functions. Then, we simplify the form of modulating function and present the mathematical relations of its hyper-parameters (Section 3.2.2). We further discuss a special case of MHEM and demonstrate its flexibility. Finally, we propose a strong moderate modulation baseline for fine-grained visual recognition (Section 3.2.3).

3.2.1 Moderate Hard Example Modulation (MHEM)

Motivated by the observations of inferior generalization on hard examples [10], we propose a Moderate Hard Example Modulation (MHEM) to modulate the losses according to the confidences of the training samples. Generally, the overall loss function $\mathscr{L}$ in MHEM can be formulated as:

$$\mathscr{L}(p) = f(p)\mathscr{L}_c(p) \tag{3.1}$$

where $p \in [0,1]$ is the probability of the sample, $\mathscr{L}_c(\cdot)$ indicates the regular cross-entropy loss. $f(\cdot)$ is a modulating function. In practice, $f(p) \geq 0$.

In MHEM, we re-weight each sample with a modulating function $f(\cdot)$. The modulating function takes the classifier prediction p as the input and outputs a weighting scalar $f(p)$ for every prediction. A proper weight is applied to each sample, which would enable a better modulation of hard examples.

To examine the properties of $f(\cdot)$, we define p_h as the probability of the hard example and p_e as the probability of the easy example. We introduce a relative coefficient ϕ and a sensitive coefficient ρ. We let $\phi = \frac{f(p_h)}{f(p_e)}$ and $\rho = \frac{f'(p_h)}{f'(p_e)}$, where $f'(\cdot)$ indicates the derivative of $f(\cdot)$. The relative coefficient ϕ measures the ratio between the weight of hard examples and the weight of easy examples. A larger ϕ indicates that the network is inclined to

solve hard examples than the easy examples. The sensitive coefficient ρ measures the ratio between the weight variation of hard examples and the weight variation of easy examples. With a small positive value Δ, from a hard example with $p_h + \Delta$ to another harder example with p_h, the variance of weights for harder examples can be formulated as $f'(p_h) = (f(p_h + \Delta) - f(p_h))/\Delta$. Similarly, the variance of the weights for easy examples is $f'(p_e) = (f(p_e + \Delta) - f(p_e))/\Delta$. The larger ρ indicates that the weights of hard examples increase more rapidly than the weights of easy examples. When ρ is large, the network will be more sensitive to the hard examples, where a slightly lower confidence may induce a significant amplification of its weight.

Our empirical study found that severe overfitting occurs when introducing high weights for hard examples. Inspired by this, we consider that the hard samples should be emphatically trained but not be penalized too much. In MHEM, we propose that the modulating function $f(\cdot)$ should retain three conditions: (1) Hard mining condition, (2) Moderate condition, and (3) Moderately sensitive condition.

(1) **Hard mining condition:** $\forall p_e > p_h, \phi > 1$. Due to the defeats in recognizing hard examples, the learning of hard samples should be emphasized. One of the efficient ways [34, 125, 162, 254] for hard mining is leveraging large punishments for hard examples. Compared to $f(p_e)$, a larger value of $f(p_h)$ imposes a larger punishment for hard examples. This enforces the network to classify the hard samples better. Given the hard mining condition, we conclude that $f(p)$ is a monotonically non-increasing function when $p \in [0, 1]$.

(2) **Moderate condition:** $\forall p_e > p_h, \exists \epsilon, \phi < \epsilon$. ϵ is a moderate factor which indicates the upper bound of the relative coefficient ϕ. In MHEM, the relative coefficient should not be too large. The moderate factor ϵ is introduced to establish the upper bound for the punishments of hard examples. It is responsible to ensure the suitability of hard example modulation. In order to satisfy the hard mining condition, we obtain $\epsilon > 1$. Meanwhile, the value of ϵ should not be too large. If $\epsilon = \infty$ and $\phi \to \infty$, the network will only learn the hard examples and ignore the easy examples, which may cause skewed memorization of hard examples. Moreover, the value of the moderate factor ϵ varies across datasets, which depends on the distribution of hard and easy examples and the inter-class and intra-class similarities.

(3) **Moderately sensitive condition:** $\forall p_e > p_h, \exists \xi, \rho < \xi$. ξ is the sensitive factor which limits the variances of the weights. In MHEM, other than the numerical limitation of $f(\cdot)$, the variance limitation should also be concerned. If the variance of weights for hard examples are far larger than that of easy examples, a slight change in the

confidences of hard examples may lead significant changes of gradients and the network will become over-sensitive to hard examples. This also should be prevented in MHEM. Since the modulating function is monotonous, we assume that $\xi > \rho > 0$. Meanwhile, ξ should not be too large. If $\xi = \infty$ and $\rho \to \infty$, the modulating function introduces too sensitive punishments for hard examples. Similar to the moderate factor ϵ, the value of ξ also depends on the datasets.

To simplify the conditions, we set a general modulate factor $\tau = min(\epsilon, \xi)$ where $min(\cdot)$ is a function returning the minimum number between ϵ and ξ. Then, we rewrite the condition (2) and condition (3) as $\forall p_e < p_h, \exists \tau, \phi < \tau, \rho < \tau$. Thus, the key to MHEM is constructing a proper modulating function $f(\cdot)$ satisfied $\phi \in (1, \tau)$ and $\rho \in (0, \tau)$.

3.2.2 Formulation of MHEM

To formulate MHEM and construct a practical baseline method, we first assume the modulating function $f(\cdot)$ as a polynomial function:

$$f(p) = Sgn((ap+b)^{\vartheta}) + d \tag{3.2}$$

where a, b, d and ϑ are the hyper-parameters. d is a global bias term preventing $f(p)$ from being too small. $Sgn(\cdot)$ is a sign function and we obtain:

$$Sgn((ap+b)^{\vartheta}) = \begin{cases} 0 & (ap+b)^{\vartheta} < 0 \\ (ap+b)^{\vartheta} & 0 \le (ap+b)^{\vartheta} \le 1 \\ 1 & (ap+b)^{\vartheta} > 1 \end{cases} \tag{3.3}$$

When $0 \le (ap+b)^{\vartheta} \le 1$ and $d = 1$, we can further simplify the function and we rewrite the function as:

$$f(p) = (\alpha p + \beta)^{\vartheta} + 1 \tag{3.4}$$

α and β denote the adaptive slope and the bias term, respectively. α and β should be adjusted by the general moderate factor τ. ϑ is usually a positive number.

Considering the hard mining condition (1), $f(\cdot)$ is monotonically non-increasing. Thus, we obtain $\alpha \le 0$. For $\alpha = 0$, $f(p)$ becomes a constant value and the overall loss function degrades to the cross-entropy loss. We set $\alpha < 0$ and we assume $\beta > 0$ because $(ap+b)^{\vartheta} \ge 0$.

According to the moderate condition (2), the relation between the general moderate factor and the modulating function can be formulated as:

$$\frac{(\alpha p_h + \beta)^{\vartheta} + 1}{(\alpha p_e + \beta)^{\vartheta} + 1} < \tau \tag{3.5}$$

We can also easily obtain:

$$\frac{(\alpha p_h + \beta)^{\vartheta} + 1}{(\alpha p_e + \beta)^{\vartheta} + 1} < \frac{(\alpha p_h + \beta)^{\vartheta}}{(\alpha p_e + \beta)^{\vartheta}} \tag{3.6}$$

To achieve a stricter condition, we rewrite Equation 3.5 with the upper bound:

$$\frac{(\alpha p_h + \beta)^{\vartheta}}{(\alpha p_e + \beta)^{\vartheta}} < \tau \tag{3.7}$$

According to the moderately sensitive condition (3), we obtain a similar condition:

$$\frac{(\alpha p_h + \beta)^{\vartheta-1}}{(\alpha p_e + \beta)^{\vartheta-1}} < \tau \tag{3.8}$$

As $f(p_h) > f(p_e)$, $\frac{\alpha p_h + \beta}{\alpha p_e + \beta} > 1$, therefore, both condition (2) and condition (3) can be satisfied under Equation 3.7. Besides, though formulating with the same equation, the moderate condition and moderately sensitive condition represent different kinds of regularization for the samples. The moderate condition requires that the moderating function's values (absolute value) should not be too large, and the moderately sensitive condition requires that the changing weights (derivatives) should not be too large.

Considering $p_e > p_h$, $p_e \in [0,1]$ and $p_h \in [0,1]$, the limitation of $\frac{(\alpha p_h + \beta)^{\vartheta}}{(\alpha p_e + \beta)^{\vartheta}}$ can be deduced:

$$\frac{(\alpha p_h + \beta)^{\vartheta}}{(\alpha p_e + \beta)^{\vartheta}} \leq \frac{\beta^{\vartheta}}{(\alpha + \beta)^{\vartheta}} \tag{3.9}$$

To make the moderate condition be always established, we let the maximum ϕ still lower than τ.

$$\frac{\beta^{\vartheta}}{(\alpha + \beta)^{\vartheta}} < \tau \tag{3.10}$$

Since the hyper-parameter ϑ is usually a positive number, the relation of α, β and τ is deduced as:

$$\alpha > (\tau^{-\frac{1}{\theta}} - 1)\beta \tag{3.11}$$

In MHEM, the general moderate factor τ reflects the degree of punishing the hard samples. τ should be properly set and the value should not too large. Meanwhile, τ depends on the datasets and the learning schedule, which is usually difficult to estimate. We discuss more details about the general moderate factor in the experiments. According to Equation 3.11 and the condition of $\tau > 1$, the necessary but not sufficient condition of α and β is $\alpha > -\beta$.

3.2.3 Moderate Modulation Baseline (M2B)

We present a moderate modulation baseline following the above relation of $\alpha > -\beta$. We empirically select the best function for the typical FGVC datasets, i.e., CUB-200-2011 [217], Stanford Cars [115] and FGVC-Aircraft [151]). We set β as 1. The best performance is achieved when α is set to -0.4, -0.8, and -0.7 on CUB-200-2011, Stanford Cars and FGVC-Aircraft, respectively.

The modulating function on CUB-200-2011 is defined as:

$$f_{CUB}(p) = (-0.4p + 1.0)^2 + 1 \tag{3.12}$$

The modulating function on Stanford Cars is defined as:

$$f_{CAR}(p) = (-0.8p + 1.0)^2 + 1 \tag{3.13}$$

The modulating function on FGVC-Aircraft is defined as:

$$f_{AIR}(p) = (-0.7p + 1.0)^2 + 1 \tag{3.14}$$

Though we only make minor modification for the loss function, significant improvements are achieved on all datasets. Our experimental evaluation proves that if the loss function satisfies the MHEM conditions, the networks trained with the loss function will achieve a better generalization ability.

Moreover, M2B is valuable for future research works on FGVC. It offers a simple way to search a good modulating function. We believe the practical value of M2B would further promote future research works about generalization and hard example modulation.

3.2.4 Discussion on Focal Loss

Focal Loss is one of the modulating functions but it does not satisfy MHEM conditions. As a useful hard mining loss, Focal Loss is widely applied in many works [34, 125, 162, 254]. Focal Loss re-weights the cross-entropy loss function with the modulating function of $(1-p)^{\vartheta}$. If ignoring the global bias term, our modulating function becomes the form of Focal Loss when $\alpha = -1$ and $\beta = 1$.

Focal Loss does not satisfy the MHEM condition (2) and condition (3). Supposing the modulating function as $f(p) = (1-p)^{\vartheta}$, the sensitive coefficient ρ is

$$\rho = \frac{(1-p_h)^{\vartheta-1}}{(1-p_e)^{\vartheta-1}}, \tag{3.15}$$

$$\lim_{\substack{p_e \to 1, \\ p_h \to 0}} \frac{(1-p_h)^{\vartheta-1}}{(1-p_e)^{\vartheta-1}} = \infty \tag{3.16}$$

Equation 3.16 may indicate that Focal Loss is more suitable for some datasets that are difficult to be fitted. In these datasets, more punishments to hard examples help the model to learn discriminative features. However, FGVC datasets contain a high inter-class similarity. On FGVC datasets, Focal Loss may impose too sensitive weights for hard examples. More punishments to hard examples may lead to severe overfitting [10, 44].

Our MHEM is more flexible than Focal Loss. MHEM introduces the adaptive slope (α) and the bias term (β). The hyper-parameter α and β are adjusted by the general moderate factor τ. However, Focal Loss can not be adaptively adjusted corresponding to different data distributions. Though ϑ can be changed, we find the variations of ϑ do not affect the performance much. This is because changing ϑ only does not enable Focal Loss with the moderate modulation property.

Empirically, we find Focal Loss does not bring improvements on FGVC datasets. In contrast, our simple moderate modulation baseline boosts the performances across all datasets.

3.3 Experiments

In this section, we experimentally evaluate the efficiency of MHEM and show the significant improvements from MHEM. First, the datasets and implementation details are introduced in Section 3.3.1 and 3.3.1. In Section 3.3.2, we compare our method with Focal Loss and discuss the validity of MHEM conditions. We compare our strong moderate

modulation baseline (M2B) with state-of-the-art methods in Section 3.3.3. We also implement M2B upon several state-of-the-art methods and validate its applicability. We provide sufficient ablation of M2B in Section 3.3.4. We provide the numerical comparisons of hard examples in Section 3.3.5. Finally, we offer the visualization of the features and hard examples in Section 3.3.6.

3.3.1 Experimental Setup

Datasets We evaluate MHEM based on three standard FGVC benchmarks, which are CUB-200-2011 (CUB) [217], Stanford Cars (CAR) [115] and FGVC-Aircraft (AIR) [151].

CUB-200-2011 (CUB) is the most widely-used dataset in FGVC. It contains 11,788 image samples and 200 categories. All images are various kinds of birds that are hard to be recognized even by a human being. It uses 5,994 samples for training and 5,794 samples for testing in experimental settings. Besides, AIR is constructed with various aircraft. This dataset includes 100 different categories of aircraft and has 10,000 images in total. Finally, as a fine-grained dataset for cars, CAR consists of 196 different kinds of cars within 16,185 images totally. There are 8,144 training images and 8,041 testing images.

Additionally, all datasets contain some meta information like location, attributes, brands, etc. We only use category labels and do not introduce any additional information.

Implementation Details

We do not introduce any overhead in both training and testing. Unlike the part-based methods [264, 277, 280], our moderate modulation baseline **does not** introduce additional learnable parameters. [58, 61] incorporate comprehensive data augmentation operations. In contrast, we only apply the typical data augmentation operations in FGVC, i.e., random rotation and randomly horizontal flip. We **do not** utilize additional data augmentations. Without extra learnable parameters and data augmentation, our method serves as a strong baseline for FGVC. It can be readily incorporated into the state-of-the-art methods and further improves the FGVC performance.

We train the proposed method M2B on the ResNet50 [88] backbone. The backbones are pre-trained on ImageNet [120]. The training scheme for M2B is identical to the baseline in [44]. In detail, we set the batch size as 16, and the initialized learning rate of the models is 0.0008. Meanwhile, the learning rate of the classifier layer is 0.008. The learning rate is decayed every 60 epoch. Overall, the models are trained 180 epoch. The optimizer is the Stochastic gradient descent (SGD) with a momentum value of 0.9 and

Table 3.1: Comparison of different values of ϑ. The Focal Loss does not show improvements over the baseline method. The baseline is trained with a standard cross-entropy loss.

Method	Dataset		
	CUB	CAR	AIR
$\vartheta = 1$	85.0	91.2	90.2
$\vartheta = 2$	85.8	91.0	90.2
$\vartheta = 3$	84.9	91.1	90.1
$\vartheta = 4$	84.6	91.1	89.9
$\vartheta = 5$	84.6	91.0	90.1
Baseline, $\vartheta = 0$	85.5	92.7	90.3

with a weight decay of 0.0001. In comparisons of recent methods [44, 61] with M2B, we follow the same implementations of those works, respectively.

3.3.2 The Effectiveness of MHEM conditions

Performances of Focal Loss: We evaluate the performance of Focal Loss on FGVC. We demonstrate that Focal Loss does not introduce any improvements on the FGVC task.

Based on the Focal Loss formulation of $FL(p) = -(1-p)^{\vartheta} log(p)$, we evaluate the classification performance by adjusting the value of ϑ. We experiment with five difference values of ϑ, i.e., $\vartheta = 1, 2, 3, 4, 5$. We compare the performances with a baseline that is trained with a standard cross-entropy loss. All results are provided in Table 3.1. The results reveal that merely adjusting ϑ does not improve the performances significantly. The MHEM conditions are not satisfied by changing ϑ.

Specifically, Focal Loss introduces larger punishments for the hard examples. In the special intra-class and inter-class variances in FGVC, the large punishments for hard examples lead the network to memorize those samples and overfit rapidly. This makes the Focal Loss achieves lower results in experiments. The variations of ϑ only influence the values of modulating function and do not change the relatively larger punishments for hard examples in Focal Loss.

In Table 3.1, we observe that adjusting ϑ of Focal Loss does not improve the standard cross-entropy loss. On CAR and AIR, the performances of Focal Loss are worse than the performances of the cross-entropy loss. For example, Focal Loss achieves the best performance on AIR [151] when $\vartheta = 1$ or $\vartheta = 2$, but the performance is still worse than the performance of the cross-entropy loss. On CAR, Focal Loss degenerates the performance by 1.5% compared to the cross-entropy loss, leading to severe overfitting. On CUB, $\vartheta = 2$ outperforms the baseline with only a small margin of 0.3%, while other values of ϑ

Table 3.2: Comparison for different conditions in hard and easy samples. Only results with lower weights for hard examples show improvements.

$f_c(\cdot)$	Hard ↑	Hard −	Hard ↓
Easy ↑	85.7	85.0	**88.2**
Easy −	84.8	85.0	**88.2**
Easy ↓	84.8	84.9	**88.1**

degenerate the performance. It it because that the adjustment of ϑ does not help Focal Loss to satisfy condition (3) of MHEM, as discussed in Section 3.2.4. Though ϑ adjusts the loss weight, it does not properly modulate hard examples for FGVC.

Additionally, to align the settings of MHEM, we perform all the experiments with the bias term in Table I. We also experiment with Focal Loss without the bias term, and there are no improvements. In CUB dataset, Focal Loss without bias term achieves 85.6 in the accuracy value when $\vartheta = 2$. The bias term does not significantly influence the performances.

The Effectiveness of Moderate Modulation: Since the Focal Loss [142] does not improve the results in FGVC, we validate if moderate modulation helps a better generalization. We showcase that considerable improvements can be obtained through piecewise combinations of modulating functions.

For the convenience of constructing modulating function, we choose Focal Loss with $\vartheta = 1$ as the baseline. We denote the modulating term of Focal Loss as $f_{FL}(p) = (1-p)$. The piece-wise combination of modulating functions $f_c(\cdot)$ is defined as:

$$f_c(p) = \begin{cases} f_1(p) & p \in (0, \mu] \\ f_2(p) & p \in (\mu, 1] \end{cases} \tag{3.17}$$

where f_1 and f_2 indicate the modulating functions for hard examples and easy examples, respectively. We define a threshold μ to differentiate hard examples and easy examples. We denote the range of hard examples as $(0, \mu]$ and the range of easy examples as $(\mu, 1]$.

We define the function of $f_c^{H\uparrow E\uparrow}(p)$ by introducing $f_1(p) > f_{FL}(p)$ and $f_2(p) > f_{FL}(p)$. The function of $f_c^{H\uparrow E-}(\cdot)$ indicates that $f_1(p) > f_{FL}(p)$ and $f_2(p) = f_{FL}(p)$. In this way, we can obtain nine modulating functions, i.e., $f_c^{H\uparrow E\uparrow}(p)$, $f_c^{H\uparrow E-}(p)$, $f_c^{H\uparrow E\downarrow}(p)$, $f_c^{H-E\uparrow}(p)$, $f_c^{H-E-}(p)$, $f_c^{H-E\downarrow}(p)$, $f_c^{H\downarrow E\uparrow}(p)$, $f_c^{H\downarrow E-}(p)$, $f_c^{H\downarrow E\downarrow}(p)$. Specifically, we design $f_c^{H\uparrow E\uparrow}(p) = 1-p/2$, $f_c^{H\uparrow E\downarrow}(p) = (1.21-p)^2$, $f_c^{H\downarrow E\uparrow}(p) = (0.95-0.5p)^2$ and $f_c^{H-E-}(p) = f_{FL}(p)$. $(1.21-p)^2$ as f_1, f_2, and f_2 of $f_c^{H\uparrow E-}$, $f_c^{H-E\downarrow}$ and $f_c^{H\downarrow E\downarrow}$, respectively. $1-p$ as f_2, f_1, f_1, and f_2 of $f_c^{H\downarrow E-}$, $f_c^{H-E\uparrow}$, $f_c^{H-E\downarrow}$ and $f_c^{H\downarrow E-}$, respectively. $(0.95-0.5p)^2$ as f_2, f_1, and f_1 of $f_c^{H-E\uparrow}$,

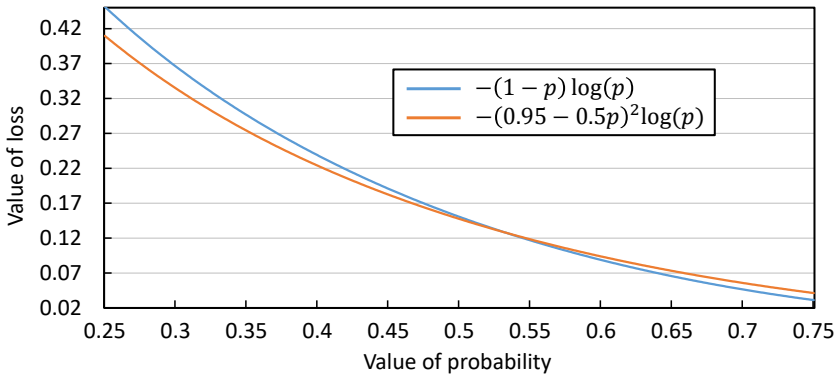


Figure 3.2: Curves of modulating functions of $(0.95-9.5p)^2$ and $1-p$. $(0.95-9.5p)^2$ presents lower punishments for hard examples and is used to formulate f_1, as a part of the piece-wise functions: $f_c^{H\downarrow E\uparrow}$, $f_c^{H\downarrow E-}$, and $f_c^{H\downarrow E\downarrow}$.

$f_c^{H\downarrow E-}$, and $f_c^{H\downarrow E\downarrow}$, respectively. We compare these modulating functions based on the classification performances on CUB [217].

In the aforementioned modulating functions, it can be validated that $f_c^{H\downarrow E\uparrow}(p)$, $f_c^{H\downarrow E-}(p)$, and $f_c^{H\downarrow E\downarrow}(p)$ satisfies the MHEM conditions. As in Table 3.2, only the three functions show significant improvements over other modulating functions. These results reveal that significant improvements are induced by reducing the weights for hard examples. Other changes do not satisfy the MHEM conditions and do not lead to any improvements.

To better clarify the difference between proper and improper modulating functions, we provide an example for the curve of MHEM and the Focal loss, as shown in Fig. 3.2. In this figure, two modulation functions are presented: $f_1(p)=(0.95-9.5p)^2$ and $f_2(p)=1-p$. These two functions correspond to the modulating function in Table II. f_1 is a part of the piece-wise functions: $f_c^{H\downarrow E\uparrow}$, $f_c^{H\downarrow E-}$, and $f_c^{H\downarrow E\downarrow}$, respectively. f_1 shows lower modulating values for hard examples with lower p. It provides slightly lower punishments for the hard examples, and all three piece-wise functions adapted f_1 for hard examples achieve higher performances in Table II. Meanwhile, f_2 is a part of piece-wise functions: $f_c^{H\downarrow E-}$, f_c^{H-E-}, and $f_c^{H\uparrow E-}$. f_1 satisfies MHEM conditions, but f_2 does not. In comparison, only $f_c^{H\downarrow E-}$ with lower punishments for hard examples obtain improvements. All the results prove that the moderate hard examples mining in our method leads to better performances.

Finally, the following conclusions can be deduced based on the results in Table 3.2 and Figure 3.2. 1) Moderate modulation is effective. Modulating functions following the MHEM conditions achieve the best performance compared to their counterparts. These

Table 3.3: Comparison of performances in three standard FGVC datasets. MA-CNN, PC, and Partial Assign use VGG19, DenseNet161, and ResNet101 as backbone networks respectively. Other methods are reported based on the results of ResNet50.

Methods	Overhead		Dataset		
	No Additional Augmentations	No Additional Parameters	CUB	CAR	AIR
MA-CNN [270]	✔	✘	86.5	92.8	89.9
PC [62]	✔	✘	86.2	92.9	89.2
DFL-CNN [226]	✔	✘	88.1	94.6	92.4
NTS-Net [245]	✔	✘	87.5	93.9	91.4
TASN [272]	✔	✘	87.9	93.8	-
ACNet [102]	✔	✘	88.1	94.6	92.4
Partial Assign [97]	✔	✘	87.7	-	-
BCN[94]	✔	✘	87.7	94.3	93.2
MC-Loss [33]	✔	✔	87.3	93.7	92.6
S3N [58]	✘	✘	88.5	94.7	92.8
LIO [277]	✔	✘	88.0	94.5	92.7
DCL[44]	✘	✔	87.8	94.5	93.0
M2B	✔	✔	**88.2**	**94.2**	**92.9**

modulating functions outperforms the typical cross-entropy loss by 2.6% in terms of recognition accuracy. 2) The performance degradation of Focal Loss is because of the unsuitable hard example mining. In $f_c^{H\downarrow E^-}$, only modifying the modulating function in hard examples induces a significant improvement of 3.2%. 3) The weighting of easy examples are not crucial for FGVC. Three variants of the modulating function, i.e., $f_c^{H\downarrow E^\uparrow}(p)$, $f_c^{H\downarrow E^-}(p)$, and $f_c^{H\downarrow E^\downarrow}(p)$, show negligible differences.

Table 3.4: Comparison of the original methods, methods augmented with Focal Loss, and methods enhanced with M2B. With the assistance of M2B, all methods demonstrate improvements in generalizing hard examples and achieve better performance. The numbers highlighted in blue indicate the performance gains relative to their corresponding baselines, whereas the numbers highlighted in red signify declines in performance compared to the baselines.

Method	CUB	CAR	AIR
ResNet50 Baseline [88]	85.5	92.7	90.3
ResNet50 + Focal Loss	85.8 (+0.3)	91.0 (-1.7)	90.2 (-0.1)
ResNet50 + M2B	**88.2** (+2.7)	**94.2** (+1.5)	**92.9** (+2.6)
DCL[44]	87.8	94.5	93.0
DCL + Focal Loss	87.7 (-0.1)	94.3 (-0.2)	92.9 (-0.1)
DCL + M2B	**88.5** (+0.7)	**94.7** (+0.2)	**93.3** (+0.3)
Single PMG[61]	88.9	95.0	92.8
Single PMG + Focal Loss	89.0 (+0.1)	95.0 (+0.0)	93.0 (+0.2)
Single PMG + M2B	**89.2** (+0.3)	**95.3** (+0.3)	**93.8** (+1.0)
Combine PMG[61]	89.6	95.1	93.4
Combine PMG + Focal Loss	89.5 (-0.1)	95.1 (+0.0)	93.0 (-0.4)
Combine PMG + M2B	**89.8** (+0.2)	**95.4** (+0.3)	**93.9** (+0.5)

3.3.3 Comparisons between M2B and the state-of-the-art methods

We compare our moderate modulation baseline (M2B) with the state-of-the-art methods. The results are shown in Table 3.3. Notably, M2B does not introduce any overhead in both training and testing. Without bells and whistles, M2B shows competitive performance to the state-of-the-art or even outperforms some recent approaches. The performances of M2B are competitive with LIO [277] and ACNet [102] among all three datasets. ACNet [102] introduces the tree attention and other networks, while M2B still outperforms the ACNet on CUB [217] and AIR [151]. DCL [44] introduces additional data augmentation in training. Our M2B uses the standard data augmentation and outperforms DCL with a clear gap of 0.4% in CUB dataset. MC-Loss [33] does not add any overhead to the baseline method, but its performances across all datasets are much lower than M2B. Compared with MC-Loss, M2B offers the improvements of 0.9% on CUB and 0.5% in CAR. The improvements are significant for fine-grained visual recognition tasks. These results clearly proves the importance of moderate hard example modulation.

Apply moderate modulation loss on state-of-the-art methods. We further combine various methods with M2B to show the effectiveness of MHEM. M2B can be easily applied to various methods and only the loss function needs to be replaced. The experimental results are shown as in Table 3.4.

M2B not only improves the performance of the ResNet50 baseline method [88] trained with cross-entropy loss, but also improves the performances of DCL [44] and PMG [61] on all three datasets. PMG and DCL have achieved considerable improvements compared to the ResNet50 baseline, benefitting from well-designed extra networks or additional augmentations. We can obtain further gain with M2B. For DCL, 0.7%, 0.2%, and 0.3% improvements are achieved on CUB, CAR, and AIR, respectively, by adapting M2B. For PMG, M2B achieves 1.0% gain on AIR, which is significant for the FGVC task. To better compare the performance of PMG, we conduct evaluations on *Single PMG* and *Combine PMG*. *Single PMG* only uses one stage in PMG. *Combine PMG* considers all PMG branches. M2B improves the performance of both *Single PMG* and *Combine PMG*. The improvements reveal that MHEM is general to various pipeline and useful to boost performances. We also test Focal Loss [142] with various methods. As shown in Table 3.4, Focal Loss does not improve the performances of DCL and PMG. The results validate that Focal Loss does not handle hard samples properly. Our M2B provides an efficient baseline to help various methods to overcome the problem of overfitting. M2B prevents

Table 3.5: Comparison of the parameter ϑ in different conditions. ϑ related to the strength of modulation. The suitable value of ϑ can help the M2B achieve better performances.

ϑ	0.5	1.0	2.0	5.0
Accuracy (%)	87.2	87.5	**88.2**	86.7

the methods from overfitting and allows them to achieve a better generalization. We believe MHEM and M2B can promote the research in further analyzing hard examples on FGVC datasets.

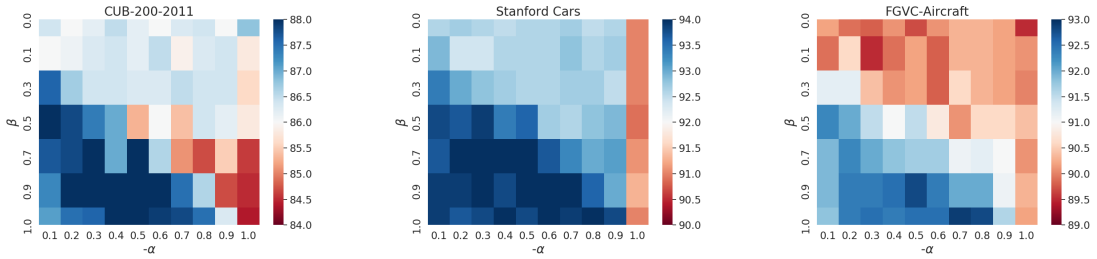


Figure 3.3: Ablation study for different α and β in three datasets. Results corresponding to $\alpha > -\beta$ are better than those dissatisfied.

3.3.4 Ablation Study for M2B

Ablation of α and β: We first compare different combinations of α and β in the second-order polynomial form in three datasets. All the results are displayed in Fig. 3.3. Since different datasets contain different properties like distributions of hard examples, numbers of classes, etc., the changing of factors leads to fluctuations in the performances of MHEM. In Figure 2, the method shows better performances in all datasets if satisfying the MHEM conditions, which reveals that MHEM is general to various datasets and provides universal improvements to various benchmarks in FGVC.

Specifically, as shown in Fig. 3.3, the combinations of α and β influence the performances. Only if the combinations follow the MHEM, significant improvements will be obtained. Among various combination, we achieve the best performance of 88.2% when $\alpha = -0.4$ and $\beta = 0.9$ in CUB [217]. In CAR [115], the best result occurs when $\alpha = -0.8$ and $\beta = 1.0$. For AIR [151], the combination of $\alpha = -0.8$ and $\beta = 1.0$ is best. We choose the best parameters for the implementation.

Based on the moderate condition described in Section 3.2.1, we derive the relation $\alpha > -\beta$. In Fig. 3.3, we obtain performance improvements from all combinations following the relation. Due to the fluctuation of training, some results dissatisfied MHEM may

also achieve slight improvements. However, all combinations of α and β that achieve significant improvements satisfy the condition of $\alpha > -\beta$. These results indicate the effectiveness of the moderate condition. The experimental results validate that MHEM is feasible in training a good model. All results with a proper combination of α and β show improvements. The condition of MHEM limits the range of hyper-parameters and provides an easier way to construct effective modulating functions.

Meanwhile, the highest performances occur in different combinations of α and β in different datasets. Considering the simplification of M2B, a more complicated modulating function with better performances may exist. This indicates that more MHEM conditions remain to be further explored. Our work aims to present the efficiency of MHEM and points out a simple solution to hard example modulation. We believe that a deeper investigation can be performed based on MHEM.

Comparisons of ϑ: We investigate the performance of M2B when ϑ is changed. We compare the performance of different ϑ on CUB. The results are shown in Table 3.5.

When $\vartheta = 0$, no modulation is applied during training and the method degrades to the typical cross-entropy loss. When $\vartheta > 0$, the modulating function works during training and better results are obtained. However, the larger value of ϑ induces the lower value of $f(\cdot)$. This leads $f(p) \to 1$ and degrades the MHEM to the standard cross-entropy. Thus, the value of ϑ should also be noticed in method designs. Compared with the results, the best performance of 88.2% appears when $\vartheta = 2$. We apply this setting in our M2B.

Besides, the larger ϑ leads to smaller values of modulating function. This would reduce the influences of MHEM and produces lower results. Meanwhile, though slight lower performances occur in the large ϑ, the result also outperforms the baseline method with a large gap of 1.2% in the accuracy. These results also reveal the robustness of MHEM.

Discussion about τ: Based on M2B, the value of τ can be estimated. However, various points can influence it (e.g., probability of hard examples, distribution of datasets, training strategy of networks, etc.). These points are hard to be quantified since most of them are uncertain during training. Thus, it may be challenging to deduce the values of τ directly. In our work, we estimate the range of τ by experiments rather than inferring the exact values. For convenience, we suppose the probability of a hard example is $p_h = 0.1$ and the probability of an easy example is $p_e = 0.9$. By simple calculation, we obtain the values of τ on CUB, CAR, and AIR:

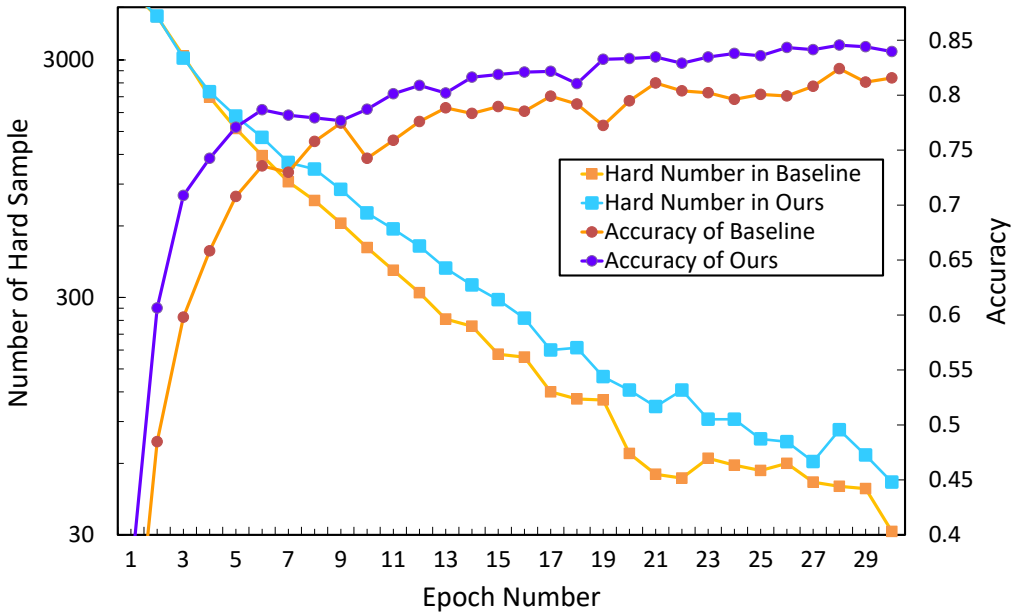


Figure 3.4: Comparison of the numbers of hard number and accuracy. We use the first 30 epochs here and suppose that the confidences of hard examples are lower than 0.5. The vertical axis indicates the epoch number. The horizontal axis indicates the values of numbers of hard examples and accuracy.

$$\begin{aligned} \tau_{CUB} &> 1.640 \\ \tau_{CAR} &> 1.764 \\ \tau_{AIR} &> 1.712 \end{aligned} \tag{3.18}$$

Thus, the comparison of the lower bound of the general moderate factor τ in different datasets is CAR > AIR > CUB. This may indicate that the CAR has a larger range for adapting hard example modulation. In other words, it contains more tolerance to various effects of hard example modulation. Interestingly, the rank of accuracy is also: CAR > AIR > CUB. CAR is easier than the others and shows lower requirements for moderate modulation.

3.3.5 Numerical Comparison of Hard Examples

In this section, we provide more evidence to show that the MHEM and M2B improve the learning of hard examples. We assume that the confidences of hard examples are lower than 0.5. We intuitively show the advantages of our method in handling hard examples by comparing the numbers of hard examples during training. We conduct all

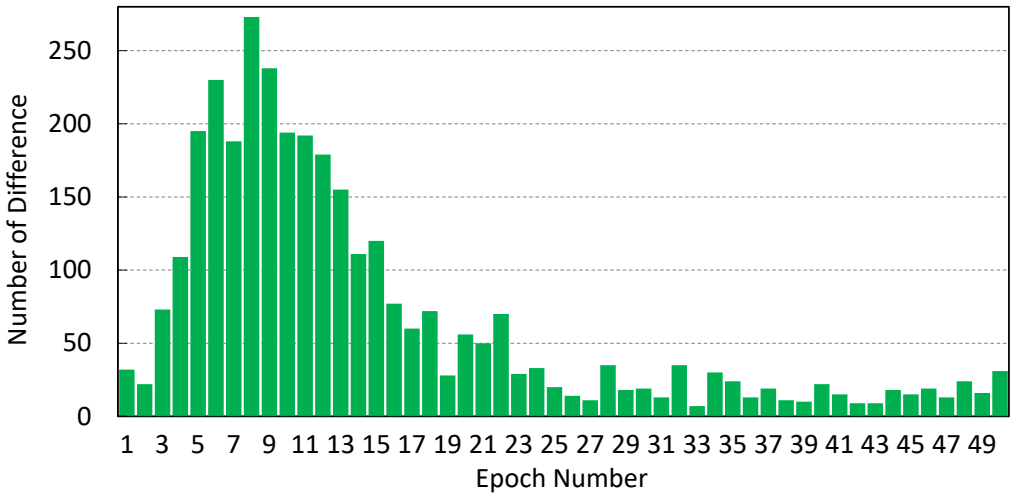


Figure 3.5: Differences of the numbers of the hard examples in the **first** 50 epochs. The horizontal axis indicates the epoch number and the vertical indicates the value of differences between M2B and the baseline.

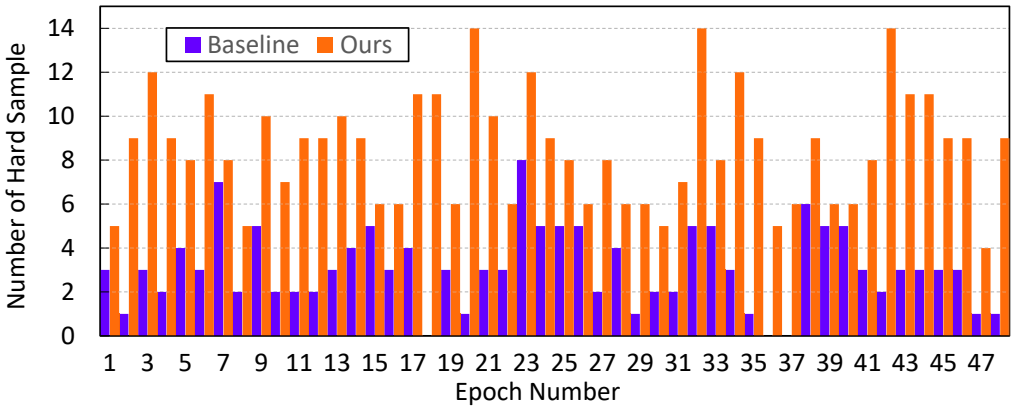


Figure 3.6: Comparison of the numbers of the hard examples in the **last** 50 epochs. The horizontal and vertical axis stand for the epoch numbers and numbers of hard examples, respectively.

comparisons on the CUB dataset. Compared to the baseline, more hard examples remain for M2B as shown in Fig. 3.5 and Fig. 3.6. The MHEM conducts moderate hard example modulation and encourages the models to maintain more hard examples in training.

The hard examples can be rapidly memorized during training and may not be thoroughly learned by the network. These samples may contain particular properties not generalized by the current network. The hard examples may include valuable information to help the networks be more generalized. In our work, MHEM reserves more hard examples for training. This makes the network more generalized and achieves better

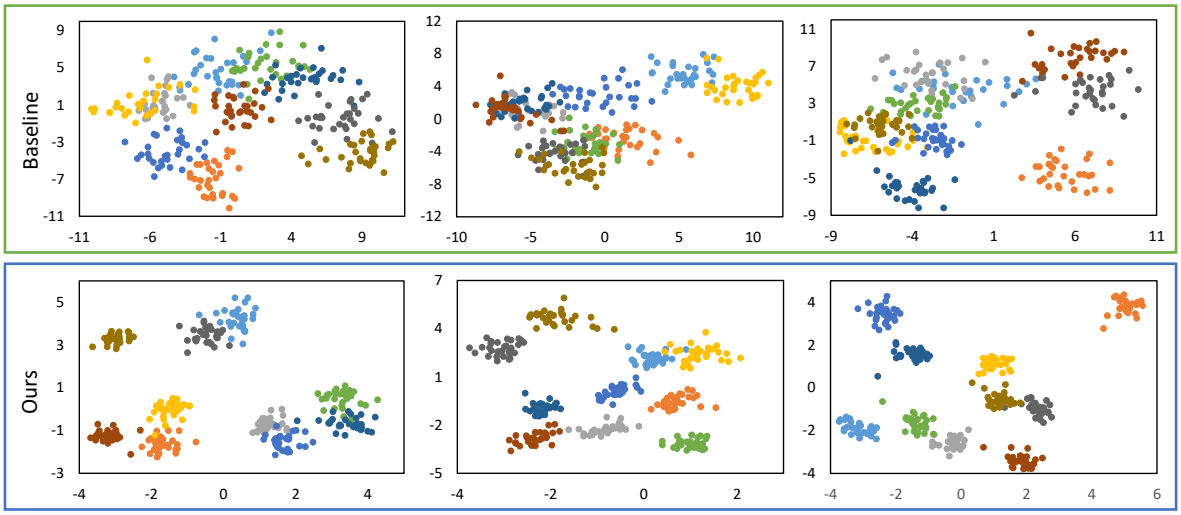


Figure 3.7: The features from the last epoch in training are visualized. We sample 10 classes randomly from CUB [217] and process the corresponding features with PCA [68]. Three different sets of classes are visualized. Figures in each column contains the same classes.

results in experiments. In detail, at the beginning of the training procedure, more hard examples are reserved in MHEM. We compare the difference of the numbers of hard examples in the first 50 epochs. The results are shown in Fig. 3.5. The baseline method fits the hard examples rapidly. The differences increases in the first ten epochs. Our method retains more hard examples for the network. Fig. 3.6 represents the comparison of the number of hard examples between our method and the baseline in the last 50 epochs. When the training comes to the last 50 epochs, the number of hard examples in our M2B is still larger than the baseline.

We compare with the accuracy and the number of hard examples between our M2B and the baseline in Fig. 3.4. M2B reserves more hard examples for training and achieves better accuracy. In the first 30 epochs, the performance gap appears and our method outperforms the baseline method. With further training, M2B achieves significantly better performance than the baseline method. This comparison reveals two points. 1) M2B prevents the network from overfitting the hard examples. More hard examples are reserved in training. 2) By avoiding overfitting the hard examples, the network achieves better performances. The moderate modulation is helpful. Preventing memorizing the hard examples enables the model to be more generalized.

3.3.6 Visual Analysis

Feature Visualization: We represent the feature visualization results in the test set in Fig. 3.7. In each figure, we random sample 10 classes from the 200 classes in CUB [217] and apply PCA [68] to reduce the dimension of features. We select three different sets of classes for visualization. Each column visualizes features from the same set of classes. The first row presents feature visualizations from the baseline method. The second row represents features from our M2B.

The baseline method and our method are trained in the same epochs. The features with M2B become more discriminative. Compared with the baseline and our method, the feature distributions of our method are more concentrated. This shows our method learns discriminative features of unseen samples from the test set. Specifically, MHEM leads the network to avoid overfitting all the examples, keep learning the hard examples and pursue more generalized representation in training. Without MHEM, the baseline method rapidly overfits all training examples and stucks in indiscriminate representation.

Meanwhile, the features of the baseline are mostly located in the range of $[-10, 10]$. In comparison, our features are located in the range of $[-4, 5]$. The distances reveal that the intra-class variances are lower in M2B. The network with MHEM generates more discriminative features and shows compact class distributions.

Example Visualization: We showcase some samples in this section. All examples (Fig. 3.8) are from the class of 'Tern', which is hard to be solved as pointed in [10]. The samples from the training set are on the left side. The samples from the test set are on the right side. All training samples are correctly recognized within 20 epochs. All test samples are misclassified by the baseline method and correctly recognized by our method. In this comparison, our M2B helps the network to become more generalized.

As shown in Fig. 3.8, the images in the training set are hard to be categorized even for a human. However, the network easily recognizes all the training cases and shows high confidence. This indicates that the common training with cross-entropy leads the network to overfit. In contrast, our moderate modulation takes advantage of learning generalized information in the class. Our M2B correctly classifies all test examples in Fig. 3.8. The presented cases show that MHEM assists the network to better handle the hard examples in training and learn more general features from the data.

Figure 3.8: Examples correctly classified by M2B and misclassified by the baseline are presented. Class 'Tern' is a hard category as in [10]. Though training examples in the left side are hardly recognized even by human, the network easily overfit and misclassifies the test samples in the right side. With MHEM, the network learns to generalize and rectify examples in test.

3.4 Conclusion and Discussion

In this chapter, we aim to improve the ability of networks to be generalized through the proper hard example mining. We propose Moderate Hard Example Modulation (MHEM) to handle the samples in training better. To operate the MHEM, we discuss three conditions and further deduce the relation of parameters. Via the deduced relationship, a powerful and efficient baseline method named Moderate Modulation Baseline method (M2B) is presented. Following the MHEM, M2B leads the network to learn samples better rather than memorizing them. In experiments, our method boosts the performances in all standard datasets. The single backbone with M2B outperforms many current methods, and the current state-of-the-art with M2B achieves further improvements. Our method is a component to be a solid baseline in FGVC.

For generalizability, MHEM formulates a simple and general solution for moderate

hard example mining in FGVC. MHEM provides a polynomial constraint between the moderate punishments and the probabilities. Then, the moderate factor controls the relation and ensures that the punishments are leveraged appropriately. This straightforward polynomial relation makes our MHEM general in various FGVC benchmarks. Meanwhile, the simple relation also implies the extensibility of MHEM. Future works can explore more adaptive or model-specific moderate punishments than MHEM. More flexible or dynamic constraints can be designed. We hope our MHEM can inspire deeper explorations and diverse extensions for the proper hard example mining. This may push the community to produce more thought collision for the relationships between data samples and adequate punishments.

CHAPTER 4

TOWARD BETTER ACCURACY FOR GENERAL VISUAL CLASSIFICATION TASKS

4.1 Introduction

Deep neural networks have achieved impressive improvements in visual recognition. The neural networks trained on large-scale visual recognition datasets, e.g., ImageNet [120], OpenImages [113], demonstrate remarkable generalization capabilities. The learned visual representations are compact and enjoy strong discriminability. Many works have been conducted to theoretically explain the rationale behind deep networks' generalization [257], but this problem is still largely unsolved and remains to be investigated.

There are a few analytical tools to probe deep neural networks' learning and generalization capabilities. Early works utilize visualization tools to understand the optimized parameters or employ dimensionality reduction techniques to visualize the quality of learned representations [193, 211, 256]. Though helpful, such visualization techniques only provide qualitative inspections on deep networks [31]. Some works develop geometric probes to analyze the geometric properties of object manifold and connect object category manifolds' linear separability with the underlying geometric properties [204]. These methods reveal the structure of memorization from different layers in deep networks but only probe layer capacity at the inference time, as shown in Fig. 4.1 (a).

Another simple strategy is to perform linear probing. One can use linear probes to

This chapter is based on joint work [135] with Linchao Zhu, Xiaohan Wang, and Yi Yang, presented primarily as it appears in the CVPR 2022 proceedings.

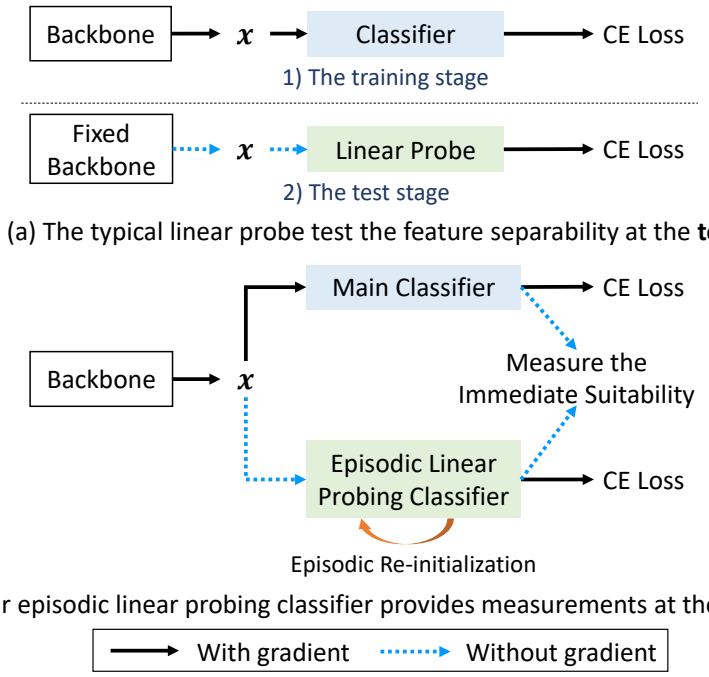


Figure 4.1: The typical linear probe in testing (a) and our ELP in training (b). Our ELP is episodically re-initialized to maintain simplicity. It effectively measures the discrimination of visual representations in an online manner.

evaluate the feature's quality quantitatively. Since the discrimination capability of linear classifiers is low, linear classifiers heavily rely on the quality of the input representation to obtain good classification accuracy [13]. Alain *et al*. [5] use linear probes to examine the dynamics of intermediate layers. The linear probe is a linear classifier taking layer activations as inputs and measuring the discriminability of the networks. This linear probe does not affect the training procedure of the model. Recently, linear probes [13] have been used to evaluate feature generalization in self-supervised visual representation learning. After representation pre-training on pretext tasks [13], the learned feature extractor is kept fixed. The linear probe classifier is trained on top of the pre-trained feature representations. Though conceptually straightforward, linear probes are effective and have been widely used in measuring the discriminability of visual representations. Noticeably, the linear probing classifier is only used in testing. A natural question arises: can we utilize linear probes during training and bring the signal from the linear probes to regularize the model training?

In this chapter, we introduce a simple strategy to regularize the network to be immediately plausible for an episodic linear probing classifier. Our simple framework (Fig. 4.1 (b)) consists of a main classifier, an episodic linear probing classifier, and a

regularization term. The regularization term considers the relation between the main classifier and the episodic linear probing classifier, which effectively penalizes examples that are not immediately plausible for episodic linear probes.

First, we propose an episodic linear probing (ELP) classifier to estimate the discrimination of visual representation in an online way. Similar to the existing linear probes [5], ELP is applied on top of the last layer of a deep network. ELP classifier is trained to classify the detached features into the same label space as a regular classifier. Different from [5], ELP is applied during model training. It is episodically re-initialized at each epoch. This maintains its simplicity, avoids classifier overfitting, and prevents the classifier from memorizing features. ELP implicitly reflects the feature discriminability and separability [190, 192]. If the ELP classifier can quickly classify the feature points, it indicates that the given features are easily separable and would potentially be more generalizable.

Second, we introduce a penalization for less suitable examples for an episodic linear probe. Intuitively, given a training example, if the episodic linear probe and the main classifier contradict each other, *e.g.*, the episodic linear probe receives a *low* prediction score while the main classifier produces a *high* prediction score, it indicates that the main network exhibits overfitting on the given instance and a larger penalty should be enforced for proper regularization. Thus we design an ELP-suitable Regularization term (ELP-SR) to mitigate the intrinsic model bias and improve the linear separability of the learned features. ELP-SR sets a re-scaling factor to each instance and adaptively modulates the cross-entropy loss to avoid overfitting. The re-scaling factor considers the deviation between an example's predictive score from the main classifier and ELP classifier, which, to a certain extent, assesses the example's suitability for linear classification.

Without bells and whistles, our method achieves significant improvements for visual recognition tasks in the wild, providing consistent gains for fine-grained, long-tailed, and generic visual recognition. The fine-grained visual recognition datasets often contain high inter-class similarities. The long-tailed visual recognition datasets exhibit long-tailed data distribution, which is realistic in real-world recognition problems. We extensively evaluate the generalization performance on six standard datasets. The results indicate that our strategy empowers various deep networks with better discrimination and mitigates the model bias.

4.2 Method

In this work, we introduce an auxiliary episodic linear probing classifier to provide additional regularization for better representation learning. As illustrated in Fig. 4.2, our framework consists of three components, i.e., a deep neural network, a main linear classifier, and an episodic linear probing classifier. We illustrate our episodic linear probing classifier in Section 4.2.1. The details of the ELP-suitable regularization are introduced in Section 4.2.2. In Section 4.2.3, we describe the training and inference strategies of the model.

4.2.1 Episodic Linear Probing Classifier

4.2.1.1 Review of The Typical Linear Probes

Training the Feature Extractor. Given a training sample $\boldsymbol{x}$, a neural network (F) extracts its feature $\boldsymbol{h}$. A linear classifier (Cls) projects the feature to a probability distribution $\boldsymbol{p}$. The cross-entropy (CE) loss calculates the cross-entropy between $\boldsymbol{p}$ and the ground-truth distribution $\boldsymbol{y}$. Formally, we denote the typical training procedure below:

$$\boldsymbol{h} = F(\boldsymbol{x}), \tag{4.1}$$

$$\boldsymbol{p} = Cls(\boldsymbol{h}), \tag{4.2}$$

$$\ell_{ce}(\boldsymbol{p}, \boldsymbol{y}) = -\sum_{j=1}^{C} y^j \log(p^j), \tag{4.3}$$

where C is the number of categories. $y^j = 1$ if j is the ground-truth label. Otherwise, $y^j = 0$. p^j is the prediction score of class j. The feature extractor and the classifier are jointly optimized end-to-end using back-propagation.

Test-time Linear Probing. Linear probing is usually built to assess the quality of deep representations after the neural network is sufficiently trained [5]. That amounts to training an auxiliary linear classifier on top of the pre-trained features. The parameters of the linear probe are randomly initialized, while the original classifier layer is neglected. The pre-trained backbone is frozen and not trained during linear probing. Since the complexity of the auxiliary classifier is not sufficient to provide additional discrimination, the classification performance heavily depends on the quality of the feature representations. Thus, predictive scores of the auxiliary linear classifier can probe the discrimination of the input features. During implementation, a linear probe

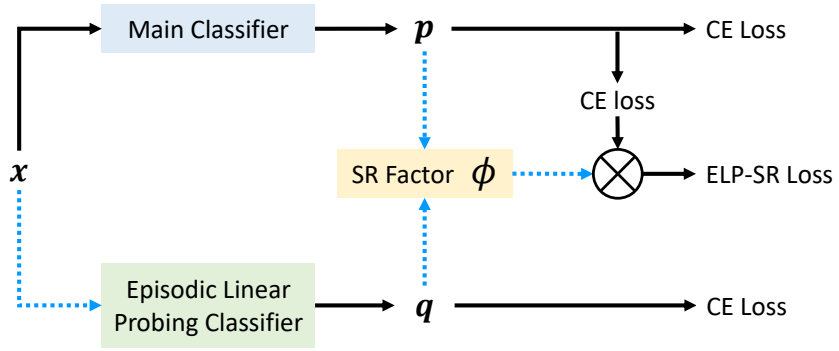


Figure 4.2: The training flow of our framework. Black lines indicate that the gradient can be back-propagated, while the blue dotted lines indicate that the gradient back-propagation is stopped.

can be extended to a Multi-Layer Perceptron (MLP) probe where the linear layer is replaced with a MLP [91].

The existing probes are mainly used during inference time, either providing quantitative evaluation on pre-trained features or interpreting intermediate layers [56]. This drives us to incorporate a linear probe during training and borrow the simple nature of the linear probe for network regularization.

4.2.1.2 Episodic Linear Probing Classifier

Motivated by the efficacy of test-time linear probe in assessing representation quality, we aim to design a linear probing classifier in training to measure the discrimination of a neural network and further leverage the probing signal to empower representation learning. We introduce an episodic linear probing (ELP) classifier and discuss its weight update scheme in training.

Detached Linear Probing Classifier in Training. When incorporating a linear probing classifier in training, we need to maintain its independence from the main classifier. While keeping the main classifier and the backbone network unchanged, we build a new episodic linear probing classifier on top of the feature extractor. We stop the linear probe classifier's gradient to back-propagate to the backbone network. This helps the linear probe not be biased by the main classifier and produce a neutral evaluation of the discrimination of the feature representations.

Formally, the episodic linear probing classifier is trained to classify the features into

C categories using the same labels assigned to the main classifier,

$$\boldsymbol{p} = Cls_{\text{main}}(\boldsymbol{h}), \tag{4.4}$$

$$\boldsymbol{q} = Cls_{\text{elp}}(\text{stop-grad}(\boldsymbol{h})), \tag{4.5}$$

$$\ell_{\text{main}}(\boldsymbol{x}, \boldsymbol{y}) = \ell_{ce}(\boldsymbol{p}, \boldsymbol{y}), \tag{4.6}$$

$$\ell_{\text{elp}}(\boldsymbol{x}, \boldsymbol{y}) = \ell_{ce}(\boldsymbol{q}, \boldsymbol{y}). \tag{4.7}$$

Cls_{main} is the main classifier, and it produces a probability prediction of $\boldsymbol{p}$. Cls_{elp} is the linear probe classifier, and it generates a probability prediction of $\boldsymbol{q}$. Cls_{elp} is trained in an online manner, but its optimization is detached from the main branch. "stop-grad" indicates that feature $\boldsymbol{h}$ is detached to train Cls_{elp}. The gradients from the ELP classifier are unavailable to the backbone and main classifier, and vice versa. The main difference between the detached linear classifier and the test-time linear probe is that the features of the detached linear classifier are adaptively changed by the network, while the features of the test-time linear classifier are always fixed.

Episodic weight re-initialization overcomes overfitting. Training the detached linear classifier with the same number of epochs as the main classifier would lead to the detached linear classifier overfitting the features. This overfitting should be avoided because the simple linear probe is supposed to reflect the discrimination of the features. If the ELP classifier memorizes all samples, it would not be competent to evaluate the features effectively. To prevent the ELP classifier from overfitting the training data, we re-initialize its parameters episodically every $\mathscr{I}$ epochs where $\mathscr{I}$ indicates episodic re-initialization interval. Specifically, given a linear classifier parameterized with W and b, where W is the projection matrix, and b is the bias, both W and b are randomly re-initialized at the interval of $\mathscr{I}$ epochs.

The episodic linear probe enables us to measure and understand the feature discriminability throughout the training process. A larger value of $\mathscr{I}$ enforces the ELP classifier to be better trained, but it makes the ELP classifier more likely to be overfitted. In contrast, the ELP classifier is under-fitted, if $\mathscr{I}$ is too small. An under-fitted ELP classifier may not well describe the generalization capabilities of the features. In practice, we set $\mathscr{I}$ as a hyper-parameter. Empirically, $\mathscr{I} = 2$ achieves consistent good probing performances across datasets.

4.2.2 The ELP-Suitable Regularization

ELP-Suitable Regularization through loss modulation. ELP assesses the features' separability in an online way. The standalone ELP is detached from the backbone and

does not influence the main network. In this chapter, we aim to utilize the prediction from the auxiliary ELP classifier to effectively improve the discriminability of the main branch. However, the design of this regularization is not straightforward. Considering the episodic nature of the ELP classifier, ELP's prediction is periodic and not as confident as the main classifier. If the regularization is not well constructed, the performance of the main branch would be severely impaired.

In this chapter, we introduce a simple formulation that modulates the cross-entropy loss with an adaptive factor ϕ,

$$\mathscr{L}_{ELP-SR} = \sum_{i=1}^{B} \text{stop-grad}(\phi_i) * \ell_{ce}(\boldsymbol{p}_i, \boldsymbol{y}_i), \tag{4.8}$$

where $\boldsymbol{p}_i$ is the prediction probability from the main classifier, B is the batch size. The scalar factor ϕ_i is assigned to each instance to modulate its cross-entropy loss adaptively. ϕ measures the main network's suitability for an ELP classifier. If an instance is not *suitable* for the ELP classifier, *e.g.*, the instance may be not discriminative, or an out-of-distribution data point, ϕ imposes a relatively large value so that the network would pay more attention to this instance. Our ELP-Suitable Regularization (ELP-SR) effectively mitigates the intrinsic model bias and regularizes the network towards better linear separability.

We detach the gradients from ϕ so that the factor only influences the magnitude of the loss gradients, but the gradient orientation is not altered. This makes the optimization progress relatively easy and stable. The strategy works surprisingly well in practice.

The instantiation of the ELP-SR factor. As aforementioned, ϕ aims to measure the main network's suitability for an ELP classifier. Given an instance $\boldsymbol{x}$ with the label c, we instantiate the ELP-SR factor by considering the prediction score of the main classifier (p^c) and the prediction of the ELP classifier (q^c). We utilize two elements when we construct the regularization factor ϕ.

First, the distance metric (D) between the prediction of the ELP classifier and the prediction of the main classifier should be concerned. The distance should reflect the main classifier's confidence gap compared to the ELP classifier. If the distance is minimized, the main classifier is pushed to act like a less-trained linear classifier. Relatively, The features would be remarkably discriminative if a less-trained classifier is already sufficient for recognizing. Therefore, this metric encourages the main classifier to become simpler, promoting the features to be more discriminative. We instantiate D by simply computing the ℓ_1 distance between p^c and q^c, *i.e.*, $D = |p^c - q^c|$.

Second, we incorporate a normalization metric (R) to reveal the discriminability of both the ELP classifier and the main classifier. The distance metric (D) measures the relative confidence gap, but we should also consider the absolute values of the confidence scores. If the distance between p^c and q^c is small, but both absolute scores are low, the network has not been well optimized to classify the instance. Thus, we should normalize the distance with a normalization metric. For simplicity, we set R as the average of p^c and q^c, *i.e.*, $R = (p^c + q^c)/2$.

We formulate the ELP-SR factor ϕ as,

$$\phi = (\frac{D}{R})^{\gamma} = (\frac{2|p^c - q^c|}{p^c + q^c})^{\gamma}, \tag{4.9}$$

where γ smoothly adjusts the rate between D and R. We empirically study other ELP-SR factor variants in the experiment section.

4.2.3 Training and inference

In the training phase, we calculate the softmax cross-entropy loss for both the main classifier and the ELP classifier. Our ELP-SR loss is summed with these losses. The overall training objective is below,

$$\mathscr{L} = \sum_{i=1}^{B} \ell_{\text{main}}(\boldsymbol{p}_i, \boldsymbol{y}_i) + \ell_{\text{elp}}(\boldsymbol{q}_i, \boldsymbol{y}_i) + \phi_i * \ell_{ce}(\boldsymbol{p}_i, \boldsymbol{y}_i) \tag{4.10}$$

In the test phase, we remove the auxiliary ELP classifier and only keep the main classifier. The final prediction is obtained only from the main classifier. Our framework does not introduce any additional overhead during testing.

4.3 Experiments

In the challenges of diverse objects of images in the wild, our method shows significant superiority for generalization. We evaluate three classification tasks, *i.e.*, fine-grained visual recognition, long-tailed recognition, and generic object recognition. First, since the classes in fine-grained recognition are similar, and samples are difficult to be recognized even by humans, the fine-grained recognition task brings extra challenges to learning discriminative features. Second, long-tailed recogni tion involves the extremely imbalanced distributions of data samples. This requests the method to possess generalization ability and recognize the tailed classes with limited samples. The evaluations of these tasks reveal the advantages of our method in improving visual representations.

We further evaluate our method on ImageNet-1K to study the generalization ability of ELP-SR. Besides the classification accuracy metric, we also report the results of a k-nearest-neighbor (KNN) classifier on the test set. This further manifests the effectiveness of our method in improving the discriminability of feature representations. Moreover, we provide ablation studies to compare different γ, $\mathscr{I}$, and formulations of the ELP-SR factor. To further demonstrate the ability of the ELP classifier, we present a comparison of the linear classifier's accuracy. The results reflect that the network with ELP-SR produces more discriminative and generalized features.

To be noticed, for all the tasks, we did **NOT** introduce any additional annotations nor incorporate extra parameters at the inference time. During testing, **only the backbone networks** are used to produce predictions.

4.3.1 Experimental Setup

Classes in fine-grained recognition are similar. They are difficult to distinguish, even for a human. Meanwhile, samples in every class are diverse [10]. Objects may be shown in various angles, illuminations, occlusions, backgrounds, etc. These induce fine-grained categories to show large intra-class variances, but small inter-class variances [10]. Samples in fine-grained classification are hard to be generalized and discriminated, which brings difficulties for learning discriminative features by networks.

Dataset and Implementation Details. To show the efficacy, we compare the performances on three standard benchmarks: CUB-200-2011 (CUB) [217], Stanford Cars (CAR) [115], and FGVC-Aircarft (AIR) [151].

Following the same training procedure in [44], we adapt ResNet-50 [88] pre-trained by ImageNet [120] as the backbone model. As the regular augmentations [44, 61, 277] in this task, resizing, random crops, rotations, and horizontal flips are applied. After operating these standard transformations, the final inputs become 448×448 resolutions. Similar to the ResNet50 baseline [44, 277], we train our method for 240 epochs and optimize the loss function by SGD. In our method, we report the results of $\gamma = 3$ for all three datasets with $D = p^c - q^c$ and $R = (p^c + q^c)/2$. For CUB, CAR, and AIR, we set $\mathscr{I} = 2$, 2, and 1, respectively. These are the best settings for parameters and will be discussed in the ablation section 4.3.4.

Experimental Results. As in Table 4.1, our method achieves significant improvements based on the ResNet50 baseline. Without bells and whistles, our results are competitive or even outperform many recent methods with complicated network designs [102], additional augmentations [44, 61], or multi-scale features [61, 277]. Merely utilizing naive

Method	Dataset		
	CUB	CAR	AIR
B-CNN [144]	84.1	91.3	84.1
HIHCA [28]	85.3	91.7	88.3
RA-CNN [69]	85.3	92.5	88.2
OPAM [172]	85.8	92.2	-
Kernel-Pooling [49]	84.7	91.1	85.7
MA-CNN [270]	86.5	92.8	89.9
MAMC [206]	86.5	93.0	-
HBP [251]	87.1	93.7	90.3
DFL-CNN [226]	87.4	93.1	91.7
NTS-Net [245]	87.5	93.9	91.4
DCL [44]	87.8	94.5	93.0
PMG [61]	88.9	95.0	92.8
ACNet [102]	88.1	94.6	92.5
LIO [277]	88.0	94.5	92.7
ResNet50 Baseline	85.5	92.7	90.3
ResNet50 Baseline + ELP-SR	88.8	94.2	92.7

Table 4.1: Comparison of three benchmarks of fine-grained classification. Without additional augmentations or network designs, our method achieves significant improvements.

backbone with ELP-SR in training, the simple backbone networks boost 3.3%, 1.5%, and 2.4% respectively in three datasets which are significant improvements in this task. Boosts in this task reveal that our method effectively improves the networks' ability to discriminate and generalize samples. To further manifest the superiority of our method, more discussions will be presented in 4.3.4.

4.3.2 Long-tailed Visual Recognition

In long-tail recognition, the data distributions of different classes show extreme imbalance. As the long-tailed distribution, a handful of 'head' classes contain considerable samples, but a large number of 'tail' classes only include limited samples. The networks are biased toward 'head' classes, and the samples in 'tail' classes are hard to be generalized. In this section, we also evaluate the performances of our method under the challenging long-tailed distribution.

Dataset and Implementation Details. The experiments are operated based on long-tailed CIFAR-10 and CIFAR-100 datasets [119]. We first produce several versions of long-tailed datasets following [29] under different imbalance ratios, which denotes the ratio between the largest and smallest numbers of samples in classes. We report the

Method	CIFAR-10			CIFAR-100		
Imbalance ratio	100	50	10	100	50	10
Focal Loss [142]	70.4	76.7	86.7	38.3	43.9	55.7
CB Focal [48]	74.6	79.3	87.1	39.6	45.2	58.0
Meta-weight [199]	75.2	80.0	87.8	42.0	46.7	58.4
CDB-CE [201]	-	-	-	42.5	46.7	58.7
Mixup [263]	73.1	77.8	88.3	39.6	45.0	58.2
ERM [29]	70.4	74.8	86.4	38.3	43.9	55.7
ERM [29] + ELP-SR	77.4	81.2	87.9	39.1	44.7	57.9
ERM [29] + ELP-SR ($\tau = 1$)	77.5	81.5	88.4	42.4	48.3	58.9
ERM [29] + ELP-SR ($\tau*$)	**78.0**	**81.5**	**88.7**	**42.4**	**48.3**	**59.1**
LDAM [29]	77.0	81.0	88.2	42.0	46.6	58.7
LDAM [29] + ELP-SR	**78.2**	**82.3**	**88.1**	**43.9**	**48.2**	**59.1**

Table 4.2: Comparison of top-1 validation accuracy of different methods on imbalanced CIFAR-10 and CIFAR-100 datasets. All results are implemented based on ResNet-32. $\tau = 1$ indicates applying τ-normalization [107] with $\tau = 1$. $\tau*$ stands for results with the best settings of τ.

results in three kinds of imbalance ratios which are 100, 50, and 10, respectively. To perform fair comparisons, we evaluate our method based on the ResNet-32 baseline from [29].

Experimental Results. As shown in Table 4.2, ELP-SR dramatically improves the performances of the baseline method in all the settings and datasets. The improvements in CIFAR-10 of imbalance ratio 100 and 50 are even larger than LDAM [29]. Moreover, after adapting the normalization from [107], the results of our method show more competitiveness in this task. All results in different settings outperform LDAM.

Besides, we further investigate our method based on the LDAM [29]. By minimizing the margin-based boundary considering the generalization [29], LDAM is well-designed for long-tailed recognition and boosts the performances dramatically. Meanwhile, our method can achieve higher performances on the foundation of LDAM. Though without specific consideration for the long-tailed distribution, ELP-SR offers general improvements to this task. These results demonstrate that our method helps the network generalize and produce discriminative features against the challenging distributions.

4.3.3 Generic Visual Recognition on ImageNet

To reveal the generalization of ELP-SR, we further investigate our method in generic object recognition on the standard benchmark for visual representation.

Dataset and Implementation Details. We evaluate ELP-SR on ImageNet-1K [120],

Backbone	Top-1 Accuracy		Top-5 Accuracy	
	Baseline	ELP-SR	Baseline	ELP-SR
ResNet50	76.13	76.82	92.86	93.32
ResNet101	77.37	77.86	93.54	94.06
ResNet152	78.31	78.77	94.04	94.42
BN-Inception	73.52†	74.05	91.56†	91.74
Inception-V3	77.45	78.12	93.56	94.04
Inception-ResNet-V2	79.63†	80.22	94.79†	95.24
SE-ResNet50	77.05	77.45	93.48	93.88
SE-ResNet101	77.62	77.94	93.93	94.38
SE-ResNet152	78.43	78.61	94.27	94.53

Table 4.3: Comparison of single-crop accuracy (%) on the ImageNet-1K validation set. Different backbones with our method show significant improvements. To perform a fair comparison, † indicates the results implemented and re-trained by ours.

containing 1.28 million images with 1000 categories. To show the effectiveness and generalization, we apply ELP-SR on different backbone networks, which are ResNet-50 [88], ResNet-101 [88], ResNet-152 [88], BN-Inception [101], Inception-V3 [208], and Inception-ResNet-V2 [207]. According to the standard implementations of these works, we adapt SGD with momentum 0.9 as the optimizer. All the networks are trained with the augmentations of random crops and horizontal flips. For ResNet-50, ResNet-101, ResNet-152, and BN-inception, we first resize the images to 256×256 resolutions and then randomly crop them to 224×224. For Inception-V3 and Inception-ResNet-V2, we resize to 320×320 and randomly crop to 299×299 as the corresponding implementations in their works [207, 208]. As in Table 4.3, we report top-1 and top-5 accuracy respectively and compare all the backbones with ELP-SR.

Experimental Results. As in Table 4.3, with ELP-SR, all backbone networks achieve performance gains. The results reveal that our method is valuable to various backbone models and generally ameliorates the representations of networks. Almost all the backbones obtain about a 0.5% percent increase in top-1 accuracy.

Furthermore, to verify the general improvements introduced by our method, we explore the performances of our method with SE-block [95]. As shown in Table 4.3, though SE-block already promotes the performances, our method leads to further boosts on the fundamental of SE-block [95].

k-nearest neighbors accuracy. To reveal the effectiveness of our method, we provide an additional evaluation with the KNN classifier [233]. For feature vector h, we select the top k nearest neighbors by the weights $\exp(h \cdot h'/t)$ corresponding to the labels, where h' indicates features from the training set and t is a temperature term. We apply $t = 0.1$

Method	20	200
ResNet50	75.04	73.21
ResNet50 + ELP-SR	**75.48**	**73.88**

Table 4.4: KNN accuracy on ImageNet-1K. Results of accuracy with 20 and 200 nearest neighbors are presented.

in our experiments.

As shown in Table 4.4, the results with 20 and 200 nearest neighbors are displayed. With the KNN classifier, our method outperforms the backbone network. This reflects that the features after training with ELP-SR become more discriminative.

In all, the general improvements in all the backbones, methods, and tasks reflect that ELP-SR is not sensitive to particular networks, designs, or visual challenges. It provides a valuable regularization for visual representation learning.

4.3.4 Ablation Studies

4.3.4.1 Ablation on Hyper-parameters

Episodic interval $\mathscr{I}$. The number of periodical intervals prevents the ELP from overfitting the features. We experiment with the different values of $\mathscr{I}$ in the CUB dataset. As shown in Table 4.5, the performances are influenced by $\mathscr{I}$. The larger $\mathscr{I}$ induces the degradation of performances. With plenty of training iterations, the ELP classifier tends to be overfitting and cannot measure generalization effectively.

Besides, we also operate comparisons on the ImageNet dataset. The model achieves 76.13, 76.82, and 76.30 when $\mathscr{I}$ equals to 1, 2, and 3, respectively. The proper value of $\mathscr{I}$ can better empower the advantages of ELP. Minor $\mathscr{I}$ may not be sufficient for the construction of ELP. The more significant $\mathscr{I}$ may induce degradation of the ability of the ELP classifier to indicate features' discrimination. Thus, we apply $\mathscr{I} = 2$ in our experiments as this condition generally shows improvements in several datasets.

γ in the SR Factor. The parameter γ is responsible for adjusting the intensity of regularization. Since $\frac{D}{R}$ is always lower than 1, the larger γ leverages smaller regularization for the inputs. As shown in Table 4.5, we compare multiple conditions of γ in fine-grained classification. The variances of γ slightly influence the performances. A proper γ leads to better performances but is not deterministic for fine-grained classification. Moreover, we evaluate different γ values under the condition of $\mathscr{I} = 2$ on ImangeNet-1K. The recognition accuracies are 76.23, 76.82, and 76.30 when γ is set to 1, 2, and 3, respectively.

The Variations of SR Factor. We further investigate our ELP-SR in different forms, as shown in Table 4.6. First, for regularization, the confidences of the ELP classifier reflect the discriminability of features. Since the main classifier tends to be overfitting, p^c is relatively higher and close to 1. Thus, a similar effect may occur for $1-q^c$ and p^c-q^c. As shown in Table 4.6, both formulations enable regularizing the networks to perform better while the model with p^c-q^c achieves a higher result. This is because p^c-q^c provides a more precise measurement of the deviation between the main classifier and the ELP classifier.

Second, to formulate the normalization term, we require both confidences of the ELP classifier and the main classifier to become higher. The higher confidence of the main classifier indicates that the sample can be correctly recognized. This is a primary requirement for better representation of the feature. If the features are hard to recognize even for the main classifier, this may indicate that the visual representation quality is relatively low. It is a primary criterion that the network should provide at least recognizable features. As shown in Table 4.6, higher performances are shown if applying the normalization terms. Both p^c+q^c and p^c*q^c are valid to normalize our ELP-SR. Third, only the regularization of higher q^c can also boost the performances. Without the normalization term, the impact of ELP-SR also guides the networks to be more generalized. However, lacking normalization, the improvements are relatively lower. Besides, simple normalization is also valuable. Since $\frac{2}{p^c+q^c}$ and $\frac{2}{p^c*q^c}$ also expect higher confidences of ELP, a similar influence may occur through leveraging the normalization term only. These results demonstrate that regularization and normalization are valuable in ELP-SR. Simultaneously, the combinations of both sides introduce a further increase in performances.

Finally, we also operate ablations for the distillation of the probability of two classifiers. Remarkable decreases are shown in Table 4.6 of both conditions for L1 and L2 regressions. The network should not be optimized to solve features' discriminability directly. Distilling can lead the main classifier to perform similarly to the ELP classifier but does not encourage the network to be more generalized. If the main classifier is optimal according to the ELP classifier, the network can 'pretend' to achieve discriminative features. However, in testing, this 'cheating' is useless. Additionally, we replace the ELP classifier with a memory bank and update the memory by a momentum-based moving average. When the momentum is 0.9 and 0.1, the results are 86.1%and 86.5%, respectively. The results show that the moving average operation helps fine-grained recognition, but it provides a weaker regularization than the episodically initialized ELP

Parameter	$\mathscr{I}=1$	$\mathscr{I}=2$	$\mathscr{I}=3$	$\mathscr{I}=4$	$\mathscr{I}=5$
$\gamma=1$	88.0	88.2	88.2	88.0	87.8
$\gamma=2$	88.0	88.5	88.2	88.0	87.8
$\gamma=3$	87.6	**88.8**	88.0	88.0	87.6
$\gamma=4$	87.5	88.0	87.8	87.8	87.5

Table 4.5: Results for different values of $\mathscr{I}$ and γ on CUB. $\mathscr{I}$ prevents the ELP from overfitting, and γ adjusts the intensity of regularization.

Formulation	D	R	Top-1 Accuracy
$\frac{D}{R}$	p^c-q^c	p^c+q^c	**76.82**
	p^c-q^c	p^c*q^c	76.75
	$1-q^c$	p^c+q^c	76.78
	$1-q^c$	p^c*q^c	76.70
D	p^c-q^c	-	76.71
	$1-q^c$	-	76.60
$\frac{1}{R}$	-	p^c+q^c	76.25
	-	p^c*q^c	76.23
Distillation	L1		76.12
	L2		76.18

Table 4.6: Comparison for variations of SR Factor on ImageNet-1K. Various conditions are presented, including different formulations of D and R, with or without D and R, and direct distillation of the main and ELP classifier.

classifier.

4.3.4.2 Visualization

To demonstrate the efficacy of our ELP, we present a visualization for the testing accuracy of our ELP based on CUB. In detail, we train the baseline method, take the features from the backbone to train ELP, but do not leverage ELP-SR for network training. Meanwhile, we take our method training with ELP-SR as the comparison. This is similar to applying linear probing for every epoch. Since ELP is re-initialized every two epochs for CUB, to better reveal the capacity of ELP under different conditions, we plot the accuracy every two epochs. As shown in Fig. 4.3, unseen features in the testing set are remarkably more recognizable. This indicates that the network with ELP-SR is more generalized and produces more discriminative features. Even for the simple classifier, the unseen samples represented by the network are easier to be classified.

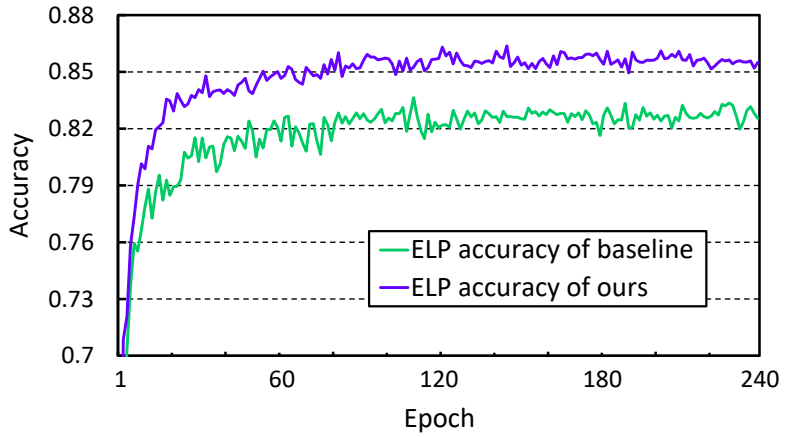


Figure 4.3: Curves of testing accuracy only with ELP classifier on CUB. Compared with our method, We utilize the baseline method that extracts the features from the backbone, trains ELP with features individually but does not leverage ELP-SR for the backbone training. Features trained with ELP-SR are more discriminative than the baseline and easier to be classified by simple ELP.

4.4 Conclusion

In this paper, we propose episodic linear probing (ELP) to estimate the generalization and discriminability of features online. By ELP, we propose an ELP-suitable Regularization term (ELP-SR) to regularize the models. Our insights are two-fold. 1). Since the main classifier may be overfitting and its confidence may not indicate the discrimination of features, the ELP classifier provides additional regularization for more discriminative features. 2). Immediate suitability is effective in measuring the discrimination of features. An intuitive hypothesis is that if the features are highly discriminative, they should be recognizable by an easily learned linear classifier. Our ELP is episodically re-initialized, effectively mitigating overfitting and regularizing the network towards better linear separability.

CHAPTER 5

IMPROVING LEARNING EFFICIENCY FOR VIDEO CAPTIONING

5.1 Introduction

Video captioning is a challenging task that requires the model to learn semantics and express through natural language. The main challenge in this task is understanding the diverse visual contents in the videos. Recently, many solutions have been proposed to solve this problem, e.g., leveraging better video representations [149, 167], complex network designs [266, 267], and end-to-end learning [141, 194]. These works facilitate a better understanding of video semantics and generate coherent descriptions of the visual contents.

Despite significant improvements, understanding video semantics remains a challenging task. A major obstacle to achieving this understanding is the semantic ambiguity in videos, caused by their visual redundancy. The contents of videos are diverse and difficult to precisely trim with specific descriptions. As illustrated in Fig. 5.1 (a), some contents, such as irrelevant and minor events, are not described by the ground truth, and without particular descriptions, they are implicit for the network to understand. In addition, contents such as transitions are related to the events described by the ground truth but do not contain valuable semantics for network learning. These contents present a challenge for neural networks. The video captioner is always hard to generalize

This chapter is based on joint work [136] with Linchao Zhu, Xiaohan Wang, and Yi Yang, presented primarily as it appears in the TMM 2023.

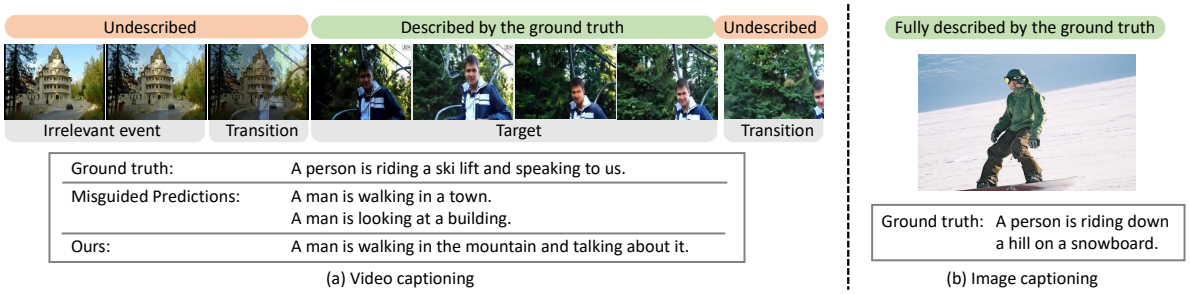


Figure 5.1: Video semantics are ambiguous. Some frames contain irrelevant events or serve as transitions. They do not provide valuable contents corresponding to the ground truth in video captioning. Meanwhile, image contents are concise and explicit. The ground truth in image captioning easily summarizes all image semantics.

redundant contents or misguided by ambiguous semantics. Meanwhile, the mismatch between descriptions and visual contents further increase the difficulties in learning video semantics. All this defeats induce the captioner may be misguided by trivial and irrelevant semantics.

Comparably, image contents are more concise, and the semantics are salient, as shown in Fig. 5.1 (b). There are no irrelevant events or transitions apart from the valuable contents in the images. Meanwhile, the image descriptions are precise. The ground truth description can summarize most contents. This makes the image samples easier to be captioned. Empirically, results from image captioning datasets [127] are often better than video captioning in various metrics.

The primary distinction between images and videos for captioners lies in content density. The redundancy and ambiguity in video data cause the network to struggle in generalizing complex video semantics. While previous works have focused on proposing better captioner designs, improved architectures to increase network capacity, which aids in learning semantics and handling redundancy. However, an another line for improving video captioning has been overlooked: modifying content density to enhance the learning process of video captioners. In this work, we propose a novel learning method called Image-Compounded learning for video Captioners (IcoCap). IcoCap compounds concise and easily-learned image semantics into video samples, diversifying the visual contents and compelling the network to learn against redundant contents. Besides, the compounded image semantics are more easily learned compared to video semantics, which are similar to a strong competitor [189, 232] for learning video semantics. To address video captioning, the captioner must investigate valuable video cues in contrast to easily-learned image contents. This further enhances the captioner's ability to learn

video semantics. Additionally, IcoCap alleviates the challenges of learning from mismatched descriptions by encouraging the network to flexibly learn descriptions based on visual semantics, rather than relying on rigidly pre-assigned captions.

Specifically, IcoCap comprises two modules: the Image-Video Compounding Strategy (ICS) and Visual-Semantic Guided Captioning (VGC). In detail, ICS is designed to compound image content into video content. This approach further diversifies the video samples, guiding the video captioner to learn against redundancy. Simultaneously, the introduction of easily-learned image content compels the network to extract valuable video cues while filtering out irrelevant elements. Additionally, IcoCap addresses the issue of ambiguous video semantics by VGC, which facilitates flexible learning of semantics based on visual content. In VGC, the ground truth is selected from relevant descriptions rather than strictly corresponding to the original video ground truth. A visual-semantic consistency factor is introduced to adjust the captioning process, promoting the network to focus on the salient visual content rather than concentrating on minor and detailed contents.

The main contributions can be summarized below:

1. We propose an Image-Compounded learning for video Captioners (IcoCap). IcoCap introduces image samples and compounds the images into video contents. IcoCap impels the network to mine valuable video cues against the semantic ambiguity in videos.

2. We propose an Image-video Compounding Strategy (ICS) and Visual-semantic Guided Captioning (VGC). ICS provides a series of operations to compound images and videos, which promotes the network's ability to learn video semantics against ambiguity. VGC helps the network to flexibly learn complex video contents from ICS, rather than rigidly following the ground truth.

3. Without complicated designs or networks, our method performs favorably or outperforms the state-of-the-art methods on various datasets, including MSR-VTT, MSVD, and VATEX.

5.2 Image-compounded video Captioner

We propose an Image-compounded video Captioner (IcoCap) to introduce image samples to improve video captioning. IcoCap contains two parts: Image-video Compounding Strategy (ICS) and Visual-semantic Guided Captioning (VGC). ICS uses image samples to augment video samples for video captioning. It contains a series of augmentation strategies to compound image contents into video contents, as shown in Fig. 5.2. Moreover,

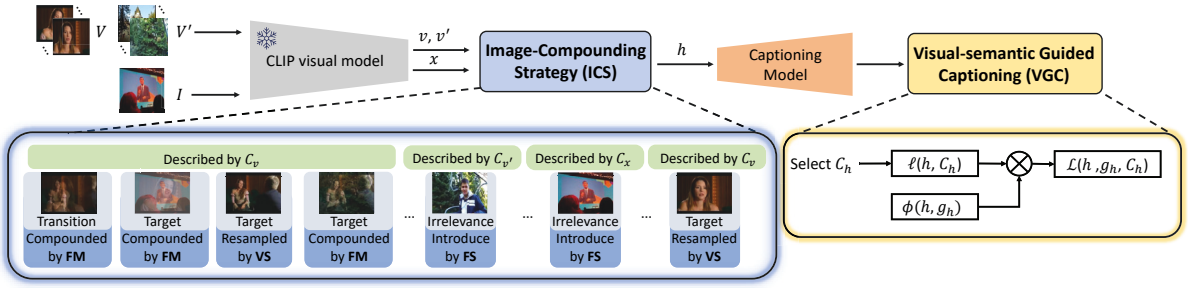


Figure 5.2: Overview of Image-video Compounding Strategy (ICS). ICS introduces image samples to help the network learn ambiguous video semantics. All features are extracted by a frozen CLIP visual model. V and V' are different video samples. I is the additional image sample. v, v' and x are features for video and image samples, respectively. C_v, $C_{v'}$, C_x are the descriptions for V, V' and I, respectively.

VGC provides a flexible learning manner for the complex and diverse semantics from ICS.

5.2.1 Image-video Compounding Strategy

The visual contents of the video data are diverse but ambiguous. It is hard to annotate all instances and events in the frames specifically. However, semantics in image data are explicit and clear. Existing works [127, 194, 205] perform joint training to learn image and video samples. They treat the images and videos individually, which are separately provided for the network as different training samples.

However, our work aims to utilize image samples to augment video samples and compound them as one training sample. Both videos and images can occur in the same training samples. This leads to the redundancy of visual content changes according to the introduction of image samples. As shown in Fig. 5.2, we proposed Image-video Compounding Strategy (ICS) to produce training samples. Specifically, for a given video V with N frames, we pre-process and represent the video as frame features v following [149, 183, 209]. In IcoCap, we randomly sample M images from [143] to construct an auxiliary image set. In every training step, we select an image sample I and extract the image feature, denoted as x. $x \in \mathbb{R}^{1\times D}$. We also additionally pre-process another video V'. We denote the frame feature from V' as v'. Then, ICS takes v, v', and x as inputs and produces compounded samples h, where $h \in \mathbb{R}^{n\times D}$.

ICS consists of three steps: Intra-Video Sampling (VS), Inter-Feature Mixup (FM), and Inter-Frame Swap (FS), represented as functions f_{VS}, f_{FM}, and f_{FS}. For the current

video sample, we construct a set of frame features as $\mathcal{V} = \{v_1, v_2, ..., v_N\}$, where v_i indicates a frame feature ($v_i \in \mathbb{R}^{1\times D}$). We also create an auxiliary set $\mathcal{A} = \{v'_1, v'_2, ..., v'_N, x_1, x_2, ..., x_N\}$, where x_i is a duplicate of the same image sample x, and $v'_i \in \mathbb{R}^{1\times D}$.

1. Intra-Video Sampling (VS): In VS strategy,

$$h = f_{\mathrm{VS}}(\mathcal{V}) \tag{5.1}$$

The frame feature h is randomly sampled from $\mathcal{V}$ and $h \in \mathbb{R}^{n\times D}$. To be noticed, previous works [39, 141, 267] usually apply uniform sampling during producing video features. Uniform sampling, with a fixed sampling interval, is insufficient in fully utilizing the information present in a video due to the redundancy of video contents. Some frames may never be sampled and used for training. Comparatively, random sampling in VS can take advantage of all frames in training and also produce more challenging and diverse samples for captioner training.

2. Inter-Feature Mixup (FM): To implement FM strategy, an auxiliary feature h' is randomly sampled from $\mathcal{A}$, and $h' \in \mathbb{R}^{n\times D}$. Then, FM strategy can be formulated as:

$$h_i = f_{\mathrm{FM}}(h_i, h'_i, J) = \begin{cases} \alpha h_i + (1-\alpha)h'_i & \text{if} \quad i \in J \\ h_i & \text{if} \quad i \notin J \end{cases} \tag{5.2}$$

where i is the index number and $i \in [1, n]$. J is the set of index numbers and $J = \{j_1, ..., j_k\}$. k is a random number and $k \in (1, n)$. Moreover, α represents the mixup ratio. $\alpha \in (0, 1)$.

3. Inter-Frame Swap (FS): In FS strategy, we replace the training feature with the auxiliary features, by given the random index Q, in which Q is also a set of index number. $Q = \{q_1, ..., q_t\}$ and $t \in (1, n)$. The operation can be written as:

$$h_i = f_{\mathrm{FS}}(h_i, h'_i, Q) = \begin{cases} h'_i & \text{if} \quad i \in Q \\ h_i & \text{if} \quad i \notin Q \end{cases} \tag{5.3}$$

where i is the index number and $i \in [1, n]$. In FS, each frame feature has a 50% probability of being mixed up and a 50% probability of being replaced by external features.

ICS blends the semantics of images and videos, generating an image-compounded video sample for captioner learning. Samples compounded with various video samples exhibit greater diversity than the original samples. This diversity sets a higher requirement for the network to handle redundancy, which in turn enhances the generalization ability of the captioner. Furthermore, some samples are constructed by compounding both video and image samples. By incorporating easily-learned image semantics, the video captioner is required to extract valuable video cues while disregarding easy but

irrelevant cues from images. This process further improves the captioner's ability to avoid being misled by irrelevant semantics.

In addition, the network architecture used in our work is a simple transformer model comprising a four-layer transformer and a fully connected layer, following the approach described in [127]. The transformer network in our work is responsible for mining valuable contents from h and understanding complex semantics. The fully connected layer produces predictions of words at every time step.

5.2.2 Visual-semantic Guided Captioning

Features from ICS are diverse and complex, which are hard to be solved by the original captioning loss. In our work, we propose Visual-semantic Guided Captioning (VGC) to encourage the network to express semantics flexibly according to the given visual semantics.

There are two kinds of visual-semantic guidance in VGC, which are visual-semantic based description selection and the visual-semantic consistent factor.

First, the ground truth for descriptions is flexibly selected from relative descriptions guided by the visual semantics of h. The training feature h may contain contents from V, V', or I, in which all the features have corresponding ground truth and are available to be expressed. For v, v' and x, the corresponding captions are C_v, $C_{v'}$, and C_x. We utilize a pre-trained language model [183] to extract the features of the descriptions and denote them as g_v, $g_{v'}$, and g_x. All the language features are in the size of $\mathbb{R}^{1\times D}$. Then, the ground truth of the current sample can be determined by the cosine similarity between visual features (h) and language features (g_v, $g_{v'}$ and g_x). We denote the language feature with the largest similarity as g_h and select the corresponding caption C_h as the ground truth. The ground truth is the corresponding caption with the maximum similarity. Besides, to calculate similarity, g is copied n times to fit the dimension of h.

To be noticed, rather than describe all possible ground truth as in [194], the network should learn to produce C_h only, which reflects most of the visual contents in h and is the most suitable for current visual semantics. This training goal requires the network to exclude expressive but minor semantics and focus on the exploration of valuable visual semantics.

Second, we further leverage the guidance of visual semantic by designing a visual-semantic consistent factor ϕ. It regularizes the learning procedure and reduces the punishments if the ground truth description possess lower consistency with current visual contents. Specifically, the factor encourages the predictions to be close to the

salient semantics in visual contents, which can be formulated as follow:

$$\phi(h, g_h) = -log(min(S/\tau, 1)) \tag{5.4}$$

where τ is a temperature coefficient and S indicates the cosine similarity between h and g_h. Regardless of whether the predictions are close to C_h, if the language features of the predictions show higher similarity to most of the visual contents in h, the factor ϕ will be relatively lower.

Finally, the overall loss function can be formulated as follows:

$$\mathscr{L}(h, g_h, C_h) = \phi(h, g_h) \cdot \ell(h, C_h) \tag{5.5}$$

where ℓ is cross-entropy loss function which is widely used as captioning loss in [2, 149].

Instead of forcing the network to generate fixed, rigid descriptions, VGC encourages the network to learn flexibly based on the visual semantics present in the video content. This is achieved by adaptively assigning the ground truth for ambiguous video semantics originating from ICS. Simultaneously, VGC introduces a flexible factor to modulate the captioning learning process, guiding the network to generate improved descriptions that better align with the diverse visual semantics. This approach not only helps the network to better adapt to varying content densities but also ensures that the generated captions are more representative of the actual video content. All modules in IcoCap contribute to enhancing the captioner's ability to learn effectively while dealing with redundant video contents and ambiguous video semantics.

5.3 Experiments

In this section, we discuss the details of our method and evaluate the captioning performances in various datasets.

5.3.1 Experimental Setup

Datasets: We evaluate our method using three established video captioning benchmarks: MSRVTT [239], MSVD [35], and VATEX [222]. In all datasets, we employ English annotations as the ground truth for experimentation.

MSR-VTT [239] is a prevalent benchmark for video captioning that consists of $10,000$ videos, each with 20 annotations. To facilitate evaluation and comparison, we adopt the standard setting used in [239], wherein the dataset is partitioned into three subsets: a

training set with 6,513 samples, a validation set with 497 samples, and a test set with 2,990 samples.

MSVD [35] is another established benchmark in video captioning, which comprises 1,970 YouTube videos, each with approximately 40 annotations. The dataset is partitioned into three subsets: a training set consisting of 1,200 videos, a validation set consisting of 100 videos, and a test set consisting of 670 videos.

In addition, VATEX [222] is another widely used dataset for video captioning, sourced from the Kinetics-600 dataset [109]. VATEX contains annotations in both English and Chinese, with 10 descriptions in each language. The dataset comprises 25,991 video clips for training, and 3,000 and 6,000 video clips for validation and testing, respectively.

Implementation Details: We extract features from all video frames and image samples using the CLIP visual model [183], which we utilize solely for representation and do not involve in network training. The CLIP model is a powerful representation method that has been widely applied in video captioning [149]. Unlike recent works [38, 141, 267] that utilize video-based backbones such as Vivit, C3D, and I3D, our work employs an image-based method through the CLIP model. Video-based backbones take into account the relationships between frames in chronological order, while external images lack such relationships and may confuse video-based models. In contrast, image-based methods do not consider sequential properties, providing more flexibility for image-video compounding. Therefore, we choose the CLIP model to represent visual content. Additionally, the pre-trained language model used in VGC is based on the CLIP language model.

We capture all the video frames and extract the features by CLIP visual model [183]. The image samples are also processed similarly. We only use CLIP model to represent video and image data, which does not participate in the network training. Besides, CLIP model is a powerful representation method and has already been widely applied in video captioning [149]. Unlike recent works [38, 141, 267] that utilize video-based backbones [12, 53, 238], our work employs an image-based method through the CLIP model. Video-based backbones take into account the relationships between frames in chronological order, while external images lack such relationships and may confuse video-based models. In contrast, image-based methods do not consider sequential properties, providing more flexibility for image-video compounding. Therefore, we choose the CLIP model to represent visual content. Additionally, the pre-trained language model used in VGC is also based on the CLIP language model.

The dimension of input features D is 512. As our captioning model, we employ a

Table 5.1: Comparison with state-of-the-art methods on the test split of MSRVTT. † indicates the results from the official implementation of [141] taking 32 frames as inputs. ViT-B/32 and ViT-B/16 stand for CLIP ViT-B/32 and CLIP ViT-B/16 models, respectively. CLIP baseline only uses the video features extracted by CLIP model and does not apply our method. Joint baseline indicates both video and image samples are jointly trained with CLIP baseline.

Method	Feature	MSR-VTT			
		BLEU-4	METEOR	ROUGE-L	CIDEr
GRU-EVE [1]	InceptionResNetV2 + C3D	38.3	28.4	60.7	48.1
STG-KD [167]	ResNet101 + I3D	40.5	28.3	60.9	47.1
POS-CG [218]	InceptionResNetV2	42.0	28.2	61.6	48.7
POS-VCT [93]	InceptionResNetV2 + C3D	42.3	29.7	62.8	49.1
ORG-TRL [267]	InceptionResNetV2 + C3D	43.6	28.8	62.1	50.9
SAAT [273]	InceptionResNetV2 + C3D	39.9	27.7	61.2	51.0
OpenBook [266]	InceptionResNetV2 + C3D	33.9	23.7	50.2	52.9
Revnet [124]	Inception-V4	42.4	28.1	62.3	53.2
HMN [247]	InceptionResNetV2 + C3D	43.5	29.0	62.7	51.5
SWINBERT† [141]	VidSwin	41.9	29.8	62.1	53.7
CLIP4Clip [149]	ViT-B/32	46.1	30.7	63.7	57.7
CLIP Baseline	ViT-B/32	43.1	29.3	61.9	54.8
Joint Baseline	ViT-B/32	43.5	29.4	62.4	55.2
Ours	ViT-B/32	46.1	30.3	64.3	59.1
Ours	ViT-B/16	**47.0**	**31.1**	**64.9**	**60.2**

simple transformer network [212], and all training settings follow [127]. For the external image subset, we randomly select samples from MSCOCO [143], with a default length of $10,000$. In addition, we set the hyper-parameters $n = 32$ and $\tau = 0.5$. Image samples, original video samples, and additional video samples have a 24.71%, 51.79%, and 23.5% probability of being the ground truth, respectively. All input samples are potentially salient content and can be learned for captioning. During evaluation, we uniformly sample frames from videos in accordance with [127, 149]. Further experiments and ablations will be presented in the next section. To ensure a fair comparison, we evaluate our method and report results for other methods based on the official test split of the corresponding datasets.

5.3.2 Performance Comparison

As shown in Table 5.1, we evaluate our method on the MSR-VTT dataset using various captioning metrics. Our method outperforms the current state-of-the-art SWINBERT [141] without bells and whistles. Using only the CLIP ViT-B/32 model, our method

improves by 5.1, 1.3, 2.8, and 5.4 in BLEU-4 [169], METEOR [55], ROUGE-L [139], and CIDEr [213], respectively, which is significant in MSR-VTT. Notably, our method does not employ multi-modal features [218, 273] or features from detectors [167, 267]. We also do not apply complex network design [38, 247], costly end-to-end training [141], or assemble operations [149].

Furthermore, we present the results of CLIP baseline (which only uses CLIP features without ICS and VGC) and joint training baselines (where CLIP baseline is jointly trained with image features without ICS and VGC) implemented with CLIP ViT-B/32 model. CLIP is a powerful representation method, and even the CLIP baseline, which utilizes only CLIP features and our network, achieves relatively higher performance than recent methods [38, 141, 267]. Additionally, introducing image samples and joint training with MSR-VTT data slightly improves the performance of the CLIP baseline. However, due to the domain gap between video and images, joint training video samples with image samples yields only marginal improvement. In comparison, our method with image samples yields a significant improvement in video captioning, with the value of the CIDEr metric increasing by 4.3.

In addition, due to the effectiveness of ICS and VGC, our method outperforms the CLIP4Clip [149] method, which is also based on the CLIP ViT-B/32 model and utilizes CLIP features. With the same features as inputs, our method achieves significantly better performance than CLIP4Clip, with gaps of 1.1, 1.2, and 3.5 in BLEU, METEOR, and CIDEr, respectively.

As presented in Table 5.2, we conducted experiments on the MSVD dataset, which further demonstrates the effectiveness of our method. In comparison with the improvements observed between our method and the joint training baseline on MSR-VTT, our method's superiority is further highlighted. Specifically, our method yields a significant increase of 7.1 in the CIDEr metric, which is a dramatic improvement compared to the joint training baseline.

Moreover, we present further results on the VATEX dataset [222], which includes more complex and diverse descriptions compared to MSR-VTT and MSVD. In addition to exploring video content, VATEX sets a higher requirement for language diversity in predictions. To achieve better performance in evaluation, the predictions should be more vivid and diverse, which is relatively challenging for the simple captioning network utilized in our method. As demonstrated in Table 5.3, we achieve comparable performance to state-of-the-art methods in VATEX. Despite not aiming to generate vivid descriptions with high linguistic complexity, IcoCap still outperforms many recent methods with more

Table 5.2: Comparison with state-of-the-art methods on the test split of MSVD. † indicates the results from the official implementation of [141] taking 32 frames as inputs.

Method	Feature	MSVD			
		BLEU-4	METEOR	ROUGE-L	CIDEr
GRU-EVE [38]	InceptionResNetV2 + C3D	47.9	35.0	71.5	78.1
POS-CG [48]	InceptionResNetV2	52.5	34.1	71.3	88.7
POS-VCT [49]	InceptionResNetV2 + C3D	52.8	36.1	71.8	87.8
SAAT [50]	InceptionResNetV2 + C3D	46.5	33.5	69.4	81.0
STG-KD [2]	ResNet101 + I3D	52.2	36.9	73.9	93.0
ORG-TRL [4]	InceptionResNetV2 + C3D	54.3	36.4	73.9	95.2
HMN [52]	InceptionResNetV2 + C3D	59.2	37.7	75.1	104.0
SWINBERT† [5]	VidSwin	55.7	**39.6**	75.7	109.4
CLIP Baseline	ViT-B/32	55.5	38.0	74.4	95.5
Joint Baseline	ViT-B/32	57.2	37.5	74.6	96.7
Ours	ViT-B/32	56.3	38.9	75.0	103.8
Ours	ViT-B/16	**59.1**	39.5	**76.5**	**110.3**

complex designs.

Additionally, the input features in our work can be summarized as image-based representations, which extract features from each frame individually. In this section, we also evaluate video-based representations, which represent multiple sequential frames as a single feature.

In detail, we first apply the ICS strategies directly to the original video frames and then extract the features using the VideoSwin Transformer following [12, 141]. Since the VGC requires calculating similarity between visual and language features, we only adapt ICS with the VideoSwin Transformer and use our captioning model for comparison. As in our work, the parameters of the VideoSwin Transformer are also fixed during training. However, after applying the video-based representations, the results on MSR-VTT are only 38.5, 27.3, 59.0, and 45.3 for BLEU-4, METEOR, ROUGE-L, and CIDEr, respectively. The performance of the ICS strategies with the VideoSwin Transformer decreases by 8.5 in the CIDEr metric compared with the reported value from SWINBERT [141]. More significantly, the gap between applying ICS with the VideoSwin Transformer and our method is 14.9. The significant drop in the captioning performance indicates that introducing augmentation strategies into video-based representation is not feasible. Methods such as the VideoSwin Transformer need to model temporal information specifically. Complex augmentations can break the connections and relationships between the original frames, causing representation methods to become confused and fail to produce valuable features for video captioning.

Table 5.3: Comparison with state-of-the-art methods on the test split of VATEX. † indicates the results from the official implementation of [141] taking 32 frames as inputs.

Method	Feature	VATEX			
		BLEU-4	METEOR	ROUGE-L	CIDEr
VATEX [222]	bi-LSTM + I3D	28.4	21.7	47.0	45.1
ORG-TRL [267]	InceptionResNetV2 + C3D	32.1	22.2	48.9	49.7
Support-set [171]	ResNet152	32.8	24.4	49.1	51.2
OpenBook [266]	InceptionResNetV2 + C3D	33.9	23.7	50.2	57.5
SWINBERT† [141]	VidSwin	37.8	26.1	53.0	71.6
CLIP Baseline	ViT-B/32	35.9	24.0	52.1	57.3
Joint Baseline	ViT-B/32	35.5	24.4	51.6	60.1
Ours	ViT-B/32	36.9	24.6	52.5	63.4
Ours	ViT-B/16	37.4	25.7	53.1	67.8

We also evaluate feature-level augmentations based on the VideoSwin Transformer [12, 141]. First, we extract frame features using a frozen VideoSwin Transformer and then augment these features using the same strategies as in ICS. However, the resulting CIDEr metric is only 20.5. Augmenting feature-level representations in video-based methods leads to a decline in performance. This is because features from video-based methods naturally contain sequential relationships and contexts. Although we can obtain features for specific frames and apply augmentations, it may be difficult for the captioning model to understand the augmented features with broken and confusing inter-frame relationships after compounding. By comparison, our image-based representations do not face this issue. Every frame is represented individually, making them flexible to various augmentations and achieving better performance.

5.3.3 Ablation Studies

In this section, we conduct a comprehensive ablation study for several details in our method. All experiments are operated based on CLIP ViT-B/32 and MVSD test split.

Comparison of Data Combinations: Our work utilizes video data v' from the current training dataset and external image samples x to augment training features. In Table 5.4, we compare the different combinations of data samples. Both combinations, v with v' or x, are useful for improving video captioning. Additional samples expand the training data and lead to better results. However, improvements from introducing image samples are more significant, with an increase of 2.9 in the CIDEr metric. Image samples possess concise contents and precise descriptions, which are easily learnable for the network. Introducing image samples formulates the more challenging samples. This

Table 5.4: Comparison for combinations of data samples and strategies in ICS. VS, FM, and FS are short for intra-video sampling, inter-feature mixup and inter-frame swap, respectively. All strategies and additional data are useful and the combinations lead to higher performances.

ICS Strategy	Data Sample	BLEU-4	METEOR	ROUGE-L	CIDEr
VS + FM + FS	v, v'	56.5	37.8	73.9	97.5
VS + FM + FS	v, x	57.5	37.8	74.1	100.4
VS	v, v', x	53.3	37.5	74.1	97.6
FM	v, v', x	53.5	37.8	74.5	97.9
FS	v, v', x	59.1	38.7	75.5	101.2
VS + FM	v, v', x	59.0	38.6	74.7	99.1
VS + FS	v, v', x	59.4	38.2	75.3	102.3
FM + FS	v, v', x	60.5	38.9	75.0	103.2

Table 5.5: Ablation for different parts in our method. Ours with all descriptions indicates taking all relative descriptions C_v, $C_{v'}$, and C_x as the ground truth for h at the same time. In comparison, all modules in our work are helpful in improving the performance of video captioning.

Method	BLEU-4	METEOR	ROUGE-L	CIDEr
Ours w/o VGC	60.4	38.7	75.1	100.8
Ours w/ all descriptions	60.1	38.4	75.3	100.0
VGC w/o ϕ	59.5	38.2	75.3	102.1
Ours w/o ICS	59.2	37.8	74.5	95.9
CLIP Baseline	55.5	38.0	74.4	95.5
Joint Baseline	57.2	37.5	74.6	96.7
Ours	56.3	38.9	75.0	103.8

forces the network to mine video semantics against the easy image semantics

Comparison of different strategies in ICS: We conducted an ablation study on the strategies in ICS, and the results are presented in Table 5.4. All strategies effectively diversify the training samples and improve performance. Inter-frame swap was found to be the most helpful strategy as it directly replaces the visual contents with additional data, offering maximum influence on the semantics of the training samples compared to other strategies. Additionally, image samples are easier to learn than video samples, and inter-frame swap introduces expressive image samples into video samples, requiring the network to ignore irrelevant semantics and refocus on the video contents. In our work, we utilized all three ICS strategies, and the combination of these strategies with additional data samples resulted in significant improvement in video captioning performance. With CLIP ViT-B/32 features, we achieved a CIDEr score of 103.8, which is an 8.3 improvement over the baseline.

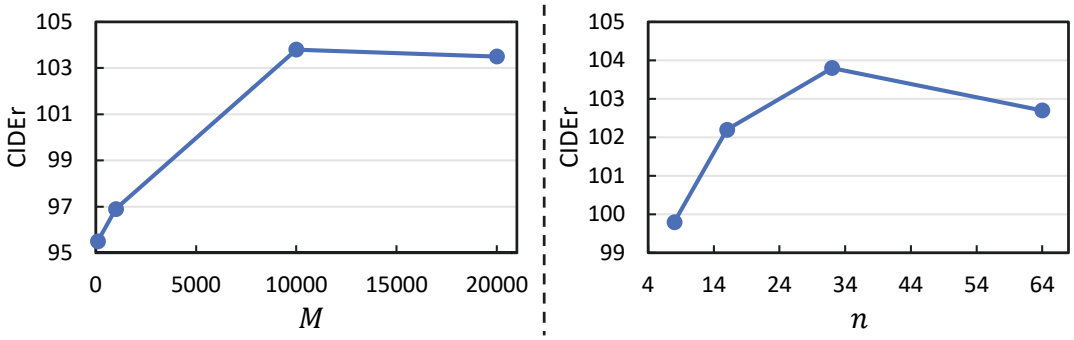


Figure 5.3: Comparison of the values of CIDEr metrics for different sizes of the external image set (M) and different numbers of frames in the video samples (n).

Efficacy of ICS and VGC: To evaluate the efficacy of VGC, we separately evaluate our captioning loss $\ell(h, C_h)$ and the factor ϕ. Firstly, we use all descriptions as the ground truth and experiment with the loss function $\mathscr{L}(h, g_h, C_v, C_{v'}, C_x) = \phi(h, g_h) \cdot (\ell(h, C_v) + \ell(h, C_{v'}) + \ell(h, C_x))$. As shown in Table 5.5, the result with all the descriptions (corresponding to VGC w/ all descriptions) is relatively lower. Giving all possible descriptions to the network at the same time may confuse the network and result in worse performance of 100.8 in CIDEr. Meanwhile, the factor ϕ is also useful in improving performance. Without the modulation of factor ϕ, the performance decreases by 1.7 in CIDEr.

Additionally, ICS serves to diversify and expand the training samples. When ICS is applied to the CLIP baseline, the network can also be significantly improved, as evidenced by a CIDEr score of 102.1. However, this improvement is not as significant as the results obtained using our method, which combines both ICS and VGC (103.8 in CIDEr). These comparisons suggest that both ICS and VGC are effective in helping the network learn useful visual content.

Moreover, Results in Table V also reveal the efficacy of VS in ICS. When employing the ICS strategy exclusively with VS, our method surpasses the baseline by 1.1% in the CIDEr metric. Conversely, when utilizing the complete ICS strategy excluding VS, there is a performance decline of 0.6% in the CIDEr metric. These comparative analyses underscore the efficacy of VS within the ICS strategy.

Ablation on frame number: The number of frames, denoted as n, plays a crucial role in determining the diversity of the training samples. As illustrated in Fig. 5.3, larger n values generally lead to better performances. However, this also increases the amount of noise introduced into the model. Interestingly, we observed that when n was set to 64, the results were slightly lower than those obtained with $n = 32$. We hypothesize that

Table 5.6: Ablation of τ. τ changes the influences of VGC and we set $\tau = 0.5$ as default.

Parameter τ	BLEU-4	METEOR	ROUGE-L	CIDEr
$\tau = 0.1$	55.0	37.8	74.5	100.0
$\tau = 0.3$	57.2	38.3	75.3	102.8
$\tau = 0.5$	56.3	38.9	75.0	103.8
$\tau = 0.7$	55.8	38.0	75.0	103.0
$\tau = 0.9$	56.6	38.3	75.3	102.4

Table 5.7: Ablation of the mixup ratio α. The ratio influences samples after augmentations, which should be set appropriately.

Parameter α	BLEU-4	METEOR	ROUGE-L	CIDEr
$\alpha = 0.01$	55.4	38.2	75.0	102.2
$\alpha = 0.05$	56.3	38.9	75.0	103.8
$\alpha = 0.5$	58.2	37.7	74.1	98.3
Random α	58.5	37.4	74.0	97.8

this is because larger n values result in more diverse and complex samples from the ICS strategy, making the training samples more difficult for the model to learn. Thus, the size of the training samples should not be too large. In our work, we set $n = 32$ to strike a balance between diversity and complexity.

In addition, we conducted an ablation study on the number of auxiliary image sets. The features x extracted from the image set provide additional visual content and semantics, which effectively enhance the video samples. As shown in Fig. 5.3, increasing the amount of image data results in better performance for our method. However, the improvements become marginal after introducing more than $10,000$ image samples. Therefore, the default value of M in our method is set to $10,000$.

5.3.4 Ablation on τ

Factor τ influences the modulation degree of VGC in IcoCap and should be properly set. As shown in Table 5.6, the results for $\tau = 0.5$ achieve the highest performances. Both larger and smaller values of τ lead to a decrease in performance. Moreover, a smaller value of τ causes S/τ in VGC to be closer to 1, resulting in lower punishments. This reduces the modulation from the consistent factor ϕ and the efficacy of our VGC. In experiments, a smaller value of τ also performs worse. These results further prove the efficacy of our method.

Table 5.8: Ablation for swapped frame ratio s in FS on MSR-VTT.

s	75%	50%	25%	10%	0%
CIDEr	58.1	59.1	58.5	57.7	57.4

5.3.5 Ablation on Mixup Ratio α

We conducted extensive experiments to analyze the impact of different mixup ratios α. As demonstrated in Table 5.7, the value of this ratio needs to be carefully determined. When setting $\alpha = 0.05$, the performance surpasses other values, yielding the best results. Furthermore, we observed that using a random value for α leads to the creation of more challenging compounded samples. It is important to note that more difficult samples do not always guarantee improved performance and may potentially confuse the networks during the learning process. The results obtained with a random ratio are lower than those achieved with $\alpha = 0.05$, exhibiting a decrease of 5.0 in CIDEr.

5.3.6 Ablation on Swap Ratio in FS

In assessing the impact of content sampling in IcoCap, we introduce a swap ratio, denoted as s, to represent the proportion of content replaced by FS. The ablation study concerning the swap ratio s is shown in Tab. 5.8, showcasing the CIDEr results on MSR-VTT.

When $s = 0$, it implies that no frames have been swapped within the visual content. On the other hand, $s = 50\%$ means that each frame has a 50% chance of being replaced by randomly selected visual content. The results illustrate that an optimal value for s diversifies inputs and enhances video learning. A lower value of s provides more straightforward input samples, reducing the need for generalization. In contrast, a very high value of s introduces more unrelated visual content, making it challenging for the network to process. In our research, we opted for $s = 50\%$, as it demonstrated the best results in our ablations.

Ablation on other baseline: We have extended our method to another baseline, VALOR [37], a large-scale pre-training model tailored for visual language tasks. Specifically, in line with the other baselines, we employed both the visual and text branches from VALOR and proceeded to fine-tune video captioning on the MSR-VTT dataset, following the settings for VALOR-B model. Through our implementation, we achieve a CIDEr metric of 61.05. Then, by integrating IcoCap with VALOR on the MSR-VTT dataset, we observed a further enhancement in performance, which are 61.53 in the CIDEr metric. With the large-scale pre-training, VALOR already exhibits remarkable

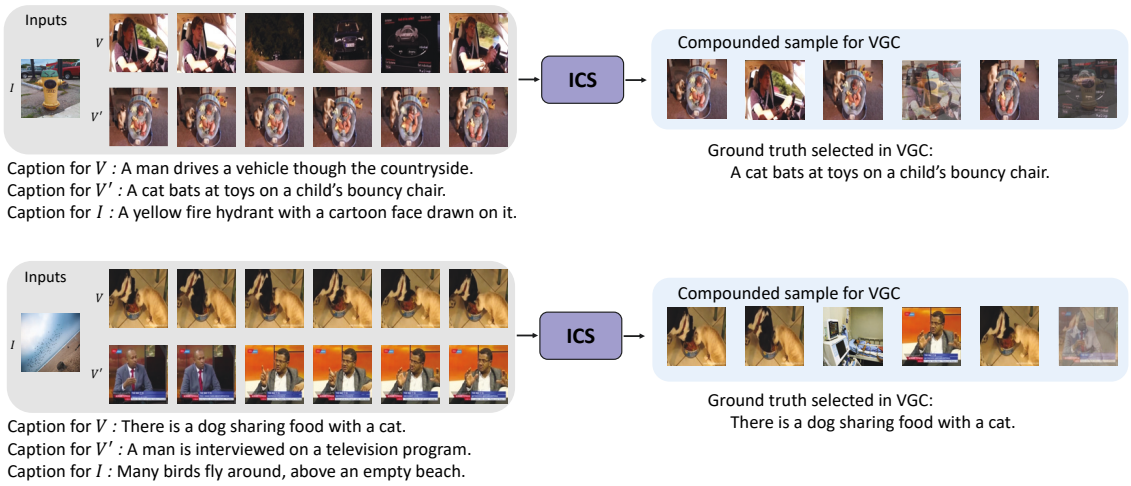


Figure 5.4: Examples for input videos and images, compounded video samples, corresponding captions, and ground truth selected by VGC in IcoCap.

Table 5.9: Performance of image captioning. The experiments are based on the test set of [143]. Only Image Set indicates only training with image set and without video set. The frame number is set as $n = 1$.

Data	BLEU-4	METEOR	ROUGE-L	CIDEr
Only Image Set	31.9	25.3	54.1	99.3
MSR-VTT + Image Set	28.0	23.5	50.5	89.1
MSVD + Image Set	32.4	25.8	54.4	101.1
VATEX + Image Set	23.3	21.4	46.4	73.9

improvements. Meanwhile, our methodology improves the captioning capability even further. The improvements underscore the generalization of our IcoCap.

5.3.7 Performance in Image Captioning

Since IcoCap introduces additional image data, we also report its performance in image captioning. To ensure a fair comparison, we train IcoCap with all the available training data and evaluate it on the test set of the MSCOCO dataset [143].

As shown in Table 5.9, training with samples compounded with image samples also empowers the model's ability in image captioning. All models trained with IcoCap can solve image captioning. However, due to the differences in video sets, the performances on the image set vary. Among the different video sets, training with MSVD [35] leads to the highest results across various metrics. On the other hand, due to the domain gap between image and video data, the performances of models trained with MSR-VTT [239] and VATEX [222] are lower. The larger scale and complexity of MSR-VTT and VATEX

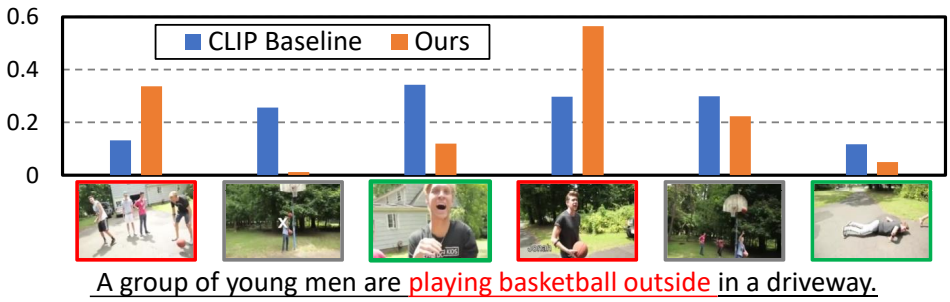


Figure 5.5: Comparison of attention weights in the captioner. The video captioner is a standard transformer model. We provide a comparison of the attention weights of the last attention layer for the video frames. keyframes, transitions, and irrelevant frames are marked with red, gray, and green borders, according to the caption below. IcoCap produces larger attention weights for the keyframes and lower weights for transitions and irrelevant frames.

may make it challenging for the network to learn complex video cues, thereby limiting the model's ability to improve semantic understanding.

5.3.8 Qualitative Analysis

Visualization for compounded samples: We provide detailed examples illustrating the application of ICS, exemplified in Fig. 5.4. This figure comprehensively presents illustrative instances, encompassing both input videos and images, the compounded samples by ICS, corresponding captions aligned with each input, and the definitive ground truth selected by VGC. These exemplars effectively spotlight two attributes of our approach: 1. The compounded samples showcase amplified diversity and reduced redundancy in comparison to the original inputs. These characteristics impose more demanding prerequisites on the captioner's learning process, thereby propelling the network to delve deeper into the realm of intricate visual content. 2. In IcoCap, the ground truth captions can be flexibly adapted based on the visual context. This phenomenon underscores the efficacy of our VGC in flexibly learning intricate visual contents.

Visualization for attention weights: In Fig. 5.5, we present a comprehensive visualization of the video frames, along with their corresponding attention weights after normalization. This illustration provides valuable insights into the attention mechanisms employed by our proposed method. Moreover, we offer a comparison of the attention weights for the video frames, specifically focusing on the last attention layer. Based on the caption provided below the figure, we have marked the keyframes, transitions, and irrelevant frames with red, gray, and green borders, respectively. Upon closer ex-

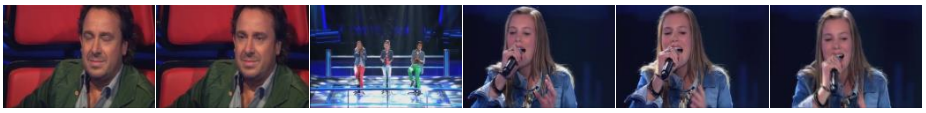
amination, it can be observed that IcoCap effectively assigns larger attention weights to keyframes, which are crucial for understanding the content, while assigning lower weights to transition frames and irrelevant frames. This demonstrates the ability of our method to effectively capture and emphasize the most relevant aspects of the video content, ultimately leading to better captioning performance.

Visualization of captioning results: Results for the baselines, SWINBERT, and our method are shown in Fig. 5.6. Due to the complexity of visual content in videos, models may be biased and produce sentences that do not holistically describe the overall content. Some results, marked in blue, only express a part of the content in the videos, which may relate to detailed and minor events in the video data but fail to describe the major and valuable events. Additionally, some inaccurate descriptions are generated due to ambiguous semantics, marked in orange, that try to describe and summarize the content but are misled by the complex content and do not correctly reflect the semantics in video frames. The diverse semantics in video samples may confuse the network, making it difficult for the network to understand video content and exclude irrelevant content. In comparison, our method effectively improves the performance of handling complex visual content. The captioning results from our method can more precisely describe the video semantics.

We present a comparison of the generated results in MSVD and VATEX datasets, as shown in Fig. 5.7. VATEX dataset is more linguistically complex, with more diverse and complex descriptions than MSVD. Although our IcoCap does not specifically address this issue, it still achieves comparable results to state-of-the-art methods.

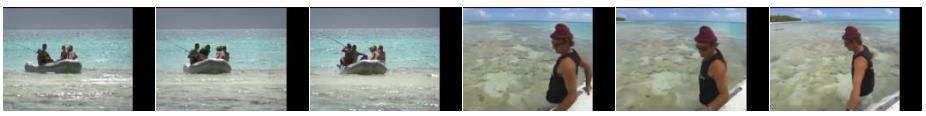
Moreover, benefiting from the compounded samples in our work, the network performs better with some complex video contents. For videos with multiple scenarios and characters (e.g., first row in Fig. 5.7 and first row in Fig. 5.8), our method is not misled by the complicated semantics and provides accurate descriptions.

Visualization of features: We employed t-SNE to visualize the features generated by ICS, as depicted in Fig. 5.9. On the left, we present the baseline features, which are from the original CLIP features. Conversely, the right showcases the features within IcoCap, formulated by the ICS. Given the inherent redundancy in the original video frames, the baseline features tend to be more compact, which are easier for network learning. Such compactness might lead a captioner towards overfitting and pose challenges in learning with intricate semantics. In contrast, the features presented in IcoCap are more diversified and intricate. Their distribution also poses a higher level of complexity compared to the baseline. This demands a more rigorous learning paradigm from the



CLIP Baseline: A women on the stage.
Joint Baseline: A man is interviewing an actor.
SWINBERT: A group of kids on stage.
Ours: A young girl is singing on the stage.

CLIP Baseline: A baby on a vehicle.
Joint Baseline: A women and baby talking about a vehicle.
SWINBERT: A women is talking about a baby trolley.
Ours: A women is giving a demo for baby trolley.



CLIP Baseline: A group of people in a boat.
Joint Baseline: A man is standing in the water.
SWINBERT: A man is walking on a boat.
Ours: A man pull a small boat in the water.

CLIP Baseline: A baby in a pink dress.
Joint Baseline: A young baby is running and playing with doll.
SWINBERT: A young girl is playing with doll.
Ours: A little girl in a pink dress and playing with toys.

CLIP Baseline: A man is cut an egg on the table.
Joint Baseline: A man is playing with an egg.
SWINBERT: A man is showing how to make a trick trick.
Ours: A man is playing with a yellow ball.

Figure 5.6: Comparison of generated captions on MSR-VTT dataset. To better illustrate the difference, we mark some results in blue, which only describe the detailed and minor semantics of the overall video. Some incorrect descriptions for the visual contents are marked in orange. Our method shows better performances against the diverse contents and ambiguous semantics in videos.

CLIP Baseline: A women and a man is working.
Joint Baseline: A woman and a man in a room.
SWINBERT: A woman is showing how a man is working on a machine.
Ours: A man is demonstrate how to use a machine.

CLIP Baseline: A boy is in a gym.
Joint Baseline: Two man are playing a game in a gym.
SWINBERT: A man is doing a leg exercise on a mat in a gym.
Ours: Two man are doing exercise in a gym.

CLIP Baseline: A person is walking down the street.
Joint Baseline: A man is talking in the street.
SWINBERT: A news reporter is talking about a news segment about a news segment.
Ours: A man is talking about a news segment about a news segment.

Figure 5.7: Comparison of generated captions on VATEX dataset.

captioner, urging it to achieve enhanced generalization for intricate visual semantics.

5.4 Conclusion

In this chapter, we propose the Image-compounded video Captioner (IcoCap), a method that introduces image samples into the training procedure of video captioning to address the issue of ambiguous semantics in video data. Due to the complexity and diversity of video contents, it is difficult for the network to learn valuable video semantics. In contrast, image samples possess concise visual contents and clear semantics, making them easier to learn. The video samples compounded with image samples possess more difficult semantics. The network should learn to mine valuable video cues to solve the complex semantics. Specifically, In IcoCap, we propose Image-Compounding Strategy (ICS), which compounds video samples with images. ICS leads the network to handle complicated

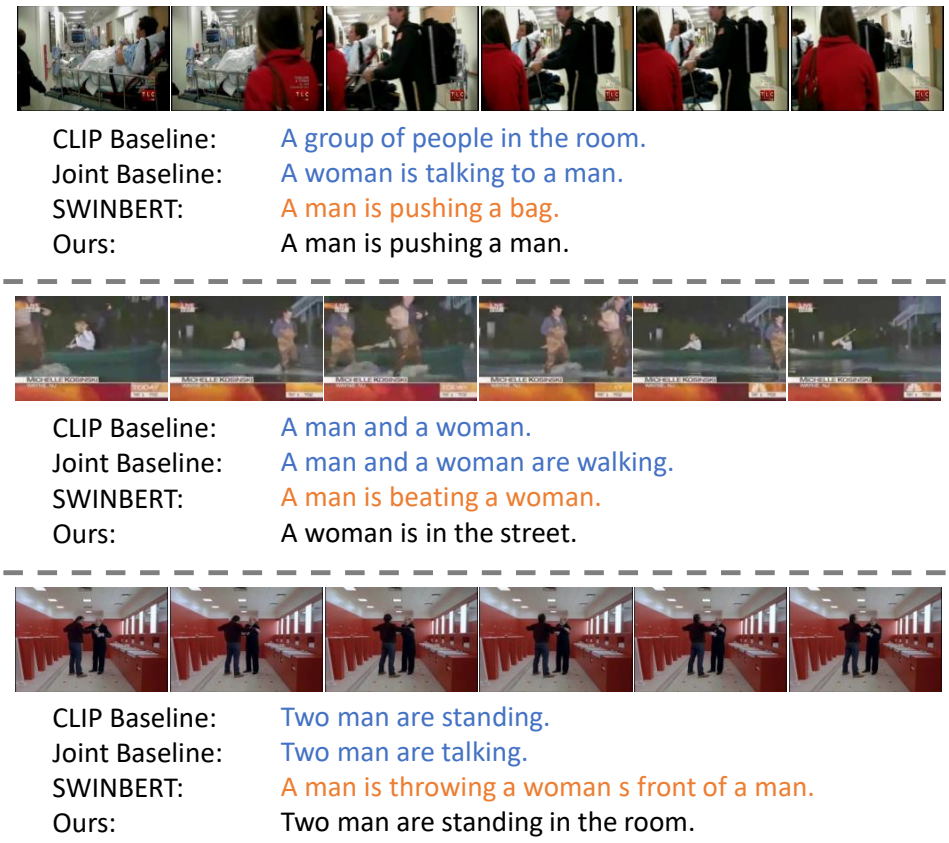


Figure 5.8: Comparison of generated captions on MSVD dataset.

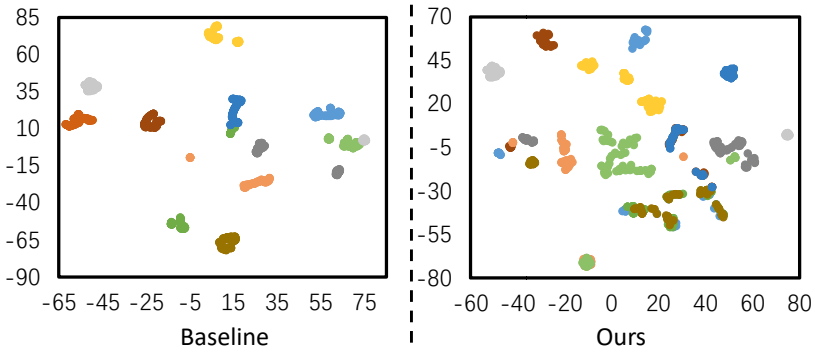


Figure 5.9: Visualization for features in baseline and IcoCap.

visual contents better and mine the valuable contents for captioning further. Besides, IcoCap also includes Visual-semantic Guided Captioning (VGC), which leads the network to learn the diverse video semantics flexibly. Experiments in various datasets prove the efficacy of our method. With a simple transformer network, we achieve comparable and even better performances in video captioning than the state-of-the-art methods.

CHAPTER 6

TOWARD BETTER ACCURACY FOR SEMANTIC-AWARE POSE GENERATION

6.1 Introduction

Recently, in synthesizing digital humans, vivid gestures can primarily improve reality, naturalness, and efficient information expression. Especially, talking gestures provide nonverbal cues of semantic expression and emphasize highlights and attitudes woven into our daily communication. Along with digital manipulation techniques, the speech-driven gesture is an emerging application, *e.g.*, digital human animation, visual dubbing in movies, online service, and education. The goal is to simulate artificial embodied agents to perform harmonious gestures aligned with the speech contents [75, 137, 182, 249]. Automated speech-driven gesture generation studies the generation of natural gesture sequences by exploring the relationships between speech and body language. It provides a new opportunity for realistic human-human interaction in virtual platforms.

Toward vivid speech-driven gestures, an intuitive expectation is to produce gestures corresponding to the speech contents. Humans naturally respond to their speeches and produce gestures to deliver specific semantics as in human ethology. As shown in Fig. 6.1, most co-speech gestures are compounded by beat and semantic gestures [32, 76]. Beat gestures are irrelevant to lexical semantics. It is independent to the content of the speech and prefers to respond to the rhythms of sounds. For example, the fast-talker tends to

This chapter is based on joint work [130] with Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang, presented primarily as it appears in the CVPR 2022 proceedings.

move more frequently in speak gestures. Semantic gestures[1] are apt to express certain speech content with body language, including iconic gestures, metaphoric gestures, and deictic gestures [32]. For example, speakers may raise their hands to emphasize their attitudes, corresponding to "clearly", "definitely", etc. Generating semantic gestures would lead to a vivid and reasonable content-based gesture rather than simply following the beat. However, the prior works of co-speech gestures synthesis [121, 182, 249, 250] do not explicitly produce semantic gestures and fail to model the lexical-semantic relevance between speech and gestures. For instance, when merely learning with the semantic-irrelevant cues, *i.e.*, the rhythms of audio and speakers' identities, we achieve a comparable score with state-of-the-art methods [249]. This indicates that the current methods are hard to learn semantics explicitly and produce semantic-aware gestures.

It is challenging to generate semantic gestures for the following two reasons. **First**, semantic cues for generating semantic gestures are hard to be mined. The styles and the movements of semantic gestures vary widely among speakers according to different contents. Meanwhile, beat gestures are inclined to intuitive and straightforward responses to the cues from sound, which commonly occur and are easier for the networks to mine. This difference induces semantic cues that are hard to be mined. The network may be relatively inclined to beat gestures and be slacked to investigate semantic cues. **Second**, semantic gestures and their corresponding texts are not well aligned temporally. As shown in Fig. 6.2, some gestures may be performed before or after the semantics they conveyed. This leads the network to unfavorably learn semantic gestures since it is hard to receive an explicit hint of semantic correlation via the given data. These two challenges hinder the generation and expression of semantics in gestures.

This chapter introduces a novel method to achieve semantic-aware co-speech gesture generation named SEmantic Energized Generation (SEEG). SEEG efficiently mines semantic and beat cues respectively and conducts semantic-aware gesture generation. Specifically, SEEG contains two components, *i.e.*, DEcoupled Mining module (DEM) and a Semantic Energized Module (SEM). **DEcoupled Mining module** decouples speech input cues into semantic-relevant cues (closely coupled to speech contents) and semantic-irrelevant cues (only beat information). Then, two separate encoders in DEM process Semantic-relevant cues and semantic-irrelevant cues to understand information for semantic and beat gestures. After input decomposition, one encoder focuses on the representation for beat gestures, while the other encoder exploits the diverse semantic

[1] We collectively refer to the three kinds of gestures as semantic gestures to distinguish them from the beat gesture.

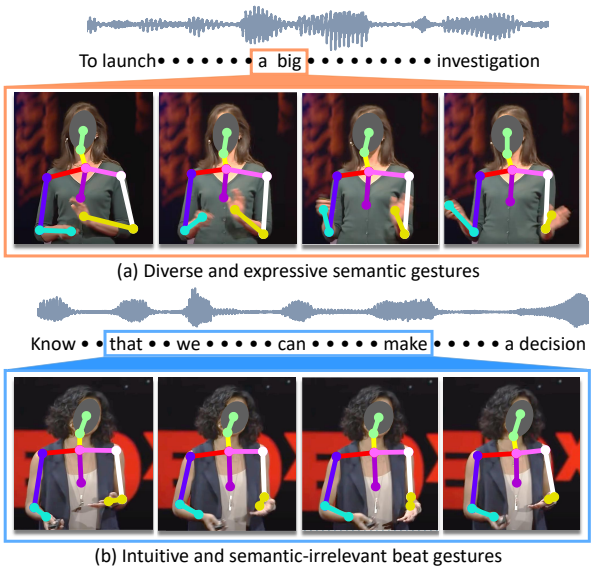


Figure 6.1: Co-speech gestures comprise semantically irrelevant beats and a variety of semantic gestures. SEEG explores both types of gestures and generates more accurate semantic gestures.

information for semantic gestures. This process eases the learning of semantic and beat gestures with huge disparities. The networks enable explicitly mine differential information for the beat and semantic gestures. If we expect the networks to learn semantics, DEM avoids forcing the networks to learn semantics from beat gestures that do not contain semantic denotations. **Semantic Energized Module** aims to avoid generation degrading to beat gestures. SEM energizes semantic learning by constraining two kinds of similarities: representational similarity and semantic similarity. Representational similarity requires the generation to be similar to the ground truth in appearances. More critically, DEM pursues semantic similarity and encourages the results to present similar semantics compared with the ground truth. In DEM, we additionally introduce a semantic prompt gallery and a semantic prompter network. The prompter is trained by the gallery and fix it in gesture generation. The prompter network is responsible for representing gestures in a semantic view. By producing similar representations under the view of the prompter, the generated gestures are regularized to align semantics conveyed from the ground truth. Rather than directly connecting speech contents to gestures that may be misaligned, SEM energizes semantic learning by restraining both representational similarity and semantic similarity.

Our main contributions can be summarized as follows:

1. We propose a new SEmantic Energized Generation (SEEG) framework for co-

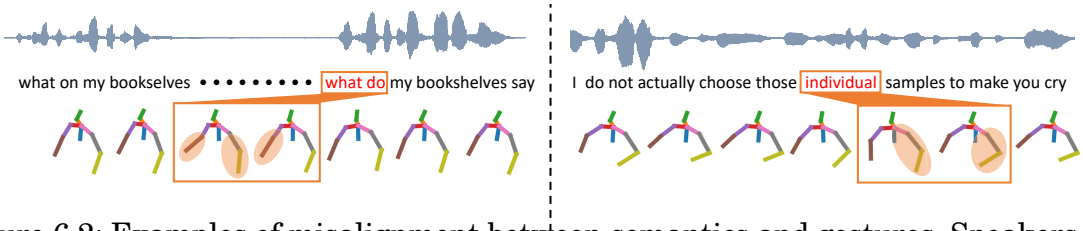


Figure 6.2: Examples of misalignment between semantics and gestures. Speakers may perform semantic gestures before (left) or after (right) the target contents. This leads to the semantic gestures being hard to match in temporary corresponding to the text or audio. We highlight the significant gestures with the orange shading.

speech gesture generation. SEEG is a semantic-aware gesture generation method that is adept at generating gestures with better semantic expressiveness.

2. We propose DEcoupled Mining (DEM) and Semantic Energized Module (SEM). DEM decouples semantic-irrelevant cues in inputs and eases the learning of disparate semantic and beat gestures. DEM encourages the network to learn semantics and produce semantic gestures.

3. In generating semantic gestures, the efficiency and advantages of our method are revealed by three subjective metrics on different datasets and objective human evaluations. We also find that the beat gestures may dominate the co-speech gesture generation. Visualizations show that SEEG achieves significant expressiveness in semantics.

6.2 SEmantic Energized Generation

We propose SEmantic Energized Generation (SEEG) to empower the learning of semantics in co-speech gesture generation. As shown in Fig. 6.3, SEEG contains two parts: DEcoupled Mining module (DEM) and Semantic Energized Module (SEM). DEM decouples semantics from inputs and contains two encoders for different inputs correspondingly. The two decoders are responsible for explicitly mining information for beat and semantic gestures. Moreover, SEM involves a semantic prompter and a gesture decoder. The decoder provides the final outputs for gesture generation. Then, the prompter network leverages an aligning loss for gestures which relieves the misalignment for semantics.

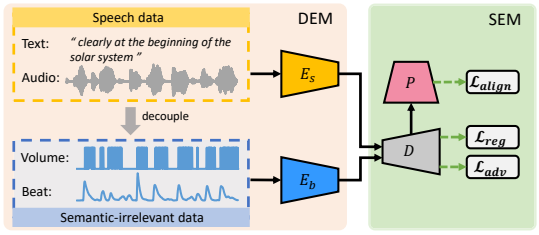


Figure 6.3: An overview of our semantic-aware gesture generation. It contains two parts: DEcoupled Mining Module (DEM) and Semantic Energized Module (SEM). Two encoder networks (E_s, E_b) and a decoder network (D) are designed to learn beat and semantic information and produce gestures comprehensively. Another prompter network (P) encourages the networks to learn and generate semantic gestures.

6.2.1 Preliminary

According to the speech data, co-speech gesture generation aims to generate vivid gestures as real speakers. Some works [137, 164, 182] synthesize body gestures, hand gestures, lips, or face key points by taking audio, text, and speaker identities as pre-processed inputs. In this work, we focus on generating upper body gestures by sequentially outputting the key points following [249, 250].

Taking the audio and text as inputs, methods are required to produce vivid speech gestures like real speakers. Generally, methods in this topic also introduce person ID and encode the ID into features. Additionally, the text is pre-processed and represented by pre-trained word vectors [22, 110, 173]. Thus, there are three parts of inputs: audio data x_a, text data x_w, and ID x_i. Then, the final output is the sequential gestural data denoted as $\hat{y}$. It contains the locations of key points for gestures in every time step. Besides, the ground truth gestures y are also extracted from videos and pre-processed [249, 250]. All x_a, x_w, y, and $\hat{y}$ correspond to the time step t.

Moreover, we focus on energizing the gestures with better semantic expressiveness in this work. Instead of generating gestures resembling the ground truth, we emphasize producing semantic gestures conveying similar semantics as the ground truth.

6.2.2 DEcoupled Mining module

In speech gestures [15, 32, 76, 98], beat gestures are intuitive and relatively simple. Semitic gestures are diverse and demand semantic understanding. These induce that the beat cues are easier to be investigated, and the semantic gestures may be ignored in the generation. Then, the method may be trapped in the beat gestures. In our work,

we first propose the DEcoupled Mining module (DEM) to learn information for semantic gestures and beat gestures separately and explicitly.

In the speech data, text corresponds to the speech content and is related to the semantics. Meanwhile, audio data reflects the pronunciations, emotions, accents, beats, volume, etc. Some factors in audio merely support semantic expression and do not convey particular semantics. Specifically, the beat and volume of the audio correspond to the rhythm and speed of the speech. They are semantic-irrelevant, and the listener cannot realize the semantics only by the beat and volume. Thus, we decouple these factors to the semantic-irrelevant information, which leads to the beat gestures.

Specifically, as shown in Fig. 6.3, we decouple the input that consists of audio amplitudes and audio onsets, which stand for volume and beat, respectively. For volume information, the audio data with large amplitude values possess large volumes. We defined the volume function as:

$$\mathscr{A}(x_a,t)=\begin{cases}1 & x_a(t)\geq \frac{1}{T}\sum_t^T x_a(t)\\ 0 & x_a(t)<\frac{1}{T}\sum_t^T x_a(t)\end{cases} \tag{6.1}$$

where x_a is the amplitude of the audio data, t is the time step, and T is the overall length. We set $\mathscr{A}(x_a,t)=1$ if the amplitude is larger than the average and vice versa. This is because the audio data contains noise and background sound. The amplitude larger than the average indicates that the speaker starts to speak apparently.

Moreover, it is difficult to capture the changing of intonation or speed of the speaker only using volume signals. We introduce the onset strength envelope [64, 65, 155] to represent the beat information. Onset [64, 65] refers to the start points of the sound. The strength envelope [21] can indicate the probabilities of the onset detected in the audio signal. This can represent the beat of the speech audio. We follow [21, 155] to extract the onset strength envelope and denote it as $\mathscr{O}(x_a)$ in our work.

In DEM, two encoders E_s and E_b are proposed to mine the information for semantic and beat, respectively. In detail, for beat gestures, E_b utilizes $\mathscr{A}(x_a,t)$ and $\mathscr{O}(x_a)$ as inputs. For semantic gestures, E_s is designed to learn from x_w and x_a. Besides, as the standard settings in [3, 249], we also add person ID x_i as inputs for encoders.

The procedure of DEM can be formulated as:

$$\begin{aligned} z_s &= E_s(x_w,x_a,x_i),\\ z_b &= E_b(\mathscr{O}(x_a),\mathscr{A}(x_a),x_i)\end{aligned} \tag{6.2}$$

where z_s and z_b are the features for semantic and beat, respectively. Moreover, both encoders possess similar network structures. They all contain three fully-connected layers

to handle the inputs. Then, two additional fully-connected layers and concatenation operations are utilized to merge three kinds of inputs. Next, a four-layer GRU network is designed to learn the sequential features produced from the above fully-connected layers. More details for the networks are displayed in the supplementary.

6.2.3 Semantic Energized Module

After mining information for semantic and beat gestures in DEM, we designed a Semantic Energized Module (SEM) to further energize semantic learning against the problem of misalignment. First, we introduce a semantic prompt gallery from the TED dataset [250]. Then, we propose a semantic prompter to learn the gallery individually. The prompter can formulate semantic representation for gestures. Through the prompter, we further leverage supervisions to predictions. This encourages the network to pursue similar representations of semantics by prompter that avoids the network learning misaligned semantics directly.

Semantic Prompt Gallery: The semantic prompt gallery is a small text-gesture collection. It contains five general classes from [15, 32, 52, 76, 98]. We take three noticeable semantics (Listing, emphasize, deictics) conveyed from gestures and two classes (negative, positive) to reflect the speakers' feelings and attitudes. The gallery is denoted as $\mathscr{G} = \{\mathscr{C}_{Listing}, \mathscr{C}_{Emphasize}, \mathscr{C}_{Deictics}, \mathscr{C}_{Negative}, \mathscr{C}_{Positive}\}$, where $\mathscr{C}_*$ is a text-gesture set, and $\mathscr{C}_* = \{[v_1, v_2, \ldots, v_M]; [g_1, g_2, \ldots, g_N]\}$. v_i and g_i denote a word and a gesture sequence, respectively. Moreover, we apply M words from [15, 32, 52, 76, 98] to construct the text set for each class as v. Besides, [98] presents a versatile collection and collecting method for semantically-congruent gestures. Following [98], we collect N gesture sequences from the TED dataset [250] for every class to formulate g. More details will be presented in supplementary.

Semantic Prompter: We propose a semantic prompter to learn the above gallery independently. As shown in Fig. 6.4, the semantic prompter P adopts gesture data as inputs and learns to classify gestures into five general semantic labels in the gallery. P consists of two fully connected layers and a four-layer GRU network, in which the fully-connected layers are utilized to process inputs and outputs. The GRU aims to model the sequential inter-connection of gestures. In all, the prompter can reflect the semantics of gestures and represent the gestures in the semantic view.

Semantic Energized Learning: As shown in Fig. 6.3, a gesture decoder D is proposed to aggregate both features from E_s and E_b and produce gestures as the final outputs, which can be described as $\hat{y} = D(z_s, z_b)$, where $\hat{y}$ denotes the final predictions.

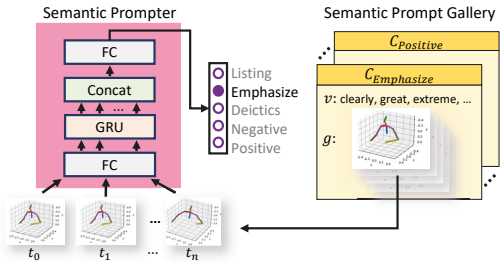


Figure 6.4: Construction and training of the semantic prompter. The semantic prompter is learned from the semantic prompt gallery. FC, Concat, and GRU denotes the fully-connected layer, concatenate operation, and GRU network, respectively. t_* indicates the time step of gesture data. The semantic prompter learns from the semantic prompts and bridges general correspondences between gestures and semantics.

D aims to decode gestures considering both information of beat and semantic. It is constructed by a single fully-connected network. Then, to energize semantic learning, SEM leverages two kinds of supervision for prediction $\hat{y}$: representational similarity and semantic similarity.

For representational similarity, we constrain P to be similar to the ground truth directly. The regression loss $\mathscr{L}_{reg}$ and adversarial loss $\mathscr{L}_{adv}$ are applied. $\mathscr{L}_{reg}$ [249] contains a smooth L1 loss to reduce the distances between y and $\hat{y}$. Meanwhile, the Kullback-Leibler (KL) divergence is included in $\mathscr{L}_{reg}$ to constrain the person ID. Besides, the same discriminator as [249] is added to perform adversarial learning for generated gestures. This also targets the representational similarity of predictions and the ground truth [249].

More important, for semantic similarity, we further propose the semantic aligned loss $\mathscr{L}_{align}$. Considering the semantic misalignment, indicating or annotating semantics to particular words may not be proper. In our work, we propose to align semantics conveyed from the gestures. In other words, we encourage the generated results to perform similar semantic representations as ground truth gestures. To this end, we apply the prompter P to represent gestures of predictions and the ground truth and propose a semantic aligned loss $\mathscr{L}_{align}$ to regularize:

$$\mathscr{L}_{align}(\hat{y}, y) = |P(\hat{y}) - P(y)| \tag{6.3}$$

where $|*|$ is the smooth L1 normalization. As P is fixed in training, to solve the above loss function, the output gestures $\hat{y}$ should reveal similar semantic representations with the ground truth y under the view of P. $\mathscr{L}_{align}$ does not regulate the predictions to be identical with the ground truth or particular gestures, and it requires similar semantics.

In all, the final loss function $\mathscr{L}$ can be formulated as:

$$\mathscr{L} = \mathscr{L}_{reg} + \mathscr{L}_{adv} + \mathscr{L}_{align} \tag{6.4}$$

6.3 Experiments

In this section, we discuss the details of SEEG and evaluate SEEG with various metrics in different datasets.

6.3.1 Experimental Setup

Implementation Details: Our network designs follow the structures of the generator in [249] and only change some fully-connected layers to fit the inputs. To perform a fair comparison, all the other settings, like the optimizer, learning rate, etc., are the same with [249]. Besides, for training the prompter network, we utilize random clipping, random resizing, and cutmix [255] to augment the gestures in the gallery. We train the prompter network with 100 epochs with the SDG optimizer and learning rate 0.001.

In addition, we collect the semantic gallery with $M = 25$ and $N = 5$. To be noticed, there are two significant differences between our semantic gallery and word-pose dictionary in previous work [137]. 1). Only general classes for semantics are defined. No specific words map particular gestures. This property avoids the misalignment between words and gestures in the gallery. 2). The gallery is only applied to train P. It is not practical and not necessary to collect a comprehensive dictionary for training. The prompter network is not responsible for recognizing all possible semantics in gestures. It only needs to reflect some generally possible semantics in the gallery.

Datasets: We test our method based on the TED dataset [250], the current largest and standard dataset for speech-driven gestures [249, 250]. As in [249], it is constructed based on TED videos and contains the 3D pose data extracted from the videos. The dataset also includes the speech audio and transcribed speech text [249].

Besides, some gestures in the TED dataset are not expressive and may not convey explicit semantics. Meanwhile, some introverted speakers may not tend to provide apparent movements in speech. To reflect the improvements in the semantic aspect, we provide a Semantic-aware test set (SatTED) based on the above dataset in [249]. Specifically, we re-rank the testing set of the TED dataset based on the confidences of P and collect about the top 50% data as SatTED. The original test set in [249] contains 25,930 samples. Our SatTED includes 12,000 samples and more than 7.5 hours. We

compare methods in the SatTED and further discuss the superiority of our method in the semantic aspect.

Evaluation Metrics: We evaluate our method based on three metrics:

1) FGD: evaluating the distances between the features of predictions and the ground truth. It robustly reflects the similarity between gestures in appearances.

2) Diversity metrics [79]: the measurement of diversity and flexibility. As expressive speakers tend to provide various gestures to support their expressions [76, 98], this metric can reflect the naturalness and semantic correlation to some degree.

3) Semantic-Aware Accuracy (SAA): we additionally propose a Semantic-Aware Accuracy (SAA) as the measurement for semantic expressiveness. With the semantic prompter, we can label the predicted gestures for semantic classes. Meanwhile, for the speech content, the semantic label can be assigned by voting. For every word in a sentence of the speech, we search the most similar description v in the gallery and assign the corresponding class C_* as the label of this word. After voting for every word, we select the class with the highest voting value as the label for the current sentence. Then, with the labels of gestures and sentences, we calculate the accuracy as SAA.

It is worth noting that $\mathscr{L}_{align}$ supervised the semantic expressions of predicted gestures and the ground truth gestures, which avoid the problem of misalignment. It does **NOT** supervise that the gestures should correspond to the text. Meanwhile, SAA describes the text-gesture correlation. This is a higher requirement since the ground truth may also not reflect the semantics closely. SAA measures the semantic expressions in an ideal condition that all gestures are semantic gestures.

Subjective Evaluation: We perform the user study through actual humans to evaluate the gestures. We random sample 20 pieces of speech audio, text, and the gestures of actual humans, Trimodal Context [249], and ours. Then, we publish these as the questionnaire for 50 different people to grade the gestures by three factors: naturalness, speech-gesture correlation, and gesture frequency. The factors are commonly used in gesture evaluation as in [228]. The range of grades is from 0 to 10. We collect all the questionnaires and calculate the average marks in experiments.

6.3.2 Quantitative Evaluation

Comparisons with state-of-the-art models: We first compare the values of FGD based on the TED dataset. We train the encoder E_s with decoder D individually, generating gestures based on semantic-irrelevant data without the prompter network. This corresponds to the generation of beat gestures. As shown in Table 6.1, With E_s + D only,

Methods	FGD (↓)
Seq2Seq [250]	18.154
Speech2Gesture [75]	19.254
Language2pose [4]	22.083
Trimodal Context [249]	3.729
Ours (E_b + D only)	3.751
Overall SEEG	6.244

Table 6.1: The performance of different methods for co-speech gesture generation in the TED dataset. We adapt FGD as the evaluating metrics. The performances are comparable even only using encoder E_b and decoder D in our method. Note that FGD may **NOT** well reflect the gesture semantics. The evaluations on gesture semantics are presented in other tables.

Dataset	Method	FGD (↓)	Diversity (↑)	SAA (↑)					
				Emphasize	Listing	Deictics	Positive	Negative	Average
TED	Real Gesture	-	$1.405^{\pm 0.058}$	52.135	41.028	65.515	19.388	27.255	37.688
	Trimodal Context [249]	3.729	$0.759^{\pm 0.029}$	32.496	43.203	51.647	17.021	29.600	30.286
	SEEG	6.244	$1.059^{\pm 0.045}$	40.438	44.465	66.116	19.004	27.246	36.851
SatTED	Real Gesture	-	$1.271^{\pm 0.056}$	54.709	64.169	82.587	22.522	29.052	43.904
	Trimodal Context [249]	4.505	$0.782^{\pm 0.037}$	32.928	55.612	61.844	12.833	21.496	30.956
	SEEG	7.451	$1.118^{\pm 0.049}$	44.518	52.322	70.461	21.322	27.763	38.457

Table 6.2: Comparison of all metrics in the TED dataset and SatTED dataset. Our method shows better performances significantly in some semantic-relevant metrics like diversity and SAA. Real Gestures indicate the gestures of real humans in the ground truth. ± means 95% confidence interval. ↑ indicates that higher values are better, and ↓ means lower values are better.

our result compares favorably to state-of-the-art methods in FGD, which utilizes comprehensive data from speech. This indicates that the network can achieve similar FGD to the recent method without mining any semantic cues. Only by mining the semantic-irrelevant data, the network can 'pretend' to produce meaningful gestures. Though we expect the network to learn semantics and produce expressive semantic gestures, the networks can also perform well without learning any semantics. This reveals two defeats in current research: 1). The beat gestures may dominate the dataset. Meanwhile, the semantic cues are hard to be mined with the comprehensive inputs. Thus, decoupled learning is valuable. DEM separately learn cues for beats and semantics, which guide the network not to be trapped in beat gestures. Besides, rather than the method side, a new sub-set with a larger ratio of semantic gestures is also required to uncover the semantic expressiveness of results. 2). FGD may be solvable in the current dataset by merely considering beat gestures. Merely measuring the distances between predictions and the ground truth is not enough. More semantic-aware measurements should be introduced. To solve the above defeats, the SatTED dataset and SAA are proposed in our

work.

Meanwhile, our overall method in FGD also outperforms previous methods with large gaps. Though slightly lower than E_s + D, our overall method also achieves competitive results than the current state-of-the-art. Since SEEG method is energized by SEM and tends to be more expressive and diverse, it may not completely follow the ground truth and focus on semantics.

Semantic-aware Evaluation: We compare all the metrics in two datasets as in Table 6.2. We also display all the semantic-aware accuracy in every class from the gallery. Results demonstrate that our method shows significant improvements in diversity and SAA than Trimodal Context [249], the current state-of-the-art in co-speech gesture generation.

Specifically, though the values of FGD are slightly lower, the diversity of our results is far better than [249]. With the SatTED dataset, the diversity of our method even approaches the real gestures of ground truth. Meanwhile, the semantics conveyed in our results are more recognizable and significant. Almost all values of SAA in every class and the average are better than Trimodal Context [249]. All these results show that SEEG is comparable in stimulating the gestures of actual humans and capable of understanding the semantics. Besides, SEEG achieves higher results than the ground truth in some categories of SAA since the ground truth may be beat gestures and do not respond to corresponding semantics.

In addition, the SatTED possesses a larger ratio of semantic gestures and is hard to be solved by the current method. As shown in Table 6.2, our method presents more significant improvements in this dataset. Results demonstrate that our method effectively boosts semantic learning for gestures and conducts a better semantic-aware generation.

Effect of Semantic Decouple: In our work, we decouple the semantics from inputs and enforce the networks to mine information for semantic and beat gestures separately. As in method design, we expect to achieve semantic gestures with E_s + SEM, beat gestures with E_b + D only, and the total outputs considering both sides (Overall). In this section, we experiment and verify the three parts as in Table 6.3. Specifically, we train E_b + D only with $\mathscr{L}_reg$ and $\mathscr{L}_adv$. E_s + SEM is trained with E_s + D with $\mathscr{L}$. Then, to show the interactions between E_s and E_b in the overall pipeline, we take the overall SEEG training from scratch with all modules and separately test each module. As Table 6.3, for E_b + D overall, we test the results by padding features z_b from E_b with zero. Similarly, E_s + D in overall pads features z_s with zero.

As shown in Table 6.3, E_b + D only achieves higher performances in FDG metrics

Dataset	Method		FGD (↓)	Diversity (↑)	SAA (↑)
TED	E_b + D only		3.751	$0.984^{\pm 0.044}$	30.022
	E_s + SEM		7.805	$1.113^{\pm 0.051}$	37.259
	Overall	E_b + D	5.472	$0.901^{\pm 0.045}$	30.597
		E_s + D	7.320	$1.127^{\pm 0.047}$	39.981
SatTED	E_b + D only		5.114	$0.922^{\pm 0.384}$	33.986
	E_s + SEM		9.291	$1.164^{\pm 0.049}$	44.218
	Overall	E_b + D	5.490	$0.990^{\pm 0.326}$	34.344
		E_s + D	6.797	$1.128^{\pm 0.049}$	46.533

Table 6.3: Comparison of different training manners. E_b + D only indicates that training individually with E_s and D without P. E_s + SEM denotes only training without encoder E_b. Overall means training with the complete method. Meanwhile, E_b + D indicates inferring the overall method with padding features from E_b as 0. E_s + D is inferring with padding features from E_s.

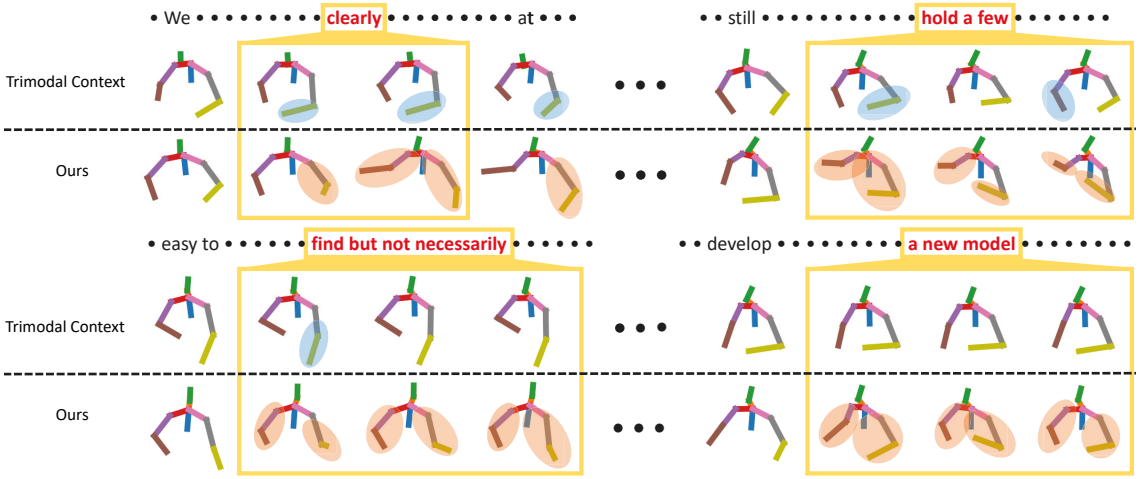


Figure 6.5: Examples of generated gestures. Our method shows better semantic expressiveness and conspicuous and reasonable responses to corresponding words. We highlight the significant gestures for [249] and ours with the blue and orange shading, respectively.

but shows significant decreases in diversity and SAA since it is unavailable to learn semantics with semantic decoupled inputs. Meanwhile, the isolated training with E_s and D tends to learn semantics only and may not perform similarly to the ground truth. This leads the results to obtain significant improvements in SAA but becomes worse in FGD. Moreover, in the overall pipeline, similar regularities also occur compared with training individually. In comparison, the learning of two parts would not be too radical. As a part of the overall pipeline, both E_b and E_s acquire improvements.

Ablation Study for Semantic Prompter: SEM relies on the semantic prompter to learn semantics in gestures. The impact of the prompter network for semantic learning

Method	Metrics		
	FGD (↓)	Diversity (↑)	SAA (↑)
Overall w/o T_s	4.937	$1.004^{\pm 0.037}$	30.920
E_s + D w/o T_s	3.915	$0.854^{\pm 0.037}$	30.216

Table 6.4: Ablation study for effect of the semantic prompter. Without the semantic prompter, the performances of diversity and SAA degrade.

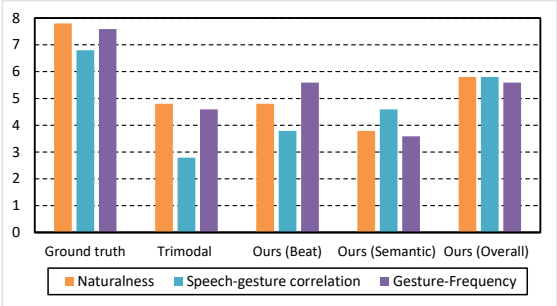


Figure 6.6: User study for synthesized gestures. The ground truth, current state-of-the-art, and our methods are compared based on three evaluating factors.

is explored in this section. We experiment with the SEM and overall pipeline with or without a semantic prompter, respectively. As shown in Table 6.4, without the semantic prompter, both semantic-aware performances like diversity and SAA degrade. Meanwhile, removing the prompter network leverages the improvements in FGD. The individual E_s + D without a prompter network performs similarly to the method in [249].

6.3.3 Qualitative Evaluation

Subjective Evaluation by User Study: We collect questionnaires from different volunteers and compute the average scores in different factors. The factors are all regular questionnaire items as in [228]. The statistical results are shown in Fig. 6.6. To investigate the performances of parts in our method, we train E_b + D only as of the beat gestures of our method (Beat), E_b + SEM as the semantic gestures of our method (Semantic), and the entire method (Overall), respectively. We compare our method with the current state-of-the-art and the ground truth. In comparison, our method shows significant improvements in all three factors. Moreover, the semantic gestures perform worse in naturalist and frequency but achieve remarkable advantages in speech-gesture correlation. This corresponds to the design of SEM, which focuses on semantic learning and may deviate from the ground truth.

Visualization: We showcase the results of our method and compare them with the current state-of-the-art [249]. In examples of generated gestures, as shown in Fig. 6.5,

significant responses occur corresponding to some words (e.g., clearly, at the beginning, quit a, available, easy, first step). The visualizations prove that our method learns semantics better and generates vivid gestures with semantic expressiveness.

6.4 Conclusion

In this chapter, a novel method for semantic-aware gesture generation is proposed. The proposed method contains two parts: DEcoupled Mining module (DEM) and Semantic Energized Module (SEM). DEM decouples semantics from inputs and forces the network to mine information for semantic and beat gestures. SEM contains a semantic prompter to leverage semantic-based supervision for the networks and produces semantic gestures. Experiments in various metrics, user study, and visualizations prove that the proposed method learns semantics better and produces semantic gestures corresponding to the speech content.

CHAPTER 7

MULTI-MODAL LEARNING FOR REAL-WORLD PROBLEMS

7.1 Introduction

Recently, robots have been widely used in various applications in manufacturing, transportation, and other industries. Toward diverse tasks, a fundamental requirement is to interact with objects by robots. To this end, the robots need to understand real-world objects, use grippers or other manipulators in the robotic system, and interact with given objects in a given scenario. As a primary problem, the object affordance problem [84, 112] is conceptualized and summarized as the first step for the interaction of robots and objects. It aims to figure out where and how to interact with an object by the robot in a given environment. Many works [23, 160] propose various solutions to solve the affordance problem. However, due to the diversity of instances and complexity of practical robotic scenarios, the problem is still far from being resolved.

Specifically, recent works focus on the affordance problem of interacting with 3D articulated objects [63, 161]. Mo et al. [159] introduce a solid benchmark for learning to manipulate articulated objects. They construct a large-scale 3D articulated object dataset and formulates a standard benchmark for the 3D articulated object affordance problem. Wang et al. [227] consider the kinematic and dynamic uncertainties of objects. They design multiple critics to improve the understanding of hidden kinematic information in articulated objects. More works [160, 269] continuously emerge, pushing the frontier of

This chapter is based on joint work [132] with Xiaohan Wang, Linchao Zhu, and Yi Yang, presented primarily as it appears in the ICCV 2023 proceedings.

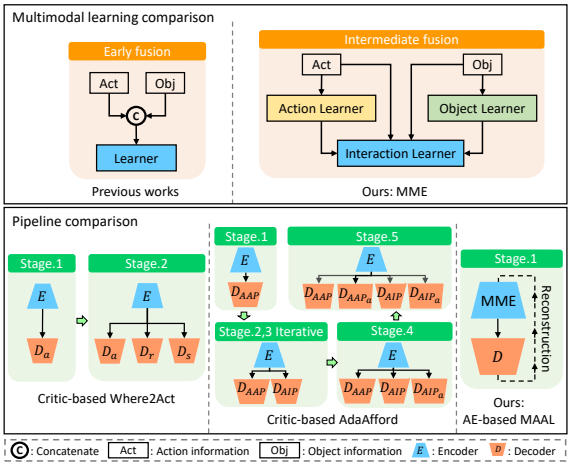


Figure 7.1: Comparison of methods. MAAL contains a MME module, which provides better multi-modal learning ability. Besides, previous methods with critics or decoders require multiple training stages. MAAL pipeline only contains one step and is trained in one go, which is more efficient.

solving the 3D object affordance problem.

Moreover, previous works can be concluded as early fusion [122] for learning multi-modal data and critic-based learning [159, 227] for 3D object affordance. Specifically, they usually concatenate all data (e.g., the point cloud of a 3D object, the robot gripper direction, etc.) as inputs. Then, multiple critics or decoders, trained by classification loss according to labels (negative or positive) initially, are introduced to leverage supervision for other networks.

The straightforward idea leads to significant advancements but still has two defeats. **First**, learning of inputs neglects the correlation between multi-modal data. In the 3D object affordance problem, the input data are from various modalities (i.e., object modality and robot modality). The relationships and interactions between objects and robots are valuable clues for understanding affordance [84, 112]. However, as shown in Fig 7.1, direct concatenation, as in [159, 227], considering as an early fusion operation [122], would miss the correlation between inputs [153, 248]. This leads that the multi-modal inputs and their interaction may not be sufficiently learned by the previous works. **Second**, the critic-based pipeline is not efficient enough. It requires adequately labeled samples to teach the critics to distinguish the difference between negative samples and positive samples [252, 259]. However, as in [159], training data of articulated object affordance are sampled from $SE(3)$ space, and most actions fail during manipulation. This means most of the samples are negative. For example, sometimes, only 1% [159]

of the data are positive samples for pulling action. Training of critic-based methods needs all the samples for training and consumes larger training time. Moreover, critics or decoders need to be trained independently. Then, they will be fixed or iteratively updated with the training of other networks, as shown in Fig 7.1. The training procedure with multiple stages further increases the overall training time.

To overcome above defeats, we present a novel solution named Multimodality-Aware Autoencoder-based affordance Learning (MAAL). In MAAL, a MultiModal Energized Encoder (MME) is introduced to handle multi-modal inputs in the affordance problem. MME energizes the multi-modal learning ability to understand 3D object affordance. Then, rather than a critic-based designation, MAAL leverages the deep autoencoder (AE) [77, 92] to solve the affordance problem and achieve better training efficiency.

Toward better multi-modal learning, MME is proposed to comprehensively understand data from various modalities and fused features at different levels. Specifically, it involves three branches, carefully designed for learning information in object modality, robot modality, and their interactions. This empowers MAAL to pursue a better understanding of affordance from different perspectives in modalities. Moreover, rather than directly concatenating all data and applying early fusion for various modalities, our encoder considers the correlation between inputs and fuses multi-level features according to the modalities. This can formulate better multi-modal learning than simply early fusion, as proved in [25, 163, 248].

Furthermore, MAAL introduces AE [77] pipeline to solve the 3D affordance problem more efficiently. AE can learn the valuable pattern [215, 252, 259] in high-dimensional data points without labeled examples [42, 80, 86]. This property leads AE can only use positive samples to learn specific valuable patterns from datasets. This also induces the better computational efficiency of the AE pipeline in solving the affordance problem. Besides, rather than learning representations with multiple critics, it only uses reconstruction loss [252, 259] as supervision. The overall pipeline can be trained in one go without multiple training steps for different parts. All these advantages lead that MAAL can achieve better training efficiency than previous critic-based works.

In addition to the above encoder, our MAAL has an action memory and an action decoder, which are used to formulate the AE pipeline. More than applying AE, MAAL specifically considers the properties of 3D object affordance, which takes object information as known conditions and aims to produce action proposals. Correspondingly, MAAL takes multi-modal data as inputs and only reconstructs action proposals as outputs. This leads the network to concentrate on learning action information and the interaction be-

tween robots and objects rather than remembering object information and overfitting to some points in objects. Overall, MAAL fully considers the multi-modal inputs, leverages the AE pipeline, and formulates a novel framework for learning 3D articulated object affordance.

Our main contribution can be summarized as follows:

1. We propose a novel pipeline named Multimodality-Aware Autoencoder-based affordance Learning (MAAL). It is an efficient framework for solving the 3D object affordance problem. MAAL does not need multiple training steps and only requires a few data samples compared to previous methods.

2. We propose MultiModal Energized Encoder (MME) to handle the multi-modal information and their interaction in the 3D object affordance problem. The proposed encoder comprehensively learns data in all modalities and provides better multi-modal learning ability.

3. Without bells and whistles, our method outperforms all current methods in both F-score and sample success rate. Visualizations also show the effectiveness of our MAAL.

7.2 Preliminary

Following the problem settings in [159], the 3D affordance problem can be generally formulated as where and how to act for a given 3D object. During training, 3D object information and interactive points are given as inputs. The methods are required to produce actionability scores for corresponding points, action proposals, and success likelihoods for proposals, respectively.

In detail, each input sample involves four kinds of data: x_o, x_p, x_a, and x_h. Specifically, x_o indicates the 3D object information represented by the 3D point cloud. $x_o \in \mathbb{R}^{\mathcal{O}\times 3}$, where $\mathcal{O}$ is the dimension of point clouds. x_p is the interactive point, and $x_p \in x_o$. x_a means an interaction proposal and can be described by gripper orientation $x_a \in SO(3)$. Finally, given gripper orientation x_a, articulated object x_o, and point x_p to the simulator, x_h is the part motion. It can indicate whether the action is successfully manipulated or not after simulation.

In this task, methods are required to:

- Given an object (x_o) and interactive point (x_p), produce an actionability score ϕ.
- Given an object (x_o) and interactive point (x_p), produce an action proposal ρ.
- Given an object (x_o), interactive point (x_p), and action proposal (x_a), produce a success likelihood score σ.

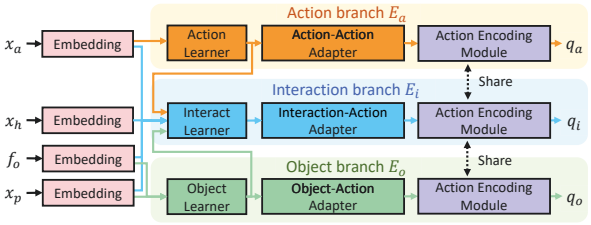


Figure 7.2: Structure of our MME. It contains three branches for learning different modalities. Features of different modalities with different levels are carefully fused in the interaction branch. MME provides better multi-modal learning for 3D object affordance. f_o is extracted by PointNet++ from x_o.

7.3 Method

We propose a Multimodality-Aware Autoencoder-based affordance Learning (MAAL) to solve the 3D object affordance problem. Specifically, MAAL contains three parts: a MultiModal Energized Encoder (MME), an action memory, and an action decoder. MME is proposed to learn multi-modal information, model the interaction and provide a comprehensive understanding of the inputs of the 3D object affordance problem. Then, action memory is used to record action information. Outputs from the encoder are taken as retrieval queries and are used to select items in the memory. Finally, given the aggregations of selected items from memory, the action decoder is proposed to reconstruct the corresponding actions.

7.3.1 MultiModal Energized Encoder

We propose MultiModal Energized Encoder (MME). MME empowers better multi-modal learning ability and solves the 3D affordance problem more effectively. Specifically, two kinds of modalities (object modality and action modality) and their interaction should be understood. Object modality mainly includes the point cloud of 3D objects and the points of the object for interaction. The action modality contains the gripper directions of the robot. Then, to model the interactions, object data, action data, and motion data from the simulator should be jointly considered. Although all the data are collected from the 3D space, there are still domain gaps among modalities: 1) Dimensional variations. The point cloud data in object modality has a dimension of $\mathbb{R}^{10000\times3}$. The gripper direction in robotic modality is a vector in $\mathbb{R}^{3\times3}$. 2) Physical property differences. Point cloud data are scalar values that indicate spatial information of objects. Robotic modality data are vectors and indicate the direction of the action. 3) Distinct networks in representation.

Different encoders or embedding layers are utilized to process various inputs, resulting in features with varying distributions, further enlarging the gaps between modalities. In our work, as shown in Fig 7.2, rather than directly processing all modalities by early fusion, MME contains multiple branches of networks to handle different modalities and carefully fuses features to learn the interaction.

First, following [159, 227], we use PointNet++ [180] network to encode the 3D point cloud of the object into feature f_o, where $f_o \in \mathbb{R}^C$ and C is the dimension of the feature. Then four embedding layers are introduced to embed action x_a, motion x_h, object feature f_o, and point x_p, respectively. All embedding layers learn individually and are built by two fully-connected layers.

Then, as shown in Fig. 7.2, we have three branches to learn multi-modal features and their interaction separately: the action branch E_a, object branch E_o, and interaction branch E_i. Each branch contains a learner module and an adapter module. Learner modules aim to learn information, particularly for each modality and interaction. Then, the adapters convert features from learners to adapt the action encoding module. Different branches in MME help the network to learn affordance with different perspectives. The network is encouraged to mine valuable clues for object affordance from every modality separately. This leads to comprehensive multi-modal modeling and would not neglect any modalities.

Specifically, in the action branch, the action learner module is proposed to learn features after embedding and is constructed by three fully-connected layers. Similarly, in the object branch, the embedded features from f_o and x_p are given to an object learner module. The object learner contains a batch normalization layer and three fully-connected layers. Moreover, the interaction branch takes all information from modalities and aims to learn the interaction between objects and robots further. It contains a bilinear network to model the interaction between features from the action learner and object learner. A residual connection block is also involved in merging features from all modalities. This designation introduces the better ability for multi-modal fusion [215, 248]. Features from different levels are considered and fused in the module. This provides a better understanding of information in multiple modalities.

Then, the adapters are introduced in the pipeline, which consists of two fully-connected layers. Finally, a shared encoding module generates query features from the different branches, denoted as q_a, q_o, and q_i, respectively. The procedure of MME

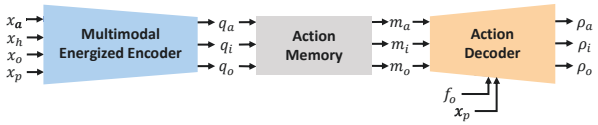


Figure 7.3: An overview of our Multimodality-Aware Autoencoder-based affordance Learning (MAAL). MAAL contains three parts: MultiModal Energized Encoder (MME), action memory, and action decoder. The encoder produces query feature q. The memory module receives queries, selects items, and aggregates them as m. Action decoder takes action information (f_o and x_p) and features m as inputs and reconstructs corresponding action x_a as ρ.

can be formulated as follows:

$$q_a = E_a(x_a), \tag{7.1}$$

$$q_o = E_o(x_o, x_p), \tag{7.2}$$

$$q_i = E_i(x_a, x_o, x_p, x_h, \theta_a, \theta_o). \tag{7.3}$$

where θ_a and θ_o are the features extracted from the action learner and interact learner, respectively. The feature dimension of all queries is C. More details are presented in the supplementary.

Moreover, other works directly use concatenated data (e.g., $[f_o, x_p, x_a, x_h]$ in [227], where $[*]$ is the concatenate operation.) as inputs. Taking all data as a whole, different modalities are learned equivalently. Comparatively, our encoder considers the learning of different modalities and their interaction. The encoder fuses multi-modal data at different levels and forms a comprehensive understanding. This leads our encoder to possess better multi-modal learning ability than the early fusion methods [159, 227].

7.3.2 Multimodality-aware Autoencoder-based Affordance Learning:

We propose Multimodality-Aware Autoencoder-based affordance Learning (MAAL). MAAL provides a more efficient pipeline to solve the affordance problem. As shown in Fig. 7.3, more than MME, we leverage a memory module M and a decoder module D to construct an AE pipeline. The memory module aims to prevent the "over-generalized" problem [77] in the original AE framework (only with an encoder and a decoder). Though only trained with positive samples, the original AE may also reconstruct negative samples with low reconstruction error during evaluation. By introducing a content-addressable memory, we do not directly provide encoder outputs to the decoder for reconstruction.

The representation from the encoder is used as a query to retrieve the most relevant item in action memory. Then, the selected memory features are aggregated and provided to the MAAL decoder. The memory module is a widely used strategy in AE, which has been applied and discussed in many works [9, 170, 187].

As shown in Fig 7.3, given q_a, q_i, and q_o, the memory module addresses memory items and aggregates them as m_a, m_i, and m_o, respectively. $m_a = M(q_a)$, $m_i = M(q_i)$, and $m_o = M(q_o)$. Finally, the decoder network is introduced to reconstruct action information. Given object information (f_o and x_p), it reconstructs the actions ρ_o, ρ_a, and ρ_i according to features m_a, m_i, and m_o, respectively. $\rho_o = D(m_a, f_o, x_p)$, $\rho_a = D(m_a, f_o, x_p)$, and $\rho_i = D(m_i, f_o, x_p)$. To be noticed, the decoder network also takes object information as inputs. This is because the 3D affordance problem treats object information as known conditions. Under the real scenario, the robots have to know the object information and then produce actions to interact. Moreover, the decoder is constructed by two batch normalization layers and five fully-connected layers. More details will be offered in the supplementary.

Generally, MAAL is not expected to memorize and reconstruct the objects precisely. The memory module only needs to record and represent action information. Given features selected by queries, the decoder is responsible for reconstructing action information according to known object information.

7.3.3 Training and Evaluation

The overall loss function $\mathscr{L}$ can be formulated as follows:

$$\mathscr{L} = \|x_a - \rho_o\| + \|x_a - \rho_a\| + \|x_a - \rho_i\| \tag{7.4}$$

where $\|*\|$ indicates the ℓ_2 distances of input actions x_a and action proposals ρ from every branch. The overall training loss consists of reconstruction losses for three queries, respectively. Only a single and end-to-end training step is required in our work, as in Fig. 7.1.

During the evaluation, the final goal of the affordance problem requires predicting action proposal ρ by given object information, actionability score ϕ by given object information, and success likelihood score σ by given action proposal and object information. The action proposal can be directly produced by reconstruction result ρ_o in MAAL. However, ϕ and σ are hard to be obtained directly through MAAL. They can be estimated according to reconstruction errors. Meanwhile, the reconstruction error in MAAL is an absolute error [152], which indicates that it may be variant by different data splits. To

overcome this problem, we additionally utilize the k-nearest-neighbor (KNN) algorithm to produce ϕ and σ.

In detail, we train the KNN algorithm using the average reconstruction error in the validation set. For every sample in the validation set, we have data x_a^{v}, x_o^{v}, x_p^{v}, and x_h^{v}, which indicate action, object, point, and motion data, respectively. Then, by MAAL, we achieve corresponding action proposals in the validation set, which are denoted as ρ_o^{v}, ρ_a^{v}, ρ_i^{v}. Thus, the reconstruction error e^{v} for a given sample in the validation set can be written as: $e^{\mathrm{v}} = (\|x_a^{\mathrm{v}} - \rho_o^{\mathrm{v}}\| + \|x_a^{\mathrm{v}} - \rho_a^{\mathrm{v}}\| + \|x_a^{\mathrm{v}} - \rho_i^{\mathrm{v}}\|)/3$. Then, we denote the KNN model as $\mathscr{K}$. $\mathscr{K}$ is trained by reconstruction error e^{v} from all the samples (including both positive and negative samples) and corresponding labels (binary labels indicate whether the actions can be successfully manipulated or not).

During the evaluation, we first achieve ρ_o^{t} by testing object data x_o^{t} and x_p^{t}. Then, the reconstructed action results of ρ_o^{t} can be calculated by:

$$m_a^{\mathrm{t}} = M(E_a(\rho_o^{\mathrm{t}})), \tag{7.5}$$

$$m_i^{\mathrm{t}} = M(\rho_o^{\mathrm{t}}, x_o^{\mathrm{t}}, x_p^{\mathrm{t}}, x_h^{\mathrm{t}}, E_a(\rho_o^{\mathrm{t}}), E_o(x_o^{\mathrm{t}}, x_p^{\mathrm{t}})), \tag{7.6}$$

$$\rho_a^{\mathrm{t}} = D(m_a^{\mathrm{t}}, x_o^{\mathrm{t}}, x_p^{\mathrm{t}}), \tag{7.7}$$

$$\rho_i^{\mathrm{t}} = D(m_i^{\mathrm{t}}, x_o^{\mathrm{t}}, x_p^{\mathrm{t}}). \tag{7.8}$$

where x_h^{t} is padded by zero. ρ_a^{t} and ρ_i^{t} are reconstruction results for ρ_o^{t} with action and interaction branches for testing. Then, for the current test sample, the actionability score $\phi = \mathscr{K}(\|\rho_o^{\mathrm{t}} - \rho_o^{\mathrm{t}}\| + \|x_o^{\mathrm{t}} - \rho_i^{\mathrm{t}}\|)/2)$. Similarly, for evaluating actions x_a^{t} in the test set, we can achieve reconstruction results ϱ_a^{t}, ϱ_i^{t}, and ϱ_a^{t} for x_a^{t}, respectively. Then, the success likelihood score can be computed as $\sigma = \mathscr{K}((\|x_a^{\mathrm{t}} - \varrho_a^{\mathrm{t}}\| + \|x_a^{\mathrm{t}} - \varrho_o^{\mathrm{t}}\| + \|x_a^{\mathrm{t}} - \varrho_i^{\mathrm{t}}\|)/3)$.

7.4 Experiment

In this section, we discuss all the details of our method design and task settings, evaluate our method with various metrics, and show the superiority and effectiveness of our work.

7.4.1 Experimental Setup

Implementation Details: Instead of training multiple critics and iterative training, all training procedures of our MAAL can be operated in one go. Specifically, the encoder, memory, and decoder modules are trained and updated at the same stage. Adam optimizer is used to optimize the networks within the learning rate 0.001 and weight decay 0.00001.

More details about the network design will be presented in the supplementary. The memory module is implemented following [77], which has been widely used in many works [9, 170, 187]. We set memory size N as 200, and the dimension C is 128. Ablations will be offered in Sec. 7.4.2. Other settings (e.g., training data generation, gripper data processing, simulator settings, etc.) follow [227]. Additionally, during evaluation, the number of nearest neighbors of the KNN classifier is 500. Due to space limitations, more details of network designs and ablations will be offered in supplementary. We will also provide more details and update the results of real-world experiments on Github .

Datasets: We experiment with all methods and operate comparisons based on PartNet-Mobility dataset [161]. It is a large-scale and standard dataset for 3D articulated object affordance problems and has been widely used in previous works [159, 160, 227, 269]. The action simulation is operated through SAPIEN simulator [234]. In this dataset, 972 articulated 3D objects within 15 object categories are used for conducting 3D object affordance tasks. There are ten classes for training and five classes for testing. Besides, the validation set is also split and contains ten categories same as the training set. For better comparison, we separately report the results for shapes with training categories and shapes with unseen novel categories, which are marked as "train cat." and " test cat." in tables, respectively. The data split is constructed following [159, 227]. Moreover, the 3D articulated object affordance task has six pre-defined actions ("pushing", "pushing up", "pushing left", "pulling", "pulling up" and "pulling left"). For a fair comparison, categories are split into "pushing all" and "pulling all" actions following [159, 227]. All actions are parameterized in the $SE(3)$ space according to the robot gripper poses. Corresponding to the actions, the training and test data samples are generated by the simulator.

Moreover, we also apply settings in [227] to evaluate some special categories and further show the effectiveness. We sample data from the doors category from pulling actions and faucet categories from pushing actions following [227]. This data split further shows the ability of methods to handle kinematic ambiguity. Besides, we also visualize the actionability scores to plot affordance heatmaps following [159, 227], which further prove the effectiveness of MAAL.

Evaluation Metrics: To evaluate and compare methods, we apply the two standard metrics in the affordance task as in [159, 227], which are F-score for success likelihood score and sample-success-rate (Sample-Succ) for action proposals. Since the generated actions are randomly sampled, the positive and negative samples may not be balanced. Thus, in [159], the authors introduced an F-score to balance precision and recall for

https://github.com/akira-l/MAAL

Dataset	Method	F-score (%)	Sample-Succ (%)
Pushing All (train cat.)	Where2Act [159]	66.29	27.33
	AdaAfford [227]	73.21	32.50
	MAAL	**76.63**	**34.25**
Pushing All (test cat.)	Where2Act [159]	52.38	21.04
	AdaAfford [227]	65.50	26.20
	MAAL	**69.88**	**28.34**
Pulling All (train cat.)	Where2Act [159]	48.76	6.40
	AdaAfford [227]	53.80	8.18
	MAAL	**59.26**	**10.47**
Pulling All (test cat.)	Where2Act [159]	40.88	5.71
	AdaAfford [227]	42.35	6.02
	MAAL	**43.57**	**6.67**

Table 7.1: The performance of the different methods for the 3D affordance problem in PartNet-Mobility dataset. Our method outperforms other methods in both data splits and metrics and also produces better action proposals than AdaAfford.

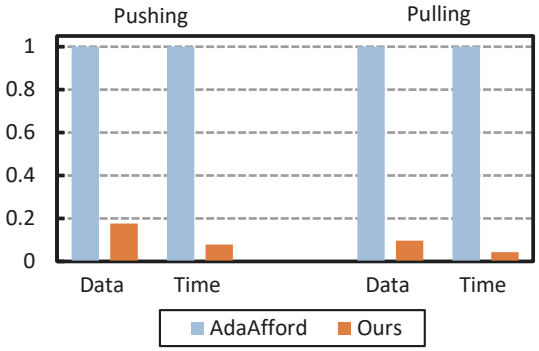


Figure 7.4: Comparison of data usage and training time. To better show the differences, we assume the data usage and training time of AdaAfford as 100% and calculate the relative percentages of MAAL compared with AdaAfford. Our method only consumes a small part of data samples and training times.

unbalanced samples. Then, Sample-Succ reflects the quality of proposals. It calculates the proportion of successfully manipulated actions among action proposals. Following [159, 227], we generate 100 candidates to compute the metric. We First select 100 points according to the actionability score ϕ in the given testing object. Then, we produce query q_o according to the object and point information and generate an action proposal. We experiment 10 times per testing object and report the average values of both metrics.

7.4.2 Results and Analysis

Comparisons with State-of-the-art Methods: As shown in Table 7.1, we first compare MAAL with previous works with four data splits following [159, 227]. Our method outperforms other methods in all data splits and metrics. The higher results reveal the

effectiveness of our method. The comparison shows the advantages of our method in two aspects. The higher values of F-score indicate that our method assesses the actions better. This proves that the reconstruction error from MAAL works well for evaluating actions. Without any critics and multiple training stages, MAAL can perform and even better complete this task. Besides, MAAL also achieves better performances in Sample-Succ. This reveals that the quality of our proposals is also better than the previous works. Moreover, in another data split from [227], our method also achieves better results, as shown in Tab 7.2. The performance gain reveals the effectiveness of our MAAL in solving the kinematic ambiguity.

Statistic for Data Usage: Due to the properties of AE, our MAAL only takes the positive samples (successfully manipulated actions in simulation) as inputs. To show the efficiency of our data usage, we statistic the percentage of positive samples in all training data. We produce data samples following [159, 227] three times and calculate the average proportion. Comparatively, our method only uses positive samples and is more efficient. As shown in Fig. 7.4, Our method only takes 17.69% data of AdaAfford for training pushing action. Meanwhile, in pulling action, the positive samples are mere 9.63%, and our method only requires such limited data samples. Moreover, our method also possesses lower training time. We compute the average time of 100 training epochs of different methods, as in Fig. 7.4. Due to the training procedure with multiple stages and more data samples, the training time of AdaAfford is 23.34 and 12.72 times than ours. All these results show the efficiency of our method.

Comparisons with Different Action Proposals: To compare the quality of action proposals, we take action proposals and actionability scores from different methods separately and combine them for comparison. Specifically, as shown in Tab. 7.3, the action proposals are provided by different methods. Where2Act-P and AdaAfford-P indicate using the action proposal parts in these methods. Where2Act-C and Adaafford-C mean using critics in these works, which are responsible for predicting confidence for action proposals. The action proposal from MAAL can be directly achieved by ρ_o, and we score the action proposals by reconstruction errors as in 7.3.3. Then, we select the top-100 action proposals by corresponding scoring modules and compute the Sample-Succ of selected actions.

Given proposals from different methods, action selections by MAAL achieve a higher or comparable success rate compared with others. This indicates that MAAL possesses a high ability to assess and score actions compared with other methods. Besides taking proposals from MAAL, other methods also achieve better Sample-Succ values. The results

Dataset	Method	F-score (%)	Sample-Succ (%)
Pulling Door	Where2Act [159]	58.26	12.84
	AdaAfford [227]	69.34	17.62
	MAAL	**70.39**	**18.27**
Pushing Faucet	Where2Act [159]	78.14	36.35
	AdaAfford [227]	81.62	39.89
	MAAL	**81.82**	**40.06**

Table 7.2: Comparison of categories selected by [227]. MAAL still achieves better results in these relatively harder categories.

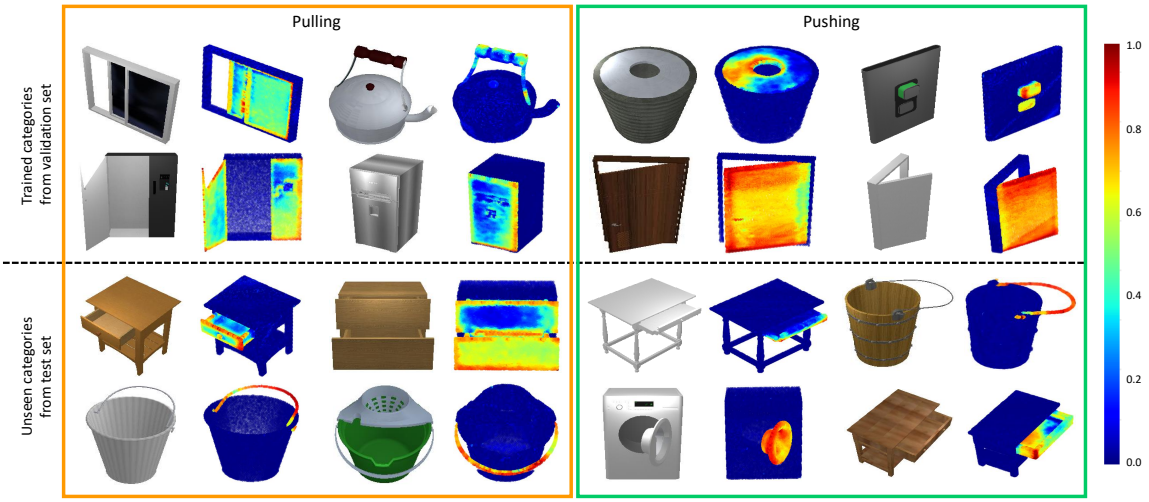


Figure 7.5: Visualization of affordance heatmap. All objects are from the test set. The heatmap is plotted by per-pixel action scores and produced by reconstruction error of action proposals from MAAL. Our method can effectively solve the 3D affordance problem and outperform the previous work.

further reflect that the proposal quality of our method is higher than others.

Ablation Study for the Multi-modal Learning: We compare different multi-modal learning as shown in Tab. 7.4. Experiments for using individual branches (only action branch, only object branch, and only interaction branch) and using the combinations of branches (action branch + object branch, action branch + interaction branch, and object branch + interaction branch) are provided.

Due to the comprehensive learning of multi-modal data, our method performs best among all the combinations. Learning with more modalities can improve the ability of the encoder. As in Tab. 7.4, the designation with only interaction outperform the designation with single modalities. Meanwhile, due to the intermediate fusion with other modalities, the interaction branch combines with another branch and outperforms the encoder only with the interaction branch. All the results prove the effectiveness

Method		Sample-Succ (%)
Action Proposal	Actionability Score	
Where2Act-P [159]	Where2Act-C [159]	27.33
	AdaAfford-C [227]	28.58
	MAAL	28.67
AdaAfford-P [227]	Where2Act-C [159]	30.90
	AdaAfford-C [227]	32.50
	MAAL	32.36
MAAL	Where2Act-C [159]	31.50
	AdaAfford-C [227]	33.44
	MAAL	34.25

Table 7.3: Comparison of different combinations of methods. The higher performances prove that MAAL possesses a higher ability to evaluate actionability scores and generate high-quality proposals.

Multi-modal Learning Method	F-score (%)	Sample-Succ (%)
only action branch	32.47	13.54
only object branch	53.42	21.75
only interaction branch	58.74	24.01
action branch + object branch	59.87	23.88
action branch + interaction branch	73.26	32.55
object branch + interaction branch	75.54	33.89
All branches	**76.63**	**34.25**

Table 7.4: Combinations of learning different modalities. MAAL jointly considers object modality and action modality and further learn the interaction from both modalities. The comprehensive multi-modal learning by MAAL achieves better performance in the comparison.

of our method design. These may also reveal the necessity of multi-modal learning in 3D affordance. With better multi-modal learning, the network can better model and understand the affordance of a given object.

Furthermore, we modify our encoder with early fusion. We remain all three branches in the encoder but do not provide features from the action and object learner to the interaction branch. This leads the encoder to degrade to an early fusion-based method but still considers multi-modal learning. Then, the performance decreases by 8.31% in F-score compared with ours. All results reveal that our encoder is effective in multi-modal learning. The idea of intermediate fusion also improves learning ability.

7.4.3 Visualization for Affordance Predictions

We showcase the affordance predictions by heatmap as Fig. 7.5. The value of each pixel is calculated by the actionability score of MAAL following [227]. The visualized results show the effectiveness of MAAL in learning 3D object affordance. The actionable point in

3D objects can be correctly predicted by MAAL. Besides, we visualize different shapes with different categories from the validation set and test set in Fig. 7.5. For the unseen categories in the test set, our method can also understand the 3D object affordance and produce high confidence for actionable points. This further reveals the generalization of our MAAL.

7.5 Conclusion

This chapter introduces a simple and data-efficient pipeline for the 3D affordance problem, named Multimodality-Aware Autoencoder-based affordance Learning (MAAL). MAAL contains three parts: MultiModal Energized Encoder(MME), action memory, and action decoder. We specifically design the encoder for multi-modal learning in 3D object affordance. The previous work usually directly applies early fusion to process multi-modal data. Comparatively, in our work, MME provides a comprehensive understanding of multi-modal learning and boosts the multi-modal learning ability for 3D affordance. In the experiment, the comparisons reveal the effectiveness of our method. MAAL outperforms former methods in different data splits, conditions, and metrics.

CHAPTER 8

CONCLUSION AND FUTURE WORK

In this thesis, we have expanded the scope of multi-modal learning research, providing valuable insights to the community, especially in understanding and applying multi-modal data. This exploration has not only enhanced the abilities of AI systems but also provide new insights for stimulating human-like perception and interaction. Through this research, we have ventured into the complexities of integrating and interpreting diverse forms of data, drawing closer to achieving AI systems with human-like sensory experiences.

Specifically, in the field of visual perception, our investigations have led to a deeper understanding of how machines can process and interpret visual data, drawing closer to the nuanced capabilities of human vision. From the complexities of fine-grained visual classification to the broader challenges of general visual representation, this research has delved into the intricate aspects of visual data processing, laying the groundwork for AI systems that see and interpret the world with enhanced accuracy and efficiency.

Moving beyond visual stimuli, the incorporation of auditory information has added another dimension to our exploration. Understanding and generating responses to audio inputs have brought us closer to simulating real human reactions and behaviors. The exploration in co-speech gesture generation, in particular, has shed light on how AI can process auditory information and respond with appropriate physical expressions, mirroring human communication dynamics. Furthermore, the integration of tactile sensory data through affordance learning for robotic grippers has marked a significant advancement in AI's interaction with the physical world. This research has not only

enabled machines to perceive and understand their environment better but also to act upon it in a manner akin to human manipulation.

Throughout this thesis, the emphasis has been on not just improving the individual capabilities of AI systems in handling different modalities but on synthesizing these modalities to create a more holistic and integrated approach to AI perception and interaction. The novel methodologies and insights derived from this research contribute significantly to the field of AI, demonstrating how multi-modal data can be leveraged to enhance the accuracy, efficiency, and human-like capabilities of AI systems.

Looking to the future, several avenues for research emerge from the findings of this thesis. The potential for further refining and expanding the MHEM strategy and ELP classifier in broader contexts of visual recognition presents a promising direction. Investigating these methodologies in more diverse and challenging datasets could yield deeper insights into their scalability and adaptability. In generative tasks, extending the application of IcoCap, SEEG, and MAAL to more intricate scenarios and multi-modal integrations offers fertile ground for research. Exploring these methods in dynamic real-world environments, augmented reality, and more complex human-computer interactions could further push the boundaries of what these technologies can achieve. Moreover, the integration of emerging technologies such as deep reinforcement learning (e.g., improving robot gripper movements in MAAL) and unsupervised learning approaches (e.g., integrating data across all modalities to construct more unified and larger models) in multi-modal learning presents a promising direction. These approaches could offer novel ways to address some of the unsolved challenges in the field, leading to more advanced and autonomous AI systems.

In conclusion, this thesis represents a substantial step forward in the quest to create intelligent, adaptable, and efficient AI systems capable of interpreting and interacting with the complex types of multi-modal data that defines our world. The insights and advancements presented are poised to inspire further research and development, reinforcing the synergy between different sensory modalities in the continual evolution of artificial intelligence. In the future, the foundations laid by this research promise to guide and inform the next generation of AI innovations, driving us closer to systems that can perceive, understand, and interact with the world in ways that are truly reminiscent of human intelligence.

BIBLIOGRAPHY

[1] N. AAFAQ, N. AKHTAR, W. LIU, S. Z. GILANI, AND A. MIAN, *Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning*, in CVPR, 2019.

[2] N. AAFAQ, A. MIAN, W. LIU, S. Z. GILANI, AND M. SHAH, *Video description: A survey of methods, datasets, and evaluation metrics*, ACM Computing Surveys (CSUR), 52 (2019), pp. 1–37.

[3] C. AHUJA, D. W. LEE, Y. I. NAKANO, AND L.-P. MORENCY, *Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach*, in European Conference on Computer Vision, Springer, 2020, pp. 248–265.

[4] C. AHUJA AND L.-P. MORENCY, *Language2pose: Natural language grounded pose forecasting*, in 2019 International Conference on 3D Vision (3DV), IEEE, 2019, pp. 719–728.

[5] G. ALAIN AND Y. BENGIO, *Understanding intermediate layers using linear classifier probes*, arXiv preprint arXiv:1610.01644, (2016).

[6] C. ALBERTI, J. LING, M. COLLINS, AND D. REITTER, *Fusion of detected objects in text for visual question answering*, arXiv preprint arXiv:1908.05054, (2019).

[7] S. ALEXANDERSON, G. E. HENTER, T. KUCHERENKO, AND J. BESKOW, *Style-controllable speech-driven gesture synthesis using normalising flows*, in Computer Graphics Forum, vol. 39, Wiley Online Library, 2020, pp. 487–496.

[8] H. ALWASSEL, D. MAHAJAN, B. KORBAR, L. TORRESANI, B. GHANEM, AND D. TRAN, *Self-supervised learning by cross-modal audio-video clustering*, Advances in Neural Information Processing Systems, 33 (2020), pp. 9758–9770.

[9] P. AN, Z. WANG, AND C. ZHANG, *Ensemble unsupervised autoencoders and gaussian mixture model for cyberattack detection*, Information Processing & Management, 59 (2022), p. 102844.

[10] C. ANDERSON, M. GWILLIAM, A. TEUSCHER, A. MERRILL, AND R. FARRELL, *Facing the hard problems in fgvc*, arXiv preprint arXiv:2006.13190, (2020).

[11] P. ANDERSON, X. HE, C. BUEHLER, D. TENEY, M. JOHNSON, S. GOULD, AND L. ZHANG, *Bottom-up and top-down attention for image captioning and visual question answering*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6077–6086.

[12] A. ARNAB, M. DEHGHANI, G. HEIGOLD, C. SUN, M. LUČIĆ, AND C. SCHMID, *Vivit: A video vision transformer*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6836–6846.

[13] Y. M. ASANO, C. RUPPRECHT, AND A. VEDALDI, *A critical analysis of self-supervision, or what we can learn from a single image*, arXiv preprint arXiv:1904.13132, (2019).

[14] K. AUSDERAU, J. SIDERIS, M. FURLONG, L. M. LITTLE, J. BULLUCK, AND G. T. BARANEK, *National survey of sensory features in children with asd: Factor structure of the sensory experience questionnaire (3.0)*, Journal of autism and developmental disorders, 44 (2014), pp. 915–925.

[15] Z. AZAR, A. BACKUS, AND A. ÖZYÜREK, *Language contact does not drive gesture transfer: Heritage speakers maintain language specific gesture patterns in each language*, Bilingualism: Language and Cognition, 23 (2020), pp. 414–428.

[16] Y. BAI, J. FU, T. ZHAO, AND T. MEI, *Deep attention neural tensor network for visual question answering*, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 20–35.

[17] M. BAIN, A. NAGRANI, G. VAROL, AND A. ZISSERMAN, *Frozen in time: A joint video and image encoder for end-to-end retrieval*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1728–1738.

[18] T. BALTRUŠAITIS, C. AHUJA, AND L.-P. MORENCY, *Multimodal machine learning: A survey and taxonomy*, IEEE transactions on pattern analysis and machine intelligence, 41 (2018), pp. 423–443.

[19] L. BARALDI, C. GRANA, AND R. CUCCHIARA, *Hierarchical boundary-aware neural encoder for video captioning*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1657–1666.

[20] K. BAYOUDH, R. KNANI, F. HAMDAOUI, AND A. MTIBAA, *A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets*, The Visual Computer, (2021), pp. 1–32.

[21] S. BÖCK AND G. WIDMER, *Maximum filter vibrato suppression for onset detection*, in Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013), vol. 7, 2013.

[22] R. BOMMASANI, K. DAVIS, AND C. CARDIE, *Bert wears gloves: Distilling static embeddings from pretrained contextual representations*, (2019).

[23] J. BORJA-DIAZ, O. MEES, G. KALWEIT, L. HERMANN, J. BOEDECKER, AND W. BURGARD, *Affordance learning from play for sample-efficient policy learning*, in 2022 International Conference on Robotics and Automation (ICRA), IEEE, 2022, pp. 6372–6378.

[24] A. BORJI AND L. ITTI, *State-of-the-art in visual attention modeling*, IEEE transactions on pattern analysis and machine intelligence, 35 (2012), pp. 185–207.

[25] S. Y. BOULAHIA, A. AMAMRA, M. R. MADI, AND S. DAIKH, *Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition*, Machine Vision and Applications, 32 (2021), p. 121.

[26] D. BOXER, *Social distance and speech behavior: The case of indirect complaints*, Journal of pragmatics, 19 (1993), pp. 103–125.

[27] H. CAI, L. ZHU, AND S. HAN, *Proxylessnas: Direct neural architecture search on target task and hardware*, arXiv preprint arXiv:1812.00332, (2018).

[28] S. CAI, W. ZUO, AND L. ZHANG, *Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization*, in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 511–520.

[29] K. CAO, C. WEI, A. GAIDON, N. ARECHIGA, AND T. MA, *Learning imbalanced datasets with label-distribution-aware margin loss*, in Advances in Neural Information Processing Systems, 2019.

[30] J. CARREIRA AND A. ZISSERMAN, *Quo vadis, action recognition? a new model and the kinetics dataset*, in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[31] A. CASANOVA, M. DROZDZAL, AND A. ROMERO-SORIANO, *Generating unseen complex scenes: are we there yet?*, arXiv preprint arXiv:2012.04027, (2020).

[32] J. CASSELL, *A framework for gesture generation and interpretation*, Computer vision in human-machine interaction, (1998), pp. 191–215.

[33] D. CHANG, Y. DING, J. XIE, A. K. BHUNIA, X. LI, Z. MA, M. WU, J. GUO, AND Y. SONG, *The devil is in the channels: Mutual-channel loss for fine-grained image classification*, IEEE Transactions on Image Processing, 29 (2020), pp. 4683–4695.

[34] N. CHAROENPHAKDEE, J. VONGKULBHISAL, N. CHAIRATANAKUL, AND M. SUGIYAMA, *On focal loss for class-posterior probability estimation: A theoretical perspective*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5202–5211.

[35] D. CHEN AND W. DOLAN, *Collecting highly parallel data for paraphrase evaluation*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, June 2011, Association for Computational Linguistics, pp. 190–200.

[36] J. CHEN, Y. PAN, Y. LI, T. YAO, H. CHAO, AND T. MEI, *Temporal deformable convolutional encoder-decoder networks for video captioning*, in Proceedings of the AAAI conference on artificial intelligence, vol. 33, 2019, pp. 8167–8174.

[37] S. CHEN, X. HE, L. GUO, X. ZHU, W. WANG, J. TANG, AND J. LIU, *Valor: Vision-audio-language omni-perception pretraining model and dataset*, ArXiv, abs/2304.08345 (2023).

[38] S. CHEN, W. JIANG, W. LIU, AND Y.-G. JIANG, *Learning modality interaction for temporal sentence localization and event captioning in videos*, in ECCV, 2020.

[39] S. CHEN AND Y.-G. JIANG, *Motion guided region message passing for video captioning*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1543–1552.

[40] S. CHEN, Q. JIN, J. CHEN, AND A. G. HAUPTMANN, *Generating video descriptions with latent topic guidance*, IEEE Transactions on Multimedia, 21 (2019), pp. 2407–2418.

[41] T. CHEN, W. WU, Y. GAO, L. DONG, X. LUO, AND L. LIN, *Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding*, in Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 2023–2031.

[42] X. CHEN, M. DING, X. WANG, Y. XIN, S. MO, Y. WANG, S. HAN, P. LUO, G. ZENG, AND J. WANG, *Context autoencoder for self-supervised representation learning*, arXiv preprint arXiv:2202.03026, (2022).

[43] X. CHEN, H. FAN, R. B. GIRSHICK, AND K. HE, *Improved baselines with momentum contrastive learning*, CoRR, abs/2003.04297 (2020).

[44] Y. CHEN, Y. BAI, W. ZHANG, AND T. MEI, *Destruction and construction learning for fine-grained image recognition*, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[45] Y. CHEN, S. WANG, W. ZHANG, AND Q. HUANG, *Less is more: Picking informative frames for video captioning*, in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 358–373.

[46] T. S. CHO, S. AVIDAN, AND W. T. FREEMAN, *A probabilistic image jigsaw puzzle solver*, in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 183–190.

[47] E. D. CUBUK, B. ZOPH, D. MANE, V. VASUDEVAN, AND Q. V. LE, *Autoaugment: Learning augmentation policies from data*, arXiv preprint arXiv:1805.09501, (2018).

[48] Y. CUI, M. JIA, T.-Y. LIN, Y. SONG, AND S. BELONGIE, *Class-balanced loss based on effective number of samples*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9268–9277.

[49] Y. CUI, F. ZHOU, J. WANG, X. LIU, Y. LIN, AND S. BELONGIE, *Kernel pooling for convolutional neural networks*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2921–2930.

[50] B. DAI, S. FIDLER, R. URTASUN, AND D. LIN, *Towards diverse and natural image descriptions via a conditional gan*, in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2970–2979.

[51] E. P. DAVIS, S. A. STOUT, J. MOLET, B. VEGETABILE, L. M. GLYNN, C. A. SANDMAN, K. HEINS, H. STERN, AND T. Z. BARAM, *Exposure to unpredictable maternal sensory signals influences cognitive development across species*, Proceedings of the National Academy of Sciences, 114 (2017), pp. 10390–10395.

[52] L. DE JONGE-HOEKSTRA, R. F. COX, S. VAN DER STEEN, AND J. A. DIXON, *Easier said than done? task difficulty's influence on temporal alignment, semantic similarity, and complexity matching between gestures and speech*, Cognitive science, 45 (2021), p. e12989.

[53] P. W. DEMPSEY, M. E. ALLISON, S. AKKARAJU, C. C. GOODNOW, AND D. T. FEARON, *C3d of complement as a molecular adjuvant: bridging innate and acquired immunity*, Science, 271 (1996), pp. 348–350.

[54] S. DENG, X. XU, C. WU, K. CHEN, AND K. JIA, *3d affordancenet: A benchmark for visual object affordance understanding*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.

[55] M. DENKOWSKI AND A. LAVIE, *Meteor universal: Language specific translation evaluation for any target language*, in Proceedings of the ninth workshop on statistical machine translation, 2014, pp. 376–380.

[56] A. DHURANDHAR, K. SHANMUGAM, R. LUSS, AND P. OLSEN, *Improving simple models with confidence profiles*, arXiv preprint arXiv:1807.07506, (2018).

[57] C. DING AND D. TAO, *Robust face recognition via multimodal deep face representation*, IEEE transactions on Multimedia, 17 (2015), pp. 2049–2058.

[58] Y. DING, Y. ZHOU, Y. ZHU, Q. YE, AND J. JIAO, *Selective sparse sampling for fine-grained image recognition*, in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6599–6608.

[59] J. DONAHUE, L. ANNE HENDRICKS, S. GUADARRAMA, M. ROHRBACH, S. VENUGOPALAN, K. SAENKO, AND T. DARRELL, *Long-term recurrent convolutional networks for visual recognition and description*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.

[60] M. DOOSTDAR, S. SCHIFFER, AND G. LAKEMEYER, *A robust speech recognition system for service-robotics applications*, in Robot Soccer World Cup, Springer, 2008, pp. 1–12.

[61] R. DU, D. CHANG, A. K. BHUNIA, J. XIE, Y.-Z. SONG, Z. MA, AND J. GUO, *Fine-grained visual classification via progressive multi-granularity training of jigsaw patches*, in European Conference on Computer Vision, 2020.

[62] A. DUBEY, O. GUPTA, P. GUO, R. RASKAR, R. FARRELL, AND N. NAIK, *Pairwise confusion for fine-grained visual classification*, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 70–86.

[63] B. EISNER, H. ZHANG, AND D. HELD, *Flowbot3d: Learning 3d articulation flow to manipulate articulated objects*, arXiv preprint arXiv:2205.04382, (2022).

[64] D. P. ELLIS, *Beat tracking by dynamic programming*, Journal of New Music Research, 36 (2007), pp. 51–60.

[65] D. P. ELLIS AND G. E. POLINER, *Identifyingcover songs' with chroma features and dynamic programming beat tracking*, in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, vol. 4, IEEE, 2007, pp. IV–1429.

[66] E. ELYAN AND M. M. GABER, *A fine-grained random forests using class decomposition: an application to medical diagnosis*, Neural computing and applications, 27 (2016), pp. 2279–2288.

[67] P. F. FELZENSZWALB, R. B. GIRSHICK, D. MCALLESTER, AND D. RAMANAN, *Object detection with discriminatively trained part-based models*, IEEE transactions on pattern analysis and machine intelligence, 32 (2009), pp. 1627–1645.

[68] K. P. F.R.S., *Liii. on lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2 (1901), pp. 559–572.

[69] J. FU, H. ZHENG, AND T. MEI, *Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4438–4446.

[70] W. GE, X. LIN, AND Y. YU, *Weakly supervised complementary parts models for fine-grained image classification from the bottom up*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3034–3043.

[71] Z. GE, D. MAHAPATRA, S. SEDAI, R. GARNAVI, AND R. CHAKRAVORTY, *Chest x-rays classification: A multi-label and fine-grained problem*, arXiv preprint arXiv:1807.07247, (2018).

[72] D. GHADIYARAM, D. TRAN, AND D. MAHAJAN, *Large-scale weakly-supervised pre-training for video action recognition*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12046–12055.

[73] A. A. GHAZANFAR AND L. R. SANTOS, *Primate brains in the wild: the sensory bases for social interactions*, Nature Reviews Neuroscience, 5 (2004), pp. 603–616.

[74] J. J. GIBSON, *The theory of affordances*, Hilldale, USA, 1 (1977), pp. 67–82.

[75] S. GINOSAR, A. BAR, G. KOHAVI, C. CHAN, A. OWENS, AND J. MALIK, *Learning individual styles of conversational gesture*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3497–3506.

[76] S. GOLDIN-MEADOW AND M. W. ALIBALI, *Gesture's role in speaking, learning, and creating language*, Annual review of psychology, 64 (2013), pp. 257–283.

[77] D. GONG, L. LIU, V. LE, B. SAHA, M. R. MANSOUR, S. VENKATESH, AND A. V. D. HENGEL, *Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1705–1714.

[78] H. GUNES AND M. PICCARDI, *Affect recognition from face and body: early fusion vs. late fusion*, in 2005 IEEE international conference on systems, man and cybernetics, vol. 4, IEEE, 2005, pp. 3437–3443.

[79] C. GUO, X. ZUO, S. WANG, S. ZOU, Q. SUN, A. DENG, M. GONG, AND L. CHENG, *Action2motion: Condite0ioned generation of 3d human motions*, in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2021–2029.

[80] M.-H. GUO, T.-X. XU, J.-J. LIU, Z.-N. LIU, P.-T. JIANG, T.-J. MU, S.-H. ZHANG, R. R. MARTIN, M.-M. CHENG, AND S.-M. HU, *Attention mechanisms in computer vision: A survey*, Computational Visual Media, 8 (2022), pp. 331–368.

[81] Y. GUO, Y. LIU, A. OERLEMANS, S. LAO, S. WU, AND M. S. LEW, *Deep learning for visual understanding: A review*, Neurocomputing, 187 (2016), pp. 27–48.

[82] Y. HAN, B. WANG, R. HONG, AND F. WU, *Movie question answering via textual memory and plot graph*, IEEE Transactions on Circuits and Systems for Video Technology, 30 (2019), pp. 875–887.

[83] A. HANJALIC AND L.-Q. XU, *Affective video content representation and modeling*, IEEE transactions on multimedia, 7 (2005), pp. 143–154.

[84] M. HASSANIN, S. KHAN, AND M. TAHTALI, *Visual affordance and function understanding: A survey*, ACM Computing Surveys (CSUR), 54 (2021), pp. 1–35.

[85] W. G. HATCHER AND W. YU, *A survey of deep learning: Platforms, applications and emerging research trends*, IEEE Access, 6 (2018), pp. 24411–24432.

[86] K. HE, X. CHEN, S. XIE, Y. LI, P. DOLLÁR, AND R. GIRSHICK, *Masked autoencoders are scalable vision learners*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[87] K. HE, H. FAN, Y. WU, S. XIE, AND R. GIRSHICK, *Momentum contrast for unsupervised visual representation learning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[88] K. HE, X. ZHANG, S. REN, AND J. SUN, *Deep residual learning for image recognition*, in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 770–778.

[89] R. HECHT-NIELSEN, *Theory of the backpropagation neural network*, in Neural networks for perception, Elsevier, 1992, pp. 65–93.

[90] G. E. HENTER, S. ALEXANDERSON, AND J. BESKOW, *Moglow: Probabilistic and controllable motion synthesis using normalising flows*, ACM Transactions on Graphics (TOG), 39 (2020), pp. 1–14.

[91] E. HERNANDEZ AND J. ANDREAS, *The low-dimensional linear geometry of contextualized word representations*, arXiv preprint arXiv:2105.07109, (2021).

[92] C. HONG, J. YU, J. WAN, D. TAO, AND M. WANG, *Multimodal deep autoencoder for human pose recovery*, IEEE transactions on image processing, 24 (2015), pp. 5659–5670.

[93] J. HOU, X. WU, W. ZHAO, J. LUO, AND Y. JIA, *Joint syntax representation learning and visual cue translation for video captioning*, in ICCV, 2019.

[94] Y.-C. HSU, C.-Y. HONG, D.-J. CHEN, M.-S. LEE, D. GEIGER, AND T.-L. LIU, *Fine-grained visual recognition with batch confusion norm*, arXiv preprint arXiv:1910.12423, (2019).

[95] J. HU, L. SHEN, AND G. SUN, *Squeeze-and-excitation networks*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[96] C. HUANG, C. C. LOY, AND X. TANG, *Local similarity-aware deep feature embedding*, arXiv preprint arXiv:1610.08904, (2016).

[97] Z. HUANG AND Y. LI, *Interpretable and accurate fine-grained recognition via region grouping*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8662–8672.

[98] S. S. HUGHES-BERHEIM, L. M. MORETT, AND R. BULGER, *Semantic relationships between representational gestures and their lexical affiliates are evaluated similarly for speech and text*, Frontiers in psychology, 11 (2020), p. 2808.

[99] F. HUTMACHER, *Why is there so much more research on vision than on any other sensory modality?*, Frontiers in psychology, 10 (2019), p. 2246.

[100] V. IASHIN AND E. RAHTU, *Multi-modal dense video captioning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 958–959.

[101] S. IOFFE AND C. SZEGEDY, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, in International conference on machine learning, PMLR, 2015, pp. 448–456.

[102] R. JI, L. WEN, L. ZHANG, D. DU, Y. WU, C. ZHAO, X. LIU, AND F. HUANG, *Attention convolutional binary neural tree for fine-grained visual categorization*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10468–10477.

[103] H. JIANG, S. LIU, J. WANG, AND X. WANG, *Hand-object contact consistency reasoning for human grasps generation*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11107–11116.

[104] L. JIANG, D. MENG, S.-I. YU, Z. LAN, S. SHAN, AND A. HAUPTMANN, *Self-paced learning with diversity*, Advances in Neural Information Processing Systems, 27 (2014), pp. 2078–2086.

[105] K. JUNG, K. I. KIM, AND A. K. JAIN, *Text information extraction in images and video: a survey*, Pattern recognition, 37 (2004), pp. 977–997.

[106] S. E. KAHOU, C. PAL, X. BOUTHILLIER, P. FROUMENTY, Ç. GÜLÇEHRE, R. MEMISEVIC, P. VINCENT, A. COURVILLE, Y. BENGIO, R. C. FERRARI, ET AL., *Combining modality specific deep neural networks for emotion recognition in video*, in Proceedings of the 15th ACM on International conference on multimodal interaction, 2013, pp. 543–550.

[107] B. KANG, S. XIE, M. ROHRBACH, Z. YAN, A. GORDO, J. FENG, AND Y. KALANTIDIS, *Decoupling representation and classifier for long-tailed recognition*, in International Conference on Learning Representations, 2020.

[108] A. KARPATHY, G. TODERICI, S. SHETTY, T. LEUNG, R. SUKTHANKAR, AND L. FEI-FEI, *Large-scale video classification with convolutional neural networks*, in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[109] W. KAY, J. CARREIRA, K. SIMONYAN, B. ZHANG, C. HILLIER, S. VIJAYANARASIMHAN, F. VIOLA, T. GREEN, T. BACK, P. NATSEV, M. SULEYMAN, AND A. ZISSERMAN, *The kinetics human action video dataset*, CoRR, abs/1705.06950 (2017).

[110] J. D. M.-W. C. KENTON AND L. K. TOUTANOVA, *Bert: Pre-training of deep bidirectional transformers for language understanding*, in Proceedings of naacL-HLT, vol. 1, 2019, p. 2.

[111] B. KHALEGHI, A. KHAMIS, F. O. KARRAY, AND S. N. RAZAVI, *Multisensor data fusion: A review of the state-of-the-art*, Information fusion, 14 (2013), pp. 28–44.

[112] M. KOKIC, J. A. STORK, J. A. HAUSTEIN, AND D. KRAGIC, *Affordance detection for task-specific grasping using deep learning*, in 2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids), IEEE, 2017, pp. 91–98.

[113] I. KRASIN, T. DUERIG, N. ALLDRIN, A. VEIT, S. ABU-EL-HAIJA, S. BELONGIE, D. CAI, Z. FENG, V. FERRARI, V. GOMES, A. GUPTA, D. NARAYANAN, C. SUN, G. CHECHIK, AND K. MURPHY, *Openimages: A public dataset for large-scale multi-label and multi-class image classification.*, Dataset available from https://github.com/openimages, (2016).

[114] J. KRAUSE, H. JIN, J. YANG, AND L. FEI-FEI, *Fine-grained recognition without part annotations*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5546–5555.

[115] J. KRAUSE, M. STARK, J. DENG, AND L. FEI-FEI, *3d object representations for fine-grained categorization*, in 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia, 2013.

[116] W. A. KRETZSCHMAR JR, W. A. KRETZSCHMAR, AND W. A. KRETZSCHMAR JR, *The linguistics of speech*, Cambridge University Press, 2009.

[117] A. KRISHNA, *Sensory marketing: Research on the sensuality of products*, Routledge, 2011.

[118] ——, *An integrative review of sensory marketing: Engaging the senses to affect perception, judgment and behavior*, Journal of consumer psychology, 22 (2012), pp. 332–351.

[119] A. KRIZHEVSKY, G. HINTON, ET AL., *Learning multiple layers of features from tiny images*, (2009).

[120] A. KRIZHEVSKY, I. SUTSKEVER, AND G. E. HINTON, *Imagenet classification with deep convolutional neural networks*, Communications of the ACM, 60 (2017), pp. 84–90.

[121] T. KUCHERENKO, P. JONELL, S. VAN WAVEREN, G. E. HENTER, S. ALEXANDERSSON, I. LEITE, AND H. KJELLSTRÖM, *Gesticulator: A framework for semantically-aware speech-driven gesture generation*, in Proceedings of the 2020 International Conference on Multimodal Interaction, 2020, pp. 242–250.

[122] D. LAHAT, T. ADALI, AND C. JUTTEN, *Multimodal data fusion: an overview of methods, challenges, and prospects*, Proceedings of the IEEE, 103 (2015), pp. 1449–1477.

[123] B. LI, Y. LIU, AND X. WANG, *Gradient harmonized single-stage detector*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8577–8584.

[124] H. LI, D. SONG, L. LIAO, AND C. PENG, *Revnet: Bring reviewing into video captioning for a better description*, in 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2019, pp. 1312–1317.

[125] X. LI, W. WANG, L. WU, S. CHEN, X. HU, J. LI, J. TANG, AND J. YANG, *Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection*, arXiv preprint arXiv:2006.04388, (2020).

[126] X. LI, B. ZHAO, X. LU, ET AL., *Mam-rnn: multi-level attention model based rnn for video captioning.*, in IJCAI, vol. 2017, 2017, pp. 2208–2214.

[127] Y. LI, Y. PAN, J. CHEN, T. YAO, AND T. MEI, *X-modaler: A versatile and high-performance codebase for cross-modal analytics*, in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 3799–3802.

[128] Y. LI, R. QUAN, L. ZHU, AND Y. YANG, *Efficient multimodal fusion via interactive prompting*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2604–2613.

[129] Y. LIANG, Y. BAI, W. ZHANG, X. QIAN, L. ZHU, AND T. MEI, *Vrr-vg: Refocusing visually-relevant relationships*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10403–10412.

[130] Y. LIANG, Q. FENG, L. ZHU, L. HU, P. PAN, AND Y. YANG, *Seeg: Semantic energized co-speech gesture generation*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10473–10482.

[131] Y. LIANG, X. QIAN, AND L. ZHU, *Towards better railway service: Passengers counting in railway compartment*, IEEE Transactions on Circuits and Systems for Video Technology, 31 (2020), pp. 439–451.

[132] Y. LIANG, X. WANG, L. ZHU, AND Y. YANG, *Maal: Multimodality-aware autoencoder-based affordance learning for 3d articulated objects*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 217–227.

[133] Y. LIANG AND W. ZHANG, *Winner solution for aliproducts challenge: Large-scale product recognition*.

[134] Y. LIANG, L. ZHU, X. WANG, AND Y. YANG, *Penalizing the hard example but not too much: A strong baseline for fine-grained visual classification*, IEEE Transactions on Neural Networks and Learning Systems, (2022).

[135] ——, *A simple episodic linear probe improves visual recognition in the wild*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[136] ——, *Icocap: Improving video captioning by compounding images*, IEEE Transactions on Multimedia, (2023).

[137] M. LIAO, S. ZHANG, P. WANG, H. ZHU, X. ZUO, AND R. YANG, *Speech2video synthesis with 3d skeleton regularization and expressive body poses*, in Proceedings of the Asian Conference on Computer Vision, 2020.

[138] R. LICKLITER, *The integrated development of sensory organization*, Clinics in perinatology, 38 (2011), pp. 591–603.

[139] C.-Y. LIN, *Rouge: A package for automatic evaluation of summaries*, in Text summarization branches out, 2004, pp. 74–81.

[140] D. LIN, Y. WANG, L. LIANG, P. LI, AND C. P. CHEN, *Deep lsac for fine-grained recognition*, IEEE Transactions on Neural Networks and Learning Systems, (2020).

[141] K. LIN, L. LI, C.-C. LIN, F. AHMED, Z. GAN, Z. LIU, Y. LU, AND L. WANG, *Swinbert: End-to-end transformers with sparse attention for video captioning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17949–17958.

[142] T.-Y. LIN, P. GOYAL, R. GIRSHICK, K. HE, AND P. DOLLÁR, *Focal loss for dense object detection*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[143] T.-Y. LIN, M. MAIRE, S. BELONGIE, J. HAYS, P. PERONA, D. RAMANAN, P. DOLLÁR, AND C. L. ZITNICK, *Microsoft coco: Common objects in context*, in European conference on computer vision, Springer, 2014, pp. 740–755.

[144] T.-Y. LIN, A. ROYCHOWDHURY, AND S. MAJI, *Bilinear cnn models for fine-grained visual recognition*, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1449–1457.

[145] W. LIN AND C.-C. J. KUO, *Perceptual visual quality metrics: A survey*, Journal of visual communication and image representation, 22 (2011), pp. 297–312.

[146] S. LIU, Z. REN, AND J. YUAN, *Sibnet: Sibling convolutional encoder for video captioning*, IEEE TPAMI, (2020).

[147] Z. LIU, Z. MIAO, X. ZHAN, J. WANG, B. GONG, AND S. X. YU, *Large-scale long-tailed recognition in an open world*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[148] J. LU, V. GOSWAMI, M. ROHRBACH, D. PARIKH, AND S. LEE, *12-in-1: Multi-task vision and language representation learning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10437–10446.

[149] H. LUO, L. JI, M. ZHONG, Y. CHEN, W. LEI, N. DUAN, AND T. LI, *Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning*, Neurocomputing, 508 (2022), pp. 293–304.

[150] R. MA, Y. LIANG, AND Y. MA, *A self-adapting method for rbc count from different blood smears based on pcnn and image quality*, in 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2016, pp. 1611–1615.

[151] S. MAJI, E. RAHTU, J. KANNALA, M. BLASCHKO, AND A. VEDALDI, *Fine-grained visual classification of aircraft*, arXiv preprint arXiv:1306.5151, (2013).

[152] O. MANGASARIAN AND R. MEYER, *Absolute value equations*, Linear Algebra and Its Applications, 419 (2006), pp. 359–367.

[153] H. P. MARTÍNEZ AND G. N. YANNAKAKIS, *Deep multimodal fusion: Combining discrete events and continuous signals*, in Proceedings of the 16th International conference on multimodal interaction, 2014, pp. 34–41.

[154] J. D. MATARAZZO, A. N. WIENS, R. H. JACKSON, AND T. S. MANAUGH, *Interviewee speech behavior under conditions of endogenously-present and exogenously-induced motivational states.*, Journal of Clinical Psychology, (1970).

[155] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, *librosa: Audio and music signal analysis in python*, in Proceedings of the 14th python in science conference, vol. 8, Citeseer, 2015, pp. 18–25.

[156] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, *Image segmentation using deep learning: A survey*, IEEE transactions on pattern analysis and machine intelligence, 44 (2021), pp. 3523–3542.

[157] S. Mittal et al., *A survey of accelerator architectures for 3d convolution neural networks*, Journal of Systems Architecture, 115 (2021), p. 102041.

[158] V. Mnih, N. Heess, A. Graves, et al., *Recurrent models of visual attention*, in Advances in neural information processing systems, 2014, pp. 2204–2212.

[159] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, *Where2act: From pixels to actions for articulated 3d objects*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6813–6823.

[160] K. Mo, Y. Qin, F. Xiang, H. Su, and L. Guibas, *O2o-afford: Annotation-free large-scale object-object affordance learning*, in Conference on Robot Learning, PMLR, 2022, pp. 1666–1677.

[161] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, *Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 909–918.

[162] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. Torr, and P. K. Dokania, *Calibrating deep neural networks using focal loss*, arXiv preprint arXiv:2002.09437, (2020).

[163] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, *Moddrop: adaptive multi-modal gesture recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 38 (2015), pp. 1692–1706.

[164] E. Ng, S. Ginosar, T. Darrell, and H. Joo, *Body2hands: Learning to infer 3d hands from conversational gesture body dynamics*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11865–11874.

[165] S. ONDÁŠ, J. JUHÁR, M. PLEVA, M. LOJKA, E. KIKTOVÁ, M. SULÍR, A. ČIŽMÁR, AND R. HOLCER, *Speech technologies for advanced applications in service robotics*, Acta Polytechnica Hungarica, 10 (2013), pp. 45–61.

[166] A. OWENS, J. WU, J. H. MCDERMOTT, W. T. FREEMAN, AND A. TORRALBA, *Ambient sound provides supervision for visual learning*, in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 801–816.

[167] B. PAN, H. CAI, D.-A. HUANG, K.-H. LEE, A. GAIDON, E. ADELI, AND J. C. NIEBLES, *Spatio-temporal graph for video captioning with knowledge distillation*, in CVPR, 2020.

[168] Y. PAN, T. MEI, T. YAO, H. LI, AND Y. RUI, *Jointly modeling embedding and translation to bridge video and language*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4594–4602.

[169] K. PAPINENI, S. ROUKOS, T. WARD, AND W.-J. ZHU, *Bleu: a method for automatic evaluation of machine translation*, in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[170] H. PARK, J. NOH, AND B. HAM, *Learning memory-guided normality for anomaly detection*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 14372–14381.

[171] M. PATRICK, P.-Y. HUANG, Y. ASANO, F. METZE, A. G. HAUPTMANN, J. F. HENRIQUES, AND A. VEDALDI, *Support-set bottlenecks for video-text representation learning*, in International Conference on Learning Representations, 2021.

[172] Y. PENG, X. HE, AND J. ZHAO, *Object-part attention model for fine-grained image classification*, IEEE Transactions on Image Processing, 27 (2017), pp. 1487–1500.

[173] J. PENNINGTON, R. SOCHER, AND C. D. MANNING, *Glove: Global vectors for word representation*, in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[174] D. POEPPEL, W. J. IDSARDI, AND V. VAN WASSENHOVE, *Speech perception at the interface of neurobiology and linguistics*, Philosophical Transactions of the Royal Society B: Biological Sciences, 363 (2008), pp. 1071–1086.

[175] T. PORCELLO, L. MEINTJES, A. M. OCHOA, AND D. W. SAMUELS, *The reorganization of the sensory world*, Annual review of anthropology, 39 (2010), pp. 51–66.

[176] S. PORIA, E. CAMBRIA, AND A. GELBUKH, *Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis*, in Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 2539–2544.

[177] R. R. PROVINE, *Laughter punctuates speech: Linguistic, social and gender contexts of laughter*, Ethology, 95 (1993), pp. 291–298.

[178] H. PURWINS, B. LI, T. VIRTANEN, J. SCHLÜTER, S.-Y. CHANG, AND T. SAINATH, *Deep learning for audio signal processing*, IEEE Journal of Selected Topics in Signal Processing, 13 (2019), pp. 206–219.

[179] Z. PYLYSHYN, *Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception*, Behavioral and brain sciences, 22 (1999), pp. 341–365.

[180] C. R. QI, L. YI, H. SU, AND L. J. GUIBAS, *Pointnet++: Deep hierarchical feature learning on point sets in a metric space*, Advances in neural information processing systems, 30 (2017).

[181] R. QIAN, T. MENG, B. GONG, M.-H. YANG, H. WANG, S. BELONGIE, AND Y. CUI, *Spatiotemporal contrastive video representation learning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6964–6974.

[182] S. QIAN, Z. TU, Y. ZHI, W. LIU, AND S. GAO, *Speech drives templates: Co-speech gesture synthesis with learned templates*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11077–11086.

[183] A. RADFORD, J. W. KIM, C. HALLACY, A. RAMESH, G. GOH, S. AGARWAL, G. SASTRY, A. ASKELL, P. MISHKIN, J. CLARK, ET AL., *Learning transferable visual models from natural language supervision*, in International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[184] D. RAMACHANDRAM AND G. W. TAYLOR, *Deep multimodal learning: A survey on recent advances and trends*, IEEE signal processing magazine, 34 (2017), pp. 96–108.

[185] J. REDMON AND A. ANGELOVA, *Real-time grasp detection using convolutional neural networks*, in 2015 IEEE international conference on robotics and automation (ICRA), IEEE, 2015, pp. 1316–1322.

[186] J. RISCHEL, *Formal linguistics and real speech*, Speech Communication, 11 (1992), pp. 379–392.

[187] L. RUFF, J. R. KAUFFMANN, R. A. VANDERMEULEN, G. MONTAVON, W. SAMEK, M. KLOFT, T. G. DIETTERICH, AND K.-R. MÜLLER, *A unifying review of deep and shallow anomaly detection*, Proceedings of the IEEE, 109 (2021), pp. 756–795.

[188] D. E. RUMELHART, R. DURBIN, R. GOLDEN, AND Y. CHAUVIN, *Backpropagation: The basic theory*, Backpropagation: Theory, architectures and applications, (1995), pp. 1–34.

[189] D. E. RUMELHART AND D. ZIPSER, *Feature discovery by competitive learning*, Cognitive science, 9 (1985), pp. 75–112.

[190] M. SABOKROU, M. KHALOOEI, AND E. ADELI, *Self-supervised representation learning via neighborhood-relational encoding*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8010–8019.

[191] N. SEBE, I. COHEN, A. GARG, AND T. S. HUANG, *Machine learning in computer vision*, vol. 29, Springer Science & Business Media, 2005.

[192] V. SEHWAG, M. CHIANG, AND P. MITTAL, *On separability of self-supervised representations*, ICML workshop on Uncertainty and Robustness in Deep Learning (UDL), (2020).

[193] R. R. SELVARAJU, M. COGSWELL, A. DAS, R. VEDANTAM, D. PARIKH, AND D. BATRA, *Grad-cam: Visual explanations from deep networks via gradient-based localization*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[194] P. H. SEO, A. NAGRANI, A. ARNAB, AND C. SCHMID, *End-to-end generative pretraining for multimodal video captioning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17959–17968.

[195] Z. SHEN, J. LI, Z. SU, M. LI, Y. CHEN, Y.-G. JIANG, AND X. XUE, *Weakly supervised dense video captioning*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1916–1924.

[196] W. SHI, Y. GONG, X. TAO, D. CHENG, AND N. ZHENG, *Fine-grained image classification using modified dcnns trained by cascaded softmax and generalized large-margin losses*, IEEE transactions on neural networks and learning systems, 30 (2018), pp. 683–694.

[197] X. SHI, J. CAI, S. JOTY, AND J. GU, *Watch it twice: Video captioning with a refocused video encoder*, in Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 818–826.

[198] C. SHORTEN AND T. M. KHOSHGOFTAAR, *A survey on image data augmentation for deep learning*, Journal of Big Data, 6 (2019), pp. 1–48.

[199] J. SHU, Q. XIE, L. YI, Q. ZHAO, S. ZHOU, Z. XU, AND D. MENG, *Meta-weight-net: Learning an explicit mapping for sample weighting*, arXiv preprint arXiv:1902.07379, (2019).

[200] K. SIMONYAN AND A. ZISSERMAN, *Two-stream convolutional networks for action recognition in videos*, Advances in neural information processing systems, 27 (2014).

[201] S. SINHA, H. OHASHI, AND K. NAKAMURA, *Class-wise difficulty-balanced loss for solving class-imbalance*, in Proceedings of the Asian Conference on Computer Vision, 2020.

[202] C. G. SNOEK, M. WORRING, AND A. W. SMEULDERS, *Early versus late fusion in semantic video analysis*, in Proceedings of the 13th annual ACM international conference on Multimedia, 2005, pp. 399–402.

[203] K. SON, J. HAYS, AND D. B. COOPER, *Solving square jigsaw puzzles with loop constraints*, in European Conference on Computer Vision, Springer, 2014, pp. 32–46.

[204] C. STEPHENSON, S. PADHY, A. GANESH, Y. HUI, H. TANG, AND S. CHUNG, *On the geometry of generalization and memorization in deep neural networks*, arXiv preprint arXiv:2105.14602, (2021).

[205] C. SUN, A. MYERS, C. VONDRICK, K. MURPHY, AND C. SCHMID, *Videobert: A joint model for video and language representation learning*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7464–7473.

[206] M. SUN, Y. YUAN, F. ZHOU, AND E. DING, *Multi-attention multi-class constraint for fine-grained image recognition*, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 805–821.

[207] C. SZEGEDY, S. IOFFE, AND V. VANHOUCKE, *Inception-v4, inception-resnet and the impact of residual connections on learning*, CoRR, abs/1602.07261 (2016).

[208] C. SZEGEDY, V. VANHOUCKE, S. IOFFE, J. SHLENS, AND Z. WOJNA, *Rethinking the inception architecture for computer vision*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[209] M. TANG, Z. WANG, Z. LIU, F. RAO, D. LI, AND X. LI, *Clip4caption: Clip for video caption*, in Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4858–4862.

[210] M. A. TANNER AND W. H. WONG, *The calculation of posterior distributions by data augmentation*, Journal of the American statistical Association, 82 (1987), pp. 528–540.

[211] L. VAN DER MAATEN AND G. HINTON, *Visualizing data using t-sne.*, Journal of machine learning research, 9 (2008).

[212] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, Ł. KAISER, AND I. POLOSUKHIN, *Attention is all you need*, in Advances in neural information processing systems, 2017, pp. 5998–6008.

[213] R. VEDANTAM, C. LAWRENCE ZITNICK, AND D. PARIKH, *Cider: Consensus-based image description evaluation*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[214] A. VEIT, N. ALLDRIN, G. CHECHIK, I. KRASIN, A. GUPTA, AND S. BELONGIE, *Learning from noisy large-scale datasets with minimal supervision*, in Proceed-

ings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 839–847.

[215] P. VINCENT, H. LAROCHELLE, Y. BENGIO, AND P.-A. MANZAGOL, *Extracting and composing robust features with denoising autoencoders*, in Proceedings of the 25th international conference on Machine learning, 2008, pp. 1096–1103.

[216] A. VOULODIMOS, N. DOULAMIS, A. DOULAMIS, E. PROTOPAPADAKIS, ET AL., *Deep learning for computer vision: A brief review*, Computational intelligence and neuroscience, 2018 (2018).

[217] C. WAH, S. BRANSON, P. WELINDER, P. PERONA, AND S. BELONGIE, *The caltech-ucsd birds-200-2011 dataset*, Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.

[218] B. WANG, L. MA, W. ZHANG, W. JIANG, J. WANG, AND W. LIU, *Controllable video captioning with pos sequence guidance based on gated fusion network*, in ICCV, 2019.

[219] H. WANG, Y. XU, AND Y. HAN, *Spotting and aggregating salient regions for video captioning*, in Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1519–1526.

[220] J. WANG, D. CHEN, Z. WU, C. LUO, L. ZHOU, Y. ZHAO, Y. XIE, C. LIU, Y.-G. JIANG, AND L. YUAN, *Omnivl: One foundation model for image-language and video-language tasks*, in NeurIPS, 2022.

[221] R. WANG, D. CHEN, Z. WU, Y. CHEN, X. DAI, M. LIU, Y.-G. JIANG, L. ZHOU, AND L. YUAN, *Bevt: Bert pretraining of video transformers*, arXiv preprint arXiv:2112.01529, (2021).

[222] X. WANG, J. WU, J. CHEN, L. LI, Y.-F. WANG, AND W. Y. WANG, *Vatex: A large-scale, high-quality multilingual dataset for video-and-language research*, in The IEEE International Conference on Computer Vision (ICCV), October 2019.

[223] X. WANG, R. ZHANG, C. SHEN, T. KONG, AND L. LI, *Dense contrastive learning for self-supervised visual pre-training*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3024–3033.

[224] X. WANG, L. ZHU, Y. WU, AND Y. YANG, *Symbiotic attention for egocentric action recognition with object-centric alignment*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2020), pp. 1–1.

[225] X. WANG, L. ZHU, Z. ZHENG, M. XU, AND Y. YANG, *Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision*, IEEE Transactions on Multimedia, (2022), pp. 1–11.

[226] Y. WANG, V. I. MORARIU, AND L. S. DAVIS, *Learning a discriminative filter bank within a cnn for fine-grained recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4148–4157.

[227] Y. WANG, R. WU, K. MO, J. KE, Q. FAN, L. J. GUIBAS, AND H. DONG, *Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions*, in Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX, Springer Nature Switzerland Cham, 2022, pp. 90–107.

[228] P. WOLFERT, N. ROBINSON, AND T. BELPAEME, *A review of evaluation practices of gesture generation in embodied conversational agents*, arXiv preprint arXiv:2101.03769, (2021).

[229] N. WOLFSON, *The bulge: A theory of speech behavior and social distance.*, Penn Working Papers in Educational Linguistics, 2 (1990), pp. 55–83.

[230] A. WU, Y. HAN, L. ZHU, AND Y. YANG, *Instance-invariant domain adaptive object detection via progressive disentanglement*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44 (2021), pp. 4178–4193.

[231] D. WU, L. PIGOU, P.-J. KINDERMANS, N. D.-H. LE, L. SHAO, J. DAMBRE, AND J.-M. ODOBEZ, *Deep dynamic neural networks for multimodal gesture segmentation and recognition*, IEEE transactions on pattern analysis and machine intelligence, 38 (2016), pp. 1583–1597.

[232] X. WU AND H. YU, *Mars-fl: Enabling competitors to collaborate in federated learning*, IEEE Transactions on Big Data, (2022), pp. 1–11.

[233] Z. WU, Y. XIONG, S. X. YU, AND D. LIN, *Unsupervised feature learning via non-parametric instance discrimination*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3733–3742.

[234] F. XIANG, Y. QIN, K. MO, Y. XIA, H. ZHU, F. LIU, M. LIU, H. JIANG, Y. YUAN, H. WANG, ET AL., *Sapien: A simulated part-based interactive environment*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11097–11107.

[235] Q. XIAO, H. LUO, AND C. ZHANG, *Margin sample mining loss: A deep learning based method for person re-identification*, arXiv preprint arXiv:1710.00478, (2017).

[236] T. XIAO, Y. XU, K. YANG, J. ZHANG, Y. PENG, AND Z. ZHANG, *The application of two-level attention models in deep convolutional neural network for fine-grained image classification*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 842–850.

[237] Q. XIE, M.-T. LUONG, E. HOVY, AND Q. V. LE, *Self-training with noisy student improves imagenet classification*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10687–10698.

[238] S. XIE, C. SUN, J. HUANG, Z. TU, AND K. MURPHY, *Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification*, in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 305–321.

[239] J. XU, T. MEI, T. YAO, AND Y. RUI, *Msr-vtt: A large video description dataset for bridging video and language*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5288–5296.

[240] C. YAN, Y. TU, X. WANG, Y. ZHANG, X. HAO, Y. ZHANG, AND Q. DAI, *Stat: Spatial-temporal attention mechanism for video captioning*, IEEE transactions on multimedia, 22 (2019), pp. 229–241.

[241] B. YANG, T. ZHANG, AND Y. ZOU, *Clip meets video captioning: Concept-aware representation learning does matter*, in Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Springer, 2022, pp. 368–381.

[242] Y. YANG, Z. MA, A. G. HAUPTMANN, AND N. SEBE, *Feature selection for multimedia analysis by sharing information among multiple tasks*, IEEE Transactions on Multimedia, 15 (2012), pp. 661–669.

[243] Y. YANG, J. SONG, Z. HUANG, Z. MA, N. SEBE, AND A. G. HAUPTMANN, *Multi-feature fusion via hierarchical regression for multimedia analysis*, IEEE Transactions on Multimedia, 15 (2012), pp. 572–581.

[244] Y. YANG, Y. ZHUANG, AND Y. PAN, *Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies*, Frontiers of Information Technology & Electronic Engineering, 22 (2021), pp. 1551–1558.

[245] Z. YANG, T. LUO, D. WANG, Z. HU, J. GAO, AND L. WANG, *Learning to navigate for fine-grained classification*, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 420–435.

[246] B. YAO, G. BRADSKI, AND L. FEI-FEI, *A codebook-free and annotation-free approach for fine-grained image categorization*, in 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3466–3473.

[247] H. YE, G. LI, Y. QI, S. WANG, Q. HUANG, AND M.-H. YANG, *Hierarchical modular network for video captioning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17939–17948.

[248] D. YI, Z. LEI, AND S. Z. LI, *Shared representation learning for heterogenous face recognition*, in 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), vol. 1, IEEE, 2015, pp. 1–7.

[249] Y. YOON, B. CHA, J.-H. LEE, M. JANG, J. LEE, J. KIM, AND G. LEE, *Speech gesture generation from the trimodal context of text, audio, and speaker identity*, ACM Transactions on Graphics, 39 (2020).

[250] Y. YOON, W.-R. KO, M. JANG, J. LEE, J. KIM, AND G. LEE, *Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots*, in 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 4303–4309.

[251] C. YU, X. ZHAO, Q. ZHENG, P. ZHANG, AND X. YOU, *Hierarchical bilinear pooling for fine-grained visual recognition*, in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 574–589.

[252] J. YU, X. ZHENG, AND S. WANG, *A deep autoencoder feature learning method for process pattern recognition*, Journal of Process Control, 79 (2019), pp. 1–15.

[253] X.-T. YUAN, X. LIU, AND S. YAN, *Visual classification with multitask joint sparse representation*, IEEE Transactions on Image Processing, 21 (2012), pp. 4349–4360.

[254] P. YUN, L. TAI, Y. WANG, C. LIU, AND M. LIU, *Focal loss in 3d object detection*, IEEE Robotics and Automation Letters, 4 (2019), pp. 1263–1270.

[255] S. YUN, D. HAN, S. J. OH, S. CHUN, J. CHOE, AND Y. YOO, *Cutmix: Regularization strategy to train strong classifiers with localizable features*, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6023–6032.

[256] M. D. ZEILER AND R. FERGUS, *Visualizing and understanding convolutional networks*, in European conference on computer vision, Springer, 2014, pp. 818–833.

[257] C. ZHANG, S. BENGIO, M. HARDT, B. RECHT, AND O. VINYALS, *Understanding deep learning requires rethinking generalization*, arXiv preprint arXiv:1611.03530, (2016).

[258] C. ZHANG, C. LIANG, L. LI, J. LIU, Q. HUANG, AND Q. TIAN, *Fine-grained image classification via low-rank sparse coding with general and class-specific codebooks*, IEEE transactions on neural networks and learning systems, 28 (2016), pp. 1550–1559.

[259] C. ZHANG, C. ZHANG, J. SONG, J. S. K. YI, K. ZHANG, AND I. S. KWEON, *A survey on masked autoencoder for self-supervised learning in vision and beyond*, arXiv preprint arXiv:2208.00173, (2022).

[260] D. ZHANG, J. HAN, G. CHENG, AND M.-H. YANG, *Weakly supervised object localization and detection: A survey*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2021), pp. 1–1.

[261] D. ZHANG, D. MENG, AND J. HAN, *Co-saliency detection via a self-paced multiple-instance learning framework*, IEEE transactions on pattern analysis and machine intelligence, 39 (2016), pp. 865–878.

[262] D. ZHANG, W. ZENG, J. YAO, AND J. HAN, *Weakly supervised object detection using proposal- and semantic-level relationships*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44 (2022), pp. 3349–3363.

[263] H. ZHANG, M. CISSE, Y. N. DAUPHIN, AND D. LOPEZ-PAZ, *mixup: Beyond empirical risk minimization*, arXiv preprint arXiv:1710.09412, (2017).

[264] L. ZHANG, S. HUANG, W. LIU, AND D. TAO, *Learning a mixture of granularity-specific experts for fine-grained categorization*, in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8331–8340.

[265] N. ZHANG, R. FARRELL, F. IANDOLA, AND T. DARRELL, *Deformable part descriptors for fine-grained recognition and attribute prediction*, in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 729–736.

[266] Z. ZHANG, Z. QI, C. YUAN, Y. SHAN, B. LI, Y. DENG, AND W. HU, *Open-book video captioning with retrieve-copy-generate network*, in CVPR, 2021.

[267] Z. ZHANG, Y. SHI, C. YUAN, B. LI, P. WANG, W. HU, AND Z.-J. ZHA, *Object relational graph with teacher-recommended learning for video captioning*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13278–13288.

[268] W. ZHAO, X. WU, AND J. LUO, *Multi-modal dependency tree for video captioning*, Advances in Neural Information Processing Systems, 34 (2021), pp. 6634–6645.

[269] Y. ZHAO, R. WU, Z. CHEN, Y. ZHANG, Q. FAN, K. MO, AND H. DONG, *Dualafford: Learning collaborative visual affordance for dual-gripper object manipulation*, arXiv preprint arXiv:2207.01971, (2022).

[270] H. ZHENG, J. FU, T. MEI, AND J. LUO, *Learning multi-attention convolutional neural network for fine-grained image recognition*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5209–5217.

[271] H. ZHENG, J. FU, Z.-J. ZHA, AND J. LUO, *Learning deep bilinear transformation for fine-grained image representation*, in Advances in Neural Information Processing Systems, 2019, pp. 4277–4286.

[272] H. ZHENG, J. FU, Z. J. ZHA, AND J. LUO, *Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5012–5021.

[273] Q. ZHENG, C. WANG, AND D. TAO, *Syntax-aware action targeting for video captioning*, in CVPR, 2020.

[274] Z. ZHENG, L. ZHENG, AND Y. YANG, *Unlabeled samples generated by gan improve the person re-identification baseline in vitro*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 3754–3762.

[275] Z. ZHONG, L. ZHENG, G. KANG, S. LI, AND Y. YANG, *Random erasing data augmentation*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 13001–13008.

[276] L. ZHOU, Y. ZHOU, J. J. CORSO, R. SOCHER, AND C. XIONG, *End-to-end dense video captioning with masked transformer*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8739–8748.

[277] M. ZHOU, Y. BAI, W. ZHANG, T. ZHAO, AND T. MEI, *Look-into-object: Self-supervised structure modeling for object recognition*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[278] L. ZHU, H. FAN, Y. LUO, M. XU, AND Y. YANG, *Temporal cross-layer correlation mining for action recognition*, IEEE Transactions on Multimedia, 24 (2021), pp. 668–676.

[279] L. ZHU AND Y. YANG, *Actbert: Learning global-local video-text representations*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8746–8755.

[280] P. ZHUANG, Y. WANG, AND Y. QIAO, *Learning attentive pairwise interaction for fine-grained classification*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 13130–13137.