MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Deep Learning for Action Recognition in Videos

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

Doctor of Philosophy
in
Computer Science

School of Mathematical and Computational Sciences,
Massey University, Albany, Auckland,
New Zealand

Yujun Ma

February 2024

Do not go gentle into that good night,
Old age should burn and rave at close of day;
Rage, rage against the dying of the light.

–Dylan Thomas

Abstract

Action recognition aims to identify human actions in videos through complete action execution. Current action recognition approaches are primarily based on convolutional neural networks (CNNs), Transformers, or hybrids of both. Despite their strong performance, several challenges persist: insufficient disentangled modeling of spatio-temporal features, (ii) a lack of fine-grained motion modelling in action representation, and (iii) limited exploration of the positional embedding of spatial tokens. In this thesis, we introduce three novel deep-learning approaches that address these challenges and enhance spatial and temporal representation in diverse action recognition tasks, including RGB-D, coarse-grained, and fine-grained action recognition.

Firstly, we develop a multi-stage factorized spatio-temporal model (MFST) for RGB-D action recognition. This model addresses the limitations of existing RGB-D approaches that rely on entangled spatio-temporal 3D convolution. The MFST employs a multi-stage hierarchical structure where each stage independently constructs spatio-temporal dimensions. This progression from low-level features to higher-order semantic primitives results in a robust spatio-temporal representation.

Secondly, we introduce a relative-position embedding based spatially and temporally decoupled Transformer (RPE-STDT) for coarse-grained and fine-grained action recognition. RPE-STDT addresses the high computational costs of Vision Transformers in video data processing, particularly due to the absolute-position embedding in frame patch tokenization. RPE-STDT utilizes dual Transformer encoder series: spatial encoders for intra-temporal index token interactions, and temporal encoders for inter-temporal dimension interactions with a subsampling strategy.

Thirdly, we propose a convolutional transformer network (CTN) for fine-grained action recognition. Traditional Transformer models require extensive training data and additional supervision to rival CNNs in learning capabilities. The proposed CTN merges CNN's strengths (*e.g.*, weight sharing, and locality) with Transformer benefits (*e.g.*, dynamic attention, and long-range dependency learning), allows for superior fine-grained motion representation.

In summary, we contribute three deep-learning models for diverse action recognition tasks. Each model achieves the state-of-the-art performance across multiple prestigious datasets, as validated by thorough experimentation.

Acknowledgements

I wish to express my profound gratitude to several individuals whose support, encouragement, and expertise have been invaluable throughout my PhD journey.

Foremost, I extend my deepest appreciation to my principal supervisor, Professor Ruili Wang. His unwavering support, guidance, and encouragement have been pivotal in my research journey. Professor Wang's expertise, insightful feedback, and constructive criticism have significantly shaped my ideas and enhanced my work. His willingness to allow me considerable freedom in exploring diverse research problems, coupled with providing an enriching collaborative environment, has been immensely beneficial.

Special thanks are owed to my co-author, Dr. Benjia Zhou, one of the kindest individuals I have had the pleasure of knowing, for his expert feedback, patience, and invaluable advice during our collaborations. His consistent readiness to offer assistance and insightful suggestions whenever needed has been deeply valued and is greatly appreciated.

I am also immensely grateful to the staff and faculty members of the School of Mathematical and Computational Sciences at Massey University. In particular, Ms. Annette Warbrooke and Ms. Sue Di Leo have provided unwavering support and assistance throughout my PhD journey. The opportunity to pursue my doctoral studies at Massey University has been an honor.

Additionally, I extend my heartfelt gratitude to my friends and colleagues from Mathematical Science Building Room 2.15. Their support, motivation, and camaraderie have been invaluable pillars of strength and encouragement, especially during the most challenging phases of my journey.

Finally, to my parents, your emotional and financial support has been the cornerstone of my journey. Thank you for instilling in me the values of respect, confidence, and proper etiquette. Your understanding and recognition of my efforts have been deeply encouraging.

Publications

The following research papers have been published in or submitted to International Journals and Conferences during my PhD study:

1. **Yujun Ma**, Benjia Zhou, Ruili Wang, and Pichao Wang, Multi-stage Factorized Spatio-Temporal Representation for RGB-D Action and Gesture Recognition, *In Proceedings of the 31st ACM international conference on Multimedia*, pp. 3149-3160, 2023, URL: https://doi.org/10.1145/3581783.3612301, (**CORE rank A***).

2. **Yujun Ma**, and Ruili Wang, Relative-position embedding based spatially and temporally decoupled Transformer for action recognition, *Pattern Recognition* 145, 109905, 2024, URL: https://doi.org/10.1016/j.patcog.2023.109905, (**CORE rank A***).

3. **Yujun Ma**, Ruili Wang, Ming Zong, Wanting Ji, et al. Convolutional transformer network for fine-grained action recognition, *Neurocomputing* 569, 127027, 2023, URL: https://doi.org/10.1016/j.patcog.2023.109905, (CORE rank B).

4. Weirong Sun, **Yujun Ma** (corresponding author), and Ruili Wang, k-NN Attention-based Video Vision Transformer for Action Recognition. *Neurocomputing*, 127256, 2024, URL: https://doi.org/10.1016/j.neucom.2024.127256, (CORE rank B).

5. Tianyu Liu, **Yujun Ma**, Wenhan Yang, Wanting Ji, Ruili Wang, and Ping Jiang, Spatial-temporal interaction learning based two-stream network for action recognition, *Information Sciences* 606 (2022): 864-876, URL: https://doi.org/10.1016/j.ins.2022.05.092, (**CORE rank A**).

6. Aihua Zhou, **Yujun Ma** (corresponding author), Wanting Ji, Ming Zong, Pei Yang, Min Wu, and Mingzhe Liu, Multi-head attention-based two-stream EfficientNet for action recognition, *Multimedia Systems* 29, 2023, 487-498, URL: https://doi.org/10.1007/s00530-022-00961-3.

7. Ming Zong, Ruili Wang, **Yujun Ma** (corresponding author), and Wanting Ji, Spatial and temporal saliency based four-stream network with multi-task learning for action recognition, *Applied Soft Computing*, 132, 2023: 109884, URL: https://doi.org/10.1016/j.asoc.2022.109884.

8. Lei Wang, Xiaoguang Yuan, Ming Zong, **Yujun Ma**, Wanting Ji, Mingzhe Liu, and Ruili Wang, Multi-cue based four-stream 3D ResNets for video-based action recognition, *Information Sciences* 575 (2021): 654-665, URL: https://doi.org/10.1016/j.ins.2021.07.079, (**CORE rank A**).

9. Jiang, Tao, Ming Zong, **Yujun Ma**, Feng Hou, and Ruili Wang, MobileACNet: ACNet-Based Lightweight Model for Image Classification, *In International Conference on Image and Vision Computing New Zealand*, pp. 361-372, 2022, URL: https://doi.org/10.1007/978-3-031-25825-1_26.

10. Yan Tian, Zhaocheng Xu, **Yujun Ma**, Weiping Ding, Ruili Wang, Zhihong Gao, *et al.*, Survey on Deep Learning in Multimodal Medical Imaging for Cancer Detection. *Neural Computing and Applications*, (2023): 1-16, URL: https://doi.org/10.1007/s00521-023-09214-4.

11. Weirong Sun, **Yujun Ma** (corresponding author), Hong Zhang and Ruili Wang, ConTrans-Detect: A Multi-Scale Convolution-Transformer Network for DeepFake Video Detection. *In 29th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, pp. 1-6. IEEE, 2023, URL: https://doi.org/10.1109/M2VIP58386.2023.10413387.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides an overview of this thesis. We introduce the background and previous studies on action recognition in Section 1.1. Then the motivation of this research is presented in Section 1.2, where we discuss issues presented by existing approaches. The research objectives are presented in 1.2. Finally, the organization of this thesis is presented in Section 1.4.

1.1 Overview of action recognition

In the realm of computer vision, action recognition entails the classification of human actions depicted in a video into a predefined set of categories [1]. Over the past decades, this field has captured significant attention for its ability to provide tailored support across various applications such as hospital patient care, human-computer interaction, and abnormal event detection [2] [3] [4].

Action recognition encompasses multiple subtasks based on modalities and action types, such as RGB-D, coarse-grained, and fine-grained action recognition. RGB-D action recognition involves two modalities, RGB and depth data, commonly utilized in applications like autonomous driving, benefiting from the detailed 3D structural information provided by depth data. Coarse-grained action recognition typically encompasses full-body motions (*e.g.*, cleaning up and cooking). Conversely, fine-grained action recognition involves subtle interactions between humans and objects (*e.g.*, turn-on tap, turn-off tap, wash spoons, and wash plates), presenting challenges due to the presence of various objects, similar backgrounds, high inter-class similarity, and low

intra-class similarity.

A primary challenge in action recognition lies in effectively representing actions within videos. Human actions exhibit variability in motion speed, camera perspective, and appearance [5]. The objective of action representation is to convert action videos into feature vectors, capturing representative and discriminative information while minimizing variations to enhance recognition performance [6].

Traditionally, hand-crafted action representation methods [5] [7] [8] involve predefined parameters set by experts. However, these traditional methods demand significant expertise and fail to capture long-term temporal information in videos.

With the advent of deep learning, action recognition approaches [9] have employed 2D convolutions to capture spatial and temporal features in videos. The breakthrough of AlexNet [10] in image classification paved the way for two-stream convolutional networks (two-stream CNNs) [9], marking a significant milestone in action recognition. From that point forward, the application of 2D convolutions on stacked RGB video frames and optical flow frames has become a mainstream method for extracting spatial and temporal features [11] [12]. Subsequently, large-scale action recognition datasets facilitated the training of spatial-temporal 3D CNNs, which requires substantial computational resources and extensive training data [13].

Inspired by the success of Transformers [14] in Natural Language Processing (NLP), Transformer models have been adapted for action recognition. Video-based action recognition shares similarities with NLP, treating videos as sequences akin to sentences. Models such as Vision Transformer (ViT) [15] and Video Vision Transformer (ViViT) [16] utilize Transformer to extract spatial and temporal features from video sequences. However, the Transformer [14] lacks certain image-specific inductive biases, prompting the exploration of hybrid CNN-Transformer structures, such as RViT [17], CVTN [18], and CTN [19], which leverage the strengths of both architectures for action recognition.

Overall, the evolution of action recognition methodologies reflects a continuous quest for robust spatiotemporal action representations and efficient spatial-temporal interaction modeling, with hybrid architectures poised to advance action recognition. Our exploration primarily encompasses three subtasks within action recognition: RGB-D, coarse-grained, and fine-grained action recognition.

1.2 Motivations of this research

Despite significant progress in action recognition, several challenges persist in the field.

- **Lack of fine-grained motion differences in RGB-D action representations**. Although many RGB-D based action and gesture recognition approaches [20] have achieved remarkable results through the utilization of highly integrated spatio-temporal representations across multiple modalities (*i.e.*, RGB and depth data), they still face several challenges. Firstly, vanilla 3D convolution struggles to capture fine-grained motion differences between local clips across different modalities. Secondly, the intricate nature of highly integrated spatio-temporal modeling can lead to optimization difficulties. Thirdly, the presence of duplicate and unnecessary information can increase complexity and complicate entangled spatio-temporal modeling.

- **Insufficient spatial relationship in coarse-grained action representations**. In recent developments, action recognition approaches have embraced architectures such as ViT and ViviT [16], which incorporate an absolute-position embedding strategy at the patch embedding stage. However, it is crucial to acknowledge a potential drawback of absolute position embedding: it may not adequately represent spatial relationships between tokens. This limitation could lead to overlooking the inherent spatial structure within a video frame, thereby constraining the ability of the model to capture fine-grained spatial dependencies and contextual information.

- **Lack of inductive biases and subtle variations in fine-grained action representations**. Fine-grained action recognition relies on capturing subtle variations in specific sub-motions, objects, and poses [21]. A common assumption in visual tasks is that neighboring pixels in video frames are always correlated. Therefore, modeling the relationship between neighboring features and low-level features in video frames is crucial for fine-grained action recognition. However, existing Transformer models [16] [22] may struggle to fully utilize these inductive biases present in videos, as they typically require a substantial amount of data to learn these nuances.

1.3 Research Objectives

- **Objective 1** is to capture fine-grained motion differences and reduce redundant information by developing a multi-stage spatial-temporal model for RGB-D action recognition. The primary aim is to learn bottom-level visual features and hierarchical spatial and temporal features from videos.
- **Objective 2** is to extract spatial and temporal dependencies among different frames by developing a novel spatially and temporally decoupled model for coarse-grained and fine-grained action recognition. Additionally, our focus involves effectively capturing subtle interactions and dependencies among different regions of the frames using an innovative positional embedding method for action recognition.
- **Objective 3** is to integrate inductive biases and subtle variations by developing a novel CNNs-transformer hybrid model for fine-grained action recognition. This hybrid model leverages the strengths of CNNs, such as weight sharing and local receptive fields, along with the advantages of Transformers, including global receptive fields and dynamic attention mechanisms.

1.4 Contributions

Throughout this thesis, we will concentrate on three sub-tasks of action recognition: RGB-D action recognition and gesture recognition (Chapter 2) and fine-grained and coarse-grained action recognition (Chapters 3 and 4). The contributions in each of these chapters are summarized as follows:

(i) Multi-stage spatio-temporal representation for RGB-D action recognition

- We propose a novel multi-stage hierarchical structure to model local spatio-temporal contexts from low-level edges and build up to higher-order semantic primitives using a multi-stage hierarchy structure. Concurrently, temporal and spatial dependencies have also been implicitly modeled in a single stage.
- We propose a central difference convolutional-based stem to capture bottom-level features (*e.g.,* patterns and colors) that form fundamental structures in video frames (*e.g.*, corners and lines).

- We establish a stack of inception-based multi-scale spatial features learning layers to extract hierarchical local spatial features, coupled with a Transformer that captures global spatial features. Moreover, we propose a temporal block composed of multiple Transformer layers for hierarchical local fine-grained and global coarse-grained temporal feature learning.

(ii) Spatially and temporally decoupled Transformer model for coarse-grained and fine-grained action recognition

- We propose a novel spatially and temporally decoupled/disentangled Transformer model to capture spatial and temporal action-recognition dependencies. The spatial and temporal Transformer encoders are constructed dimension-independently.

- We enhance the spatial Transformer encoders by introducing relative-position embeddings, which replace the traditional absolute-position embeddings in self-attention. This modification allows the model to capture subtle interactions and dependencies among different regions of the frames.

- We incorporate a deviation-based temporal subsampling layer into the temporal Transformer encoders. This layer selectively maps the most informative components to features over temporal indices while reducing the computational cost.

(iii) Convolutional transformer network for fine-grained action recognition

- We propose a novel Convolutional enhanced Transformer Network (CTN) model for fine-grained action recognition. The proposed CTN model takes all the advantages of CNNs, including sharing weights and local receptive fields, while inheriting all the benefits of Transformer, including the global receptive field and dynamic attention.

- Experimental results indicate that the proposed CTN shows better results compared to models constructed on CNNs and models constructed on pure-Transformer on two fine-grained action recognition datasets while utilizing a similar number of FLOPs.

1.5 Organization of this thesis

Literature reviews of the deep learning based action recognition methods are presented in each chapter corresponding to the proposed three novel deep learning approaches.

The rest of this thesis is organized as follows:

Chapter 2 presents the proposed multi-stage factorized spatio-temporal model, which is based on our work published in the *Proceedings of the 31st ACM International Conference on Multimedia 2023*, titled "Multi-stage factorized spatio-temporal representation for RGB-D action and gesture recognition" [23].

Chapter 3 presents the proposed relative-position embedding based spatially and temporally decoupled Transformer model, which is based on our work published in the journal of *Pattern Recognition* 145 (2024), titled "Relative-position embedding-based spatially and temporally decoupled Transformer for action recognition" [24].

Chapter 4 presents the proposed convolutional transformer network, based on our work published in the journal of *Neurocomputing* 569 (2024), titled "Convolutional transformer network for fine-grained action recognition" [19].

Chapter 5 concludes this thesis and discusses future work and directions.

Note that references related to each chapter are listed at the end of each chapter, and the Statements of Contributions are inserted at the beginning of each relevant chapter.

References

[1] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[2] Weirong Sun, Yujun Ma, and Ruili Wang. k-nn attention-based video vision transformer for action recognition. *Neurocomputing*, 574:127256, 2024.

[3] Yan Tian, Yifan Cao, Jiachen Wu, Wei Hu, Chao Song, and Tao Yang. Multi-cue combination network for action-based video classification. *IET Computer Vision*, 13(6):542–548, 2019.

[4] Tianyu Liu, Yujun Ma, Wenhan Yang, Wanting Ji, Ruili Wang, and Ping Jiang. Spatial-temporal interaction learning based two-stream network for action recognition. *Information Sciences*, 606:864–876, 2022.

[5] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[6] Ming Zong, Ruili Wang, Yujun Ma, and Wanting Ji. Spatial and temporal saliency based four-stream network with multi-task learning for action recognition. *Applied Soft Computing*, 132:109884, 2023.

[7] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3169–3176. IEEE, 2011.

[8] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[9] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[11] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.

[12] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[13] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2012.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

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

[16] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[17] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14063–14073, 2022.

[18] Youcef Djenouri and Ahmed Nabil Belbachir. A hybrid visual transformer for efficient deep human activity recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 721–730, 2023.

[19] Yujun Ma, Ruili Wang, Ming Zong, Wanting Ji, Yi Wang, and Baoliu Ye. Convolutional transformer network for fine-grained action recognition. *Neurocomputing*, 569:127027, 2024.

[20] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, Fan Wang, and Du Zhang. Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20154–20163, 2022.

[21] Miao Ma, Naresh Marturi, Yibin Li, Ales Leonardis, and Rustam Stolkin. Region-sequence based six-stream cnn features for general and fine-grained human action recognition in videos. *Pattern Recognition*, 76:506–521, 2018.

[22] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13505–13515, 2021.

[23] Yujun Ma, Benjia Zhou, Ruili Wang, and Pichao Wang. Multi-stage factorized spatio-temporal representation for rgb-d action and gesture recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3149–3160, 2023.

[24] Yujun Ma and Ruili Wang. Relative-position embedding based spatially and temporally decoupled transformer for action recognition. *Pattern Recognition*, 145:109905, 2024.

Chapter 2

Multi-stage Factorized Spatio-Temporal Representation for RGB-D Action and Gesture Recognition

*RGB-D action and gesture recognition remain an interesting topic in human-centered scene understanding, primarily due to the multiple granularities and large variation in human motion. Although many RGB-D based action and gesture recognition approaches have demonstrated remarkable results by utilizing highly integrated spatio-temporal representations across multiple modalities (*i.e., *RGB and depth data), they still encounter several challenges. Firstly, vanilla 3D convolution makes it hard to capture fine-grained motion differences between local clips under different modalities. Secondly, the intricate nature of highly integrated spatio-temporal modeling can lead to optimization difficulties. Thirdly, duplicate and unnecessary information can add complexity and complicate entangled spatio-temporal modeling. To address the above issues, we propose an innovative heuristic architecture called Multi-stage Factorized Spatio-Temporal (MFST) for RGB-D action and gesture recognition. The proposed MFST model comprises a 3D Central Difference Convolution Stem (CDC-Stem) module and multiple factorized spatio-temporal stages. The seamless integration of these innovative designs results in a robust spatio-temporal representation that outperforms state-of-the-art approaches on RGB-D action and gesture recognition datasets. Note that the work presented in this chapter has been published in the 31st ACM International Conference on Multimedia.*

2.1 Introduction

Action and gesture recognition has garnered significant attention in video understanding owing to its wide-ranging application scenarios, such as intelligent driving, human-machine interaction, and virtual reality [1] [2] [3]. This field has made significant progress by leveraging a wide range of data representations, such as visual appearance, skeleton, depth, and optical flows [4] [5] [6]. Among those, the RGB-D-based action and gesture recognition has drawn lots of interest due to its strong adaptability to dynamic circumstances and complex backgrounds [7]. Compared to the RGB modality, the depth modality is less sensitive to illumination, invariant to color and texture changes [8] and can provide detailed 3D structural information about the scene [9].

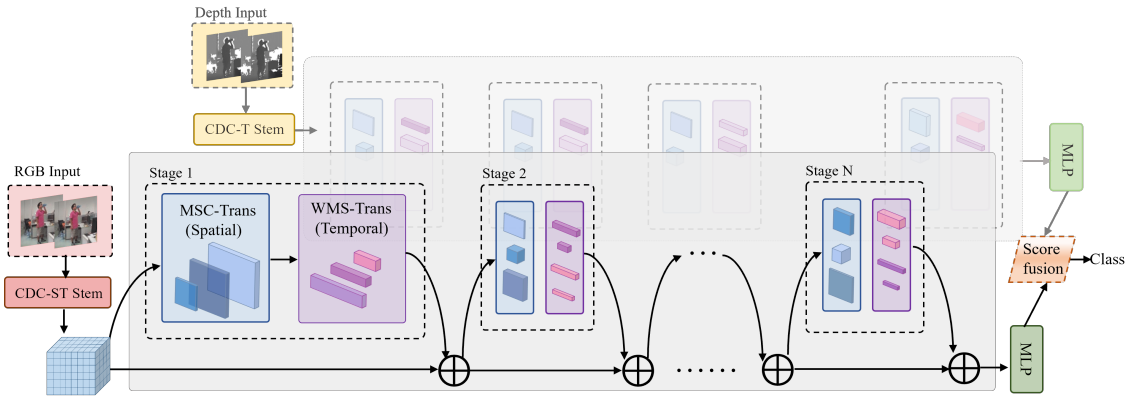


Figure 2.1: Illustration of the proposed MFST model: CDC-Stem, multiple spatio-temporal stages. Each stage starts with MSC-Trans block for spatial features, followed by WMS-Trans layers for temporal features. Residual connections optimize the model, and $\oplus$ indicates element-wise addition.

The existing RGB-D action and gesture recognition methods [10] [11] [12] commonly employ convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to extract features from RGB and depth cues [13]. However, in recent years, Transformer-based methods [14] [15] [16] have achieved surprising performance for modeling long-range dependencies regarding temporal dynamics for RGB-D action and gesture recognition. Transformer [17] is suitable for processing sequential data as it can effectively model the relationships among different frames, which enables the model to recognize and interpret the same motion occurring at different time indexes. However, most Transformer-based methods [18] [19] [20] process spatial and temporal information using the same strategy or combine spatio-temporal information without

explicitly analyzing the differences between the spatial and temporal dimensions in RGB-D data.

Specifically, there are still some problems in the following three aspects. (i) Vanilla 3D convolution (e.g., C3D [21] and I3D [22]) is widely used in action recognition due to its effective modeling of spatiotemporal features. However, **fine-grained motion differences might be overlooked when employing vanilla 3D convolution, which operates on the entire input volume simultaneously, without considering the subtle motion differences that may arise within local clips or across different modalities. This can limit the ability of the model to distinguish between similar actions or events that differ only in subtle ways.**

(ii) **Optimization difficulties** often arise when working with limited video data, primarily due to the highly interdependent and intertwined nature of spatio-temporal modeling. The intricate relationships between the spatial and temporal information can make optimizing the model parameters challenging, resulting in reduced performance.

(iii) **Duplicate and unnecessary information** can create additional complexity and make it difficult to deal with in entangled spatio-temporal modeling. To solve the above-mentioned problems, some decoupled models [23] [24] [25] have been proposed, which model spatial and temporal representations separately. For example, Bertasius *et al.* [26] presented a divided space-time attention model for action recognition, which separately applied temporal and spatial attention within each block. However, such divided space-time attention could potentially result in losing essential spatio-temporal relationships since there is no explicit mechanism for spatial and temporal feature interaction or spatio-temporal feature coupling. Later, Zhou *et al.* [27] proposed to utilize a distillation based recouple strategy to establish the space-time dependency of decoupled spatial and temporal representations for RGB-D based action and gesture recognition. However, such decoupling and recoupling structures could reintroduce bias or errors and require additional computation. Furthermore, distilling the inter-frame correlations from the time dimension into the space dimension could potentially guide the model prone to temporal information.

(iv) **Insufficient exploration** of RGB-D multi-modality for action and gesture recognition based on videos. Most existing approaches focus on RGB action and gesture recognition, which may not be sufficient to capture a person's subtle movements and

poses during an action. However, cost-effective depth sensors like Microsoft Kinect can capture depth information in addition to color information, enabling more accurate tracking of body movements and poses, providing additional pixel-level positional cues of the scene for recognizing actions and gestures. Thus, there is still a significant research gap in exploring effective multi-modal spatio-temporal representation for dynamics RGB-D action and gesture recognition.

Given the above-mentioned concerns, as illustrated in Figure 2.1, we introduce a new Multi-stage Factorized Spatio-Temporal (MFST) learning architecture for RGB-D-based action and gesture recognition. Broadly, it begins with a lightweight 3DCDC-based [28] convolutional Stem (CDC-Steam) module to capture fine-grained spatio-temporal features. Then, inspired by the computationally efficient manner of CNN's hierarchical structure, we partition the spatio-temporal learning process into multiple factorized spatio-temporal stages with residual connections to learn hierarchical spatial and temporal features via the proposed Multi-Scale Convolution and Transformer (MSC-Trans) hybrid block and Weight-shared Multi-Scale Transformer (WMS-Trans) block.

In more detail, as shown in Figure 2.1, the gist of our proposed model is composed of three aspects: (1) **Convolutional lightweight Stem**. Firstly, we introduce a Central Difference Convolutional (CDC) [28] based Stem to capture bottom-level features (*e.g.*, patterns and colors) that form fundamental structures in video frames (*e.g.*, corners and lines). More specifically, we use the Spatio-Temporal CDC based Stem (CDC-ST Stem) to process the RGB data and the Temporal CDC based Stem (CDC-T Stem) to process the depth data. This consideration of stems from the fact that the former is sensitive to both temporal and spatial priors, while the latter is only sensitive to temporal priors. (2) **Multi-stage hierarchical structure**. The multi-stage process contains spatial feature embeddings and temporal embeddings simultaneously. This hierarchical manner allows the model to model local spatio-temporal contexts from low-level edges and build up to higher-order semantic primitives using a multi-stage hierarchy structure, similar to Convolutional Neural Networks (CNNs). Concurrently, temporal and spatial dependencies have also been implicitly modeled in a single stage. (3) **Factrized spatio-temporal learning**. As mentioned before, each stage begins with a Multi-Scale Convolution and Transformer (MSC-Trans) hybrid spatial block, and a Weight-shared Multi-Scale Transformer (WMS-Trans) temporal block comprises the remainder of each stage. Concretely, the MSC-Trans spatial block

consists of a stack of inception-based multi-scale spatial features learning layers to extract hierarchical local spatial features, coupled with a Transformer that captures global spatial features. The WMS-Trans temporal block is composed of multiple Transformer layers for hierarchical local fine-grained and global coarse-grained temporal feature learning. Specially, we adopt a Transformer structure based on Kvt [29], which reduces the redundant marginal information in extracted temporal features.

Our work is inspired by the De-Recouple model [27], but we introduce the MFST approach, which differs in spatial-temporal representation. Our method employs a fine-grained multi-stage process, capturing spatio-temporal features individually in each stage. In contrast, De-Recouple uses a global way, potentially leading to temporal dominance due to asymmetries in the modeling capabilities of sub-networks. Additionally, we propose a weight-shared single Transformer branch for more efficient temporal multi-scale feature modeling compared to their multi-branch approach. With these designs, our MFST model effectively learns spatio-temporal features within each modality for action and gesture recognition. To the best of our knowledge, we are the first to propose an efficient factorized spatio-temporal representation at the fine-grained stage level for RGB-D action and gesture recognition rather than a coarse-grained model level.

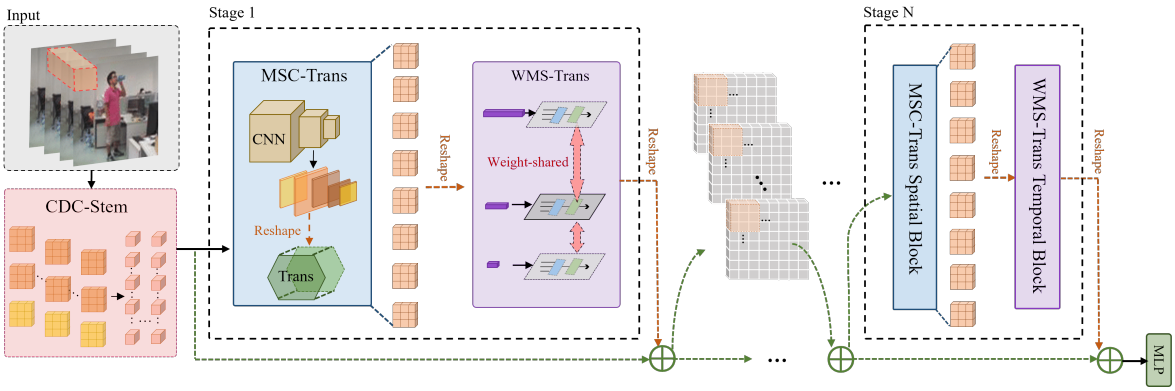


Figure 2.2: Illustration of our proposed MFST model for RGB-D action and gesture recognition includes CDC-Stem, multiple stages with MSC-Trans hybrid spatial blocks, and WMS-Trans temporal blocks. Residual connections are utilized between consecutive stages to optimize the model, and $\oplus$ indicates element-wise addition.

2.2 Related work

2.2.1 RGB-D Data based Action and Gesture Recognition

With the recent advancements in cost-effective RGB-D sensors, such as Microsoft Kinect and Asus Xtion, there has been a growing interest in RGB-D-based action and gesture recognition. This is mainly due to the fact that the additional dimension (depth) is unaffected by changes in illumination and provides comprehensive 3D structural information of the motion [30]. Recent works in RGB-D gesture and action recognition have proposed different techniques to capture and enhance temporal and spatial information from RGB-D data. Wang *et al.* [31] utilized scene flow for compact RGB-D representation learning, which considered the modalities as one entity. Later, Yu *et al.* [28] proposed a temporal enhanced multi-modality model for gesture recognition, which utilized three different types of 3D central difference convolution to capture rich temporal dynamics information. Wang *et al.* [32] proposed a single convolutional neural network based on cooperative training for RGB-D action recognition, which enhanced the discriminative power of the deeply learned features. Cheng *et al.* [10] proposed a cross-modality compensation model for RGB-D action recognition, which employed a compensation block to capture interaction information between different modalities. Liu *et al.* [11] presented a dual-stream Transformer model for RGB-D action recognition that incorporates dual streams and a cross-modality fusion layer to extract spatio-temporal features simultaneously. Unlike global-based spatio-temporal decoupling methods [24] [27], we focus on decomposing spatio-temporal dependencies in each local stage. Our approach can achieve the goal of spatio-temporal decoupling and reduce the risk of spatio-temporal connection weakening caused by the complete decoupling of space and time.

2.2.2 Disentangled spatio-temporal representation

Disentangled spatio-temporal models decouple the spatial and temporal information in videos, enabling the use of spatial and temporal architectures optimized for their respective domains. Shi *et al.* [33] present a decoupled spatio-temporal attention network for skeleton-based action recognition, emphasizing the specific characteristics of space/time and different motion scales. Bertasius *et al.* [26] introduced a divided space and time attention Transformer model for RGB action recognition, which first computed self-attention across the spatial dimensions, followed by computing self-

attention across the temporal dimension. Zhang *et al.* [24] proposed performing spatial and temporal attention for RGB action recognition, respectively, by decoupling the 3D self-attention to a spatial attention and a temporal attention. Liu *et al.* [34] proposed a disentangled spatio-temporal Transformer for video inpainting. The model attended to temporal features on different frames at the same spatial pixels and then to similar background textures on the same frame at all spatial positions. However, coarse-grained or global-level model decoupled spatio-temporal networks have a limitation in effectively incorporating dependencies between spatial and temporal information, which can result in incomplete decoupling. In this study, we aim to acquire both spatial and temporal characteristics, as well as their dependencies, for video-based RGB-D action and gesture recognition. To achieve this, we propose a hierarchically multi-stage factorized spatio-temporal representation that incorporates residual connections.

2.3 Proposed MFST model

2.3.1 Overview

In this work, we propose a hierarchical multi-stage factorized spatio-temporal feature representation learning architecture for RGB-D action and gesture recognition. Multi-stage hierarchy design borrowed from CNNs is adopted, where four stages in total are used in this work (described in Section 2.3.2. The proposed MFST model comprises a 3D Central Difference Convolutional Stem (CDC Stem) module (detailed in Section 2.3.3, a sequence of factorized spatio-temporal feature learning stages (described in Section 2.3.4, and an MLP-based prediction head.

2.3.2 Hierarchical multi-stage factorized spatio-temporal representation

The MFST model (Figure 2.2) comprises a 3D Central Difference Convolutional Stem (CDC-Stem) module as its first step. This module generates visual embeddings that are fed into cascaded factorized spatio-temporal stages, consisting of four stages in total. To capture multi-scale spatial features both locally and globally, each stage uses a hybrid spatial block (MSC-Trans) that consists of a Multi-Scale Convolution module followed by a Transformer. Then, the spatial features obtained from each

stage are input into a Weight-shared Multi-Scale Transformer block (WMS-Trans) to capture multi-scale temporal features. Additionally, we also introduce the residual connection [35] between consecutive stages to ensure stable gradients. Finally, an MLP layer makes the final predictions based on the output obtained from the final stage.

Let $V \in \mathbb{R}^{T\times 3\times H\times W}$, where T is the length of the video and H and W represent the height and width of each frame, respectively. The embeddings Z after the first stage can be formally expressed as:

$$Z = WSM\text{-}Trans(MSC\text{-}Trans(CDC\text{-}Stem(V))). \tag{2.1}$$

As depicted in Figure 2.2, the output of the MSC-Trans block is reshaped to serve as input for the WMS-Trans block.

The factorized spatio-temporal representation stage combines the MSC-Trans spatial block and WMS-Trans temporal block, enabling the simultaneous learning of visual patterns that incorporate both local spatial information and higher-order temporal information. Moreover, integrating these two types of blocks within a single stage allows the model to effectively capture spatio-temporal context at different levels of complexity. The factorized spatio-temporal representation stage also facilitates the effective modeling of spatio-temporal dependencies without recoupling or reconstructing spatio-temporal representations, unlike the approach [27].

2.3.3 Central Difference Convolutional Stem

We propose a lightweight Central Difference Convolutional Stem (CDC-Stem) module that embeds bottom-level visual features such as lines, edges, and corners before passing them to subsequent multi-stage representations. The Stem is constructed using the first five layers of the Inflated 3D Convolutional Neural Network (I3D) [22], similar to the design used in [36]. However, vanilla 3D convolution makes it hard to capture fine-grained motion differences between local clips across different modalities. To further enhance the spatio-temporal representation and fully exploit rich local motion information, we adopt the 3D Central Difference Convolution (3D-CDC) [28] in our Stem module.

Specifically, we replace the vanilla 3D operation with 3D-CDC in the Stem and employ different CDC operations for RGB modality and Depth data. As illustrated in Figure 2.3, the 3D-CDC module consists of two distinct steps: sampling and aggregation. The sampling step is similar to vanilla 3D convolution, while the aggregation step differs. Figure 2.3 (a) shows the Spatio-Temporal Central Difference Convolution (CDC-ST) in the Stem module for the RGB modality, and Figure 2.3 (b) shows the Temporal Central Difference Convolution (CDC-T) in the Stem module for the depth modality.

The input tensor V is first passed through a CDC-Stem module. In general, the CDC-Stem module is written as follows:

$$E \;=\; CDC\text{-}Stem(V, \Psi),\; E \;\in \mathbb{R}^{T\times C\times H\times W}, \tag{2.2}$$

where Ψ indicates the learnable parameter matrices. After passing the input tensor V through the CDC-Stem module, the resulting embedded spatio-temporal features E are further processed by hierarchical factorized spatio-temporal feature learning stages.

2.3.3.1 CDC-ST Stem for RGB

We adopt CDC-ST in the Stem module for the RGB modality because RGB data provides detailed appearances, such as local shape changes of the hands and fingers, which are better suited to spatial gradient cues and appearance details. CDC-ST combines spatio-temporal gradient information into a single 3D convolution operator, effectively capturing subtle changes across the frame sequence. Thus, our proposed CDC-ST Stem leverages the rich local motion information to enhance the spatial-channel representation, which is critical for identifying the fine-grained temporal dynamics of motions and differentiating between similar actions and gestures.

Explicitly, the vanilla 3D convolution operation involves sampling a local receptive field cube $\mathcal{Z}$ over the input feature map x and aggregating the sampled values via a weighted summation with learnable weights w. The output feature map y_{3D} can be formulated as:

$$y_{3D}(p_0) = \sum_{p_n \in \mathcal{Z}} w(p_n) \cdot x(p_0 + p_n), \tag{2.3}$$

where p_0 indicates the current location on both input and output feature maps while p_n enumerates the locations in $\mathcal{Z}$. The output feature map of CDC-ST can be formulated as:

$$y_{CDC-ST}(p_0) = \sum_{p_n \in \mathcal{Z}} w(p_n) \cdot (x(p_0 + p_n) - x(p_0)). \tag{2.4}$$

To enhance the robustness and discriminative modeling capacity in the proposed CDC-ST Stem module, we utilized the setting described in [28], which involves a combination of vanilla 3D convolution and 3D-CDC-ST. The generalized 3D-CDC-ST operation can be formulated as follows:

$$y_{3D-CDC-ST}(p_0) = \theta \cdot y_{CDC-ST}(p_0) + (1 - \theta) y_{3D}(p_0), \tag{2.5}$$

where the hyperparameter θ ranges from *0* to *1*, determines the trade-off between the contribution of intensity-level and gradient-level information.

2.3.3.2 CDC-T Stem for Depth

We adopt CDC-T in the Stem module for the depth modality because depth data directly captures the distance between objects and the camera, which is more sensitive to temporal changes in the environment than RGB data. However, depth data is often noisier and less detailed than RGB data, making it challenging to extract accurate temporal information. By proposing CDC-T Stem, we can effectively capture central temporal differences and filter out some of the noise in the depth data, resulting in a sequence of filtered depth frames that are more sensitive to dynamic temporal changes.

As shown in Figure 2.3 (b), in 3D-CDC-T operation, the sampled local receptive field cube $\mathcal{Z}$ is divided into two regions: 1) the region in the current time step $\mathcal{T}'$, and 2) the regions in the adjacent time steps $\mathcal{T}''$. In the context of a 3D-CDC-T, the central difference term is only computed from the region $\mathcal{T}''$. Thus the generalized 3D-CDC-T can be formulated as follows:

$$y_{3D-CDC-T}(p_0) = y_{3D}(p_0) + \theta \cdot \left(-x(p_0) \cdot \sum_{p_n \in \mathcal{T}''} w(p_n) \right). \tag{2.6}$$

The results of the ablation experiments in Section 2.3.2 confirm that 3D-CDC-ST is more effective in processing RGB data, while 3D-CDC-T performs better in handling depth data. The experimental results align with our hypothesis that 3D-CDC-T is better suited for depth data due to its ability to reduce sensor noise and pre-processing artifacts between frames of depth data. Moreover, processing depth data requires more emphasis on temporal reasoning context rather than spatial gradient cues and appearance details, which is why 3D-CDC-ST is better suited for RGB data.

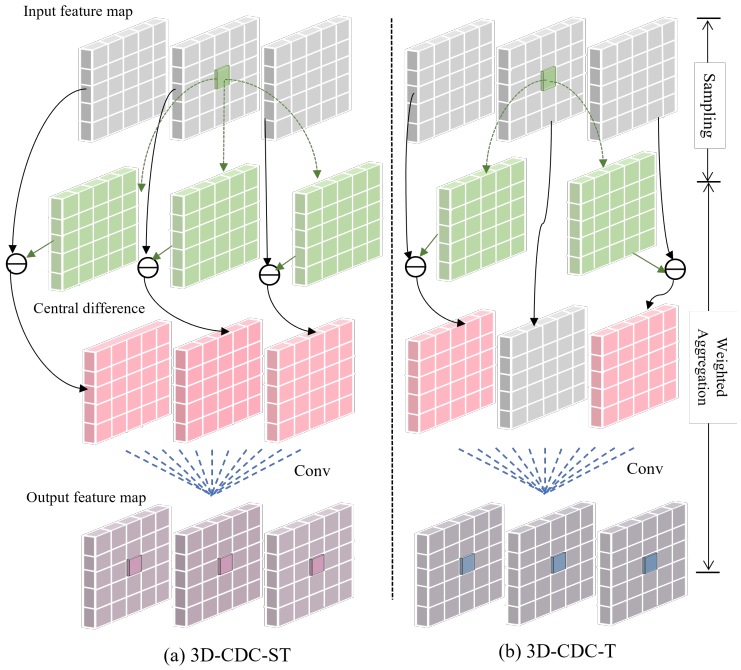


Figure 2.3: (a) 3D-CDC-ST incorporates central difference information from the entire local spatio-temporal regions; (b) 3D-CDC-T only utilizes central difference clues from the local spatio-temporal regions of the adjacent frames.

2.3.4 Factorized Spatio-Temporal Representation

Our proposed model involves the factorization of spatio-temporal representation in each stage. Each stage commences with a Multi-Scale Convolution and Transformer (MSC-Trans) hybrid spatial block, followed by a Weight-shared Multi-Scale Transformer (WMS-Trans) temporal block that comprises the remainder of the stage.

2.3.4.1 MSC-Trans block

As shown in Figure 2.4, the MSC-Trans spatial block consists of a stack of inception-based multi-scale spatial features learning layers to extract hierarchical local spatial features, coupled with a k-NN Transformer [29] that captures global spatial features. Multi-Scale Convolution (MSC) operation comprises a space-centric 3D Inception Module and a 3D Max Pooling layer, which models local multi-scale spatial features, as denoted by:

$$F_{MSC} = MaxPool(InC_{1\times k\times k}(X,\theta)), \tag{2.7}$$

where $InC_{1\times k\times k}(X,\theta)$ refers to the Inception Module with a kernel size of $1 \times k \times k$. The notion of capturing hierarchical spatial features arises from the recognition that spatial patterns vary at multiple scales when an action takes place.

The multi-scale local spatial feature from MSC is then reshaped into a size that can be accepted by the Transformer. To reduce the redundant marginal information in captured temporal features, here we utilize a Transformer structure based on K-NN multi-head self-attention layer [29] for global spatial feature modeling. Therefore, the modeling process of the K-NN Transformer can be computed through:

$$L_{Trans} = MLP(LN(MSA_{knn}(F_{MSC}))), \tag{2.8}$$

where F_{MSC} denotes the output of the MSC operation, MSA_{knn} refers to k-NN multi-head self-attention layer and LN represents layer normalization. In order to maintain a balance between the MSC-Trans block and WMS-Trans block and prevent the factored spatio-temporal stage from being biased towards any dimension, we employ four Transformer layers with a head number of eight in the proposed MSC-Trans block.

2.3.4.2 WMS-Trans block

Video data contains both spatial and temporal information, and existing approaches typically use 2D convolutions to capture spatial information and 3D components in the higher layers to extract temporal information [37] [38]. Building on this concept, we propose a Weight-shared Multi-scale Transformer (WMS-Trans) temporal block after the MSC-Trans spatial block. The WMS-Trans block comprises a Trans-

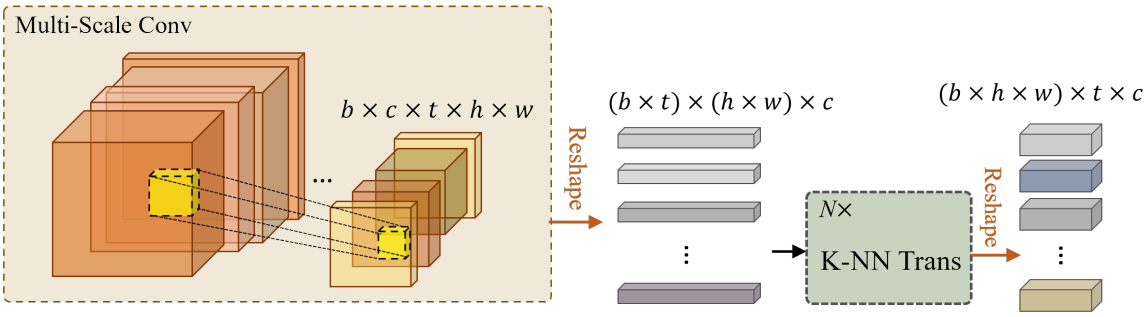


Figure 2.4: The MSC-Trans block is composed of two main parts: the Multi-Scale Convolution (MSC) layers and a K-NN Transformer.

former branch that performs hierarchical local fine-grained and global coarse-grained temporal feature learning.

To obtain multiple temporal scales, we reshape the multi-scale spatial features from the previous MSC-Trans block into various temporal dimensions. As shown in Figure 2.5, the Transformer branch in the WMS-Trans block models multi-scale temporal features by handling different temporal dimensions of the input sequence. For example, the multi-scale input can be represented as $(\mathbb{R}^{b,t,c})$, $(\mathbb{R}^{b,\frac{1}{2}t,c})$, and $(\mathbb{R}^{b,\frac{1}{4}t,c})$, which cover short-term to long-term temporal dependencies. A weight-shared strategy enables saving the number of parameters and promotes interaction between different temporal scales. The WMS-Trans block can effectively learn a more generalizable representation of the input embeddings by sharing the same parameters across different temporal scales. Our weight-shared multi-scale Transformer differs from the De+Recouple model [27]. They use a multi-branch Transformer for capturing multi-scale temporal information without weight-sharing, which is computationally heavier and may neglect dependencies among different scales of temporal information.

Furthermore, we have incorporated positional embeddings into the input embeddings of the WMS-Trans block. As shown in Figure 2.5, the positional encodings share the same dimensions as the input embeddings, enabling their direct addition. Following [18], we have utilized sine and cosine functions with varying frequencies to achieve this:

$$PE(pos, 2i) = sin\left(pos/10000^{2i/d_{model}}\right), \tag{2.9}$$

$$PE(pos, 2i+1) = cos\left(pos/10000^{2i/d_{model}}\right), \tag{2.10}$$

where *pos* denotes the position in the sequence, and i represents the dimension index within the embeddings. The WMS-Trans block can learn spatial relationships and capture meaningful patterns in the input embeddings by adding positional embedding. To sum up, the WMS-Trans block is capable of capturing temporal features at different time scales hierarchically, allowing it to model complex and wide range temporal patterns in the input embeddings.

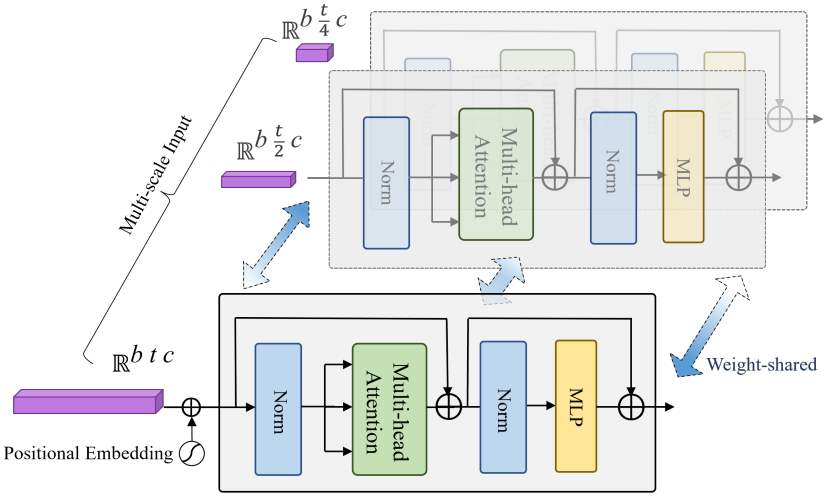


Figure 2.5: Illustration of the WMS-Trans block, consisting of a Transformer branch with multi-scale temporal dimension inputs, and a weight-shared strategy is applied.

2.4 Experiments

2.4.1 Implementation Details

All experiments were conducted using PyTorch and were run on 4 × A100 GPUs. Input sequences were randomly or center-cropped to 224 × 224 during both training and inference. We utilized SGD as the optimizer with a weight decay of 0.0003 and a momentum of 0.9. The learning rate was linearly ramped up to 0.01 during the first three epochs and then decayed with a cosine schedule. The training process lasted 100 epochs, and we adopted MixUp [39] as a data augmentation strategy.

We evaluated our MFST baseline model (MFST-BM) on the NTU RGB-D dataset, which does not incorporate CDC, residual connections, and weight-shared Transform-

ers. However, it still employs a hierarchical multi-stage factorized spatio-temporal approach using vanilla 3D convolutions and Transformers. Furthermore, we conducted experiments with two variations of our model that incorporate CDC, residual connections, and weight-shared Transformers: MFST-Base, which has three stages, and MFST-Large, which has four stages. In the basic configuration, we used four Transformer layers ($L = 4$) in the temporal feature learning block and eight heads ($N_H = 8$). Unless specified otherwise, we refer to this setting as the basic configuration of our network.

2.4.2 Comparison with State-of-the-art Methods

The proposed method achieves state-of-the-art (SOTA) performance on both large-scale RGB-D action and gesture recognition datasets (*e.g.*, NTU RGB-D, and IsoGD). More SOTA comparison performance can be found in the supplementary material.

2.4.2.1 NTU RGB-D Dataset

NTU RGB-D [40] is a highly utilized large-scale indoor action recognition dataset that includes over 56 thousand action clips and 4 million frames across 60 action classes. The dataset consists of clips performed by 40 volunteers captured in a constrained laboratory setting using three KinectV2 cameras, each providing a different viewpoint. NTU RGB-D presents a significant challenge due to the considerable intra-class and viewpoint variations. Many recent studies have utilized the skeleton information to perform 2D/3D convolutions for action recognition. However, skeleton data provide limited information about the human body and its movements, are more sensitive to occlusions, and inability to capture fine-grained details. In this work, we focus solely on the RGB and depth modalities for action recognition. Our results, as presented in Table 2.1, demonstrate SOTA performance on both the Cross-View (CV) and Cross-Subject (CS) protocols. Compared to the SOTA skeleton-based method Ta-CNN [41], our proposed MFST-Large model achieves 4.2% performance improvement on the CS protocol and 3.4% on the CV protocol. Compared to the SOTA RGB-D based method De+Recouple [27], our proposed MFST-Large model achieves approximately 0.4% on the CS protocol and 0.9% on the CV protocol. Our results demonstrate superior performance over the De+Recouple [27] model on both the RGB and depth modalities. The proposed MFST model shows strong motion perception abilities and robustness to noisy backgrounds, as evidenced by the comparable or superior

Table 2.1: Comparison of the SOTA methods on NTU RGB-D.

Methods	Publisher	Modality	CS (%)	CV(%)
S-GCN[42]	CVPR20	SKL	90.7	69.2
DC-GCN[43]	ECCV20	SKL	90.8	69.4
MST-GCN[44]	AAAI21	SKL	91.5	96.6
ResGCN[45]	ACMMM20	SKL	90.9	96.0
STST[46]	ACMMM21	SKL	91.9	96.8
Ta-CNN[41]	AAAI22	SKL	90.4	94.8
STAR-Trans[15]	CVPR23	SKL+RGB	92.0	96.5
D-Bilinear[47]	ECCV18	RGB-D+SKL	85.4	90.7
Coop-CNN[32]	AAAI18	RGB+DEP	86.4	89.1
P4Trans[48]	CVPR21	Points	90.2	96.4
De+Recouple[27]	CVPR22	RGB	90.3	95.4
De+Recouple[27]	CVPR22	DEP	92.7	96.2
De+Recouple[27]	CVPR22	RGB+DEP	94.2	97.3
Ours(MFST-Large)	-	RGB	**92.1**	**95.8**
Ours(MFST-Large)	-	Depth	**93.8**	**97.3**
Ours(MFST-Large)	-	RGB+DEP	**94.6**	**98.2**

performance of the depth modality.

2.4.2.2 Chalearn IsoGD Dataset

The Chalearn IsoGD dataset [49] comprises 47,933 RGB-D gesture videos categorized into 249 different types of gestures, and 21 individuals performed these videos. This dataset is considered more challenging due to two main factors: firstly, it covers gestures in various fields and ranges of motion, from delicate fingertip movements to broad arm swings. Secondly, many of the gestures exhibit high similarity. However, as shown in Table 2.2, our model has demonstrated impressive performance on this dataset despite these challenges. In terms of RGB-D gesture recognition, our MFST-Large model achieves more than 2% accuracy improvement compared to the SOTA methods using RGB and depth modalities. This success can be attributed to utilizing hierarchical and consolidated features learned through multi-stage factorized spatio-temporal representation coupled with multi-scale feature modeling, which can effectively capture subtle dynamic differences between similar gestures.

2.4.3 Ablation Study and Analysis

In the ablation study, we utilized the NTU RGB-D dataset and conducted all experiments using the Cross-Subject (CS) protocol. For additional ablation studies, we refer the readers to our supplementary material.

Table 2.2: Comparison of the SOTA methods on IsoGD.

Methods	Modality	Accuracy(%)
3DDSN[50]	RGB	46.08
AttebtionLSTM[51]	RGB	57.42
NAS[28]	RGB	58.88
De+Recouple[27]	RGB	60.87
Ours(MFST-Large)	RGB	**61.26**
AttentionLSTM[51]	DEP	54.18
3DDSN[50]	DEP	54.95
NAS[28]	DEP	55.68
De+Recouple[27]	DEP	60.17
Ours(MFST-Large)	DEP	**61.29**
AttentionLSTM[51]	RGB+DEP	61.05
NAS[28]	RGB+DEP	62.47
De+Recouple[27]	RGB+DEP	66.79
Ours(MFST-Large)	RGB+DEP	**68.47**

Table 2.3: Ablation study on multi-stage factorized spatio-temporal learning. MFST-Base has three stages, and MFST-Large has four stages.

Models	Acc(%)	#params	#frames
De-Recouple[27]	90.3	38.0MB	64
MFST-Base	90.5	32.6MB	32
MFST-Large	92.1	49.7MB	32

2.4.3.1 Effectiveness of Multi-stage Factorized Spatio-Temporal learning

Table 2.3 demonstrates that our proposed MFST-Base already surpasses De+Recouple [27] model, achieving a 0.2% increase in top-1 accuracy on the RGB data while utilizing fewer parameters and input frames. Moreover, our MFST-Large model has outperformed our MFST-Base model by 1.6% and has also outperformed De-recouple [27] by 1.8%. This is because, unlike the other decoupled methods [27] [46], which disentangle spatio-temporal features globally at the whole model level. Our multi-stage structure with factorized spatio-temporal learning can effectively capture local spatio-temporal contexts from low-level edges and gradually build up to higher-order semantic primitives in videos. In addition, performance continually improves as more stages use the design, validating the MFST model as an effective modeling strategy. Additional ablation experiments have been conducted on the MFST-Large model unless mentioned.

Table 2.4: Comparison among different types of 3D convolutions in the CDC-Stem module, including vanilla 3D, 3D-CDC-ST, and 3D-CDC-T.

Stem-Setting	RGB(%)	DEP(%)	θ
Vanilla-3D	90.4	91.2	0
3D-CDC-ST	**92.1**	91.6	0.6
3D-CDC-T	89.6	**93.8**	0.6
3D-CDC-ST	91.0	91.0	0.8
3D-CDC-T	89.9	92.4	0.8

2.4.3.2 3D Central Difference Convolution based Stem

Table 2.4 presents a comparison between vanilla 3D convolution and 3D central difference convolution within the context of our proposed lightweight Stem module. We have discovered that 3D-CDC-ST performs better when applied to RGB data, while 3D-CDC-T yields better results with depth data. This is because 3D-CDC-T can help to reduce sensor noise and pre-processing artifacts in depth data by leveraging the temporal context between frames. Depth data is often noisier than RGB data, and using temporal context can help to smooth out these fluctuations [28]. On the other hand, RGB data provides information about both spatial details and temporal motion patterns. This means that spatial gradient cues and temporal dynamics are beneficial for RGB data processing, hence the higher performance with 3D-CDC-ST.

Additionally, we evaluated various values of theta to assess the impact of the 3D-CDC on performance. Notably, the results indicate that when compared to the 3D vanilla convolution ($\theta = 0$), the 3D-CDC-ST ($\theta = 0.6$) significantly enhances the performance in RGB data, while the 3D-CDC-T ($\theta = 0.6$) improves the performance for depth data. These findings suggest that the optimal trade-off between the vanilla 3D and 3D-CDC is achieved at $\theta = 0.6$.

2.4.3.3 Effect of Residual Connection

In order to thoroughly investigate the effectiveness of residual connections within our proposed MFST-Base and MFST-Large model across various stages, we conducted experiments that included or excluded this component on RGB data. As shown in Table 2.5, the model with residual connections outperforms the model without, providing evidence of the effectiveness of applying residual connections among different factorized spatio-temporal stages. These connections allow for the direct flow of infor-

Table 2.5: Recognition accuracy comparison of our proposed MFST-Base and MFST-Large model with and without residual connection.

Res. Connect	MFST-Base	MFST-Large
×	90.2	91.1
✓	90.5	92.1

Table 2.6: Recognition accuracy comparison between vanilla Transformer and weight-shared Transformer.

Trans Type	RGB(%)	DEP(%)
V-Trans	91.0	92.2
WMS-Trans	92.1	93.8

mation from one stage to the subsequent stage, enabling the network to reuse earlier spatio-temporal features smoothly.

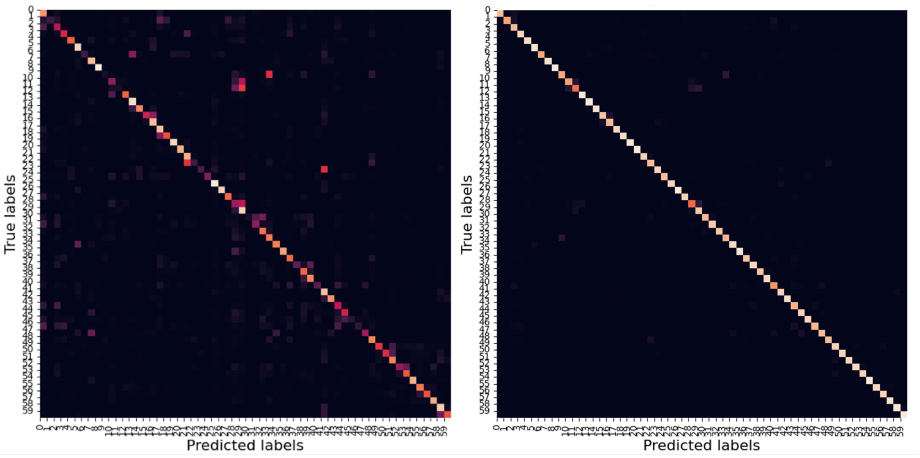
2.4.3.4 Effect of Weight-shared Multi-Scale Transformer

We conduct an ablation study to evaluate the effectiveness of the proposed weight-shared multi-scale Transformer implemented in the WMS-Trans block. Our results are presented in Table 2.6. We compared the performance of the vanilla Transformer (V-Trans) with that of the WMS-Trans, using the same layer number, head number, and dimension. Our results indicate that the WMS-Trans achieved superior performance compared to the V-Trans, with improvements of 1.1% on RGB data and 1.6% on depth data. This validates the efficiency of the WMS-Trans design, which enables it to capture more complex patterns and dependencies and learn richer feature representations.

2.4.3.5 Qualitative Analysis

The confusion matrices for the RGB modality and CS protocol on the NTU RGB-D dataset are shown in Figure 2.6 for the MFST baseline model and MFST-Large models. Figure 2.6 (a) reveals some misclassifications, such as "#10 clapping" being confused with "#34 rub two hands," "#25 reach into pocket" being confused with "#42 staggering," and "#30 type on a keyboard" being confused with "#12 writing." However, the confusion matrix of the MFST-Large model shows significant improvements in these subtle actions. This indicates the effectiveness of our proposed strategies for categorizing subtle actions through the learned hierarchical local

fine-grained and global coarse-grained spatio-temporal features.



(a) MFST-baseline (b)MFST-Large

Figure 2.6: Confusion matrices for RGB modality and CS protocol on the NTU RGB-D dataset of MFST-BM and MFST-Large models.

2.5 Conclusion

We propose a multi-stage factorized spatio-temporal learning model for RGB-D action and gesture recognition that aims to individually capture spatial and temporal features in each stage while also modeling their dependencies without additional recoupling operations. Additionally, we introduce a lightweight CDC Stem that can efficiently capture coarse- and fine-grained temporal patterns. At each factorized spatio-temporal learning stage, our CNN-Transformer hybrid block and weight-shared Transformer block effectively capture multi-scale spatial and temporal features. As a result, our model achieves new state-of-the-art performance on both RGB-D action and gesture recognition benchmarks, including NTU RGB-D and IsoGD. These results demonstrate the effectiveness of our proposed multi-stage factorized spatio-temporal representation. In the future, we aim to leverage the multi-modality fusion method to represent more comprehensive spatio-temporal information from different modalities in an end-to-end manner.

References

[1] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera. Rgb-d-based human motion recognition with deep learning: A survey. *Computer*

Vision and Image Understanding, 171:118–139, 2018.

[2] Tianyu Liu, Yujun Ma, Wenhan Yang, Wanting Ji, Ruili Wang, and Ping Jiang. Spatial-temporal interaction learning based two-stream network for action recognition. *Information Sciences*, 606:864–876, 2022.

[3] Qiguang Miao, Yunan Li, Wanli Ouyang, Zhenxin Ma, Xin Xu, Weikang Shi, and Xiaochun Cao. Multimodal gesture recognition based on the resc3d network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3047–3055, 2017.

[4] Zhimin Gao, Peitao Wang, Pei Lv, Xiaoheng Jiang, Qidong Liu, Pichao Wang, Mingliang Xu, and Wanqing Li. Focal and global spatial-temporal transformer for skeleton-based action recognition. In *Proceedings of the Asian Conference on Computer Vision*, pages 382–398, 2022.

[5] Benjia Zhou, Jun Wan, Yanyan Liang, and Guodong Guo. Adaptive cross-fusion learning for multi-modal gesture recognition. *Virtual Reality & Intelligent Hardware*, 3(3):235–247, 2021.

[6] Lei Wang, Xiaoguang Yuan, Ming Zong, Yujun Ma, Wanting Ji, Mingzhe Liu, and Ruili Wang. Multi-cue based four-stream 3d resnets for video-based action recognition. *Information Sciences*, 575:654–665, 2021.

[7] Huogen Wang, Pichao Wang, Zhanjie Song, and Wanqing Li. Large-scale multimodal gesture recognition using heterogeneous networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3129–3137, 2017.

[8] Yunan Li, Qiguang Miao, Kuan Tian, Yingying Fan, Xin Xu, Rui Li, and Jianfeng Song. Large-scale gesture recognition with a fusion of rgb-d data based on saliency theory and c3d model. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2956–2964, 2017.

[9] Benjia Zhou, Yunan Li, and Jun Wan. Regional attention with architecture-rebuilt 3d network for rgb-d gesture recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3563–3571, 2021.

[10] Jun Cheng, Ziliang Ren, Qieshi Zhang, Xiangyang Gao, and Fusheng Hao. Cross-modality compensation convolutional neural networks for rgb-d action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1498–1509, 2021.

[11] Zhen Liu, Jun Cheng, Libo Liu, Ziliang Ren, Qieshi Zhang, and Chengqun Song. Dual-stream cross-modality fusion transformer for rgb-d action recog-

nition. *Knowledge-Based Systems*, 255:109741, 2022.

[12] Liang Zhang, Guangming Zhu, Peiyi Shen, Juan Song, Syed Afaq Shah, and Mohammed Bennamoun. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3120–3128, 2017.

[13] Pradyumna Narayana, Ross Beveridge, and Bruce A Draper. Gesture recognition: Focus on the hands. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5235–5244, 2018.

[14] Xiangyu Li, Yonghong Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li. Trear: Transformer-based rgb-d egocentric action recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):246–252, 2021.

[15] Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. Star-transformer: A spatio-temporal cross attention transformer for human action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3330–3339, 2023.

[16] Zhen Liu, Qin Cheng, Chengqun Song, and Jun Cheng. Cross-scale cascade transformer for multimodal human action recognition. *Pattern Recognition Letters*, 168:17–23, 2023.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[18] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[19] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208:103219, 2021.

[20] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14063–14073, 2022.

[21] Yunan Li, Qiguang Miao, Kuan Tian, Yingying Fan, Xin Xu, Rui Li, and Jianfeng Song. Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model. In *2016 23rd International Conference on Pattern Recognition*

(ICPR), pages 25–30. IEEE, 2016.

[22] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[23] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3163–3172, 2021.

[24] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13577–13587, 2021.

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

[26] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[27] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, Fan Wang, Du Zhang, Zhen Lei, Hao Li, and Rong Jin. Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20154–20163, 2022.

[28] Zitong Yu, Benjia Zhou, Jun Wan, Pichao Wang, Haoyu Chen, Xin Liu, Stan Z Li, and Guoying Zhao. Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. *IEEE Transactions on Image Processing*, 30:5626–5640, 2021.

[29] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin. Kvt: k-nn attention for boosting vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 285–302. Springer, 2022.

[30] Liang Zhang, Guangming Zhu, Lin Mei, Peiyi Shen, Syed Afaq Ali Shah, and Mohammed Bennamoun. Attention in convolutional lstm for gesture recognition. *Advances in Neural Information Processing Systems*, 31, 2018.

[31] Pichao Wang, Wanqing Li, Zhimin Gao, Yuyao Zhang, Chang Tang, and Philip Ogunbona. Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 595–604, 2017.

[32] Pichao Wang, Wanqing Li, Jun Wan, Philip Ogunbona, and Xinwang Liu. Cooperative training of deep aggregation networks for rgb-d action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[33] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[34] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Decoupled spatial-temporal transformer for video inpainting. *arXiv preprint arXiv:2104.06637*, 2021.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[36] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[37] Tingzhao Yu, Lingfeng Wang, Cheng Da, Huxiang Gu, Shiming Xiang, and Chunhong Pan. Weakly semantic guided action recognition. *IEEE Transactions on Multimedia*, 21(10):2504–2517, 2019.

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

[39] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[40] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016.

[41] Kailin Xu, Fanfan Ye, Qiaoyong Zhong, and Di Xie. Topology-aware convolutional neural network for efficient skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2866–2874, 2022.

[42] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing

Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 183–192, 2020.

[43] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 536–553. Springer, 2020.

[44] Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1113–1122, 2021.

[45] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *proceedings of the 28th ACM International Conference on Multimedia*, pages 1625–1633, 2020.

[46] Yuhan Zhang, Bo Wu, Wen Li, Lixin Duan, and Chuang Gan. Stst: Spatial-temporal specialized transformer for skeleton-based action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3229–3237, 2021.

[47] Jian-Fang Hu, Wei-Shi Zheng, Jiahui Pan, Jianhuang Lai, and Jianguo Zhang. Deep bilinear learning for rgb-d action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351, 2018.

[48] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14204–14213, 2021.

[49] Jun Wan, Yibing Zhao, Shuai Zhou, Isabelle Guyon, Sergio Escalera, and Stan Z Li. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–64, 2016.

[50] Jiali Duan, Jun Wan, Shuai Zhou, Xiaoyuan Guo, and Stan Z Li. A unified framework for multi-modal isolated gesture recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1s):1–16, 2018.

[51] Guangming Zhu, Liang Zhang, Lu Yang, Lin Mei, Syed Afaq Ali Shah, Mo-

hammed Bennamoun, and Peiyi Shen. Redundancy and attention in convolutional lstm for gesture recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 31(4):1323–1335, 2019.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Yujun Ma
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 2

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Yujun Ma, Benjia Zhou, Ruili Wang, and Pichao Wang. "Multi-stage factorized spatio-temporal representation for rgb-d action and gesture recognition." In Proceedings of the 31st ACM International Conference on Multimedia, pp. 3149-3160. 2023.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Yujun Ma Digitally signed by Yujun Ma DN: cn=Yujun Ma, c=NZ, o=Massey University, ou=School of Mathematical and Computational Science, email=yma1@massey.ac.nz Date: 2024.02.22 14:22:46 +13'00'
Date:	20-Feb-2024
Primary Supervisor's Signature:	Prof Ruili Wang Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, c=NZ, email=ruili.wang@massey.ac.nz Location: School of Natural and Computational Sciences Date: 2024.07.23 15:59:49 +12'00'
Date:	

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 3

Relative-position embedding based spatially and temporally decoupled Transformer for action recognition

*Recognition of human actions is to classify actions in a video. Recently, Vision Transformer (ViT) has been applied to action recognition. However, the Vision Transformer is unsuitable for high-resolution input videos due to the constraint of computing power since ViT splits frames into fixed-size patches embedded (*i.e.*, tokens) with absolute-position information and adopts a pure Transformer encoder to model the relationships among these tokens. To address this issue, we propose a relative-position embedding based spatially and temporally decoupled Transformer (RPE-STDT) for action recognition, which can capture spatial-temporal information by stacked self-attention layers. The proposed RPE-STDT model consists of two separate series of Transformer encoders. The first series of encoders is the spatial Transformer encoders, which model interactions between tokens extracted from the same temporal index. The second series of encoders is the temporal Transformer encoders, which model interactions across time dimensions with a subsampling strategy. Furthermore, we replace the absolute-position embeddings in the Vision Transformer encoders with the proposed relative-position embeddings to capture the order of the embedded tokens to reduce computational costs. Finally, we conduct thorough ablation studies. Our RPE-STDT achieves state-of-the-art results on multiple action recognition datasets, exceeding prior convolution and Transformer-based networks. Note that the work presented in this chapter has been published in the journal of Pattern Recognition.*

3.1 Introduction

With the popularity of surveillance and increased network bandwidth for video streaming, the recognition of human actions from videos has attracted much attention [1] [2] [3]. However, action recognition remains challenging due to the complexity of visual contents and deteriorated video frames, such as motion blur, occlusion induced by camera movement, and large object motions [4] [5].

The primary challenge in video based action recognition is to efficiently learn the most informative spatial and temporal features and compactly represent them [6]. For example, as shown in Figure 3.1, the spatial attention during actions (*e.g.,* "somersault" or "flic_flac") often focuses on the main object (highlighted blue boxes), rather than the entire frame. Similarly, within the temporal sequence, certain frames—marked by red boxes—are more crucial for recognizing actions. Therefore, action recognition approaches need to consider both spatial and temporal attention to effectively focus on the most informative regions and frames [7] [8].

Early action recognition approaches [9] [10] used hand-crafted features to encode spatial and temporal information in videos. Later, deep neural networks [11] have been widely used for action recognition. Two-stream CNNs [12] [13] [14] independently learned features from RGB video frames and optical flow frames, which were then fused [15]. Other approaches [16] [17] [18] adopted CNNs via 3D convolutions, extracting both spatial and temporal features over consecutive video frames for action recognition. However, 3D convolutions require expensive computing and large-scale training datasets.

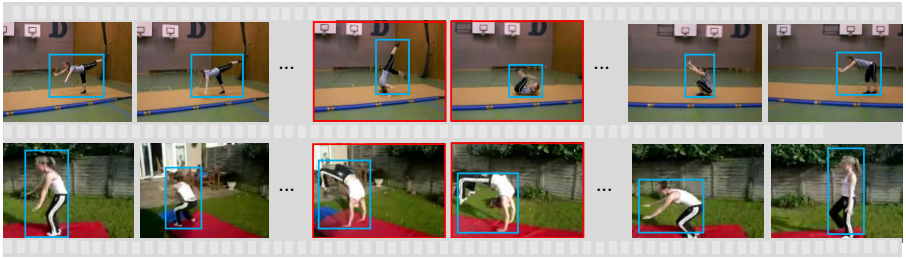


Figure 3.1: Two examples are presenting spatial and temporal attention over consecutive video frames from the "somersault" (upper row) and "flic_flac" (lower row) action classes, respectively. The blue box refers to the spatial focus in a frame, while the red box relates to the keyframes in the frame sequence.

Transformer [19] has revolutionized natural language processing (NLP) by using the

self-attention mechanism. Transformer [19] is a novel deep neural network that mainly adopted self-attention layers to extract discriminative features from long sequences. Inspired by the remarkable achievements of Transformer in NLP, Transformer has been introduced for action recognition [20] [21] [22] [23]. Video-based action recognition shares similar charterers with NLP, in which a sentence and a video are both sequences of small units [24]. Each patch in a frame is analogous to a word in a sentence. Inspired by this, Dosovitskiy *et al.* [24] proposed a Vision Transformer (ViT) for image classification, which adopted a pure Transformer encoder that directly learns from the sequences of split image patches. ViT [24] broke down an image into fix-sized non-overlapping patches and then linearly project the patches into Transformer acceptable tokens [24]. Specifically, the grids were flattened to create tokens, and each token position was mapped to a 1D position.

The essence of the Transformer is the self-attention mechanism [19], which can learn from a sequence of embedded tokens. However, self-attention has an inherent drawback: incapable of capturing the order information of input tokens [25]. Therefore, it is essential to incorporate explicit positional representations for Transformer. To solve this problem, an absolute-position embedding strategy [19] was utilized in Transformer to present the positional information of the tokens. Similar to calculating word embeddings in NLP, the Vision Transformer [24] [21] adopted the absolute-position embeddings to notice Transformer that the top left patch is the 1^{st} token and the bottom right patch is the last token for an image.

Recently, approaches leveraged ViT as a foundational architecture for action recognition, integrating absolute-position embedding strategy at the patch embedding stage [21]. For example, Chen *et al.* [26] introduced a multi-modal video Transformer for action recognition, incorporating self-attention across space, time, and modality dimensions. Their approach employed spatio-temporal absolute position encoding by adding it to each patch token, aiming to preserve the positional information within the video frames. However, it is important to note that absolute position embedding may lack explicit representation of spatial relationships between tokens. This limitation can result in neglecting the inherent spatial structure within a video frame and restrict the model's ability to capture fine-grained spatial dependencies and contextual information. Ahn *et al.* [27] utilized absolute position embedding on the joint mapping of multi-class tokens created by aggregating spatio-temporal skeleton features and RGB video features. However, absolute position embedding assigned fixed

positional embeddings to each token, which might not provide enough flexibility to capture the semantic differences and complexities in the multi-class tokens.

In general, the absolute-position embedding strategy is unsuitable for high-resolution input successive video frames [21], which involves a substantial number of tokens. Motivated by these observations, we propose a relative-position embedding (RPE) based patch tokenization for the video Transformer, which is suitable for extremely long sequences while capturing nuanced interactions and dependencies between different regions of the frames by considering the relative distances and orientations among tokens. Concretely, we extend the self-attention mechanism to consider representations of the directional relative positions or distances among patches in a video frame rather than absolute positions. The relative-position embedding strategy can foster interaction between the queries, keys, and positional embedding within the self-attention mechanism allowing the model to effectively attend to relevant context and capture spatially related information. Furthermore, different heads of multi-head self-attention can share the same relative-position embeddings for parameter saving.

On the other hand, videos inherently contain both spatial and temporal information [13]. Consequently, many approaches [7] [28] have focused on the space-time entanglement model, aiming to capture the intricate dependencies between spatial and temporal dimensions. For example, Yang *et al.* [29] proposed a recurring Transformer for action recognition, which captured spatio-temporal features by a recurrent execution. However, such entangled spatio-temporal modeling could pose optimization challenges due to the highly interdependent and intertwined nature of spatio-temporal information. Zhang *et al.* [30] proposed decoupling the 3D self-attention into temporal attention and spatial attention for action recognition. However, processing temporal information before spatial information could potentially lead to a lack of spatial context for interpreting the action.

Building upon these insights, we propose a spatially and temporally decoupled Transformer for action recognition, consisting of two series: spatial Transformer encoders and temporal Transformer encoders. Concretely, the spatial Transformer encoders compute self-attention between spatial tokens extracted from the same temporal indices to capture comprehensive spatial information. Then the temporal Transformer encoders calculate self-attention between tokens from the different temporal indexes to extract discriminate temporal information. Additionally, we apply a temporal

subsampling layer in the temporal Transformer encoders to pick out the frames with strong localized attention and filter out non-informative frames.

The rest of this paper is organized as follows. Section 3.2 reviews the related work. Section 3.3 presents the proposed RPE-STDT model. Experimental results and analyses are provided in Section 3.4. Section 3.5 presents the conclusion, and Section 3.6 reviews the limitations.

3.2 Background and related work

Action recognition has attracted intensive research interests recently. This section briefly reviews action recognition models, Transformer-based models, and positional representation methods.

3.2.1 Action recognition

With the emergence of deep learning, approaches for action recognition utilized 2D convolutions to capture spatial and temporal features in videos [13]. The breakthrough of AlexNet [12] on ImageNet initially led to the success of two-stream convolutional networks (two-stream CNN) [13] for action recognition. These approaches [14][31] applied 2D convolutions on stacked RGB video frames and optical flow frames to capture spatial and temporal features in videos. However, large-scale action recognition datasets [32] subsequently facilitated the training of spatial-temporal 3D CNNs [33], which required significant computation and enormous training data.

As 3D convolutional models required more calculation than 2D convolutional models, many approaches decomposed 3D convolutions over spatial and temporal dimensions or used separable convolutions. For example, Feichtenhofer *et al.* [34] proposed an X3D network for action recognition, which adopted a feature selection method to adjust the parameters, such as input resolution, temporal dimensions, network depth, and width. Tran *et al.* [35] proposed an R3D network for action recognition, which decomposed the 3D convolutions into two independent convolutions to extract spatial and temporal features in videos. Yang *et al.* [36] proposed an asymmetric 3D CNNs model for action recognition, which adopted cascaded one-directional asymmetric 3D convolutions to approximate the traditional 3D convolution. The model consisted of multiple sub-asymmetric 3D convolutional MicroNets to integrate multi-scale spatial-

temporal features in videos. Tu *et al.* [37] proposed a human-related region-enhanced multi-stream model for action recognition, consisting of a spatial, a temporal, and a human-related stream. The proposed model could focus on a secondary region containing an actor's major moving part based on motion saliency to improve action recognition. Wang *et al.* [33] proposed a multi-cue-based 3D residual model for action recognition, which fused an appearance cue, a direct motion cue, and a salient motion cue straight from the input frames instead of a single cue.

Recently, some approaches applied attention mechanisms to action recognition. Li *et al.* [38] proposed a unified spatio-temporal attention network for action recognition, which utilized an attention neural cell to generate the attention weights on both spatial and temporal dimensions. Kalfaoglu *et al.* [39] integrated a hybrid of 3D convolutions and a Transformer model for action recognition, which applied a Transformer to act on the output feature vectors of 3D convolutions.

Li *et al.* [7] developed a spatio-temporal attention model for action recognition, which could characterize valuable information at both the spatial and channel levels. 3D CNN-based models had demonstrated effectiveness in action recognition, but achieving higher efficiency often necessitates larger kernels or deeper structures for improved performance. Li *et al.* [40] proposed an attention-based spatio-temporal deformable 3D CNNs model for action recognition, which adopted the attention mechanism to capture long-range dependencies over the temporal dimensions and long-distance grid dependencies in the spatial dimension. Wang *et al.* [41] proposed a symbiotic attention-based multi-branch model for egocentric action recognition, which identified the most action-relevant candidates for classification by promoting mutual interaction between the verb and noun branches. However, unlike egocentric action recognition, where the focus is primarily on the actions performed by the person, coarse-grained and fine-grained action recognition sometimes benefit from low-level features such as backgrounds and contextual information.

A recent approach [42] indicated that the depth-wise convolution significantly lowered the computation cost for action recognition, but the depth-wise convolution also increased the model inference latency. However, Transformer [19] did not rely on heavily piled convolutional layers to learn features but efficiently extracted features globally by self-attention. Motivated by these findings, we decouple spatial and temporal indices of videos to improve efficiency and leverage the Transformer independently on

both the spatial and temporal dimensions within the context of a pure-Transformer structure.

3.2.2 Vision Transformers

Transformer has become the dominant network in natural language processing (NLP) and its performance exceeded previous language translation, and text generation approaches. Transformer was first proposed by Vaswani *et al.* [19], which consisted of stacked encoders and decoders. The encoder and decoder consisted of a multi-head self-attention layer and a feed-forward layer. The residual skip connections were adopted at each layer, followed by layer normalization.

The Vision Transformer (ViT) was first proposed by Dosovitskiy *et al.* [24], which applied a pure-Transformer encoder directly to the sequences of image patches for image classification. Then, following the paradigm of ViT, Arnab *et al.* [21] proposed a Video Vision Transformer (ViviT) for action recognition, which extracted spatial-temporal tokens from the input video, and then encoded them by a series of Transformer encoder blocks.

Recently, Transformer-based approaches have achieved promising results in action recognition. Liu *et al.* [43] proposed a Video Swin Transformer for action recognition, which applied a non-overlapping 3D shifted window in a vanilla Transformer block to extract both spatial and temporal features in videos. Zhang *et al.* [30] proposed a separable attention-based video Transformer for action recognition, which expanded the ViT Transformer model for 3D feature learning. The proposed spatial-temporal separable attention video Transformer aggregated spatial-temporal features from raw videos and captured 3D motion features on a sequence of local patches.

Later, Bertasius *et al.* [20] proposed a divided space-time attention model for action recognition, which applied temporal and spatial attention separately in each Transformer block. The model first computed temporal attention by comparing each embedded token with all the tokens in the same spatial position in other frames. Temporal attention was then fed into spatial attention instead of the multi-layer perceptron (MLP). Zhang *et al.* [44] applied a novel zero-parameter, zero-FLOPs token shift operator at the Transformer encoder for action recognition. The token shift operator only processed global frame representation, which could decrease computations of pair-wise attention in managing a sequence of flattened patches.

Li *et al.* [45] proposed a motion guided video Transformer for action recognition, which adopted a motion self-attention to highlight motion information patches. The proposed motion self-attention globally concatenated the query token and the neighborhood tokens in other frames over the temporal index when modelling the token's temporal dependencies. Yan *et al.* [22] proposed a multiview Transformer for action recognition, which consisted of individual parallel encoders to learn different views of a video clip. A linearized cross-view attention layer fused the extracted features from different views. Neimark *et al.* [28] proposed a video Transformer network for video classification, which consisted of a 2D spatial feature extractor and a temporal attention-based Transformer encoder.

Different from a recent approach [28] that extracted spatial features with a Transformer encoder on every single frame and then aggregated features with self-attention, and other entangled spatio-temporal models, our RPE-STDT model spatio-temporal independent representation from raw videos. This is achieved through the utilization of a spatially and temporally decoupled Transformer without demanding larger video resolution or extra long clip length, which also differs from very recent work [24].

3.2.3 Positional representations in Transformer

Initially, self-attention mechanisms had a limitation in capturing the sequential order of input tokens [19]. Therefore, incorporating precise positional representations is essential for Transformer. Vaswani *et al.* [19] proposed an absolute-position embedding strategy with dimensions to match the token embeddings in Transformer. The absolute-position embeddings provided order information to the sequence of tokens, which encoded the tokens from *1* to the maximum sequence length. In addition, each token had an individual position embedding obtained by sine and cosine functions at different frequencies.

Besides the absolute-position embeddings, recent approaches developed the relative-position representations to consider the pairwise relationships between tokens. Shaw *et al.* [25] proposed a relative-position embedding strategy for self-attention, which considered arbitrary relations between two tokens. Huang *et al.* [46] proposed a relative-position embedding strategy to simultaneously boost interaction between query and key in self-attention. Ramachandran *et al.* [47] introduced a positional representation for 2D images, which decomposed the 2D relative embedding horizon-

tally and vertically. However, these positional embedding strategies were only applied to keys.

Absolute-position embedding strategy was proposed for model 1D word sequences. However, unlike 1D sentences, the pixels are highly spatially structured in the 2D video frame, forming a grid-like arrangement [25]. Therefore, directly applying the same absolute-position embedding strategy used for 1D sequences may not be optimal for capturing the spatial relationships between pixels in a video frame. Additionally, the absolute-position embedding strategy is unsuitable for high-resolution input videos involving a substantial token number, such as video-based action recognition. On the other hand, in the context of video frames, the relative position of tokens holds significant importance as it governs the relative order and distance between them. Therefore, we introduce a relative-position embedding method to the video Transformer, which considers the relationship between two tokens based on Euclidean distance and quantifies the distance by an index function.

3.3 Proposed RPE-STDT model

This section details the proposed relative-position embedding based spatially and temporally decoupled Transformer (RPE-STDT). The proposed RPE-STDT model is based on pure-Transformer architecture. As shown in Figure 3.2, Transformer is decoupled into two series: spatial Transformer encoders and temporal Transformer encoders. Firstly, spatial Transformer encoders compute the attention between tokens in the same video frame to generate a hidden representation over a temporal index. Relative-position embedding is utilized for patch tokenization, capturing the relative positions of tokens within the frame. Secondly, the obtained representations from the spatial Transformer encoders are passed into the temporal Transformer encoders. These encoders compute attention over sequences of frames, capturing temporal dependencies and interactions between frames. In addition, we apply a temporal subsampling layer in temporal Transformer encoders to filter out non-informative frames while reducing the temporal dimension. Finally, a learnable classification token is applied to gather information from all the output tokens, which is the final representation to serve the multi-layer perceptron (MLP) classification head.

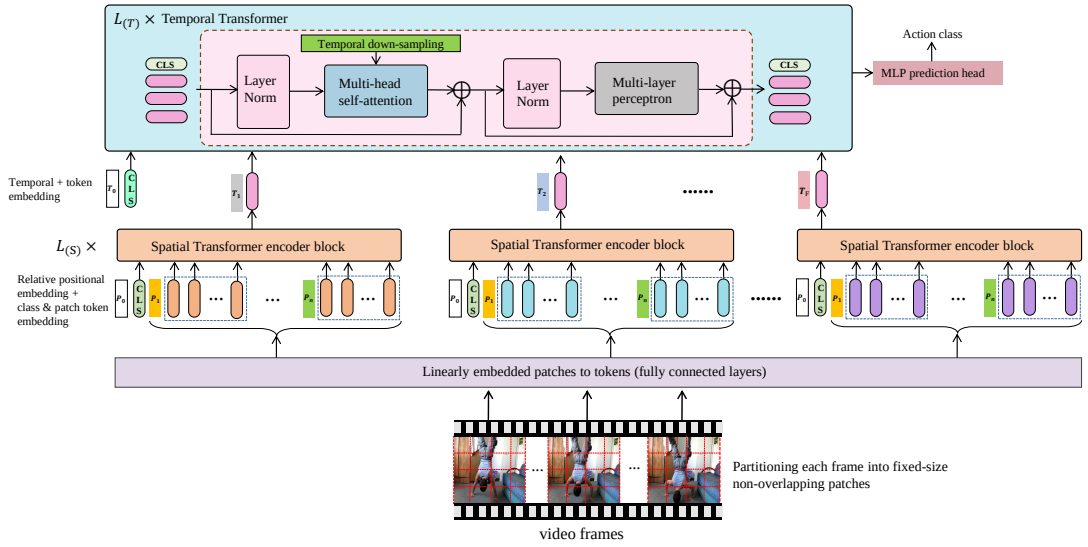


Figure 3.2: The proposed model comprises two components: spatial Transformer encoders and temporal Transformer encoders. The spatial Transformer encoders initially compute attention between tokens within the same video frame, generating a hidden representation over a temporal index. These representations are subsequently inputted into the temporal Transformer encoders, which compute attention over sequences of frames. This two-step decoupled process enables the model to effectively capture spatial and temporal dependencies, resulting in a more holistic and informative representation of videos.

3.3.1 Preliminaries: ViT and ViViT

We represent the input clip as $V \in \mathbb{R}^{T\times H\times W\times C}$, where T refers to the clip length; W and H denote width and height of an RGB video frame, and C indicates the number of channels. Since Transformer cannot directly process images, ViT [24] partitions an image into non-overlapping patches and flattens, and then patches are linearly embedded into tokens, which are the acceptable modality for Transformer. ViViT [21] extends this to video by decomposing each frame into numerous non-overlapping patches. Concretely, $n_H \times n_W$ non-overlapping patches are produced from each frame, then a total number of $n_H \times n_W \times n_F$ embedded tokens from F RGB frames are fed into Transformer. Furthermore, a class token $Z_{cls} \in \mathbb{R}^d$ is prepended with the input embeddings, which can accumulate information from the other tokens in the input sequence as a placeholder cell. When the vision Transformer eventually performs the final classification of the sequence, it utilizes a multi-layer perceptron (MLP) head that only uses information from the last layer's class token.

Transformer was originally developed for natural language processing [19], where the positional information of each word in the input sequence is encoded. This is accomplished by adding absolute-position embeddings to each word, which indicates the order of the words in the sequence. Likewise, ViviT [21] also adds an absolute-position embedding to each patch since the permutation invariant of Transformer, which helps Transformer determine that the top left patch is the 1^{st} token and the bottom right patch is the last token for one frame. In general, this tokenization process can be represented as:

$$Z = [Z_{cls}, Ex_1, Ex_2, ...Ex_n] + P, \tag{3.1}$$

where x_i denotes each patch; the patches are then projected to tokens by E, a linear operator equal to a 2D convolution. Also, the $P \in \mathbb{R}^{(N+1)\times d}$ refers to absolute-position embeddings. Concretely, sinusoids with varying frequencies obtain the absolute-position embeddings [19]. Therefore, each dimension of the positional embedding corresponds to a sinusoid.

3.3.2 Absolute-position embeddings

In vision Transformer, each multi-head attention layer operates on a sequence of embedded patches $x = (x_1, x_2, ..., x_n)$, including n tokens, where $x_n \in \mathbb{R}^{d_x}$. The tokenization generates a new sequence $z = (z_1, z_2, ..., z_n)$, where $z_n \in \mathbb{R}^{d_z}$. Each output element z_i is computed as a weighted sum of input elements:

$$z_i = \sum_{j=1}^{n} a_{ij} \left(x_j W^V\right), \tag{3.2}$$

where the weight coefficient a_{ij} is calculated by a softmax function:

$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{n} exp(e_{ik})}, \tag{3.3}$$

where e_{ij} is the attention weight from position j to i; a scaled dot production is applied to compare two tokens after a linear projection:

$$e_{ij} = \frac{\left(x_i W^Q\right)\left(x_j W^K\right)^\top}{\sqrt{d_z}}. \tag{3.4}$$

Linear projections of the inputs bring sufficient expressive power. The scaling factor $\sqrt{d_z}$ is essential to keep the training process stable. W^Q, W^K, $W^V \in \mathbb{R}^{d_x \times d_z}$ are learnable parameter matrices.

3.3.3 Relative-position embeddings

Absolute-position embedding [19] is unsuitable for high-resolution and multi-frame input videos, where the token number might be enormous. Meanwhile, the relationship between neighboring pixels plays a vital role in action recognition, primarily due to the spatial structure exhibited by pixels within a frame. Therefore, we propose a relative-position embedding strategy to replace the absolute-position embedding strategy in the vision Transformer [21].

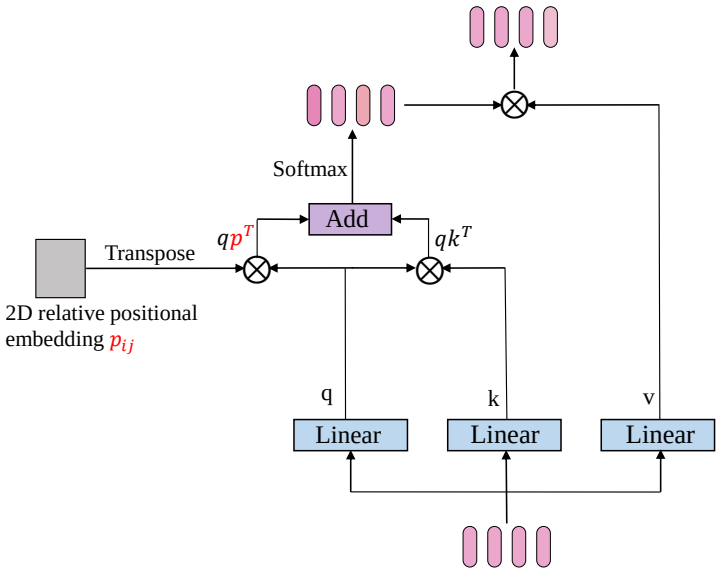


Figure 3.3: Self-attention with relative-position representation on the query, where the p_{ij} denotes the newly added relative-position embedding tensor.

The relative-position embedding strategy maps the relative distance between the input token x_i and x_j into vectors $R^Q_{ij}, R^K_{ij}, R^V_{ij} \in \mathbb{R}^{d_R}$, where $d_R = d_z$. These relative-position embeddings can be shared across self-attention heads. We modify Equation 3.1 and Equation 3.3 to encode the vectors into the self-attention:

$$z_i = \sum_{j=1}^{n} a_{ij} \left(x_j W^V + R^V_{ij} \right), \tag{3.5}$$

$$e_{ij} = \frac{\left(x_i W^Q + R_{ij}^Q\right)\left(x_j W^K + R_{ij}^K\right)^T}{\sqrt{d_z}}. \tag{3.6}$$

In such manner, the pairwise position relationship is captured during the training of Transformer. Figure 3.3 shows the architecture of self-attention modules with relative-position embeddings on queries and keys in the self-attention module, where p_{ij} refers to the relative-position representation. We consider the interaction of positional embeddings with the query, and then the self-attention with the scaled dotted product is computed as:

$$e_{ij} = \frac{\left(x_i W^Q\right)\left(x_j W^K\right)^T + \left(x_i W^Q\right) p_{ij}^T}{\sqrt{d_z}}, \tag{3.7}$$

where $p_{ij} \in \mathbb{R}^{d_z}$ denotes the relative-position weights and is interacted with the query embedding.

Moreover, instead of the sine and cosine function used in absolute-position embedding [19], we present a many-to-one function as the mapping function for the relative-position embedding strategy. The mapping function can project a relative distance between two patches to an integer set. Therefore, the p_{ij} can be indexed by the integer, allowing the different patches in different positions to share positional embeddings. The mapping function can lower the computational cost and decrease the number of tokens for long input sequences (*i.e.*, high-resolution multi-frames).

In [25], a clip function is introduced as $clip(x, k) = max(-k,\ min\ (k, x))$, where x represents the position coordinates of the patches. This clip function enables the patches with a relative distance longer than k to share the same positional embedding. However, this clip function inevitably ignores some contextual information in the long-range patches. Inspired by [48], we acknowledge that spatially neighboring pixels in a frame usually exhibit high correlation, which plays an essential role in recognizing actions. Therefore, the adjacent patches are better assigned together to ensure that the part of spatial information is intact. Based on this intuition, we introduce a sectioned function as the mapping function to project a relative distance between two patches into a corresponding embedding and distribute the attention by the relative distance, which is indicated as:

$$s(x) = [x], |x| \leqslant b, \tag{3.8}$$

$$s(x) = \pm min\left(y, \left[b + \frac{In\left(\frac{|x|}{b}\right)}{In\left(\frac{a}{b}\right)}(y-b)\right]\right), |x| > b, \tag{3.9}$$

where $[x]$ defers the rounding function; $\pm$ indicates the sign of x; b refers to the partition point; y determines the output range [-y,y], and a alters the curvature of the logarithmic function.

Euclidean distances are utilized to measure the relative distance between the patches, and then we project the distances into the corresponding embeddings by the above-introduced mapping function. The calculation is defined as:

$$D(i,j) = s\left(\sqrt{(\tilde{x}_i - \tilde{x}_j)^2 + (\tilde{y}_i - \tilde{y}_j)^2}\right). \tag{3.10}$$

The relative-position embeddings are formulated as:

$$p_{ij} = B_{D^{\tilde{x}}(i,j), D^{\tilde{y}}(i,j)}. \tag{3.11}$$

We consider the trainable vector B as a cluster, which holds the relative-positional weight. From an abstract point of view, a cluster consists of multiple patches that share the same relative-position embeddings. The number of clusters is $(2y+1)^2$, which is defined in Equation 3.9.

3.3.4 Spatially and temporally decoupled Transformer

We decouple Transformer into spatial and temporal encoders modularly, which are constructed dimension-independently. Additionally, the decoupled encoders naturally leverage the benefits of the self-attention mechanism compared to convolutional methods regarding more useful utilization of temporal features in videos.

As shown in Figure 3.2, Transformer is decoupled into two series: spatial Transformer encoders and temporal Transformer encoders. Firstly, spatial Transformer encoders

compute the attention between tokens in the same video frame to generate a hidden representation over a temporal index. Relative-position embedding is utilized for patch tokenization, capturing the relative positions of tokens within the frame. Secondly, the obtained representations from the spatial Transformer encoders are passed into the temporal Transformer encoders. These encoders compute attention over sequences of frames, capturing temporal dependencies and interactions between frames.

We adopt the same input embeddings and class token as presented in [24]. Each representation of the temporal index is obtained after spatial encoder layers $L_{(s)}$. The spatial representations $h_i \in \mathbb{R}^d$, are concatenated into $H \in \mathbb{R}^{n_t \times d}$, then fed into a temporal encoder consisting of $L_{(t)}$ temporal encoder layers to compute attention between the tokens from the different temporal indexes.

The decomposed spatial and temporal encoders have several benefits. Firstly, this reduces the split of the video into the short computation of sequences. Each patch is compared to N other patches within a frame during spatial encoding. The temporal encoders compare each frame representation to F others, resulting in less overall calculation than comparing each patch to $N \times F$ other patches. Secondly, the temporal features are more useful when they are modeled on a more abstract level of the entire network. The frame-level embeddings offer more explicit clues about what is happening in videos than separated patches.

3.3.5 Temporal subsampling

The information in temporal episodes may not always be informative for action recognition. Recent studies [49] [50] have shown that an action recognition model can achieve better performance by picking the informative temporal segments as input rather than taking uniformly split frames as input. To address this issue, we adopt the temporal subsampling layer in the temporal Transformer encoder to reduce the redundant temporal information, as shown in Figure 3.2. We also compare different temporal subsampling strategies, including temporal max-pooling with a kernel size of 2, temporal average-pooling, and 1D temporal convolution with stride 2.

A drawback of these subsampling strategies is that they uniformly extract features across time instances, but the informative frames are distributed non-uniformly in video clips. We are inspired by the concept of non-uniform temporal aggregation

from previous work [51]. Unlike this work that directly subsamples the query tensor using temporal average pooling, we propose a standard deviation-based Top-K pooling strategy for the attention maps, which is named after [52]. When a temporal frame is informative, the temporal attention score is highly activated on a few temporal segments. Conversely, if a temporal frame is non-informative, the attention score is more likely to be equally distributed over the whole video clip. The proposed Top-K pooling selects the *top k* highest standard deviation rows in the attention matrix, focusing on the most significant temporal features:

$$Pool\left(Attn_t^{(1:,:)}\right) = Attn_t^{\left(topk\left(\sigma\left(Attn_t^{(1:,:)}\right)\right),:\right)}, \tag{3.12}$$

where $Attn_t$ represents temporal attention obtained by performing matrix multiplication between q_t and k_t (*i.e.*, query and key features after applying independent linear layer), and $\sigma \in \mathbb{R}^T$ refers to standard deviation of $Attn_t^{(1:,:)}$, which can be calculated by:

$$\sigma^{(i)} = \frac{1}{T}\sqrt{\sum_{i=1}^{T}\left(Attn_t^{(i,:)} - \mu\right)^2}, \tag{3.13}$$

$$\mu^{(i)} = \frac{1}{T}\sum_{i=1}^{T} Attn_t^{(i,:)}. \tag{3.14}$$

The σ obtained by standard deviation function, and μ indicates the mean average of $Attn_t^{(1:,:)}$. The experiments show that Top-K pooling results in higher accuracy than temporal average pooling and convolutional pooling. The Top-K pooling strategy can be comprehended as choosing the frames with strong localized attention and filtering frames with uniform attention.

3.4 Experiments

3.4.1 Datasets

UCF-101 [53] UCF-101 is a challenging coarse-grained action recognition dataset of real videos collected from YouTube, which has significant variations in views, lights,

scenes, and resolutions. UCF-101 includes 13320 videos from 101 action classes with an average frame of 160 per video. This paper reports the top-1 classification results on the validation set based on split one of UCF-101.

HMDB-51 [54]: HMDB-51 is a more challenging coarse-grained action recognition dataset of real videos selected from various sources, including films and internet videos. The dataset contains 6,849 videos from 51 action classes (such as "drink", "punch", and "golf"), with each class including at least 100 clips. This paper presents the classification accuracy of the first split for a fair comparison with the other approaches.

Diving 48 [55]: Diving 48 is a fine-grained action recognition dataset with 18404 diving videos from 48 categories collected from worldwide competitions. In addition, diving 48 shows a standard diving action where subtle details of the diving process clarify the different classes instead of coarse moves. We adopt the traditional split, 16067 training videos and 2337 test videos. We use the updated Diving 48 dataset (marked 'v2') released in Oct 2020.

3.4.2 Implementation Details

Model instantiating. The proposed RPE-STDT model consists of two components: spatial encoders and temporal encoders. The relative-position embedding is adopted into all self-attention layers. We set $b : y : a = 1 : 2 : 8$ for the sectioned function *s(x)* and modify the number of clusters by adjusting y. An extra block is utilized to hold the relative-position embeddings of the extra class token. For training, the frames are decomposed into 14×14 non-overlapping patches.

We follow the model weights of the ViT-Base approach in our spatial Transformer encoder, which consists of 8 multi-head self-attention (MHSA) layers (*i.e.*, $L_{(S)} = 8$), each with 12 self-attention heads. For the temporal Transformer encoder, we use a lighter version of the Transformer, which contains 4 MHSA layers (*i.e.*, $L_{(T)} = 4$), each with 8 self-attention heads. The temporal encoders are initialized randomly and trained from scratch in our experiments.

The frames are sampled uniformly from raw videos with frame intervals set to 32. Then, we resize the resolution of each frame to a scale range of [256, 320] and randomly crop them to 224×224 in the same position for all frames in the same video. We

adopt a random horizontal flip with a factor of 0.5 on all frames as the augmentation strategy to address overfitting issues.

We crop the frames to 224×224 from the centre for model inference. We adopt the same frame sampling for model training and inference. We use SGD as the optimizer with a momentum of 0.9, and a cosine learning rate schedule with linear warm-up during the fine-tuning is based on [21] for 100 epochs. All experiments are all run on $4 \times$ A100 GPUS.

3.4.3 Ablation Experiments

Analysis on relative-position embeddings. Table 3.1 provides a comparison between absolute-position embeddings and relative-position embeddings, along with the comparisons between relative-position embeddings when they are shared or unshared among multiple heads. To study different positional embeddings, we utilize the original absolute-position embeddings [21] for comparison. The results demonstrate that relative-position embeddings outperform absolute-position embeddings across all three datasets, including coarse-grained and fine-grained datasets. This can be attributed to the fact that relative-position embeddings effectively capture subtle interactions and dependencies among different frame regions while maintaining the contextual relationship with neighboring pixels, thereby preserving the high possibility of context relevance.

Regarding relative-position embeddings, each head can compute its relative-position embedding (RPE) using Equation 3.7. Notably, the accuracy difference between unshared and shared RPE is negligible across all three datasets. As a result, we opt for the shared strategy in our final model to achieve parameter-saving benefits.

Comparison of the clip function and sectioned function. We compare the effectiveness of the sectioned function *s(x)* represented in Equation 3.9 and the clip function on the HMDB-51 dataset. Table 3.2 shows that the sectioned function is better than the clip function for action recognition. The reason is that the sectioned function is more practical when the input sequence is long. Multiple video frames lead to a more extended input sequence. Therefore, we presume that the sectioned function can distribute different attention scores to the positions with a relatively large distance.

Number of clusters. The number of clusters primarily affects the computational cost, total parameters and classification accuracy. Therefore, we conduct ablation experiments to explore the performance of different cluster numbers on three datasets. Figure 3.4 reports the changing Top-1 classification results and the number of clusters per frame. We can observe that the overall accuracy increased along with the number of clusters. However, there is no noticeable improvement in accuracy after 50 clusters. Therefore, we consider that 50 clusters are the best choice for a trade-off between accuracy and computational cost.

Analysis of attention schemes. Table 3.3 evaluates different attention settings: the spatial attention-only Transformer encoder (*i.e.*, the ViT-B backbone) and the spatially and temporally decoupled Transformer encoder (*i.e.*, the ViT-B backbone and the temporal encoder). Table 3.3 indicates that the spatially and temporally decoupled encoder provides a gain of 3.7%, 2.1%, and 3.3% over the spatial attention-only model on the HMDB-51, UCF-101, and Diving 48 datasets with fewer FLOPs (Floating-Point Operations), respectively. The results on the UCF-101, HMDB-51 and Diving 48 datasets demonstrate temporal attention's positive impact, proving that temporal modeling is also critical for the Transformer-based model for action recognition.

Analysis of different subsampling strategies. We evaluate four subsampling strategies in the temporal encoder: temporal max-pooling with a kernel size of 2, temporal average-pooling, 1D temporal convolution with stride 2, and the proposed Top-K subsampling. Table 3.4 indicates that the proposed Top-K subsampling exceeded other subsampling strategies. The reason is that Top-K sampling can non-uniformly aggregate temporal information in videos. We also utilized temporal subsampling at different encoder layers to explore the best position. As a result, the temporal dimension is reduced by half. Table 4 indicates that performing temporal subsampling after the first and third temporal encoder layers performs best. Adopting subsampling at the beginning of the temporal encoder leads to a computational saving but has dropped performance among the three datasets.

To further evaluate the effectiveness of Top-K subsampling, we conducted a comparison between our baseline STDT model (without RPE) and a baseline-STDT model that integrated Top-K subsampling on the UCF-101 dataset. The results presented in Table 5 demonstrate that the Top-K subsampling approach outperforms the baseline

model while utilizing fewer parameters. This can be attributed to reducing duplicated or non-informative temporal instances. The difference between our model and the previous work [30] is the application of Top-K subsampling on the generated spatial features obtained from the spatial Transformer rather than initially focusing on temporal dynamics.

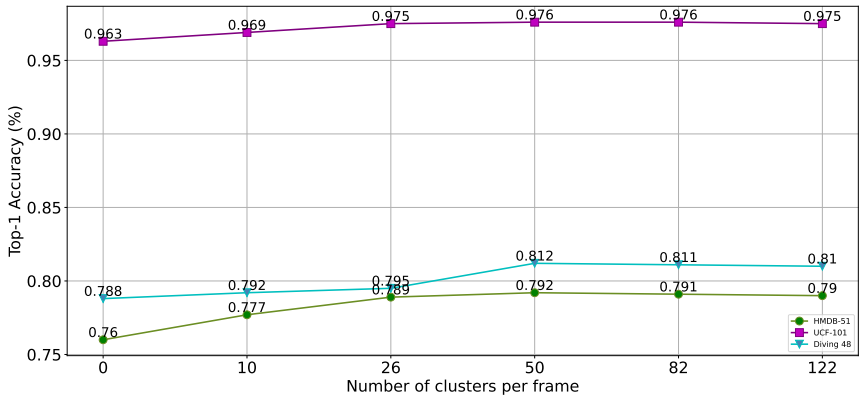


Figure 3.4: Comparison of the different numbers of clusters per frame in the proposed RPE-STDT model with shared relative-position embeddings over UCF-101, HMDB-51 and Diving 48.

3.4.4 Qualitative Results

Relative-position Visualization: We visualize the relative-position embeddings in Figure 3.5. In the ViT model, a frame is split into 14 × 14 non-overlapping patches, leading to the number of tokens being 196 (except for the extra classification token). Thus, each frame has 196 relative positions theoretically.

For visualization, we chose the centre position (7, 7) as the reference position (*i.e.*, the red flag in Figure 3.5). Then, the relative distance between the reference position and the remaining 14 × 14 - 1 patch is computed by Equation (3.10). The $(x_i - x_j, y_i - y_j)$ refers a relative-position. We plot the map of the relative encoding p_{ij}, which is defined in Equation (3.7) and Equation (3.11). Specifically, i denotes the reference position, and j refers to one of the 14 × 14 positions. Multiple p_{ij} may share a cluster, shown in the same color in Figure 5. A different color represents the different clusters.

Visualizing feature embeddings. Figure 3.6 and Figure 3.7 report the visualization of the high-level features after the Transformer encoder in ViT and the pro-

Table 3.1: Ablation of different position embeddings and shared/unshared relative position encoding across attention heads over three datasets. The cluster number of each frame is 50.

Strategies	#Param [M]	UCF-101 Top-1 Acc(%)	HMDB-51 Top-1 Acc(%)	Diving 48 Top-1 Acc(%)
Absolute-position embedding	100.7	96.26±0.08	78.08±0.04	79.82±0.04
Shared RPE	97.06	97.53±0.05	79.12±0.04	81.60±0.04
Unshared RPE	97.26	97.61±0.12	79.25± 0.16	81.82±0.08

Table 3.2: Comparison of clip function and sectioned function. The experiments are conducted using shared-head relative-position embedding on our proposed RPE-STDT model over the HMDB-51 and UCF-101 datasets. The cluster number of each frame is 50.

Function	Mode	HMDB-51 Top-1 Acc(%)	UCF-101 Top-1 Acc(%)
Clip	Shared	78.6	97.2
Sectioned	Shared	79.3	97.7

Table 3.3: Comparison of different attention schemes on the UCF-101, HMDB-51 and Diving 48 datasets.

Model	Input Frames	FLOPs [G]	HMDB-51 Top-1 Acc(%)	UCF-101 Top-1 Acc(%)	Diving 48 Top-1 Acc(%)
ViT(spatial attention)	16	269	75.6	95.5	78.5
RPE-STDT	16	273	79.3	97.6	81.8

Table 3.4: Comparison of different subsampling strategies and positions on the UCF-101, HMDB-51 and Diving 48 datasets.

Configurations	HMDB-51 Top-1 Acc(%)	UCF-101 Top-1 Acc(%)	Diving 48 Top-1 Acc(%)
Temporal max-pooling	76.2	96.1	78.3
Temporal average-pooling	76.3	96.3	78.8
1D temporal convolution	78.6	96.6	79.0
Top-k subsampling [Layer 0, 2]	78.6	96.5	79.7
Top-k subsampling [Layer 1, 3]	79.3	97.6	81.8

Table 3.5: Comparison of different subsampling strategies on the baseline-STDT model and positions on the UCF-101 datasets.

Configurations	#Param [M]	UCF-101 Top-1 Acc(%)
Baseline-STDT	88.9	96.5
Baseline-STDT+Top-K subsampling	82.7	97.0

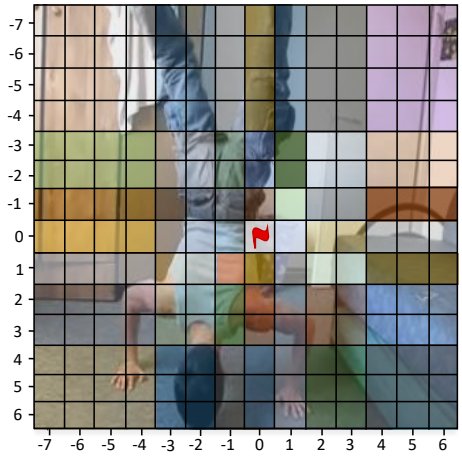


Figure 3.5: Visualization of relative-position embeddings. The red flag refers to the reference position. The positions in the same color share the same relative-position embedding.

posed RPE-STDT model on the HMDB-51 dataset. We use the t-SNE method [56] and project the features into a 2D space. We randomly selected 1000 videos from HMDB-51 to estimate whether the learned features are discriminative. For clarity, we only show videos from the ten action categories in the validation set, which are "sword_exercise", "flic_flac", "dribble", "ride_horse", "laugh", "shake_hands", "shoot_gun", "hit", "sit", "wave". Each star represents a single video, and the stars with the same colors depict the same action classes. Based on Figure 3.7, we can observe that the feature embedding of the proposed RPE-STDT model is more semantically separable than the ViT [24] model, which is a spatial attention-only model.

Confusion Matrices. We report the visualized confusion matrices from different attributes of the Diving 48 dataset in Figure 3.8. Diving 48 is a challenging fine-grained dataset due to the severe occlusions of objects and motions, view-point variations, and different times in performing the same action, such as "Forward, 1.5 somersault, no twist, PIKE" and "Back, 2.5 somersault, no twist, TUCK". As shown in Figure 3.8, the proposed RPE-STDT performs well in recognizing "take off" and "flight pose". However, the primary misrecognizing is the number of "somersaults" and "twists", particularly those with similar counts. This is likely attributed to the model's reliance on spatial information. The bias towards spatial cues, possibly due to the proportion of the spatial domain being larger than the temporal domain, guides the model to

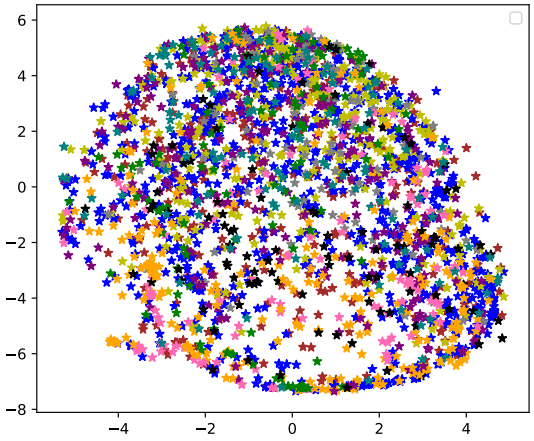


Figure 3.6: ViT: Feature visualization using the t-SNE method on the HMDB-51 dataset.

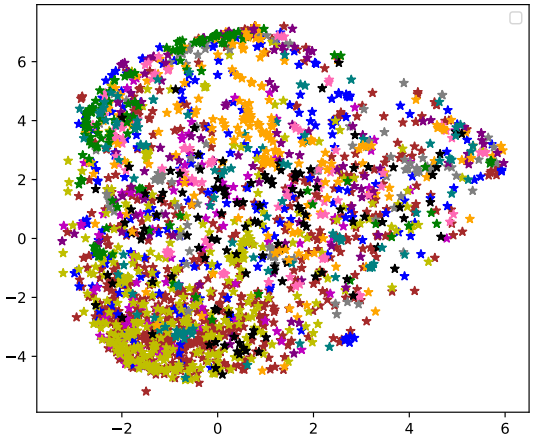
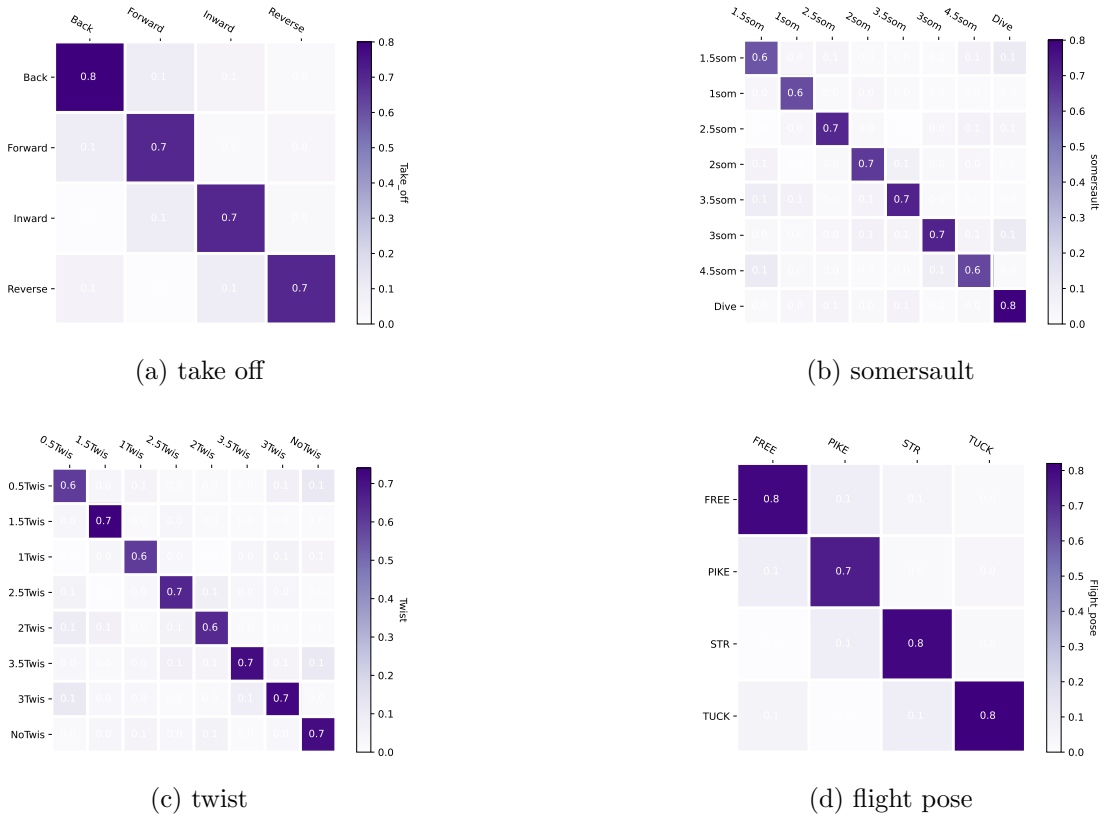


Figure 3.7: RPE-STDT: Feature visualization using the t-SNE method on the HMDB-51 dataset.

prioritize spatial cues over temporal cues. As a result, the model may struggle to discern the subtle temporal variations associated with different counts of "somersaults" and "twists", leading to misrecognition over the Diving 48 dataset.



(a) take off

(b) somersault

(c) twist

(d) flight pose

Figure 3.8: The confusion matrix of the proposed RPE-STDT model from different attributes of the Diving 48 dataset.

3.4.5 Comparison to the State-of-the-Art

Table 3.6 shows the comparison of our proposed RPE-STDT model with state-of-the-art models, including models constructed on pure-Transformer and models built on CNNs on the UCF-101 dataset, HMDB-51 dataset and Diving 48 dataset. The proposed RPE-STDT achieves better accuracy with fewer input modalities than CNN-based models. For example, the proposed RPE-STDT obtains 97.6% accuracy on the UCF-101 dataset, which surpasses the I3D model by 4.2%, surpasses the Two-stream fusion model by 4.1%, surpasses the TSN model by 3.3%, surpasses the TSM model by 3.3%, surpasses the TesNet model by 2.3%, surpasses the SGM model by 7.1%, surpasses the VidTr-M model by 1.0%, and surpasses the MM-ViT model by 1.6%. In

Table 3.6: Comparison of Top-1 classification accuracies of the state-of-the-art approaches on the UCF-101, HMDB-51, and Diving 48 datasets.

Models	Pretrained dataset	Modality	Frames	Params (M)	Accuracy (%)		
					UCF-101	HMDB-51	Diving 48
I3D[57]	Kinetics 400	RGB	64	25	93.4	66.4	-
Two-Stream Fusion[14]	ImageNet	RGB+flow	8	-	93.5	69.2	-
TSN[58]	ImageNet	RGB+flow	3	-	94.2	69.4	52.5
TRNms[59]	ImageNet	RGB+flow	3	-	-	-	54.4
Asymmetric 3D-CNN[36]	FCVID	RGB+flow	16	57.82	92.6	65.4	-
TSM[60]	ImageNet	RGB+flow	3	24.3	94.3	-	51.1
TesNet[61]	ImageNet	RGB+flow	10	-	95.2	71.5	-
STDA-ResNeXt[40]	ImageNet	RGB+flow	-	16	95.1	74.1	-
SGM[62]	Sport1M	RGB	16	30	90.5	60.2	-
TimeSformer[20]	ImageNet	RGB	8	121.4	-	-	74.9
VidTr-M[30]	ImageNet	RGB	16	-	96.6	74.4	-
MM-ViT[26]	ImageNet	RGB	16	158.9	95.9	-	-
ViviT-B[21]	ImageNet	RGB	16	88.9	-	-	77.1
RPE-STDT	ImageNet	RGB	16	**97**	**97.6**	**79.3**	**81.8**

addition, the proposed RPE-STDT model obtains a 79.3% accuracy on the HMDB-51 dataset and 81.8% on the Diving 48 dataset, which outperforms several state-of-the-art models. Our proposed RPE-STDT model shows superior advantages compared to concurrent models built on CNNs and models constructed on pure-Transformer.

3.5 Conclusion

In this paper, we propose a relative-position embedding based spatially and temporally decoupled Transformer (RPE-STDT) for action recognition, which can capture spatial and temporal information by stacked self-attention layers. Extensive experiments indicate that the proposed RPE-STDT achieves state-of-the-art or comparable performance on three public action recognition datasets while retaining computational efficiency. Furthermore, the proposed RPE-STDT model enables long-term video modelling due to temporal attention and subsampling strategy. Going beyond video classification toward more complex tasks is a clear next step in the future, such as video captioning, video storytelling and visual question-answering.

3.6 Limitations

While our proposed RPE-STDT method achieves state-of-the-art results on three public action recognition datasets, it is important to acknowledge its limitations. Firstly, the current version may exhibit an overemphasis on spatial information, as the reliance on relative position embeddings and the relatively larger weight assigned to the spatial domain may overshadow the importance of temporal cues. This bias can hinder the model's ability to capture and utilize subtle temporal variations effectively.

Addressing this issue and achieving a better balance between spatial and temporal cues is a focus of our future work. Secondly, our exploration has been limited to RGB modalities, and the applicability and performance of the method with other modalities, such as optical flow and depth, remain to be further validated.

References

[1] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, 2022.

[2] Tianyu Liu, Yujun Ma, Wenhan Yang, Wanting Ji, Ruili Wang, and Ping Jiang. Spatial-temporal interaction learning based two-stream network for action recognition. *Information Sciences*, 606:864–876, 2022.

[3] Miao Ma, Naresh Marturi, Yibin Li, Ales Leonardis, and Rustam Stolkin. Region-sequence based six-stream cnn features for general and fine-grained human action recognition in videos. *Pattern Recognition*, 76:506–521, 2018.

[4] Zhe Chen, Ruili Wang, Zhen Zhang, Huibin Wang, and Lizhong Xu. Background–foreground interaction for moving object detection in dynamic scenes. *Information Sciences*, 483:65–81, 2019.

[5] Xiaohan Wang, Linchao Zhu, Zhedong Zheng, Mingliang Xu, and Yi Yang. Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision. *IEEE Transactions on Multimedia*, 2022.

[6] Chen Liang, Wenguan Wang, Tianfei Zhou, and Yi Yang. Visual abductive reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15565–15575, 2022.

[7] Jun Li, Xianglong Liu, Wenxuan Zhang, Mingyuan Zhang, Jingkuan Song, and N. Sebe. Spatio-temporal attention networks for action recognition and detection. *IEEE Transactions on Multimedia*, 22:2990–3001, 2020.

[8] Xingguo Chen, Yang Gao, and Ruili Wang. Online selective kernel-based temporal difference learning. *IEEE Transactions on Neural Networks and Learning Systems*, 24(12):1944–1956, 2013.

[9] Ivan Laptev. On space-time interest points. *International journal of Computer Vision*, 64:107–123, 2005.

[10] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories.

In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[11] Pourya Shamsolmoali, Xiaofang Li, and Ruili Wang. Single image resolution enhancement by efficient dilated densely connected residual network. *Signal Processing: Image Communication*, 79:13–23, 2019.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[13] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 27, 2014.

[14] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.

[15] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:221–231, 2013.

[17] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[18] Ming Zong, Ruili Wang, Yujun Ma, and Wanting Ji. Spatial and temporal saliency based four-stream network with multi-task learning for action recognition. *Applied Soft Computing*, page 109884, 2022.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[20] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[21] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. *2021 IEEE/CVF Interna-*

tional Conference on Computer Vision (ICCV), pages 6816–6826, 2021.

[22] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3333–3343, 2022.

[23] Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, 124:108487, 2022.

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021.

[25] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *NAACL*, 2018.

[26] Jiawei Chen and Chiu Man Ho. Mm-vit: Multi-modal video transformer for compressed video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1910–1921, 2022.

[27] Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. Star-transformer: A spatio-temporal cross attention transformer for human action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3330–3339, 2023.

[28] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3156–3165, 2021.

[29] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14063–14073, 2022.

[30] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13577–13587, 2021.

[31] Tam V. Nguyen, Zheng Song, and Shuicheng Yan. Stap: Spatial-temporal attention-aware pooling for action recognition. *IEEE Transactions on Circuits*

and Systems for Video Technology, 25:77–86, 2015.

[32] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *ArXiv*, abs/1705.06950, 2017.

[33] Lei Wang, Xiaoguang Yuan, Ming Zong, Yujun Ma, Wanting Ji, Mingzhe Liu, and Ruili Wang. Multi-cue based four-stream 3d resnets for video-based action recognition. *Information Sciences*, 575:654–665, 2021.

[34] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[35] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[36] Hao Yang, Chunfeng Yuan, Bing Li, Yang Du, Junliang Xing, Weiming Hu, and Stephen J Maybank. Asymmetric 3d convolutional neural networks for action recognition. *Pattern recognition*, 85:1–12, 2019.

[37] Zhigang Tu, Wei Xie, Qianqing Qin, Ronald Poppe, Remco C Veltkamp, Baoxin Li, and Junsong Yuan. Multi-stream cnn: Learning representations based on human-related regions for action recognition. *Pattern Recognition*, 79:32–43, 2018.

[38] Dong Li, Ting Yao, Ling yu Duan, Tao Mei, and Yong Rui. Unified spatio-temporal attention networks for action recognition in videos. *IEEE Transactions on Multimedia*, 21:416–428, 2019.

[39] M Esat Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 731–747. Springer, 2020.

[40] Jun Li, Xianglong Liu, Mingyuan Zhang, and Deqing Wang. Spatio-temporal deformable 3d convnets with attention for action recognition. *Pattern Recognition*, 98:107037, 2020.

[41] Xiaohan Wang, Linchao Zhu, Yu Wu, and Yi Yang. Symbiotic attention for egocentric action recognition with object-centric alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

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

[43] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[44] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. Token shift transformer for video classification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 917–925, 2021.

[45] Li Li and Liansheng Zhuang. Mevit: Motion enhanced video transformer for video classification. In *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part II*, pages 419–430. Springer, 2022.

[46] Zhiheng Huang, Davis Liang, Peng Xu, and Bing Xiang. Improve transformer models with better relative position embeddings. *arXiv preprint arXiv:2009.13658*, 2020.

[47] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019.

[48] A. Criminisi, P. Perez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004.

[49] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6222–6231, 2019.

[50] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6232–6242, 2019.

[51] Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Hao Chen, and Joseph Tighe. Nuta: Non-uniform temporal aggregation for action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3683–3692, 2022.

[52] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin. Kvt: k-nn attention for boosting vision transformers. In *Computer*

Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV, pages 285–302. Springer, 2022.

[53] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[54] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[55] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018.

[56] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

[57] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[58] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[59] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.

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

[61] Guoxi Huang and Adrian G Bors. Learning spatio-temporal representations with temporal squeeze pooling. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2103–2107. IEEE, 2020.

[62] Tingzhao Yu, Lingfeng Wang, Cheng Da, Huxiang Gu, Shiming Xiang, and Chunhong Pan. Weakly semantic guided action recognition. *IEEE Transactions on Multimedia*, 21(10):2504–2517, 2019.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Yujun Ma
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 3

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Yujun Ma, and Ruili Wang. "Relative-position embedding based spatially and temporally decoupled Transformer for action recognition." Pattern Recognition 145 (2024): 109905.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Yujun Ma Digitally signed by Yujun Ma DN: cn=Yujun Ma, c=NZ, o=Massey University, ou=School of Mathematical and Computational Science, email=yma1@massey.ac.nz Date: 2024.02.20 14:22:25 +13'00'
Date:	20-Feb-2024
Primary Supervisor's Signature:	Prof Ruili Wang Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, c=NZ, email=ruili.wang@massey.ac.nz Location: School of Natural and Computational Sciences Date: 2024.07.23 16:00:15 +12'00'
Date:	

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

Chapter 4

Convolutional transformer network for fine-grained action recognition

Fine-grained action recognition is one of the critical problems in video processing, which aims to recognize similar actions of subtle interactions between humans and objects. Inspired by the remarkable performance of the Transformer in natural language processing, Transformer has been applied to the fine-grained action recognition task. However, Transformer needs abundant training data and extra supervision to achieve comparable results with convolutional neural networks (CNNs). To address these issues, we propose a Convolutional Transformer Network (CTN), which integrates the merits of CNNs (e.g., sharing weights, capturing low-level features in videos and locality) and the benefits of Transformer (e.g., dynamic attention and learning long-range dependencies). In this paper, we propose two modifications to the original Transformer: (i) We propose a video-to-tokens module that can extract tokens from extracted spatial-temporal features in videos by 3D convolutions instead of the direct token embedding from raw input video clips; (ii) We completely replace the linear mapping in multi-head self-attention layer with depth-wise convolutional mapping, which applies a depth-wise separable convolution operation on embedded token maps. With these two modifications, our approach can extract effective spatial-temporal features from videos and process the long sequences of tokens encountered in videos. Note that the work presented in this chapter has been published in the journal of Neurocomputing.

4.1 Introduction

Increased camera coverage and constantly growing network bandwidth for video streaming are making action recognition more demanding than before in many domains, such as elderly behavior monitoring, video index, and human-computer interaction [1] [2] [3]. There are two types of action recognition: coarse-grained action recognition and fine-grained action recognition. The coarse-grained actions commonly involve full-body motions (*e.g.*, cleaning up and cooking), which are recognized easily due to high intra-class similarity. The fine-grained actions (*e.g.*, turn-on tap, turn-off tap, wash spoons, and wash plates) involve subtle interactions between humans and objects, which are more difficult to differentiate because of the existence of various objects, similar backgrounds, high inter-class similarity, and low intra-class similarity [4].

Fine-grained action recognition can be widely used in homes or a particular environment industry as assistive technology. As shown in Figure 4.1, the objects/actors and background information from only a few frames can inform the classification results in a coarse-grained action recognition dataset UCF101 [5]. However, fine-grained action recognition depends on distinguishing subtle differences in actions and temporal ordering of objects such as "Forward,15 somersault, no twist, PIKE" and "Back, 25 somersault, no twist, TUCK" in the Diving 48 dataset [6]. Therefore, fine-grained action recognition is challenging due to the occlusion of objects and motions, view-point variations, and different times in performing the same action [4].

In image classification, Convolutional Neural Networks (CNNs) have been utilized for action recognition by 3D convolutions, which can extract both spatial and temporal features in videos [7] [8]. However, 3D convolutions are constantly involved with expensive computations. In recent years, most performant approaches for action recognition were based on two-stream 2D convolutional Convolutional Neural networks Networks (CNNs). These approaches [9] [10] utilized 2D CNNs to extract spatial and temporal features from stacked RGB video frames and optical flow frames individually and then fuse them in the last layer.

Inspired by the recent success of Transformer in the field of Natural Language Processing (NLP) [11], Transformer has been used in computer vision tasks such as image classification and action recognition [9] [12] [13] [14] [15]. Like an NLP model, the input words are represented as a sequence of tokens. Images can be represented as

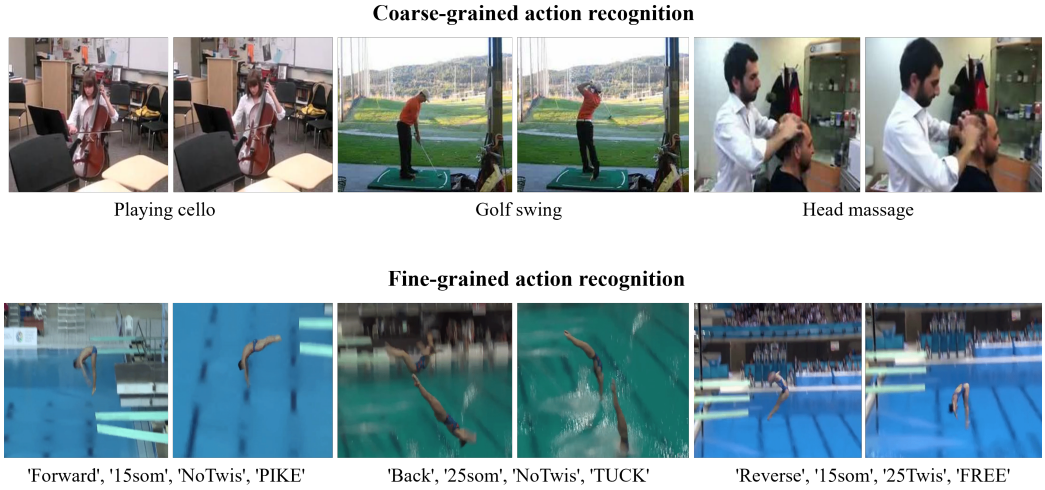


Figure 4.1: Top: three examples selected from the coarse-grained action recognition dataset UCF101. Bottom: three examples selected from the fine-grained action recognition dataset Diving 48.

a sequence of image patches, and videos can be represented as RGB frames. Several attempts [13] [14] completely replaced CNNs with Transformer. The Transformer [11] builds on multi-head self-attention layers, which can learn global attention from the input sequence.

Dosovitskiy *et al.* [16] proposed a Vision Transformer (ViT) model for image classification, which directly utilized a pure-Transformer to capture visual features from images. The ViT outperformed several state-of-the-art CNNs for large-scale image classification tasks. Later, Arnab *et al.* [17] proposed a Video Vision Transformer (ViViT) model for action recognition, which utilized a pure-Transformer to extract spatial and temporal features from a sequence of spatial-temporal maps that were generated from the videos.

However, compared to CNNs, Transformer lacks some image-specific inductive bias, which requires a significant amount of data to learn these inductive biases. To solve this problem, many recent approaches, such as CoAtNet [18], BoT [19], and CoTr [20], attempted to incorporate the inductive biases of CNNs into Transformer, which could impose local receptive fields for self-attention.

The two most essential characteristics of CNNs are locality and translation equivariance [18]. Locality in CNNs is a basic acknowledgment that neighboring pixels are always related [19]. On the other hand, translation equivariance is relevant to the

shared weights, improving the generalization of CNNs based models with limited-size datasets [21]. However, pure- Transformer models cannot take advantage of these prior biases in video frames. Firstly, ViviT [17] performs tokenization of tubelets directly from the input video clips with a size of 16162, which ignores the low-level features such as edges and corners in videos. Secondly, the self-attention mechanism in Transformer mainly focuses on modeling long-range relationships among the tokens while ignoring the relationship of neighbored pixels in the video frames.

To address the issues mentioned above, we propose a Convolutional Transformer Network (CTN) for fine-grained action recognition, which takes advantage of CNNs in capturing spatial-temporal features and sharing weights, and the benefit of the Transformer in modeling long-range relationships. We made two changes to the original ViviT [17] model: (i) We propose a novel video-to-tokens module that can extract tokens from the extracted spatial-temporal features in videos by light 3D convolutions instead of the direct token embedding from raw input video clips; (ii) We completely replace the fully-connected layer based linear mapping in multi-head self-attention layer with depth-wise convolutional mapping, which applies a depth-wise separable convolution operation on embedded token maps. We make the following contributions:

- We propose a new Convolutional enhanced Transformer Network (CTN) model for fine-grained action recognition. The proposed CTN model takes all the advantages of CNNs, including sharing weights and local receptive fields, while inheriting all the benefits of Transformer, including the global receptive field and dynamic attention.
- Experimental results indicate that the proposed CTN shows better results compared to models constructed on CNNs (*e.g.*, I3D [22]) and models constructed on pure-Transformer (*e.g.*, ViviT [17] and TimeSformer [23]) on two fine-grained action recognition datasets (*i.e.*, Epic-Kitchens [24] and Diving 48 [6]) while utilizing with a similar number of FLOPs.
- Additionally, we show that the positional embedding in ViviT [17] can be removed without any degradation to the results performed in the proposed CTN model.

The rest of this chapter is organized as follows: The related works on action recogni-

tion and Transformer are reviewed in Section 4.2. Section 4.3 presents the proposed Convolutional Transformer Network (CTN). The experimental setting details and experimental results are provided and discussed in Section 4.4. The conclusion and future work are represented in Section 4.5.

4.2 Related work

Action recognition has attracted intensive research interest recently. This section briefly provides a literature review of coarse-grained action recognition models, fine-grained action recognition models, and Transformer-based models.

4.2.1 Coarse-grained action recognition

Approaches for action recognition usually utilized 2D or 3D convolutions on spatial and temporal information in videos [25] [26] [27]. The breakthrough of AlexNet [28] in image classification initially guides the success of two-stream convolutional networks (two-stream CNN) for action recognition [9]. Concretely, one stream extracted appearance-based features, and the other stream extracted motion-based features. Experimental results demonstrated that motion-based features were particularly useful for fine-grained action recognition tasks (*e.g.*, flic_flac and somersault), where small differences in motion can be critical for accurate recognition. Later, the emergence of larger action recognition datasets such as Kinetics thereupon promoted the training of spatial-temporal 3D CNNs [22]. Feichtenhofer *et al.* [7] proposed an X3D network for action recognition, which used network research to obtain parameters such as input resolution, temporal demission, network depth, and width. Tran *et al.* [29] proposed an R3D network for action recognition, which decomposed the 3D convolutions into two independent convolutions to extract spatial and temporal features in videos. Fan *et al.* [30] proposed to model optical flow-like features from videos in an end-to-end manner, which could learn both spatial and temporal representations jointly in a single model.

Recently, some approaches applied the attention mechanism and Transformer in action recognition. Li *et al.* [26] proposed a unified spatio-temporal attention network for action recognition, which utilized an attention neural cell to generate the attention weights on both spatial and temporal dimensions. Kalfaoglu *et al.* [31] proposed a hybrid of 3D convolutions and a Transformer model for action recognition, which ap-

plied a Transformer to process the 3D CNN's output feature vectors. Zhu *et al.* [32] proposed a temporal cross-layer correlation model for action recognition, a two-stage operation that uncovered both local and global fields from video. Girdhar *et al.* [33] proposed a video action Transformer network for action recognition, which utilized a Transformer to aggregate the spatiotemporal context features. Wang *et al.* [10] proposed a spatial-temporal pooling-based network for action recognition consisting of temporal and spatial attention blocks. The network was trained via a spatial-temporal loss function to enhance the sparsity of weight from spatial and temporal dimensions.

4.2.2 Fine-grained action recognition

Fine-grained action recognition refers to distinguishing between subtle actions that are similar in appearance but have distinct differences in execution [34]. For example, differentiating between different types of diving moves or sports playing techniques. In this case, low-level features such as motion vectors, joint levels, and texture patterns can be critical for accurately recognizing the subtle differences between these actions [35]. However, coarse-grained action recognition refers to distinguishing between broad categories of actions such as walking, eating, or fencing. In this case, high-level features such as body posture, motion direction, and overall motion patterns are often sufficient to recognize the actions [36] [37] [38].

Zhu *et al.* [39] proposed a multi-view attention mechanism-based model for fine-grained action recognition, which assisted the network focus on the essential clues. Liu *et al.* [40] proposed a motion saliency and mid-level patches-based model for fine-grained action recognition, which could model mid-level patches' spatial and temporal features and the motion relationship among patches.

Later, Munro *et al.* [34] proposed a different modality domain adaptation model for fine-grained action recognition, which utilized multi-model self-supervision. Ma *et al.* [41] proposed a region-sequence-based six-stream CNN model for fine-grained action recognition, which contained six independent streams. The input of six streams had both spatial and temporal cues by cropping the video frames at different scales (*e.g.*, human region and operation region). Zhang *et al.* [35] proposed a temporal query network for fine-grained action recognition, which obtained a query-response mechanism and a structural understanding of fine-grained actions. The proposed network

applied multiple layers of the decoder in Transformer. Despite these achievements, challenges remain in distinguishing subtle differences between actions and objects and recognizing fine-grained actions without actors and background biases.

4.2.3 Vision Transformers

The transformer was first proposed by Vaswani *et al.* [9] for Natural Language Processing (NLP), where it achieved remarkable performance and has been widely applied in many NLP tasks such as language translation and text generation. It was composed of an encoder module and a decoder module. Both the encoder module and decoder module consisted of a self-attention layer and a feed-forward network.

Inspired by the significant success of Transformer in NLP tasks [9], Transformer has been applied to computer vision (CV) studies [18] [19]. The Vision Transformer (ViT) was first proposed by Dosovitskiy et al. [16], which was the first attempt to show that a pure-Transformer could achieve remarkable results on image classification tasks when dealing with large-scale datasets (*i.e.*, ImageNet22k and JFT300M). Dai *et al.* [18] proposed a CoAtNet for image classification, combining convolutional layers and self-attention layers to inherit the strengths from CNNs and Transformer. Srinivas *et al.* [17] proposed a bottleneck Transformer model for visual recognition, which replaced the convolution layers with self-attention layers in the last three bottleneck blocks of the residual network.

Yuan *et al.* [42] introduced a layer-wise Tokens-to-Token projection that progressively structured an image into tokens by recursively aggregating neighboring tokens into one. This process enabled the modeling of the local structure represented by surrounding tokens and reduced the overall length of the tokens. However, each image patch was treated as a separate token in a purely transformer-based model, which might limit its ability to capture fine-grained spatial information. Later, Wang *et al.* [43] proposed an overlapping patch embedding to tokenize images with a pyramid-shaped vision Transformer model, where lower layers captured local features while higher layers captured global features. However, the complexity of pyramid-shaped models could make them challenging to train and optimize. Chen *et al.* [44] presented a self-attention-based spatial branch and a convolutional channel brunch to interact with one another, which provided complementary clues in the channel and spatial dimensions for image classification.

Recently, Transformer-based approaches have achieved promising results in action recognition. Girdhar *et al.* [45] proposed an anticipative video Transformer model for action recognition, which applied a Transformer as a frame feature encoder to predict future frames' features. Liu *et al.* [14] proposed a video swin Transformer for action recognition, which applied a non-overlapping 3D shifted window in a vanilla Transformer block to extract both spatial and temporal features in videos. Li *et al.* [46] proposed a separable attention-based video Transformer for action recognition, which expanded the ViT Transformer model for 3D feature learning. The proposed spatial-temporal separable attention video Transformer aggregated spatial-temporal features from raw videos and captured 3D motion features on a sequence of local patches.

Sharir *et al.* [15] developed a space-time attention model for video classification, which factorized the Transformer encoder into the spatial encoder and the temporal encoder. The spatial encoder processed attention to the input embedded patches and then the temporal encoder processed attention to the sequence of frame embedding vectors. Neimark *et al.* [47] proposed a video Transformer network for video classification, which consisted of a 2D spatial feature extractor and a temporal attention-based encoder. The temporal attention-based encoder applied attention mechanisms to make global dependencies in a long sequence. Yan *et al.* [48] proposed a multi-view Transformer for action recognition, which consisted of individual parallel encoders to learn different views of a video clip. The extracted features from different views were fused by a linearized cross-view attention layer. Long *et al.* [49] investigated a pure attention cluster-based model for video classification, which concatenated the output of multiple attention components applied in parallel to model a global representation.

However, self-attention mechanisms lack certain desirable inductive biases of convolutional models, requiring more data and computational resources to compensate for these limitations. Furthermore, fine-grained action recognition often occurs in the same environment or a limited number of settings, making it hard to obtain an extremely large-scale dataset. To address the above issues, we propose a Convolutional Transformer Network (CTN) for fine-grained action recognition, which inherits all the benefits of CNNs and the Transformer. Both the spatial and temporal information in videos are considered.

4.3 Proposed CTN model

Fine-grained action recognition depends on subtle variations in a specific sub-motion, object, and pose. The common assumption in a visual task is that neighboring pixels of video frames are always correlated. Therefore, modelling the relationship of neighboring features and low-level features in video frames is important for fine-grained action recognition such as "turn-on tap" and "turn-off tap". However, pure-Transformer cannot fully utilize these prior biases in videos. To use the advantages of CNNs (e.g., weight sharing and local receptive field) while inheriting all the benefits of Transformer (e.g., global receptive field and dynamic attention). We propose a novel video-to-tokens based Convolution enhanced Transformer Network (CTN) for fine-grained action recognition. The details of the video-to-tokens module and the convolutional Transformer blocks are introduced in this Section. An overview of CTN is shown in Figure 4.2 (a).

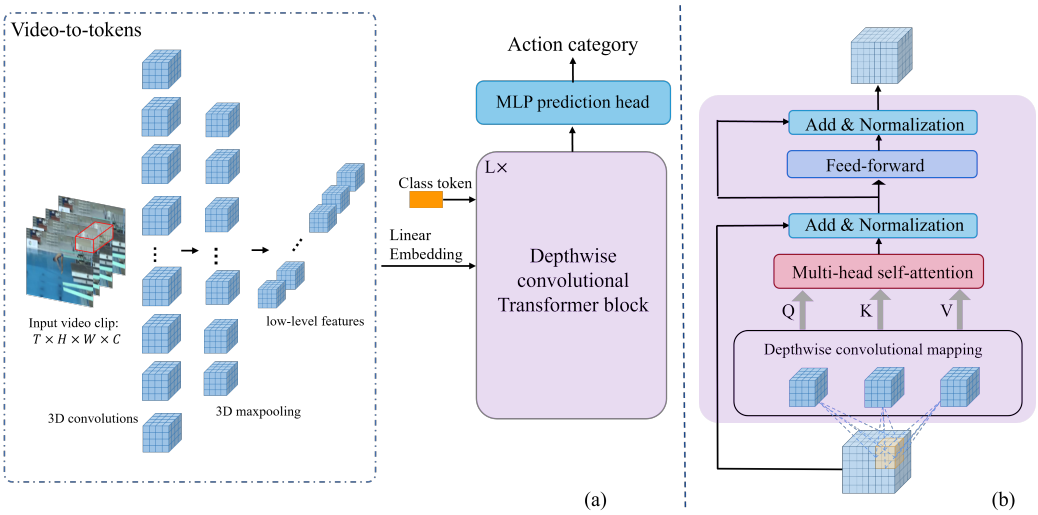


Figure 4.2: (a) The overall architecture of the proposed Convolutional Transformer Network. The proposed CTN contains a video-to-tokens module, multiple Transformer encoders, and an MLP classification head; (b) The detailed architecture of the depth-wise convolutional Transformer encoder.

4.3.1 Preliminaries: ViviT

We begin with an overview of the vision Transformers, ViT [16], and its extension to video, ViViT [17], on which our model is based. ViviT [17] includes token embedding, Transformer encoder blocks, and a classification head. The Transformer block

contains a multi-head self-attention (MHSA) layer and a feed-forward network (FFN) layer.

1) Tokenization/Embedding video clips. The vanilla Transformer encoder receives a sequence of token embeddings as input. ViT [16] segments an image into multiple non-overlapping patches and linearly projects them into embedded tokens. To handle 3D videos, ViviT [17] maps a video clip $V \in \mathbb{R}^{T\times H\times W\times C}$ to a sequence of tokens following ViT [16]. Then the positional embedding is added to the tokens to retain the positional information. In the meantime, an optional learned classification token is concatenated to the embedded sequence. The class token serves as a placeholder data structure for the final classification representation used by the MLP classification head. The final obtained reshaped sequence is regarded as the input of the Transformer encoder.

ViviT [17] extends the patch-based partitioning of images from ViT to handle video inputs. Specifically, it extracts N non-overlapping, spatio-temporal tubelets of size 16×16×2 from the input video. This technique can be regarded as an extension of the ViT embedding technique into the 3D domain, where it is utilized as a form of 3D convolution. The convolution operates with a kernel size of $t \times h \times w$ and a stride of (t, h, w) in the time, height, and width dimensions, respectively. However, directly tokenizing input videos with large tubelets can have two limitations. Firstly, it may be difficult to capture low-level information in videos such as edges, backgrounds, and corners. Secondly, large kernels may be over-parameterized and difficult to optimize, since they may require a substantial amount of training samples or training iterations. Additionally, since the temporal information in videos may not be fully considered, accurate action recognition may not be achieved.

2) Transformer Encoder blocks. ViviT [17] consists of multiple stacked Transformer encoders. Each encoder block has two sub-layers: a multi-head self-attention (MHSA) layer and a feed-forward network (FFN) layer. A residual skip connection is applied with each sublayer, followed by layer normalization (LN) [50]. The Transformer block is computed as:

$$x' = \text{LayerNorm}(x + \text{MHSA}(x)) \tag{4.1}$$

$$y = \text{LayerNorm}(x' + \text{FFN}(x')), \tag{4.2}$$

where x denotes the input of one Transformer block and y denotes the output of one Transformer block.

3) Multi-head self-attention. For a self-attention layer, the sequence of input tokens $x_t \in \mathbb{R}^{((N+1)\times C)}$ is linearly mapped into Q (query), K (key), and V (value) spaces. Then a weighted sum of overall values in the sequence is computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{C}}\right)V. \tag{4.3}$$

Then a linear transformation is committed to the weighted values. Multi-head self-attention is an extension of self-attention, which splits Q, K, and V for h (*i.e.*, the number of heads) times and computes the attention function in parallel. The final output is the concatenation of the output from each head. The similarity among different tokens is calculated by computing self-attention. However, the spatial relationship among tokens is not considered leading to the original ViviT needing to learn a mass of training data.

4) Feed-forward network (FFN). Each Transformer encoder block contains a fully connected feed-forward network. FFN is composed of two linear transformations separated by a non-linear activation GELU (Gaussian error linear Unit), which can be computed as:

$$FFN(x) = \ \sigma(xw_1 + b_1\ )\ w_2 + b_2, \tag{4.4}$$

where w_1 is the weight matrix of the first layer, and w_2 is the weight matrix of the second layer; b_1 and b_2 denote biases; σ is the non-linear activate function GELU. The feed-forward network can enhance the representation ability of tokens by expanding or reducing the dimension and performing non-linear projection on each token.

However, the spatial relationship among pixels is ignored, which is important in video frames. This leads that the original ViviT [17] needs abundant training data to learn these inductive biases while obtaining comparable performance with CNNs.

4.3.2 Overall architecture of CTN

The overall architecture of the proposed Convolutional Transformer Network (CTN) is shown in Figure 4.2. We present two operation types that incorporate convolution designs into the ViviT model, including a video-to-tokens module and depth-wise convolutional mapping blocks. Figure 4.2 (a) shows the proposed CTN contains a video to tokens module, multiple Transformer encoders, and an MLP classification head. We present the video-to-tokens module to extract tubelets from generated feature maps instead of raw input video clips. A tensor then embeds the tokens and passes through a Transformer encoder structure consisting of a sequence of L Transformer encoders. We append a classification token following previous work [17], whose purpose is to aggregate features from the whole sequence for classification.

Specifically, we stack 12 convolutional-Transformer encoders, with each encoder consisting of a 12-head self-attention layer following [17]. The detailed convolutional Transformer encoder block is shown in Figure 4.2 (b). Rather than the standard position-wise linear mapping in the original Transformer encoder [17], we use a depth-wise separable convolution to generate query (Q), key (K), and value (V) embeddings, respectively. Finally, a linear multilayer perceptron (MLP) prediction head is applied to the classification token to predict the action category.

4.3.3 Video-to-Tokens

To address the issues mentioned above in tokenization, we propose a simple but effective video-to-tokens module that extracts tokens from the spatial-temporal feature maps. Our video-to-tokens module is a lightweight feature extractor that performs feature capture and pooling, distinguishing it from the 3D embedding utilized in the ViviT model, which decomposes a video into multiple tubelets. As shown in Figure 4.2, the video-to-tokens module is a lightweight convolutional stem consisting of a 3D convolutional layer and a 3D max-pooling layer. The batch normalization layer follows the convolutional layer to benefit the training process. Given a video clip $V \in \mathbb{R}^{T \times H \times W \times C}$, where T denotes the length of the clip; W and H denote the width and the height of the video frame, and C represents the number of channels. The video-to-tokens can be represented as:

$$Z' = \text{3D-MaxPool}(\text{BN}(\text{3D-Conv}(z))) \tag{4.5}$$

where z refers to a token tensor representing an input video clip; $z' \in \mathbb{R}^{\frac{T}{s} \times \frac{H}{s} \times \frac{W}{s} \times C'}$ represents the output tokens of video-to-tokens module; s is the stride, and C' is the number of enriched channels. The extracted features are embedded into a sequence of tokens in both the spatial and temporal dimensions. We set $s = 4$ following the setting in [17].

The embedding filter is a 3D tensor because of the 3D feature matrixes. The video-to-tokens module fully utilizes the advantages of CNNs, which can extract low-level features and model the relationship of neighboring pixels in videos.

4.3.4 Depth-wise convolutional mapping

Video-based action recognition approaches distil features from the spatial and temporal dimensions of an input video. To integrate the benefit of CNNs to extract low-level features with the power of a Transformer to model long-range relationships, we propose a depth-wise convolutional mapping for multi-head self-attention. Basically, the proposed Transformer encoder block with depth-wise convolutional mapping is a generalization of the original Transformer encoder block. Existing approaches [18] [19] offer a more flexible mixture of a self-attention mechanism and convolution or add specific convolution layers to the Transformer encoder block for image classification and machine translation. However, these approaches often show a more complicated architecture and expensive computational cost. Therefore, to incorporate convolution designs into Transformer and given the input tokens, we propose a depth-wise convolutional mapping for multi-head self-attention instead of the original linear mapping using a fully connected layer.

The original linear mapping applied in vanilla Transformer and ViviT is shown in Figure 4.3, which obtains the query, key, and value matrices, respectively. The detailed operation of our proposed depth-wise convolutional mapping is shown in Figure 4.4. The input tokens are first reshaped into a token map. Then a depth-wise separable convolution layer is used as the convolutional mapping while the kernel size is $s \times s$.

In depth-wise convolution mapping, the convolution operation is decomposed into two steps using two types of filters. Firstly, separated s×s filters are first utilized for each channel, then a 1×1 filter is used to obtain a pointwise linear combination on feature maps generated by the first step. Finally, the tokens are flattened and projected into the initial dimension.

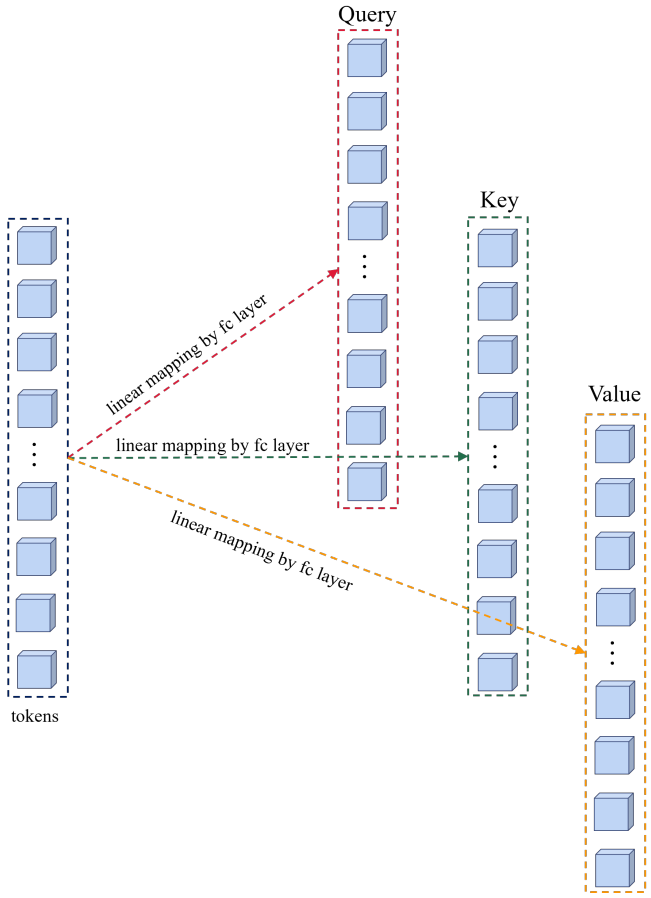


Figure 4.3: Linear mapping in Transformer and ViviT.

Batch normalization is applied after the first step. The mapping can be calculated as:

$$z_l^{q,k,v} = Flatten(Conv2d(Reshaped\ 2d(z_l), s)), \tag{4.6}$$

where z_l represents the unperturbed token before the convolutional projection; $z_l^{q,k,v}$ indicates the token input for query (Q), key (K), and value (V) matrices at layer l. Conv2d is a depth-wise separable convolution obtained by: Depthwise Conv $\rightarrow$ BatchNorm $\rightarrow$ Pointwise Conv, and s denotes the kernel size of convolution. We set s=3 in practice.

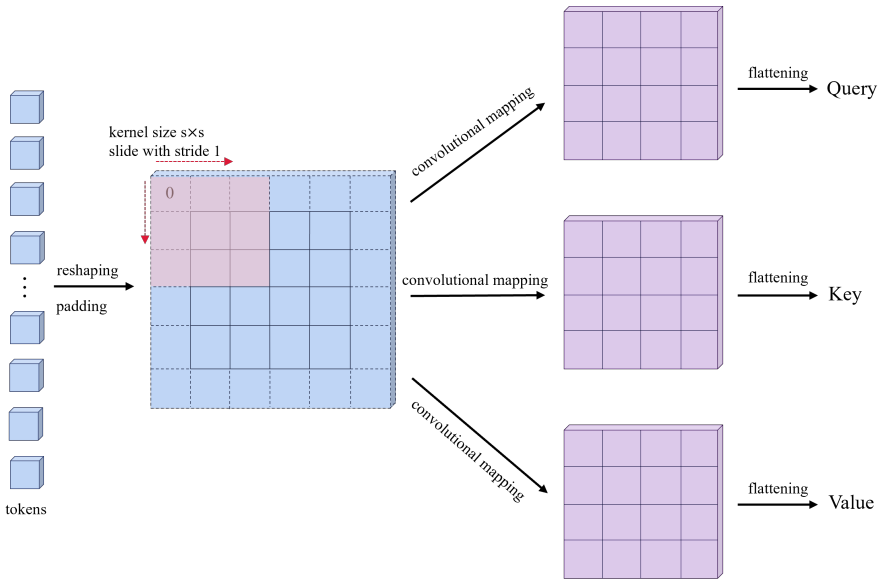


Figure 4.4: The detailed operation of the proposed depth-wise convolutional mapping.

4.3.5 Computational Complexity Analysis

We give the analysis of the extra computational cost of our proposed CTN model by the value of floating-point operations (FLOPs). Generally, the proposed CTN model can efficiently inherit the advantages of both CNN and Transformer and result in a better performance with a slight computational increase.

1) Video-to-tokens vs Tubelets with positional embedding. The method of tokenization can significantly affect the computational complexity of embeddings. Tokenization in ViviT decomposes the video clip into multiple tubelets with the size of 16×16×2, and the total FLOPs are $3C(THW)^2$. In the ViviT model, token embeddings are extracted from the height, width, and temporal dimensions of the input video using a 3D convolution with a kernel size of 16×16×2 and a stride of 16, 16, 2. Smaller tubelet sizes equal more tokens and, thus, more computation.

For the video-to-tokens module, the FLOPs include two parts which are feature extraction and embedding. The 3D convolution FLOPs of video-to-tokens are computed as:

$$7 \times 7 \times 2 \times 7 \times \frac{H}{2} \times \frac{W}{2} \times \frac{T}{2} \times D = \frac{343}{4} DHWT, \tag{4.7}$$

where the kernel size of the max-pooling is 3.

The final operation is equivalent to utilizing a convolution to split the tubelets. The kernel size and stride are the same (p/s). The amount of calculation is:

$$\frac{P}{S} \times \frac{P}{S} \times \frac{P}{S} \times \frac{H}{P} \times \frac{W}{P} \times \frac{T}{P/S} \times D \times C = \frac{1}{16} DHWTC, \tag{4.8}$$

where $S = 4$; the resolution of tubelets is reduced to $\left(\frac{P}{S}, \frac{P}{S}, 2\right)$. The total FLOPs of video-to-tokens are $\left(\frac{343}{4} + \frac{9}{16}\right) DHWT + \frac{1}{16} DHWTC$. For the ViviT-B/16×2 model, the ratio between the video-to-tokens module and the one in ViviT is around 1.1. Therefore, the additional computational cost of our proposed CTN model is negligible.

2) Linear mapping vs Convolutional mapping. We utilize efficient depth-wise separable convolutions for convolutional mapping. The standard convolutions would require. The standard convolutions would require $\mathcal{O}\left(k^2 C^2 T\right)$ FLOPs, where k denotes the size of the convolutional kernel; C denotes the number of token channel dimensions, and T denotes the number of tokens for modeling. We replace the convolution with a depth-wise separable convolution. Specifically, a convolution operation is decomposed into two steps using two types of filters: (i) applying separate filters for each individual channel; (ii) using a 1×1 filter to obtain a pointwise linear combination on the feature maps output by the first step.

Finally, the total FLOPs of our proposed convolutional mapping are $\mathcal{O}\left(k^2 CT + CT\right)$. The FLOPs of a fully connected layer with C input channels and T tokens for processing are $\mathcal{O}\left(CT\right)$. Therefore, each proposed convolutional mapping only brings an additional $\mathcal{O}\left(k^2 CT\right)$ FLOPs compared to the original fully connected layer based linear mapping in multi-head self-attention, which are negligible concerning the total FLOPs of the model.

4.4 Experiments

In this section, we evaluate the proposed CTN on two fine-grained action recognition datasets. We perform ablation studies to validate the performance of the proposed network. Besides, we compare our proposed CTN model with several state-of-the-art fine-grained action recognition models including CNN-based models and Transformer-based models.

4.4.1 Network architecture and experimental details

We build the proposed Convolutional Transformer Network (CTN) in a similar configuration to ViviT [17], which details are given in Figure 4.2. We use a video clip of 8 frames as input, which are cropped into 224×224. The video-to-tokens module includes a 3D convolutional layer. The kernel size is 7×7×7 and has a stride of 2. The number of enriched channels is 32. To enhance the stability of the training process, a batch normalization layer is added. Finally, a 3D max-pooling layer is followed. The kernel size is 3×3×3 and has a stride of 2. We utilize a patch size of 4×4×2 to generate a sequence of tokens based on the patch size of 16×16×2 in ViviT-Base. Specifically, we obey the same setting in [17] about the number of encoder blocks, which is 12. Each encoder block consists of a 12-head self-attention layer. For the convolutional mapping in the multi-head self-attention block, the kernel size of the depth-wise convolution is set to 3×3.

We initialize the weights of the proposed CTN from a ViviT vanilla model pre-trained on ImageNet [51]. Table 2.1 shows the detailed hyper-parameters of our experiments. We use SGD [17] as the optimizer with a momentum of 0.9, and a cosine learning rate schedule with linear warm-up during the fine-tuning is based on [17] for 100 epochs. The batch size is set to 16. We apply data augmentation (*e.g.*, random crop, random flip, and color jitter) on both the Epic-Kitchens dataset [24] and the Diving 48 dataset [6] during training. The models were trained on 4 A100 GPUs using PyTorch.

4.4.2 Datasets and Metrics

Diving 48. Diving 48 [6] is a fine-grained action recognition dataset, which contains 18k diving video clips from 48 categories. Three examples that have been selected from the Diving 48 dataset are shown in Figure 4.5. Diving 48 similarly evaluates action recognition by having a standard diving setting where subtle details of diving sequences define the various categories instead of coarse actors or backgrounds. The video clips of Diving 48 are obtained by segmenting online videos of major diving competitions worldwide. The ground-truth labels are transcribed from the information board before the start of each dive. As shown in Figure 4.5, the label of each dive is defined by a combination of takeoff, pose in flight (*i.e.*, somersaults and twists), and an entry position. We use the standard split which contains 16k training clips and 2k test clips. The lengths of videos have a wide range: from 24 frames to 822

Table 4.1: The details of our proposed CTN model

Model	Video-to-tokens	Encoder	Dimension	Heads	MHSA
CTN	Conv: [7 7 7], stride 2 Max-pool: [3 3 3], stride 2 Channels: 32	12	768	12	12

frames, the average number of frames is 158. We use the cleaned-up labels released in Oct 2020, which are denoted as 'v2'.

Epic-Kitchens 100. Epic-Kitchens 100 [24] is a popular dataset for fine-grained action recognition, which contains 90,000 egocentric video clips recording daily kitchen actions spanning 100 hours in 45 environments. We report the performance following the standard "action recognition" protocol. As shown in Figure 4.5, each video has two types of labels, which are "verb" and "noun". Specifically, Epic-Kitchens 100 contains 97 "verb" classes and "300" noun classes. Therefore, we construct a single network with two "heads" to predict both "verb" and "noun" categories. The predictions are then merged to construct an "action" which is used to calculate the accuracy as the primary metric.

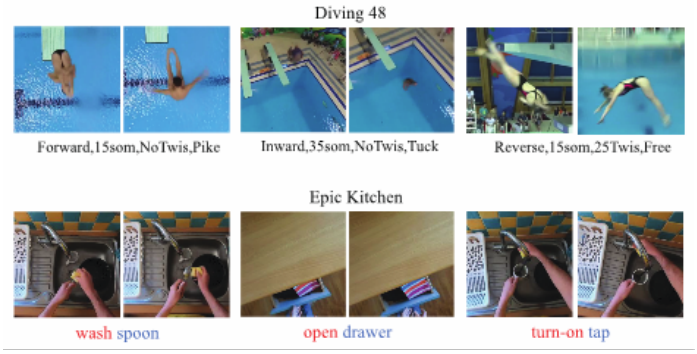


Figure 4.5: Examples in the Diving 48 and Epic-Kitchens datasets.

Metrics. The evaluation metric is Top-1 classification accuracy for the Diving 48 and Epic-Kitchens datasets.

4.4.3 Ablation Study

To further investigate the effectiveness of the CTN model, we conduct various ablation experiments. Firstly, we show the difference between ViviT's 3D input embedding and our proposed video-to-tokens module. We also demonstrate that incorporating

Table 4.2: The detailed training hyperparameters for the experiments

Optimization methods	Epic-Kitchens	Diving 48
Optimizer	SGD	SGD
Momentum	0.9	0.9
Batch size	16	16
Learning rate schedule	Cosine linear warmup	Cosine linear warmup
Linear warmup epochs	2.5	2.5
Base learning rate	0.005	0.005
Epochs	100	100

Table 4.3: Ablation study results on the input embedding and positional embedding.

Model	Input Embedding	Pos. Embedding	Accuracy (%)		FLOPs ($\times 10^9$)
			EK	D-48	
ViviT-B	Tube. embedding	Default	43.1	77.1	325
ViviT-B	Tube. embedding	N/A	42.0	76.2	318
CTN	Tube. embedding	Default	43.9	77.1	353
CTN	*Video-to-tokens*	Default	44.5	77.7	370
CTN	*Video-to-tokens*	N/A	45.9	78.2	358

convolutions eliminates the need for positional embedding. Finally, we investigate the effectiveness of the proposed video-to-tokens and convolutional mapping.

Removing positional embedding. The local context in videos can be captured given that we have incorporated convolutions into the model, we conduct the ablation study to prove whether the positional embedding is still necessary for CTN. Table 4.3 shows the comparison results of the Epic-Kitchens dataset and the Diving 48 dataset. The comparation results between different input embeddings (*i.e.*, Tubelets embedding and video-to-tokens) show the efficiency of our proposed convolutional lightweight stem. This is because the video-to-tokens module maps the low-level visual embedding before the Transformer encoder, while the tubelets embedding is directly decomposed from the video. The results also indicate that removing positional embedding from our proposed video-to-tokens module does not degrade the performance. Moreover, despite the modest increase in floating-point operations (FLOPs) relative to ViviT-B [17], the performance of the proposed method improved significantly.

As a comparison, dropping the positional embedding of ViviT-B [17] would lead to a 1.1% drop in accuracy in the Epic-Kitchens dataset, as it does not learn the spatial relationships other than by concatenating the positional embedding. The results

Table 4.4: Ablations experiment results on the different types of kernel size in the *video-to-tokens* module

Video-to -tokens				EK	D-48
Conv	Max-pool	Batch norm	channels	Top-1 Acc(%)	Top-1 Acc(%)
[7 7 7], stride 4	×	×	64	41.9	75.6
[5 5 5], stride 4	×	×	64	41.6	75.2
[7 7 7], stride 2	[3 3 3], stride 2	×	32	43.6	76.2
[7 7 7], stride 2	[3 3 3], stride 2	✓	32	45.9	78.2

Table 4.5: Ablations study performance on convolutional mapping and position-wise linear mapping.

Method	Conv. Mapping	Linear Mapping	EK (Acc %)	D-48 (Acc %)
CTN-a		✓	44.1	77.3
CTN-b	✓		45.9	78.2

further indicate the effectiveness of our incorporated convolutions. Furthermore, the positional embedding can be removed entirely in our proposed CTN model. This shows that our proposed CTN can provide the possibility of simplifying adaption to action recognition based on videos without a re-designing of the embeddings.

Different types of video-to-tokens modules. The influencing cues in the video-to-tokens module contain the kernel size of 3D convolutions, the stride of 3D convolutions, the existence of the batch normalization layer, and the max-pooling layer. ViviT extends the patch-based partitioning of images from ViT to handle video inputs. Specifically, it extracts N non-overlapping, spatio-temporal tubelets of size 16×16×2 from the input video. This technique can be regarded as an extension of the ViT embedding technique into the 3D domain, where it is utilized as a form of 3D convolution. The convolution operates with a kernel size of t × h × w and a stride of (t, h, w) in the time, height, and width dimensions, respectively.

However, directly tokenizing input videos with large tubelets can have two limitations. Firstly, it may be difficult to capture low-level information in videos such as edges, backgrounds, and corners. Secondly, large kernels may be over-parameterized and difficult to optimize, since they may require a substantial amount of training samples or training iterations. Additionally, since the temporal information in videos may not be fully considered, accurate action recognition may not be achieved. The results are shown in Table 4.4. Under the non-max-pooling circumstance, the 3D convolution layer with a kernel size of 7×7×7 and 5×5×5 (stride = 4) decreases the accuracy

Table 4.6: Ablation study performance on the different types of convolutional mapping.

Convolutional mapping-type		EK	D-48
Kernel size	Batch Norm		
3×3	×	43.1	75.8
3×3	✓	45.9	78.2
5×5	×	43.6	76.1
5×5	✓	45.6	77.7

results. On the other hand, batch normalization and the max-pooling layers can benefit the training results. Therefore, we adopt the best arrangement in our proposed CTN model, which is indicated in the last row of Table 4.4.

Convolutional mapping. First, we investigate how the proposed depth-wise convolutional mapping affects the performance by deciding whether to apply convolutional mapping or vanilla linear mapping for the Transformer blocks. In Table 2.5, CTN-a refers to the linear mapping with a fully connected layer and CTN-b denotes the convolutional mapping with a depth-wise convolutional layer. Table 2.5 shows that using convolutional mapping instead of linear mapping improves the Top-1 accuracy on the Epic-Kitchens dataset from 43.1% to 45.9% and improves the Top-1 accuracy on the Diving-48 dataset from 75.8% to 78.2%. Compared to sample linear mapping, the proposed convolutional mapping can extract local spatial features in video frames because of the locality of convolutions.

Then, we show the proposed convolutional mapping using different kernel sizes and strides in Table 4.6. Specifically, we adopt kernel sizes of 3×3 and 5×5 in Table 4.6. When increasing the size of the convolution kernel, each token can accumulate with the neighboring tokens by the non-linear mapping. The kernel sizes of 3×3 and 5×5 both achieve the same good performance. The mapping receives a better performance when the batch normalization layer is applied. We choose the kernel size 3×3 as the final experimental setting because of the trade-off between the number of FLOPs and accuracy. The batch normalization layer can significantly improve the performance.

4.4.4 Comparison to the state-of-the-art

Table 4.7 shows the comparison of our proposed CTN with state-of-the-art models, including models constructed on pure-Transformer [17] [23] and models built on CNNs [22] [54] [53] [52] on the Epic-Kitchens dataset and Diving 48 dataset. Compared

Table 4.7: Comparisons to state-of-the-art models across the Epic-Kitchen and Diving 48 dataset.

Models		Pretrained dataset	Modality	Frames in training	Params	EK	D-48
CNNs based	I3D [22]	Kinetics 400	RGB	8	28	N/A	48.3
	TSN [52]	ImageNet	RGB + flow	3	-	33.2	52.5
	TRNms [53]	ImageNet	RGB + flow	3	-	35.3	54.4
	TSM [54]	ImageNet	RGB + flow	3	24.3	38.3	51.1
	GSM [55]	ImageNet	RGB + flow	8	-	33.45	40.27
	SlowFast [56]	Kinetics 400	RGB	8	34.6	38.5	77.6
	GSF [57]	ImageNet	RGB	16	-	44.48	75.31
Transformer based	X-ViT [58]	ImageNet	RGB	8	-	41.5	N/A
	Timesformer [23]	ImageNet	RGB	8	121.4	42.5	74.9
	ViviT-B [17]	ImageNet	RGB	16	88.9	43.1	77.1
	Mformer [59]	Kinetics 400	RGB	16	109.1	44.5	-
CNNs+ Transformer	**CTN**	ImageNet	RGB	8	88.6	**45.9**	**78.2**

to CNN-based models, the proposed CTN achieves better accuracy with fewer input modalities. The proposed CTN obtains 45.9% accuracy on the Epic-Kitchens dataset, which surpasses the TSN [52] model by 10.7%, surpasses TRN [53] by 8.6%, surpasses the TSM [54] model by 5.6%, surpasses SlowFast [56] by 5.4%, surpasses GSF [57] model 1.5%, surpasses the X-ViT model [58] by 2.4%, surpasses the TimeSformer model [23] by 1.4%, exceeds ViviT [17] by 0.8% and exceeds Mformer model [59] 1.4%. CTN obtains a 78.2% accuracy on the Diving 48 dataset, which outperforms I3D [22] by 29.9%, exceeds TSN [52] by 25.7%, TRN [53] by 23.8% and surpasses SlowFast [56] by 0.6%, surpasses GSF model [57] 2.9%, and beats TimeSformer [23] by 3.3%. Our proposed CTN model shows superior advantages compared to concurrent models built on CNNs and models constructed on pure-Transformer on both the Epic-Kitchens dataset and the Diving 48 dataset.

The proposed Convolutional Transformer Network (CTN) model takes advantage of CNNs in capturing spatial-temporal features and sharing weights, and the benefit of the Transformer in modeling long-range relationships. Therefore, the CTN model can capture discriminative spatial features and informative temporal features for action recognition.

4.5 Conclusion

In this work, we propose a Convolutional Transformer Network (CTN) model, which introduces convolutions into the video Transformer network to integrate the advantages of Transformer in learning long-range dependencies with the advantages of

CNNs in extracting low-level features for fine-grained action recognition. Extensive experiments indicate that the proposed video-to-tokens module and convolutional mapping make the proposed CTN model incomparable results while retaining computational efficiency. Additionally, due to the internal local context relationship brought by CNNs, the proposed CTN model no longer needs a positional embedding.

Our future work focuses on removing the dependence on image pretrained models while maintaining the remarkable performance of the model for fine-grained action recognition. In addition, we are going to extend fine-grained action recognition into more complex video-related tasks.

References

[1] Tianyu Liu, Yujun Ma, Wenhan Yang, Wanting Ji, Ruili Wang, and Ping Jiang. Spatial-temporal interaction learning based two-stream network for action recognition. *Information Sciences*, 606:864–876, 2022.

[2] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alex G Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2568–2577, 2015.

[3] Amin Ullah, Khan Muhammad, Tanveer Hussain, and Sung Wook Baik. Conflux lstms network: A novel approach for multi-view action recognition. *Neurocomputing*, 435:321–329, 2021.

[4] Mahshid Majd and Reza Safabakhsh. Correlational convolutional lstm for human action recognition. *Neurocomputing*, 396:224–229, 2020.

[5] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[6] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018.

[7] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[8] Mengyuan Liu, Hong Liu, and Chen Chen. Robust 3d action recognition through

sampling local appearances and global distributions. *IEEE Transactions on Multimedia*, 20(8):1932–1947, 2017.

[9] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 27, 2014.

[10] Jiaming Wang, Zhenfeng Shao, Xiao Huang, Tao Lu, Ruiqian Zhang, and Xianwei Lv. Spatial–temporal pooling for action recognition in videos. *Neurocomputing*, 451:265–278, 2021.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[12] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 579–588, 2021.

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[14] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[15] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. An image is worth 16x16 words, what is a video worth? *arXiv preprint arXiv:2103.13915*, 2021.

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

[17] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[18] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

[19] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel,

and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16519–16529, 2021.

[20] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 171–180. Springer, 2021.

[21] Zhouyong Liu, Shun Luo, Wubin Li, Jingben Lu, Yufan Wu, Shilei Sun, Chunguo Li, and Luxi Yang. Convtransformer: A convolutional transformer network for video frame synthesis. *arXiv preprint arXiv:2011.10185*, 2020.

[22] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[23] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[24] Will Price and Dima Damen. An evaluation of action recognition models on epic-kitchens. *arXiv preprint arXiv:1908.00867*, 2019.

[25] Earnest Paul Ijjina and Chalavadi Krishna Mohan. Hybrid deep neural network model for human action recognition. *Applied soft computing*, 46:936–952, 2016.

[26] Dong Li, Ting Yao, Ling-Yu Duan, Tao Mei, and Yong Rui. Unified spatio-temporal attention networks for action recognition in videos. *IEEE Transactions on Multimedia*, 21(2):416–428, 2018.

[27] Ming Zong, Ruili Wang, Xiubo Chen, Zhe Chen, and Yuanhao Gong. Motion saliency based multi-stream multiplier resnets for action recognition. *Image and Vision Computing*, 107:104108, 2021.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[29] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[30] Lijie Fan, Wenbing Huang, Chuang Gan, Stefano Ermon, Boqing Gong, and Junzhou Huang. End-to-end learning of motion representation for video un-

derstanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6016–6025, 2018.

[31] M Esat Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 731–747. Springer, 2020.

[32] Linchao Zhu, Hehe Fan, Yawei Luo, Mingliang Xu, and Yi Yang. Temporal cross-layer correlation mining for action recognition. *IEEE Transactions on Multimedia*, 24:668–676, 2021.

[33] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

[34] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 122–132, 2020.

[35] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021.

[36] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[37] Zhenbing Liu, Zeya Li, Ruili Wang, Ming Zong, and Wanting Ji. Spatiotemporal saliency-based multi-stream networks with attention-aware lstm for action recognition. *Neural Computing and Applications*, 32:14593–14602, 2020.

[38] Ming Zong, Ruili Wang, Yujun Ma, and Wanting Ji. Spatial and temporal saliency based four-stream network with multi-task learning for action recognition. *Applied Soft Computing*, 132:109884, 2023.

[39] Yisheng Zhu and Guangcan Liu. Fine-grained action recognition using multi-view attentions. *The Visual Computer*, 36(9):1771–1781, 2020.

[40] Fang Liu, Liang Zhao, Xiaochun Cheng, Qin Dai, Xiangbin Shi, and Jianzhong Qiao. Fine-grained action recognition by motion saliency and mid-level patches. *Applied Sciences*, 10(8):2811, 2020.

[41] Miao Ma, Naresh Marturi, Yibin Li, Ales Leonardis, and Rustam Stolkin.

Region-sequence based six-stream cnn features for general and fine-grained human action recognition in videos. *Pattern Recognition*, 76:506–521, 2018.

[42] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.

[43] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[44] Qiang Chen, Qiman Wu, Jian Wang, Qinghao Hu, Tao Hu, Errui Ding, Jian Cheng, and Jingdong Wang. Mixformer: Mixing features across windows and dimensions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5249–5259, 2022.

[45] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13505–13515, 2021.

[46] Xinyu Li, Yanyi Zhang, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. *arXiv e-prints*, pages arXiv–2104, 2021.

[47] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3163–3172, 2021.

[48] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3333–3343, 2022.

[49] Xiang Long, Gerard De Melo, Dongliang He, Fu Li, Zhizhen Chi, Shilei Wen, and Chuang Gan. Purely attention based local feature integration for video classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2140–2154, 2020.

[50] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[51] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[52] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[53] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.

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

[55] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift networks for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1102–1111, 2020.

[56] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019.

[57] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift-fuse for video action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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

[59] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in Neural Information Processing Systems*, 34:12493–12506, 2021.

Chapter 5

Summary

This chapter offers concluding remarks on the thesis. Throughout our research, we have contributed to multiple subtasks of action recognition. In this final chapter, we will review the proposed methods in Section 5.1 and provide insights into future research directions in Section 5.2.

5.1 Research Summary

In this thesis, we have investigated multiple sub-tasks of action recognition, including fine-grained action recognition, coarse-grained action recognition, and RGB-D action recognition. A summary of our methods and contributions is provided below:

Chapter 2 presented a multi-stage factorized spatio-temporal learning model for RGB-D action and gesture recognition that aims to individually capture spatial and temporal features in each stage while also modeling their dependencies without additional recoupling operations. Additionally, we introduce a lightweight CDC Stem that can efficiently capture coarse- and fine-grained temporal patterns. At each factorized spatio-temporal learning stage, our CNN-Transformer hybrid block and weight-shared Transformer block effectively capture multi-scale spatial and temporal features. As a result, our model achieves new state-of-the-art performance on both RGB-D action and gesture recognition benchmarks, including NTU RGB-D and IsoGD. These results demonstrate the effectiveness of our proposed multi-stage factorized spatio-temporal representation.

Chapter 3 presented a relative-position embedding-based spatially and temporally decoupled Transformer (RPE-STDT) for action recognition, which can capture spatial and temporal information by stacked self-attention layers. Extensive experiments indicate that the proposed RPE-STDT achieves state-of-the-art or comparable performance on three public action recognition datasets while retaining computational efficiency. Furthermore, the proposed RPE-STDT model enables long-term video modeling due to temporal attention and subsampling strategy.

Chapter 4 presented a novel Convolutional Transformer Network (CTN) model, which introduces convolutions into the video Transformer network to integrate the advantages of Transformer in learning long-range dependencies with the advantages of CNNs in extracting low level features for fine-grained action recognition. Extensive experiments indicate that the proposed video-to-tokens module and convolutional mapping make the proposed CTN model incomparable results while retaining computational efficiency.

To sum up, in this thesis, we present three innovative deep-learning approaches for action recognition. By building upon established deep models, such as I3D [1] and Transformer [2], our research emphasizes enhancements in spatiotemporal learning, multimodal representation, and positional embedding techniques. Our approaches have improved accuracy across six public action recognition datasets significantly. Our works have been published in top-tier conferences and journals (*e.g.*, *ACM Multimedia, Pattern Recognition*, and *Neurocomputing*), marking a significant contribution to action recognition.

5.2 Future work and directions

In this section, we discuss two certain directions of potential research: self-supervised learning, as detailed in Section 5.1, and multimodal fusion and alignment, discussed in Section 5.2.

5.2.1 Self-supervised learning

In our research, we employed supervised learning for action recognition [3] [4]. However, as the reliance on annotated datasets poses challenges, we will explore self-supervised learning methodologies [5] [6]. Leveraging techniques like Contrastive

Language-Image Pre-training (CLIP) [7] can enable the extraction of meaningful representations from unlabeled data. By leveraging temporal and spatial structures within videos, self-supervised approaches enhance the robustness of action recognition systems [8] [5]. Integration of contrastive learning and predictive modeling facilitates the discovery of discriminative features, paving the way for more efficient and scalable action recognition frameworks in diverse real-world scenarios.

5.2.2 Multimodal fusion and alignment

In our research, we focused on spatial and temporal feature extraction for action recognition rather than feature fusion [9] [10]. However, we will explore enhancing action recognition performance by improving multimodal feature fusion methods and incorporating alignment techniques [8] [11] [12]. Investigating novel approaches to fuse features from diverse modalities [13], such as RGB, depth, and optical flow, while also addressing alignment challenges, can lead to more robust and accurate action recognition [14].

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[3] Yujun Ma, Benjia Zhou, Ruili Wang, and Pichao Wang. Multi-stage factorized spatio-temporal representation for rgb-d action and gesture recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3149–3160, 2023.

[4] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, and Fan Wang. A unified multimodal de-and re-coupling framework for rgb-d motion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[5] Jinghao Zhou, Li Dong, Zhe Gan, Lijuan Wang, and Furu Wei. Non-contrastive learning meets language-image pre-training. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 11028–11038, 2023.

[6] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Xiang Wang, Yuehuan Wang, Yiliang Lv, Changxin Gao, and Nong Sang. Mar: Masked autoencoders for efficient action recognition. *IEEE Transactions on Multimedia*, 2023.

[7] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

[8] Yizhou Zhao, Zhenyang Li, Xun Guo, and Yan Lu. Alignment-guided temporal attention for video action recognition. *Advances in Neural Information Processing Systems*, 35:13627–13639, 2022.

[9] Yujun Ma and Ruili Wang. Relative-position embedding based spatially and temporally decoupled transformer for action recognition. *Pattern Recognition*, 145:109905, 2024.

[10] Tianyu Liu, Yujun Ma, Wenhan Yang, Wanting Ji, Ruili Wang, and Ping Jiang. Spatial-temporal interaction learning based two-stream network for action recognition. *Information Sciences*, 606:864–876, 2022.

[11] Muhammad Waleed Gondal, Jochen Gast, Inigo Alonso Ruiz, Richard Droste, Tommaso Macri, Suren Kumar, and Luitpold Staudigl. Domain aligned clip for few-shot classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5721–5730, 2024.

[12] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan 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*, pages 17949–17958, 2022.

[13] Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. Star-transformer: a spatio-temporal cross attention transformer for human action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3330–3339, 2023.

[14] Xiang Wang, Shiwei Zhang, Jun Cen, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Clip-guided prototype modulating for few-shot action recognition. *arXiv preprint arXiv:2303.02982*, 2023.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Yujun Ma
Name/title of Primary Supervisor:	Professor Ruili Wang
In which chapter is the manuscript /published work:	Chapter 4

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Yujun Ma, Ruili Wang, Ming Zong, Wanting Ji, Yi Wang, and Baoliu Ye. "Convolutional transformer network for fine-grained action recognition." Neurocomputing 569 (2024): 127027.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Yujun Ma Digitally signed by Yujun Ma DN: cn=Yujun Ma, c=NZ, o=Massey University, ou=School of Mathematical and Computational Science, email=yma1@massey.ac.nz Date: 2024.02.22 14:23:25 +13'00'
Date:	20-Feb-2024
Primary Supervisor's Signature:	Prof Ruili Wang Digitally signed by Prof Ruili Wang DN: cn=Prof Ruili Wang, c=NZ, email=ruili.wang@massey.ac.nz Location: School of Natural and Computational Sciences Date: 2024.07.23 16:00:32 +12'00'
Date:	

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.